Universiteit Leiden

# Master Computer Science

Mining Physical Activity Behaviours in Older Adults


| | |
|---|---|
| Name: | Abhishek Akshat |
| Student ID: | S2581418 |
| Email ID: | a.akshat@umail.leidenuniv.nl |
| Date: | 29/08/2022 |
| Specialisation: | CS: Data Science |
| 1st Supervisor: | Arno Knobbe |
| 2nd Supervisor: | Marian Beekman |
| 3rd Supervisor: | Stylianos Paraschiakos |


Master's Thesis in Computer Science


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

**Abstract**

Our project focuses on mining physical activity behaviours by identifying and exploring physical activities (PA) between different individuals. We aim to test and develop unsupervised methods to recognise different physical activity behaviours among older adults. For this research project, we are using the data provided by the Growing Old Together Study, which consists of 164 participants whose daily activities were monitored using wearable sensors over the span of two weeks, one before and one after a lifestyle intervention. We aggregated the obtained data at various scales, i.e. we combined different window lengths and aggregation methods to obtain the best possible combination for our goals. After some initial analysis and pre-processing of the provided data, we selected 114 participants for our experiments. For every participant, we constructed frequency features per 30-second aggregation window. Afterwards, we extracted the features over all the hours and the days. Then, using the constructed feature dataset, we applied two unsupervised methods, namely the $k$-means clustering algorithm and the hierarchical clustering algorithm, to identify differences in activity behaviours for the concluded clusters. Based on these, we expect to identify active or less active behaviours among the participants. We also consider ways to optimise the feature dataset by reducing the dimensions of our data for our clustering tasks, such as principal component analysis. This research would help us to stimulate vital and healthy ageing among older people along with inspiration and motivation to get started and stay active for a longer, healthier and happier life.

# Contents

# Chapter 1

# Introduction

Physical activity refers to any bodily movement produced by skeletal muscles that require energy expenditure. Physical activity holds all types of activities, i.e. activities performed throughout the day or night with any level of intensity. It includes activities carried out in our daily life routine (e.g., walking, running, cycling, etc.) along with the movements performed during leisure, excluding sitting still or lying down. The different types, duration, frequency and intensity of activities performed over a day constitute the physical activity behaviour of a person. Active behaviour over the span of a day is vital for healthy ageing [7, 2]. Being physically active can help to reduce health risks, improve brain function, help manage weight, etc. Proper monitoring of physical activities can help identify high-risk subgroups from a sample population and promote healthy ageing.

Technological advances such as artificial intelligence play an important role in healthcare. Using a wide range of data collected from healthcare communities such as hospitals, pathology centres, medical universities, et cetera; machine learning which is a division of artificial intelligence has helped us tackle many health-related problems such as predicting illnesses and treatments [4, 34, 9], aiding in drug development [10], neurological disease detections [20], promoting healthy ageing [7, 2], forecasting health risks of various communities [40] and helping pathologists to develop more accurate diagnosis of numerous diseases [15]. With the recent development of coronavirus that caused a worldwide pandemic, people started paying more attention to their health and well-being. Machine learning and artificial intelligence also played a crucial role in helping us save lives by identifying patients at high risk of critical illness [25] and predicting its severity in specific populations [31].

Wearable sensor technologies have recently gained much popularity among many people. It has helped us to enhance the quality of life with people tracking and maintaining their daily body activities and energy consumption. With the help of various machine learning algorithms and data collected from these sensors, we can stimulate health benefits.

In this research project, we use the data provided by the GOTO study [33] which consists of

165 participants whose daily activities were measured using wearable sensors over the span of two weeks, one before and one after a lifestyle intervention. With the information obtained, we can identify specific behaviours of the participants from the baseline, i.e. data collected before the lifestyle intervention and compare them with the changes per group with the data obtained after the intervention. This can help us monitor health changes per group in the future and get some more insight into a healthy way of living.

