Universiteit Leiden

Master Computer Science

A Deeper Dive into the Relations Between Physical Activity and Cardiovascular Disease Using Subgroup Set Discovery and a Smartphone-Based Dataset

Name: Kamand Hajiaghapour
Student ID: S3058107
Date: 26/01/2023

Specialisation: Data Science

1st supervisor: Matthijs van Leeuwen
2nd supervisor: Tobias Bonten

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Leiden University

Abstract

Physical inactivity is considered one of the risk factors of cardiovascular disease(CVD). However, the relation between CVD and physical activity is not simple enough to put in one sentence. The purpose of this study is to investigate potential relations between these two factors in more detail. In general we want to answer these questions: what is the best time during the day for physical activity if the goal is to reduce the risk of getting a CVD? What about the intensity of the activity? What about different types of physical activity? Are there also lower/upper bounds for the duration or intensity of the exercise? How will the answers change based on the gender, age, ethnicity, height, and weight of the participants? To this end we apply the SSD++ algorithm, a subgroup discovery approach, on a smartphone-based dataset encompassing physical activity, demographic, and CVD-related attributes of 12,043 participants with 91 attributes overall. This results in 15 patterns describing parts of the data that deviate from the rest based on having CVD. For evaluating our model we implement it for prediction on a test set and compare our result with Random Forest, Naïve Bayes, and Decision Tree classifiers. In addition, we compare the rules we get with state-of-the-art studies. Eventually, this comparison points out that the results we got are reliable and that the patterns recognized further are worth deeper examination.

Key words: Cardiovascular Disease, Physical Activity, Subgroup Set Discovery, Smartphone-Based Data

Contents

Figures

Tables

Appendix Contents

Chapter 1

Introduction

In the entire world, cardiovascular diseases (CVDs) are the primary cause of death [1]. In 2019, the cause of 32% of death (17.9 million) globally was CVDs [1]. Even in countries like the Netherlands, where CVDs are not the leading cause of death anymore, the number of hospitalizations because of CVDs is still a burden [2]. Physical inactivity is one of the risk factors for getting CVDs. Previous studies depicted an adverse relation between getting CVDs and physical activity [3-6]. Moreover, many studies demonstrated that intensity, frequency, duration and other attributes of physical activity can have various effects on different groups of people. For example, authors in [7] indicated a positive relation between evening physical activity and CVDs. In [8] it is explained that afternoon exercise can be more beneficial to diabetic people in comparison to morning exercise. These all show that the connection between CVDs and physical activity is more complicated than it may seem at first glance and requires more investigation.

In this study, we aim to dive deeper into these relations. More specifically, we want to answer multiple questions, such as what is the best time during the day for physical activity in relation to the risk of CVD? What about the intensity of the activity? What about different types of physical activity? Are there also lower/upper bounds for the duration or intensity of the exercise? How do the answers change based on the gender, age, ethnicity, height, and weight of persons? As we mentioned earlier each of these questions was the topic of various studies. Some studies tried to answer these questions from the metabolic and molecular levels [9, 10]. There is also a different approach to this problem through statistical analysis of a moderately small controlled group [11-13]. In [12], authors utilized the clustering method to a relatively big (86,657 participants) dataset to find the most proper timing for physical activity for multiple groups of participants based on age and sex.

In most of these studies, the control group only include people with CVDs or its risk factors without containing healthy people. Another significant characteristic of the mentioned studies is that the number of features is usually less than 50. Therefore, classical statistical approaches work properly in this regard. In addition, none of these studies can separately answer all our questions at the same time. It means the applied approaches do not have the potential to consider all physical activity attributes together and recognize interesting relations among them and having cardiovascular diseases or their risk factors.

The focus of the studies we mentioned so far were on a control group of people with specific characteristics usually due to difficulty of data collection or inaccessibility of the data. The widespread availability of smartphones allows collection of physical activity data in various aspects on a larger and more diverse population. One example of such datasets is My Heart Counts Cardiovascular Health Study [14] dataset which we use in this study to answer the questions at hand. This smartphone-based dataset was collected from March 10 to October 28, 2015, in the United States. Collected using the iPhone application MyHeart Counts iOS, it is based on the data of participants who consented to use their data in research. Some parts of this dataset are sensor data recorded by the iPhone, such as movement and sleep data. Some other parts were collected using questionnaires inside the application, such as questionnaires about well-being, risk factors of cardiovascular disease, activity and sleep, diet, and physical activity readiness. One related study to both our topic and this dataset is [15]. The focus of this study was on assessing whether it is possible to discover fitness measures based on a smartphone dataset in addition to discovering patterns between physical activity and life satisfaction and self-reported diseases. The method applied here was clustering data based on physical activity patterns and then comparing them using chi-square. One interesting finding of this study is that there were relations between changing state from stationary to active or vice versa with self-reported diseases. In this study again, the patterns were compared, separately and the decision for investigating specific patterns was based on the authors' intuition and state-of-the-art studies.

We go through all available tables in this dataset and by extensive preprocessing to clean the data and extract demographic, physical activity, and CVD attributes. The difference between this dataset and the ones used in mentioned studies is here, our final dataset includes 12,043 participants with 91 attributes. Now the challenge will be to choose an approach to answer our questions. To this regard, we will use the subgroup discovery approach.

Subgroup discovery [16] gives us the opportunity to respond to all our questions at once. This technique finds interesting subgroups in the dataset that deviates from the rest based on the specified target variable(s) [17]. Utilizing the Subgroup Discovery technique, we can obtain clusters of data with unique behavior from the rest based on the possibility of having CVDs or its risk factors.

These clusters are called subgroups and described using conditions in Boolean format on the defined attributes.

Another difference of this approach in comparison to mentioned studies is that there is no need to choose specific attributes to compare with each other one by one. This creates a chance to discover unexpected relations or recognize perspectives not being noticed so far instead of just examining the state-of-the-art hypotheses.

Subgroup discovery is the task of finding interesting subsets of the dataset that deviate from the rest given one or more target variables based on a local measure [18-20]. The task of aggregating these subgroups in a set that describes all the deviations in the target variable(s) distribution(s) is called subgroup set discovery. This final set that makes the global model of the dataset can either comprise of a sequential list of these subgroups (subgroup list) or an unordered set of them(subgroup set).

In this study, we implement the SSD++ algorithm [18] to obtain a global model of the dataset based on subgroup set discovery. The output of this algorithm is a subgroup list comprises of subgroups ordered from most to least relevant. This state-of-the-art approach introduced in 2021 showed better performance regarding statistical robustness and subgroups redundancy in comparison to top-k subgroup discovery [21], seq-cover, CN2-SD [22], Diverse Subgroup Set Discovery (DSSD)[23], Mont Carlo Tree Search for Data Mining (MCTS4DM) [24], and FSSD [25].

The first step in our study is to preprocess the dataset extensively to extract as many relevant features as possible. In this stage, we also deal with noisy data, duplicates and invalid values. In addition, we transform the data into a format applicable to our algorithm. After that, we implement the SSD++ algorithm on our dataset which leads to the discovery of a list of 15 patterns in the dataset. We evaluate our model both locally and globally. By local evaluation, we mean looking at each of the rules individually and comparing them based on Coverage, Probability and Weighted Kullback-Leibler (WKL) measures. We also compare these rules with state-of-the-art studies in this domain.

At the global level, we look at some subgroup discovery measures: the Sum of Weighted Kullback-Leibler and Length Ratio, Number of Subgroups, Average Number of Conditions, and Jaccard Index. In addition, to examine the complication of the problem and power of our model in dealing with that, we try our SSD++ algorithm for predicting the labels of a test set and compare the results with Random Forest, Decision Tree and Naïve Byes algorithms based on Accuracy and F1-Score. This study is the first study using subgroup discovery in this domain and on the MHC dataset.

In the following chapter we will review the history of subgroup discovery and its application in the healthcare domain. In Chapter 3 we will describe our dataset and the preprocessing steps we take. Chapter 4 explains our methodology and is a careful study of the SSD++ algorithm [18]. Chapter 5 describes our results and discussion. Finally, Chapters 6 is the conclusion.

Chapter 2

Background

In this chapter, we describe the related studies to our work both from medical and methodology perspectives. Section 2.1 elaborates more on studies related to physical activity and its relation to Cardiovascular Disease. Section 2.2 focus on the approach we use in this study meaning subgroup discovery, its definition, evolution during time and applications.

2.1. Physical Activity and Cardiovascular Diseases

In the 21st century we, as humans, look for evidence, facts and science for answering questions in every decision in every aspect of our life. From general ones to detailed questions regarding every routine we have on a daily basis. We started suspecting how we can optimize our behavior toward a healthier lifestyle. One of these detailed questions has been about physical activity and its effect on our health condition. There are many studies showing there is a positive relation between being physically active and living a healthy life both physically and mentally [26-29] The focus of these studies is on different aspects and health conditions. One that got a lot of attention is cardiovascular diseases and heart conditions [3-6].

The majority of the studies for detecting a relation between physical activity and cardiovascular diseases focus on a certain control group or/and specific pattern in doing physical activity. This is because all these factors can cause diverse effects on the results. For example, the focus of the study can be on the elderly population such as in [3, 30] or children [31], or it can be related to participants with certain conditions for example [32-33]. It can also focus on participants' gender [34] or ethnicity [35], or a combination of different factors. In addition, it can exclusively focus on a specific pattern of physical activity. For instance, the focus can be on leisure-time physical activity [36] or on the intensity of the activity [37]. Some studies also focus on the timing of the physical activity, meaning at what time of the day it takes place [9-13], [38].

Another aspect of looking at the related studies is their approach towards resolving the question. Studies such as [9,11] see the problem through the molecular and metabolic view. In [9] intra- and inter- tissue metabolite responses are mapped and compared after exercise at different times of the day on mice. Authors in [11] compared the metabolic effect of exercise during the morning and afternoon in a controlled group of 32 male adults at risk of or diagnosed with diabetes.

Some other studies applied statistical analysis to determine a relation between exercise attributes and cardiovascular diseases or its risk factors [12], [15], [39]. Authors in [12] made distinct clusters based on the timing of physical activity and then used multivariable-adjusted Cox-proportional hazard models to compare different clusters based on sex and age using Hazard ratios. In [39], two separate cohorts of 26 men and 36 women were analyzed separately based on a comparison of their pre and post-training muscular strength, endurance, power, body composition, systolic/diastolic blood pressure, respiratory exchange ratio, profile of mood states, and dietary intake.

Since the exploration of the aforementioned attributes individually led to interesting associations, it is well worth exploring a larger group of variables at once. This can give us the chance to have a holistic view at all the possible relations or potentially discover new patterns that might deviate from our expectation. To this regard, we can use smartphone-based datasets. The widespread availability of smartphones allows researchers to reach a larger and more diverse population, enabling better follow up. Studies such as [14] and [15] are related to the applicability of this type of data in relation to finding a healthier lifestyle. However, these methods are fairly new and need further exploration in order to be able to utilize the potential of all these data.

2.2. Subgroup Discovery

Subgroup discovery is an exploratory data analysis approach for finding interesting relations among features in a dataset and one or multiple target variables [40]. If we consider supervised and unsupervised learning in machine learning as two edges of a spectrum, subgroup discovery is in the middle of this spectrum. It is close to clustering in unsupervised learning since the purpose is not prediction but to divide the data space to multiple subgroups. It is also close to classification in supervised learning because we have some target variables that we want to find subgroups that deviates from the rest of the data based on them.

Therefore, in subgroup discovery the input is a dataset with explanatory and one or more target variables. The output, based on the subgroup discovery technique, can be a list of patterns that together define the interesting subgroups, the number of cases these patterns are true in and the probability of each. As an example based on a simplified version of our problem, assume we have a dataset with average duration of vigorous physical activity per week(in minutes), age, average duration of physical activity during noon (in minutes) and a binary variable indicating whether a person has CVD or not, for 12043 participants. Consider we want to discover interesting patterns in this dataset regarding having CVD(our target variable). If we run a subgroup discovery algorithm on this dataset an imaginary result can be a table like Table 1.

Table 1: Simple Example of Subgroup Discovery Output

Subgroup	Rules	Probability	Usage
1	age >= 48 AND vigorous activity < 30	Having CVD: 0.56 Not Having CVD:0.44	154
2	age >= 40 AND noon duration >= 31	Having CVD: 0.49 Not Having CVD:0.51	309

In Table 1, the Rules column is related to the patterns found in the dataset and probability shows the probability of each category of data. Usage means the number of cases in each of the subgroups. Rule 1 indicates if a 48 year-old or older participant has 30 minutes or less vigorous physical activity during week the probability of having CVD for him/her is 56 percent. This pattern is seen in 154 participants. It is worth mentioning that in the whole dataset the probability of having CVD is 25%. Therefore, this subgroup behavior deviates from the whole dataset by having 31% higher probability of having CVD. In the case of having a continuous variable, instead of probability of each class we will have the mean and standard deviation for the target variable distribution based on each discovered pattern.

In our simple example we only had one target variable which was binary, however in general we have four categories of problems based on type and number of target variable(s): 1) single-nominal; 2) single numeric; 3) multi-nominal; and 4) multi-numeric [18].

Each subgroup contains two parts. One is the description of the subgroup(in our example the rule section of Table 1). This part consists of some conditions on our explanatory variable set that together form a Boolean function. The second part of a subgroup is the cover. Cover is a set of all the instances that the subgroup description is true in relation to them [18]. For example, subgroup 1 in our example covers 50 instances.

In our example, the output of the algorithm is a list of subgroups that are ranked based on a specific criterion(we will discuss this criterion in the next chapter). The global model here is a ranked list of subgroups found. We call this output a subgroup list. In subgroup list, the rules are aggregated in an if-then-else format, meaning the former rule can only be true if previous ones are not true. Another approach is to have a set of subgroups(subgroup set), without ranking them and letting them have mutual instances. In our study we use the first approach. This approach has its advantages and disadvantages in comparison to the subgroup set strategy. The advantage is that the subgroups are listed and ranked, so it is more straightforward to make a comparison and have an impression of the model with a quick glance. However, these ranked subgroups can only be interpreted as an else-if rule meaning each can be true if the formers in the list are false. This makes the analysis of the results sophisticated, especially regarding the last subgroups in the list.

2.2.1. Subgroup Discovery Applications and Algorithms

Subgroup discovery has applications in different domains including but not limited to fraud detection [41], flight delay identification [40], bioinformatics [42, 43] and marketing [44] . It also has been used in the healthcare domain in studies such as [45], [46-49]. In [50] the writers used the SD algorithm to find interesting rules concerning brain ischemia. In [48] subgroup discovery was used to find the patterns regarding surviving breast cancer in the short-term and long-term using Rule Induction Algorithm for Subgroup discovery (RIAS). In [45] the multi-objective evolutionary algorithm MESDIF is used to find the subgroups of patients based on their arrival time to the psychiatric emergency department. All these studies confirm that subgroup discovery can be a helpful approach to answering medical questions and it can lead to relevant results that are not achievable either through classification or using clustering.

In general, subgroup discovery consists of three main steps. The first step is the exploration for finding interesting candidates, typically using exhaustive search, beam search, or some other form of heuristic search. In the second step, the algorithm prunes the chosen candidates to only preserve the most relevant ones. The main pruning strategies are minimum support or coverage pruning, optimistic estimate pruning, and constraint pruning. Ultimately, the candidates are ranked based on a quality measure [51]. Choosing the quality measure is also a critical step and it depends on the problem at hand. Quality measures can be classified into four groups based on their objective meaning complexity, generality, precision, and interest [52]. The most popular quality measures are the number of rules, coverage [22], weighted relative accuracy (WRAcc) [53] and Weighted Kullback-Leibler divergence(WKL) [54].

Subgroup discovery was first introduced by Kloesgen [16], Wrobel [55] and Siebes [56] as Data Surveying. Since then many different algorithms were introduced based on a variety of strategies for searching, pruning and ranking subgroups [51]. These algorithms can be put into three groups:

algorithms that are an extension of classification, algorithms that are an extension of association rules, and lastly algorithms that are based on evolutionary algorithms [52].

EXPLORA [16] and MIDOS [55] were the first algorithms developed in this domain. These algorithms are extensions of classification tree search algorithms and they can use both exhaustive and heuristic exploration approaches. Other classification-based algorithms that are extensions of these two algorithms are SD [57] and CN2-SD [22]. As we mentioned before, some subgroup discovery algorithms are an extension of association rule algorithms [58]. In the sense that in association rule mining, the purpose is to find the interesting relations in the dataset. However, these relations are not based on the target variable/variables as it is in subgroup discovery. The most famous algorithms in this category are Apriori-SD [58] and SD-Map [59]. Finally, there are some subgroup discovery algorithms that use evolutionary algorithms for exploration [60]. In other words, in these algorithms, an evolutionary algorithm is implemented to generate new candidates. The most famous algorithms of these groups are SDIGA(Subgroup Discovery Iterative Genetic Algorithm) [61] and MESDIF (Multiobjective Evolutionary Subgroup Discovery Fuzzy rules) [45].

2.2.2. Subgroup Discovery Algorithms Limitations

Overall, since the introduction of the subgroup discovery approach in 1995 [16, 55, 56] many different algorithms were developed based on that, each of which tried to improve part of the process or to extend the approach applicability in more domains. However, there have always been three issues that remained unsolved [18]. The first one is related to using exhaustive search to find the candidates to be considered in the subgroup. Even though exhaustive search can result in the global optimum solution, it can also be computationally expensive and inefficient [23, 24]. The second issue is related to the redundancy of the extracted subgroups [23], meaning it is possible that the candidate sets actually cover mutual parts of the dataset. The third issue concerns the reliability of the subgroups and the lack of generalization [62]. This issue is about how we can guarantee that the model built is robust and the combination of the subgroups can reliably describe the dataset at a global level.

In this study, we will implement the SSD++ algorithm [18]. SSD++ is a heuristic algorithm which means it will not give us the one and only best solution (global optimum) but one possible good option. In Chapter 3, we will describe this algorithm in more detail. The SSD++ algorithm addresses two mentioned limitations. It solves the problem of robustness by using the Minimum Description Length (MDL) principle [63] and applying the WKL quality measure, which together guarantee that the algorithm will result in an improvement in each iteration and gets closer to a model that has both good quality and simplicity in global and local level. Concerning redundancy, since the result of this algorithm is a subgroup list, this issue is not a concern anymore. Based on [18], this algorithm is successful in dealing with these issues and can get better results in comparison to the state-of-the-art algorithms for subgroup discovery. This is why we chose to implement it in this study.

Chapter 3

Methodology

In this study, we will implement subgroup discovery to find interesting relations between physical activity attributes like duration of physical activity, the timing of the physical activity, hardness of it and so on, and the possibility of having cardiovascular disease or its risk factors. Subgroup discovery represents an exploratory statistical approach that finds interesting subsets of the data based on one or more target variables. The interestingness of the subset defines by some statistical quality measures that usually indicate how different it is from the whole dataset [40]. We describe various quality measures in Section 3.4.

In general, the input of the subgroup discovery is a dataset with some features, in our case physical activity attributes, and one or more target variables, in our case having or not having cardiovascular disease or its risk factors. Both feature and target variables can be numeric or nominal [18], [20]. However, not all subgroup discovery algorithms can be applied to both types of data [18]. One of the strengths of the algorithm we implement here, SSD++, is its applicability to all types of data. In Section 3.2 we will see more reasons for choosing this algorithm and in Section 3.3 we go through the algorithm steps and see how it actually works. The output of subgroup discovery algorithms is formed by rules that describe unusual subsets. In Section 3.1 we present the notation of these rules.

3.1. Notation

Our dataset $D = (X,Y) = \{(x^1,y^1),(x^2,y^2),\ldots,(x^n,y^n)\}$ consists of $n=12043$ rows. Each instance of the dataset (x,y) encompasses the information regarding one particular participant. This information includes a vector of explanatory variables(x), in addition to, in our case, one binary target variable(y). The number of instances is described using superscriptions. x represents a vector of the explanatory attributes as follow: $x = (x_1,x_2,\ldots,x_i)$, where $i=0,1\ldots,89$. The types of these variables are numeric or categorical and they include 90 physical activity and demographic features of the participants. Target variable y is a Boolean value indicating whether the participant has a cardiovascular disease or its risk factors. In Chapter 4 we describe the variables in more detail.

The outcome of the SSD++ algorithm is a list of subgroups. As we mentioned in Chapter 2, each subgroup is composed of two parts: description(pattern) and cover. Descriptions, which are in the form of Boolean functions, are followed by the probability of the target variable, being, in our case, true or false. In the case of having a numeric target variable, the pattern is followed by the distribution of the target variable. As an example, we can consider description S as S = { age >= 40 AND noon duration >= 31}. Therefore, a description S is a query, formed by a conjunction of intervals or values of variables. Now, we can have the following notation for S:

$$D_S = \{(X,Y) \in D \mid S(X) = true\}, \qquad (1)$$

where D_S means pattern S over the dataset D, S(X) indicates whether the conditions of the pattern S are satisfied by tuple X.

Each pattern links a query of explanatory variables to a probability of the target variable. For each pattern, the empirical probability of our binary target variable over subgroup D_S is shown as $\widehat{p}_s(y)$.

$$S \rightarrow \widehat{p}_s(y) \qquad (2)$$

Table 2 Shows all the notation implemented in this study.

Table 2: Variable Notations in This Study

Notation	Meaning
D	Dataset
S	Pattern
$\mathbf{D_S}$	Pattern S over the dataset D
$\widehat{p}_S(y)$	Empirical probability of the target variable y over subgroup D_s
X	Vector of all explanatory variables
y	Target variable
$\mathbf{x_i}$	Explanatory variable i
$\widehat{\theta}^i$	The maximum likelihood estimation of the probability distribution parameters over y
M	Model, i.e., a rule list
L(D,M)	The length of the encoded model M for dataset D

3.2. SSD++ Algorithm

The purpose of our study is to find a set of subgroups that jointly form a global model of the dataset; each describes an interesting part of the dataset based on our target variable. This process is addressed as subgroup set discovery. This set aims to show all fundamental deviations in the target distribution [18]. Therefore, we want to transform the local models that we determine into a global model. Based on LeGo(from Local Patterns to Global Models) [64] to achieve this goal, we need to undertake three steps: 1) find local subgroup candidates; 2) make a set of the candidates we found in step 1 that is solid and encompasses as much information as possible; and 3) make a global model from the candidates chose in step 2 [18].

There are three approaches for combining the candidates for making the final global model: top-k ranking, subgroup list, and subgroup set discovery. The SSD++ algorithm uses the subgroup list paradigm. The format of the results in subgroup lists is as follows, where $\widehat{\theta}^i$ is the maximum likelihood estimation of the probability distribution (Dist) parameters over y [18]:

$$
\begin{array}{llllll}
S_1: & \text{IF} & a_1 \subset x & \text{THEN} & y \sim \text{Dist}(\widehat{\theta}^1) & \\
 & & \vdots & & & \\
S_w: & \text{ELSE IF} & a_w \subset x & \text{THEN} & y_1 \sim \text{Dist}(\widehat{\theta}^w) & \\
\text{Dataset:} & \text{ELSE} & & & y_1 \sim \text{Dist}(\widehat{\theta}^d) &
\end{array}
$$

The SSD++ algorithm includes two steps that iteratively repeat. In each iteration, it generates a new candidate subgroup using beam search and adds this new candidate to the list using the Separate-and-Conquer (SaC) [65] strategy [20]. For choosing the best subgroup among others in each iteration compression gain based on MDL is being used. Algorithm 1 [18] shows the pseudocode of the SSD++ algorithm.

Algorithm 1: SSD++ algorithm [18]

Input: Dataset D, number of cut points n_{cut}, beam width w_b, depth max. d_{max} and normalisation β
Output: Subgroup list S
$M \leftarrow [\theta_d(Y)]$;
$subgroup \leftarrow BeamSearch(M,D,w_b,n_{cut},d_{max})$;
while $\Delta_b L(D, M \oplus subgroup) > 0$ **do**
    $subgroup \leftarrow BeamSearch\ (M,D,w_b,n_{cut},d_{max})$;
    $M \leftarrow M \oplus subgroup$;
end
return $S \in M$

In this algorithm number of cut points (n_{cut}) is used for discretizing numeric attributes for generating new conditions. Beam width (w_b) and depth max. (d_{max}) are beam search parameters that we discuss in more detail in Section 3.1.2. β (also called alpha gain) is a parameter of compression gain criterion explained in Section 3.1.3.

In the following sections, we initially see what the separate-and-conquer strategy is, then we explain beam search and at the end, we describe compression gain.

3.1.1. Separate-and-Conquer (SaC)

Separate-and-conquer [65] is a rule-learning strategy that can generate if-then rules sequentially. It starts by adding the best local rule to the set or list of rules and then removing or re-weighting parts of the data set related to the rule added to the list. These two steps repeat one after another until there is no more data to cover in the dataset. This approach has been adopted in many subgroup discovery algorithms for adding new subgroups to a set or list of subgroups [22- 25]. SSD++ is one of these algorithms.

3.1.2. Beam Search

Beam search is a heuristic algorithm that has applications in various domains. In subgroup discovery, this approach can be pragmatic in generating new subgroups to be added to the list of subgroups. Its advantage in comparison to exhaustive search is being less computationally expensive. However, it does not guarantee to reach the global optimum solution.

Beam search has three different parameters: width(w), d(depth), and a quality measure. In subgroup discovery, the width of the beam search indicates the number of rules investigated in each iteration for selecting the best candidate to be added to the list. Depth is related to the number of conditions that can be attached to the query. In the SSD++ algorithm, the quality measure is compression gain, explained in Section 3.3.3.

In each iteration, the SSD++ algorithm generates w candidates by considering one condition. Subsequently, the generated candidates are refined by adding another condition to the query. This process repeats until d conditions have been added. Then based on the compression gain of the w generated candidates one of them is elected to append to the subgroup list in each iteration.

3.1.3. Compression Gain

As we mentioned in the previous section, one crucial ingredient of beam search is a quality measure that enables us to select the right candidate among w generated candidates. In the SSD++ algorithm, compression gain based on MDL is the used quality measure [18]. The formula for calculating this measure can be seen below:

$$s=\operatorname{argmax}_{s\in f}\Delta_{\beta}L(D,M\oplus s)=\operatorname{argmax}_{s\in f}\left[\frac{L(D,M)-L(D,M\oplus s)}{(n_s)^{\beta}}\right],\beta\in[0,1] \quad (3)$$

Here, $\mathrm{L(D,M)}$ represents the length of the encoded model M for the dataset D, $L(D, M \oplus s)$ is the length of the encoded model M after attaching subgroup s, $\mathrm{n_s}$ is the number of subgroups, f is a set of all subgroup candidates, and β is a hyperparameter to trade-off between having more subgroups each encompassing a small number of instances or having fewer subgroups that include more instances. Therefore, the purpose is to find the subgroup that can maximize the reduction in the encoded length of the model. It means we are seeking a model that can encode the dataset in the most compressed format. The idea behind this criterion stems from the Minimum Description Length (MDL) [63] principle in model selection, which considers the shortest model in describing the dataset as the best one.

In [18], it is proven that this criterion can guarantee the statistical robustness of the algorithm since it is equivalent to Bayesian testing. This way, the third mentioned issue in classical subgroup discovery algorithms is not the case in this approach.

3.3. Evaluation Measures

The evaluation of the final model in this study is two-folded. We both evaluate the model at the local (subgroup) and global levels. At the local level we want to be able to assess each subgroup in the model or compare it with its companions. At the global level we want to know how good our model is as a whole and how successful it is in describing the data.

3.3.1. Local Level Evaluation Measures

Regarding local level assessment we consider these measures: Coverage, Weighted Kullback-Leibler (WKL) [23] and Support. Our focus in this study is on WKL since it is suggested by the SSD++ algorithms' developers [18]. As we mentioned before, the SSD++ algorithm is based on the MDL principal, meaning that the purpose is to find the model that can encode the dataset in the shortest way possible. In [18] the authors proved that the MDL-optimal solution and discovering the subgroup that maximizes WKL are the same in practice.

Coverage: This measure indicates the number of instances in each subgroup. In other words, it is the count of the instances that the subgroup is based on. For each subgroup in the final model, we have one value for coverage.

Support: This measure indicates the number of potential instances in each subgroup. In other words, it is the count of the instances that follow the subgroup's description without considering former subgroups or by also considering the mutual incidents in former subgroups in the subgroup list.

Weighted Kullback-Leibler (WKL): For a univariate target variable, such as our problem, this measure define as:

$$WKL(\widehat{\Theta}^a;\widehat{\Theta}^d) = n_a KL(\widehat{\Theta}^a;\widehat{\Theta}^d) \qquad (4)$$

Where $KL(\widehat{\Theta}^a;\widehat{\Theta}^d)$ is the Kullback-Leibler divergence between the subgroup and dataset for target variable Y. $\widehat{\Theta}^a$ is the empirical target distribution of the subgroup pattern, $\widehat{\Theta}^d$ is the empirical target distribution of the dataset and n_a is the coverage of the subgroup. The formula for calculating the Kullback-Leibler divergence is:

$$KL(\widehat{\Theta}^a;\widehat{\Theta}^d) = \Sigma_{y \in Y^a} Pr(y|\widehat{\Theta}^a_j) log(\frac{Pr(y|\widehat{\Theta}^a_j)}{Pr(y|\widehat{\Theta}^d_j)}) \qquad (5)$$

Therefore, like every other subgroup discovery quality measure, WKL also includes a measure of coverage which shows in how many instances the pattern is found and a measure of distinction indicating how much the subgroup distribution is different from the whole dataset. For each subgroup in the final model, we have one WKL.

3.3.2. Global Level Evaluation Measures

For global level we implement: Number of Subgroups, Average Number of Conditions, Jaccard Index, Accuracy, Precision, Recall and R-squared. In Chapter 5 we will use our model for prediction and compare it to three other baseline models: Random Forest, SVM, and Decision Tree. For this comparison, we will use Accuracy, Precision and Recall. The rest of the measures are some general measures that are used for explaining the rules in the best way possible.

Number of Subgroups: This measure shows the number of subgroups in the model. It can be helpful because the model with fewer numbers of subgroups is less complicated and easier to interpret.

Average Number of Conditions: This measure calculates the average number of conditions based on all subgroups in the model. This one, again, can show how complicated our model is and how easy it is to interpret the result.

Jaccard Index: This measure is also known as the Jaccard Similarity measure. It calculates the similarity of sample sets by dividing the intersection of the sample sets on their union.

$$J(A,B) = \frac{A \cap B}{A \cup B} \qquad (6)$$

Accuracy: This measure indicates how accurate the model is if it is used in prediction on an unseen dataset. Accuracy can be defined as:

$$\text{Accuracy} = \frac{\text{number of correctly classified points}}{\text{total number of points}} \qquad (7)$$

Precision: This measure is also applicable for finding out how good our model is in prediction. It is equal to proportion of the classified labels that are correct.

$$\text{precision} = \frac{\text{relevant retrieved instances}}{\text{retrieved instances}} \quad (8)$$

Recall: This measure, as well, is applicable for finding out how good our model is in prediction. It measures the proportion of correct labels that are classified.

$$\text{recall} = \frac{\text{relevant retrieved instances}}{\text{relevant instances}} \quad (9)$$

R-squared: This measure is used in regression models to investigate how much of the dataset variance is explained by the model built based on regression. The formula for this measure is as follow:

$$R^2 = 1 - \frac{Residual\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares} \quad (10)$$

Residual sum of squares is the sum squared of the model errors in predicting the target variable for each instance(residual errors). Total sum of square is the sum square of errors of the simplest possible model for describing the data(the mean model).

Chapter 4

Data

The dataset we concentrate on in this study is MyHeart Counts Cardiovascular Health Study [15]. This smartphone-based dataset collected from March 10, 2015 to October 28, 2015 in the United States. It was collected using the iPhone application MyHeart Counts iOS and based on the data of participants who consented to use their data in research [14]. Some parts of this dataset are sensor data recorded by iPhone, such as movement and sleep data, and some other parts are being collected using questionnaires inside the application. Figure 1 is taken from [14] and shows the number of participants in this study**.** Table 3 indicates all the tables of this dataset that we are interested in and some explanation about them.

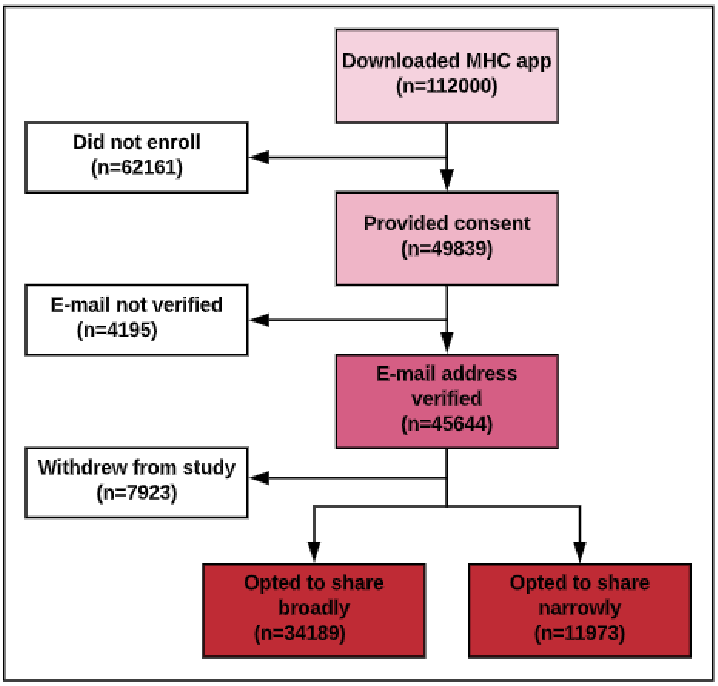


Figure 1: Number of Participants in the Study(adapted from [14])

Table 3: Tables of the dataset

Tables	Description	Unique Users
Physical Activity Readiness	Questionnaire about how ready the participant is for physical activity(survey)	22136
Daily Check-in	Questionnaire about sleep and activity in a daily basis(survey)	16593
Activity and Sleep Survey	Questionnaire about activity and sleep in general(survey)	21382
Risk Factor Survey	Questionnaire about different risk factors for cardiovascular disease(survey)	14485
Heart/Stroke Risk Score + Heart Age	Questionnaire for calculating heart age(survey)	4569
6-Minute Walk Test	Used guided 6 minutes walk test result(sensor)	3639
HealthKit Data	Sensor data from application regarding user movement(sensor)	4320
HealthKit Workout	Sensor data from application regarding workout(sensor)	881
Demographics	Demographic questionnaire(survey)	3320
Motion Tracker	Core motion phone data	21,382

In Section 4.1 we see the process of accessing the data. Section 4.2 is about our first stage of preprocessing, in which we go through all interesting tables. In Section 4.3 we describe our final dataset.

4.1. Data Acquisition

My Heart Count data set is not a public dataset. It is accessible through the dHealth portal [66]. The focus of this portal is on digital health data and tools. To access the My Heart Count dataset we made an account on the website, defined our project and agreed to the data-specific conditions. We used the Python Synapse client (https://pypi.org/project/synapseclient/) for accessing and pre-processing the dataset.

4.2. Pre-Processing

Since the dataset includes different tables, in the first step we go through each of them to see what information we can gain from the dataset and find out available features and features we can extract. Generally, in this step, we study the data using extensive pre-processing. All ten tables we study are mentioned in Table 3. In this section, we go through pre-processing steps of all these tables.

Features we are interested in can be put into four groups. First are the sleep-related features, such as sleep time or wake-up time, duration of sleep or being diagnosed with sleep problems. The second category is physical activity-related attributes including but not limited to duration of physical activities, part of the day it takes place, energy burn, distance, weekend activity etc. The third group are demographic features of the participants such as age, gender, height, weight and ethnicity. Lastly, we also want to extract information regarding having cardiovascular disease or its risk factors.

In our dataset, we have two types of data tables: cross-sectional tables including Risk Factor Survey, Physical Activity Readiness(PAR), Heart Age, Activity and Sleep Survey, and Demographics. Time series data: 6-Minute Walk Test, HealthKit Data, HealthKit Workout, Motion Tracker and Daily Check-in. The overall steps for preprocessing each group are almost the same. Therefore, in the upcoming subsections, we go through preprocessing of the tables based on these two groups of data tables and at last, we take a look at preprocessing the merged dataset.

4.2.1. Cross-Sectional Tables Preprocessing

For tables of this group we generally need to take these steps: dropping unnecessary columns, handling duplicate values, finding and dealing with noisy values, changing the format of the data, extracting new features. There are four unnecessary columns in all tables of this group, including 'recordId', 'appVersion', 'phoneInfo' and 'createdOn'. Since we do not need the information in these columns, we drop them in all data tables. All these tables include huge number of duplicates, since they were filled once every 90 days. For dealing with this issue, we replace numeric duplicate values with the “average”. Regarding nominal values for all cross-sectional tables except for PAR, we keep the first value. Our reason is for these attributes it was not reasonable that the user fills the form twice; For example, regarding their smoking history or their family’s early heart age diagnosis history. In the PAR table, we only keep the last entry since this table includes questions about having specific symptoms such as chest pain, dizziness, or heart problems; Therefore, the last entry in this table is more important. Now, we take a look at each tables’ specific preprocessing steps.

Activity and Sleep Survey: This table is related to the activity and sleep questionnaire answered by the participants every 90 days. The number of items in this table is 24,966. After preprocessing the data we have a table with 21,570 unique users and 19 attributes.

In this table, other than dropping unnecessary columns and replacing duplicates, the way we mentioned earlier, we drop noisy values for vigorous activity, moderate activity and sleep time1 attributes. We distinguish the noise by visualization of the data columns in the form of box plots. Vigorous and moderate activity are the average amount of weekly vigorous and moderate activity of the user respectively, reported by users. For these attributes, we only keep items less than 4,200 minutes. Sleep time1 is the users’ answer to this question: How much sleep do you think you need every night to be rested? (in hours). There were no predefined restrictions for participants in entering this data. To be able to represent the general population, it is assumed that more than 15 hours of sleep were entered erroneous.

In addition, we extract some new features from this table: mostly_sit_stand, mostly_walk, mostly_lift, hard_physical_activity, not_much_physical_activity, once_or_twice_physical_activity, three_times_physical_activity, daily_physical_activity, three_times_vigorous_activity, daily_vigorous_activity. These features are the options for some questions in this table that seem useful for our future analysis and modelling. More details about this table are available in Appendix A.1.

PAR-Q Survey (physical activity readiness survey): The second table we analyze is the PAR-Q or physical activity readiness survey. This table includes participants' answers to questions regarding their readiness for physical activity. The original number of items in this table is 25,815. After dropping the duplicate values and only keeping the last item, we get 23,990 items. The attribute we use in the rest of this study from this table is the heart condition. More details about this table are available in Appendix A.2.

Heart Age: The heart age table is based on a questionnaire filled out every 90 days; The number of rows in this table is 10,772 before dropping duplicate values and 4,760 after that. This table is related to attributes needed for calculating the Heart age of a participant such as cholesterol, blood pressure, age, etc. Heart age, as it is obvious from its name, is a measure for calculating the age of the heart and its vessels based on heart disease risk factors. More details about this table are available in Appendix A.5.

In this table, after general steps for preprocessing, we see that there are invalid values for some attributes including a systolic blood pressure more than 180 millimeters of mercury (mmHg), or less than 95(mmHg), Diastolic blood pressure more than 120 (mmHg) or less than 55(mmHg), blood glucose greater than 15 milligrams or less than 3 milligrams, Hdl cholesterol level larger than 7 mmol/L(millimoles per liter),or smaller than 0.8 (mg/Dl), Ldl cholesterol level more than 7 mmol/L, or less than 1 mmol/L. We first replace these values with a null value and use the misforest imputation [67] to find the best replacement for the empty cells. Features from this table that we use in the rest of this study include age, ethnicity, gender, hypertension, cholesterol and diabetes.

Risk Factor Survey: This table concerns the risk factors participants might have. This is again, based on a questionnaire filled out by participants on the first day of participation. The number of items in this table is 14,277. After dropping duplicates, this number reaches 13,852. We also extract some features from this table based on the options of the question related to medication to treat, in which participants indicated what medications they were using; This could reveal medical conditions they were struggling with. The features of this table used in our analysis include lower blood pressure, lower cholesterol treatment, diabetes, heart disease, and vascular disease. More details about this table are available in Appendix A.3.

Demographic: This table is about participants demographic information such as weight, height, age, and sex, in addition to waking up and sleeping time. This table has 12,439 rows. In addition to dropping unnecessary columns and replacing duplicate values with averages, we change the unit of weight attribute from pounds to kilograms; We also transform the height unit to centimeters from inch; There are some rows with only null values that we drop. For waking-up time and sleeping time attributes in this table, we change the format of these two to date time. The attributes of this table used in the final dataset are patientWeightPounds, patientHeightInches, age, waking_time and Gender; After all these steps, we will have the data of 3,320 participants in this table. More details about this table are available in Appendix A.6.

Figure 2 shows the necessary steps for each data table. In general, we have four potential preprocessing steps here: handling duplicates, handling noise, unit transformation and feature extraction. Here we see which steps are applied to each table.

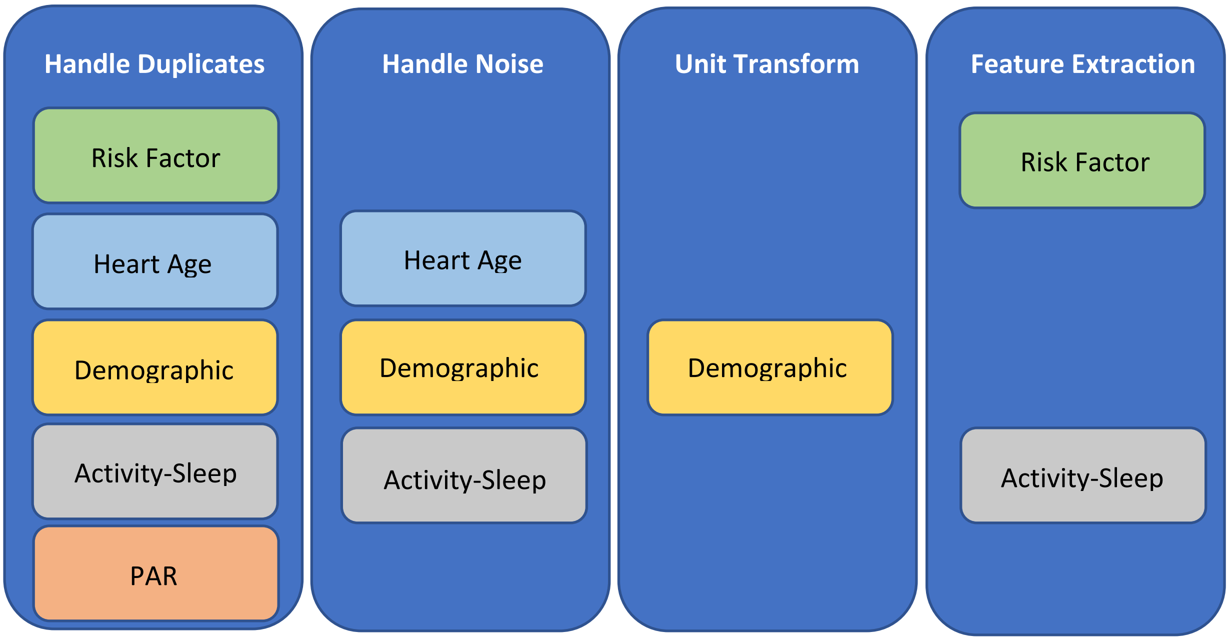


Figure 2: Cross-Sectional Tables Preprocessing Steps

4.2.2. Time Series Data Preprocessing

Tables HealthKit Data, HealthKit Workout, 6-Minute Walk Test and Motion Tracker are related to the sensor data based on participants' activity. Table Daily Check-in does not contain sensor data; however, it is in time series format(a form of a questionnaire filled by users every day). In the next subsection, we go through the sensor-time series and questionnaire time series preprocessing steps.

Sensor Time Series: in this subsection, we focus on the HealthKit Workout , HealthKit Data, Motion Tracker and 6-Minute Walk Test tables.

HealthKit Workout includes information about the duration, distance, energy consumed and type of physical activity recorded by wearable sensors. A list of different possible activity types can be found at [68]. It consists of the data of 880 unique participants. The original table includes excel files about the physical activity data of participants in a time series format. Each excel file encompasses these columns: startTime, endTime, type, workoutType, total.distance, unit, energy.consumed, unit.1 , source, sourceIdentifier.

The original HealthKit Data includes 4,920 unique users. This table also includes physical activity data recorded by wearables. There is an excel file for each participant and activity in a specific period of time. These excel files have startTime, endTime, type, value, unit, source, sourceIdentifier. One difference between this table and the HealthKit Workout is that we do not know activity types here. In addition, each file might include information related to one of the energy, heart rate, count of the steps or distance attributes. This information is not mentioned in different columns. we can distinguish it based on the unit specified for the specific row.

Motion Tracker data is recorded based on the users' phone motion sensors [69]. It includes some JSON files with information on users' changes of state. There are five different possible states: stationary, unknown, running, walking and cycling.

The general flow chart for preprocessing these three tables is drawn in Figure 3. For each table, there might be some differences in the details of every step. Our purpose is to make cross-sectional data out of them. We do this by calculating the average per day for numeric values and the mode for nominal values.

In this regard, we first download data files of each table, then convert all files related to one participant to a data frame and make a list of data frames from all participants' data. Since this data is recorded using sensors there are a lot of noisy values. We first find these values and drop them. The process of finding invalid values is more of a trial and error since the data is too huge to investigate all of it. An example of an issue is start time of activities having a value of zero which is not acceptable. Dropping rows with invalid values might result in an empty data frame. Therefore, after this step, we check whether the data frame still contains information. After this, we convert time-related columns to date-time format to be able to do our next calculations. We extract the duration of each activity using the start and end times. In HealthKit Data and HealthKit Workout tables we have separate columns for start and end time of activities. However, 6-Minute Walk Test and Motion Tracker files are just series of activities. In these tables, the start time of one activity is the end time of the previous one.

Sometimes, the columns are moved to the right resulting in values in some cells not being valid anymore. We move the values to their correct place in these cases. In addition, these steps might result in having NaN values in the start time column. We also drop these rows.

After these general steps, we add some auxiliary variables to help us extract the features we look for. These variables can be different in each table, though some are common. Common variables include hour (the start hour of physical activity), day of the week (a value the 0-6 range for showing which day of the week an activity take place), day part(this variable label each activity based on the part of the day it takes part in including early morning (5-9), morning (9-11), noon(11-13), afternoon(13-17), evening(17-21), late evening (21-23:59) and night(later than 23:59)), number of days(how many days the user data been recorded), weekdays(number of weekdays the user was active in), weekend(number of weekends the user was active in) and duration(duration of each activity).

In HealthKit Workout, for energy and distance, we have separate columns. However, in the HealthKit Data, the data is recorded in a different way. For each activity, this information is saved in various files. Therefore, we also create energy and distance columns for each of these variables extracted from multiple files. Moreover, in HealthKit Data table, we have the data related to heart rate and steps.

After adding these auxiliary columns to our tables, we are able to extract the features we like. The common extracted features in all these three tables are the average duration and count of physical activity in different parts of the day, weekends and weekdays. For HealthKit Workout and HealthKit Data tables, we also extract average energy and distance in different parts of the day and different parts of the week. We also calculate the average amount of energy, distance and duration for each user. In addition, we find the day part the user is most active in.

In the HealthKit Workout table, we also have the information regarding different types of activity. Average duration, count, energy consumed, and distance are calculated for each activity.

In the HealthKit Data table, we also extract features related to heart rate and steps. However, in this table, we do not have information about the types of activity.

For core motion, the focus is on whether the user is active. Here again, we have information about types of activities, but they are not the same as HealthKit Workout. We have five activity types meaning: running, cycling, unknown, stationary and walking. So we calculate the average duration and count of each of these activities as well. In addition, we extract features regarding the number of times users change their position from active to stationary and the average duration of being active and stationary per day. Further, we calculate the duration of being active (not stationary) on average per day in different parts of the day, weekends and weekdays.

After dropping non-valid values, some data frames will be empty. Therefore, for making a data frame of all the values for each table we do not consider those empty data frames.

After preprocessing the HealthKit Data table and HealthKit Workout, we merge the information of the users in these tables together. For mutual columns such as duration and energy, we consider the average value for numeric attributes and for nominal, we consider information in the HealthKit Data table since it seems to be less noisy and include more data.

Regarding the 6-Minute Walk Test table, the format of the files is closer to core motion data meaning there are not two sperate columns for start and end time, but it is just a series of events. This table includes the data from the six-minutes-walk(6mw) [70] test for different participants. For each participant, there is a JSON file with this information inside it direction unit(always equal to meters), vertical Accuracy, horizontal Accuracy, displacement Unit, direction, displacement, altitude, and timestamp. After preprocessing the data we will have 339 unique users' information. From this file, we only calculate the distance the user traversed during this test based on the column called displacement. Concerning pre-processing of this table, there are two possibilities for each file, either it is only one dictionary, meaning it includes only one timestamp, or it includes more than one.

In the beginning, we merge the data of each user into a data frame. Then for pre-processing each data frame for each participant, we change the format of time-related values to date time. In addition, some users tried this test more than once. So in each data frame, we first calculate the difference between the maximum and minimum of the time they tried this test, if it is larger than 7, then there is a possibility that the user tried it more than once since the test itself only takes 6 minutes. In this case, we calculate the displacement based on the last time user tried it. Otherwise, we compute it by summing all the values. In addition, there is another possibility, and that is test taking less than 6 minutes. This data is not valid and we consider that as a null value. Figure 4 shows the flowchart for calculating this step. More details about these tables are available in Appendix A.7,A.8 and A.10 to .

Questionnaire Time Series(Daily Check-in): The daily check survey includes information regarding 17622 unique users. Since in this table, participants' data was recorded daily we calculate the mean and extracted some extra features from it. The features in this table are related to phone-use duration, physical activity duration and type of it(light or intense) during the day.

The preprocessing of data in this table is straightforward. We find all the data related to each user and extract the interesting features from it by filtering data. Th extracted features include activity_dasys, Light_intensity_count, Moderate_intensity_count, Vigorous_intensity_count, Light_intensity_time, Moderate_intensity_time and Vigorous_intensity_time. These features are related to the number of times a user-declared physical activity and its duration based on its type. More details about this table are available in Appendix A.9.

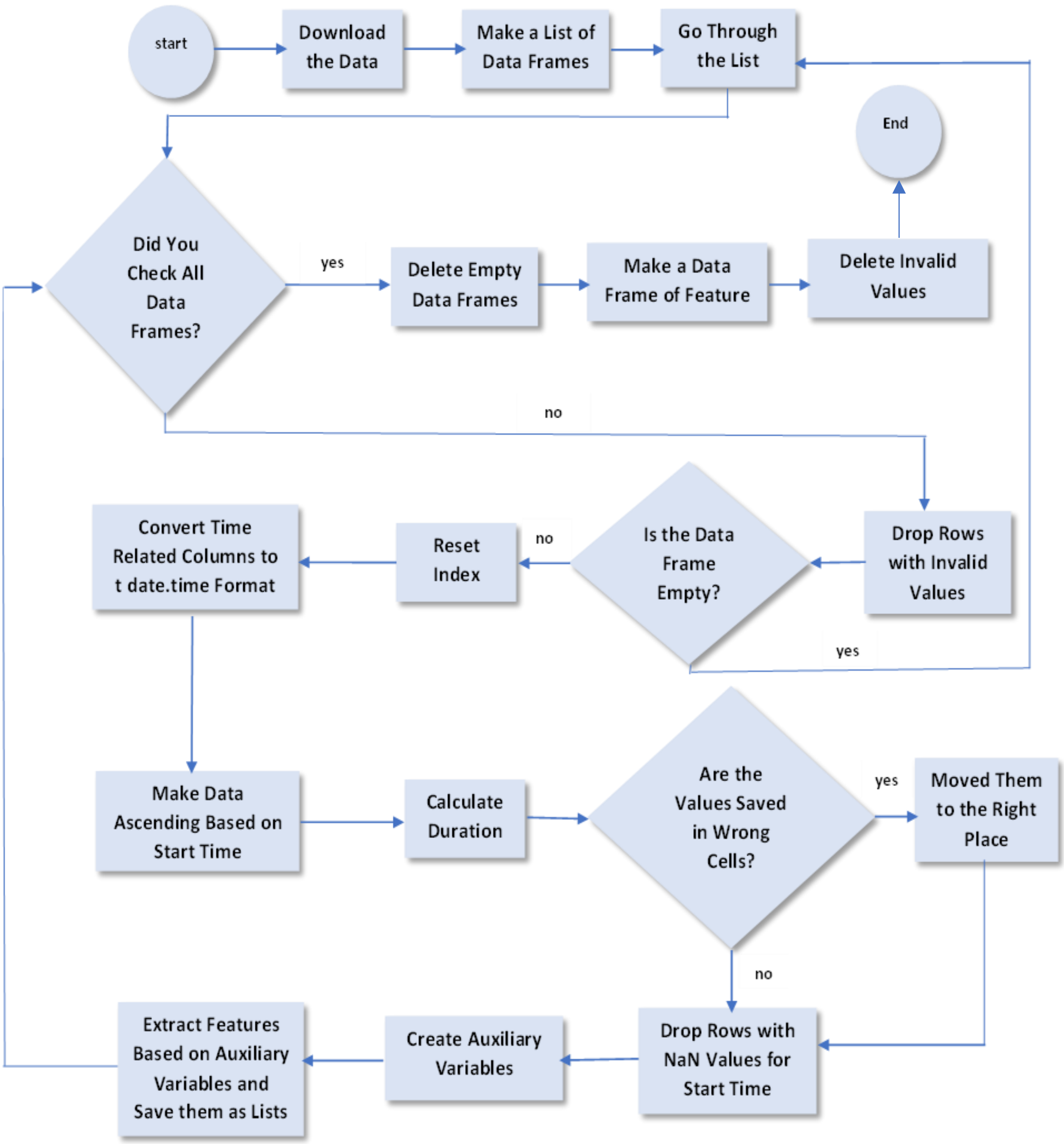


Figure 3: 'HealthKit Data', 'HealthKit Workout' and 'Motion Tracker' Pre-Processing Flowchart

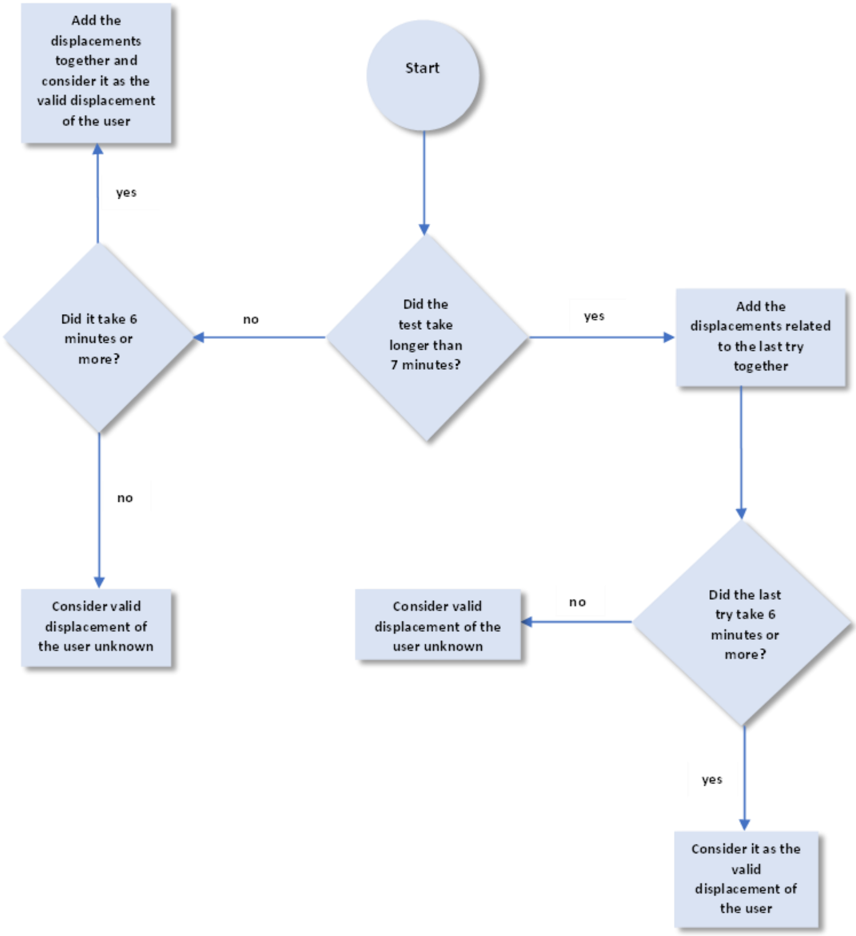


Figure 4:6-Minute Walk Test Pre-Processing Flowchart

4.2.3. Joined Data Preprocessing

After analyzing all available tables, we distinguished the features that can be helpful in our study. It is good to mention again that the purpose of the study is to find interesting relations between physical activity attributes and having cardiovascular disease. Therefore, we eventually do not consider features that are not related to either of these factors.