We employed two unsupervised machine learning algorithms, namely $k$-means [26, 8], and Hierarchical Clustering [18, 42] for the features that we constructed based on the average of activities per hour and period of hours in order to identify different physical activity patterns. To find the optimal number of clusters, we used the Elbow method [5, 27, 47] and Silhouette score [35, 39] as evaluation metrics for $k$-means clustering algorithm, whereas for Hierarchical clustering algorithm, we used a tree diagram known as dendrogram [44, 37]. This research would help us to stimulate vital and healthy ageing among older people along with inspiration and motivation to get started and stay active for a longer, healthier and happier life.
Our work focuses on activity pattern mining by identifying and exploring physical activities between different individuals.

**Our work aims to investigate the following:**

1. Test unsupervised methods in order to recognise different physical activity behaviours among participants.
2. Identify active, non-active and irregular profiles using wearable sensors from our constructed feature dataset.
    - **Research Question 2.1:** Can wearable sensor data be relied upon while determining profiles of participants based on their activity behaviour?
3. Deal with the high dimensionality and high correlations of our generated feature-dataset, i.e. optimising the feature-dataset for the clustering task.
    - **Research Question 3.1:** Can dimensionality reduction techniques be precisely applied on the wearable sensor data?

The structure of this report is as follows: In Chapter 2, we outline the related work associated with our research. In Chapter 3, we discuss the data we have used. In Chapter 4, we present our approaches and methods used. In Chapter 5, we discuss the results, and finally, Chapter 6 presents some concluding remarks and the future scope of our work.

# Chapter 2

# Related Work

Our work relates to subject matters such as activity behaviour mining, physical activity profiling, clustering time-series data, etc. The following sections discuss the related work that motivated and influenced our work.

## 2.1 Physical Activity Profiling

Wearable sensors have seen a considerable increase in popularity and interest lately in people's daily life. These sensors are electronic devices that can measure physical movements, body temperature, blood pressure, breathing rate, blood oxygen, and electrical signals of the heart, brain and muscles. With the large amount of data collected by these sensors, several types of research have been conducted that helped us to promote and stimulate various health benefits.

Recent developments and researches show a significant impact on health consequences associated with physical activity and sleep behaviours. A study by Willetts et al. [46] uses random forests with Hidden Markov Models to detect and classify physical activity and sleep behaviours using data collected from wrist-worn activity sensors. The trained model was later used to find different activity patterns of the given population. A similar study [28] uses raw activity data collected from wrist actigraphy to distinguish behavioural activity rhythms (BAR) in older adults. Older adults must have a balanced level of physical activity and sedentariness for healthy ageing. Trumpf et al. [45] also used data from multiple wearable sensors (i.e. a wrist-worn actigraph and a hybrid motion sensor attached to the lower back) to quantify the activities in older adults accurately.

An accelerometer is a type of sensor that captures the intensity of physical activity. In [23], the authors propose a data-driven approach using a Hidden Semi-Markov Model (HSMM) for clustering the accelerometer data collected from 500 participants to categorise activity intensity as sedentary, light, moderate and vigorous. The limitation of such a data-driven approach is that the trained classifier may overfit the experimental setup under which it was trained. In [21] by Jones et al., an essential advantage of using the k-means model for accelerometer data is

discussed. The main idea is to store the centroid of each cluster and apply it again to multiple accelerometer datasets, enabling similar clusters to be fitted from multiple datasets, making it possible to compare different populations. Therefore, we decided to use the k-means clustering algorithm as one of the unsupervised machine learning method choices.

## 2.2 Time-series Clustering

Time series data is a collection of specific observations at constant intervals. In [1], the authors provided a detailed literature review of multiple approaches for clustering time-series data. Time series accelerometer-based data is extensively used in the fields of machine learning. Accelerometer-based on-body sensors have been widely used for monitoring health and various medical applications. In the paper by Amini et al. [3], the authors used both supervised and unsupervised time series analysis methods to capture the motion data using accelerometers. Accelerometer-based data is also being used in the applications of gesture recognition. In [19] by Jang et al., the authors used a Dynamic time warping (DTW) based K-means clustering algorithm for recognising handwriting.