In the final preprocessing step, we join all the mentioned tables together based on a left-join on the Motion Tracker table, hence the number of unique users in this table is more than in other tables. In addition, we check our final data frame for invalid values again and replace noisy values with Nan.

Our target variable is a binary variable indicating whether the participant has a cardiovascular disease or one of its risk factors. The issues we consider as cardiovascular disease include heart disease and vascular disease. For risk factors, we consider Diabetes, hyper/hypo tension and high cholesterol.

By heart disease, we mean Heart Attack/Myocardial Infarction, Heart Bypass Surgery, Coronary Blockage/Stenosis, Coronary Stent/Angioplasty, Angina (heart chest pains), High Coronary Calcium Score, Heart Failure or Congestive Heart Failure, Atrial fibrillation and Congenital Heart Defect.

Vascular disease means Stroke, Transient Ischemic Attack (TIA), Carotid Artery Blockage/Stenosis, Carotid Artery Surgery or Stent, Peripheral Vascular Disease (Blockage/Stenosis, Surgery, or Stent), and Abdominal Aortic Aneurysm.

For extracting the final target variable, we first extract four different variables for cardiovascular, cholesterol, hyper-hypo, and diabetes, respectively. We merge the information in different tables to get the final label. This is because not all participants filled all the tables. The cardiovascular variable indicates whether a participant has cardiovascular disease or not. It is based on merging users' answers to questions in the risk factor table and PAR table regarding having a heart condition or cardiovascular disease. Cholesterol is related to having high cholesterol. This variable is partly extracted based on the heart age table and participants entering their cholesterol levels. We consider participants with higher cholesterol than 240 mg/dL (milligrams per deciliter) as positive for this criterion. The other part of the data is based on the risk factor table and participant indication of using medication for treating high cholesterol in this table. Hyper-Hypo is related to having hypertension or hypotension. We extract labels for this variable based on heart age and risk factor tables. In heart age, we both use the users' answers to the question of having hypertension and the data entered for their systolic and diastolic blood pressure. If they have a systolic blood pressure higher than 140(mmHg) and diastolic blood pressure lower than 90(mmHg) we consider them as positive for this issue. We also use the medication to treat variable in the risk factor table for this attribute. For diabetes, we use the heart age and risk factor table and participants' answers to the question about having diabetes in these tables.

Another attribute that we extract after joining all tables together is gender. Participants indicated their gender in two tables: heart age and demographic. We merge the information in these tables for the final data.

Finally, we encode data in a way to be able to use it in the RuleList algorithm. For selecting features, in the beginning, we counted on our knowledge based on previous studies. Then by running the algorithm multiple times we optimized our feature selection procedure.

4.3. Exploratory data analysis

The final dataset includes 12,043 participants with 91 attributes. Table 4 And Table 5 show categorical and numeric attributes in this study with related statistics. Information regarding the original questions and variables in each table is based on MyHeart Counts Public Researcher Portal available at: https://www.synapse.org/#!Synapse:syn11269541/wiki/588018. In Table 4 we have a column describing the attribute and its function(description), count column shows the number of participants the feature is available for. Mean and std are mean and standard deviation of feature values respectively. For nominal attributes, instead of mean and std we have the number of related items for each category.

Table 4: Numeric attributes in the study

name	description	count	mean	std
unkown_time_core	Average duration of unknown activity per day(in seconds)	8783	16978.91	9777.73
walking_time_core	Average duration of walking per day(in seconds)	10745	3574.47	4732.56
running_time_core	Average duration of running per day(in seconds)	12000	112.81	868.09
stationary_time_core	Average duration of stationary state per day (in seconds)	8782	47728.12	11473.80
cycling_time_core	Average duration of cycling per day(in seconds)	11498	1098.69	2385.60
morning_time_core	Duration of being active during morning per day (in seconds)	11863	725.24	984.09
noon_time_core	Duration of being active during the noon per day (in seconds)	11595	1418.23	1257.37
afternoon_time_core	Duration of being active during the afternoon per day (in seconds)	10658	4932.45	2223.48
evening_time_core	Duration of being active during the evening per day (in seconds)	10459	6008.21	2270.48
night_time_core	Duration of being active during the nigth per day (in seconds)	10486	6674.41	6417.10
active_time_core	Duration of being active on average per day (in seconds)	7414	27096.65	10646.20
change_of_position	Number of times users change their positions per day	12043	700.6	258.44
weekend_duration_core	Average duration of being active during weekends (in seconds)	12036	1617.50	6539.80
weekday_duration_core	Average duration of being active during weekdays (in seconds)	12001	4499.30	10812.44
early_morning_time_core	Duration of being active during the early morning per day (in seconds)	11617	1896.85	2015.87
late_evening_time_core	Duration of being active during the late evening per day (in seconds)	10568	4580.19	1753.13
patientWeightPounds	Weight of the participants	1006	85.77	20.90
patientHeightInches	Height of the participants	1023	175.78	9.50
moderate_act	Minutes of moderate activity in a week	10710	150.59	229.50
phys_activity	Leisure Time Activity	10710	2.66	1.89
sleep_time	Sleep time	12043	17.12	9.13

sleep_time1	Amount of sleep the participant usually get at night on weekdays or workdays	10710	6.85	1.11
vigorous_act	Minutes of vigorous activity the participant gets in a week	10710	69.36	133.23
Age	Age of the participants	3770	42.24	14.86
Displacement	Displacement of the participants during six minutes walk test(meters)	128	829.59	909.26
Sleep	Average sleep duration entered by the user(Minutes)	432	443.43	3203.833
activity_dasys	Number of days the user filled at least one activity	10927	0.54	2.19
Light_intensity_count	Number of times the user added at least one light intensity activity	10927	0.65	2.55
Moderate_intensity_count	Number of times the user added at least one moderate intensity activity	10927	0.21	1.13
Vigorous_intensity_count	Number of times the user added at least one vigorous intensity activity	10927	0.23	1.61
Light_intensity_time	Sum of the light intensity activity duration added by user	10927	2320.51	20820.76
Moderate_intensity_time	Sum of the moderate intensity activity duration added by the user	10927	428.59	4826.22
Vigorous_intensity_time	Sum of the vigorous intensity activity duration added by the user	10927	771.59	8549.57
Duration	Average minutes of activity per day (based on health kit data and health kit workout tables)	1556	16.34	77.54
Steps	Mean number of steps per day	1288	40.51	127.12
Energy	Mean amount of energy burnt per day(kcal)	1556	831.91	9917.34
Distance	Mean amount of distance passed per day(meters)	1556	742.07	3907.05
night_steps	Average number of steps at the night	1288	9.98	92.55
evening_steps	Average number of steps the evening	1288	8.35	10.18
afternoon_steps	Average number of steps in the afternoon	1288	7.50	11.01
noon_steps	Average number of steps at noon	1288	1.67	3.80
morning_steps	Average number of steps in the morning	1288	0.71	4.51
night_distance	Average amount of distance passed at night	1556	68.71	408.13
evening_distance	Average amount of distance passed in the afternoon	1556	143.18	1060.86
afternoon_distance	Average amount of distance passed in the afternoon	1556	264.10	2469.25
noon_distance	Average amount of distance passed at noon	1556	106	1525.98
morning_distance	Average amount of distance passed in the morning	1556	22.01	279.35
night_energy	Average amount of energy burnt activity at night	1556	115.12	1488.23
evening_energy	Average amount of energy burnt in the evening	1556	95.01	592.08
afternoon_energy	Average amount of energy burnt in the afternoon	1556	120.72	1475.19
noon_energy	Average amount of energy burnt at noon	1556	51.61	680.75

morning_energy	Average amount of energy burnt in the morning	1556	21.86	384.63
night_time	Average amount of time spent on physical activity at the night	1556	4.10	47.29
evening_time	Average amount of time spent on physical activity in the evening	1556	2.16	8.79
afternoon_time	Average amount of time spent on physical activity in the afternoon	1556	2.52	12.20
noon_time	Average amount of time spent on physical activity at noon	1556	0.77	6.87
morning_time	Average amount of time spent on physical activity in the morning	1556	0.24	2.18
weekend_energy	Average amount of energy burnt at weekends	1556	1341.87	22184.64
weekend_distance	Average amount of distance passed in the weekends	1543	342.54	1577
weekend_duration	Average amount of time spent on physical activity in the weekends	1556	20.91	123.46
weekend_steps	Average number of steps in weekdays	1288	45.77	300.23
weekday_energy	Average amount of energy burnt at weekends	1556	755	7979
weekday_distance	Average amount of distance passed in the weekdays	1556	460.40	2683.97
weekday_duration	Average amount of time spent on physical activity in the weekdays	1556	15.97	88.72
weekday_steps	Average number of steps in weekdays	1288	45.32	188.78
early_morning_steps	Average number of steps in the early mornings	1288	6.46	84.70
late_evening_steps	Average number of steps in the evenings	1288	5.84	7.89
early_morning_time	Average amount of time spent on physical activity in the early morning	1550	1.75	13.28
late_evening_time	Average amount of time spent on physical activity in the late evening	1556	1.71	13.49
early_morning_energy	Average amount of energy burnt in the early morning	1556	374.96	9306.95
late_evening_energy	Average amount of energy burnt in late evening	1556	52.62	303.88
early_morning_distance	Average amount of distance passed in the early morning	1556	18.87	198.49
late_evening_distance	Average amount of distance passed in the late evening	1556	118.72	1344.33
waking_time	Waking up time	12043	7.02	2.23

Table 5: Categorical attributes in the study

Name	Description	count	categories
heartAgeDataEthnicity	Ethnicity	3702	0: White(2881) 1: Asian(276) 2: Hispanic(274) 7: Other(122) 3: Black(119) 4: American Indian(18) 5: Pacific Islander(10) 6: Alaska Native(2)

Atwork	Work Time Activity	10710	0: I spent most of the day sitting or standing(8252) 1: I spent most of the day walking or using my hands and arms in work that required moderate exertion (2133) 3: I spent most of the day doing hard physical labor(270) 4: None(55) 2: I spent most of the day lifting or carrying heavy objects or moving most of my body in some other way (0)
phys_activity	Leisure Time Activity	10710	1: Once or twice a week, did light activities (3030) 3: Almost daily, that is five or more times a week, did moderate activities (2543) 4: About three times a week, did vigorous activities (1534) 0: did not do much physical activity (1433) 5: Almost daily, that is, five or more times a week, did vigorous activities(1342) 6: None(828) 2: About three times a week, did moderate activities (0)
sleep_diagnosis1	Being diagnosed with sleep disorder	10702	0: False(9479) 1: True(1223)
mostly_sit_stand	Whether the user chose the first option in 'atwork' section	10710	1: True(7115) 0: False(3595)
mostly_walk	Whether the user chose the second option in 'atwork' section	10710	0: False(8873) 1: True(1837)
mostly_lift	Whether the user chose the third option in 'atwork' section	10710	0: False(10486) 1: True(224)
hard_physical_activity	Whether the user chose the fourth option in 'atwork' section	10710	0: False(10663) 1: True(47)
not_much_physical_activity	Whether the user chose the first option in 'phys_activity' section	10710	0: False(9279) 1: True(1431)
once_or_twice_physical_activity	Whether the user chose the second option in 'phys_activity' section	10710	0: False(7682) 1: True(3028)
three_times_physical_activity	Whether the user chose the third option in 'phys_activity' section	10710	0: False(8171) 1: True(2539)
daily_physical_activity	Whether the user chose the fourth option in 'phys_activity' section	10710	0: False(9181) 1: True(1529)
three_times_vigorous_activity	Whether the user chose the fifth option in 'phys_activity' section	10710	0: False(9369) 1: True(1341)
daily_vigorous_activity	Whether the user chose the sixth option in 'phys_activity' section	10710	0: False(9884) 1: True(826)
day_part	Part of the day the user was mostly active in	1556	4: evening(486) 3: afternoon(451) 6: night(331) 5: late_evening(190) 0: early morning (45) 2: noon(34) 1: morning(19)
Gender	Gender of the participant	1077	0: Male(901) 1: Female(176)
Any of the issues(OR)	Whether the participant has any of the issues 9cardiovascular issues)	11691	0: False(8713) 1:True(2978)

As we can see in the tables, the average age of participants in this study is 42 years. In this study, the options for ethnicity include: White, Asian, Hispanic, Black, Other, Prefer not to indicate, American Indian, Pacific Islander, Alaska Native, or None. The majority of participants are male(84%) and belong to the white ethnicity(78%). Figure 5 illustrates the mean duration of being active and stationary per day and the mean active time in different parts of the day based on gender. Participants tend to be more in a stationary position. Female participants are less active than male participants. Total active time is around 7 hours for female participants and close to 8 hours for male participants. Participants are generally more active during the night. They are the least active during the morning. In Figure 6 we see how different is the activity duration among different ethnicities. The mean duration of being active is almost the same in all ethnicities(around 7.5 hours). Hispanic participants have a longer activity duration(almost 8 hours).

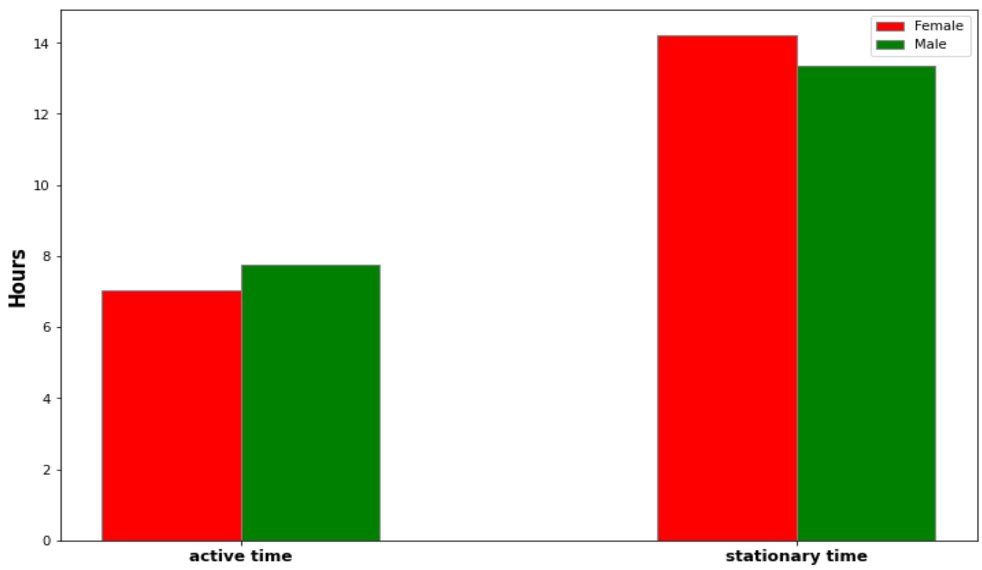


Figure 5: Mean Duration of Being Active and Stationary.

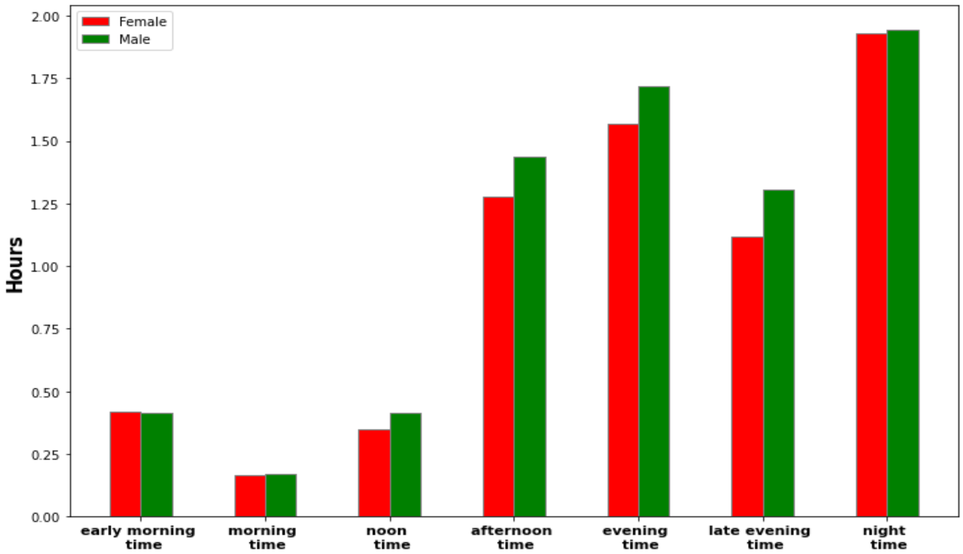


Figure 6: Mean Active Time in Different Parts of the Day

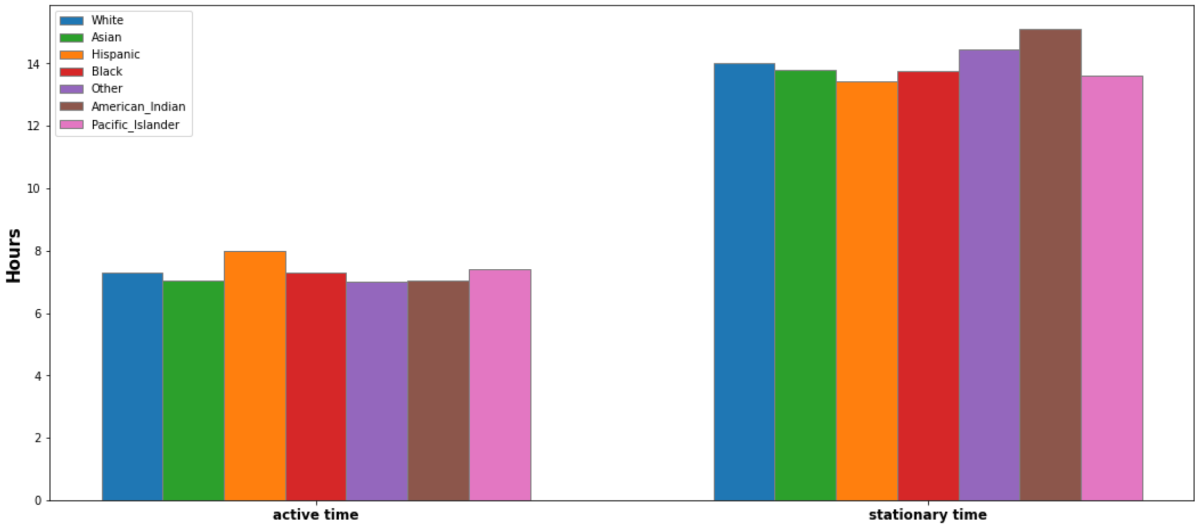


Figure 8: Average active time for different ethnicities

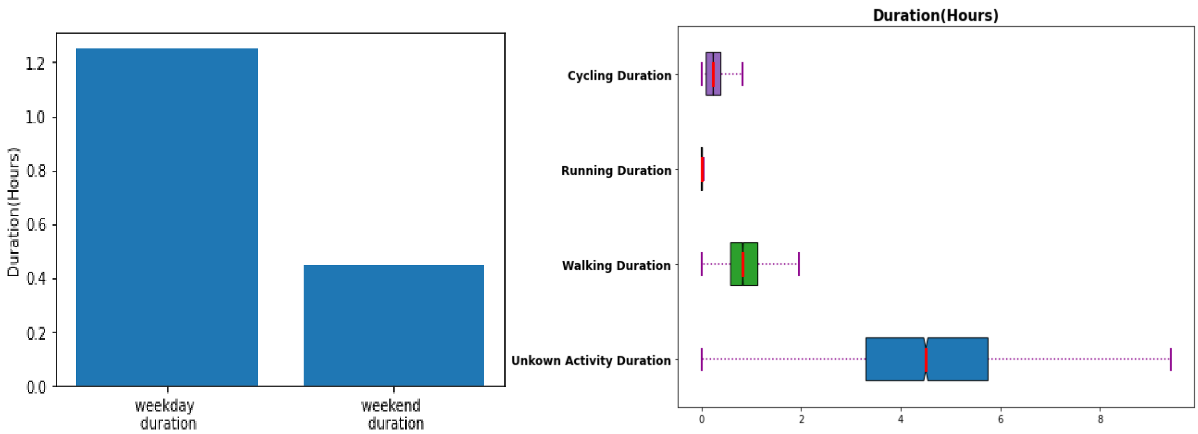


Figure 7: Left: Mean Duration of Being Active During Weekends and Weekdays; Right: Mean Duration of Different Activies

In addition, participants are more active during weekdays. They are on average two times more active in weekdays in comparison to weekends. In our study we have data related to cycling, running and walking. When the participant is active but the type of the activity is not obvious it is recorded as unknown. We can see in Figure 7 that the majority of the time the user is active, the activity is recorded as unknown. The second most popular activity is walking with an average of an hour per day.

Participants also answered a question regarding how active they are at their work. In Figure 9 we see that most of the participants(77%) indicated they mostly sit or stand during work. Around 20% of the participant mentioned they are mostly walking. 2.5% work in jobs with hard physical labour.

At last, there is a bar chart Figure 9 showing the number of participants with distinct compelling issues of this study. In general, the number of participants with hyper/hypo tension is higher in comparison to other issues (1,735). The least true label is related to diabetes(356). For our experiment, we have 2978 participants with at least one of the issues of the study. This means, in contrast to similar studies [12], [13], we have more healthy participants rather than sick ones.

Tables related to this section are available in more detail in Appendix A.11.

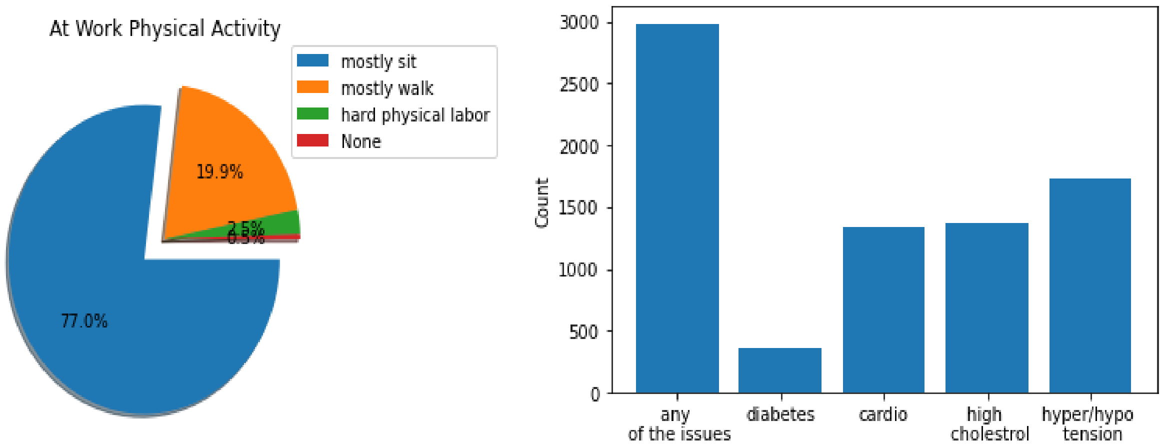


Figure 9: Left: At Work Physical Activity; Right: Number of Participants with Different Issues

Chapter 5

Results

After preprocessing the data and preparing it for implementation, we applied the SSD++ algorithm using the RuleList python package [18]. In this chapter, we go through the results obtained from our experiment. To begin with, we explain the general result and its evaluation. Then to investigate the complication of the problem and the strength of the SSD++ model, we use this model in a prediction task and compare its result with the Naïve Bayes, Random Forest, and Decision Tree classifiers results on the same task. After that, we dive deeper into the most interesting rules and their interpretation.

5.1. Experiment

As we mentioned in Chapter 3, SSD++ has 7 parameters: max_depth, beam_width, min_support, n_cutpoints, discretization, max_rules and alpha_gain. We tried different values for these parameters and realized the default values get better results in our case with a trivial difference. Therefore, the following results are based on default values for these parameters mentioned in Table 6.

Table 6: SSD++ Parameters and their values

Parameters	**Definition**	**Default Values**
max_depth	Maximum description size	5
beam_width	Number of selected patterns to be expanded in each iteration of beam search	100
n_cutpoints	Number of cut points in discretizing a numeric attribute	5
alpha_gain(β)	Normalize or absolute gain in expanding rules in beam search algorithm	1

Implementing the SSD++ algorithm on our data set results in 15 different rules. Table 7 includes all these rules in addition to the probability of suffering from cardiovascular disease(CVD) or its risk factors, usage and support of each rule and WKL. The description of WKL, usage and support can be found in Section 3.3. Rules one to four have the highest WKL. Rule 13 has the lowest WKL(25.86). In addition, rules one to eight are related to when the probability of having CVD or its risk factors is relatively high(above 50%), while rules 9 to 15 indicate rules concerning low probability(under 50%) of having these diseases.

Table 7: Result Table

Rank	Rule	Pr(1)	Usg	Support	WKL
1	age >= 59 AND early_morning_time < 0.19 AND unkown_time_core < 19096.46	0.92	109	109	156.30
2	running_time_core < 5.2 AND age >= 59 AND walking_time_core < 4664.097 AND late_evening_time_core >= 3047.38	0.87	154	182	187.07
3	age >= 48 AND vigorous_act < 10 AND active_time_core < 23308.87	0.83	87	135	90.72
4	age >= 59 AND early_morning_time_core < 1213.09	0.67	249	479	139.21
5	48 <= age < 59 AND 167.64 <= patientHeightInches < 180.34	0.82	40	46	41.32
6	48 <= age < 59 AND 4016.09 <= late_evening_time_core < 6230.24 AND 0 <= activity_dasys < 1	0.64	144	164	70.08
7	age >= 59	0.56	114	661	34.72
8	age >= 48 AND vigorous_act < 30	0.56	154	666	47.95
9	age >= 40 AND noon_time_core >= 1890.69	0.49	309	617	59.05
10	16223.51 <= unkown_time_core < 22821.48 AND 13.47 <= running_time_core < 139.63 AND vigorous_act >= 60	0.06	352	371	67.04
11	early_morning_time_core >= 2170.08 AND late_evening_time_core >= 4016.09 AND walking_time_core >= 2990.5 AND running_time_core >= 13.47	0.09	700	815	85.43
12	phys_activity >= 5 AND 3047.38 <= late_evening_time_core < 5404.69	0.13	730	983	48.38
13	age >= 40	0.38	477	1925	25.86
14	night_time_core >= 6167.19 AND running_time_core >= 1.5 AND noon_time_core < 543.43 AND atwork >= 0	0.11	934	1559	80.86
15	running_time_core >= 5.2 AND cycling_time_core >= 539.15	0.18	2678	4894	58.74

Table 8 is a summary of all the measures related to this experiment. The average support of each subgroup is 907. It means there are on average 907 instances following the patterns of each subgroup by considering the mutual instances in multiple subgroups. The average usage(coverage) of each subgroup is 482. This is the number of instances the pattern is based on without considering the mutual instances. WKL calculated based on support and usage are 210.92 and 79.52, respectively. The average Jaccard similarity is 7% which means the average similarity among different subgroups is less than 10%. This can show our model encompasses different parts of the data since the number of mutual instances in different subgroups is small. The average items is a measure of number of conditions in a rule. This measure is corresponding to 2.5 in our study. Another interesting measure is the length ratio which is 0.93. This measure is equal to the fraction of the final on the original encoded data length. Therefore, our proposed model reduces the encoded length of the data by 7%.

Table 8: Overall Measures of the Experiment

measure	values
Average Support	907.07
WKL based on Support	210.92
Average Usage	482.07
WKL based on Usage	79.52
Average Jaccard Similarity	0.07
Number of Rules	15
Average Items	2.53
Summation of WKLs	1192.75
Normalized Summation of WKLs	0.10
Original Length	9571.22
Final Length	8883.02
Length Ratio	0.93

In total 16 out of 90 features of the dataset are part of the patterns of the final model. The most frequent of all is age. This attribute is part of 10 out of 15 rules conditions, which means more than 50% of the rules have a condition about this attribute. The second popular attribute is running_time_core. This attribute appeared in five rules. It is related to the running time of the participants recorded by motion tracker sensors. The next popular attribute is Late_evening_time_core by appearing in four rules. vigorous_act appeared in three rules. Unknown_time_core, early_morning_time_core, walking_time_core and noon_time_core are used in two patterns. Lastly, there are some attributes only applied in one rule including cycling_time_core, at_work, night_time_core, activity_days, active_time_core, early_morning_time, phys_activity and patientHeightInches.

In addition, we investigate how much of the target variable variance is explained by these 15 rules. To this end, we calculate the R-square of the model which is 78%. Therefore, 78% of the variations in the target variable can be explained by this model.

5.2. Prediction

In this section we generate two random independent uniform sample of data: train and test set. Train set consist of 80% of the dataset(9352 instances of the dataset). Test set includes 2339 instances (Table 9). In this section, we examine our model for prediction on the test and train sets to see how good it can be in prediction. Even though, the purpose of our experiment is not to make predictions, still this experiment can give us an intuition of how good/bad our model generalized the dataset. Moreover, we train three classification models using SVM, Decision Tree and Random Forest algorithms on the train set. After that, we compare the prediction result of these models on the test set with the SSD++ algorithm. This result can show us how complicated the problem at hand is. We also consider a naive model as the baseline. In this naive model a set of the most frequent label (zero) is considered the prediction result for all instances.

Table 9: Number of True and False Labels in Train and Test Sets

	True Labels(1)	False Labels(0)	Total
Train Set	2358	6994	9352
Test Set	620	1719	2339

The results of these experiments can be observed in Table 10. Figure 10 also shows the confusion matrix related to these experiments. The SSD++ model accuracy is 76%, meaning out of all the unknown labels of the test dataset(2,445 instances), our model is able to predict 76% of them correctly. The precision of this model is 64%, indicating that out of all the times that the model predicts the label of an unknown item to be true, 64% of the time the label is actually true. The lowest value is related to the recall of the model(21%). It means out of all the instances having true labels our model only could predict 21% of them as true. The confusion matrix can make it more clear. We see that our model is more successful in predicting False(0) labels in comparison to True(1) which makes sense since there are more data with False labels (the dataset is biased toward healthier people). If the purpose of the study were the prediction, we would have concentrated on making the data balance. However, this is not the case here.

Table 10: Prediction Measures

Model	Precision	Recall	Accuracy
Naive Model	0	0	0.73
SSD++	0.64	0.21	0.76
SSD++ on Train Set	0.67	0.24	0.78
Naïve Bayes	0.50	0.13	0.73
Decision Tree	0.57	0.60	0.78
Random Forest	0.91	0.50	0.86

We also try predicting labels of the train set using the SSD++ model. In general, the result on train set is better based on all the measures as it was expected. However, there is no more than 3% difference between any of the measures. This result shows that our model is actually working without any overfitting. Regarding the naive model, since 73% of the target variables are zero, this model has a 73% accuracy on the test set which is even better than the Naïve Bayes model; however, the precision and recall of the model are zero since we do not have any true positive values in the predictions of the model. Based on Table 10, we can see that the SSD++ model enjoys better performance regarding precision(64%) in comparison to Naïve Bayes and Decision Tree classification methods. In general, Naïve Bayes has the worst performance in comparison to three other models based on all three criteria. Decision Tree model shows a better performance regarding recall and accuracy in comparison to SSD++ model; however, the accuracy of this model is only around 2% better. This model has the best performance considering the recall (60%). Random Forest has the best performance based on precision(91%) and accuracy(86%). The precision of this model is significantly 27% better than the SSD++ precision which is the second highest precision.

These results depict that the task of prediction on this dataset is complicated since the classical classification algorithms could not obtain better accuracy than 86%. This is also partly because of the unbalance dataset. In addition, even though the propose of the SSD++ algorithm is not classification, it nevertheless has a good performance in this task which is even better than some classic classification models such as Naïve Bayes and very close to Decision Tree. This guarantees that the patterns we discovered using this algorithm are not just some random outputs and they demonstrate the validity to be studied in more depth.

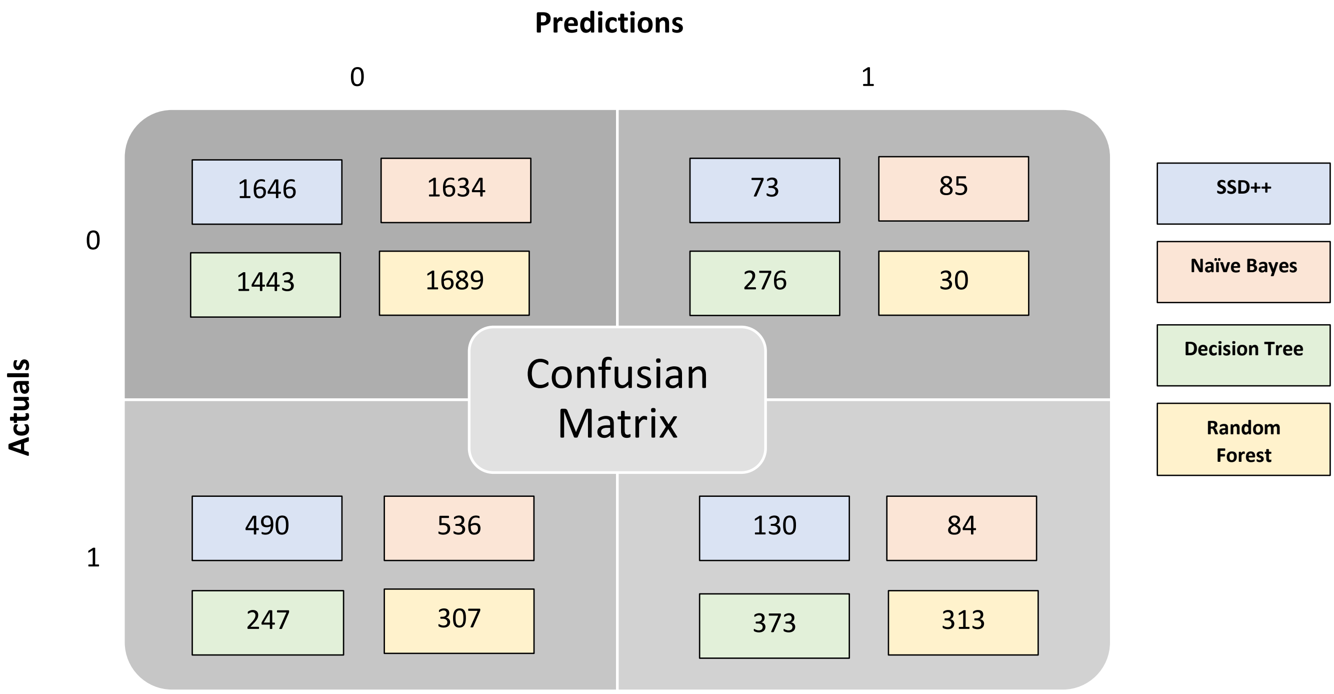


Figure 10: Confusion Matrix for Predictions

5.3. Rules

In this section, we go through five interesting rules found by the model and take a closer look at them using visualization. We compare the distribution of each subgroup with two datasets. One that we entitle “whole” is related to the whole population, and the other one that we mention as “healthy” is related to the dataset with only the healthy participants meaning participant who does not have any of the mentioned diseases. Visualization and interpretations related to 10 other rules are documented in Appendix B.

5.3.1. Rule1

Rule 1 implies that if a participant age is above 59 years, they are less active than 0.19 minutes during early morning and the unknown activity duration per day for them is below 5 hours and 30 minute then, the probability of having CVD or its risk factors is 91% for this user. There are 109 instances in the dataset that follow this rule. The WKL for this subgroup is 156.3.

In Figure 11, we compare the conditions of rule 1 with a box plot of the mentioned attributes. As we can see, more than 75% of the population is younger than 59, both in the healthy population and whole data. The mean and median for age attributes are lower in the healthy population. Regarding early morning time lower than 0.19 minute, it is clear that this amount is less than the mean for both the healthy and the whole population. In addition, the 75 percentile of unknown time activity is around 6 hours which is higher than the boundary of the third condition.

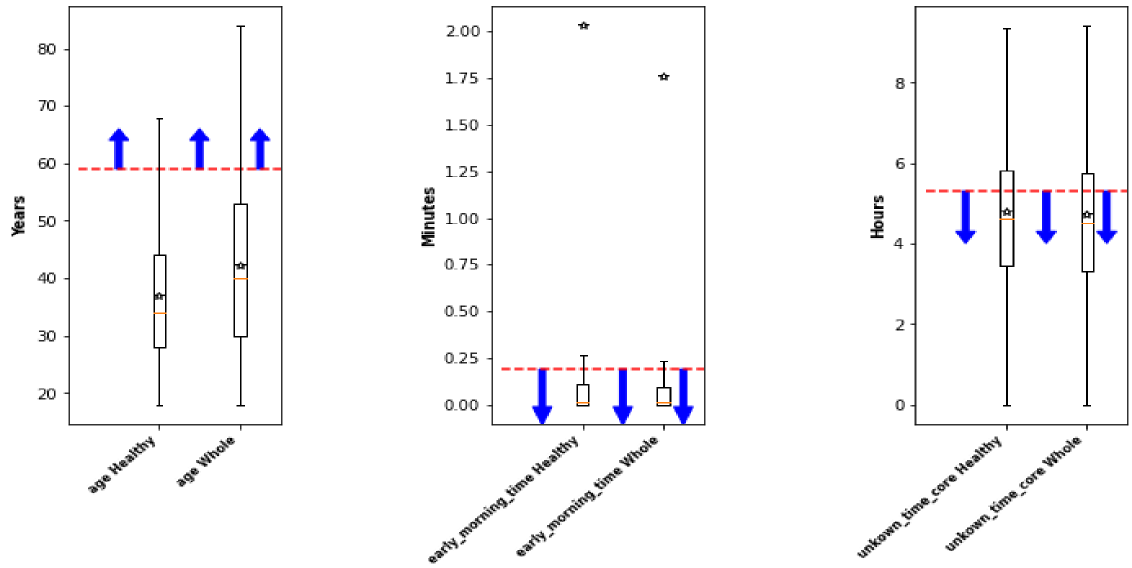


Figure 11: Rule 1 patterns in Comparison to the Healthy and Whole data distribution

In Figure 12, we compare the distribution of our subgroup with the healthy and whole population. We can see that subgroup 1 encompasses a broader range of people regarding age, and it has a higher median and mean in comparison to the two other populations. Notably, this gap is more significant regarding the healthy population. This trend is also true concerning early morning time. There is not much difference among unknown time activity distributions of the three categories of data.

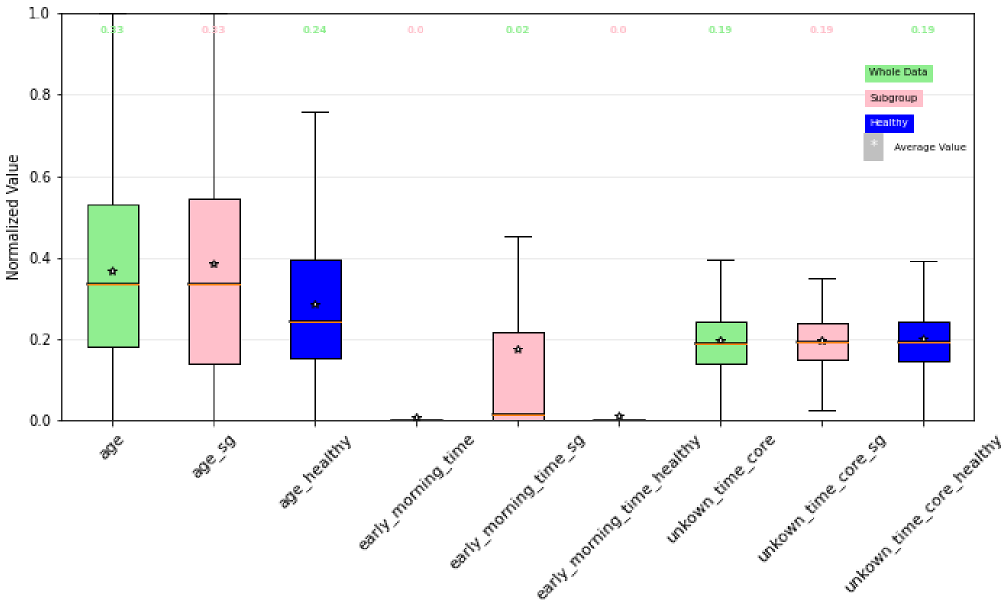


Figure 12: Distribution Comparison of Subgroup 1 with the Healthy and Whole Population

Figure 13 and Figure 14 illustrate two demographic attributes of subgroup 1, gender and ethnicity, as to the healthy and whole population. The percentage of female participants is 10% lower in this subgroup. In addition, the percentage of Hispanic ethnicity people is also lower in comparison to the two other datasets. There is no person from the American Indian and Pacific Islander ethnicities in this subgroup.

Regarding the height and weight of the participants(Figure 15), in subgroup one participants have a lower average weight in comparison to healthy and whole data. However, they are taller, and the height distribution is less scattered.

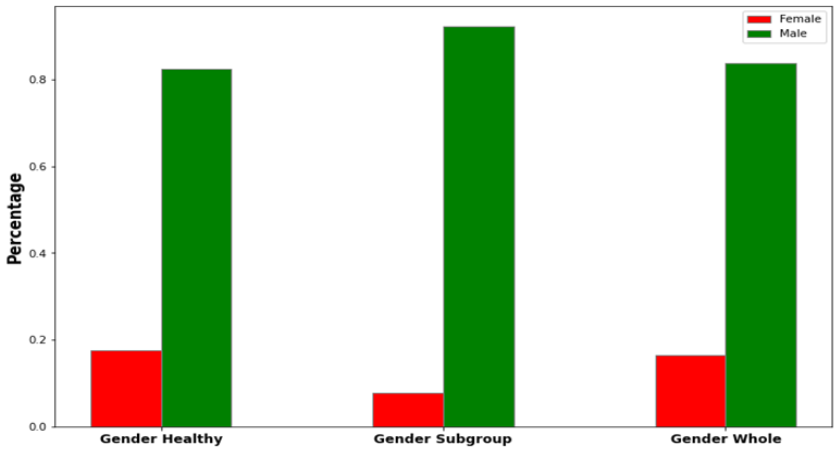


Figure 13: Gender Distribution in the Healthy, Subgroup1 and Whole Populations

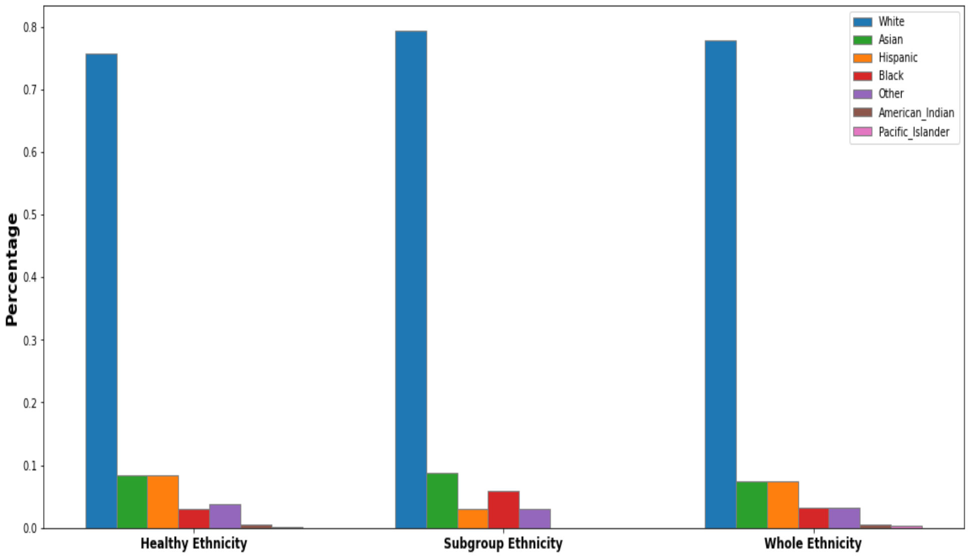


Figure 14: Ethnicity Distribution in the Healthy, Subgroup 1 and Whole Populations

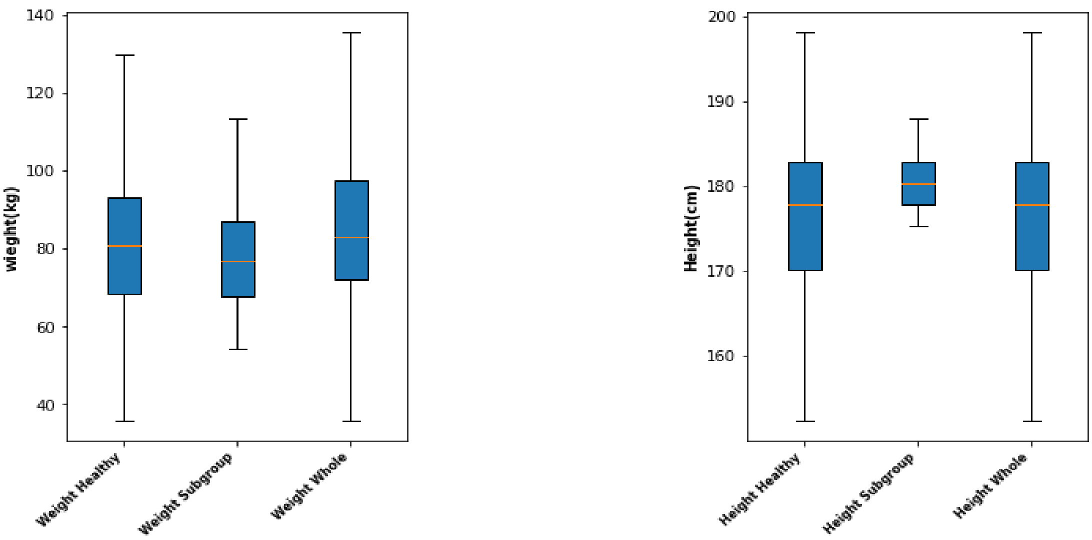


Figure 15:Height and Weight of Participants in Different Groups of the Data

5.3.2. Rule3

The third rule indicates if a participant is older than 48, they have less weekly vigorous activity than 10 minutes, and they are less active than 6 hours and 28 minutes per day they have an 83% chance of having CVD or its risk factors. When we look at Figure 16, we see that more than 75% of the healthy population is younger than 48 years old. In addition, 10 minutes of weekly vigorous physical activity is less than the mean and median of this attribute for both healthy and whole populations. This is also equally accurate about activity duration, implying this pattern is realistic.

In Figure 17, we see subgroup 3 has a larger median and mean for all these three attributes. Regarding the age of the population, in general, the population in this subgroup is older than the two other data groups. The weight of the participant is also heavier in this subgroup but, the median for the height attribute is smaller. In this subgroup, the proportion of women is around 40%, which is two times more than the two other subgroups. There is no participant from Pacific Islander, American Indian and Asian ethnicities in this subgroup(Figure 19).

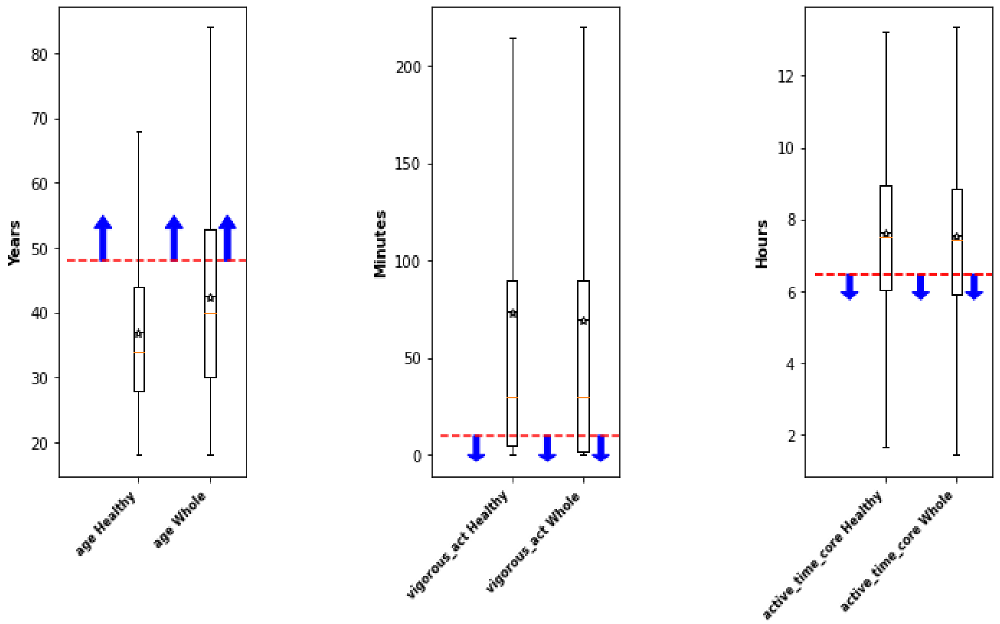


Figure 16: Rule 3 patterns in Comparison to the Healthy and Whole data distribution

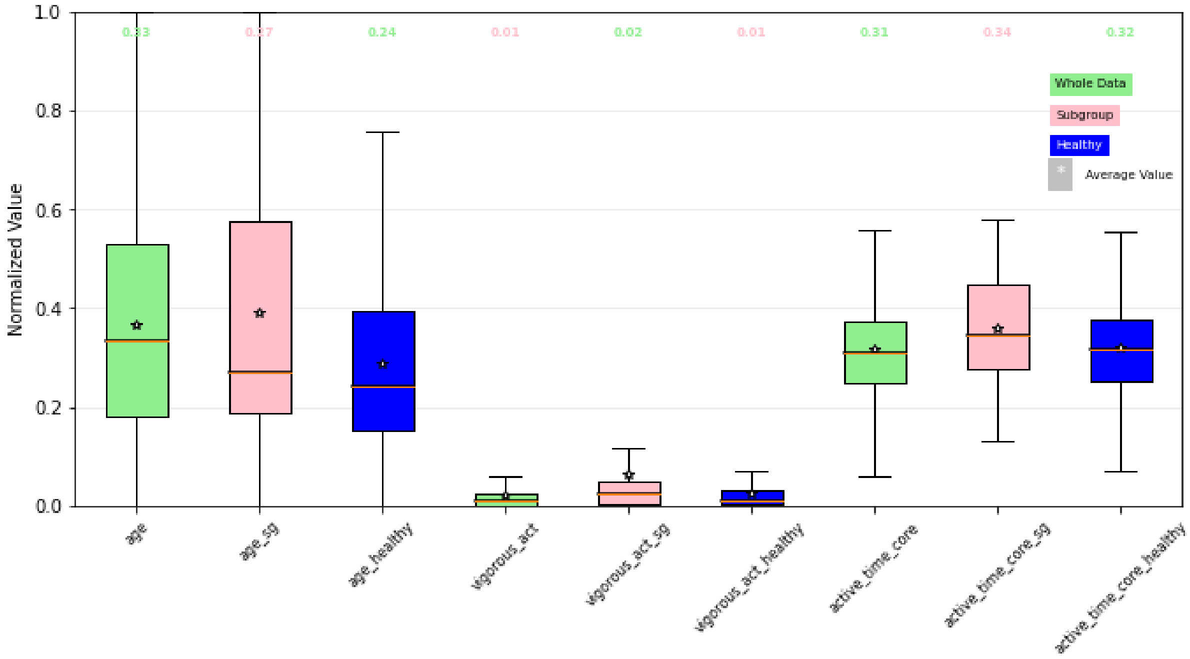


Figure 17: Distribution Comparison of Subgroup 3 with the Healthy and Whole Population

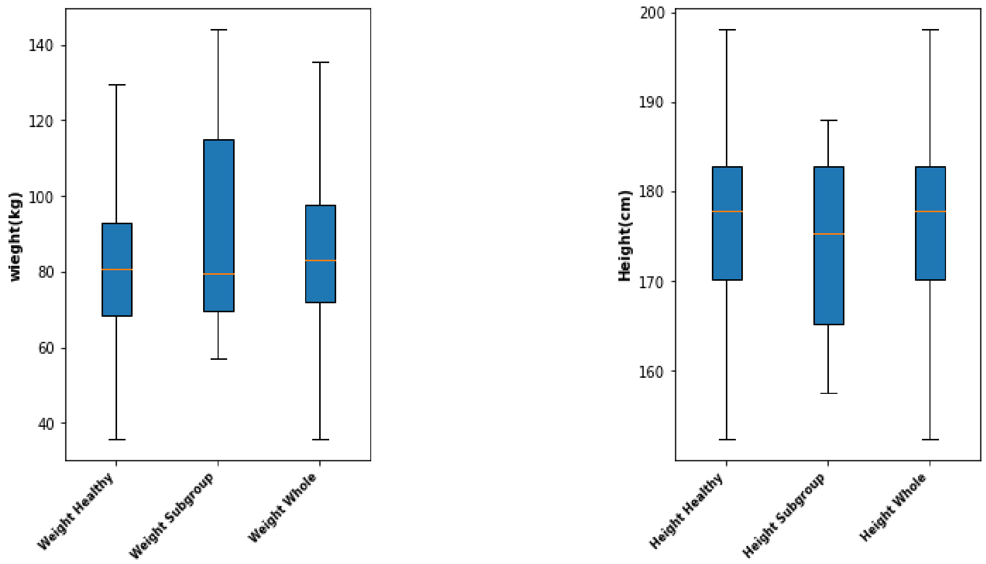


Figure 18: Distribution of the Height and Weight Attributes in Subgroup 3, the Healthy and Whole Populations

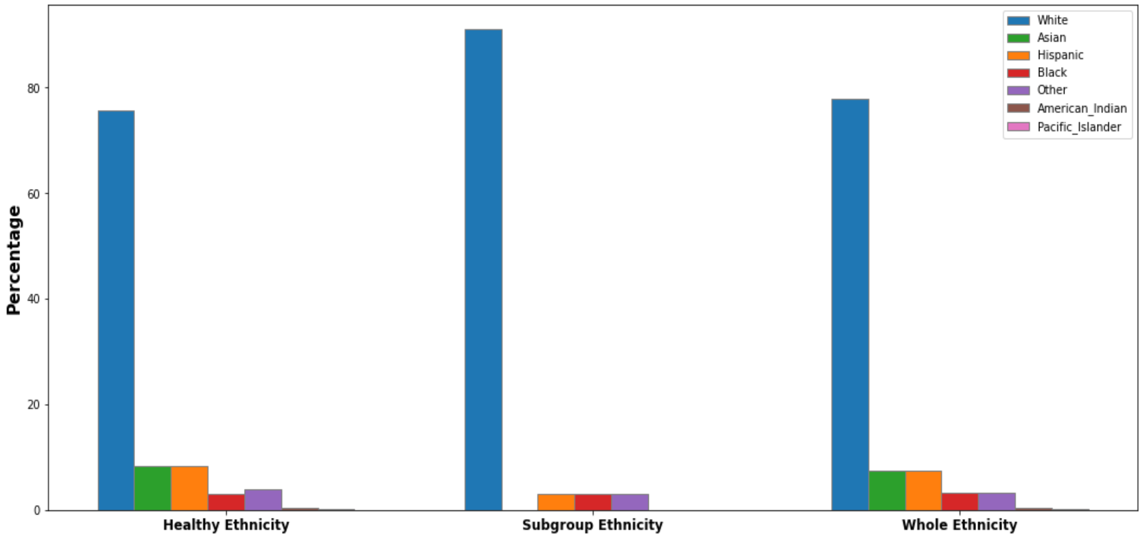


Figure 19: Ethnicity Distribution in the Healthy, Subgroup 3 and Whole Populations

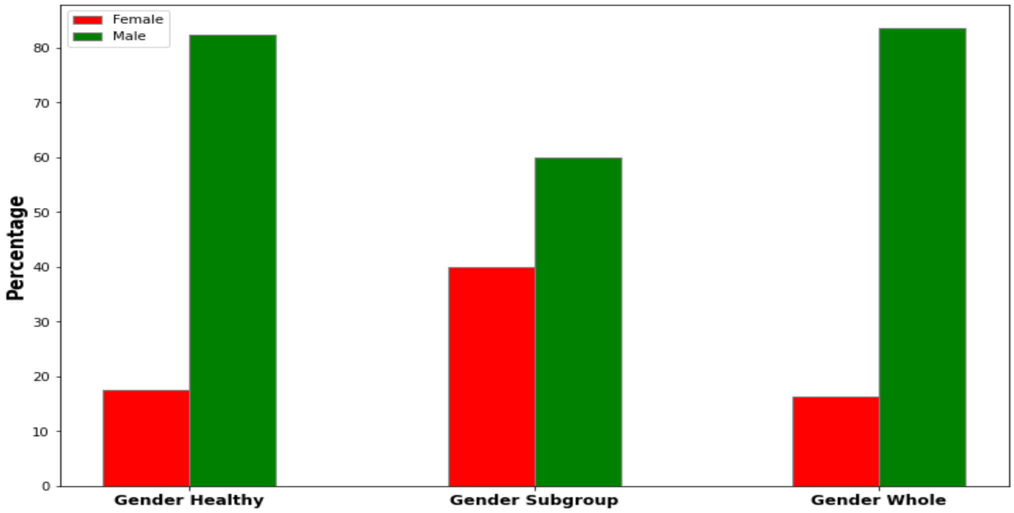


Figure 20: Gender Distribution in the Healthy, Subgroup3 and Whole Populations

5.3.3. Rule10

The 10th rule in our study implies that if a participant has an unknown activity duration between the median and maximum score of the whole and healthy population(between 4.5 and 6 hours per day), and has at least one hour of vigorous physical activity during the week, which is more than the weekly vigorous activity of 50% of both healthy and the whole population and their running duration per day is also more than 50% of the healthy and whole population, the probability of having CVD or its risk factor is remarkably low in this participant(6%)(Figure 21). This pattern is recognized in 371 items, and the usage is 352.