Dynamic time warping (DTW) is a similarity measure which is considered useful when we have time-series data of different lengths. It finds the optimal non-linear alignment between different time-series data with distortion in the time axis. A majority of research have used Euclidean distance metric for clustering of time series [24, 11, 14]. The Euclidean distance is more efficient concerning time and space efficiency than the DTW similarity measure. Therefore, considering the efficiency of our experiments, we decided to use Euclidean distance as our similarity measure for clustering. We were also interested in detecting similar patterns in our data without considering when the pattern occurred. Therefore, we can use the Euclidean distance measure for k-means clustering since we are more interested in finding the active and non-active groups rather than the time of occurrence of those physical activities.

In the paper by Jothi [22], the author presents a comparative study of three clustering algorithms, namely K-means, Hierarchical Agglomerative Clustering and Fuzzy C-means for clustering time-series data generated by smart devices for Human Activity Recognition (HAR). Providing a sufficient amount of labelled data representing free-living conditions is sometimes quite challenging. In [13] by Domingo et al., the authors propose motif discovery as an unsupervised activity recognition approach to tackle these limitations and detect habitual activity from accelerometer data. This influenced us to choose Hierarchical Agglomerative Clustering as our subsequent unsupervised machine learning approach for finding clusters.

# Chapter 3

# Data

For this research project, we used the data provided by the Growing Old Together Study (GOTO) [33], which consists of 164 participants whose daily activities were measured using wearable sensors over the span of two weeks, one before and one after a lifestyle intervention. The aim of the GOTO study is to assess the effect of lifestyle intervention on the older adults. The intervention is a reduced energy balance by 25% for a 13-week duration, which includes a targeted 12.5% reduction in caloric intake and a 12.5% increase in physical activity. We are dealing with datasets in a sequence of data points indexed in time order i.e. time-series data. In the following sections, a detailed description of the data, the pre-processing steps and analysis is provided.

## 3.1 Data Description

The table 3.1 represents the personal attributes of the participants in the GOTO study. A total of 164 older adults participated in this study out of which 83 participants are male and 81 are female, with an average age of 64 for male participants and 62 for female participants, and an average Body Mass Index (BMI) of approximately 26 for both male and female individuals.

| | **Male** | **Female** |
|---|---|---|
| **Number** | 83 | 81 |
| **Age** (mean $\pm$ std) | 63.84 $\pm$ 5.41 | 62.05 $\pm$ 5.88 |
| **BMI** (mean $\pm$ std) | 26.18 $\pm$ 2.19 | 26.60 $\pm$ 2.84 |

Table 3.1: Personal Attributes of the Participants in GOTO Study

As mentioned, the GOTO dataset consists of two large chunks of data, the wearables data collected before and after the intervention (referred to as $\textbf{Data}_{pre}$ and $\textbf{Data}_{post}$). In each chunk of data, we have accelerometer data for about a week (the duration varies per participant, ranging from 5 to 7 days), for a total of 164 participants. The raw data is recorded at roughly 1 Hz, but we will be looking at aggregation levels of once per second to once every 30 seconds. By means of an Activity Recognition model [29] developed in the GOTOv study (specifically

meant to provide such a model for further interpreting the GOTO data), the raw data was interpreted and classified according to 15 distinct activities, at a granularity of one label per second (1 Hz). These round-the-clock activity labels will be the core of our analysis in this thesis, since they provide a detailed picture of a participant's daily routines and level of activity. The original activity classes were specifically designed to be quite detailed, in fact more detailed than we require for our purpose of activity profiling. The following activity classes were originally distinguished:

- lyingDownLeft
- lyingDownRight
- sittingSofa
- sittingChair
- walkingSlow
- walkingNormal
- walkingFast
- walkingStairsUp
- step
- standing
- stakingShelves
- vacuumCleaning
- dishwashing
- sittingCouch
- cycling

Along with the original design came a taxonomy that organized the 15 activities into various different levels of granularity, such that the activities can be distinguished at different levels of detail. For example, the low-level activities of lyingDownLeft and LyingDownRight (which from a sensor point of view look quite different) can be aggregated to a more general class of lyingDown, and so on.

## 3.2 Data Pre-processing

We are more interested in general classes of activity than the level of detail provided by the 15 classes. Therefore, we decided to move to an intermediate aggregation level in the taxonomy that provides 6 distinct activities: **lyingDown**, **sitting**, **walking**, **household**, **standing** and **cycling**. The figure 3.1 indicates the low-level classes that were merged in order to obtain these six classes:

- ❏ **lyingDown** ← lyingDownLeft, lyingDownRight
- ❏ **sitting** ← sittingSofa, sittingChair, sittingCouch
- ❏ **walking** ← walkingSlow, walkingNormal, walkingFast, walkingStairsUp, step
- ❏ **household** ← stakingShelves, vacuumCleaning, dishwashing
- ❏ **cycling**
- ❏ **standing**

Figure 3.1: Converted Classes

The comparison between frequency of the accelerometer data that was collected before the lifestyle intervention, represented by $\textbf{Data}_{pre}$, with the frequency of accelerometer data that was collected after the lifestyle intervention, represented by $\textbf{Data}_{post}$ is given in table 3.2. In almost all the activities, we can observe that there was no significant change between aggregate frequency of $\textbf{Data}_{pre}$ and $\textbf{Data}_{post}$.
In this work, we focus on the baseline activity data i.e. $\textbf{Data}_{pre}$.

| Activity | $\textbf{Data}_{pre}$ (s) | $\textbf{Data}_{post}$ (s) |
|---|---|---|
| **Walking** | 4.33 | 4.47 |
| **Cycling** | 2.62 | 2.61 |
| **Household** | 17.43 | 16.53 |
| **Sitting** | 45.93 | 45.97 |
| **Lying Down** | 27.52 | 28.20 |
| **Standing** | 2.17 | 2.22 |

Table 3.2: Comparison of data distribution between before and after the intervention

We also checked for average weight gain or loss of all the participants using both $\textbf{Data}_{pre}$ and $\textbf{Data}_{post}$. We found the average weight loss after lifestyle intervention over all participants to be 4.19%.

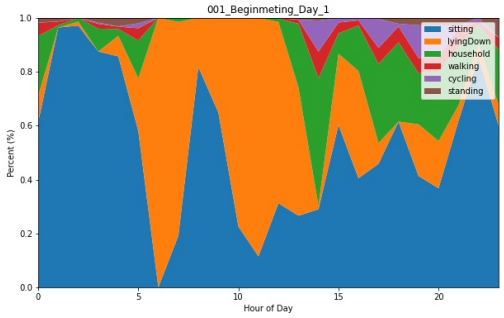
### 3.2.1 Quality Control

The data we obtained might contain some noisy data due to partial accelerometer wear or some classification error. Therefore, we further performed quality control of our data on all the participants. Using visual inspection from stacked area plots, we removed unwanted and incomplete days from our datasets. An example of the stacked area plots used for filtering the data is shown in the figure 3.3. The x-axis represents hours of day and the y-axis represents the corresponding total percentage for each activity.

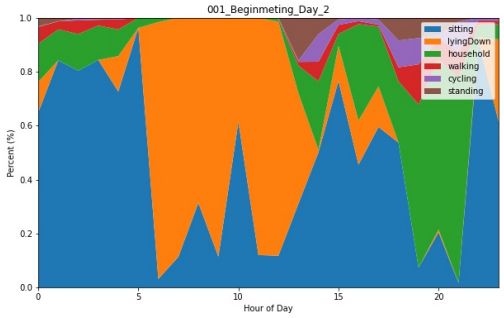
In figure 3.3, we have stacked area plots for eight days of a random participant. In day 8, we can observe some abnormality in the plot in certain hours of day i.e. from 04:00 to 14:00. It shows that the given participant was only lying down without performing any other activity for that complete period followed by complete sitting in the further periods, which is not a normal behavior of this person. In plots for day 1 to day 7, we can observe normal behaviour of lying down and sitting during those periods with combination of other activities i.e. these days represent normal everyday activity behaviour of that person. Therefore, we will disregard day 8 in this case for this particular participant. The reason behind this could be that day 8 contains low involvement in physical activities, which might be due to lack of device-wearing time. We performed this task for all the participants in our study so we don't have irregular results while performing the clustering tasks.

Figure 3.2 represents the legend for the given plots in figure 3.3.