The distribution of the three datasets is almost the same for unknown_time_core and running_time_core attributes; however, the mean and median of subgroup10 are bigger than the other two groups for weekly vigorous activity duration. This subgroup also has a more dispersed distribution(Figure 22).

Regarding demographic attributes, subgroup 10 has a more scattered distribution for weight and height attributes in comparison to the two other data groups. The median height in this subgroup is lower than the two other data sets. The proportion of female participants is around 5% higher than the healthy and whole population. Concerning ethnicity, there is not any pacific Islander in this subgroup. The percentage of Asian people in this subgroup is less than the two other subgroups. In addition, Hispanic and Black ethnicities have a higher percentage(Figure 23).

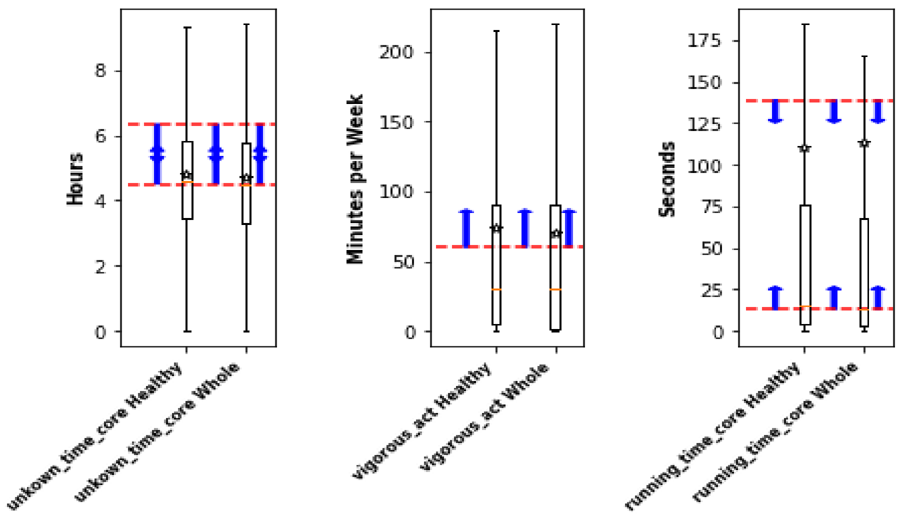


Figure 21: Attributes of Rule10

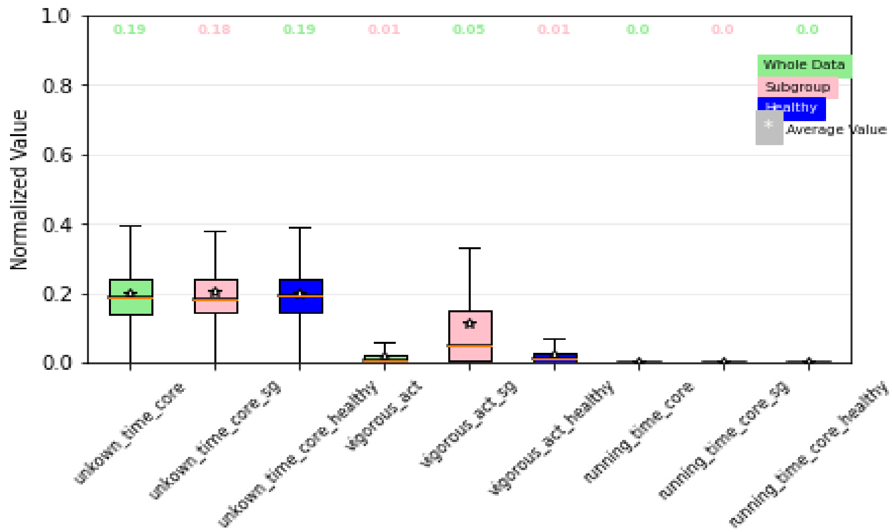
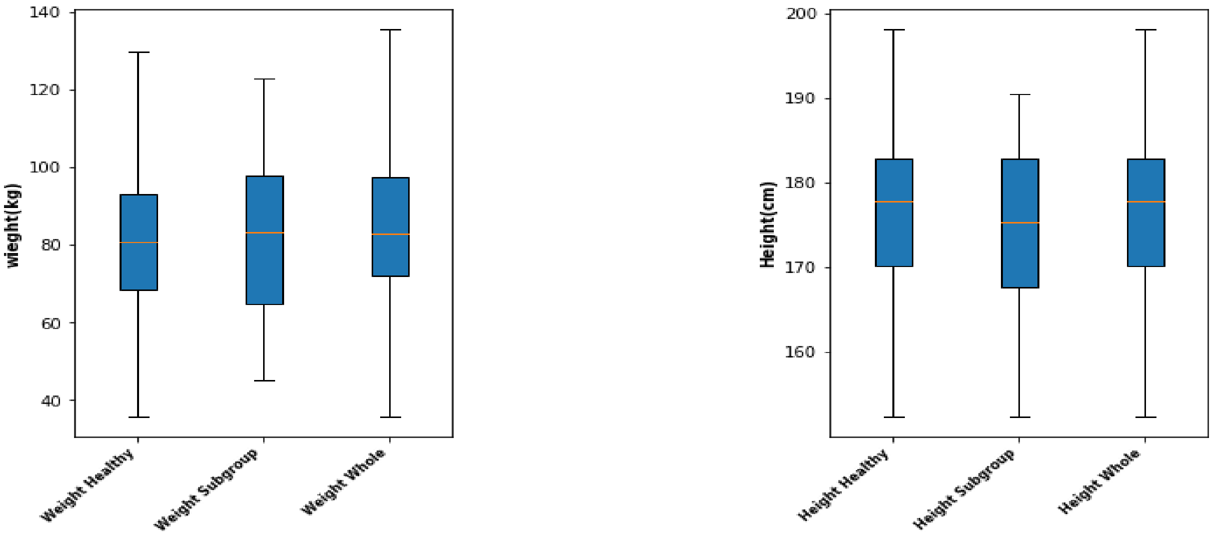
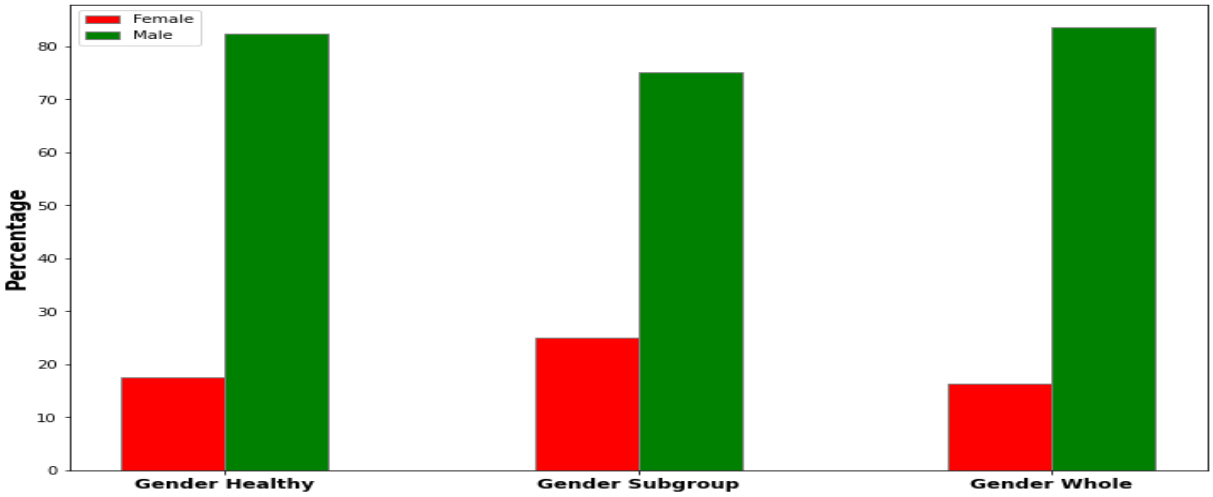


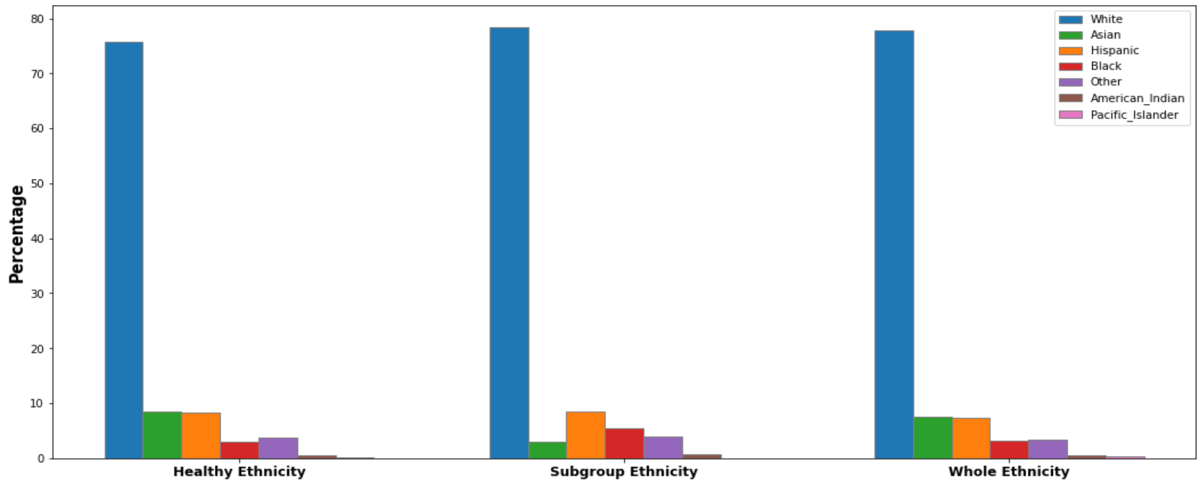
Figure 22: Distribution Comparison for Rule10



a. Weight and Height in Different Data Groups



b. Gender in Different Data Groups



c. Gender in Different Data Groups

Figure 23: Demographic Attributes of subgroup10

5.3.4. Rule11

Rule 11 includes four conditions. The first one is related to early morning(5-9 A.M.) activity duration per day being more than 36 minutes. It means more than the average and median of the whole and healthy population. The second constraint is about the duration of running per day being bigger than 13 seconds. This means the participant runs more than 50% of the whole population per day. The third condition concerns about late evening activity(9-11:59 P.M.) duration being more than one hour per day, meaning more than late evening activity of 25% of the whole and healthy population. The last condition is about walking more than 50% of the whole and healthy data population(50 minutes) per day. These conditions lead to a probability of 9% for having CVD or its risk factors(Figure 24).

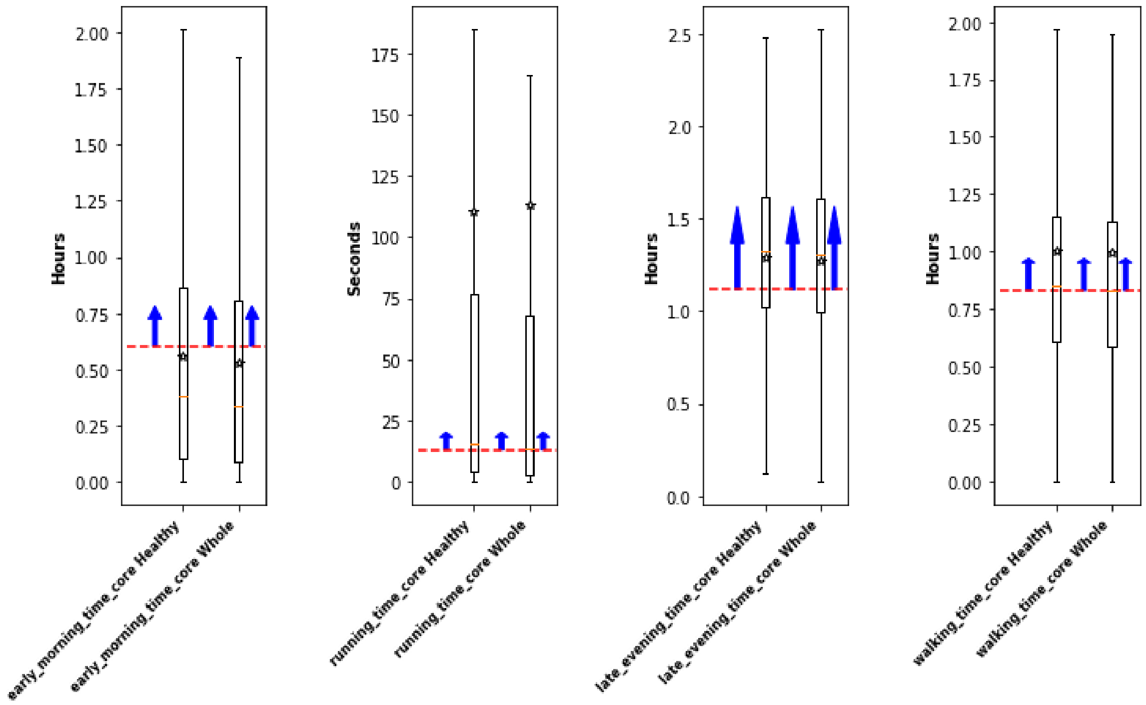


Figure 24: Attributes of Rule11

In relation to demographic attributes, the median for height and weight attributes for all three groups of data is almost the same. Subgroup 11 distribution is more scattered regarding weight and more skewed to the right concerning height.

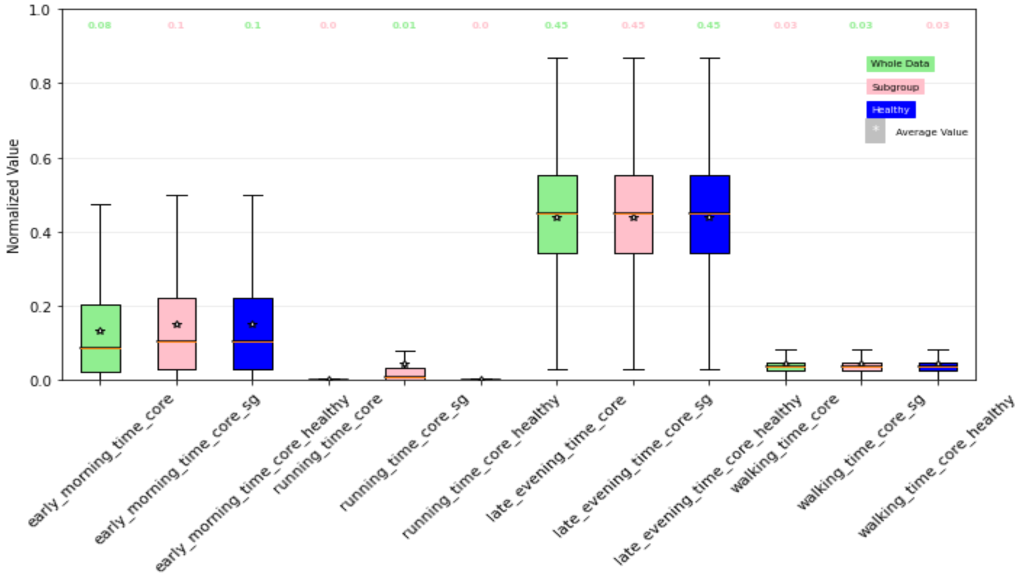


Figure 25: Distribution Comparison for Rule11

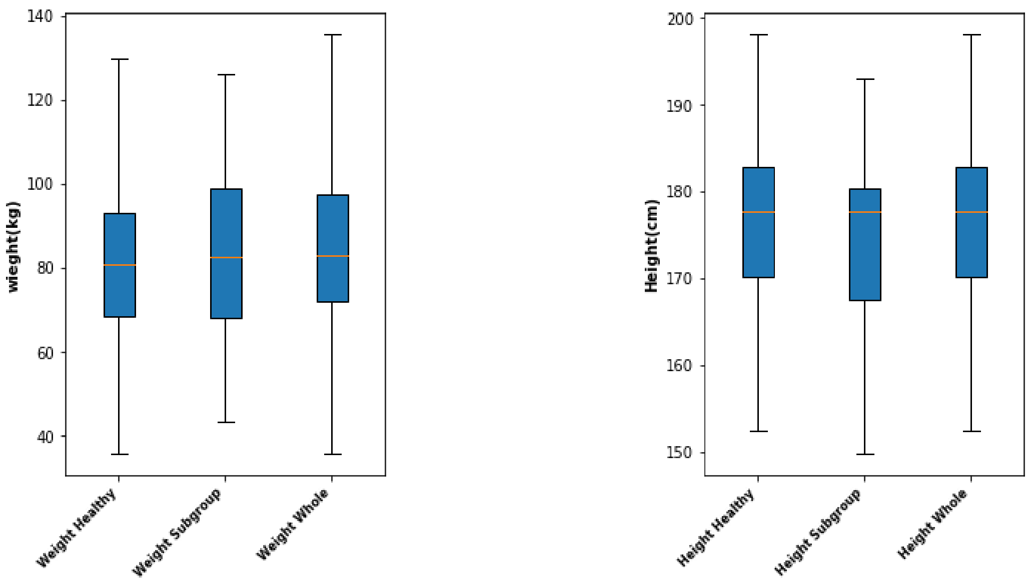


Figure 26: Distribution of the Height and Weight Attributes in Subgroup 11, the Healthy and Whole Populations

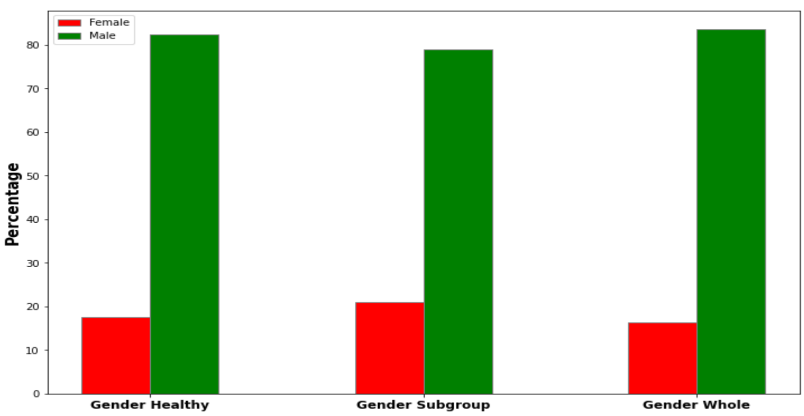


Figure 27: Gender Distribution in the Healthy, Subgroup11 and Whole Populations

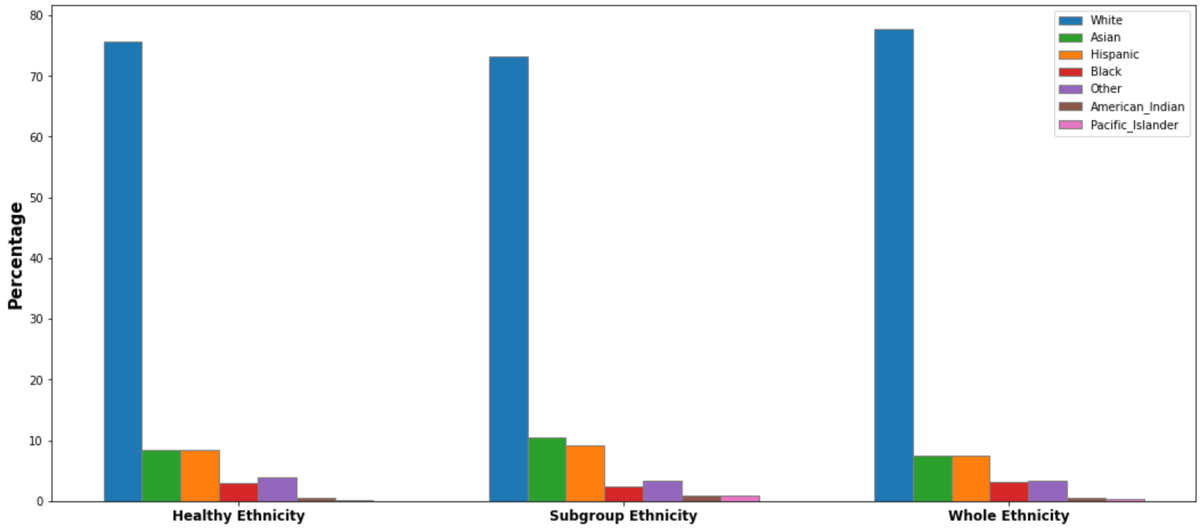


Figure 28: Ethnicity Distribution in the Healthy, Subgroup 11 and Whole Populations

5.3.5. Rule15

Rule 15 focuses on running and cycling duration. It implies that if a participant runs between 5 to 37 seconds each day, which is between the lower quartile and upper quartile of running time for both healthy and whole population and has a cycling time of more than nine minutes per day(more than lower quartile for both healthy and whole data groups) the probability of having CVD for that participant is around 16 percent. This is an interesting rule since it provides an upper bound for the running duration(*Figure 29*).

Concerning demographic attributes, this subgroup has a median of 80 kg for weight which is almost the same as the healthy population. The difference is the lower quartile of this subgroup is higher than the healthy population for this attribute, and it is more skewed to the left. The median(175 cm) for height is lower than both other datasets. 50% of the subgroup population is older than 40. Gender and Ethnicity are the attributes that have almost the same features in all three datasets (*Figure 30*).

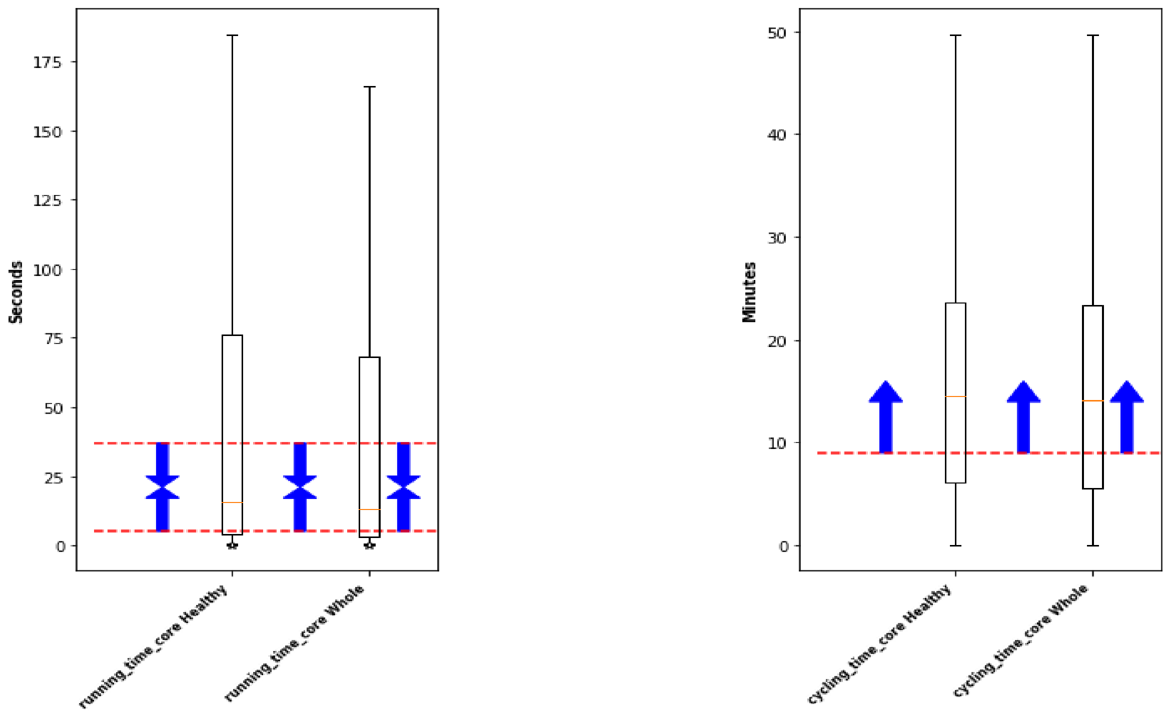


Figure 29: Numeric Attributes of Rule15

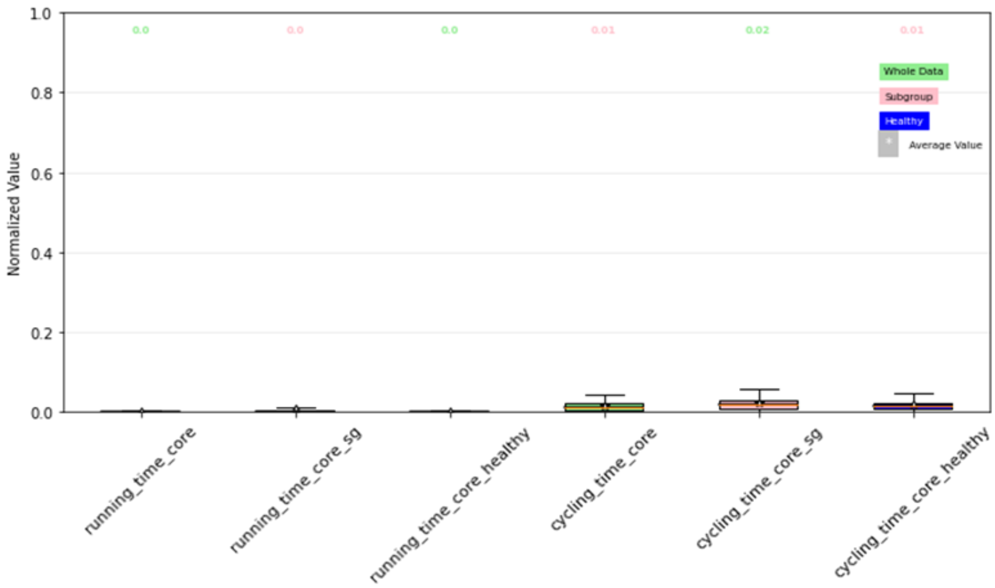
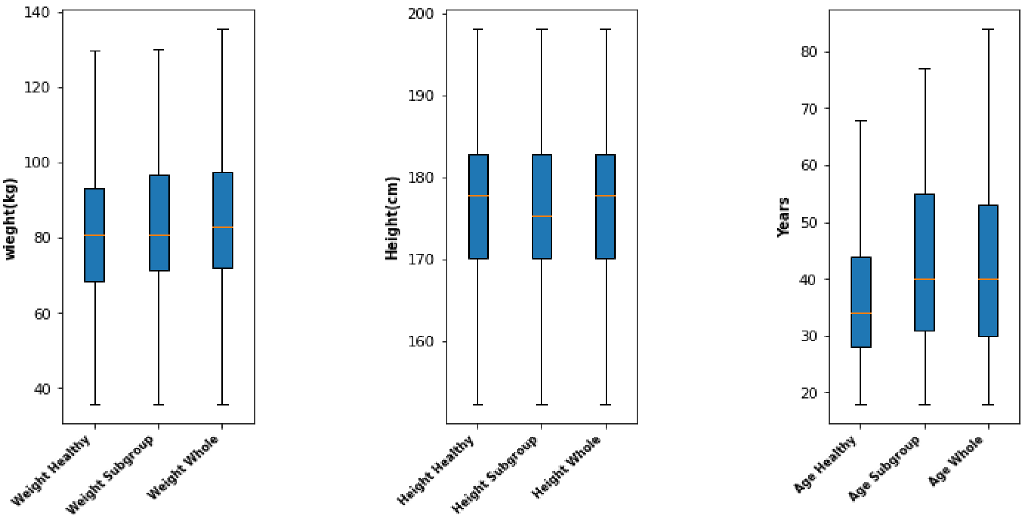
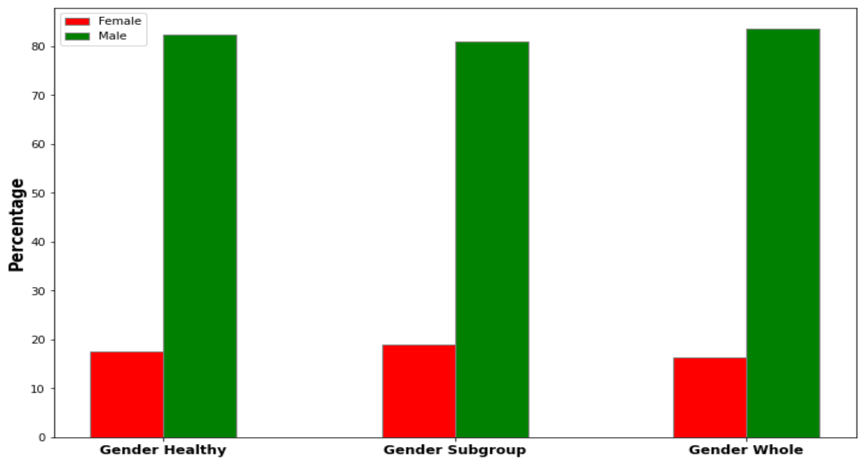


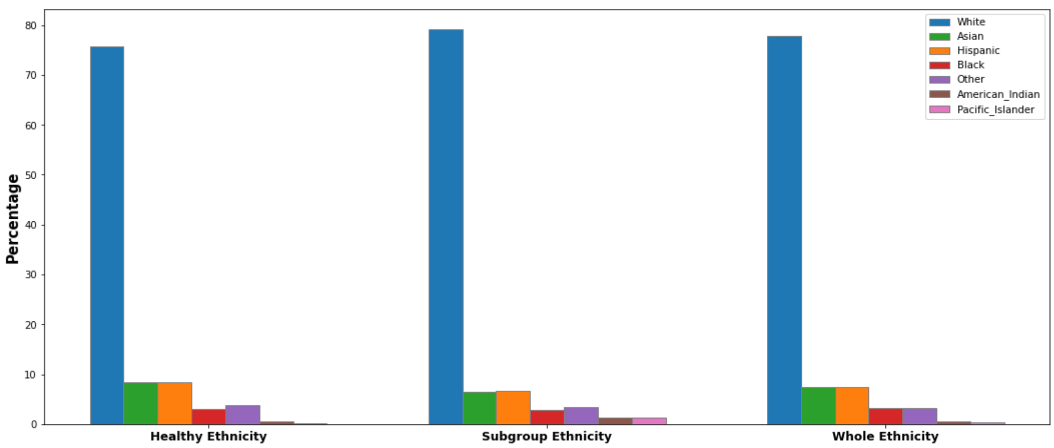
Figure 30: Distribution Comparison of Subgroup, Whole and Healthy Populations



a. Weight, Height and Age in Different Data Groups



b. Gender in Different Data Groups



c. Ethnicity in Different Data Groups

Figure 31: Demographic Attributes of subgroup15

5.4. Discussion and Future Works

After preprocessing the data by omitting non-related features, deleting noisy data, changing the format of some features, extracting new ones, and dealing with null values and time series data, we got our data set ready for trying the SSD++ subgroup discovery algorithm. Applying this algorithm to our dataset resulted in a subgroup list with 15 rules. By having the R-square equal to 78%, we can say that the variance of the target variable is properly described by these 15 rules in this model.

We can evaluate our model from two perspectives, local and global. local evaluation is related to comparing different rules and seeing how they are, considering the rest of the rules or how much they make sense based on available studies and our knowledge about CVD and its risk factors. For this purpose, we focused mainly on four measures: the probability of the rule, its usage, its support, and the WKL of the rule, which sums up the probability and usage of it. In addition, we compared the distribution of each subgroup and its demographic attributes with two datasets. One is our original dataset, and the other is a dataset with only healthy people in the study, meaning participants without any True label for our target variables. Moreover, we visualized what the rule implied based on the distribution of the whole and healthy population.

Regarding this evaluation approach, we saw that the extracted rules were in line with current knowledge of CVD and its risk factors. For example, the age attribute appeared in more than 50% of the rules. All these rules indicate a high chance of having cardiovascular disease in case of being older than a certain age or vice versa. Another example is rule 5 which is about a relation between the height of the participant and the probability of having CVD or its risk factors. [71], [72] are studies that focused on this relation. We also found a link between having more physical activity and a lower chance of having CVD or its risk factors in rule 3. This is in line with the findings in [12]. Rules 1, 4 and 11 have a condition emphasizing adverse relation between the duration of physical activity during early morning and the chance of having cardiovascular disease in complete agreement with the results of studies such as [12], [73], In [7], the authors found an adverse relation between evening activity and overall cardiovascular health(CVH). Rule 2 in our subgroup list consists of conditions leading to an 87% chance of having CVD or its risk factors. One of which is having late evening activity of more than 50 minutes per day. Rule 6 and rule 12 also have some upper bound for late evening activity duration. It is also interesting that the focus of [7] is on women, and in both subgroups 6 and 12 the proportion of female participants is larger than the whole and healthy population. The boundaries of the conditions have also been interesting since most of the time they were bigger or smaller than the mean or median of the healthy and whole population.

The second aspect of our evaluation, global evaluation, took place by calculating some subgroup discovery measures indicating how powerful our model was (Table 8). We also implemented our model for prediction on test data(part of the data that is not used in the process of training the model) which result in 76% accuracy. Next, we compared the prediction result using SSD++model with three classification models meaning Random Forest, Decision Tree and Naïve Bayes. Our model had a better performance in comparison to Naïve Bayes based on all measures (precision, recall and accuracy). The precision was also better than Decision Tree. The best model was Random Forest with 86% accuracy and a recall of 50%. This means classification in this dataset is a complicated task. Therefore, based on the performance of our model in such a complicated problem we can say that our model is not just a combination of some random rules.

Based on these evaluations, we can declare that our results are valid and therefore, worth considering. It means the conditions that seem unexpected at first glance are worth examination. This is actually the purpose of this study. To find astonishing relations that pave the way for future studies focusing on specific situations. Examples of this in our study are when we found an upper bound for running time in rule 10 or for the duration of physical activity in certain parts of the day; for instance, noon or late evening.

The novelty of this study is related to applying subgroup discovery for finding interesting relations between CVD and physical activity. This gave us the chance to find relations that are not being mentioned in previous studies and worth closer look. In addition, applying this approach made it possible to have a holistic view for answering our questions and be able to look at the problem from a different perspective. Moreover, using a smartphone based data set gave us the ability to have various variables and including different aspects and attributes regarding physical activity at once.

Even though, we got good results based on SSD++ algorithm, we only tested our model on one dataset. It is always beneficial to examine the machine learning models on multiple datasets to see how they work on a completely new space. In addition, it is true that the dataset we used here included different aspects but still it was pretty noisy and unbalanced. The majority of the participants in this dataset were healthy.

Therefore one aspect of future works can focus on examining interesting rules found in this study, especially the ones that there are not related study about them, in more detail on other datasets to see to which degree these outcomes are generalizable. Specifically, using a less noisy dataset that is more balanced regarding attributes such as gender, ethnicity and different health conditions.

In addition, the result of our algorithm is a list of ranked subgroups. In this algorithm, when one part of the data examined for one rule, it will not be considered for another subgroup generation. Therefore, for subgroup 2 onward, the rules evaluation can become more and more complicated since each rule is only generated based on the instances not being considered in prior patterns. Therefore, it will be interesting to look at each rule independently and see how it will work if it is the first rule.

Moreover, our target variable is both based on the participants declaration about having specific diseases and extracting this information based on medical measurements entered. Therefore, we did not differentiate between participants who knew about having CVD and who did not know about it. However, awareness of having CVD can affect participants behavior. This is another aspect that can be investigated in future studies.

Chapter 6

Conclusion

This study explored the existence of patterns between different aspects of physical activity (timing, intensity, duration, etc.) and having (risk factors for) cardiovascular disease. To this extent, the SSD++ algorithm, a subgroup discovery technique, was used on the My Heart Counts USA dataset, including data from up to 50,000 users from the USA who joined one of the first remotely conducted medical trials in 2015. The subgroup discovery resulted in a list of 15 different subgroups, each indicating one interesting rule found based on the dataset. In 13 out of 15 rules, there was at least one condition regarding the duration of physical activity in a specific part of the day. We evaluated our outputs from two aspects. One is comparing rules with each other, locally by their probability, usage and WKL. We also compared our results with the state-of-the art knowledge about CVD. We found relations that were mentioned in previous studies such as the relation between age and CVD, morning physical activity and CVD, and afternoon physical activity and CVD. For global evaluation we looked at the problem as a classification problem and used our model for prediction on a test set. This assessment demonstrated the complication of the problem since the best accuracy gained was 86% based on the Random Forest model. It also revealed the power of our model in understanding the dataset by having 76% accuracy. This alignment with previous studies and showing comparable performance in prediction with classical classification algorithms showed us that this model is reliable, and accordingly, the relations found without any related studies about them are worth examination in more detail. To conclude, we think that there is a huge potential in analyzing and modeling the medical datasets; there is so much to learn and explore. Further connecting and understanding these two 'worlds' would be very valuable and this study was one example of this endless opportunities.

Bibliography

[1] (2021) WHO fact sheets. [online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] M. J. G. Leening, S. Siregar, I. Vaartjes, M.L. Bots, M. I. M. Versteegh, et al., “Heart disease in the Netherlands: A quantitative update.” Netherlands Heart Journal, vol. 22, no. 1. Bohn Stafleu van Loghum, pp. 3–10, Jan. 01, 2014. doi: 10.1007/s12471-013-0504-x.

[3] G. David Batty, “Physical activity and coronary heart disease in older adults A systematic review of epidemiological studies.” [Online]. Available: https://academic.oup.com/eurpub/article/12/3/171/497847

[4] D. Verdaet, P. Dendale, D. de Bacquer, J. Delanghe, P. Block, et al., “Association between leisure time physical activity and markers of chronic inflammation related to coronary heart disease.” Atherosclerosis, vol. 176, no. 2, pp. 303–310, Oct. 2004, doi: 10.1016/j.atherosclerosis.2004.05.007.

[5] C. B. Eaton, “Relation Of Physical Activity And Cardiovascular Fitness To Coronary Heart Disease, Part I: A Meta-Analysis Of The Independent Relation Of Physical Activity And Coronary Heart Disease”, doi: 10.3122/jabfm.5.1.31.

[6] C. B. Eaton and B. Eaton, “Relation Of Physical Activity And Cardiovascular Fitness To Coronary Heart Disease, Part II: Cardiovascular Fitness And The Safety And Efficacy Of Physical Activity Prescription”, doi: 10.3122/jabfm.5.2.157.

[7] N. Makarem, J. Paul, E. G. v. Giardina, M. Liao, and B. Aggarwal, “Evening chronotype is associated with poor cardiovascular health and adverse health behaviors in a diverse population of women.” Chronobiol Int, vol. 37, no. 5, pp. 673–685, May 2020, doi: 10.1080/07420528.2020.1732403.

[8] M. Savikj, M. B. Gabriel, S. P. Alm, J. Smith, K. Caidahl, et al., “Afternoon exercise is more efficacious than morning exercise at improving blood glucose levels in individuals with type 2 diabetes: a randomised crossover trial.” Diabetologia, vol. 62, no. 2, pp. 233–237, Feb. 2019, doi: 10.1007/s00125-018-4767-z.

[9] S. Sato, K. A. Dyar, T. J. Treebak, S. L. Jepsen, A. M. Ehrlich, et al., “Atlas of exercise metabolism reveals time-dependent signatures of metabolic homeostasis.” Cell Metab, vol. 34, no. 2, pp. 329-345.e8, Feb. 2022, doi: 10.1016/j.cmet.2021.12.016.

[10] G. B. Ehret, P. B. Munroe, K. M. Rice, M. Bochud, A. D. Johnson, et al., “Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk.” Nature, vol. 478, no. 7367, pp. 103–109, Oct. 2011, doi: 10.1038/nature10405.

[11] R. Mancilla, B. Brouwers, V. B. Schrauwen-Hinderling, M. K. C. Hesselink, J. Hoeks, et al., “Exercise training elicits superior metabolic effects when performed in the afternoon compared to morning in metabolically compromised humans.” Physiol Rep, vol. 8, no. 24, Jan. 2021, doi: 10.14814/phy2.14669.

[12] G. Albalak, M. Stijntjes, D. van Bodegom, J. W. Jukema, D. E. Atsma, et al., “Setting your clock: associations between timing of objective physical activity and cardiovascular disease risk in the general population.” Eur J Prev Cardiol, Nov. 2022, doi: 10.1093/eurjpc/zwac239.

[13] J. H. P. M. van der Velde, S. C. Boone, E. Winters-van Eekelen, M. K. C. Hesselink, V. B. Schrauwen-Hinderling, et al., “Timing of physical activity in relation to liver fat content and insulin resistance.” Diabetologia, 2022, doi: 10.1007/s00125-022-05813-3.

[14] S. G. Hershman, B.M. Bot, A. Schcherbina, M. Doerr, Y. Moayedi, et al., “Physical activity, sleep and cardiovascular health data for 50,000 individuals from the MyHeart Counts Study Background and Summary”, doi: 10.1038/s41597-019-0016-7.

[15] M. v McConnell et al., “Feasibility of Obtaining Measures of Lifestyle From a Smartphone App The MyHeart Counts Cardiovascular Health Study.” JAMA Cardiol, vol. 2, no. 1, pp. 67–76, 2017, doi: 10.1001/jamacardio.2016.4395.

[16] W. Klösgen, “Explora: A Multipattern and Multistrategy Discovery Assistant.” Advances in Knowledge Discovery and Data Mining, 1996.

[17] M. Scholz, "Knowledge-based sampling for subgroup discovery." In Local Pattern Detection, pp. 171-189. Springer, Berlin, Heidelberg, 2005.

[18] H. M. Proença, P. Grünwald, T. Bäck, and M. van Leeuwen, “Robust subgroup discovery.” Mar. 2021, doi: 10.1007/s10618-022-00856-x.

[19] H. M. Proença, "Robust rules for prediction and description." PhD diss., PhD thesis, Leiden University, 2021.

[20] D. N. Gunjate, and B. R. Kanawade, "Subgroup Discovery a Data Mining Technique: Immense Survey."

[21] M. Atzmueller, “Subgroup discovery.” WIREs Data Mining Knowl Discov, vol. 5, pp. 35–49, 2015, doi: 10.1002/widm.1144.

[22] P. Flach, N. Lavrač, B, Kavšek, L. Todorovski, “Subgroup Discovery with CN2-SD.” J. Mach. Learn. Res. 5, no. 2, pp. 153-188, 2004.

[23] M. van Leeuwen and A. Knobbe, “Diverse subgroup set discovery,” in Data Mining and Knowledge Discovery, Sep. 2012, vol. 25, no. 2, pp. 208–242. doi: 10.1007/s10618-012-0273-y.

[24] G. Bosc, J. F. Boulicaut, C. Raïssi, and M. Kaytoue.“Anytime discovery of a diverse set of patterns with Monte Carlo tree search,” Data Min Knowl Discov, vol. 32, no. 3, pp. 604–650, May 2018, doi: 10.1007/s10618-017-0547-5.

[25] A. Belfodil et al., “FSSD - A fast and efficient algorithm for subgroup set discovery.” in Proceedings - 2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2019, Oct. 2019, pp. 91–99. doi: 10.1109/DSAA.2019.00023.

[26] S. N. Blair, H. W. Kohl, N. F. Gordon, and R. S. Paffen, “HOW MUCH PHYSICAL ACTIVITY IS GOOD FOR HEALTH?” 1992. [Online]. Available: www.annualreviews.org

[27] P. Kokkinos, “Physical Activity, Health Benefits, and Mortality Risk.” ISRN Cardiol, vol. 2012, pp. 1–14, Oct. 2012, doi: 10.5402/2012/718789.

[28] D. Macauley, “A history of physical activity, health and medicine.” Journal of the Royal Society of Medicine 87, no. 1 (1994): 32.

[29] W. L. Haskell, S. N. Blair, and J. O. Hill, “Physical activity: Health outcomes and importance for public health policy.” Preventive Medicine, vol. 49, no. 4. pp. 280–282, Oct. 2009. doi: 10.1016/j.ypmed.2009.05.002.

[30] S. Banerjee, P. Kumar, S. Srivastava, and A. Banerjee, “Association of anthropometric measures of obesity and physical activity with cardiovascular diseases among older adults: Evidence from a cross-sectional survey, 2017–18.” PLoS One, vol. 16, no. 12, Dec. 2021, doi: 10.1371/journal.pone.0260148.

[31] C. C. Cesa et al., “Physical activity and cardiovascular risk factors in children: Meta-analysis of randomized clinical trials.” Preventive Medicine, vol. 69. Academic Press Inc., pp. 54–62, Dec. 01, 2014. doi: 10.1016/j.ypmed.2014.08.014.

[32] F. B. Hu et al., “Physical Activity and Risk for Cardiovascular Events in Diabetic Women Background: Increased physical activity has been associated.” 2001. [Online]. Available: https://annals.org

[33] N. Shields, J. Hussey, J. Murphy, J. Gormley, and H. Hoey, “An exploratory study of the association between physical activity, cardiovascular fitness and body size in children with Down syndrome.” Dev Neurorehabil, vol. 20, no. 2, pp. 92–98, Feb. 2017, doi: 10.3109/17518423.2015.1077901.

[34] Y. Oguma and T. Shinoda-Tagawa, “Physical activity decreases cardiovascular disease risk in women: Review and meta-analysis.” Am J Prev Med, vol. 26, no. 5, pp. 407–418, 2004, doi: 10.1016/j.amepre.2004.02.007.

[35] P. Jousilahti, E. C. Barengo, Q. Qiao, T. A. Lakka, and J. Tuomilehto, “Physical Activity, Cardiovascular Risk Factors, and Mortality Among Finnish Adults With Diabetes.” 2005. [Online]. Available: http://diabetesjournals.org/care/article-pdf/28/4/799/566336/zdc00405000799.pdf

[36] S. Savela et al., “Leisure-time physical activity, cardiovascular risk factors and mortality during a 34-year follow-up in men.” Eur J Epidemiol, vol. 25, no. 9, pp. 619–625, Sep. 2010, doi: 10.1007/s10654-010-9483-z.

[37] J. F. Sallis, W. L. Haskell, P. D. Wood, S. P. Fortmann, and K. M. Vranizan.“ VIGOROUS PHYSICAL ACTIVITY AND CARDIOVASCULAR RISK FACTORS IN YOUNG ADULTS” 1986.

[38] S. Sato et al., “Time of Exercise Specifies the Impact on Muscle Metabolic Pathways and Systemic Energy Homeostasis.” Cell Metab, vol. 30, no. 1, pp. 92-110.e4, Jul. 2019, doi: 10.1016/j.cmet.2019.03.013.

[39] P. J. Arciero et al., “Morning Exercise Reduces Abdominal Fat and Blood Pressure in Women; Evening Exercise Increases Muscular Performance in Women and Lowers Blood Pressure in Men.” Article, vol. 13, p. 1, 2022, doi: 10.3389/fphys.2022.893783.

[40] H. M. Proença, R. Klijn, T. Bäck, and M. van Leeuwen, “Identifying flight delay patterns using diverse subgroup discovery.” In 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 60-67. IEEE, 2018.

[41] R. M. , D. W. , K. W. and K. A. Konijn, “Discovering local subgroups, with an application to fraud detection.” Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 1–12, 2013.

[42] M. Mueller, R. Rosales, H. Steck, S. Krishnan, B. Rao, and S. Kramer, “Subgroup Discovery for Test Selection: A Novel Approach and Its Application to Breast Cancer Diagnosis.” In International Symposium on Intelligent Data Analysis, pp. 119-130. Springer, Berlin, Heidelberg, 2009.

[43] J. Schmidt et al., “Interpreting PET scans by structured patient data: A data mining case study in dementia research,” Knowl Inf Syst, vol. 24, no. 1, pp. 149–170, 2010, doi: 10.1007/s10115-009-0234-y.

[44] N. Lavrač, “Subgroup discovery techniques and applications.” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2005, vol. 3518 LNAI, pp. 2–14. doi: 10.1007/11430919_2.

[45] C. J. Carmona, P. González, M. J. del Jesus, M. Navío-Acosta, and L. Jiménez-Trevino, “Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department.” Soft comput, vol. 15, no. 12, pp. 2435–2448, Dec. 2011, doi: 10.1007/s00500-010-0670-3.

[46] C. Esnault, M. L. Gadonna, M. Queyrel, A. Templier, and J. D. Zucker, “Q-Finder: An Algorithm for Credible Subgroup Discovery in Clinical Data Analysis — An Application to the International Diabetes Management Practice Study.” Front Artif Intell, vol. 3, Dec. 2020, doi: 10.3389/frai.2020.559927.

[47] C. J. Carmona et al., “A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans.” Inf Sci (N Y), vol. 298, pp. 180–197, Mar. 2015, doi: 10.1016/j.ins.2014.11.030.

[48] J. V. Park, S. J. Park, and J. S. Yoo, “Finding characteristics of exceptional breast cancer subpopulations using subgroup mining and statistical test.” Expert Syst Appl, vol. 118, pp. 553–562, Mar. 2019, doi: 10.1016/J.ESWA.2018.10.016.

[49] D. Gamberger, N. Lavrač, and G. Krstačić, “Active subgroup mining: a case study in coronary heart disease risk group detection.” Artif Intell Med, vol. 28, no. 1, pp. 27–57, May 2003, doi: 10.1016/S0933-3657(03)00034-4.

[50] D. Gamberger et al., “Clinical data analysis based on iterative subgroup discovery: experiments in brain ischaemia data analysis.” Appl Intell, vol. 27, pp. 205–217, 2007, doi: 10.1007/s10489-007-0068-9.

[51] S. Helal, “Subgroup discovery algorithms: A survey and empirical evaluation.” J Comput Sci Technol, vol. 31, no. 3, pp. 561–576, 2016, doi: 10.1007/s11390-016-1647-1.

[52] F. Herrera, · Cristóbal, J. Carmona, P. González, · María, and et. al, “An overview on subgroup discovery: foundations and applications.” Knowl Inf Syst, vol. 29, pp. 495–525, 2011, doi: 10.1007/s10115-010-0356-2.

[53] N. Lavrač, P. Flach, and B. Zupan, “Rule evaluation measures: A unifying view.” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1999, vol. 1634, pp. 174–185. doi: 10.1007/3-540-48751-4_17.

[54] M. van Leeuwen and A. Knobbe, “Non-redundant Subgroup Discovery in Large and Complex Data.” In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 459-474. Springer, Berlin, Heidelberg, 2011.

[55] S. Wrobel GMD and S. Birlinghoven, “An Algorithm for Multi-relational Discovery of Subgroups.” In European symposium on principles of data mining and knowledge discovery, pp. 78-87. Springer, Berlin, Heidelberg, 1997.

[56] A. Siebes, “Data Surveying Foundations of an Inductive Query Language.” 1995. [Online]. Available: www.aaai.org

[57] D. Gamberger, N. Lavrac, “Expert-guided subgroup discovery: Methodology and application.” Journal of Artificial Intelligence Research, vol. 17, pp.501-527, 2002

[58] B. Kavšek and N. Lavrač, “APRIORI-SD: Adapting association rule learning to subgroup discovery.” Applied Artificial Intelligence, vol. 20, no. 7, pp. 543–583, Sep. 2006, doi: 10.1080/08839510600779688.

[59] M. Atzmueller and F. Puppe, “SD-Map-A Fast Algorithm for Exhaustive Subgroup Discovery.” In European Conference on Principles of Data Mining and Knowledge Discovery, pp. 6-17. Springer, Berlin, Heidelberg, 2006.

[60] D. B. F. Z. M. Thomas Baeck, “Handbook of Evolutionary Computation”. Release 97, no. 1,1997.

[61] M. J. del Jesus, P. González, F. Herrera, and M. Mesonero, “Evolutionary fuzzy rule induction process for subgroup discovery: A case study in marketing.” IEEE Transactions on Fuzzy Systems, vol. 15, no. 4, pp. 578–592, Aug. 2007, doi: 10.1109/TFUZZ.2006.890662.

[62] M. van Leeuwen and A. Ukkonen, “Expect the Unexpected-On the Significance of Subgroups.” In International Conference on Discovery Science, pp. 51-66. Springer, Cham, 2016.

[63] J. Rissanent, “Modeling By Shortest Data Description.” Automatica, vol. 14, no. 5, pp. 465-471, 1978

[64] A. Knobbe, B. Crémilleux, J. Fürnkranz, and M. Scholz, “From Local Patterns to Global Models: The LeGo Approach to Data Mining.”2007

[65] J. F. Urnkranz, “Separate-and-Conquer Rule Learning.” Artificial Intelligence Review, vol. 13, no. 1, pp. 3-54, 1999.

[66] dHealth. [online]. Available: “https://dhealth.synapse.org/.”

[67] D. J. Stekhoven and P. Bühlmann, “Missforest-Non-parametric missing value imputation for mixed-type data,” Bioinformatics, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/bioinformatics/btr597.

[68] Apple Developer. [online]. Available: “https://developer.apple.com/documentation/healthkit/hkworkoutactivitytype”

[69] Apple Developer. [online]. Available: “https://developer.apple.com/documentation/coremotion”

[70] P. L. Enright, “The Six-Minute Walk Test.” Respiratory care, vol. 48, no. 8, pp. 783-785, 2003.

[71] J. W. Rich-Edwards et al., “Height and the Risk of Cardiovascular Disease in Women.” 1995. [Online]. Available: https://academic.oup.com/aje/article/142/9/909/88229

[72] P. R. Hebert et al., “Height and Incidence of Cardiovascular Disease in Male Physicians.” [Online]. Available: http://ahajournals.org

[73] J. Qian et al., “Association of Objectively Measured Timing of Physical Activity Bouts With Cardiovascular Health in Type 2 Diabetes.” Diabetes Care, vol. 44, no. 4, pp. 1046–1054, Apr. 2021, doi: 10.2337/DC20-2178.

Appendices

A. Data Appendix

Information regarding the original questions and variables in each table of this section is based on MyHeart Counts Public Researcher Portal available at:
https://www.synapse.org/#!Synapse:syn11269541/wiki/588018

A.1. Activity and Sleep Survey Table

Table A. 1: Activity and Sleep Table Attributes

Column Name	Question	Answers and statistics
work	Do you do regular work?	True (85.2%) False(14.8%)
atwork	Work Time Activity	1: I spent most of the day sitting or standing(64.3%) 2: I spent most of the day walking or using my hands and arms in work that required moderate exertion(17.3%) 3: I spent most of the day lifting or carrying heavy objects or moving most of my body in some other way (2.3%) 4: I spent most of the day doing hard physical labor (0.5%) None: 15.5%
phys_activity	Leisure Time Activity	1: I did not do much physical activity (15.57%) 2: Once or twice a week, I did light activities (28.26%) 3: About three times a week, I did moderate activities (22.85%)

		4: Almost daily, that is five or more times a week, I did moderate activities (13.55%) 5: About three times a week, I did vigorous activities (11.7%) 6: Almost daily, that is, five or more times a week, I did vigorous activities (7.75%) None: 0.3%
moderate_act	Overall, how many minutes of moderate activity do you get in a week?	count 21570 mean 148.38 std 215.25 min 0 25% 40 50% 90 75% 180 max 4096
vigorous_act	Overall, how many minutes of vigorous activity do you get in a week?	count 21570 mean 70.59 std 130.81 min 0 25% 2 50% 30 75% 90 max 3600
sleep_time1	How much sleep do you usually get at night on weekdays or workdays?	Count: 22841 Mean: 6.88 Std: 1.17 Min: 0 25%: 6

		50%: 7 75%: 8 Max: 15
sleep_time	How much sleep do think you need every night to be rested? (in hours)	Count: 22841 Mean: 7.77 Std: 1.13 Min: 0 25%: 7 50%: 8 75%: 8 Max: 15
sleep_diagnosis1	Have you ever been told by a doctor or other health professional that you have a sleep disorder?	True: 11.07% False: 88.92%
Extracted Attributes		
mostly_sit_stand	Whether the user chose the first option in ‘atwork’ section	True: 64.33% False: 35.67%
mostly_walk	Whether the user chose the second option in ‘atwork’ section	True: 17.30% False: 82.70%
mostly_lift	Whether the user chose the third	True: 2.34% False: 97.66%

	option in 'atwork' section	
hard_physical_activity	Whether the user chose the fourth option in 'atwork' section	True: 0.51% False: 99.49%
not_much_physical_activity	Whether the user chose the first option in 'phys_activity' section	True: 15.57% False: 84.43%
once_or_twice_physical_activity	Whether the user chose the second option in 'phys_activity' section	True: 28.26 % False: 71.74 %
three_times_physical_activity	Whether the user chose the third option in 'phys_activity' section	True: 22.85% False: 77.15%
daily_physical_activity	Whether the user chose the fourth option in 'phys_activity' section	True: 13.55 % False: 86.45%
three_times_vigorous_activity	Whether the user chose the fifth	True: 11.71% False: 88.29 %

	option in 'phys_activity' section	
daily_vigorous_activity	Whether the user chose the sixth option in 'phys_activity' section	True: 7.75% False: 92.25%

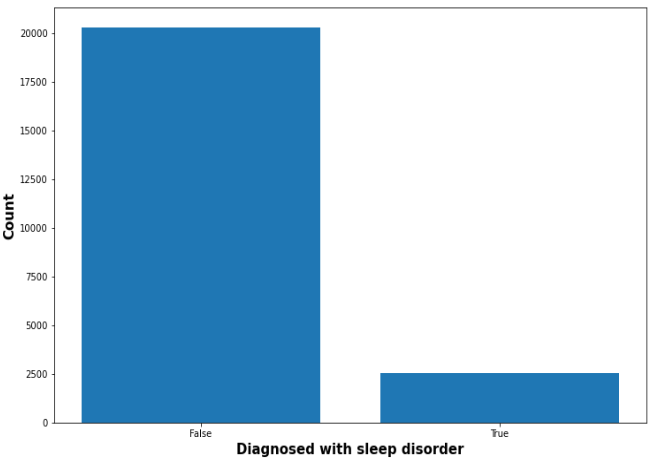


fig.a

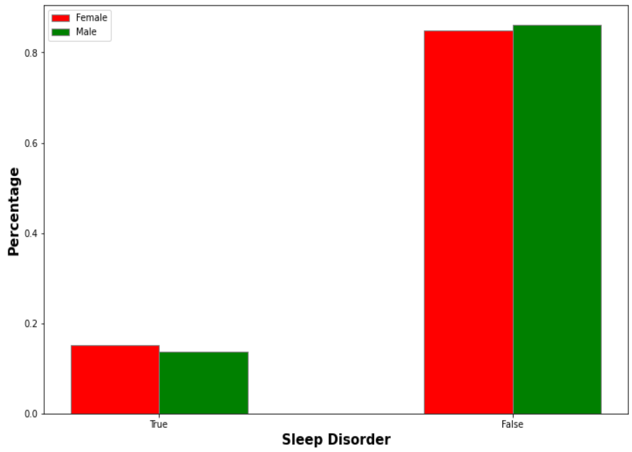


fig.b

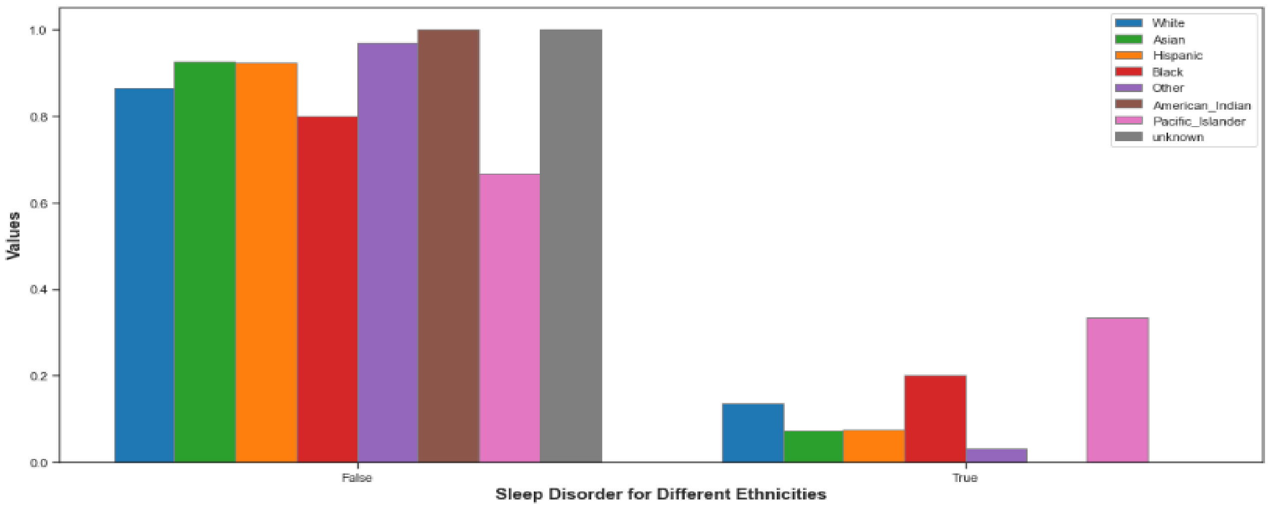


fig.c

Figure A. 1: a: Counts of Participant with and without Sleep Disorder. b: Percentage of Men and Women with and without Sleep Disorder. c: Percentage of Sleep Disorder among Different Ethnicities

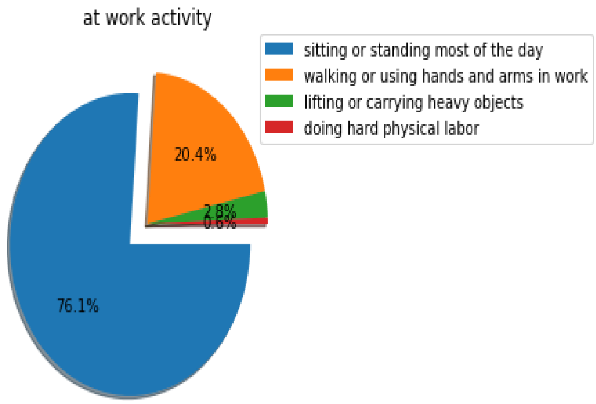


fig.a

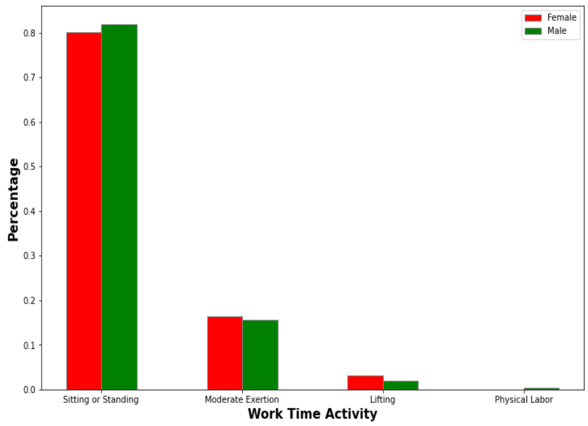


fig.b

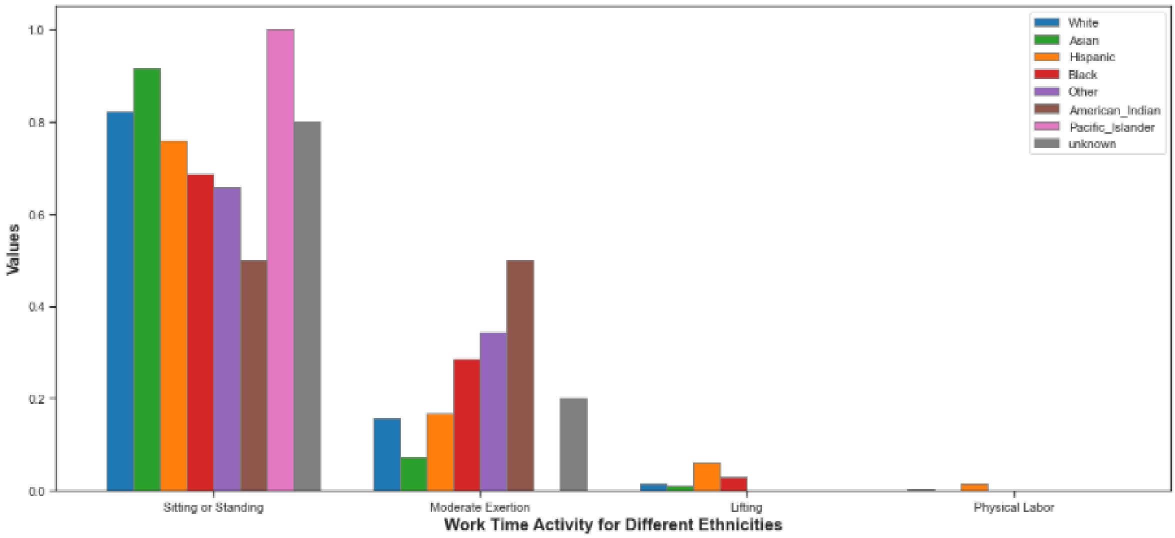


fig.c

Figure A. 2: a: At Work Physical Activity of Participants. b: At Work Physical Activity for Men and Women. c: At Work Physical Activity for Different Ethnicities

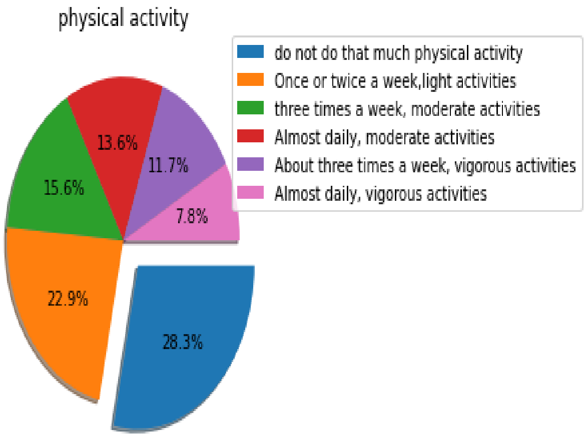


fig. a

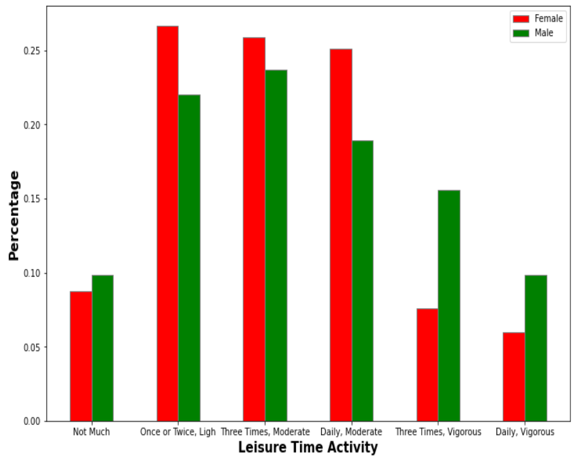


fig. b

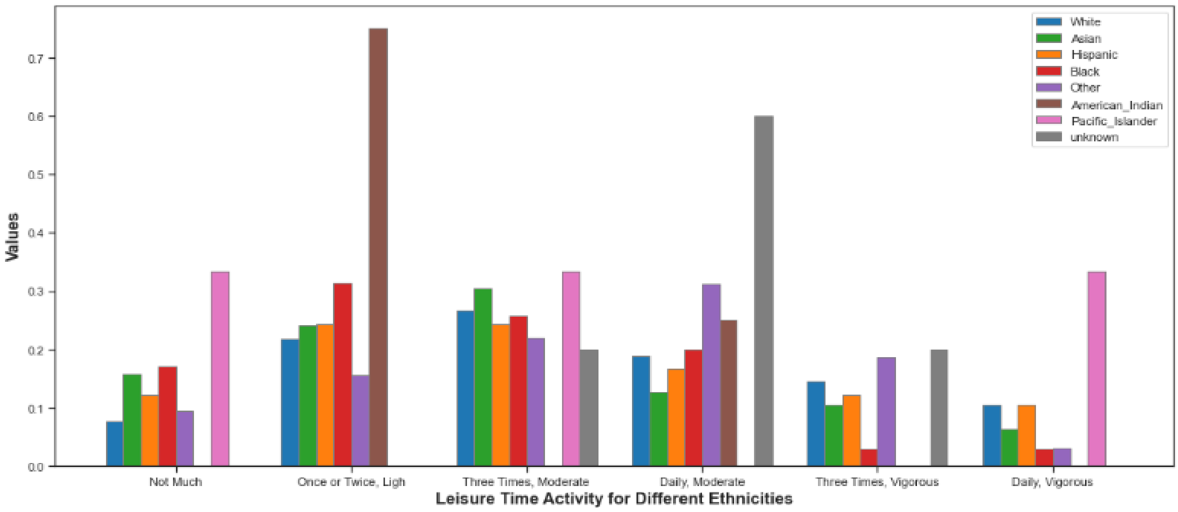


fig. c

Figure A. 3: a: Amount of Physical Activity. b: Physical Activity in Men and Women. c: Physical Activity in Dfferent Ethnicity Groups

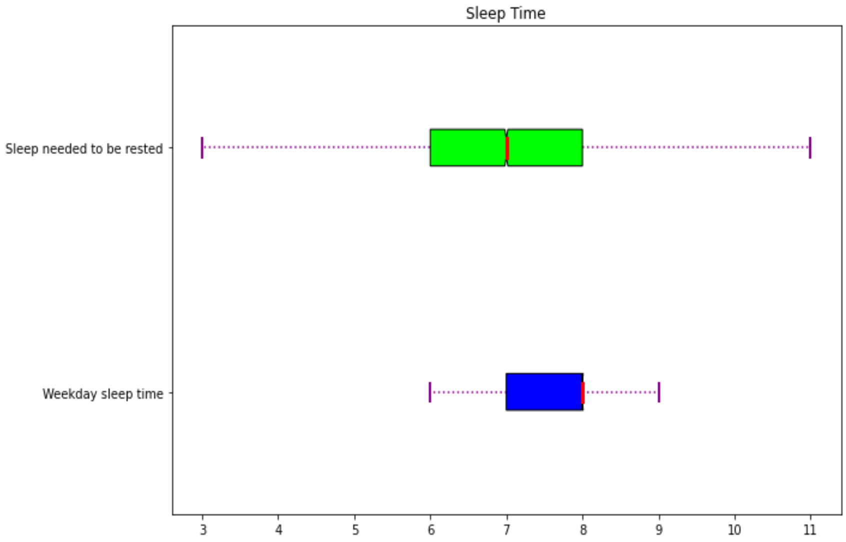


Figure A. 4: Sleep duration

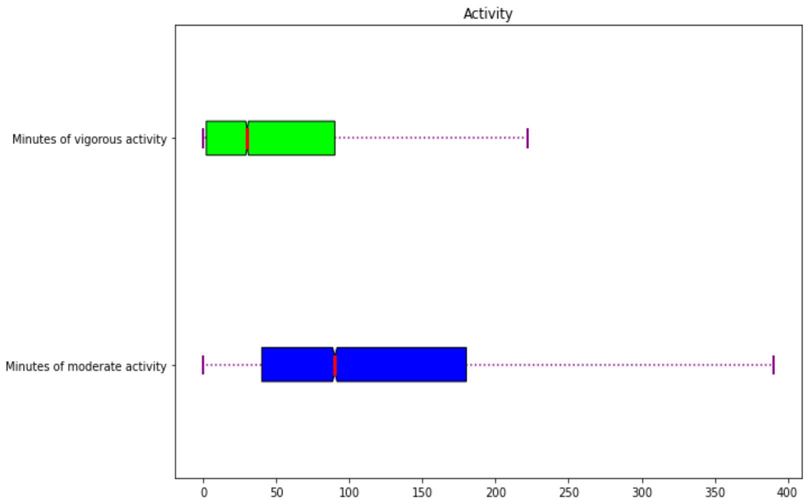


Figure A. 5: Moderate and vigorous physical activity duration

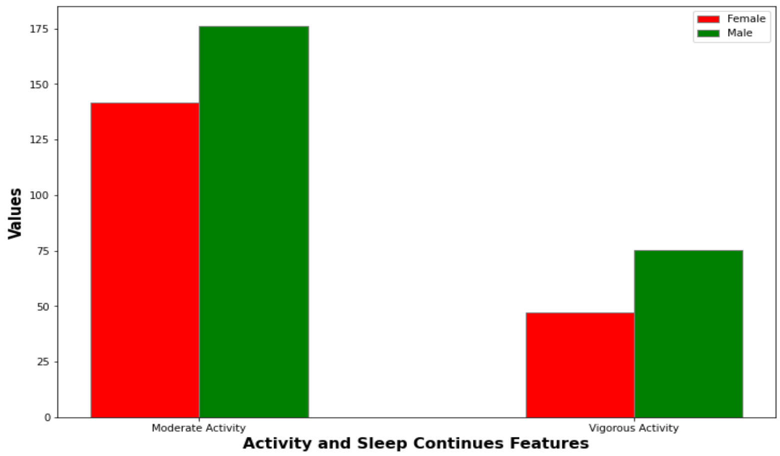


Figure A. 6: Moderate and vigorous physical activity amount between men and women

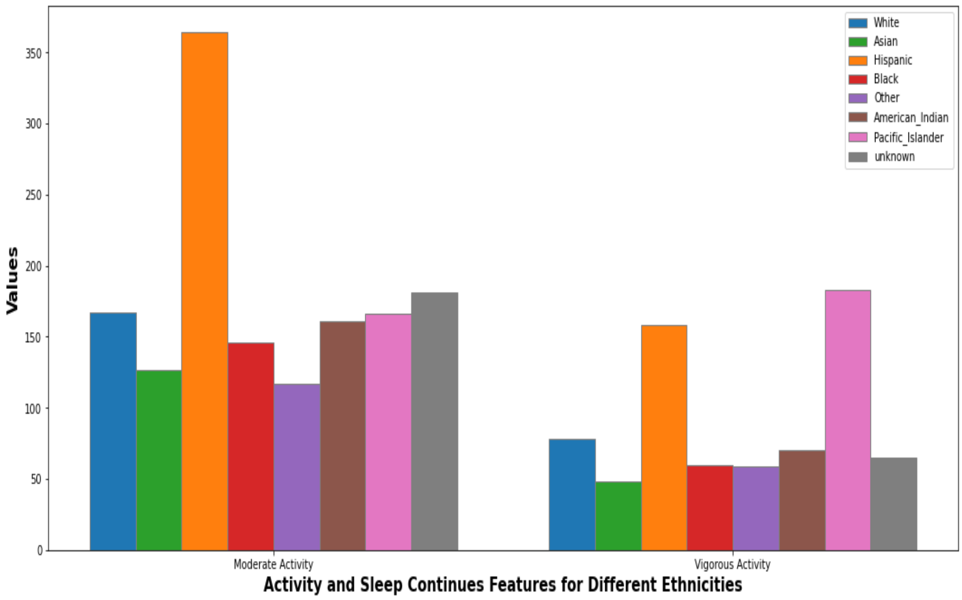


Figure A. 7: Moderate and vigorous physical activity amount among different ethnicities

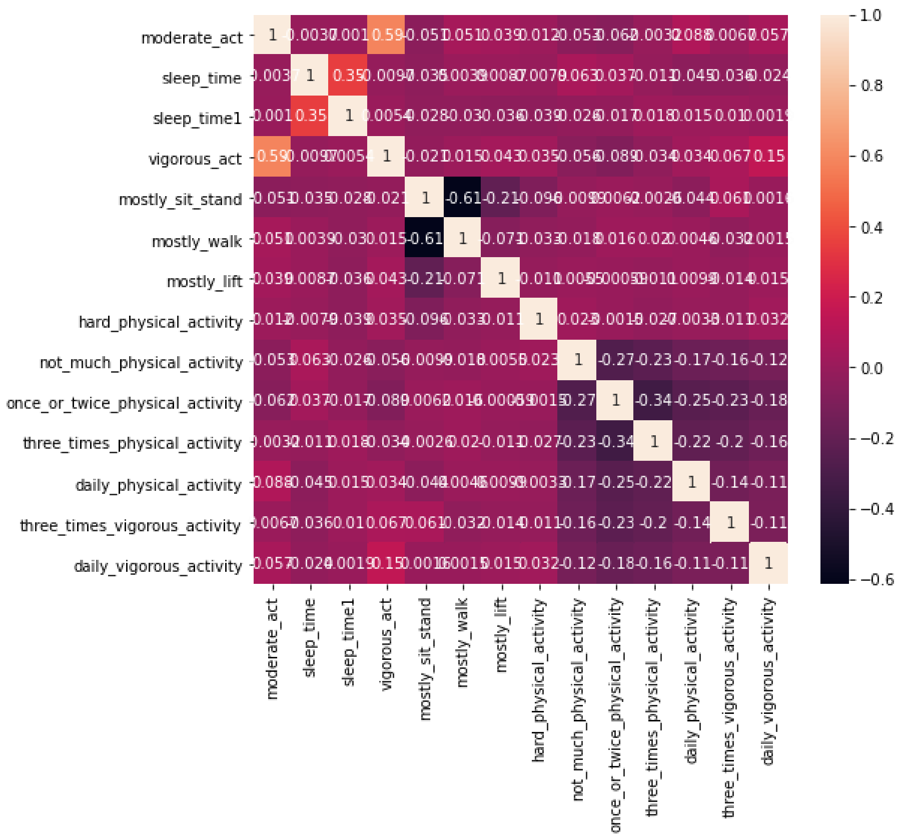


Figure A. 8: Correlation between Different Variables of Activity and Sleep Table

A.2. Physical Activity Readiness (PAR)

Table A. 2: Physical Activity Readiness Survey Attributes

Column Name	Question	Options
chestPain	Do you feel pain in your chest when you do physical activity?	True (9%) False (91%)
chestPainInLastMonth	In the past month, have you had chest pain when you were not doing physical activity?	True (15%) False (85%)
dizziness	Do you lose your balance because of dizziness or do you ever lose consciousness?	True (13%) False (87%)
heartCondition	has your doctor ever said that you have a heart condition and that you should only do physical activity recommended by a doctor?	True (6%) False (94%)
jointProblem	Do you have a bone or joint problem that could be made worse by a change in your physical activity?	True (20%) False (80%)
physicallyCapable	Do you know of any reason why you should not do physical activity?	True (3%) False (97%)
prescriptionDrugs	Is your doctor currently prescribing drugs (for example water pills) for your blood pressure or heart condition?	True (15%) False (85%)

Table A. 3: Percentage of Participant without Specific Issues

Percentage of participants without:	
Chest pain	91%
Chest pain in last month	85%
dizziness	87%
Heart condition	94%
Joint problem	80%
Physically capability	97%
Using prescribed drugs	85%

Table A. 4: Number of Participants with One or More Issues

Number of participants with at least one issue	**11428**
Number of participants with more than one issue	**4839**
Number of participants with all the issues	**77**
Number of participant with exactly one issue(first minus the second)	**6512**

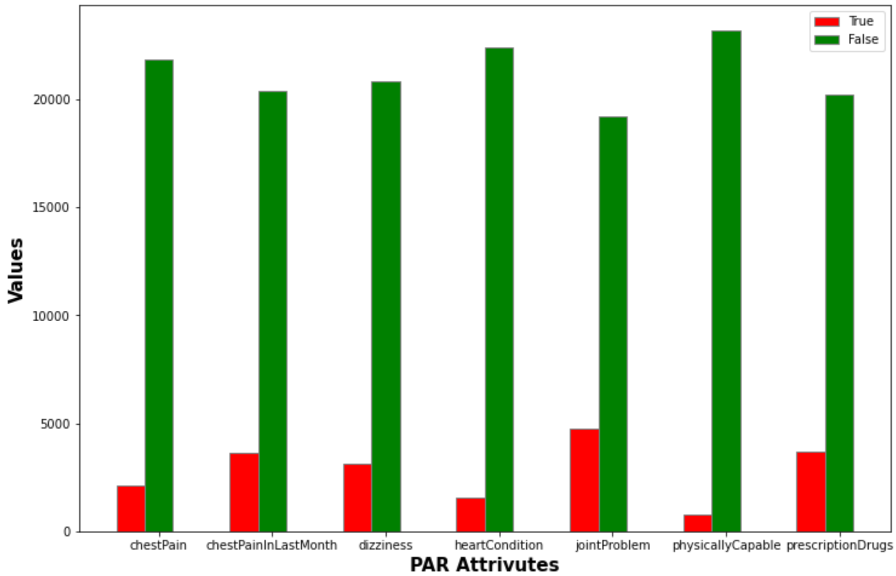


Figure A. 9: Participants with and without Specific Conditions

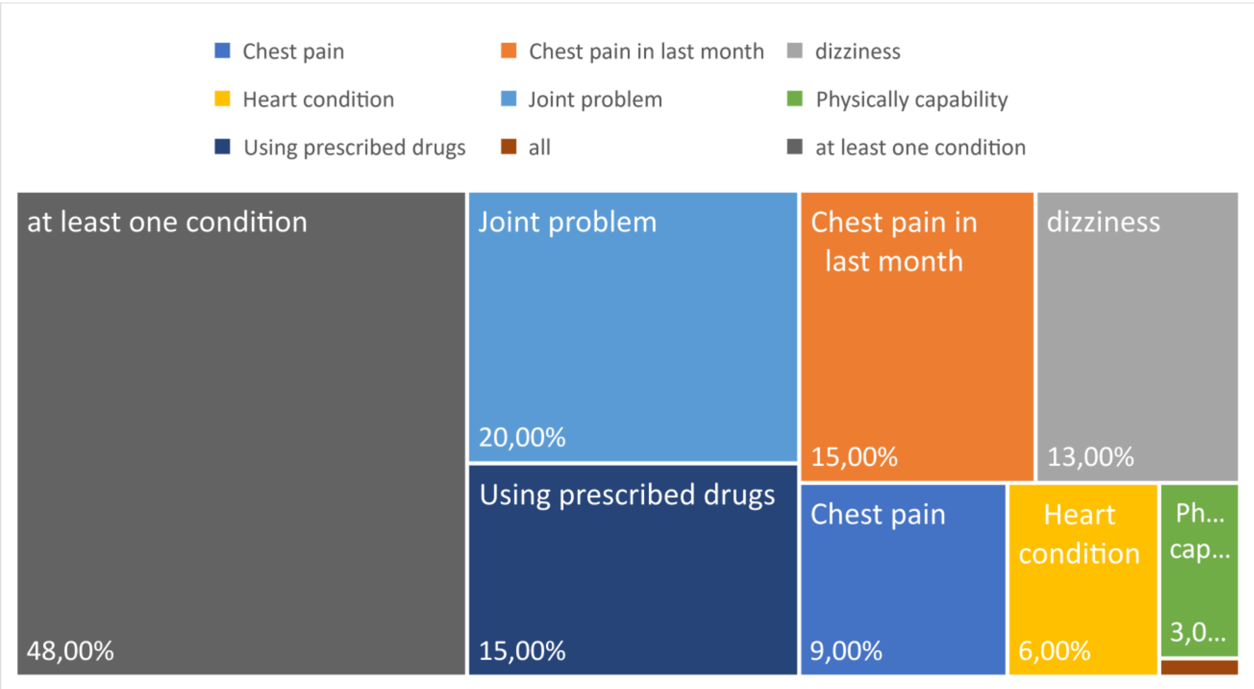


Figure A. 10: Percentage of Prevalence of each Condition in PAR Table

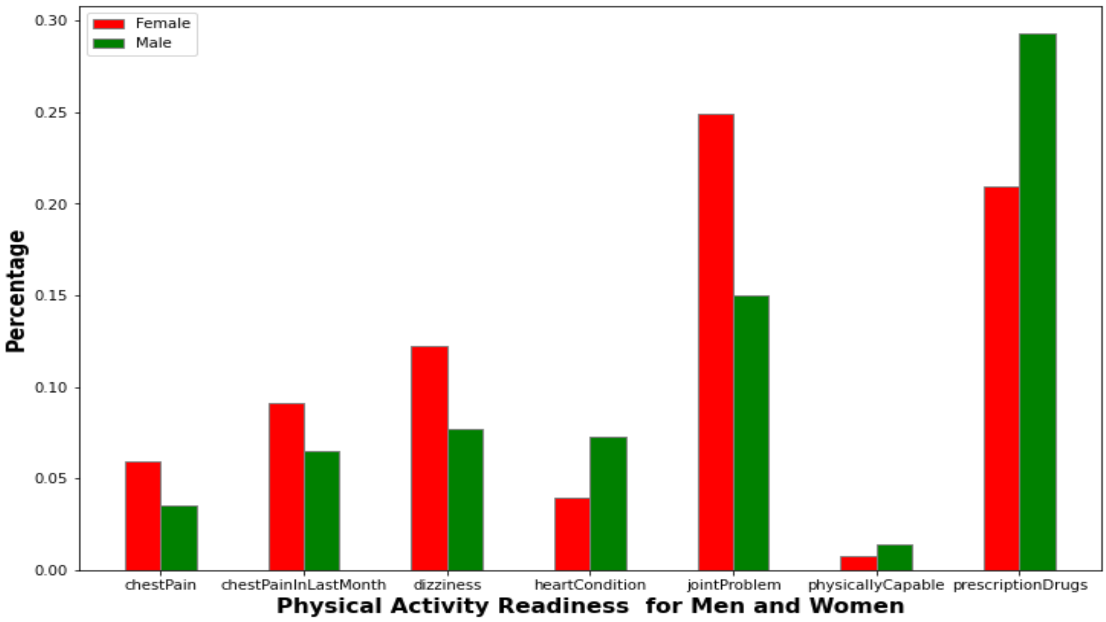


Figure A. 11

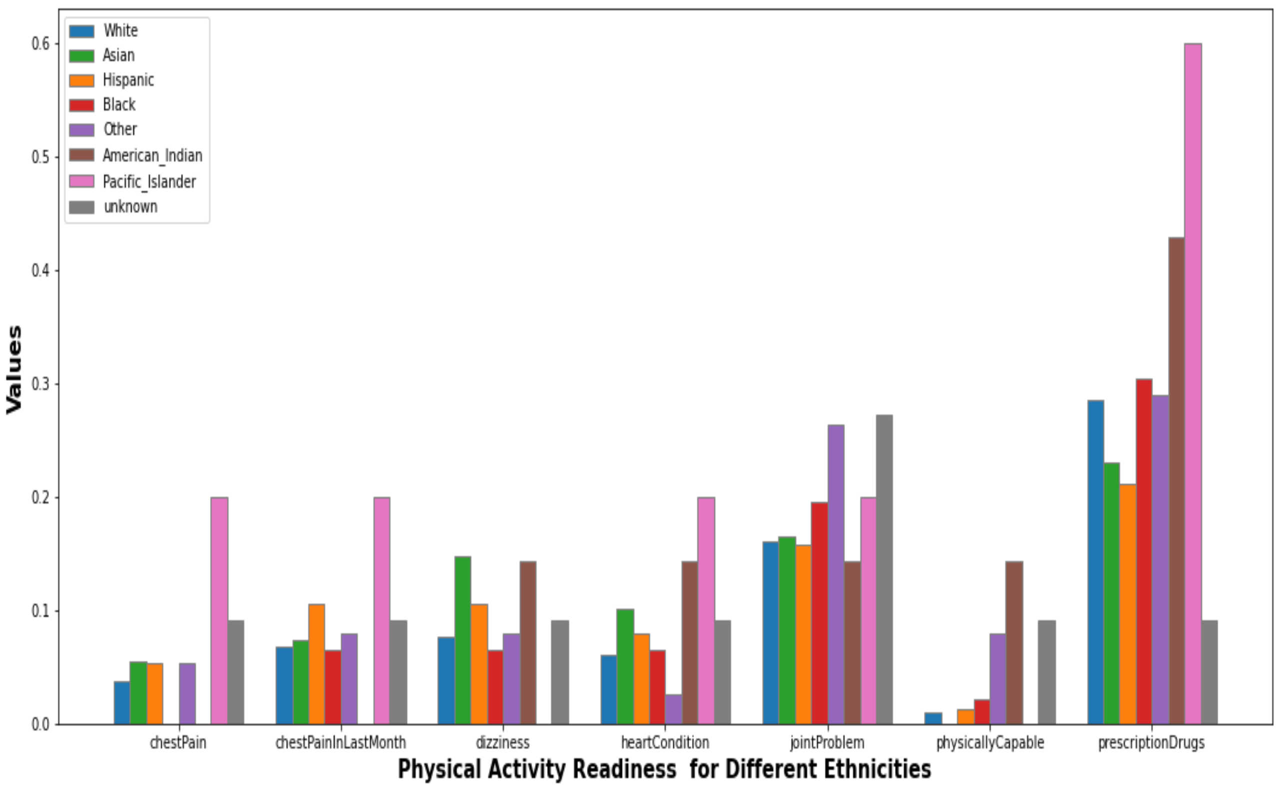


Figure A. 12

A.3. Risk Factor Survey

Table A. 5: Risk Factor Survey Attributes

Description	Values	Meaning of Each Value	Percentage
family_history	1	Father or brother with early heart disease	15.83%
	2	Mother or sister with early heart disease	5.23%
	3	None	76.62%
	1 and 2	both	2.32%
	None	Count of None values	106
heart_disease	1	Heart Attack/Myocardial Infarction	0.38%
	2	Heart Bypass Surgery	0.17%
	3	Coronary Blockage/Stenosis	0.24%
	4	Coronary Stent/Angioplasty	0.4%
	5	Angina (heart chest pains)	0.76%
	6	High Coronary Calcium Score	0.3%
	7	Heart Failure or Congestive Heart Failure	0.3%
	8	Atrial fibrillation	1.65%
	9	Congenital Heart Defect	1.42%
	10	None of the above	91.13%
	More than one	Diagnosed with more than one	36.5%
	None	Count of None values	59
medications_to_treat	1	To treat and lower cholesterol	5.51%
	2	To treat hypertension and lower blood pressure	8.04%
	3	To treat diabetes/pre-diabetes and lower blood sugar	0.84%
	4	None of the above	77.35%
	More than one	More than one medication is being used	8.25%
	None	Count of None values	26
vascular	1	Stroke	0.45%
	2	Transient Ischemic Attack (TIA)	0.41%
	3	Carotid Artery Blockage/Stenosis	0.45%
	4	Carotid Artery Surgery or Stent	0.83%
	5	Peripheral Vascular Disease (Blockage/Stenosis, Surgery, or Stent)	0.67%
	6	Abdominal Aortic Aneurysm	0.25%
	7	None of the above	95.98%
	More than one	More than one vascular disease	0.96%
	None	Count of None values	70
Extracted Features			
father_or_brother	True	Participants who chose option 1 regarding family history question	18.14%

	False	Participants who did not choose option 1 regarding family history question	81.86%
	None	Number of participants who did not choose any option for this question	106
mother_or_sister	True	Participants who chose option 2regarding family history question	7.56%
	False	Participants who did not choose option 2 regarding family history question	92.44%
	None	Number of participants who did not choose any option for this question	106
Heart_Attack	True	Participants who chose option 1 regarding heart disease question	93.34%
	False	Participants who did not choose option 1 regarding heart disease question	6.65%
	None	Number of participants who did not choose any option for this question	59
Bypass_Surgery	True	Participants who chose option 2 regarding heart disease question	1.06%
	False	Participants who did not choose option 2 regarding heart disease question	98.93%
	None	Number of participants who did not choose any option for this question	59
Coronary_Blockage	True	Participants who chose option 3 regarding heart disease question	1.69%
	False	Participants who did not choose option 3 regarding heart disease question	98.31%
	None	Number of participants who did not choose any option for this question	59
Coronary_Stent	True	Participants who chose option 4 regarding heart disease question	97.85%
	False	Participants who did not choose option 4 regarding heart disease question	2.15%
	None	Participants who chose option 4 regarding heart disease question	59
Angina	True	Participants who chose option 5 regarding heart disease question	1.97%
	False	Participants who did not choose option 5 regarding heart disease question	98.03%
	None	Number of participants who did not choose any option for this question	59
High_Coronary_Calcium_Score	True	Participants who chose option 6 regarding heart disease question	99.53%
	False	Participants who did not choose option 6 regarding heart disease question	0.47%
	None	Number of participants who did not choose any option for this question	59
Heart_Failure	True	Participants who chose option 7 regarding heart disease question	0.77%

	False	Participants who did not choose option 7 regarding heart disease question	99.23%
	None	Number of participants who did not choose any option for this question	59
Atrial_fibrillation	True	Participants who chose option 8 regarding heart disease question	0.23%
	False	Participants who did not choose option 8 regarding heart disease question	97.64%
	None	Number of participants who did not choose any option for this question	59
Congenital_Heart_Defect	True	Participants who chose option 9 regarding heart disease question	1.95%
	False	Participants who did not choose option 9 regarding heart disease question	98.05%
	None	Number of participants who did not choose any option for this question	59
lower_cholesterol_treatment	True	Participants who chose option 1 regarding using medication question	13.19%
	False	Participants who did not choose option 1 regarding using medication question	86.81%
	None	Number of participants who did not choose any option for this question	26
hypertension_lower_blood_pressure	True	Participants who chose option 2 regarding using medication question	15.82%
	False	Participants who did not choose option 2 regarding using medication question	84.18%
	None	Number of participants who did not choose any option for this question	26
diabetes	True	Participants who chose option 3 regarding using medication question	3.35%
	False	Participants who did not choose option 3 regarding using medication question	96.65%
	None	Number of participants who did not choose any option for this question	26
stroke	True	Participants who chose option 1 regarding vascular disease question	99.97%
	False	Participants who did not choose option 1 regarding vascular disease question	0.63%
	None	Number of participants who did not choose any option for this question	70
TIA	True	Participants who chose option 2 regarding vascular disease question	0.65%
	False	Participants who did not choose option 2 regarding vascular disease question	99.35%

	None	Number of participants who did not choose any option for this question	70
Carotid_Artery_Blockage	True	Participants who chose option 3 regarding vascular disease question	1.01%
	False	Participants who did not choose option 3 regarding vascular disease question	98.99%
	None	Number of participants who did not choose any option for this question	70
Carotid_Artery_Surgery	True	Participants who chose option 4 regarding vascular disease question	1.41%
	False	Participants who did not choose option 4 regarding vascular disease question	98.85%
	None	Number of participants who did not choose any option for this question	70
Peripheral_Vascular_Disease	True	Participants who chose option 5 regarding vascular disease question	1.09%
	False	Participants who did not choose option 5 regarding vascular disease question	98.90%
	None	Number of participants who did not choose any option for this question	70
Abdominal_Aortic_Aneurysm	True	Participants who chose option 6 regarding vascular disease question	0.35%
	Fales	Participants who did not choose option 6 regarding vascular disease question	99.65%
	None	Number of participants who did not choose any option for this question	70

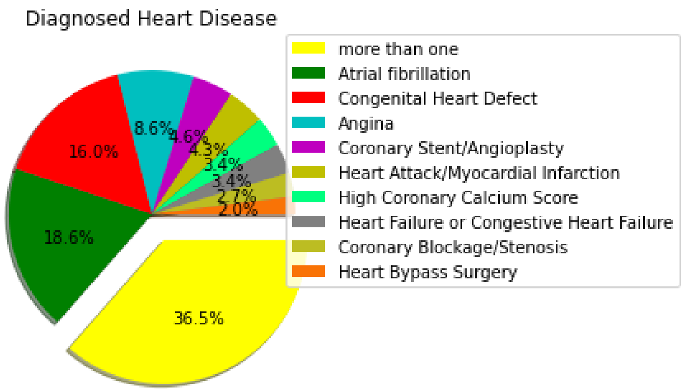


Figure A. 13: Proportion of Different Heat Diseases among Participants

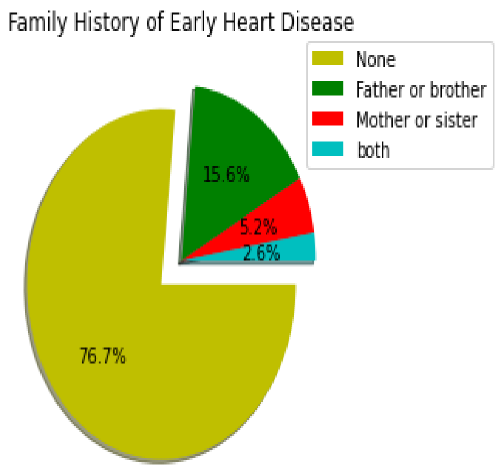


fig.a

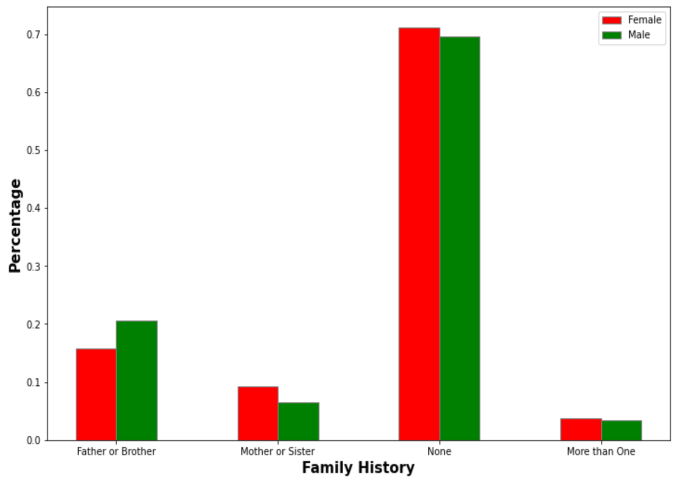


fig.b

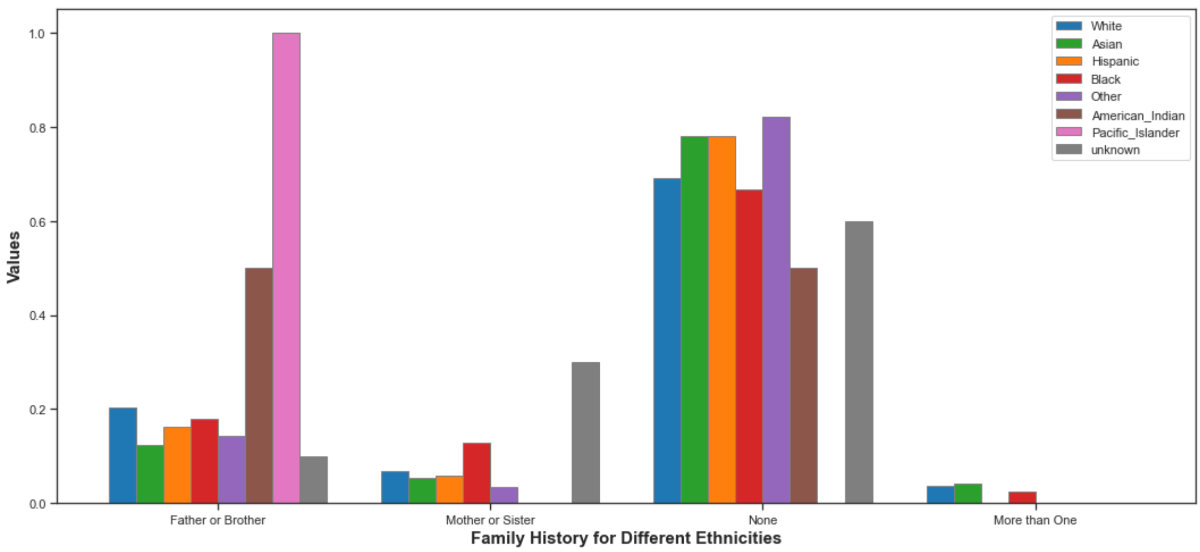


fig.c

Figure A. 14: a: Family History of Early Heart Disease. b: Family History of Early Heart Disease in Men and Women. c: Family History of Early Heart Disease in Different Ethnicities

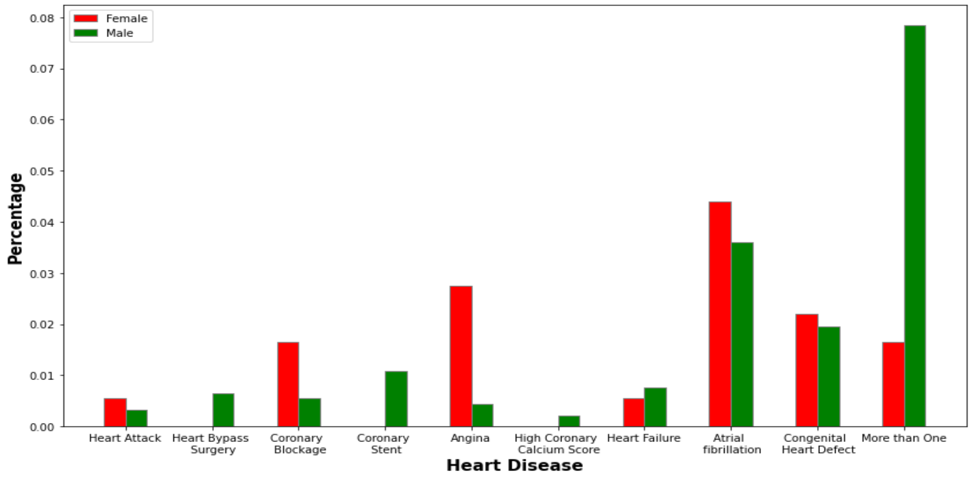


Figure A. 15: Heart Disease in Men and Women

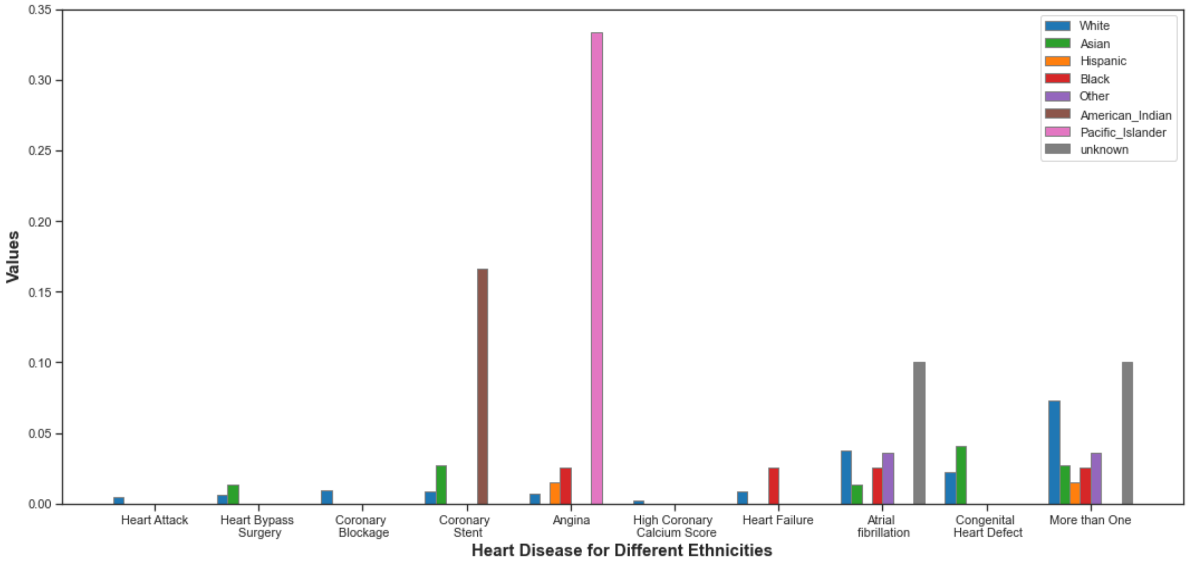


Figure A. 16

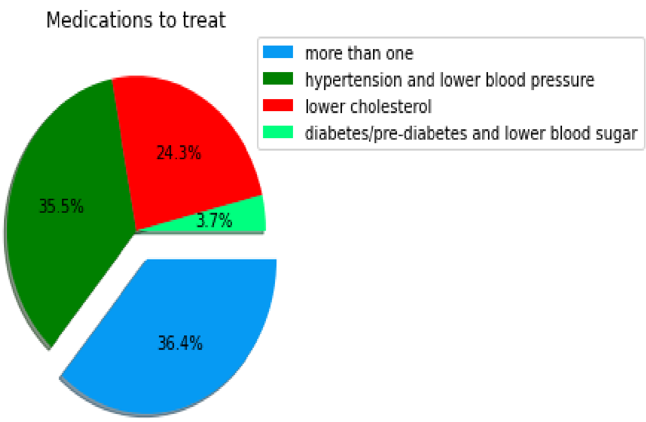


Figure A. 17

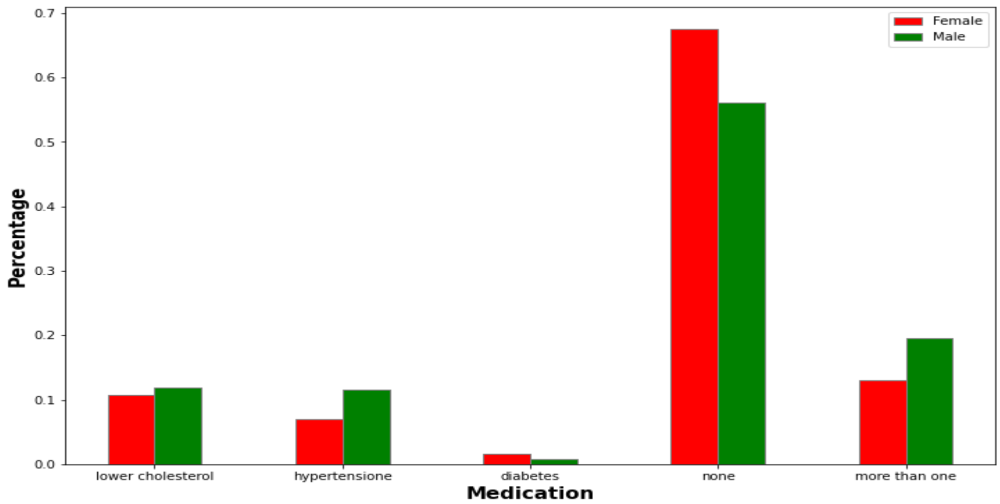


Figure A. 18: Medications among Men and Women

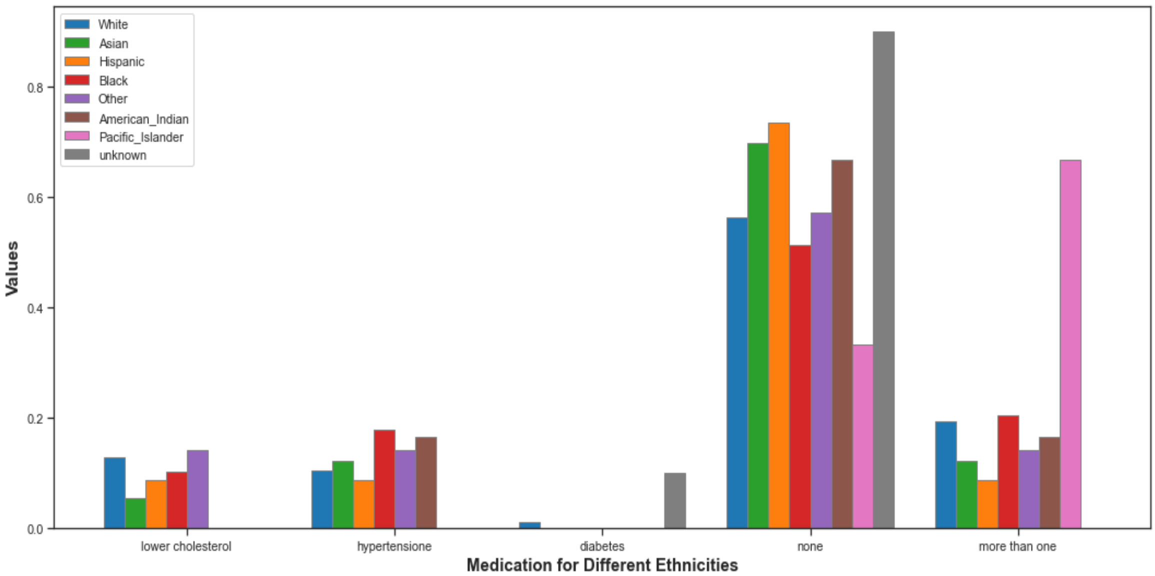


Figure A. 19

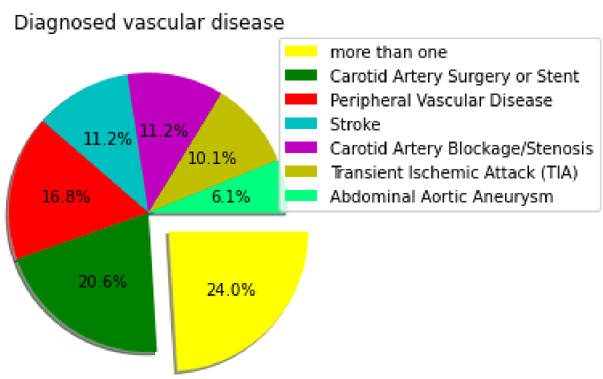


Figure A. 20

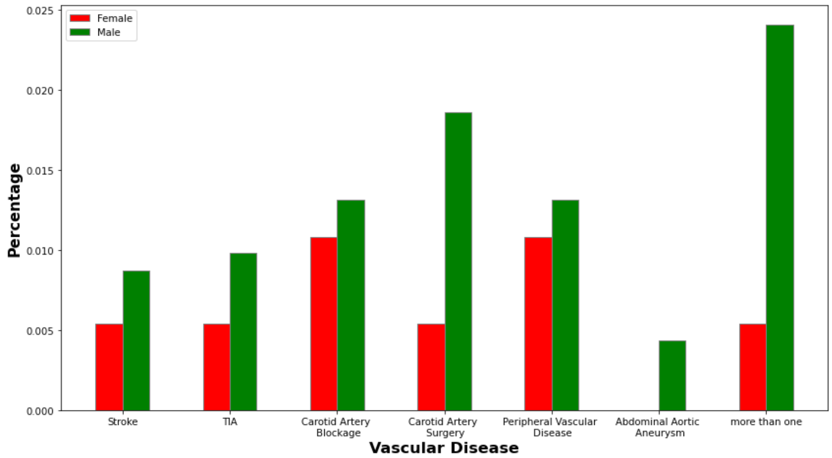


Figure A. 21: Vascular Disease in Men and Women

Table A. 6: Correlation among Different Risk Factors

Correlation between:	Value
vascular and heart disease	0.4
vascular and medication	0.22
vascular and family history	0.07
heart disease and medication	0.35
heart disease and family history	0.08
medication and family history	0.06

A.4. Cardio-Diet Survey

Table A. 7: Cardio Diet Survey Attributes

Column	Question	Options	Description
fish	How many servings of fish do you eat on an average week?	Servings of fish per week	Count 15411 mean 1.21 std 1.49 min 0 25% 0 50% 1 75% 2 max 50
fruit	How many cups of fruit do you eat in an average day?	Cups of fruit per day	Count 15381 mean 1.33 std 1.42 min 0 25% 1 50% 1 75% 2 max 50
grains	How many servings of whole grains do you eat on an average day?	Servings of whole grain per day	Count 15213 mean 2.26 std 2.44 min 0

			25% 1 50% 2 75% 3 max 50
sodium	Select the statements that apply to you:	a.I avoid eating prepackaged and processed foods.	33.03%
		b.I avoid eating out, but when I do, I seek out low-sodium options.	6.96%
		c.I avoid salt when I'm cooking at home.	20.50%
		*d. none of the above	12.54%
		A and b	4.87%
		A and c	9.77%
		B and c	2.83%
		A, b, c	9.48%
		Not filled	1903
sugar_drinks	How many beverages with added sugar do you drink every week?	Beverages with added sugar per week	Count 15428 mean 4.12 std 6.26 min 0 25% 0 50% 2 75% 5 max 50
vegetable	How many cups of vegetables do you eat in an average day?	Cups of vegetables per day	Count 15412 mean 1.91 std 1.75 min 0 25% 1 50% 2 75% 2 max 50
Extracted Features			
avoid_pre_packed	True	Participants who chose option 1 regarding sodium consumption	57.16%
	False	Participants who did not choose option 1 regarding sodium consumption	42.84%
	None	Number of participants who did not choose any option for this question	1903
avoid_eating_out	True	Participants who chose option 1 regarding sodium consumption	75.84%
	False	Participants who did not choose option 1 regarding sodium consumption	24.15%
	None	Number of participants who did not choose	1903

		any option for this question	
avoid_salt	True	Participants who chose option 1 regarding sodium consumption	42.58%
	False	Participants who did not choose option 1 regarding sodium consumption	57.41%
	None	Number of participants who did not choose any option for this question	1903
not_avoiding	True	Participants who chose option 1 regarding sodium consumption	12.79%
	False	Participants who did not choose option 1 regarding sodium consumption	87.21%
	None	Number of participants who did not choose any option for this question	1903