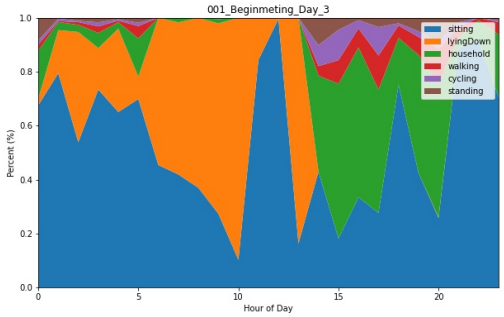

Figure 3.2: Legend for plots in figure 3.3



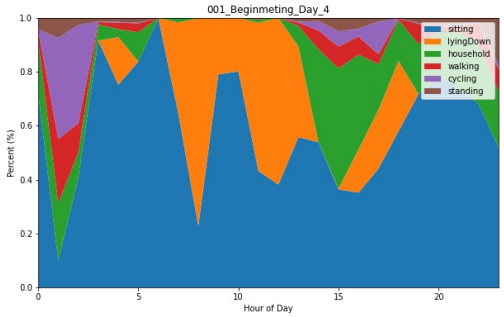

(a) Day 1



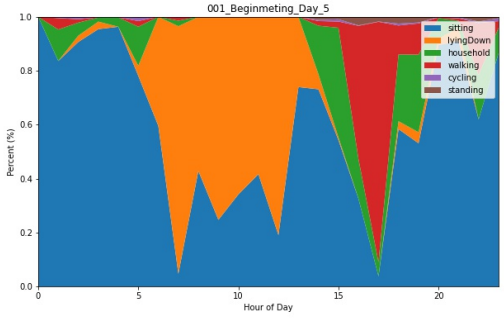

(b) Day 2



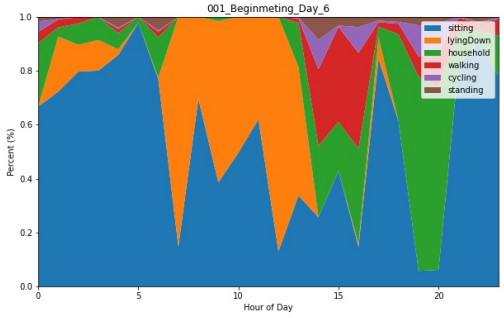

(c) Day 3



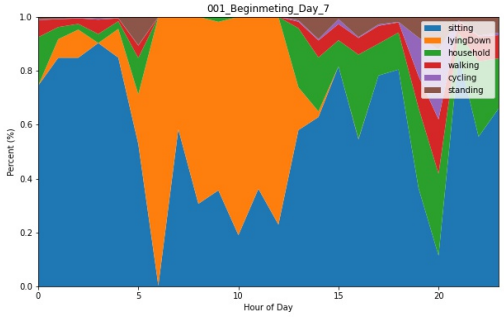

(d) Day 4



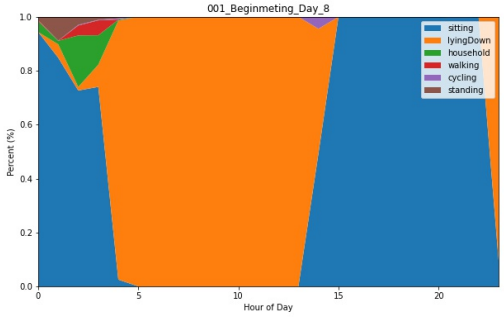

(e) Day 5




(f) Day 6




(g) Day 7




(h) Day 8

Figure 3.3: Example - Stacked area plot of one of the participants

### 3.2.2 Accelerometer Data Comparison

A total of 164 elderly individuals participated in the GOTO study [33]. We explored and compared statistics of all GOTO participants (N=164) with statistics of the subset of participants with both ankle and wrist accelerometer data (N=114), and also the subset of participants with only ankle accelerometer data (N=132) to investigate whether the data we are using for our experiments is representative of the data provided by the GOTO study. Table 3.3 shows the comparison between different datasets explored. While using just the ankle accelerometer, the movement of only the ankle is being considered. Since we want to have fewer anomalies and more clear patterns in order to find better clusters, we decided to use data that was collected for participants using both ankle and wrist accelerometers. As we can observe from the demographic results of 114 participants, the mean age and BMI is almost similar to that of 164 participants.