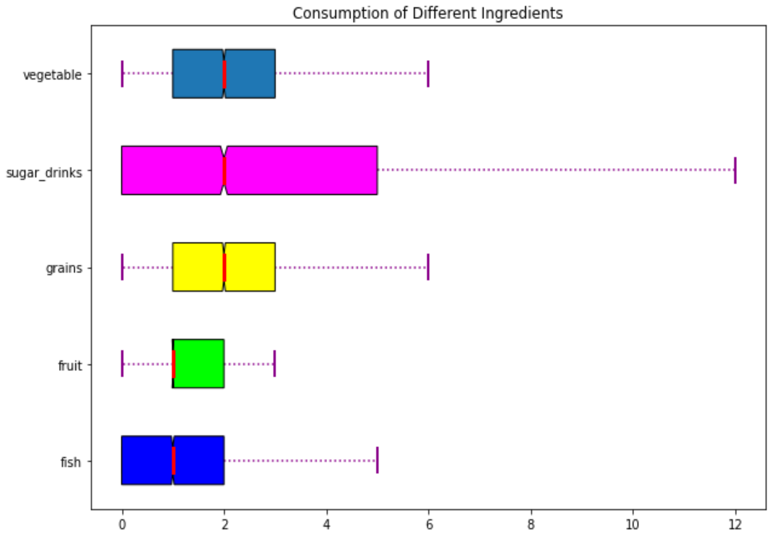


Figure A. 22

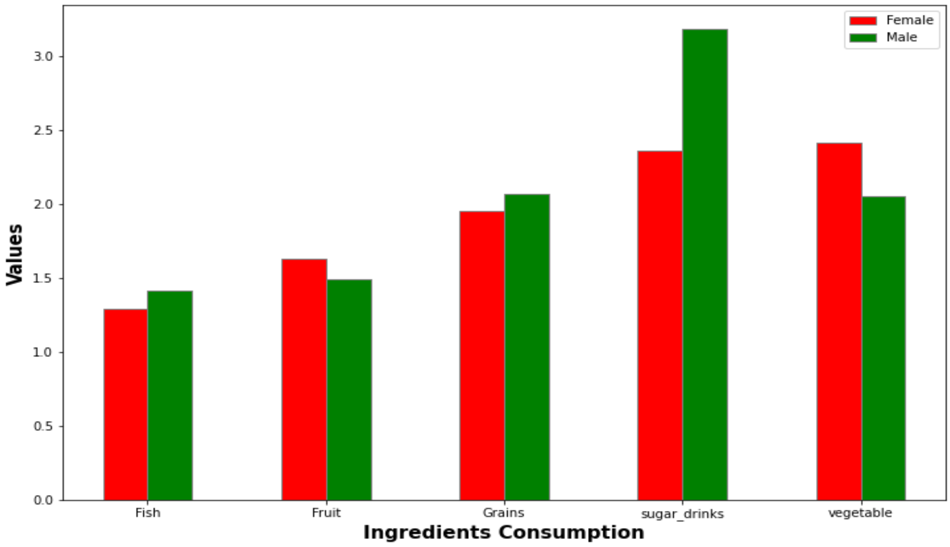


Figure A. 23

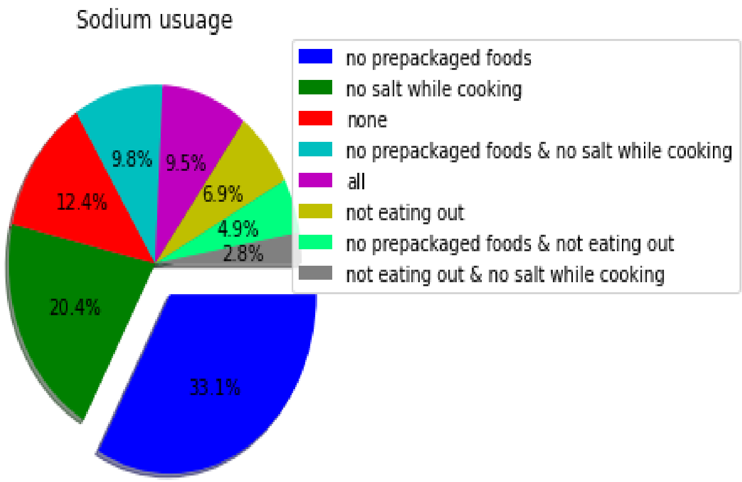


Figure A. 24

Table A. 8: Cardio Diet Table Correlations

	fish	fruit	grains	sugar_drinks	vegetable
fish	1.000000	0.200794	0.115036	-0.054145	0.279695
fruit	0.200794	1.000000	0.171117	-0.099039	0.479012
grains	0.115036	0.171117	1.000000	0.081852	0.185980
sugar_drinks	-0.054145	-0.099039	0.081852	1.000000	-0.097459
vegetable	0.279695	0.479012	0.185980	-0.097459	1.000000

A.5. Wellbeing Survey

Table A. 9: Wellbeing (satisfied) Survey Attributes

Columns	Question	Values		Description
feel_worthwhile1	Overall, to what extent do you feel the things you do in your life are worthwhile?	0-10		Count 14122 mean 7.35 std 2.05 min 0 25% 6 50% 8 75% 9 max 10
feel_worthwhile2	How about happy?	0-10		Count 14120 mean 7.05 std 2.06 min 0 25% 6 50% 7 75% 8 max 10
feel_worthwhile3	How about worried?	0-10		Count 14118 mean 4.60 std 2.73 min 0 25% 2 50% 4 75% 7 max 10
feel_worthwhile4	How about depressed?	0-10		Count 14036 mean 2.52 std 0.9 min 1 25% 0 50% 2 75% 4 max 10
riskfactors1	Over the next 10 years, how likely do you think it is that you personally will have a heart attack, stroke, or die due to cardiovascular disease?	a.	Not at all	48.80%
		b.	A little	33.45%
		c.	Moderately	13.10%
		d.	A lot	3.51%
		e.	Extremely	1.14%

riskfactors2	Over the next 10 years, compared to others your age and sex, how would you rate your risk of having a heart attack, stroke, or dying due to cardiovascular disease?	Much lower than average	23.88%
		Lower than average	29.26%
		Average	25.8%
		Higher than average	18.75%
		Much higher than average	2.21%
riskfactors3	Over your lifetime how likely do you think it is that you personally will have a heart attack, stroke, or die due to cardiovascular disease?	a. Not at all	15.52%
		b. A little	36.98%
		c. Moderately	30.12%
		d. A lot	12.06%
		e. Extremely	5.15%
Riskfactors4	Over your lifetime, compared to others your age and sex, how would you rate your risk of having a heart attack, stroke, or dying due to cardiovascular disease?	a. Much lower than average	18.21%
		b. Lower than average	27.83%
		c. Average	28.51%
		d. Higher than average	21.62%
		e. Much higher than average	3.57%
satisfiedwith_life	Overall, how satisfied are you with life as a whole these days?	0-10	Count 14133 mean 7.08 std 1.97 min 0 25% 6 50% 7 75% 8 max 10

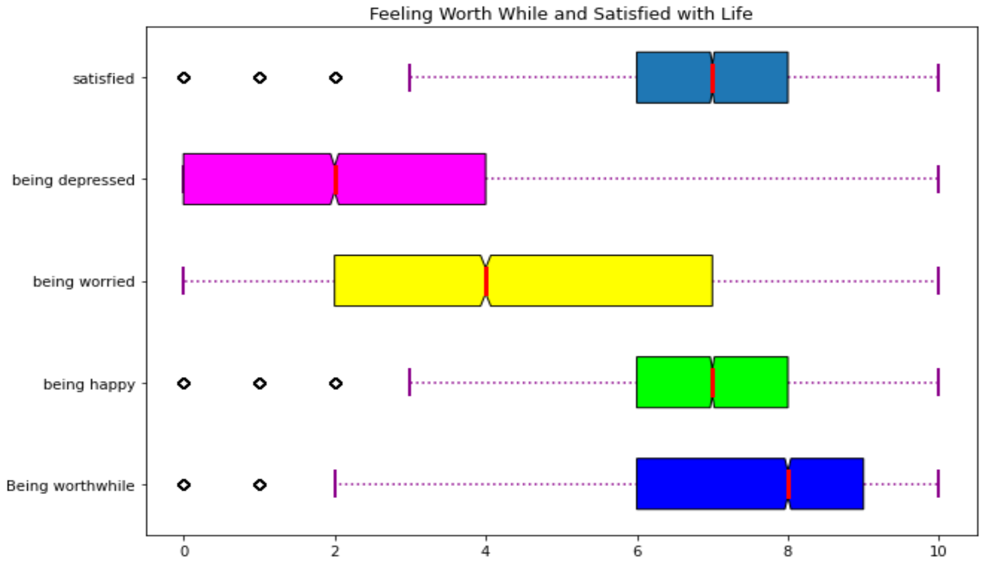


Figure A. 25

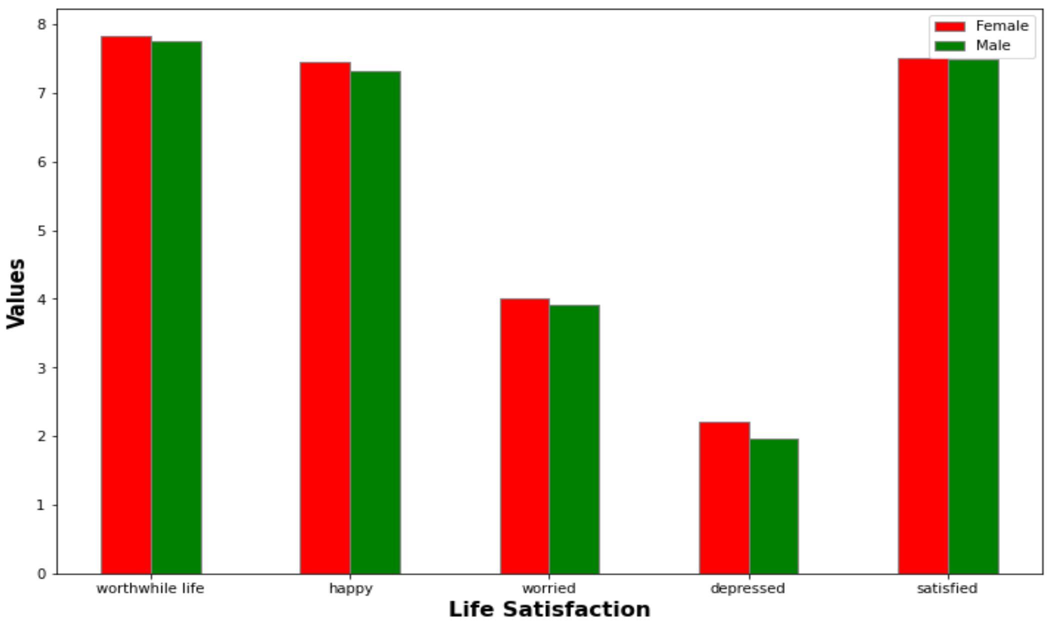


Figure A. 26

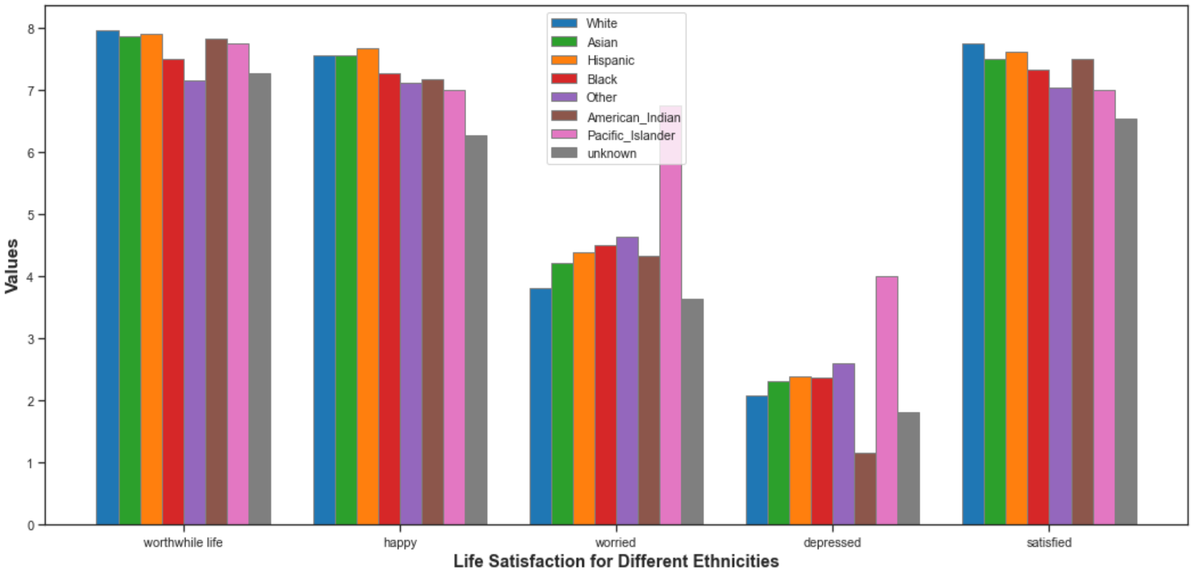


Figure A. 27

Figure17: 17.

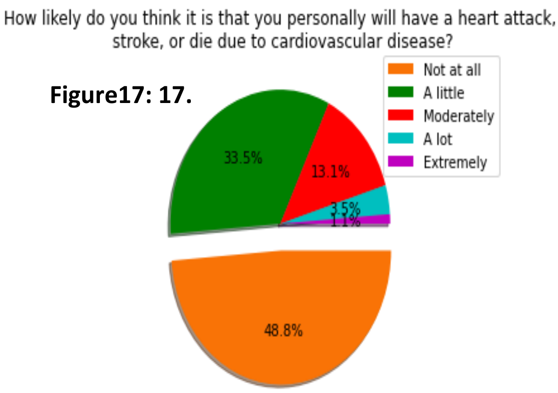


fig.a

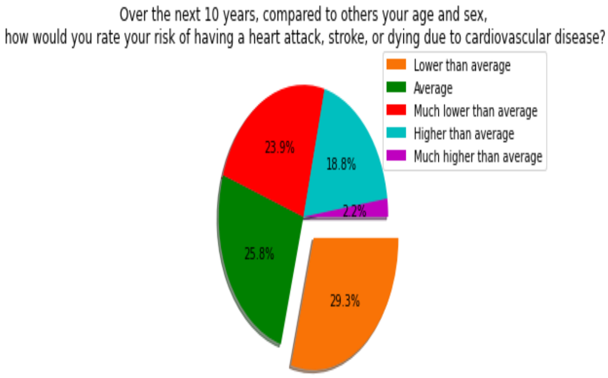


fig.b

Figure A. 28: a: Over the Next 10 Years, How Likely Do You think It Is That You Personally will Have a Heart Attack, Stroke, or Die Due to Cardiovascular Disease? b: Over the Next 10 Years, Compared to Others Your Age and Sex, How Would You Rate Your Risk of Having

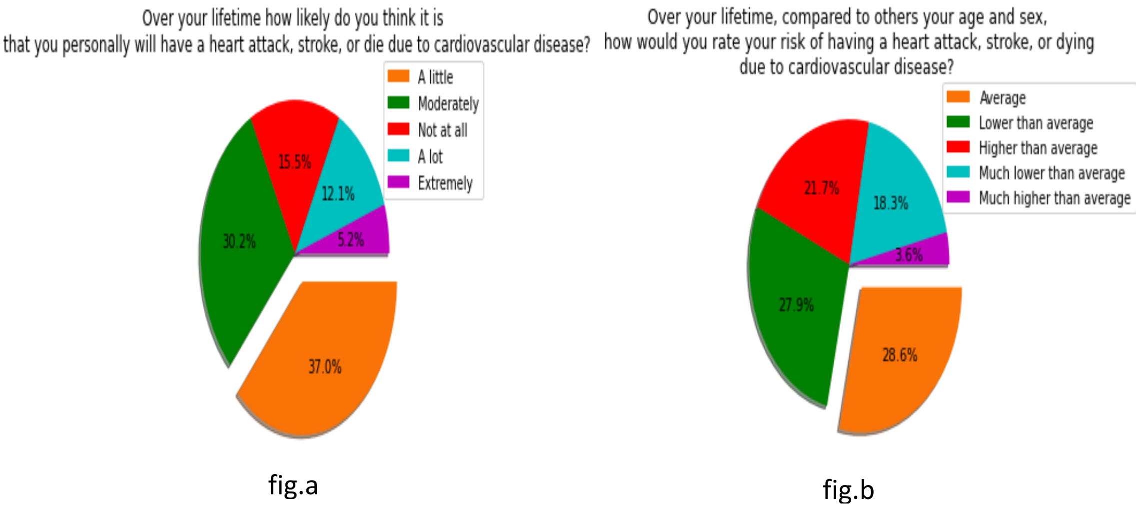


Figure A. 29:a: Over Your Lifetime How Likely Do You Think It Is That You Personally will Have a Heart Attack, Stroke, or Die due to Cardiovascular Disease? b: Over Your Lifetime, Compared to Others Your Age and Sex, How Would You Rate Your Risk of Having a Heart

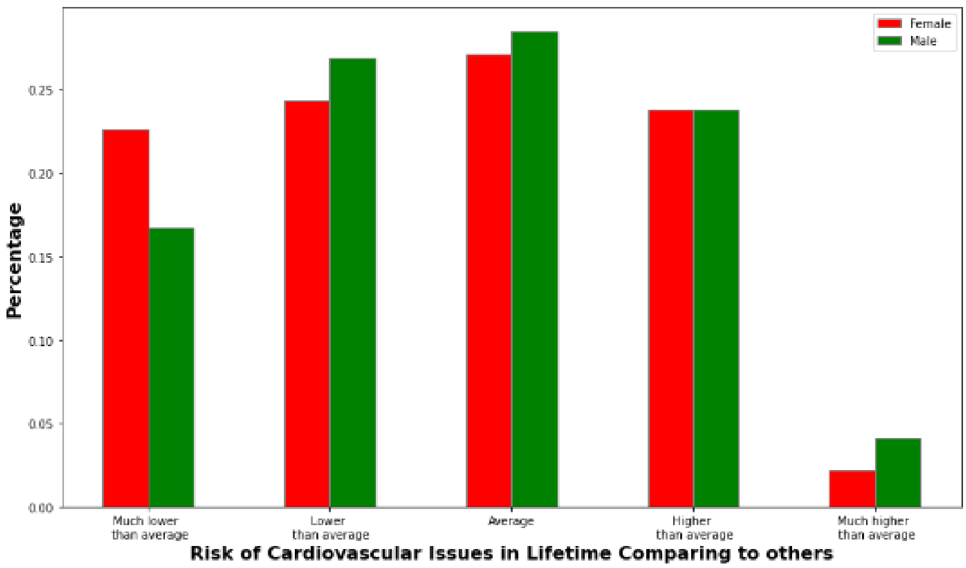


Figure A. 30

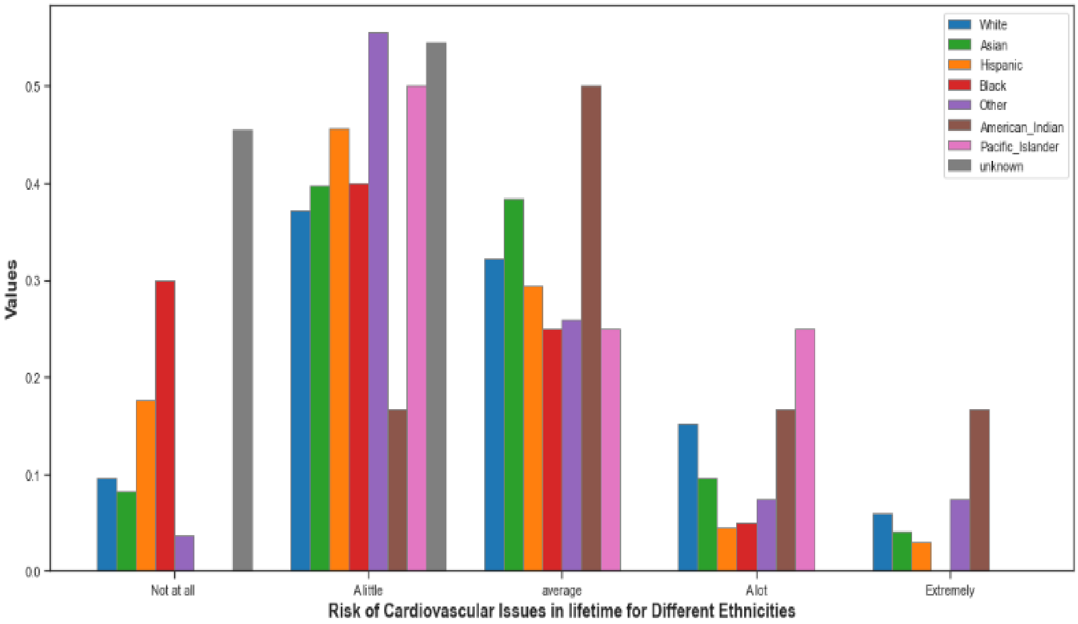


Figure A. 31

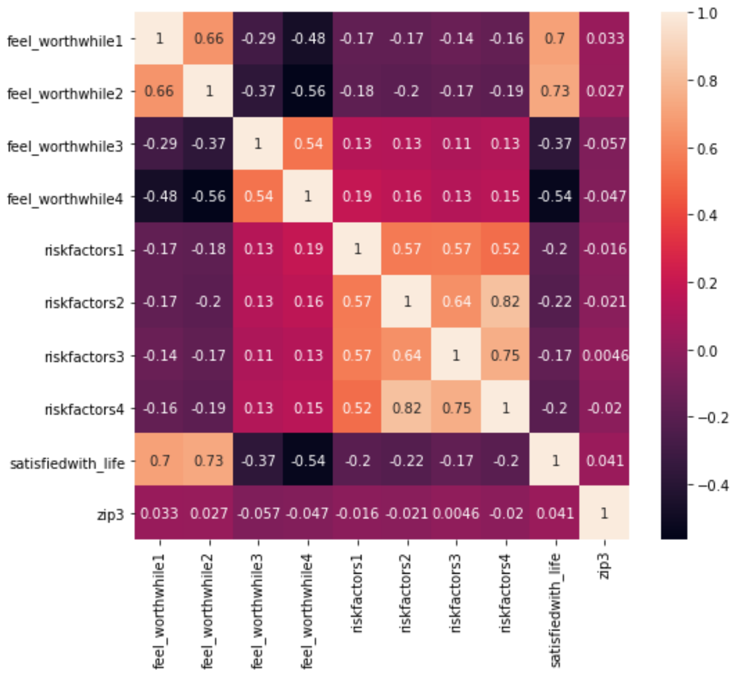


Figure A. 32: correlation between different risk factors and satisfaction factors

A.5. Heart Age

Table A. 10: Heart Age Attributes Survay

columns	question	values	description
bloodPressureInstruction	Enter your diastolic blood pressure	Continues(0-200 mmHg)	Count: 4760 Mean: 107.50 Std: 8.51 Min: 58 25%: 102.01 50%: 102.01 75%: 116 Max: 120
heartAgeDataBloodGlucose	If available, enter your fasting blood glucose	Continues(0-1890 mmHg)	Count: 4760 Mean: 6,17 Std: 2.26 Min: 3 25%: 5.04 50%: 5.04 75%: 5.63 Max: 15
heartAgeDataHdl	Enter your HDL Cholesterol	Continues(0-96104 mg/dl)	Count: 4760 Mean: 2.41 Std: 1.68 Min: 1 25%: 1.27 50%: 1.43 75%: 2.72 Max: 7
heartAgeDataLdl	If available, enter your LDL Cholesterol	Continues(0-7989 mg/dl)	Count: 4760 Mean: 2.08 Std: 1.42 Min: 1 25%: 1.09 50%: 1.67 75%: 2.04 Max:6
heartAgeDataSystolicBloodPressure	Enter your systolic blood pressure	Continues(0-851 mmHg)	Count: 4760 Mean: 116.48 Std: 19.37 Min: 95 25%: 110 50%: 111.36 75%: 120 Max: 180
heartAgeDataTotalCholesterol	Enter you total cholesterol	Continues(0-400 mg/dl)	Count: 4760 Mean: 7.19 Std: 1.24 Min: 5 25%: 6.19 50%: 6.19 75%: 8.15 Max: 12

heartAgeDataAge	What is your age?	Continues(18-86 years)	Count: 4723 Mean: 42.53 Std: 15.01 Min: 18 25%: 31 50%: 40 75%: 54 Max: 86
heartAgeDataDiabetes	Do you have Diabetes?	Boolean	True: 4.18% False: 95.82% None: 1
heartAgeDataGender	What is your gender?	Categorical	Male: 81.89% Female: 17.88% Other: 0.23% None: 50
heartAgeDataEthnicity	Ethnicity	Categorical	White: 3632 Asian: 351 Hispanic: 321 Black: 144 Other: 116 Prefer not to indicate: 36 American Indian: 21 Pacific Islander: 14 Alaska Native: 3 None: 1
heartAgeDataHypertension	Are you being treated for Hypertension?	Boolean	True: 22.43% False: 77.57% None: 2
smokingHistory	Are you currently smoking cigarettes?	Boolean	True: 4.72% False: 95.28% None: 36

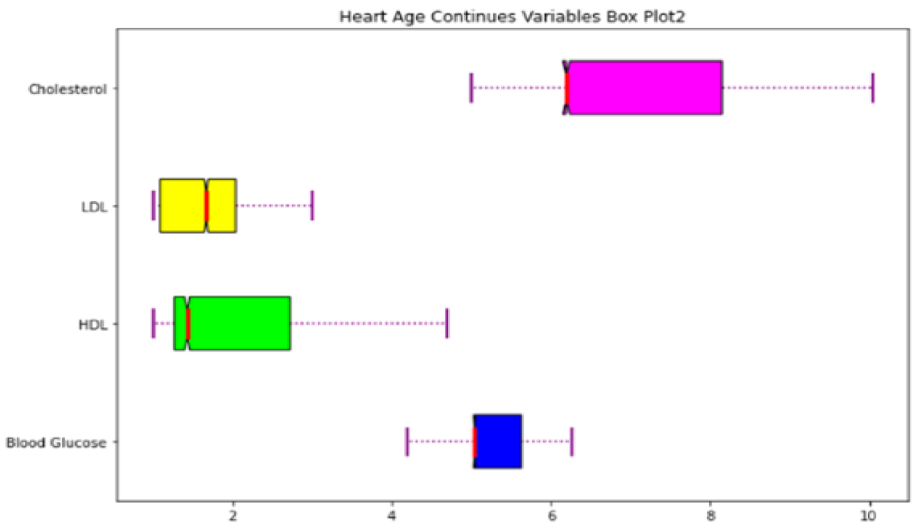


Figure A. 33

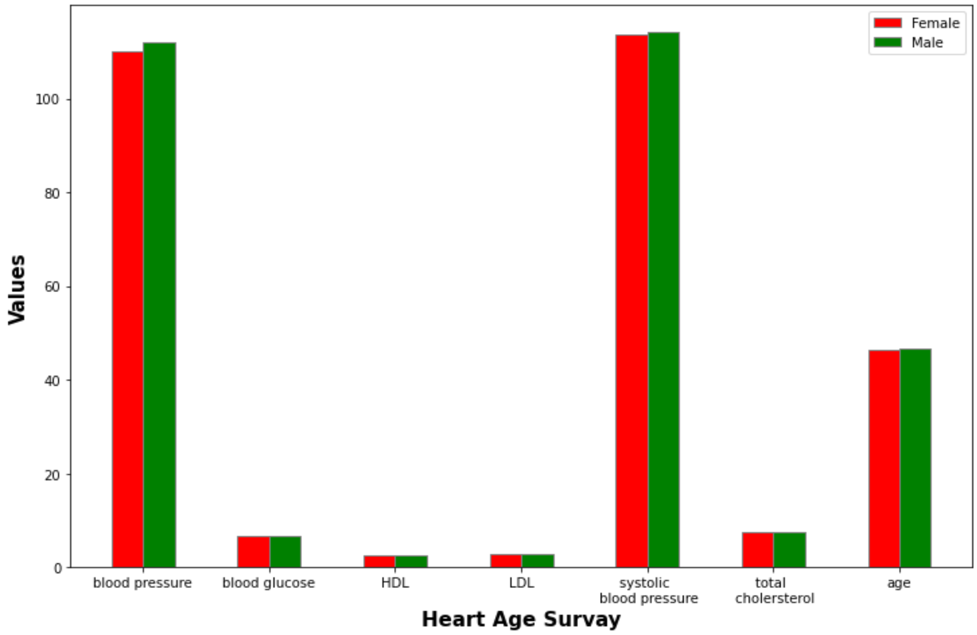


Figure A. 34: Heart Age Table Continues Variables in Men and Women

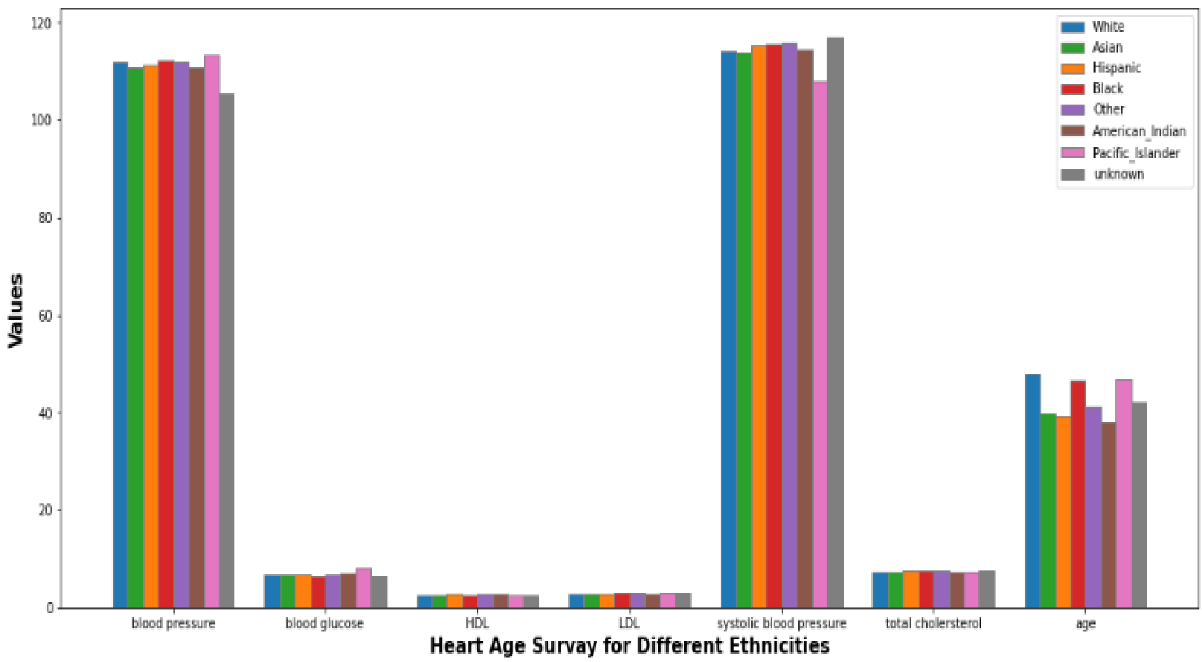


Figure A. 35

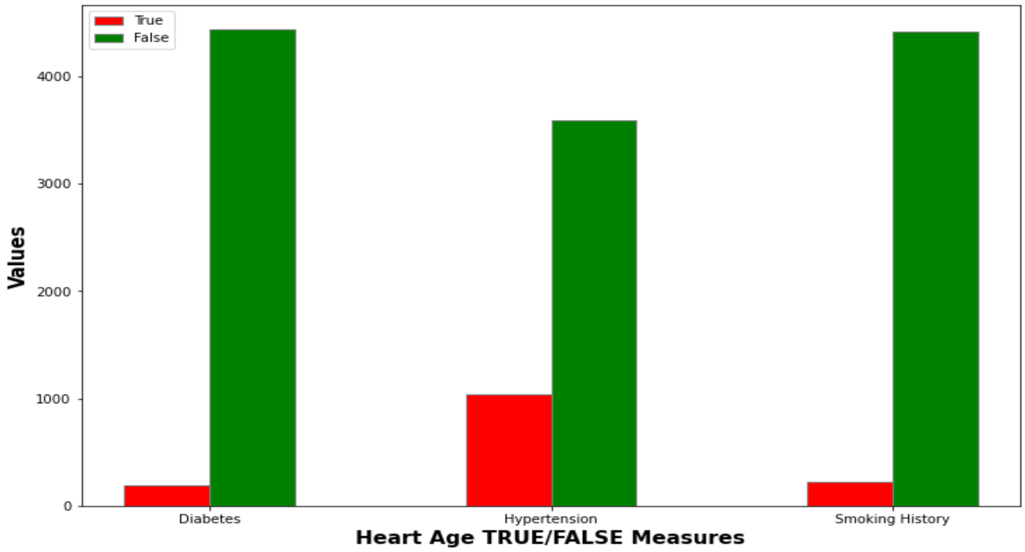


Figure A. 36

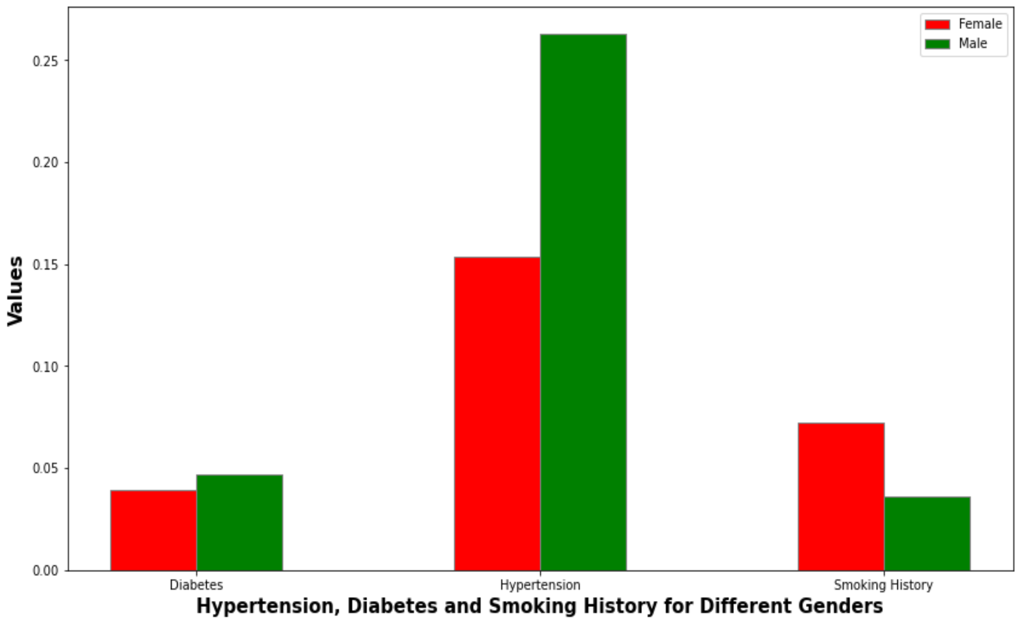


Figure A. 37

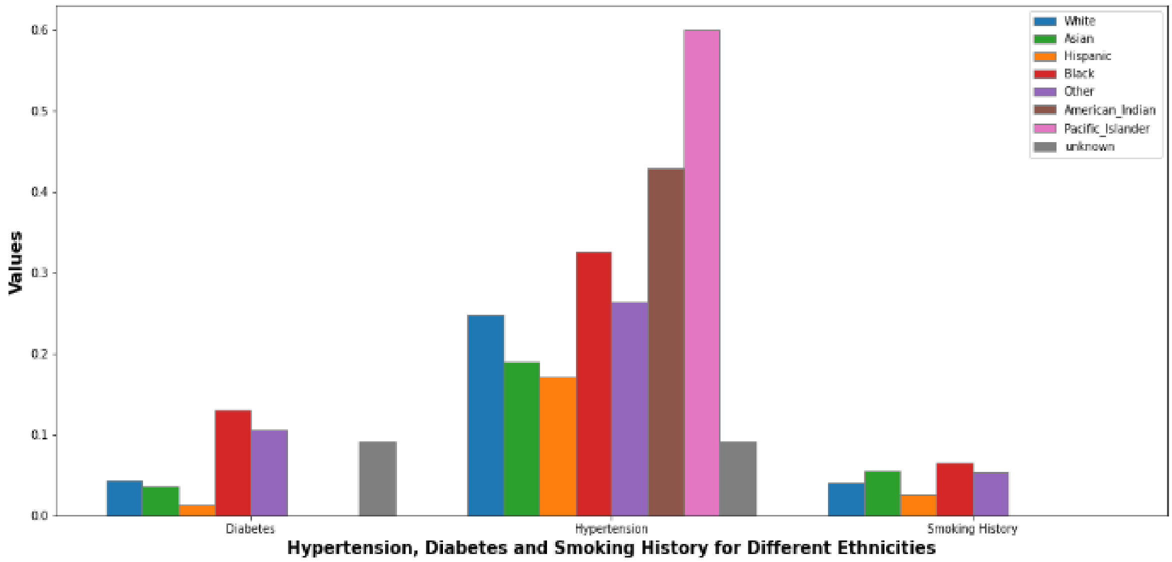


Figure A. 39: Heart Age Table Continues Variables in Different Ethnicities

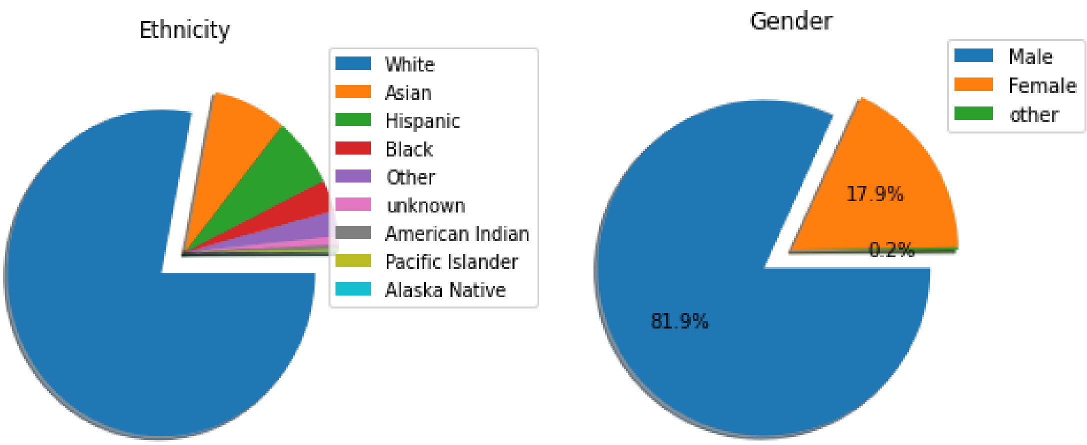


Figure A. 38: Heart Age Categorical Variables Pie Charts

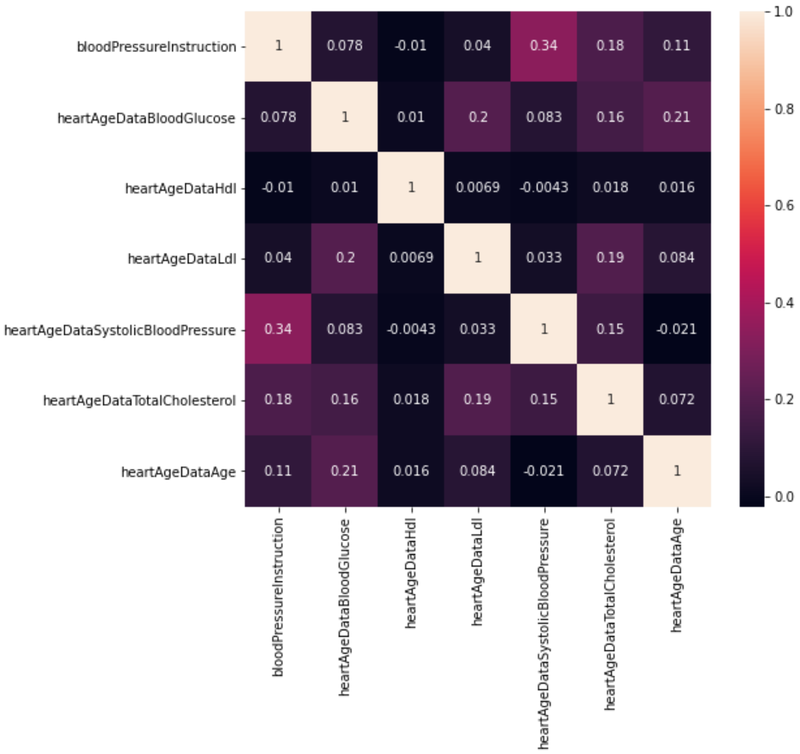


Figure A. 40: The Correlation Between Different Features in Heart Age Table

A.6. Demographic

Table A. 11: Demographic Survey Attributes

Columns	Question	Values	Description
patientWeightPounds	Weight of the participant	Continues(79-351 pounds)	Count: 2987 Mean: 84.09 Std: 21.23 Min: 35.83 25%: 70.31 50%: 81.65 75%: 95.48 Max: 159.21
patientHeightInches	Height of the participant	Continues(59-79 inches)	Count: 3060 Mean: 175.4 Std: 9.52 Min: 149.86 25%: 170.18 50%: 177.8 75%: 182.88 Max: 200.66
patientCurrentAge	Age of the participant	Continues(18-89 years)	Count: 1652 Mean: 40.06 Std: 14.95 Min: 18

			25%: 30 50%: 36 75%: 50 Max: 89
patientBiologicalSex	Sex of the participant	Categorical	Female Male
patientWakeUpTime	Waking up time of the participant	Categorical	
patientGoSleepTime	Sleeping time of the participant	categorical	

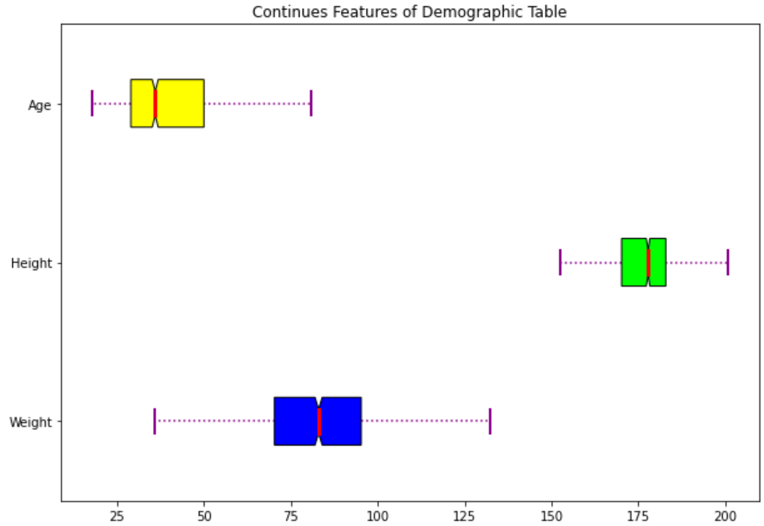
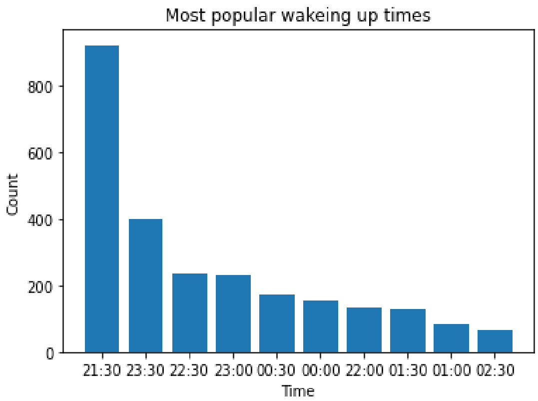


Figure A. 41



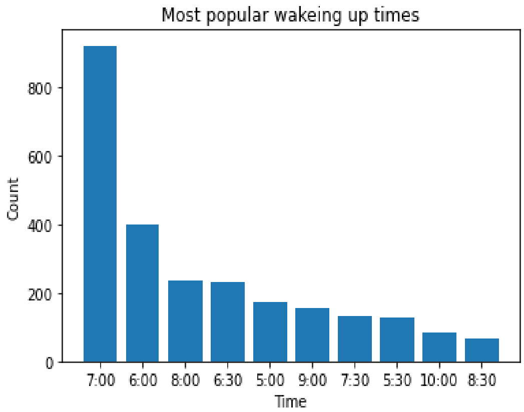


Figure A. 42: Most Popular Waking and Sleeping Times

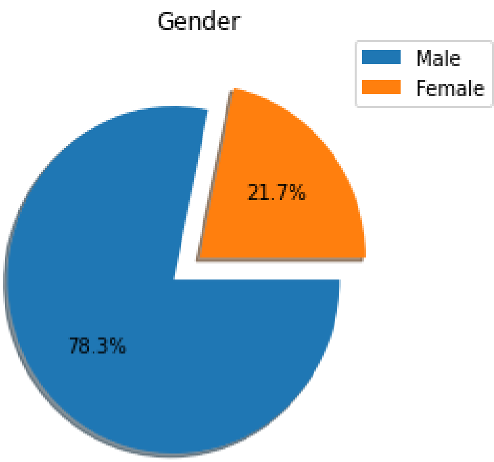


fig.a

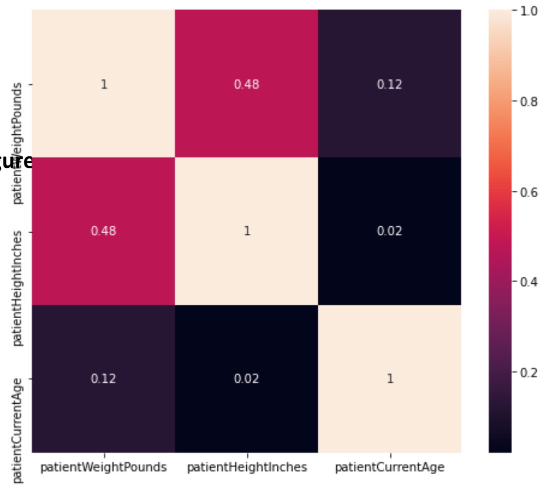


fig.b

Figure A. 43: a: Gender of Participants and b: Correlation among Continues Variables

A.7. Healthkit Workout

Table A. 12: HealthKit Workout Attributes

name	definition	values
Numeric attributes		
duration	Average minutes that the users are active per day	Count: 833 Mean: 39.65 Std: 44.28 Min: 0 25%: 15.79 50%: 30.57 75%: 49.63 Max: 507.69
energy	Average kcal that the users burn per day	Count: 833 Mean: 430.23 Std: 3256.51 Min: 0 25%: 77.41 50%: 176.02 75%: 320.39 Max: 76561.33
Distance	Average distance that the pass per day	Count: 833 Mean: 4174.47 Std: 8034.86 Min: 0 25%: 256.48 50%: 1862.46 75%: 4353.19 Max: 78429

freq	Number of times in a day that the user has a physical activity loner than 15 minutes	Count: 797 Mean: 1.15 Std: 1.60 Min: 0 25%: 0.81 50%: 1 75%: 1.26 Max: 38.44
'work out'_duration	Total time spent on a specific workout (85 workouts in list[4].	
'work out'_count	Total number of times a participant did a specific workout (85 workouts in list[4].	
weekend_duration	Average physical activity duration during weekend for each participant	count 833 mean 28.94 std 52.69 min 0 25% 0 50% 4.67 75% 38.5 max 507.69
weekday_duration	Average physical activity duration during weekdays for each participant	count 833 mean 29.93 std 47.84 min 0 25% 0 50% 18.97 75% 41.53 max 635.5
early_morning_time	Average physical activity duration during early morning for each participant	count 833 mean 3.32 std 13.85 min 0 25% 0 50% 0 75% 0 max 221.53
morning_time	Average physical activity duration during morning for each participant	count 833 mean 3.10 std 12.21 min 0 25% 0 50% 0 75% 0 max 150.08
noon_time	Average physical activity duration during noon for each participant	count 833 mean 4.60 std 16.03 min 0 25% 0 50% 0 75% 0.34 max 220.60
afternoon_time	Average physical activity duration during afternoon for each participant	count 833 mean 9.88 std 24.58 min 0

		25% 0 50% 0 75% 8.40 max 341.32
evening_time	Average physical activity duration during evening for each participant	count 833 mean 9.23 std 22.36 min 0 25% 0 50% 0 75% 10.06 max 375.97
late_evening_time	Average physical activity duration during late evening for each participant	count 833 mean 4.98 std 20.42 min 0 25% 0 50% 0 75% 2.12 max 461.59
night_time	Average physical activity duration during night for each participant	count 833 mean 4.54 std 17.72 min 0 25% 0 50% 0 75% 0.15 max 384
weekend_count	Average number of time a physical activity is recorded for the participant at weekend	count 833 mean 0.58 std 0.49 min 0 25% 0 50% 1 75% 1 max 1
weekday_count	Average number of time a physical activity is recorded for the participant in the weekdays	count 833 mean 0.79 std 0.59 min 0 25% 0 50% 1 75% 1 max 2
early_morning_count	Average number of time a physical activity is recorded for the participant in early morning	count 833 mean 0.09 std 0.23 min 0 25% 0 50% 0 75% 0 max 1
Morning_count	Average number of time a physical activity is recorded for the participant in the morning	count 833 mean 0.07 std 0.20 min 0

		25% 0 50% 0 75% 0 max 1
noon_count	Average number of time a physical activity is recorded for the participant at noon	count 833 mean 0.1 std 0.22 min 0 25% 0 50% 0 75% 0.07 max 1
afternoon_count	Average number of time a physical activity is recorded for the participant in the afternoon	count 833 mean 0.22 std 0.31 min 0 25% 0 50% 0.03 75% 0.33 max 1
evening_count	Average number of time a physical activity is recorded for the participant in the evening	count 833 mean 0.24 std 0.33 min 0 25% 0 50% 0.07 75% 0.33 max 1
late_evening_count	Average number of time a physical activity is recorded for the participant in late evening	count 833 mean 0.13 std 0.25 min 0 25% 0 50% 0 75% 0.17 max 1
night_count	Average number of time a physical activity is recorded for the participant at night	count 833 mean 0.14 std 0.28 min 0 25% 0 50% 0 75% 0.12 max 1
weekend_energy	Average amount of energy user burnt during weekend	count 833 mean 683.35 std 12209.49 min 0 25% 0 50% 26.81 75% 220.60 max 343771.37
weekday_energy	Average amount of energy user burnt during weekdays	count 833 mean 270.25 std 1767.93 min 0

		25% 0 50% 96.26 75% 260.10 max 39934.84
early_morning_energy	Average amount of energy user burnt during early morning	count 833 mean 24.16 std 90.60 min 0 25% 0 50% 0 75% 0 max 893.4
morning_energy	Average amount of energy user burnt during morning	count 833 mean 21.23 std 81.12 min 0 25% 0 50% 0 75% 0 max 841.76
noon_energy	Average amount of energy user burnt during noon	count 833 mean 29.03 std 108.32 min 0 25% 0 50% 0 75% 2.06 max 1743.05
afternoon_energy	Average amount of energy user burnt during afternoong	count 833 mean 192.18 std 2945.37 min 0 25% 0 50% 0 75% 47.94 max 76393.64
evening_energy	Average amount of energy user burnt during evening	count 833 mean 60.44 std 221.94 min 0 25% 0 50% 1.25 75% 52.03 max 5135.27
late_evening_energy	Average amount of energy user burnt during late evening	count 833 mean 70.75 std 1197.04 min 0 25% 0 50% 0 75% 11.84 max 34442.84
night_energy	Average amount of energy user burnt during night	count 833 mean 32.45 std 97.06 min 0

		25% 0 50% 0 75% 4.06 max 1029.94
weekend_distance	Average distance the user passed during weekend	count 833 mean 3716.19 std 10716.92 min 0 25% 0 50% 0 75% 2870.03 max 124690.64
weekday_distance	Average distance the user passed during weekday	count 833 mean 2936.53 std 6610.18 min 0 25% 0 50% 483.23 75% 3282.45 max 66541.73
early_morning_distance	Average distance the user passed during early morning	count 833 mean 518.23 std 2933.62 min 0 25% 0 50% 0 75% 0 max 46928.48
morning_distance	Average distance the user passed during morning	count 833 mean 388.43 std 1916.82 min 0 25% 0 50% 0 75% 0 max 31172.66
noon_distance	Average distance the user passed during noon	count 833 mean 632.31 std 3719.61 min 0 25% 0 50% 0 75% 0 max 59104.54
afternoon_distance	Average distance the user passed during afternoon	count 833 mean 1088.24 std 4255.51 min 0 25% 0 50% 0 75% 497.01 max 67450.59
evening_distance	Average distance the user passed during evening	count 833 mean 823.23 std 2437.90 min 0

		25% 0 50% 0 75% 508.44 max 28165.80
late_evening_distance	Average distance the user passed during late evening	count 833 mean 405.15 std 1956.14 min 0 25% 0 50% 0 75% 36.80 max 47611.69
night_distance	Average distance the user passed during night	count 833 mean 318.87 std 1195.01 min 0 25% 0 50% 0 75% 0 max 15080.15
number_of_days	Number of days the user data is being recorded	count 833 mean 118.88 std 793.97 min 1 25% 1 50% 4 75% 14 max 13047
Categorical attributes		
day_part	Part of the day that the user was mostly active in	afternoon 223 evening 207 night 109 late_evening 95 early_morning 83 noon 65 morning 51

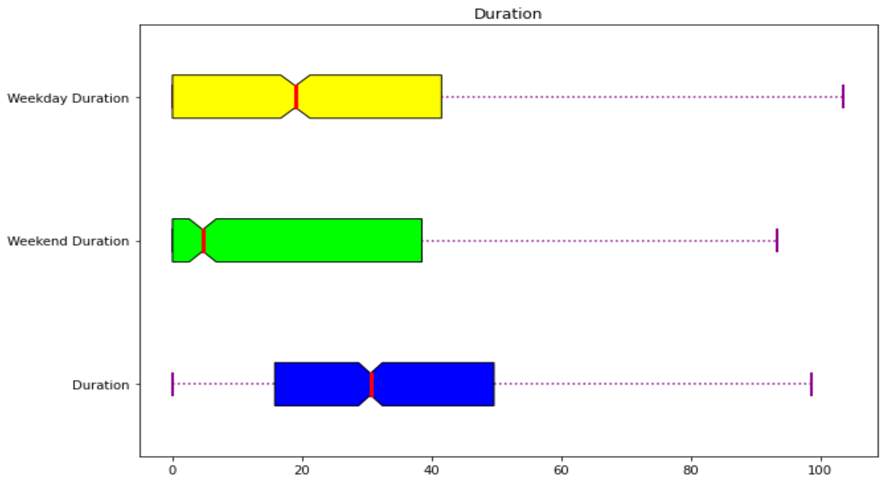


Figure A. 44: Average Physical Activity Duration During Weekend and Weekday in Minutes

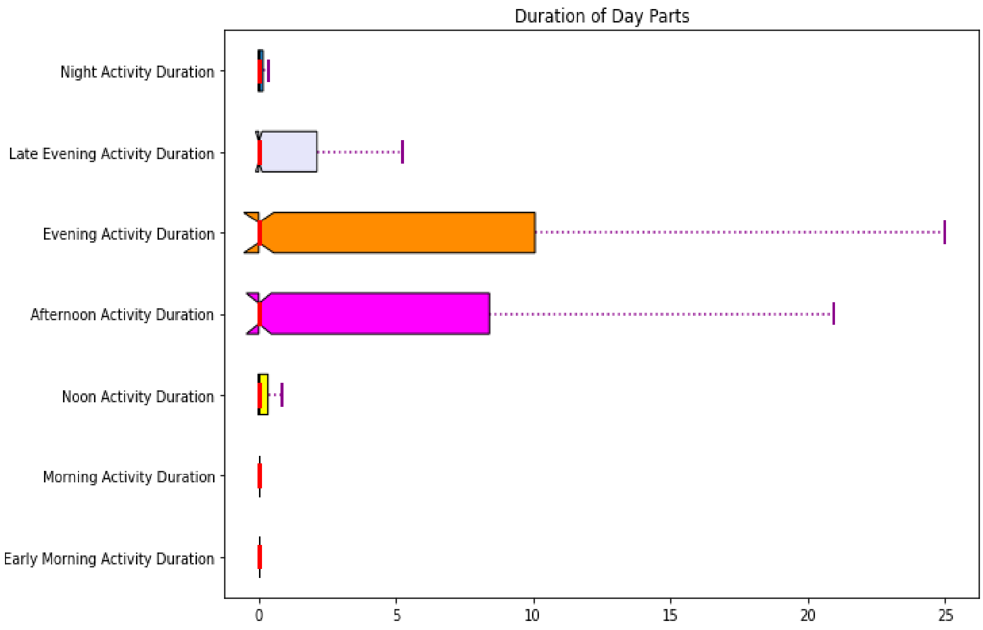


Figure A. 45: Average Physical Activity Duration in Different Parts of the Day in Minutes

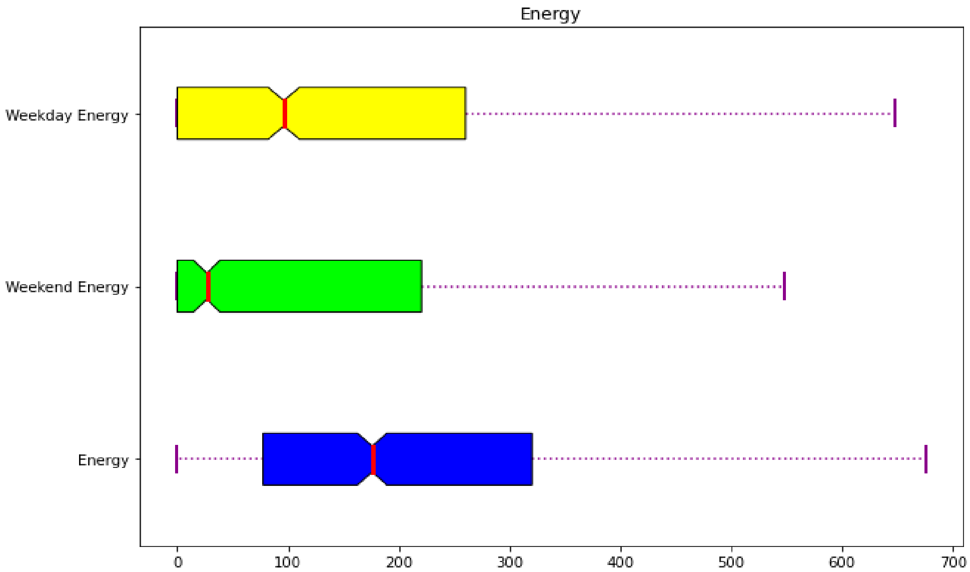


Figure A. 46: Average Energy Burnt During Physical Activity in Weekend and Weekday in Minutes

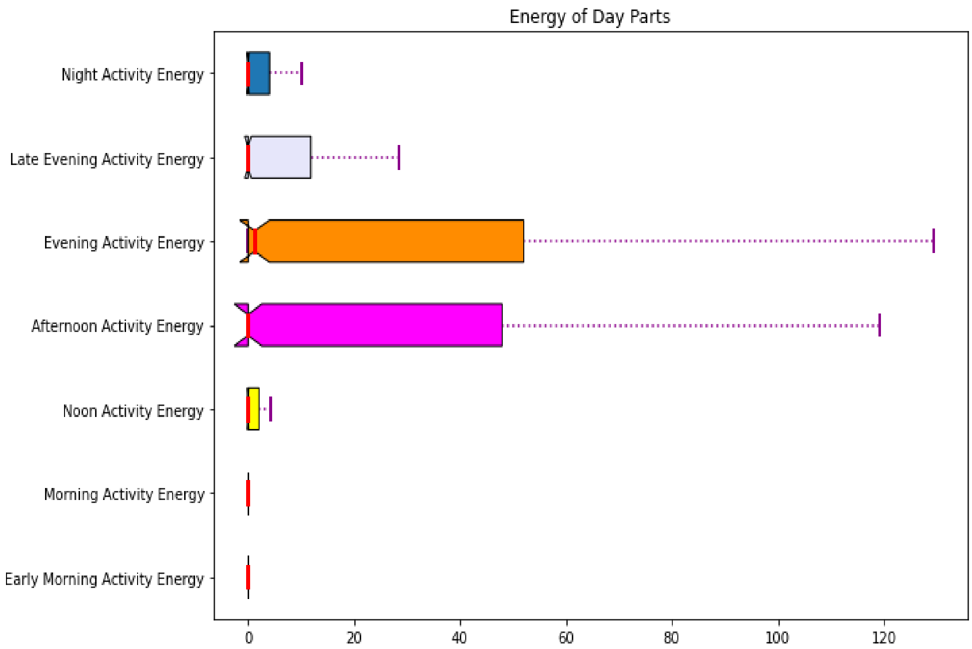


Figure A. 47: Average Energy Burnt During Physical Activity in Different Parts of the Day in Minutes

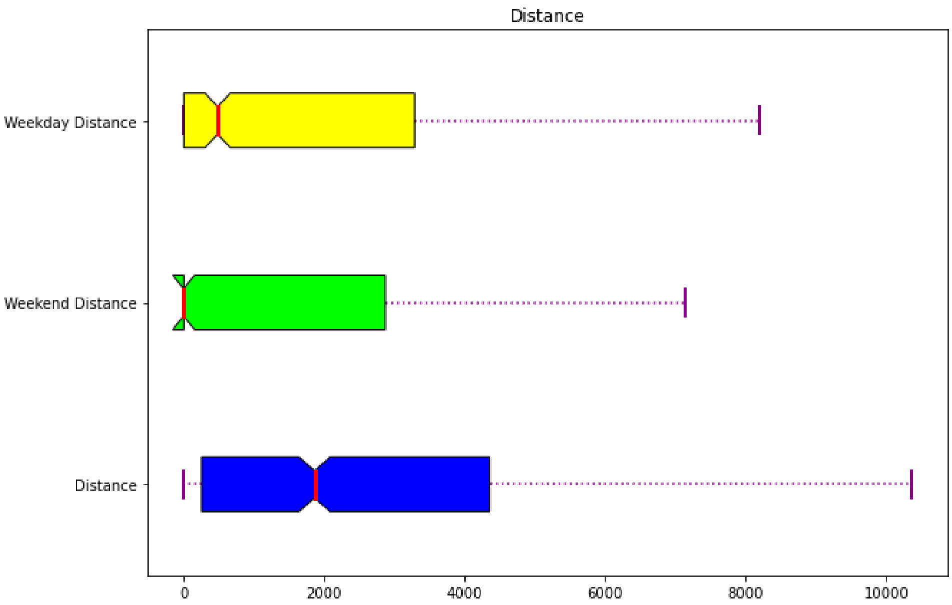


Figure A. 48: Average Distance Passed During Physical Activity in Weekend and Weekday in Minutes

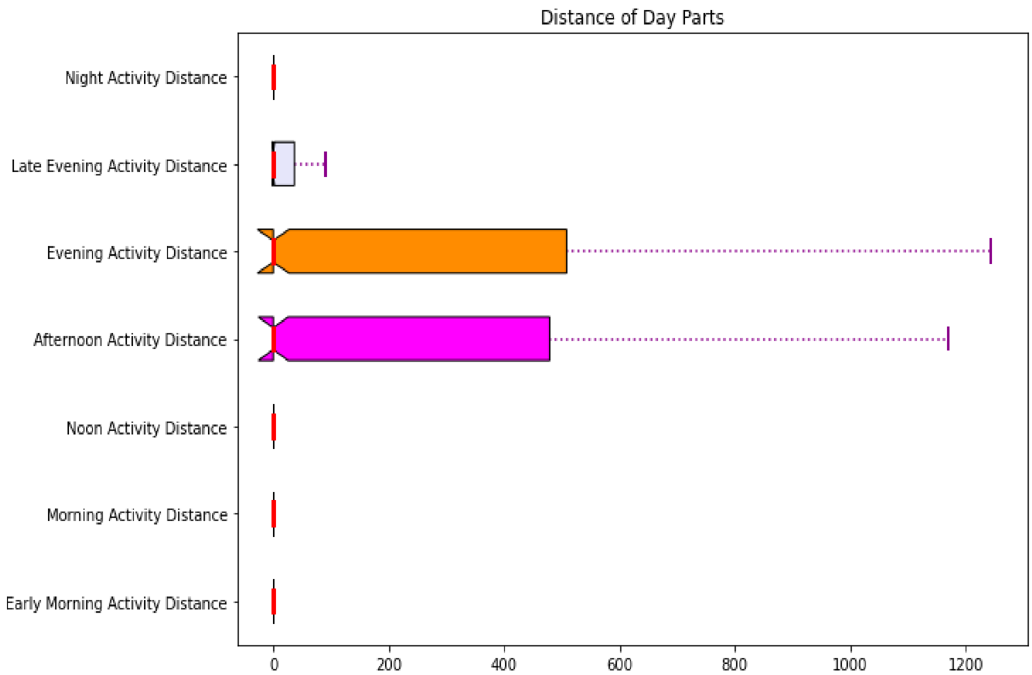


Figure A. 49: Average Distance Passed During Physical Activity in Different Parts of the Day in Minutes

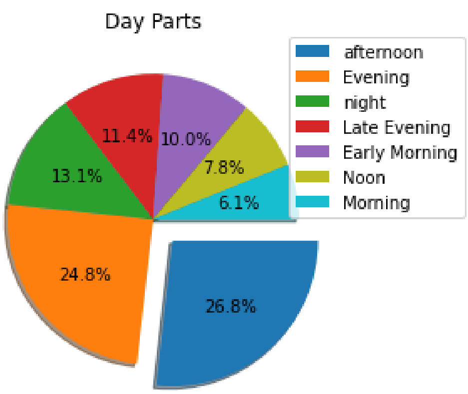


Figure A. 50: Proportion of Being Mostly Physically Active in Different Parts of the Day

A.8. Healthkit Data

Table A. 13: HealthKit Data Attributes

name	definition	values
Numeric attributes		
duration	Mean duration of being physically active per day	count 4919 mean 9.66 std 83.53 min 0 25% 1.07 50% 1.59 75% 2.57 max 1439.98
heart_rate	Mean heart rate during physical activity per day	count 4919 mean 0.10 std 0.24 min 0 25% 0 50% 0.01 75% 0.09 max 2.15
steps	Mean number of steps during physical activity per day	count 4919 mean 36.37 std 124.29 min 0 25% 9.47 50% 20.10 75% 38.18 max 3589.56
energy	Mean amount of energy burnt during physical activity per day	count 4919 mean 1229.99 std 14391.79 min 0 25% 0 50% 0.35 75% 295.28 max 506113.87
distance	Mean amount of distance passed during physical activity per day	count 4919 mean 109.43 std 5447.57 min 0 25% 6.94 50% 15.22 75% 28.75 max 381976.67
freq	Mean number of times the user had a physical activity per day	count 4919 mean 0.29 std 3.79 min 0 25% 0 50% 0 75% 0 max 117

weekend_steps	Average number of steps during physical activity in the weekend	count 4919 mean 36.36 std 226.06 min 0 25% 0 50% 11.33 75% 30.38 max 7951.83
weekday_steps	Average number of steps during physical activity in weekdays	count 4919 mean 37.36 std 148.78 min 0 25% 7.93 50% 18.5 75% 37.02 max 4381
early_morning_steps	Average number of steps during physical activity in early mornings	count 4919 mean 5.67 std 68.36 min 0 25% 0 50% 0.15 75% 1.70 max 2606.85
morning_steps	Average number of steps during physical activity in the morning	count 4919 mean 1.63 std 7.43 min 0 25% 0 50% 0.04 75% 1.16 max 412.18
noon_steps	Average number of steps during physical activity at noon	count 4919 mean 2.45 std 5.4 min 0 25% 0 50% 0.64 75% 2.61 max 130.75
afternoon_steps	Average number of steps during physical activity in the afternoon	count 4919 mean 7.35 std 15.06 min 0 25% 1.17 50% 3.76 75% 8.62 max 516.88
evening_steps	Average number of steps during physical activity in the evening	count 4919 mean 6.56 std 9.31 min 0 25% 1.01 50% 3.59 75% 8.72 max 142.6

late_evening_steps	Average number of steps during physical activity at evenings	count 4919 mean 4.14 std 10.97 min 0 25% 0.20 50% 1.61 75% 4.98 max 558.36
night_steps	Average number of steps during physical activity at night	count 4919 mean 8.57 std 97.66 min 0 25% 0.02 50% 1.19 75% 4.81 max 3585.48
weekend_distance	Average amount of distance passed during physical activity in the weekends	count 4919 mean 42.54 std 609.81 min 0 25% 0 50% 8.07 75% 22.53 max 32985.13
weekday_distance	Average amount of distance passed during physical activity in the weekdays	count 4919 mean 133.92 std 6933.41 min 0 25% 5.43 50% 13.88 75% 27.62 max 486089.06
early_morning_distance	Average amount of distance passed during physical activity in early morning	count 4919 mean 5.76 std 64.21 min 0 25% 0 50% 0.09 75% 1.29 max 1937.65
morning_distance	Average amount of distance passed during physical activity in the morning	count 4919 mean 1.54 std 11.99 min 0 25% 0 50% 0.01 75% 0.93 max 558.44
noon_distance	Average amount of distance passed during physical activity at noon	count 4919 mean 63.27 std 4296.98 min 0 25% 0 50% 0.4 75% 1.95 max 301372.72

afternoon_distance	Average amount of distance passed during physical activity in the afternoon	count 4919 mean 23.28 std 1150 min 0 25% 0.72 50% 2.77 75% 6.4 max 80594.54
evening_distance	Average amount of distance passed during physical activity in the afternoon	count 4919 mean 6.14 std 39.4 min 0 25% 0.57 50% 2.65 75% 6.43 max 2089.16
late_evening_distance	Average amount of distance passed during physical activity at late evening	count 4919 mean 3.08 std 12.85 min 0 25% 0.09 50% 1.06 75% 3.50 max 441.11
night_distance	Average amount of distance passed during physical activity at night	count 4919 mean 6.37 std 60.43 min 0 25% 0 50% 0.73 75% 3.46 max 2116.34
weekend_energy	Average amount of energy burnt during physical activity at weekend	count 4919 mean 1923.21 std 29608.45 min 0 25% 0 50% 0 75% 107.21 max 1348600
weekday_energy	Average amount of energy burnt during physical activity at week day	count 4919 mean 1232.17 std 13981.35 min 0 25% 0 50% 0.23 75% 277.11 max 513083.86
early_morning_energy	Average amount of energy burnt during physical activity at early morning	count 4919 mean 378.2 std 8987.16 min 0 25% 0 50% 0 75% 1.98 max 434989.74

morning_energy	Average amount of energy burnt during physical activity in the morning	count 4919 mean 47.88 std 1634.65 min 0 25% 0 50% 0 75% 0.42 max 113340.50
noon_energy	Average amount of energy burnt during physical activity at noon	count 4919 mean 42.1 std 473.32 min 0 25% 0 50% 0 75% 11.58 max 19740.59
afternoon_energy	Average amount of energy burnt during physical activity in the afternoon	count 4919 mean 111.91 std 1306.68 min 0 25% 0 50% 0 75% 55.79 max 54021.67
evening_energy	Average amount of energy burnt during physical activity in the evening	count 4919 mean 81.09 std 735.61 min 0 25% 0 50% 0 75% 44.45 max 43000
late_evening_energy	Average amount of energy burnt during physical activity in late evening	count 4919 mean 52.11 std 1070.84 min 0 25% 0 50% 0 75% 9.24 max 72092.43
night_energy	Average amount of energy burnt during physical activity at night	count 4919 mean 516.71 std 10607.74 min 0 25% 0 50% 0 75% 1.85 max 505536.88
weekend_duration	Average amount of time spent on physical activity in the weekend	count 4919 mean 14.45 std 119.58 min 0 25% 0 50% 1.27 75% 2.39 max 1439.98

weekday_duration	Average amount of time spent on physical activity in the weekday	count 4919 mean 11.01 std 93.26 min 0 25% 1 50% 1.53 75% 2.55 max 1439.98
early_morning_time	Average amount of time spent on physical activity in early morning	count 4919 mean 3.75 std 54.44 min 0 25% 0 50% 0.03 75% 0.18 max 1436.01
morning_time	Average amount of time spent on physical activity in the morning	count 4919 mean 0.11 std 0.35 min 0 25% 0 50% 0.01 75% 0.14 max 16.11
noon_time	Average amount of time spent on physical activity at noon	count 4919 mean 0.17 std 0.32 min 0 25% 0 50% 0.08 75% 0.21 max 5.9
afternoon_time	Average amount of time spent on physical activity in the afternoon	count 4919 mean 0.62 std 4.18 min 0 25% 0.15 50% 0.34 75% 0.62 max 221.26
evening_time	Average amount of time spent on physical activity in the evening	count 4919 mean 0.48 std 0.86 min 0 25% 0.15 50% 0.34 75% 0.62 max 31.47
late_evening_time	Average amount of time spent on physical activity in late evening	count 4919 mean 0.32 std 1.45 min 0 25% 0.05 50% 0.18 75% 0.39 max 74.31

night_time	Average amount of time spent on physical activity at night	count 4919 mean 4.21 std 60.13 min 0 25% 0.01 50% 0.16 75% 0.43 max 1439.98
weekend_count	Average number of times of physical activity in the weekend	count 4919 mean 0.73 std 0.44 min 0 25% 0 50% 1 75% 1 max 1
weekday_count	Average number of times of physical activity in the weekday	count 4919 mean 0.95 std 0.22 min 0 25% 1 50% 1 75% 1 max 2
early_morning_count	Average number of times of physical activity in early morning	count 4919 mean 0.08 std 0.13 min 0 25% 0 50% 0.02 75% 0.10 max 1
morning_count	Average number of times of physical activity in the morning	count 4919 mean 0.06 std 0.09 min 0 25% 0 50% 0.01 75% 0.09 max 0.95
noon_count	Average number of times of physical activity at noon	count 4919 mean 0.08 std 0.10 min 0 25% 0 50% 0.06 75% 0.13 max 1
afternoon_count	Average number of times of physical activity in the afternoon	count 4919 mean 0.24 std 0.17 min 0 25% 0.13 50% 0.23 75% 0.33 max 1

evening_count	Average number of times of physical activity in the evening	count 4919 mean 0.24 std 0.16 min 0 25% 0.13 50% 0.23 75% 0.32 max 1
late_evening_count	Average number of times of physical activity in the late evening	count 4919 mean 0.14 std 0.13 min 0 25% 0.04 50% 0.12 75% 0.21 max 1
night_count	Average number of times of physical activity at night	count 4919 mean 0.15 std 0.17 min 0 25% 0.01 50% 0.11 75% 0.23 max 1
number_of_days	Average number of days a physical activity was recorded for the user	count 4919 mean 7043.03 std 18930.13 min 2 25% 534 50% 1534 75% 5488.5 max 352249
Categorical attributes		
day_part	Part of the day the user was mostly active in	afternoon 1646 evening 1476 night 804 late evening 390 early morning 311 noon 161 morning 131

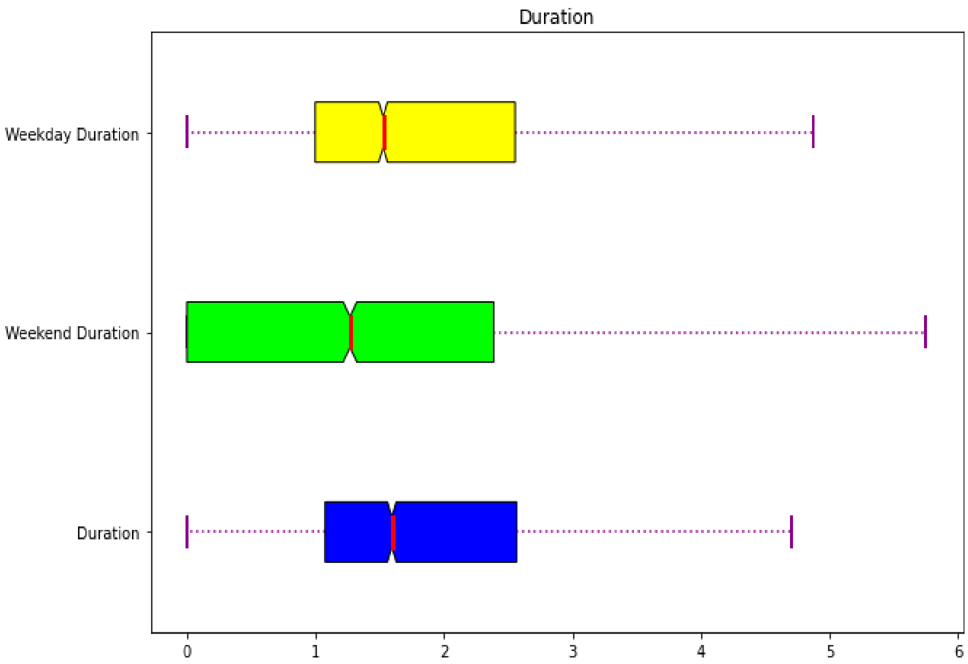


Figure A. 51: Average Physical Activity Duration During Weekend and Weekday in Minutes

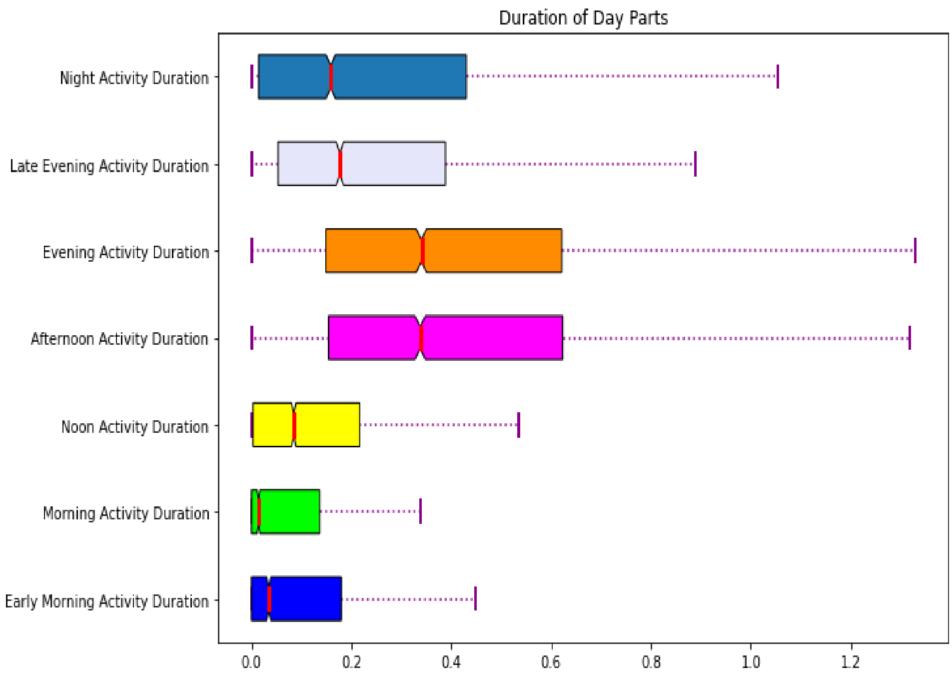


Figure A. 52: Average Physical Activity Duration in Different Parts of the Day in Minutes

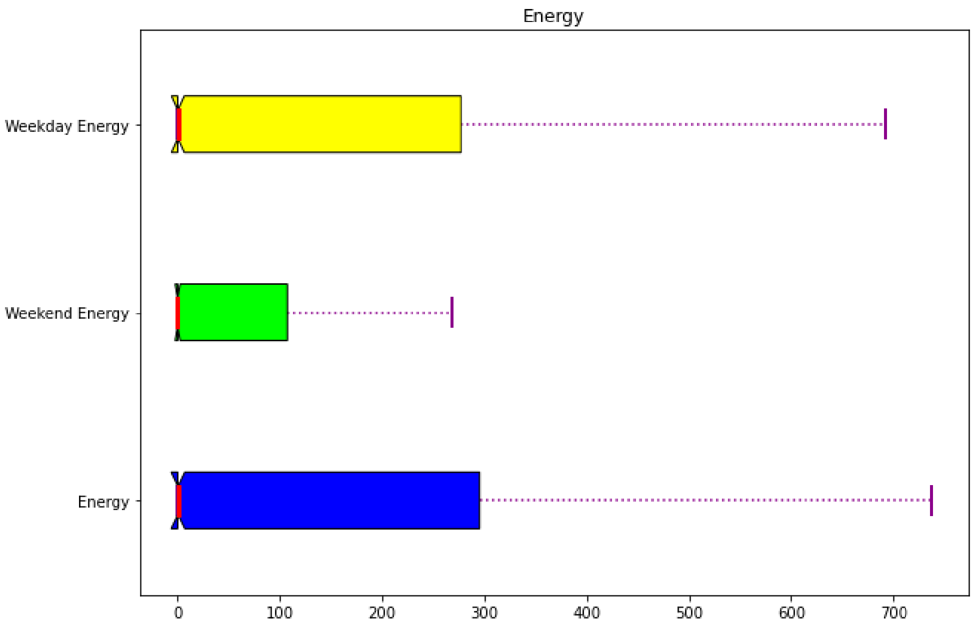


Figure A. 53: Average Energy Burnt During Physical Activity in Weekend and Weekday in Minutes

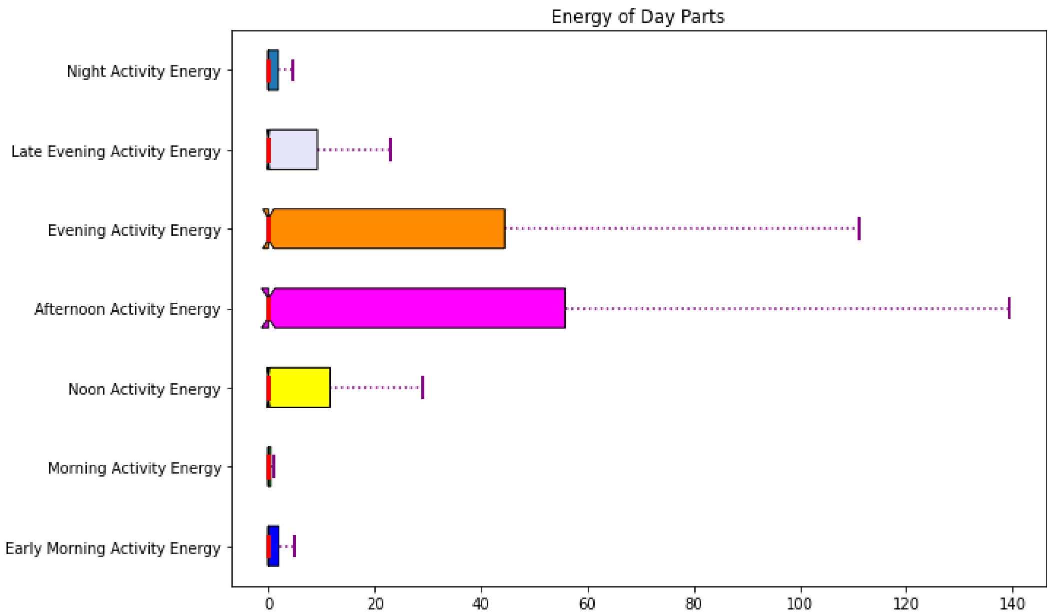


Figure A. 54: Average Energy Burnt During Physical Activity in Different Parts of the Day in Minutes

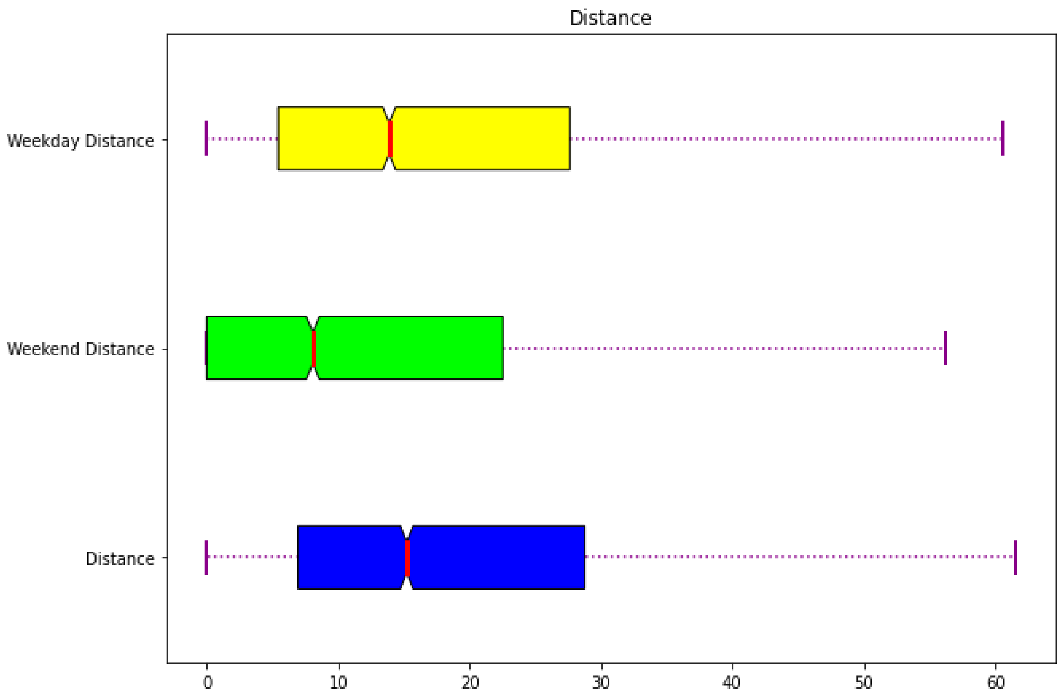


Figure A. 55: Average Distance Passed During Physical Activity in Weekend and Weekday in Minutes

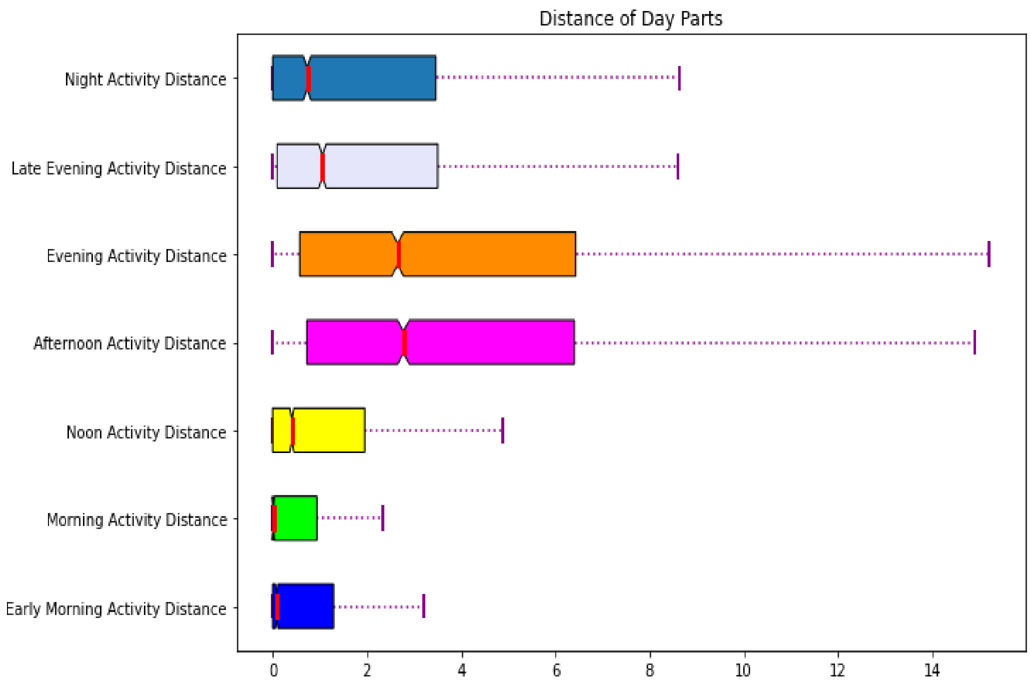


Figure A. 56: Average Distance Passed During Physical Activity in Different Parts of the Day in Minutes

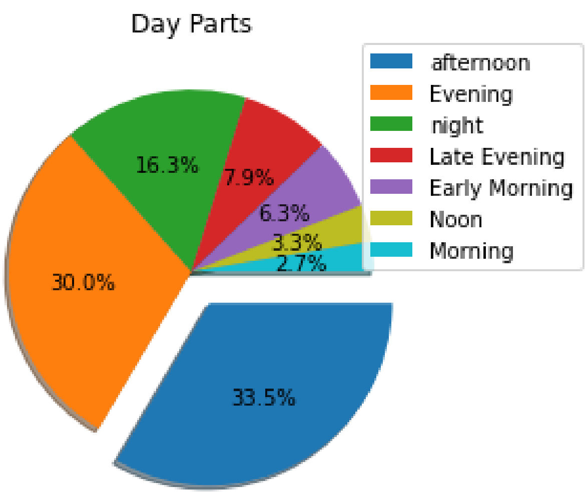


Figure A. 57: Proportion of Being Mostly Physically Active in Different Parts of the Day

A.8. Six Minute Walk

Table A. 14: Six Minutes Walk Attributes

Name	Explanation	Values
Displacement	Distance passed during 6 minutes walk test	count 339 mean 939.483908 std 994.475863 min 0.778227 25% 589.300614 50% 699.532600 75% 936.443774 max 9408.320327

A.9. Daily Check

Table A. 15: Daily Check Attributes

Column name	Question	Values
phone_on_user	In the last 24 hours, how often did you have your phone or wearable device with you?	1=All day and all night; 2=All day, but not at night; 3=About half of the time; 4=Rarely if at all
activity1_option	Did you perform any physical activities yesterday that you think were not recorded by your phone or wearable device?	boolean
activity1_type	Which activity did you do that may have been improperly recorded?	1=Walking; 2=Jogging; 3=Cycling; 4=Tennis or other racquet sport; 5=Soccer, basketball, or other team sport; 6=Weight-lifting; 7=Swimming
activity1_time	How long did you do the activity?	duration

activity1_intensity	How intense was the activity?	1=Light; 2=Moderate; 3=Vigorous
activity2_option	Did you perform any additional physical activities yesterday that you think were not recorded by your phone or wearable device?	boolean
activity2_type	Which activity did you do that may have been improperly recorded?	1=Walking; 2=Jogging; 3=Cycling; 4=Tennis or other racquet sport; 5=Soccer, basketball, or other team sport; 6=Weight-lifting; 7=Swimming
activity2_time	How long did you do the activity?	duration
activity2_intensity	How intense was the activity?	1=Light; 2=Moderate; 3=Vigorous
sleep_time	How many hours of sleep did you get last night?	duration

Table A. 16: Daily Activity Final Table

Column name	Explanation	Statistics
sleep	Sum of sleep times entered by the user	Mean: 174816,6 Std: 360596,9 Min: 0 25%: 28800 50%: 77460 75%: 180120 Max: 6539940
number_of_days	Number of days the user filled the form	Mean: 7,75 Std: 14.54 Min: 1 25%: 1 50%: 4 75%: 8 Max: 227
activity_dasys	Number of days the user filled at least one activity	Mean: 0,45 Std: 1,89 Min: 0 25%: 0 50%: 0 75%: 0 Max: 99
filled_twice	Number of times the user added two activities	Mean: 0,03 Std: 0,32 Min: 0 25%: 0 50%: 0 75%: 0 Max: 21
Light_intensity_count	Number of times the user added at least one light intensity activity	Mean: 0,53 Std: 2,21 Min: 0 25%: 0 50%: 0 75%: 0 Max: 104
Moderate_intensity_count	Number of times the user added at least one moderate intensity activity	Mean: 0,16 Std: 0,958597 Min: 0 25%: 0 50%: 0

		75%: 0 Max: 69
Vigorous_intensity_count	Number of times the user added at least one vigorous intensity activity	Mean: 0,18 Std: 1,4 Min: 0 25%: 0 50%: 0 75%: 0 Max: 58
Light_intensity_time	Sum of the duration the user added for light intensity activity	Mean: 1969,59 Std: 17416,95 Min: 0 25%: 0 50%: 0 75%: 0 Max: 1763220
Moderate_intensity_time	Sum of the duration the user added for moderate intensity activity	Mean: 346.7 Std: 4012.01 Min: 0 25%: 0 50%: 0 75%: 0 Max: 425940
Vigorous_intensity_time	Sum of the duration the user added for vigorous intensity activity	Mean: 574.52 Std: 935.72 Min: 0 25%: 0 50%: 0 75%: 0 Max: 384660
All_day_night_phone_use	Number of times user indicated using their phone all day and night	Mean: 3.17 Std: 9.15 Min: 0 25%: 0 50%: 1 75%: 3 Max: 193
All_day_phone_use2	Number of times user indicated using their phone all day long	Mean: 3.49 Std: 9.72 Min: 0 25%: 0 50%: 1 75%: 3 Max: 186
half_of_the_time_phone_use	Number of times user indicated using their phone half of the time	Mean: 0.89 Std: 3.56 Min: 0 25%: 0 50%: 0 75%: 1 Max: 132
Rarely_phone_use4	Number of times user indicated using their phone rarely	Mean: 0.16 Std: 1.29 Min: 0 25%: 0 50%: 0

		75%: 0 Max: 91

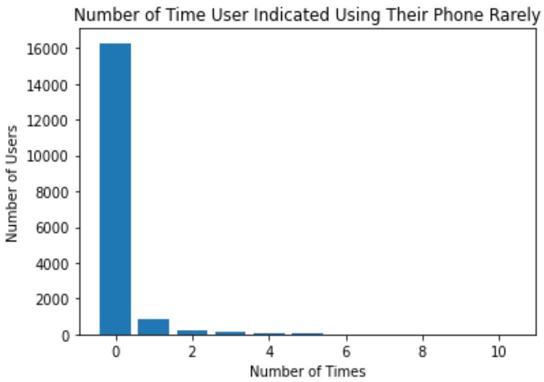


fig.a

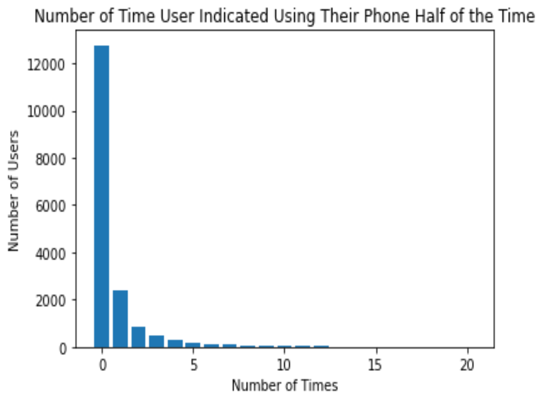


fig.b

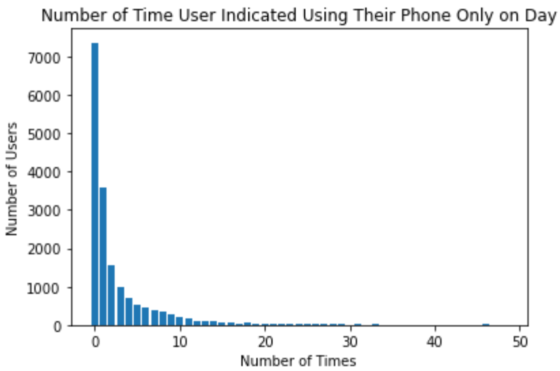


fig.c

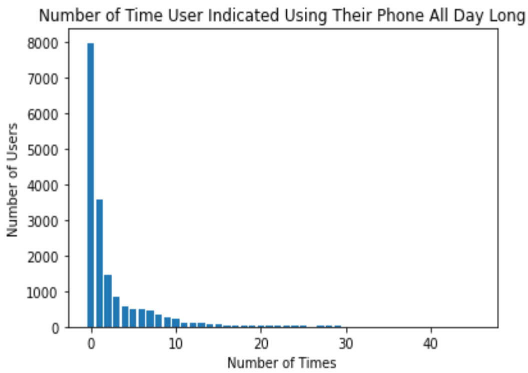


fig.d

Figure A. 58: Pattern of Using Phone During the Day in Participants

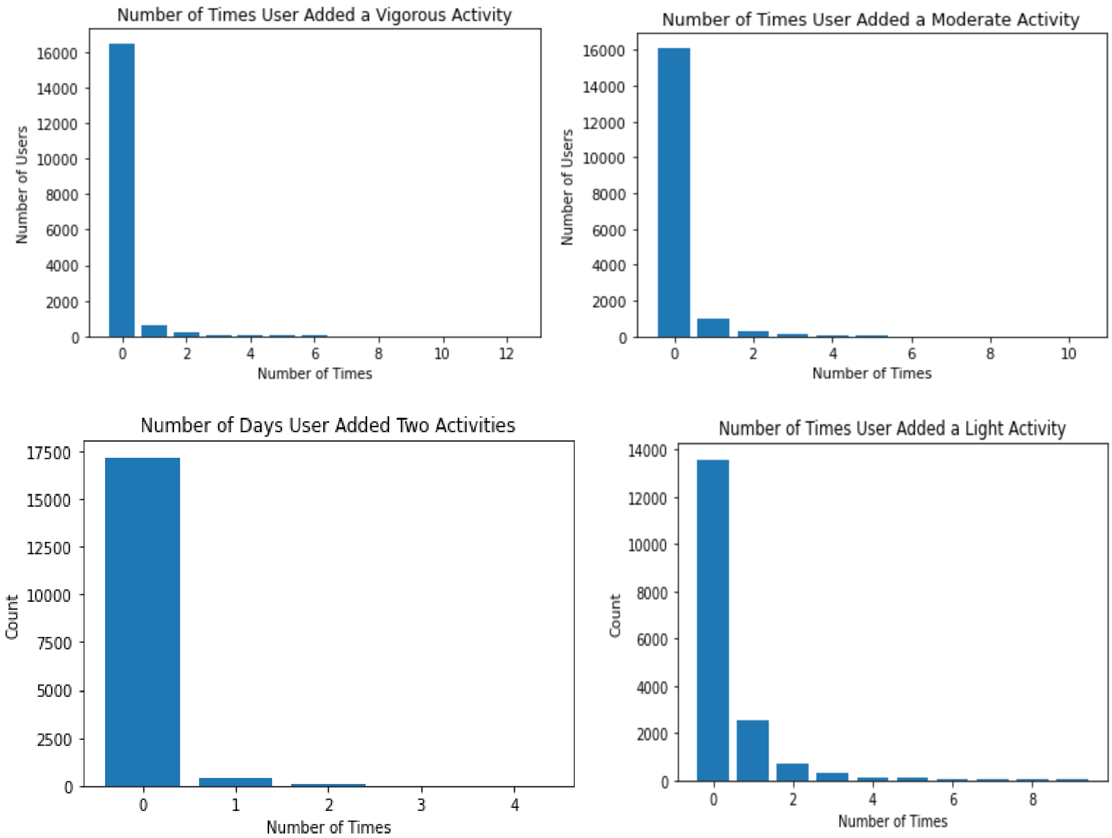


Figure A. 59: Patterns of Entering Physical Activity to the Daily Check Survey in Participants

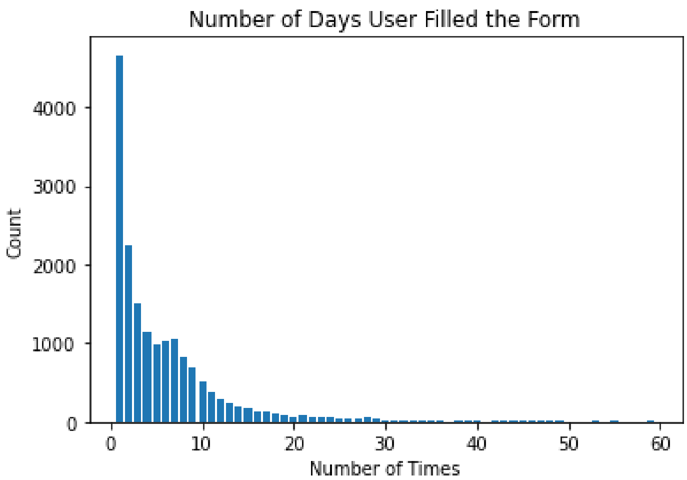


Figure A. 60

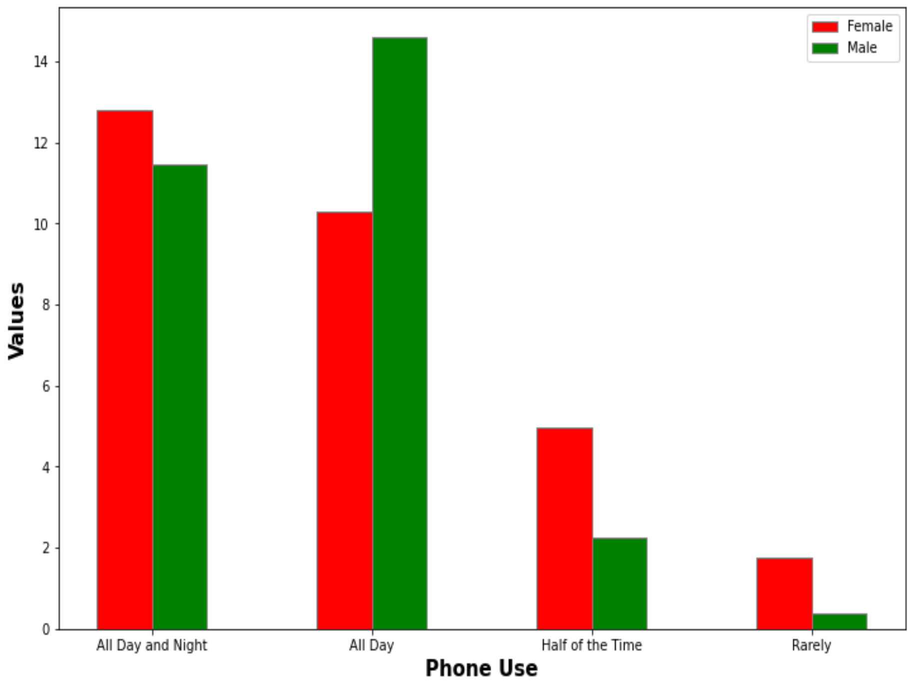


Figure A. 61: Average Number of Times Participants Indicated Using Their Phone "All Day and Night", "All Day", "Half of the Time" or "Rarely" based on Gender

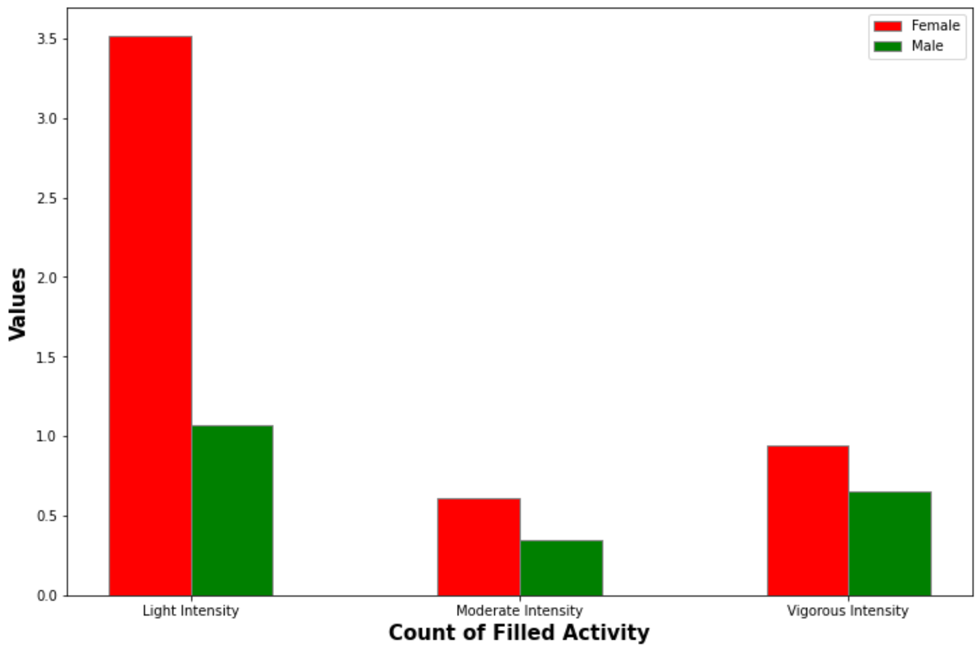


Figure A. 62: Average Number of Times Participants Indicated Having "Light Intensity Activity", "Moderate Intensity Activity" or "Vigorous Intensity Activity" based on Gender

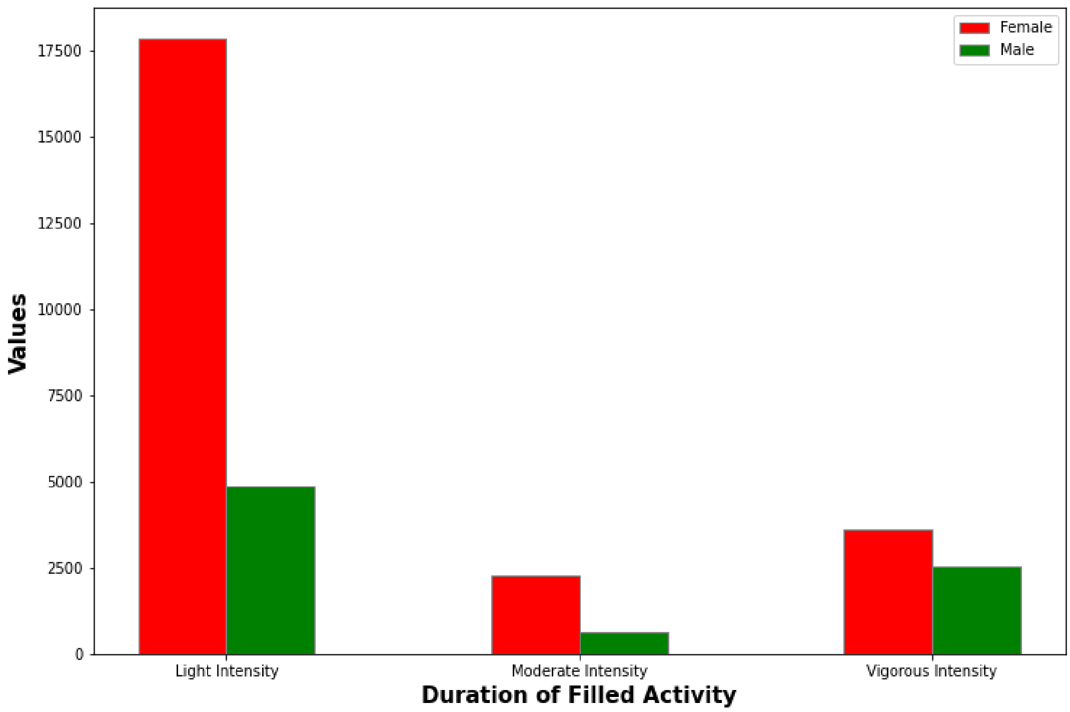


Figure A. 63: Figure A. 62: Average Duration of Physical Activity per Week in Minutes based on Gender

A.10. Motion Tracker

Table17: Motion Tracker Attributes

Name	Explanation	Statistics
unknown_count	Number of times an unknown state been recorded for the user per day	count 12043 mean 19166.05 std 15919.63 min 0 25% 9921 50% 13972 75% 23791.50 max 245124
unkown_time	Average duration of an unknown state been recorded for the user per day(in seconds)	count 12043 mean 2551343 std 1.6293600 min 0 25% 13680.46 50% 19467.57 75% 192493.8 max 1659283000
walking_count	Number of times a walking state been recorded for the user per day	count 12043 mean 6416.72 std 6327.04 min 0 25% 2847 50% 4523

		75% 7886 max 126613
walking_time	Average duration of a walking state been recorded for the user per day(in seconds)	count 12043 mean 846558.3 std 3726300 min 0 25% 2220.90 50% 3224.80 75% 4706.886 max 46679690
running_count	Number of times a running state been recorded for the user per day	count 12043 mean 167.13 std 388.05 min 0 25% 18 50% 60 75% 165 max 12860
running_time	Average duration of a running state been recorded for the user per day(in seconds)	count 12043 mean 24948.81 std 637949.6 min 0 25% 3.01 50% 13.38 75% 69.58 max 40479840
stationary_count	Number of times a stationary state been recorded for the user per day	count 12043 mean 13519.70 std 11482.27 min 0 25% 6798 50% 9909 75% 16525 max 146208
stationary_time	Average duration of a stationary state been recorded for the user per day(in seconds)	count 12043 mean 2400505 std 6070389 min 0 25% 44729.74 50% 52797.20 75% 177360.5 max 56400560
cycling_count	Number of times a cycling state been recorded for the user per day	count 12043 mean 2550.70 std 3309.13 min 0 25% 625.50 50% 1678 75% 3289.50 max 70266
cycling_time	Average duration of a cycling state been recorded for the user per day(in seconds)	count 12043 mean 352410 std 2428342 min 0 25% 370.17 50% 894.75

		75% 1512.54 max 419545700
weekend_count_core	Number of times a user was active during weekends	count 12043 mean 0.70 std 0.12 min 0 25% 0.67 50% 0.71 75% 0.75 max 1
weekend_duration_core	Average duration of being active during weekends (in seconds)	count 12043 mean 1682.37 std 7085.67 min 0 25% 9.60 50% 12.51 75% 19.57 max 144427.97
weekday_count_core	Number of times a user was active during week days	count 12043 mean 2.00 std 62.55 min 753.00 25% 2.39 50% 1.86 75% 1.47 max 3705.67
weekday_duration_core	Average duration of being active during weekdays (in seconds)	count 12043 mean 132752.4 std 15120190 min 5880081 25% 2.575.59 50% 40.49 75% 24.97 max 1659283000
early_morning_count	Number of times user was active in the early morning in each day	count 12043 mean 119.33 std 148.32 min 0 25% 17.44 50% 66.16 75% 168 max 1422.89
early_morning_time	Duration of being active during early morning in each day (in seconds)	count 12043 mean 242123.4 std 2090777 min 0 25% 326.35 50% 1309.75 75% 3176.31 max 45456840
morning_count	Number of times user was active in the morning in each day	count 12043 mean 46.27 std 74.03 min 0 25% 3.12 50% 18.42

		75% 54.20 max 843
morning_time	Duration of being active during morning in each day (in seconds)	count 12043 mean 91127.10 std 1263133 min 0 25% 45.92 50% 326.07 75% 1056.20 max 56363770
noon_count	Number of times user was active at noon in each day	count 12043 mean 97.16 std 104.94 min 0 25% 18.58 50% 67.55 75% 144.55 max 1434.76
noon_time	Duration of being active during noon in each day (in seconds)	count 12043 mean 244171.8 std 2061246 min 0 25% 322.13 50% 1229.15 75% 2407.84 max 46675680
afternoon_count	Number of times user was active in the afternoon in each day	count 12043 mean 364.90 std 238.82 min 0 25% 209.94 50% 328.81 75% 472.68 max 4578.19
afternoon_time	Duration of being active during afternoon in each day (in seconds)	count 12043 mean 917906.3 std 15526530 min 0 25% 3644.441 50% 5249.231 75% 7206.093 max 1659283000
evening_count	Number of times user was active in the evening in each day	count 12043 mean 449.67 std 258.18 min 0 25% 291.34 50% 415.83 75% 563.30 max 4715.71
evening_time	Duration of being active during evening in each day (in seconds)	count 12043 mean 940529 std 3906218 min 0 25% 4892.76 50% 6438.50

		75% 8408.17 max 43035160
late_evening_count	Number of times user was active in the late evening in each day	count 12043 mean 330.65 std 195.27 min 0 25% 212.36 50% 306.50 75% 417.97 max 3131.90
late_evening_time	Duration of being active during late evening in each day (in seconds)	count 12043 mean 820431 std 3615377 min 0 25% 3782.11 50% 4980.25 75% 6358.91 max 50358560
night_count	Number of times user was active at night in each day	count 12043 mean 414.91 std 276.38 min 0 25% 232.22 50% 383.15 75% 545.25 max 4627.19
night_time	Duration of being active during afternoon in each day (in seconds)	count 12043 mean 948201 std 3804378 min 0 25% 4425.21 50% 6728.79 75% 9380.18 max 47620770
active_time	Number of times user was active in each day	count 12043 mean 4204490 std 16986300 min 0 25% 24733.35 50% 33805.33 75% 5002495 max 1659283000
change_of_position	Number of times user changed their position in each day	count 12043 mean 700.62 std 258.44 min 0 25% 537.19 50% 692.64 75% 853.18 max 2238.85
day_filled	Number of days users data been recorded	count 12043 mean 17.85 std 11.31 min 1 25% 10 50% 15

		75% 22 max 141
data_collection_duration	Duration of data collection (in month)	count 12043 mean 62.22 std 80.85 min 0 25% 0.39 50% 13.77 75% 109.89 max 630.97

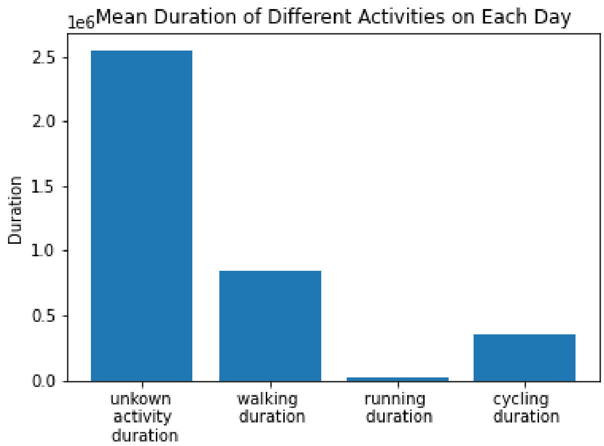


fig.a

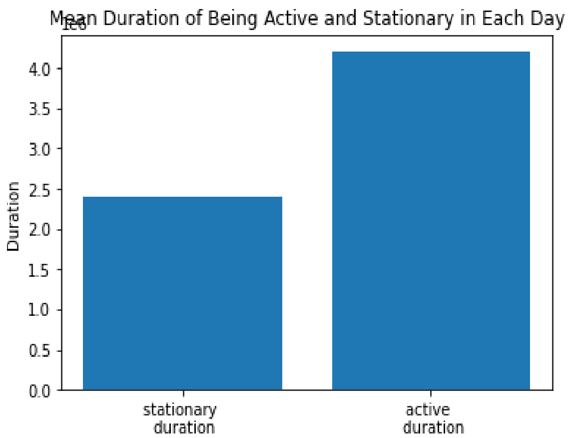


fig.b

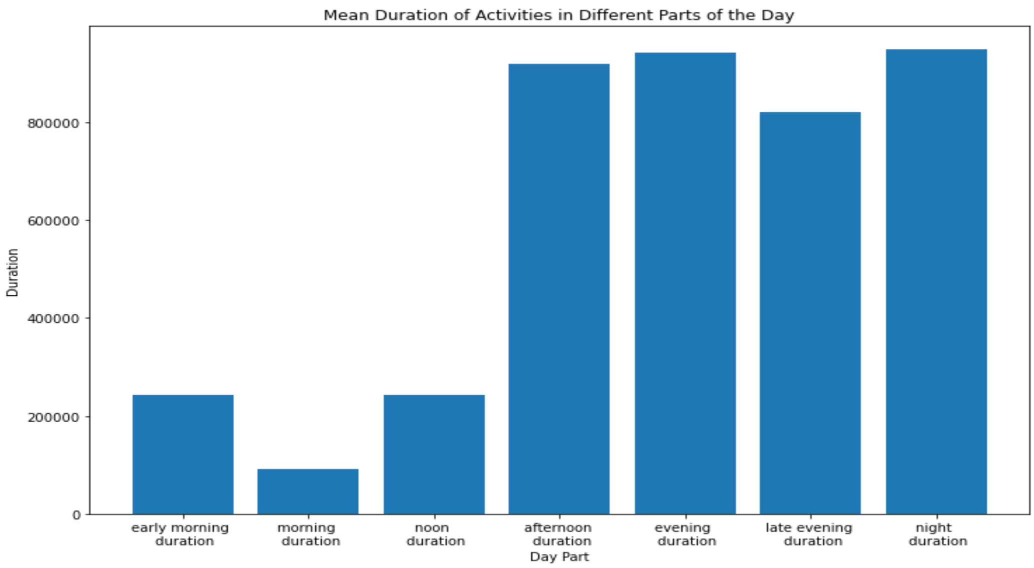


fig.c

Figure A. 64: a: Mean duration of different activities on each day, b: mean duration of activities in different parts of the day, c: mean duration of being active and stationary in each day

A.11. Joined Data

Table A. 17: Categorical Attributes of Final Data

Name	count	unique	top	freq	categories
heartAgeDataEthnicity	3702	8	0	2881	0(White) 2881 1(Asian) 276 2(Hispanic) 274 7(Other) 122 3(Black) 119 4(American Indian) 18 5(Pacific Islander) 10 6(Alaska Native) 2
atwork	10710	5	0	8252	0: I spent most of the day sitting or standing) 8252 1: I spent most of the day walking or using my hands and arms in work that required moderate exertion 2133 3: I spent most of the day doing hard physical labor 270 4: None 55 2: I spent most of the day lifting or carrying heavy objects or moving most of my body in some other way 0
phys_activity	10710	7	1	3030	1: Once or twice a week, did light activities (3030) 3: Almost daily, that is five or more times a week, did moderate activities (2543) 4: About three times a week, did vigorous activities (1534) 0: did not do much physical activity (1433) 5: Almost daily, that is, five or more times a week, did vigorous activities(1342) 6: None(828) 2: About three times a week, did moderate activities (0)
sleep_diagnosis1	10702	2	0	9479	0: False 9479 1: True 1223
mostly_sit_stand	10710	2	1	7115	1: True 7115 0: False 3595
mostly_walk	10710	2	0	8873	0: False 8873 1: True 1837
mostly_lift	10710	2	0	10486	0: False 10486 1: True 224
hard_physical_activity	10710	2	0	10663	0: False 10663 1: True 47
not_much_physical_activity	10710	2	0	9279	0: False 9279 1: True 1431
once_or_twice_physical_activity	10710	2	0	7682	0: False 7682

					1: True	3028
three_times_physical_activity	10710	2	0	8171	0: False 1: True	8171 2539
daily_physical_activity	10710	2	0	9181	0: False 1: True	9181 1529
three_times_vigorous_activity	10710	2	0	9369	0: False 1: True	9369 1341
daily_vigorous_activity	10710	2	0	9884	0: False 1: True	9884 826
day_part	1556	6	4	486	4: evening 3: afternoon 6: night 5: late_evening 0: early morning 2: noon 1: morning	486 451 331 190 45 34 19
Gender	1077	2	0	901	0: Male 1: Female	901 176
Any of the issues(OR)	11691	2	0	8713	0: False 1:True	8713 2978

Table A. 18: Numeric Attributes of the Final Data

name	count	mean	std	min	25%	50%	75%	max
unkown_time_core	8783	16978.91	9777.73	0	11888.76	16218.56	20714.08	85946.70
walking_time_core	10745	3574.47	4732.56	0	2104.36	2987.75	4069.24	86208.53
running_time_core	12000	112.81	868.09	0	3	13.25	68.25	64377.45
stationary_time_core	8782	47728.12	11473.80	0	42234.26	48473.29	54597.69	86251.61
cycling_time_core	11498	1098.69	2385.60	0	335.92	850.24	1398.45	68599.96
morning_time_core	11863	725.24	984.09	0	43.98	313.60	1001.34	6920.55
noon_time_core	11595	1418.23	1257.37	0	297.70	1150	2269.10	7141
afternoon_time_core	10658	4932.45	2223.48	0	3437.91	4896.71	6388.16	13879.89
evening_time_core	10459	6008.21	2270.48	0	4661.21	6044.76	7521.27	17957
night_time_core	10486	6674.41	6417.10	0	4086.15	6156.85	8172.35	86313.14
active_time_core	7414	27096.65	10646.20	0	21237.96	26672.90	31968.49	85963.37
change_of_position	12043	700.6	258.44	0	537.19	692.64	853.18	2238.85

weekend_duration_core	12036	1617.50	6539.80	0	9.60	12.51	19.54	84183.19
weekday_duration_core	12001	4499.30	10812.44	0	25.92	42.20	3089.97	86277.14
early_morning_time_core	11617	1896.85	2015.87	0	303.67	1222.11	2903.25	14391.45
late_evening_time_core	10568	4580.19	1753.13	0	3584.90	4714.36	5787.16	10470.18
patientWeightPounds	1006	85.77	20.90	35.83	72.12	83.01	97.52	159.21
patientHeightInches	1023	175.78	9.50	149.86	170.18	177.80	182.88	198.12
moderate_act	10710	150.59	229.50	0	40	90	180	4096
sleep_time	12043	17.12	9.13	0	11	22	23	23.98
sleep_time1	10710	6.85	1.11	0	6	7	8	15
vigorous_act	10710	69.36	133.23	0	2	30	90	3600
age	3770	42.24	14.86	18	30	40	53	84
displacement	128	829.59	909.26	0	577.99	652.39	784.20	9151.05
Sleep	432	443.43	3203.833	3180.59	264.46	392.29	611.72	896.96
Light_intensity_count	10927	0.65	2.55	0	0	0	1	104
Moderate_intensity_count	10927	0.21	1.13	0	0	0	0	69
Vigorous_intensity_count	10927	0.23	1.61	0	0	0	0	54
Light_intensity_time	10927	2320.51	20820.76	0	0	0	0	1763220
Moderate_intensity_time	10927	428.59	4826.22	0	0	0	0	425940
Vigorous_intensity_time	10927	771.59	8549.57	0	0	0	0	384660
duration	1556	16.34	77.54	0	1.41	2.24	3.50	1439.98
steps	1288	40.51	127.12	0	13.80	25.16	40.75	2835.13
energy	1556	831.91	9917.34	0	0	0.03	149.12	354001.19
distance	1556	742.07	3907.05	0	10.10	21.36	40.28	78596.22
night_steps	1288	9.98	92.55	0	0.80	3	7.65	2835.13
evening_steps	1288	8.35	10.18	0	1.98	5.83	11.00	142.60
afternoon_steps	1288	7.50	11.01	0	1.17	4.44	9.60	195.25

noon_steps	1288	1.67	3.80	0	0	0.18	1.81	45.22
morning_steps	1288	0.71	4.51	0	0	0	0.13	120.50
night_distance	1556	68.71	408.13	0	0.41	2.12	6.21	5422.64
evening_distance	1556	143.18	1060.86	0	1.15	4.19	8.70	28165.80
afternoon_distance	1556	264.10	2469.25	0	0.62	3.21	7.96	67573.35
noon_distance	1556	106	1525.98	0	0	0.09	1.19	53743.98
morning_distance	1556	22.01	279.35	0	0	0	0.06	7481.42
night_energy	1556	115.12	1488.23	0	0	0	0.49	43140.58
evening_energy	1556	95.01	592.08	0	0	0	0.90	12646.89
afternoon_energy	1556	120.72	1475.19	0	0	0	1.01	54021.67
noon_energy	1556	51.61	680.75	0	0	0	0	19740.59
morning_energy	1556	21.86	384.63	0	0	0	0	14312.69
night_time	1556	4.10	47.29	0	0.12	0.31	0.64	1439.98
evening_time	1556	2.16	8.79	0	0.23	0.48	0.83	139.03
afternoon_time	1556	2.52	12.20	0	0.15	0.39	0.73	187.80
noon_time	1556	0.77	6.87	0	0	0.03	0.17	221.60
morning_time	1556	0.24	2.18	0	0	0	0.03	45.91
weekend_energy	1556	1341.87	22184.64	0	0	0	6.01	830099.83
weekend_distance	1543	342.54	1577	0	0	13.57	30.13	19577.47
weekend_duration	1556	20.91	123.46	0	0.30	1.76	3.10	1439.98
weekend_steps	1288	45.77	300.23	0	0	18.15	35.38	7951.83
weekday_energy	1556	755	7979	0	0	0.01	75.30	272153.72
weekday_distance	1556	460.40	2683.97	0	6.79	18.58	34.46	53967.8
weekday_duration	1556	15.97	88.72	0	1.22	2.08	3.23	1439.98
weekday_steps	1288	45.32	188.78	0	11.11	24.17	40.43	4381
early_morning_steps	1288	6.46	84.70	0	0	0.03	0.57	2606.85
late_evening_steps	1288	5.84	7.89	0	1.22	3.73	7.59	148.55
early_morning_time	1550	1.75	13.29	0	0	0.01	0.09	215.58
late_evening_time	1556	1.71	13.49	0	0.14	0.32	0.61	462.09
early_morning_energy	1556	374.96	9306.95	0	0	0	0	354001.22
late_evening_energy	1556	52.62	303.88	0	0	0	0.22	5672.44

early_morning_distance	1556	18.87	198.49	0	0	0.01	0.42	5399.28
late_evening_distance	1556	118.72	1344.33	0	0.55	2.59	6	47635.20
waking_time	12043	7.02	2.23	0	6	7	7.5	22

B. Result Appendix

B.1. Rule2

Before taking a closer look at rule 2, it is good to recall since the SSD++ algorithm results in a ranked list of subgroups, each subgroup can be true if the prior subgroups in the list are not true. For example, rule 2 can only apply if rule 1 is relaxed.

This rule indicates if a participant is older than 59 years(older than the upper quartile of the healthy and whole population), their running time is less than one minute per day(less than the median of the healthy and whole population), they are more active during the late evening than whole population minimum and walks less than the maximum of the whole population, the probability of having CVD or its risk factor for them is 87%. This rule is true about 182 items of the dataset. And the WKL for this rule is 187.07. Figure B 1 shows how the conditions of this rule are in comparison to the distribution of the whole and healthy population.

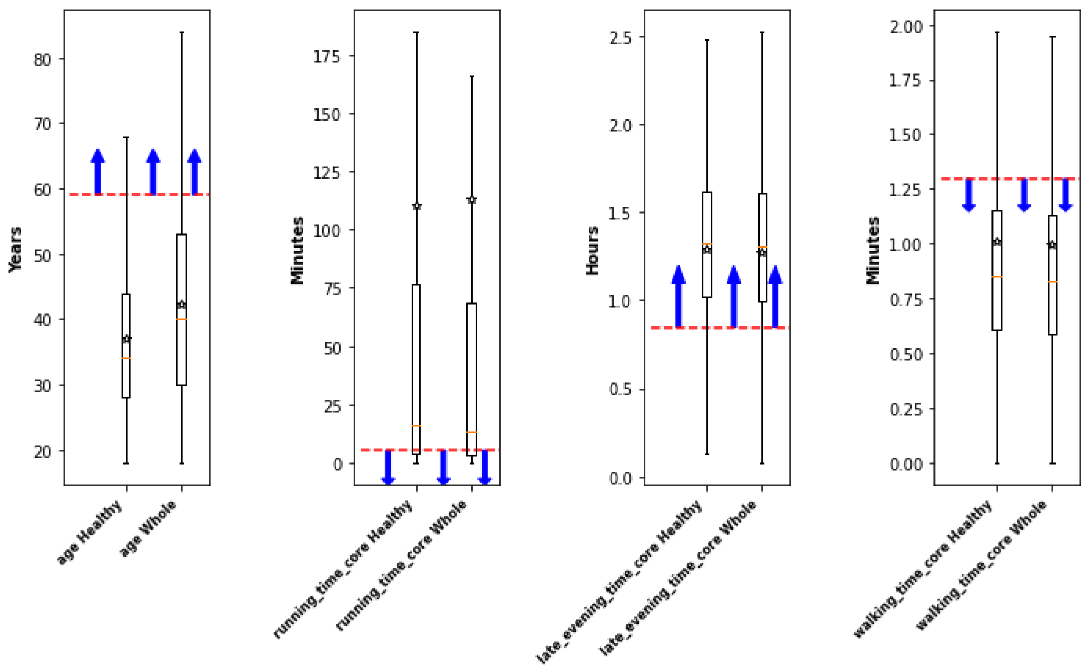


Figure B 1: Rule 2 patterns in Comparison to the Healthy and Whole data distribution

We can see in Figure B 2 that the mean and median age of subgroup 2 is higher than the healthy population. In addition, participants in subgroup 2 ran and walk more than the whole and the healthy population. They are also more active (higher mean, median and lower quartile) during the late evening (21-23:59).

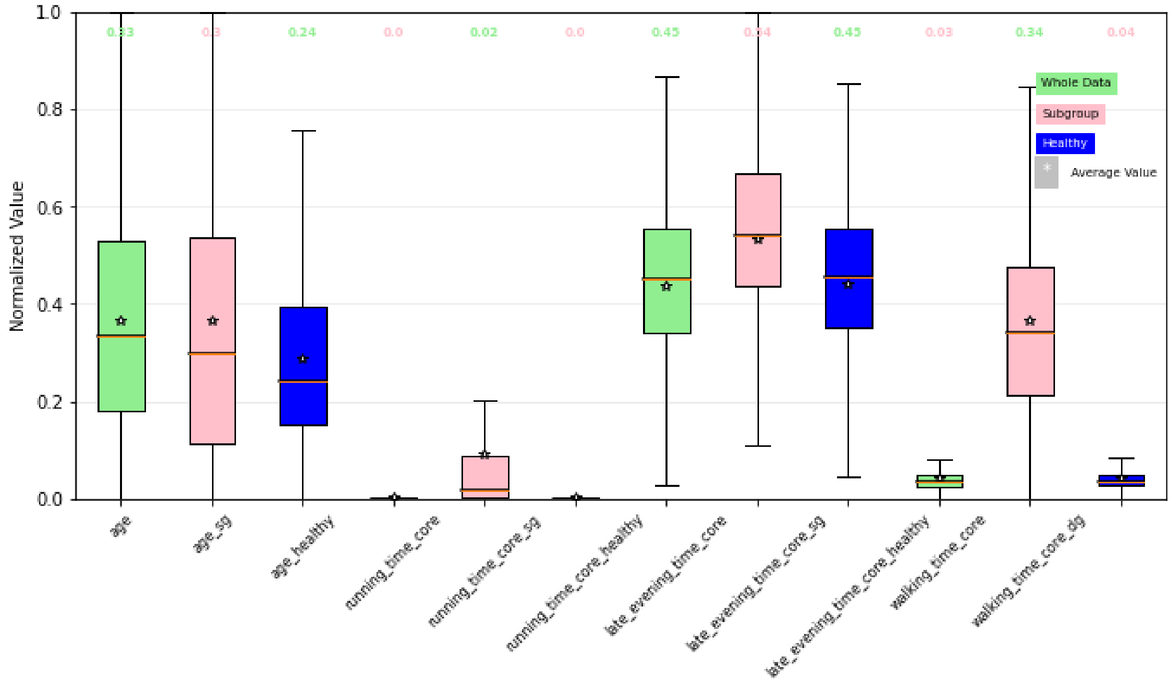


Figure B 2: Distribution Comparison of Subgroup 2 with the Healthy and Whole Population

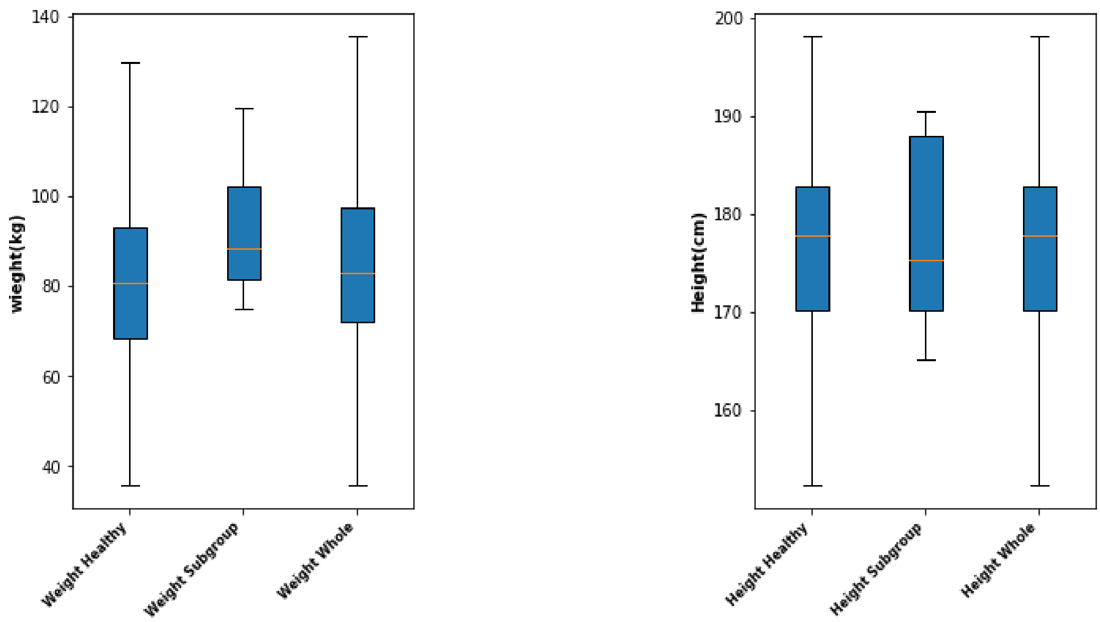


Figure B 3: Height and Weight of Participants in Different Groups of the Data

Figure B 3, **Figure B 4** and **Figure B 5** compare the demographic attributes of participants in subgroup 2, with the whole and healthy population. The median for weight attribute in this subgroup is higher than the other two groups. But this is the opposite of height. The percentage of the female participant is lower in this subgroup. In addition, participants are not from the American Indian ethnicity group in this subgroup.

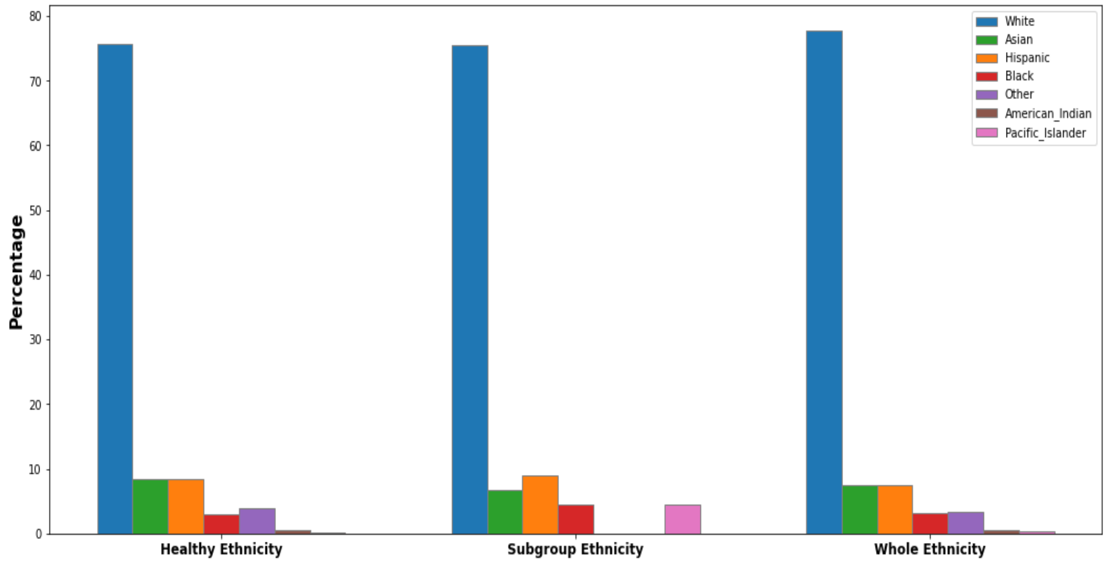


Figure B 5: Ethnicity Distribution in the Healthy, Subgroup 2 and Whole Populations

Figure B 4: Gender Distribution in the Healthy, Subgroup1 and Whole Populations

B.2. Rule4

Rule 4 has one more condition in addition to a condition on age. It indicates for a participant older than 59 and lower average duration of physical activity during the early morning(5-9 A.M.) than 20 minutes per day, there is a 67% possibility of having CVD or its risk factors. This amount of being active during the early morning is less than the median of both healthy and the whole population. Meaning 50% of the population in these two data groups have more activity during this period. The median and mean age in this subgroup are larger than the healthy and the whole population. This is also true about the early morning activity duration(**Figure B 6**)

The median of both height and weight attributes in this subgroups is lower than the two other datasets. Weight attribute distribution is skewed to the right. The proportion of female participants in this subgroup is around 2% more than the two other groups. Participant in subgroup4 are only from white ethnicity(**Figure B 9**).

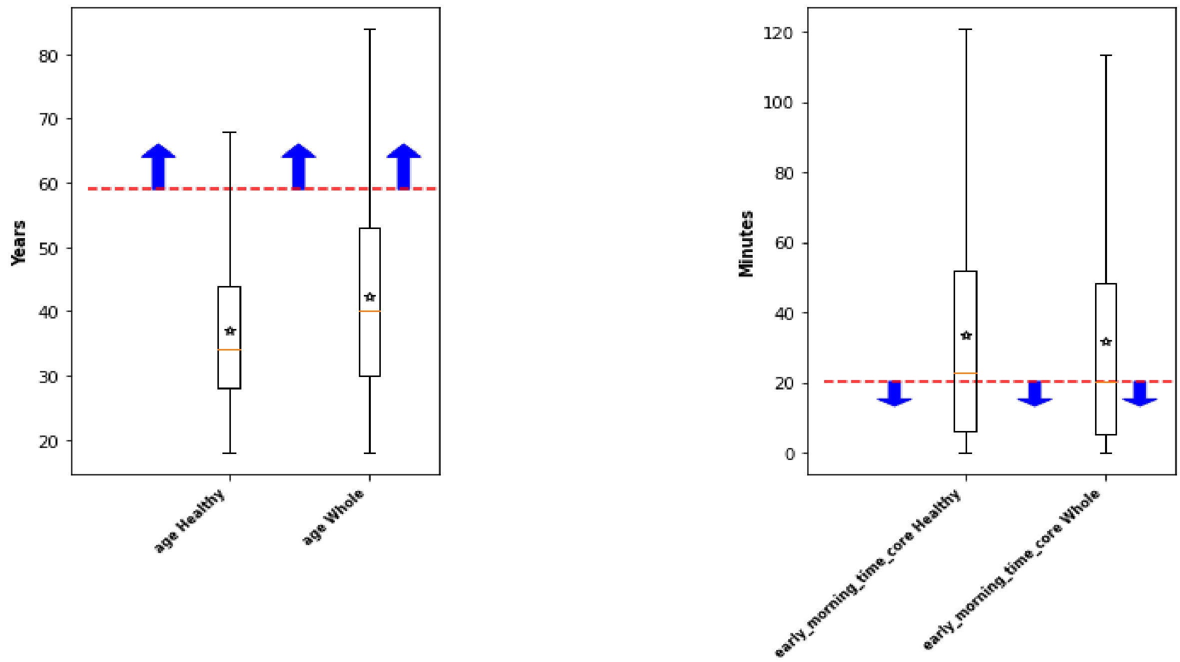


Figure B 6: Rule 4 patterns in Comparison to the Healthy and Whole data distribution

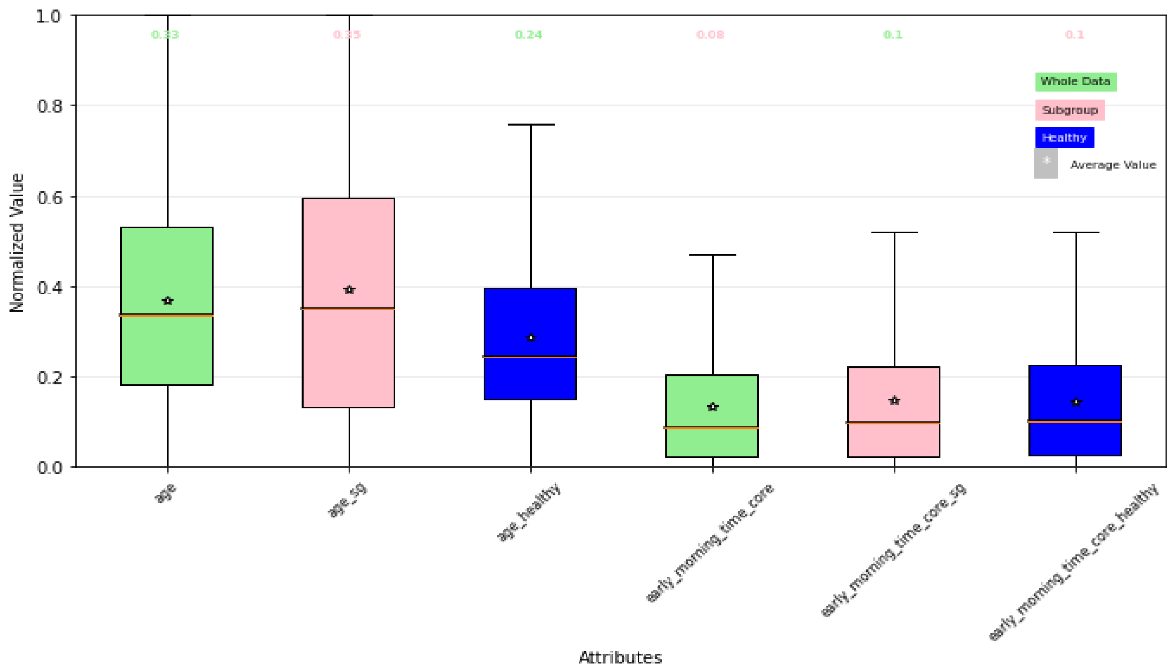


Figure B 7: Distribution Comparison of Subgroup 4 with the Healthy and Whole Population

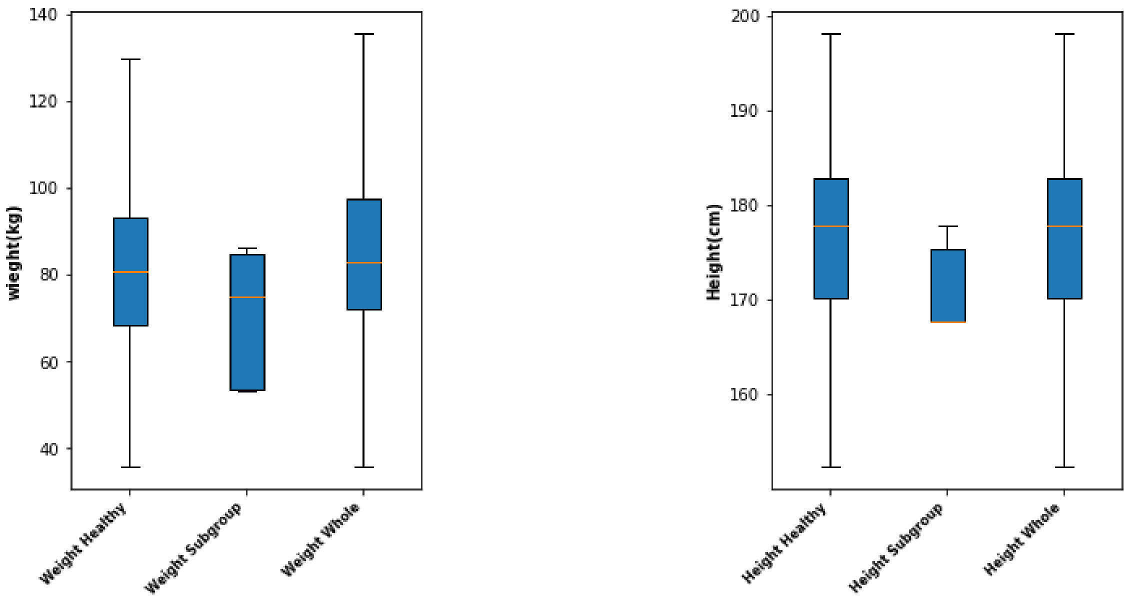


Figure B 10: Height and Weight of Participants in Different Groups of the Data

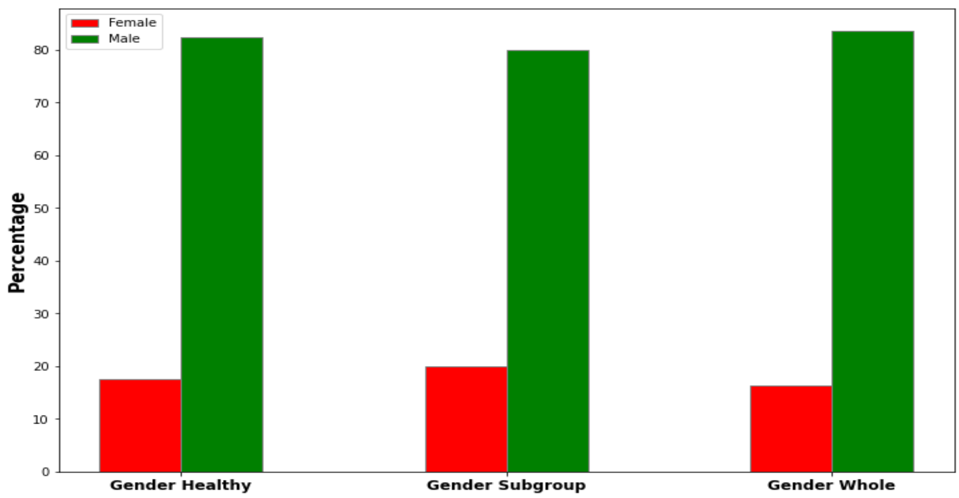


Figure B 8: Gender Distribution in the Healthy, Subgroup 4 and Whole Populations

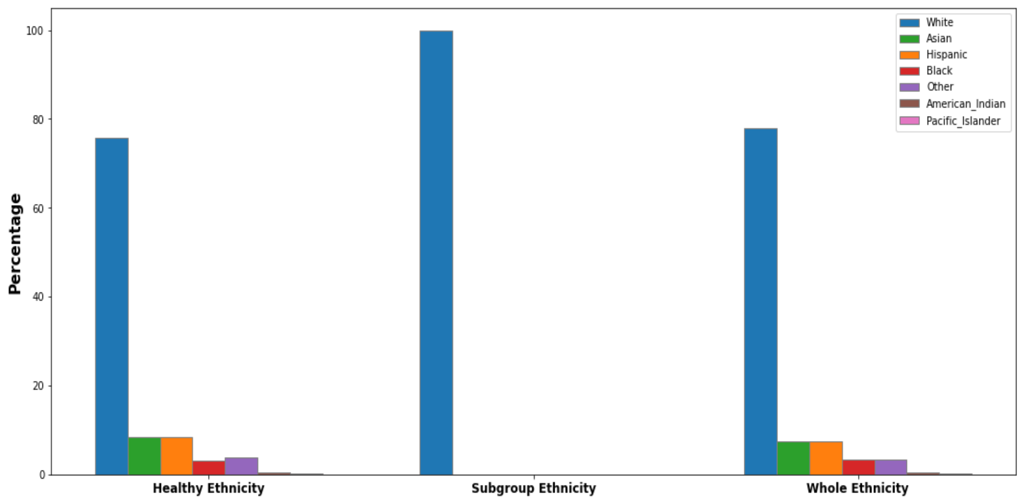


Figure B 9: Ethnicity Distribution in the Healthy, Subgroup 4 and Whole Populations

B.3. Rule5

Rule 5 only includes two conditions. One is related to age another one is related to participants' height. It implies if a participant is older than 48 and younger than 59 and they are taller than 167 cm and shorter than 180 cm it is 77% probable that they have CVD or its risk factors if the three previous rules are not true. This pattern is fined in 40 items if we do not consider previous subgroups and in total in 46 items. Even though the usage of this subgroup is only 40, it is an interesting rule since it might seem surprising to find a relation between the height of the participants and the probability of having CVD. However, there are some studies validating this relation[2], [3].

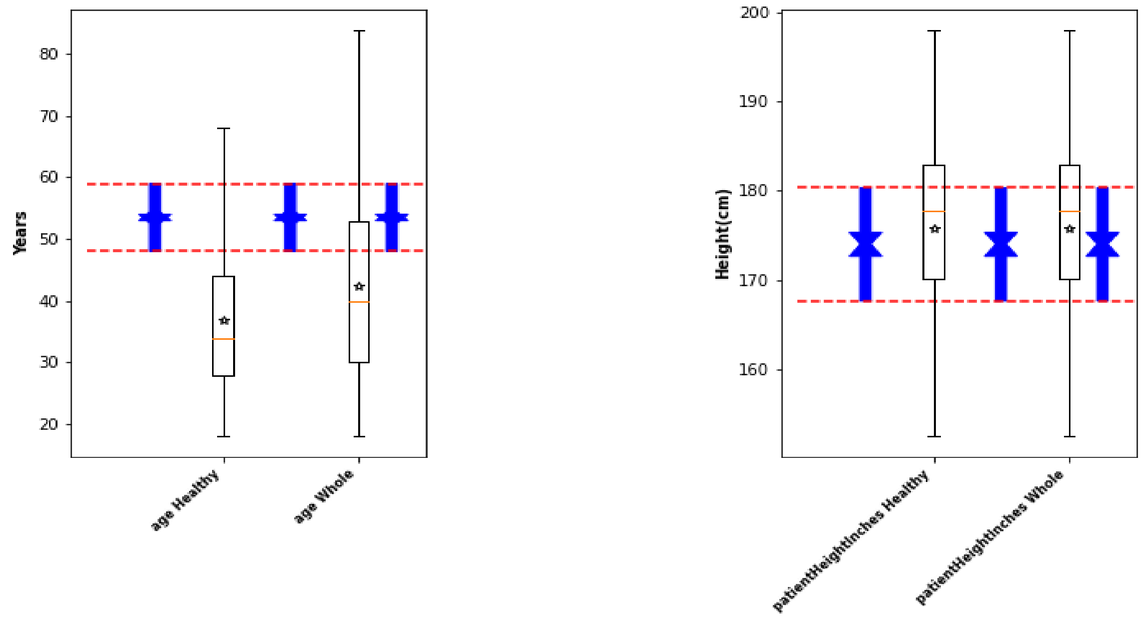


Figure B 11: Rule 5 patterns in Comparison to the Healthy and Whole data distribution

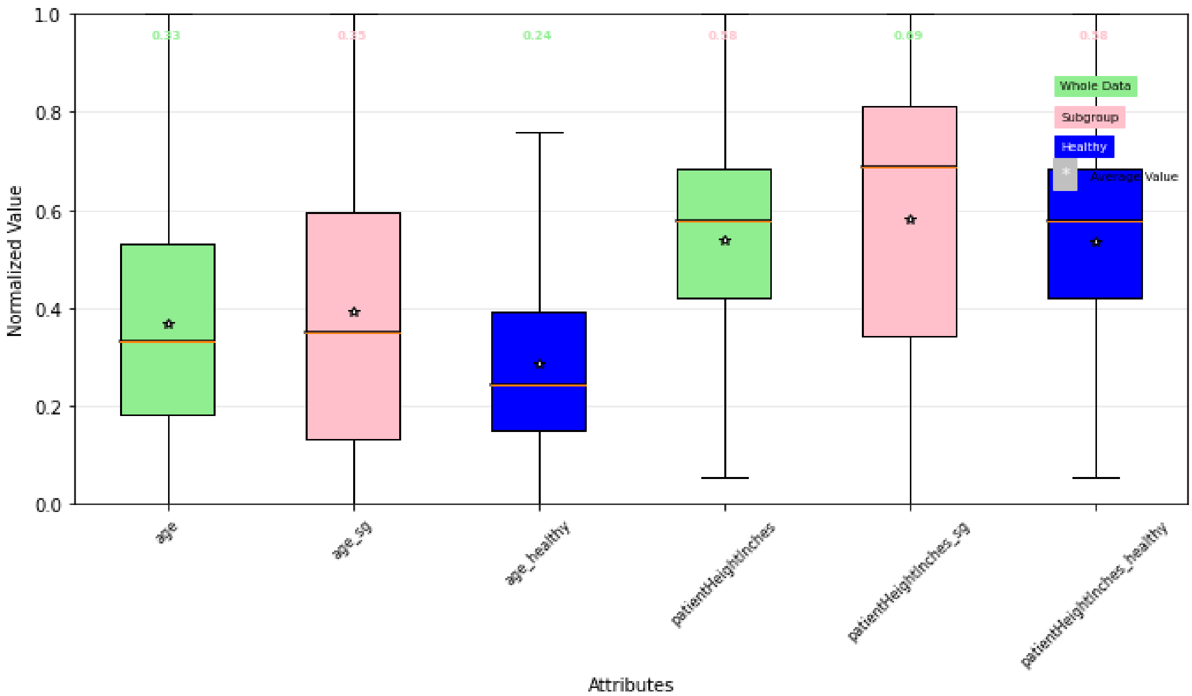


Figure B 12: Distribution Comparison of Subgroup 1 with the Healthy and Whole Population

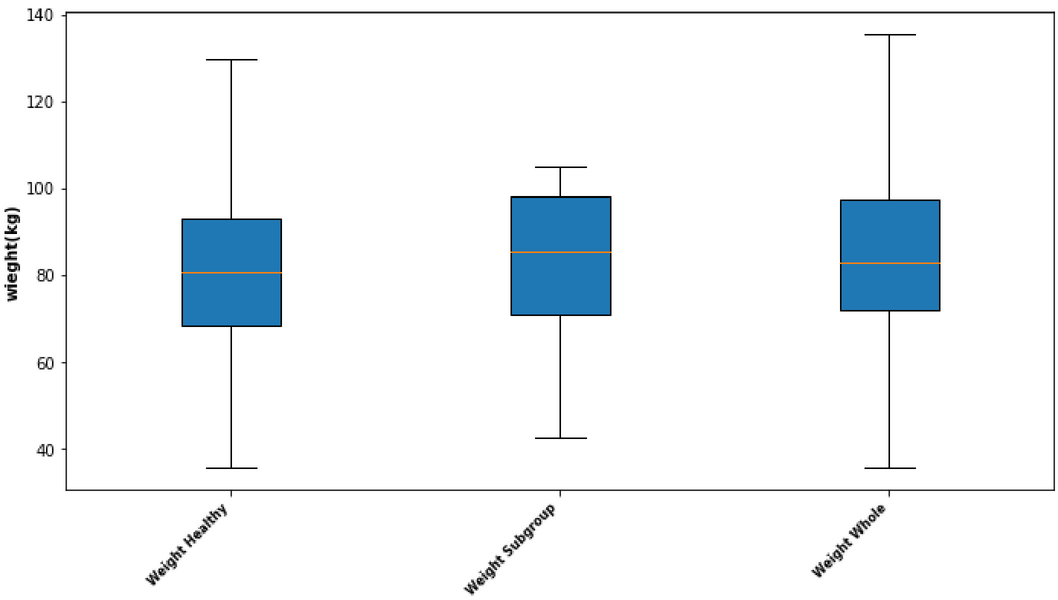


Figure B 13: Height and Weight of Participants in Different Groups of the Data

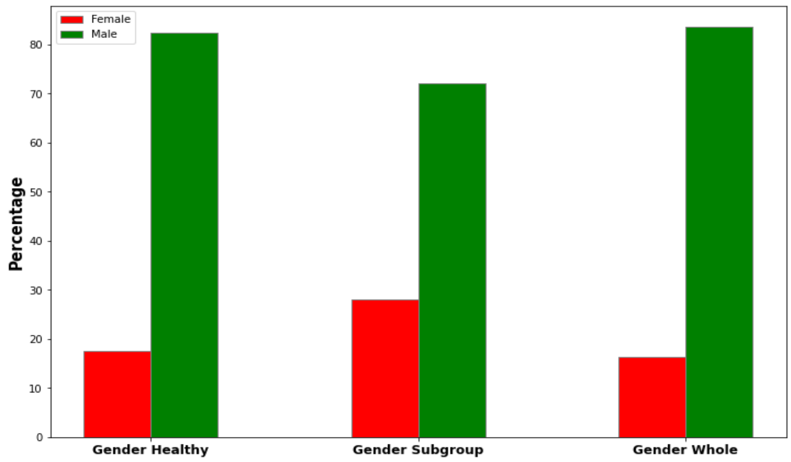


Figure B 14: Gender Distribution in the Healthy, Subgroup1 and Whole Populations

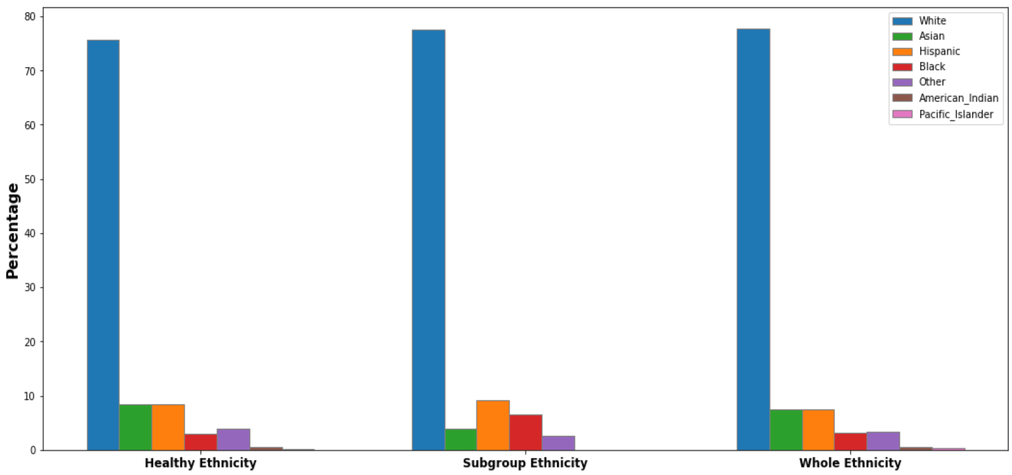


Figure B 15: Ethnicity Distribution in the Healthy, Subgroup 5 and Whole Populations

Regarding demographic attributes, the median weight is greater in subgroup 5 in comparison to the two other data sets. The proportion of female participants is also larger in this subgroup. This is also true regarding the proportion of Hispanic and Black people. However, the percentage of participants of Asian ethnicity is lower in comparison to other datasets.

B.4. Rule6

Rule 6 is related to a circumstance when the participant's age is between 48 and 59 (older than the whole and healthy dataset average and younger than their maximum value), late evening activity duration per day is between 1 hour and 1 hour and 45 minutes (more than lower quartile of whole and healthy datasets and less than their maximum) and the participant entered on average between 0 to 1 activity in daily survey. In this situation, the probability of having CVD or its risk factors is 64%. This pattern is seen in 144 items without considering mutual ones with previous subgroups(**Figure B 16**) The mean and median for the distribution of all these attributes in subgroup 6 are larger than other data groups(**Figure B 18**).

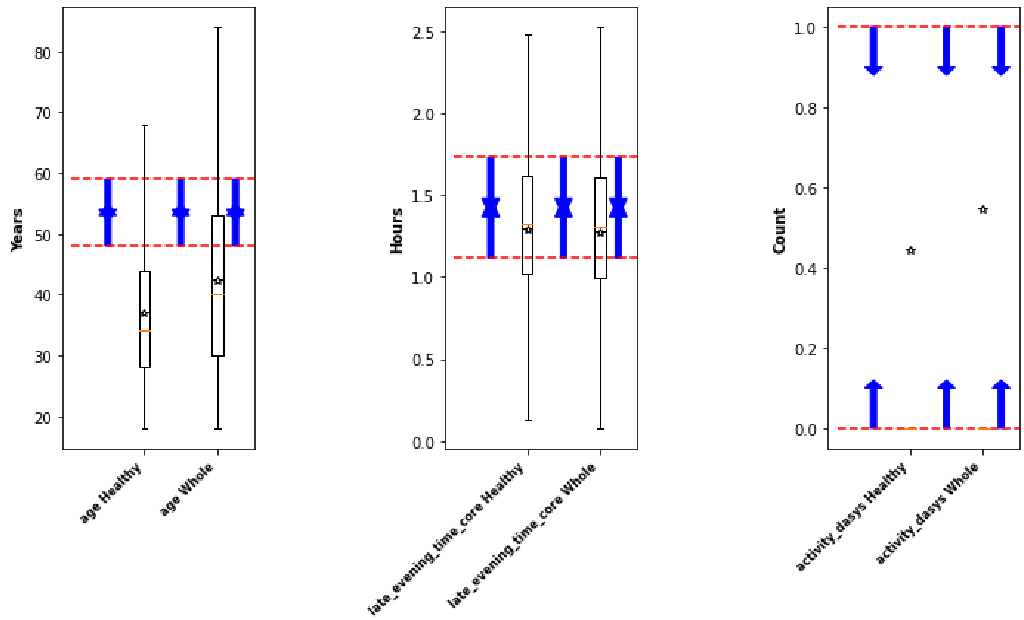


Figure B 16: Rule 6 patterns in Comparison to the Healthy and Whole data distribution

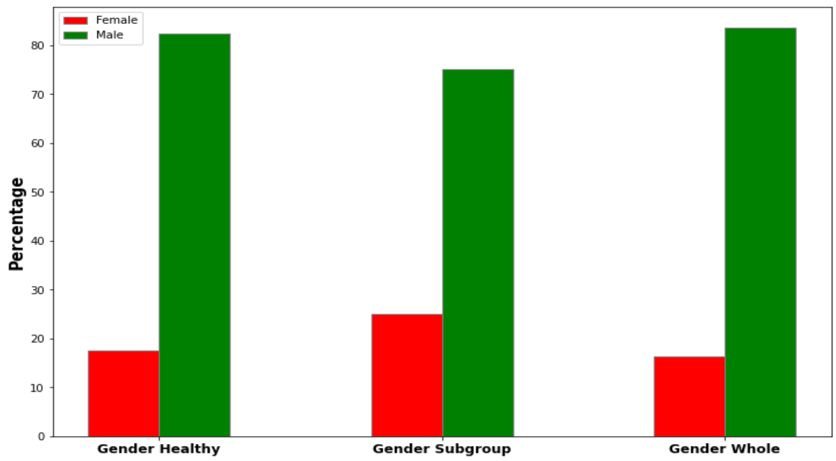


Figure B 17: Gender Distribution in the Healthy, Subgroup 6 and Whole Populations

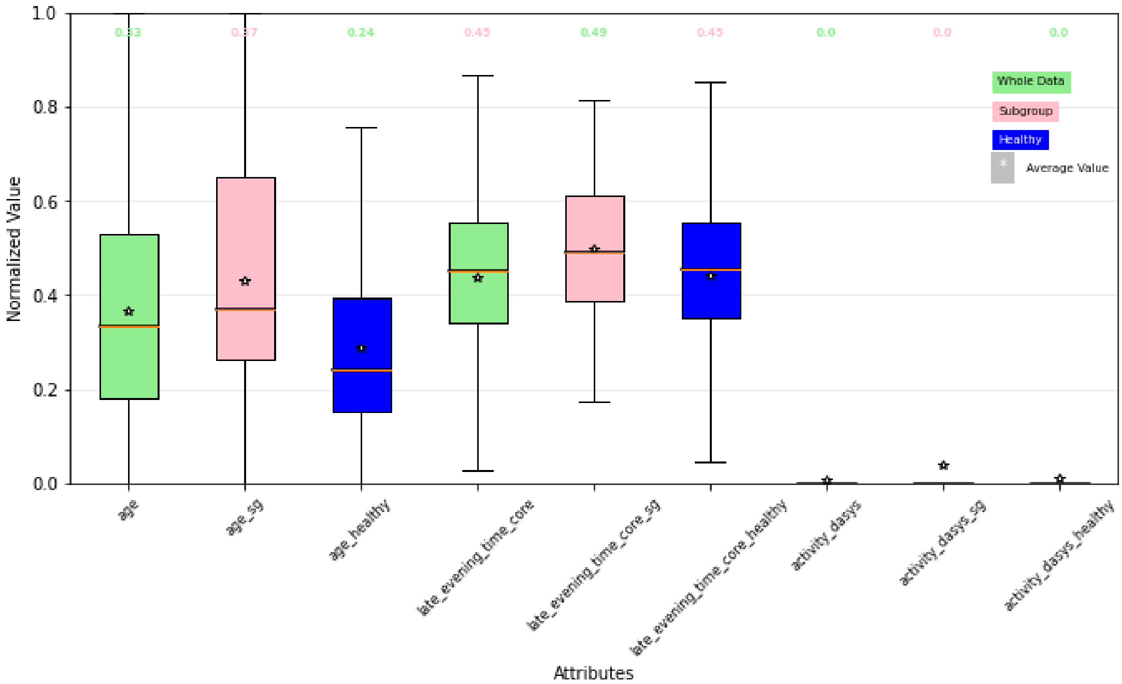


Figure B 18: Distribution Comparison of Subgroup 6 with the Healthy and Whole Population

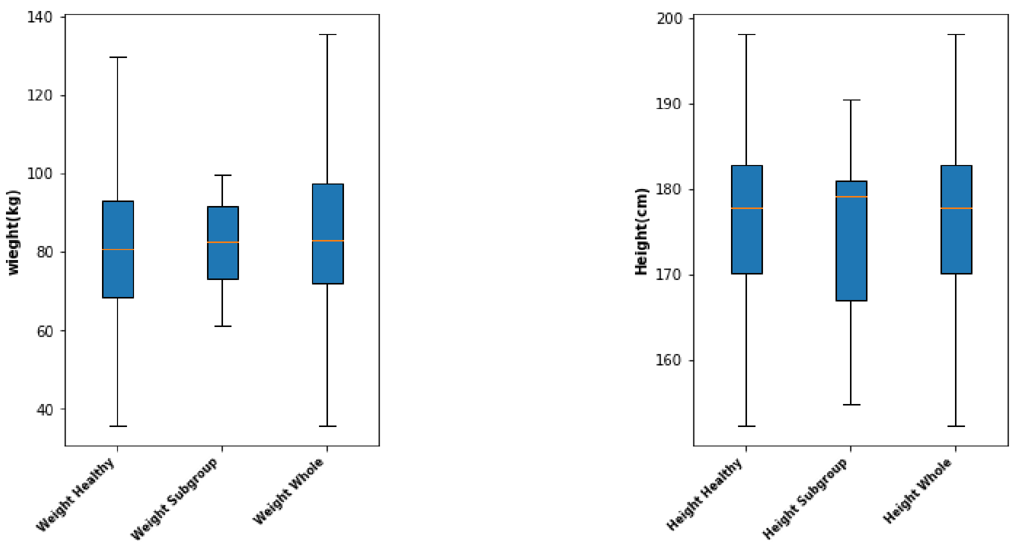


Figure B 19: Height and Weight of Participants in Different Groups of the Data

Participants' weight in subgroup 6 has the same median as the two other datasets; however, it has a denser distribution. The median for the height attribute is larger than the other datasets and the distribution is skewed to the right. Concerning the proportion of female participants(24%) it is almost 5% more than whole and healthy datasets. There are no Hispanic or American Indian people in subgroup 6; however, the percentage of black participant is higher.

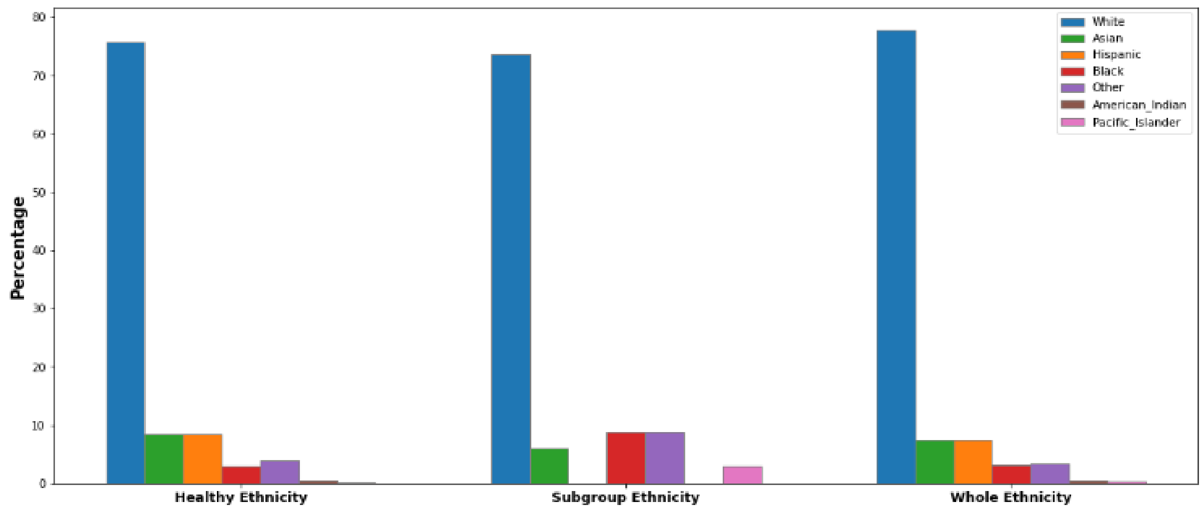
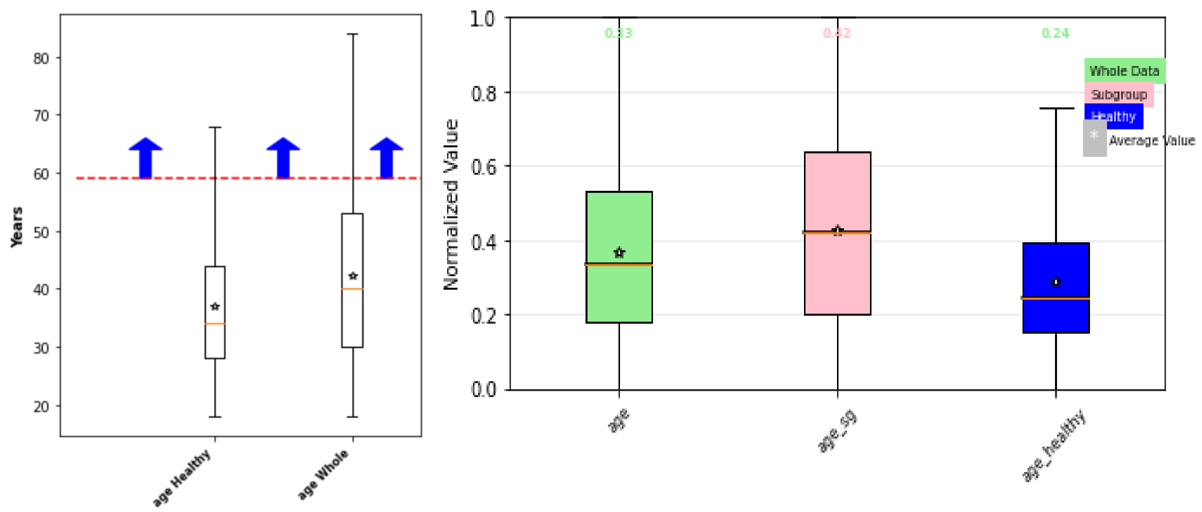


Figure B 20: Ethnicity Distribution in the Healthy, Subgroup 6 and Whole Populations

B.5. Rule7

Rule 7 only has one condition related to age. It indicates if a participant is older than 59, the probability of having CVD or its risk factors will be 56% for them. This is in line with the fact that CVD is more common in people over 50. In addition, as we can see in **Figure B 21**.a, this number is bigger than the upper quartile for both the whole and healthy population.

The distribution of this attribute is also more dispersed in comparison to other groups of data, with a larger average and median score(**Figure B 21**.b).



a. Attribute of Rule7

b. Distribution Comparison for Rule15

Figure B 21

Concerning demographic attributes, the height and weight of subgroup 7 have a less varied interquartile range. Weight attribute also has a larger median but for height, it is almost the opposite. The proportion of men and women in the three datasets are almost the same. There is not any participant from American Indian and Pacific Islander ethnicities. In addition, there are around

10% more Hispanic(15%) in subgroup 7 compared to the healthy and whole population and 5% fewer Asians(**Figure B 24**).

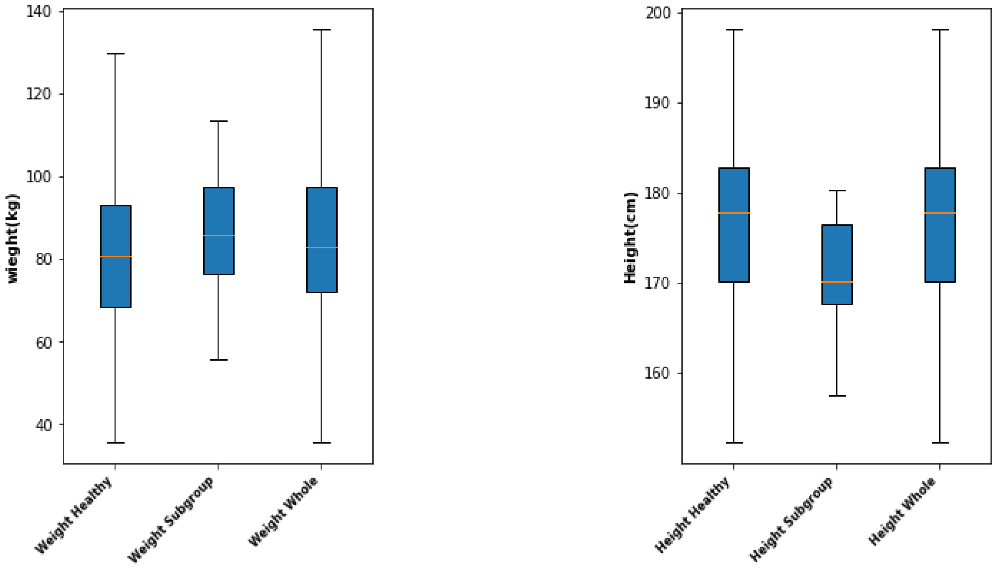


Figure B 23: Height and Weight of Participants in Different Groups of the Data

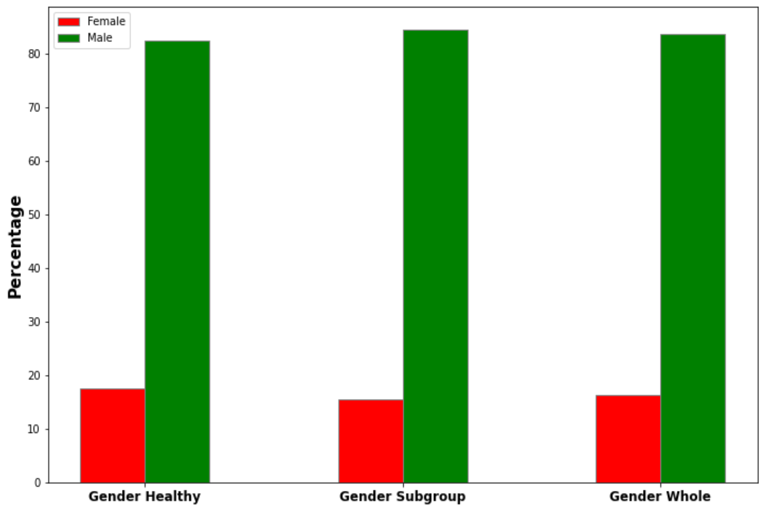


Figure B 22: Gender Distribution in the Healthy, Subgroup 7 and Whole Populations

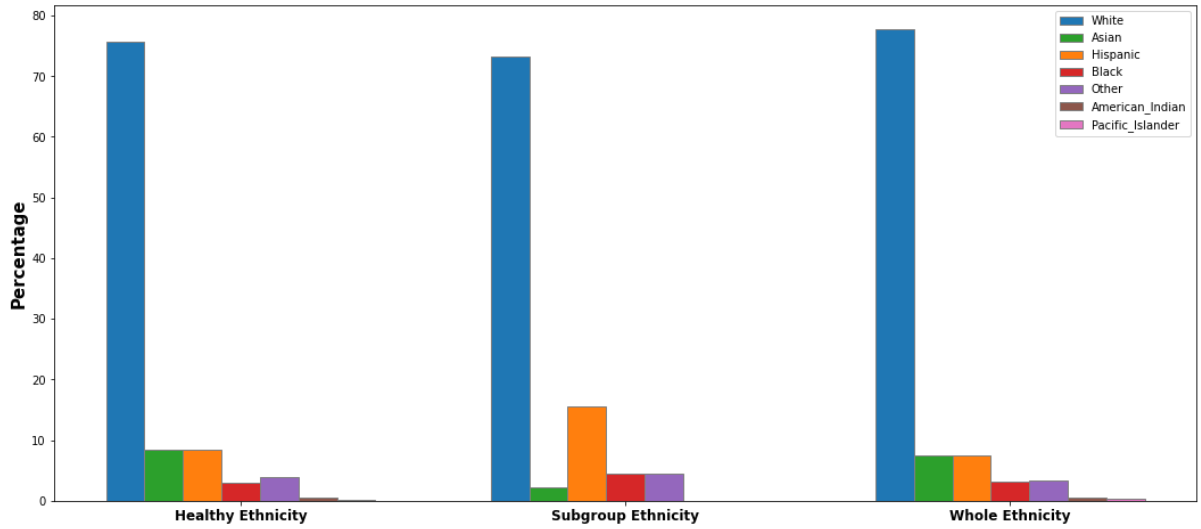


Figure B 24: Ethnicity Distribution in the Healthy, Subgroup 7 and Whole Populations

B.6. Rule8

Rule 8 is about when participant's age is more than 48 years old, and the related weekly vigorous physical activity is less than 30 minutes. In this case, the probability of having CVD or its risk factors is 56%. This pattern is true in 154 items without considering the mutual items with prior subgroups and 666 items totally. 30 minutes weekly vigorous physical activity is less than the vigorous activity that 50% of the healthy and whole population has(**Figure B 25**). In subgroup 8, the median and mean for both attributes is larger than the two other datasets. In addition, the distribution is more dispersed as well (**Figure B 26**).

Regarding demographic attributes, both height and weight distributions are denser in comparison to the whole and healthy dataset. The mean weight is also larger, and the distribution skewed to the right. However, concerning height, the median is smaller. The proportion of women and men participants is almost the same in the three datasets. There is no person from Blac or American Indian in this subgroup. However, the proportion of Pacific Islanders and Alaska Natives is bigger(**Figure B 29**).

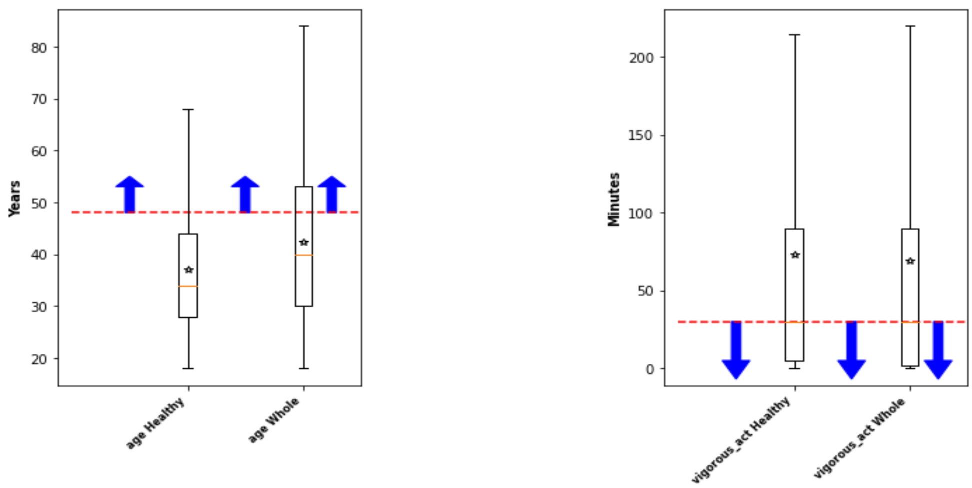


Figure B 25: Rule 8 patterns in Comparison to the Healthy and Whole data distribution

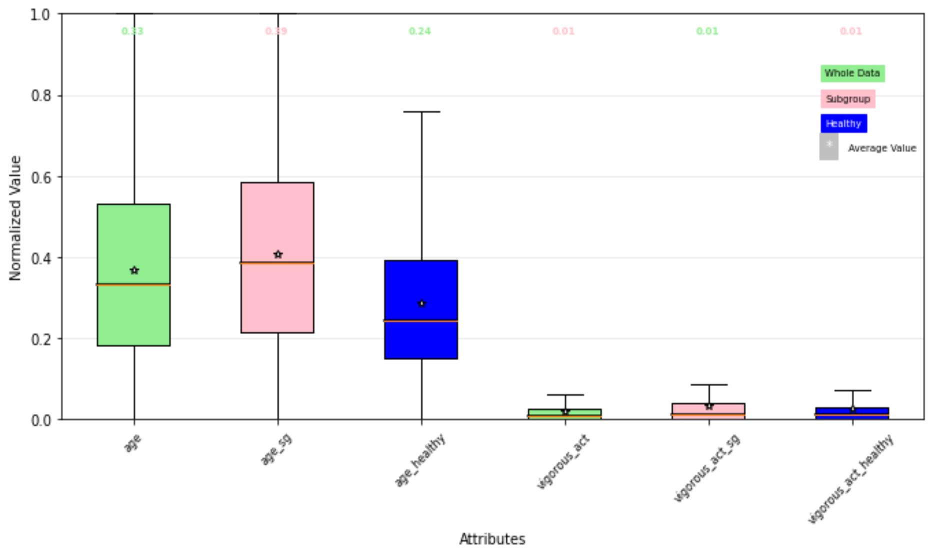


Figure B 26: Distribution Comparison of Subgroup 8 with the Healthy and Whole Population

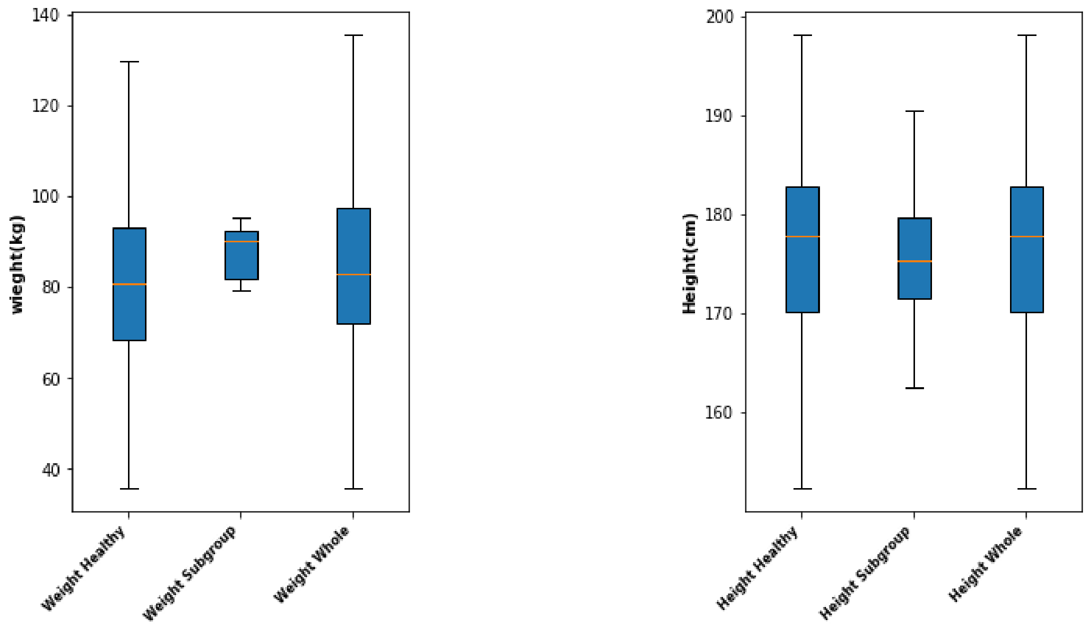


Figure B 27: Height and Weight of Participants in Different Groups of the Data

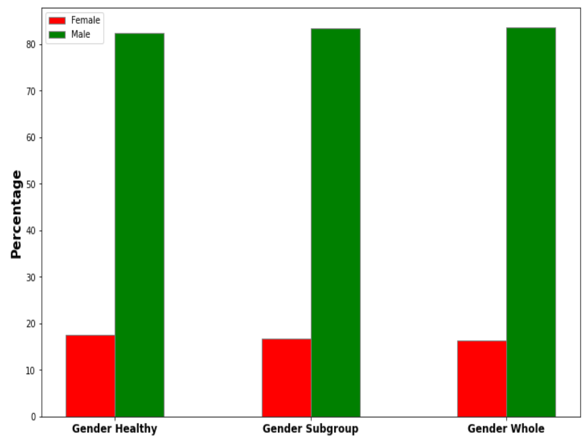


Figure B 28: Gender Distribution in the Healthy, Subgroup 8 and Whole Populations

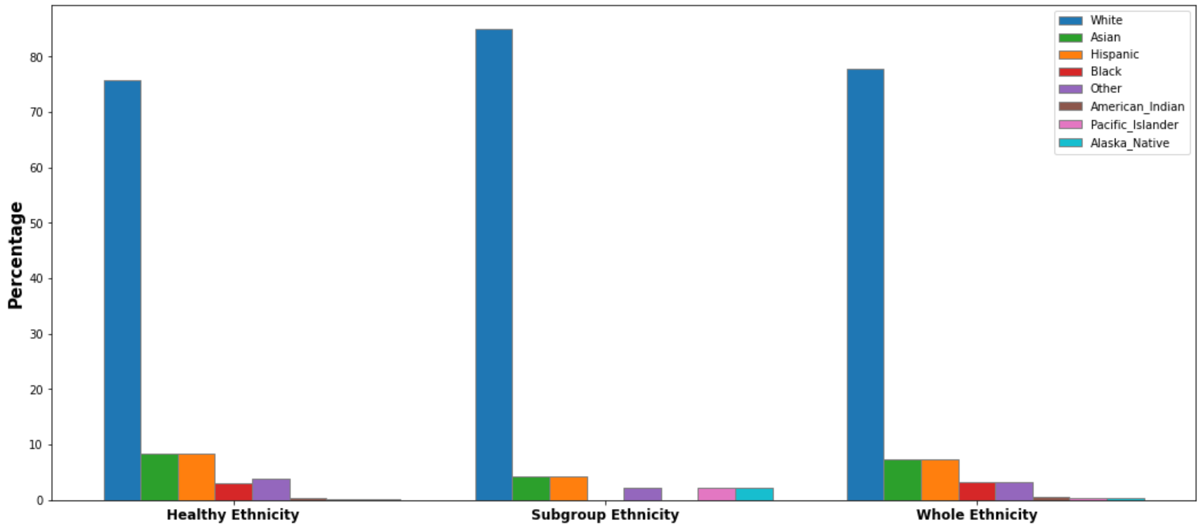
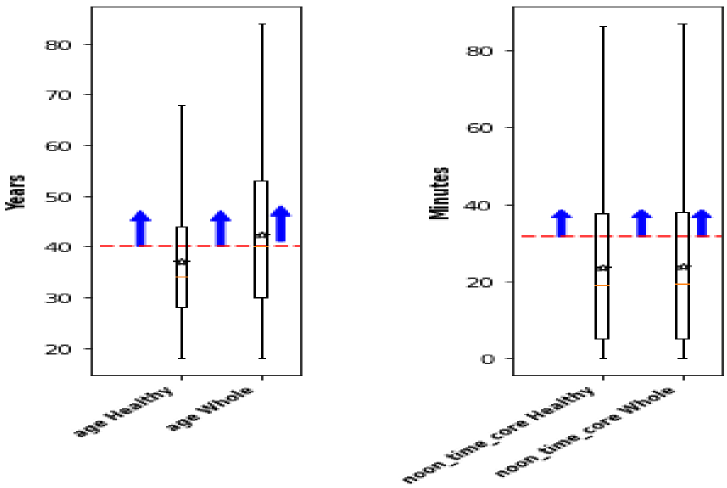


Figure B 29: Ethnicity Distribution in the Healthy, Subgroup 8 and Whole Populations

B.7. Rule9

Rule 9 includes two conditions age>=40 and noon_time_core>=35 minutes. The probability of having CVD or its risk factors under this pattern is 49% which is neutral. The usage of this subgroup is 309, and over all, there are 617 rows that follow this pattern. The age distribution in this subgroup has a higher median score in comparison to the healthy population, and it is more dispersed. For the noon duration of the activity, the median and average are larger that the two other groups, and it is again more dispersed(**Figure B 30**).

There is not much difference between the three dataset groups regarding demographic attributes other than the weight attribute being less scattered in subgroup9 and not having any participants from American Indian ethnicity. In addition, there are around 5% more participants from black ethnicity in this subgroup compared the whole and healthy population.



.

Figure B 30: Rule 9 patterns in Comparison to the Healthy and Whole data distribution

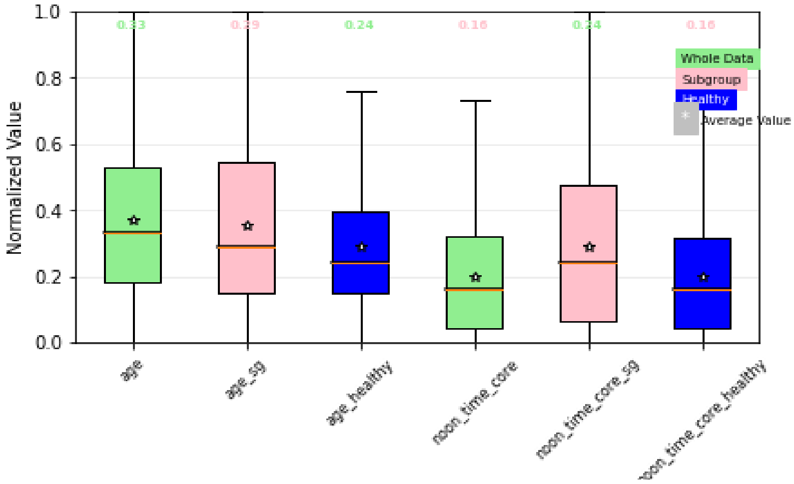


Figure B 31: Distribution Comparison of Subgroup 9 with the Healthy and Whole Population

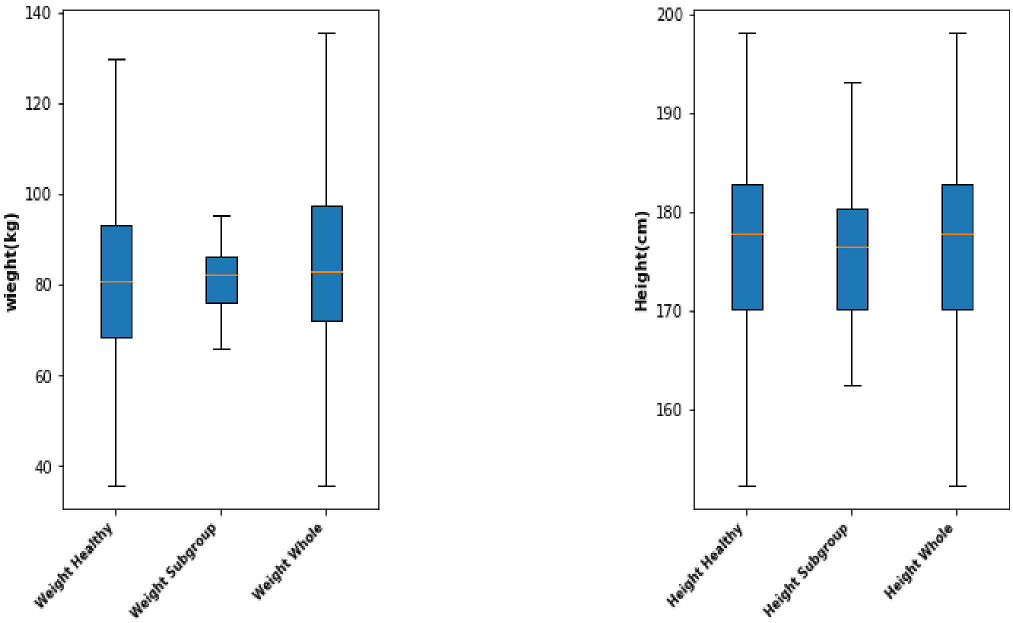


Figure B 32: Height and Weight of Participants in Different Groups of the Data

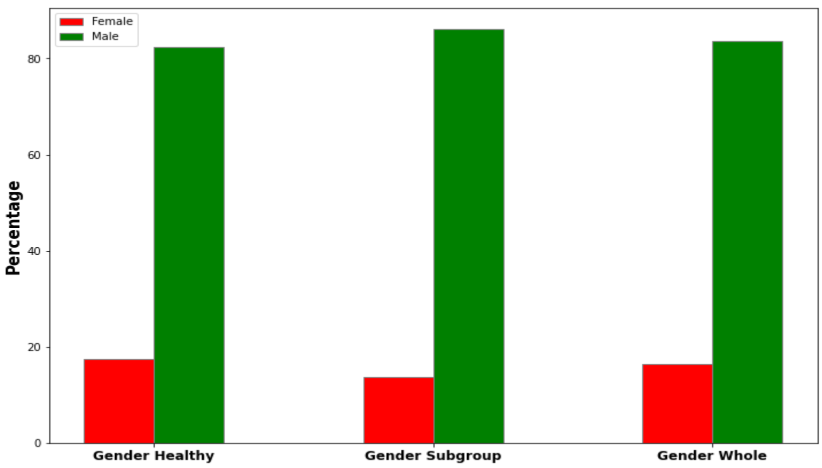


Figure B 33: Gender Distribution in the Healthy, Subgroup 9 and Whole Populations

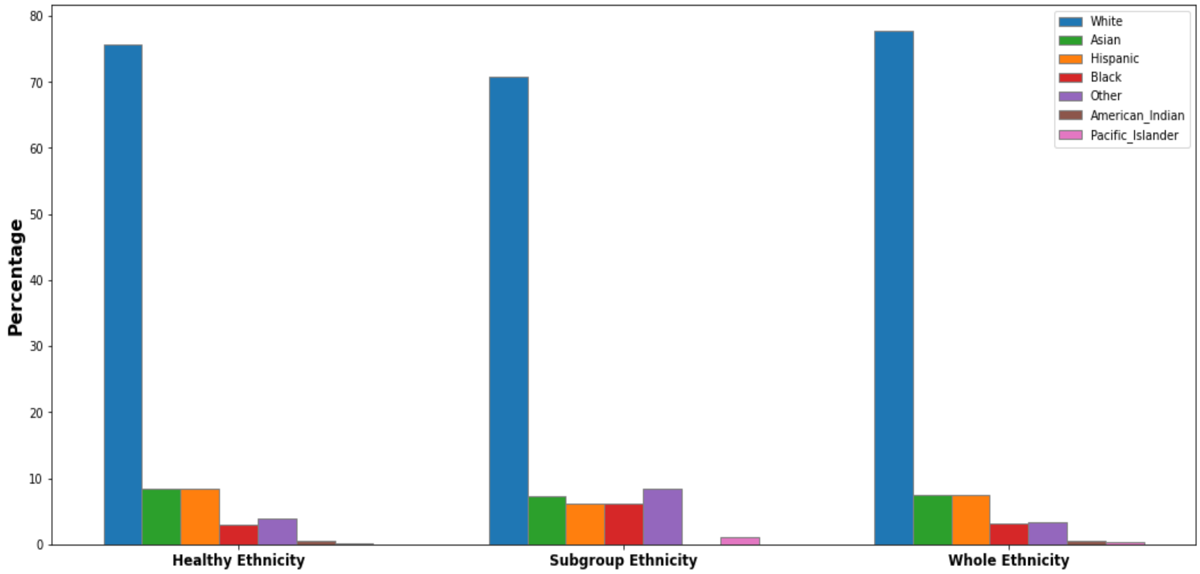


Figure B 34: Ethnicity Distribution in the Healthy, Subgroup 9 and Whole Populations

B.8. Rule12

Rule 12 indicates if a participant does daily vigorous physical activity, and the duration of physical activity during the late evening for them is less than the upper quartile of both the healthy and whole population, then the probability of having CVD or its risk factor for them will be 13%. This pattern is based on 730 items, and 983 items follow this pattern altogether. This is an interesting rule since we usually assume it is better to have more physical activity, no matter which part of the day it is taking place; however, this rule implies that it is better not to have more than a certain amount of physical activity during the late evening. The percentage of participants who has daily vigorous activity in this subgroup(14%) is almost 1% more than the healthy and whole population(**Figure B 37**).

Regarding gender, the percentage of female participants is around 5% more in this subgroup in comparison to other data groups. The ethnicity of the three data groups is not that different other than the fact that there are more Asian and fewer Hispanic participants in subgroup 12 compared to the other two data groups.

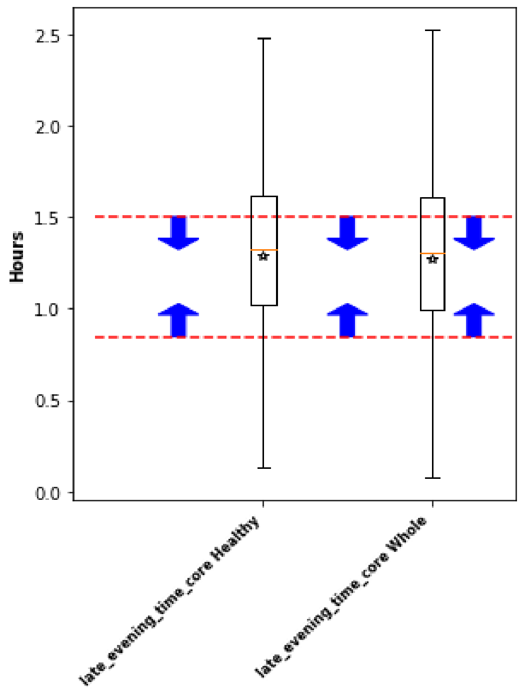


Figure B 35: Rule 12 patterns in Comparison to the Healthy and Whole data distribution

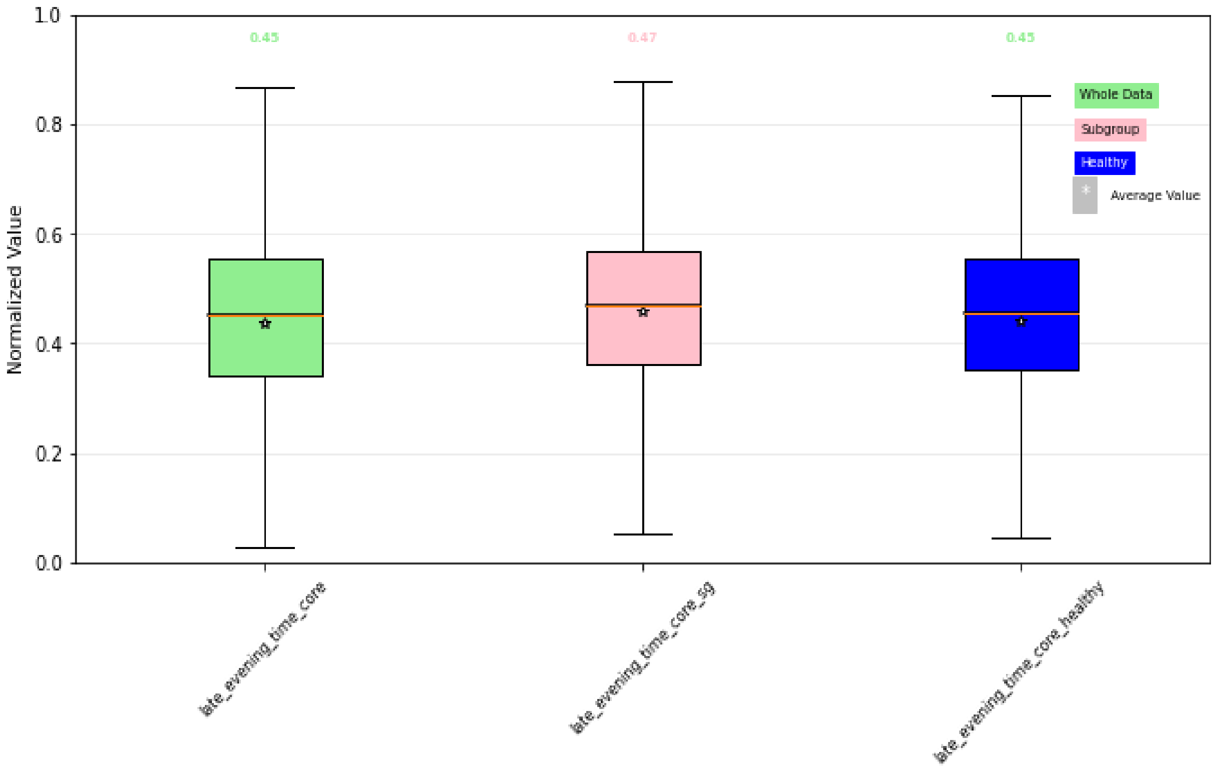


Figure B 36: Distribution Comparison of Subgroup 12 with the Healthy and Whole Population

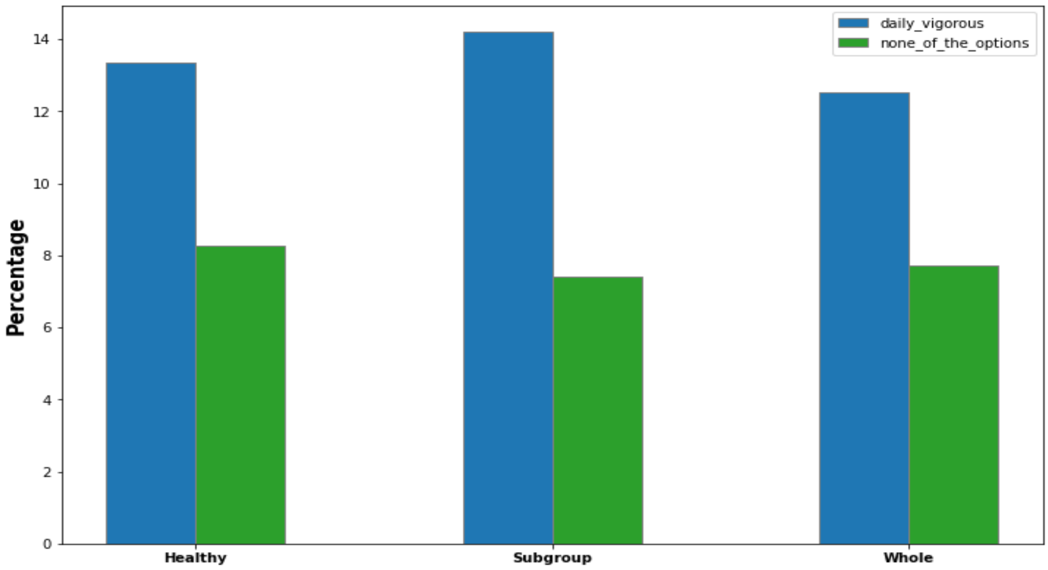


Figure B 37: Categorical Attributes of Subgroup 12 in Comparison to the Healthy and Whole Population

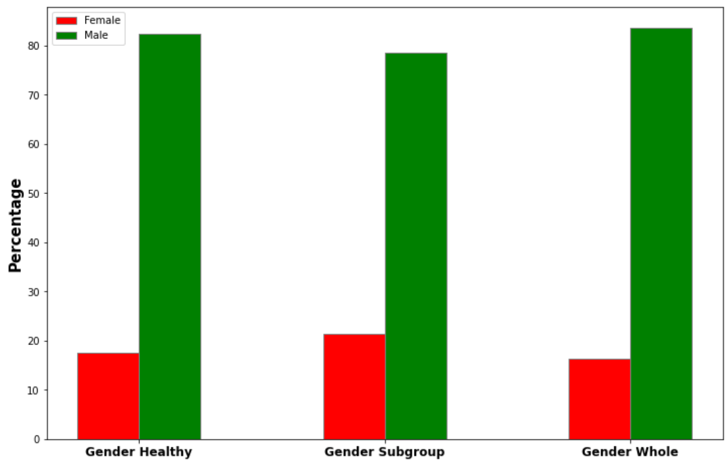


Figure B 38: Gender Distribution in the Healthy, Subgroup 12 and Whole Populations

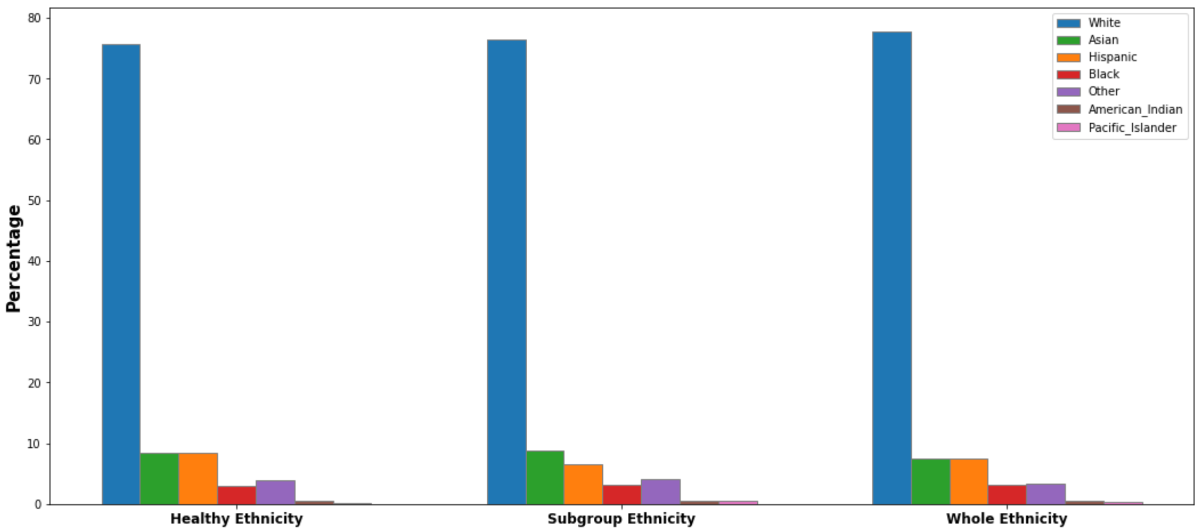


Figure B 39: Ethnicity Distribution in the Healthy, Subgroup 12 and Whole Populations

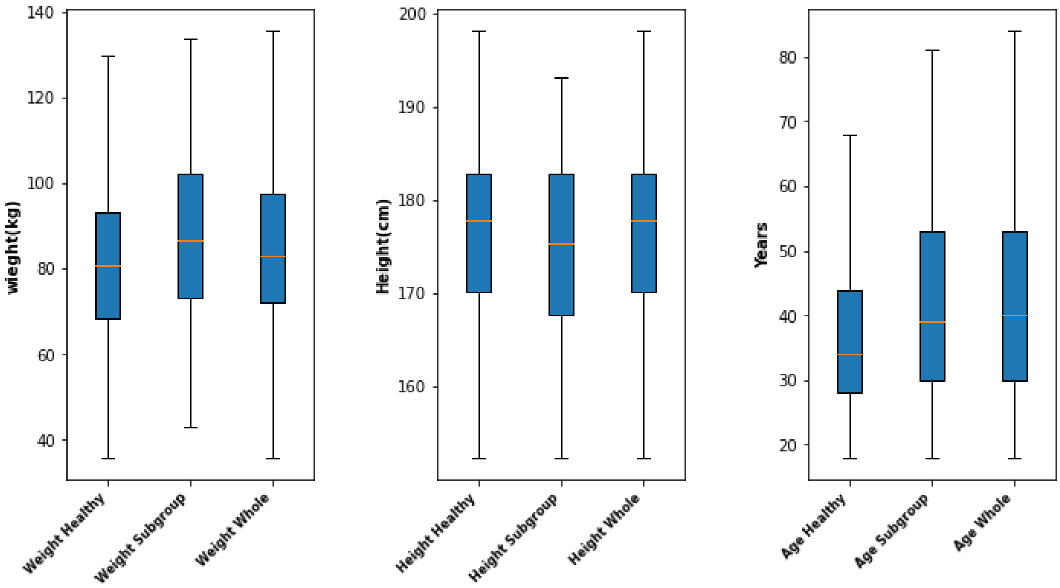
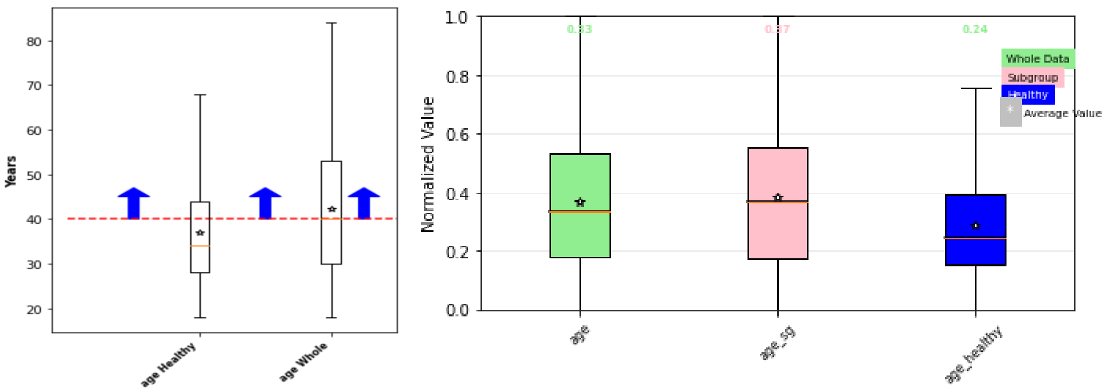


Figure B 40: Height and Weight of Participants in Different Groups of the Data

B.9. Rule13

Rule 13, such as rule 7, only includes one condition on age attribute. It indicates if a participant is older than 40, meaning older than 50% of both the healthy and whole population, there will be a 38% chance of having CVD or its risk factors. This pattern is been seen in 1925 items in the dataset.



a. Attributes of Rule13

b. Distribution Comparison for Rule13

Figure B 41

The weight attribute in subgroup 13 has a larger median in comparison to the two other datasets. However, height distribution is almost the same as a whole and healthy population. This is also true regarding gender attributes. Regarding ethnicity, there is no one with American Indian or Pacific Islander ethnicity in this subgroup. In addition, the proportion of Black and Hispanic ethnicities is also smaller.

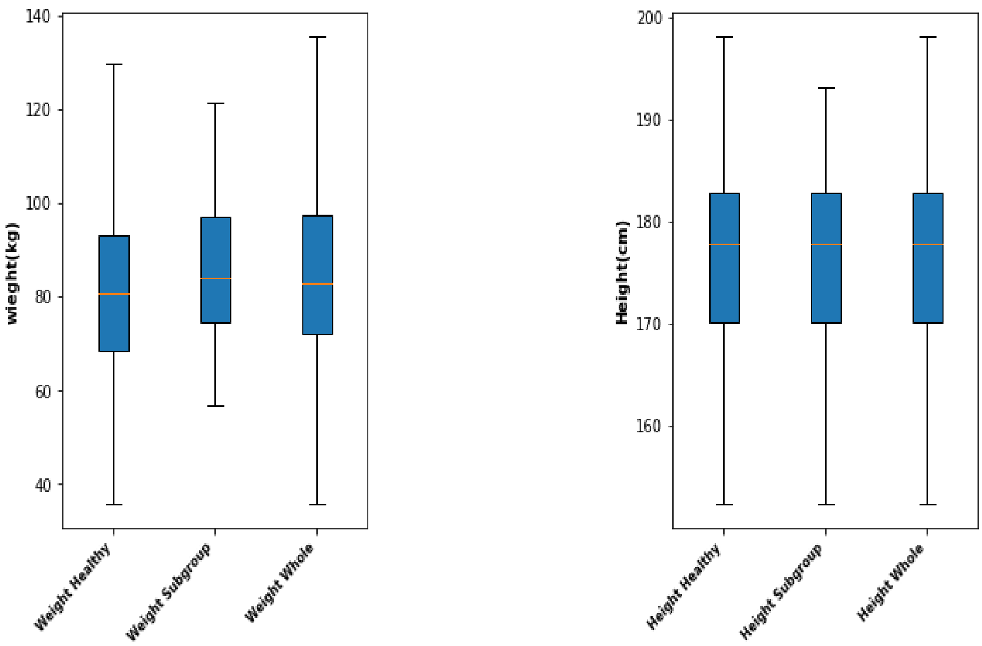


Figure B 42: Height and Weight of Participants in Different Groups of the Data

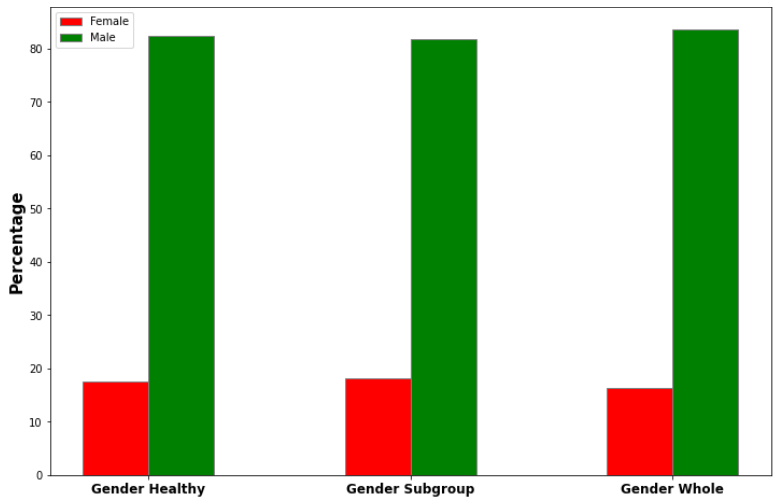


Figure B 43: Gender Distribution in the Healthy, Subgroup 13 and Whole Populations

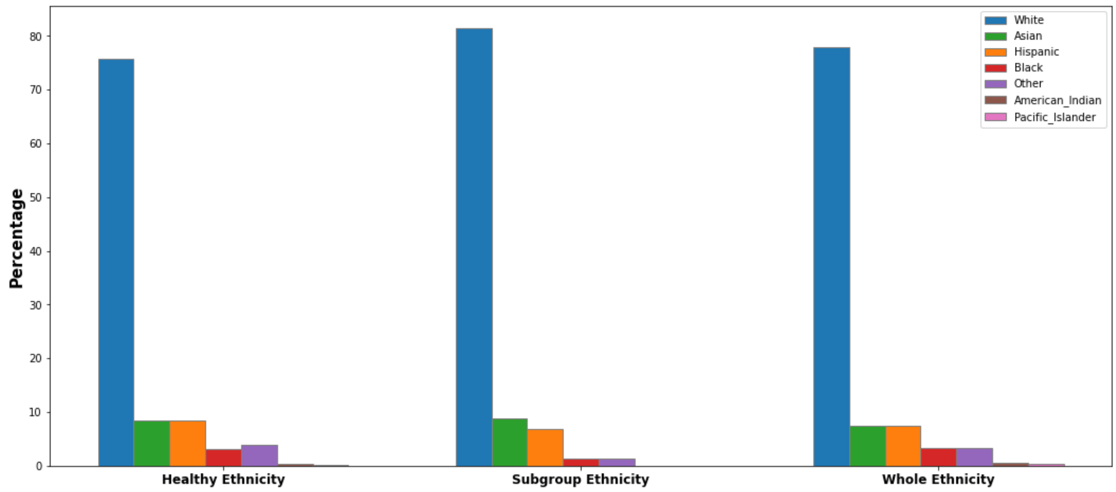


Figure B 44: Ethnicity Distribution in the Healthy, Subgroup 1 and Whole Populations

B.10. Rule14

Rule 14 includes four conditions. The first condition is about having more than one hour and 42 minutes of physical activity during the night(later than 23:59). It means being more active than 50% of the population of both healthy and whole population during the night time. The second condition is related to running more than 1.5 seconds per day on average. This is more than the lower quartile of the whole population and more than the minimum for the healthy population. The third condition, which is interesting, implies having less than nine minutes activity during noon,meaning less than the mean and median of the two other data sets. Lastly, it is related to choosing one option for physical activity at work (**Figure B 46**). Under this circumstance, the probability of having CVD and its risk factors is 11%. The usage for this rule is 934. It is an interesting rule since the first two conditions focus on having more physical activity, but the third one is making an upper bound for it during noon which is unexpected.

Subgroup14 has a larger median and mean in comparison to the other datasets for all these three numeric attributes(**Figure B 47**). Regarding the last condition, percentage of the participants who did not answer the question regarding the amount of physical activity during work is 10% in this subgroup. This is 11% for the whole population.

Concerning demographic attributes, the median for weight in subgroup 14(84 kg) is bigger than the median of healthy and whole population. The median height (177 cm) is equal to the median of other datasets and skewed to the right. However, the distribution is less dispersed. The median for age(41) is bigger than the median of both datasets. The proportion of male and female participants is the same in all three datasets. Regarding ethnicity, proportions of Black and Hispanic people are larger in subgroup14, but it is the opposite for Asian ethnicity.

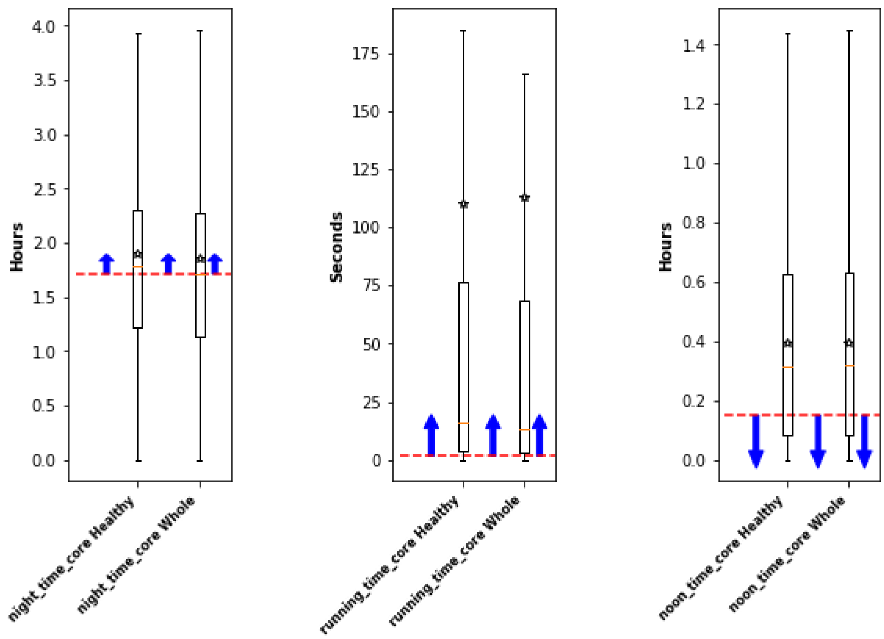


Figure B 45: Rule 1 patterns in Comparison to the Healthy and Whole data distribution

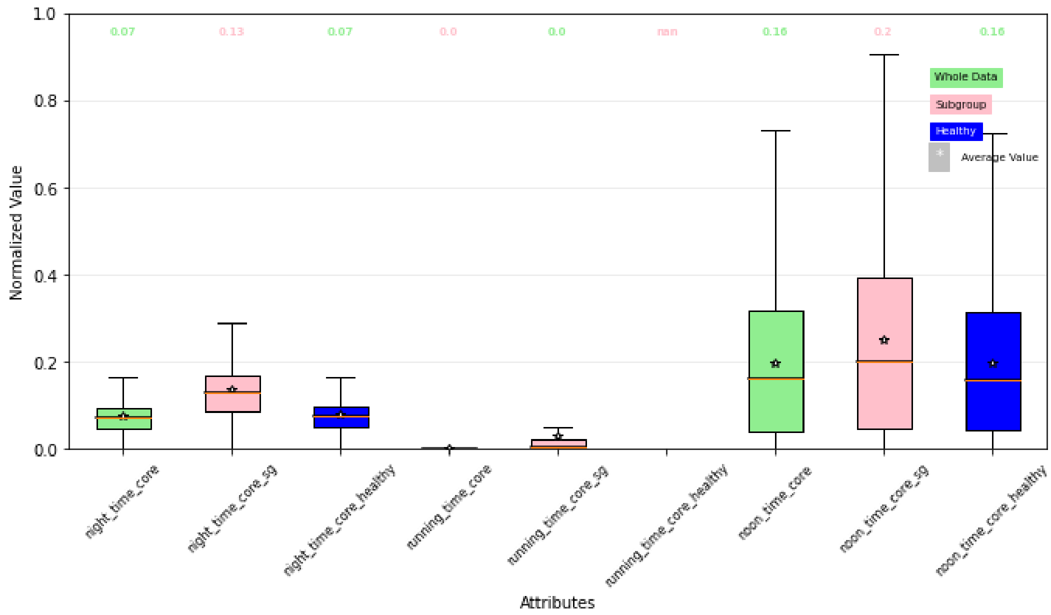


Figure B 46: Distribution Comparison of Subgroup 14 with the Healthy and Whole Population

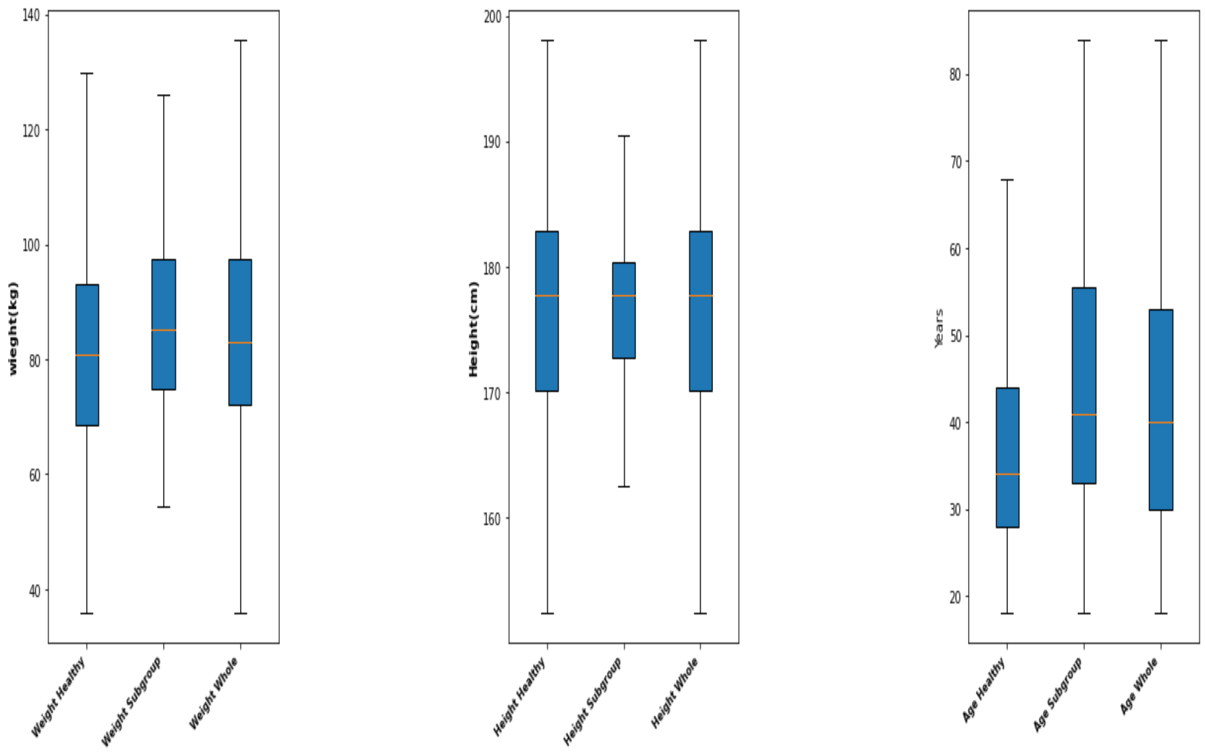


Figure B 47: Height and Weight of Participants in Different Groups of the Data

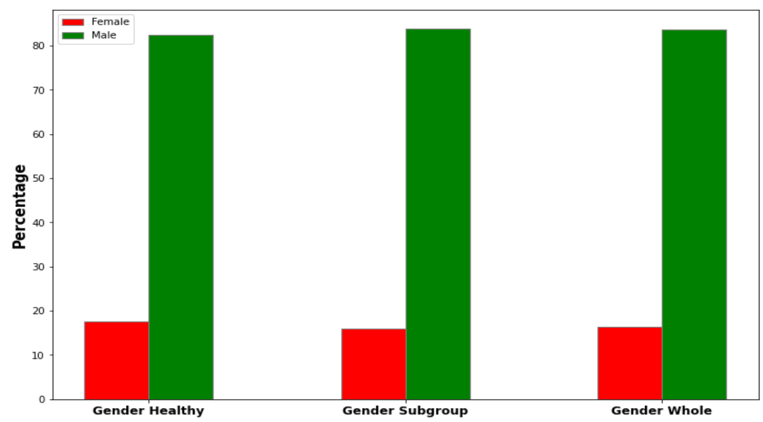


Figure B 48: Gender Distribution in the Healthy, Subgroup 12 and Whole Populations

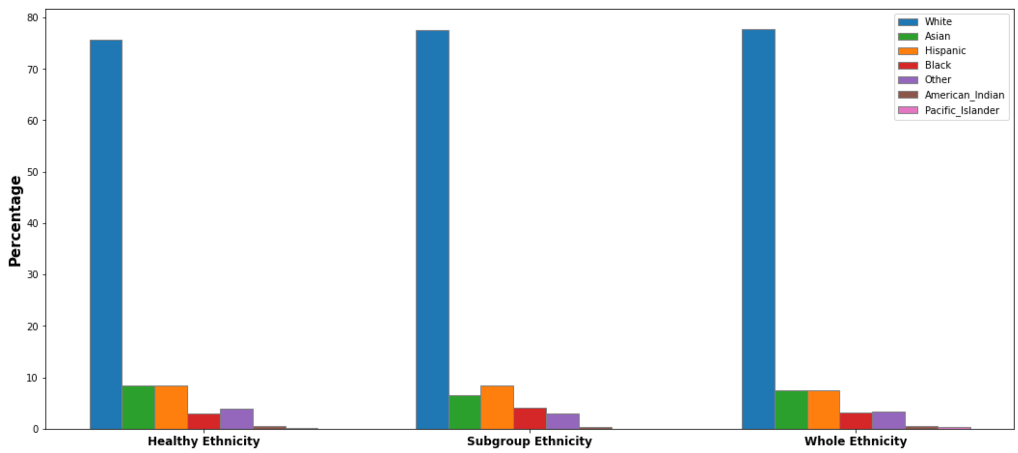


Figure B 49: Ethnicity Distribution in the Healthy, Subgroup 12 and Whole Populations