| **All GOTO Participants (N=164)** | | |
|---|---|---|
| | **Male** | **Female** |
| **Number** | 83 | 81 |
| **Age** (mean ± std) | 63.84 ± 5.41 | 62.05 ± 5.88 |
| **BMI** (mean ± std) | 26.18 ± 2.19 | 26.60 ± 2.84 |
| **Participants with only Ankle Accelerometer Data (N=132)** | | |
| **Number** | 68 | 64 |
| **Age** (mean ± std) | 63.57 ± 5.20 | 61.55 ± 5.88 |
| **BMI** (mean ± std) | 26.64 ± 2.06 | 27.25 ± 2.81 |
| **Participants with both Ankle and Wrist Accelerometer Data (N=114)** | | |
| **Number** | 60 | 54 |
| **Age** (mean ± std) | 63.68 ± 5.33 | 61.70 ± 5.63 |
| **BMI** (mean ± std) | 26.63 ± 2.15 | 27.09 ± 2.74 |

Table 3.3: Accelerometer Data Comparison - Demographic Results

### 3.2.3 Aggregation

Although the original data is classified at 1 Hz, we are interested in aggregating the data at higher temporal scales. This is because we are interested in a broad understanding of how someone spends their day, rather than what they are doing (or what the AR algorithm thinks they are doing) second to second. Furthermore, it can be expected that the AR algorithm makes some minor mistakes, and the activity classes are perhaps more reliable when aggregated over a longer time window. For example, if the algorithm decides that there is a sequence of five minutes of cycling, but within that sequence, there is one anomalous classification of lyingDown for a second, this is quite unlikely, and labeling the entire sequence as cycling is probably justified.

We will consider aggregated versions of the label data for different time windows. These include

windows of 1 s (no aggregation), 10 s and 30 s. Each window will end up having a single label (and thus aggregates 1, 10 or 30 labels). Additionally, if a window is $n$ seconds long, the next window will be computed in $n$ seconds also (the window length equals the stride length).

Having now specified the various window lengths, we still need to decide on the method for aggregation the $n$ labels into a single label per window. In fact, we will consider different options for aggregation. An obvious choice for computing the window label from $n$ detailed labels is to take the majority class (breaking ties arbitrarily). This means that infrequent labels in the sequence are deemed anomalies, and thus are ignored in favor of the majority label. Our cycling case is a good example of this. An alternative choice is to assign a window the label that has at least some minimum fraction of frequency in the sequence of labels, for example at least 80% or 90% frequency. This introduces the possibility that the aggregation method actually doesn't assign a label to all windows, since some windows might have quite mixed labels. This behavior is intentional, since you might opt to base your analysis only on windows that are reliable, and ignore the windows that are ambiguous. Where the former method assigns a label to all windows, the latter only assigns labels to the unambiguous windows.

Combining the different window lengths and aggregation methods, we opt for the following aggregations:

**i. Baseline: per Second**

This method simply takes windows of 1 s, and no aggregation is required. As argued, this method does not address any misclassification in the AR.

**ii. 10 s: Majority Class**

We extract the time window with intervals of 10 s, and the majority label.

**iii. 10 s: 80% Threshold**

Next, with the extracted time window of 10-seconds intervals, we only consider the single activity that has a frequency of more than 80%. Note that this method will discard any windows for which no such activity exists (mixed sequence of labels).

**iv. 30 s: 90% Threshold**

Finally, we extract the time window with intervals of 30-seconds and only considered the single activity that has a frequency of more than 90%. An example of this aggregation method is demonstrated in figure 3.4.

| TimeStamp | Activity | Count | Count (%) | |
|---|---|---|---|---|
| 2013-01-09 00:00:00 | walking | 28 | 93.0 | |
| 2013-01-09 00:00:00 | cycling | 2 | 7.0 | $1_{st}$ |
| 2013-01-09 00:00:30 | sitting | 17 | 55.0 | |
| 2013-01-09 00:00:30 | cycling | 10 | 35.0 | $2_{nd}$ |
| 2013-01-09 00:00:30 | household | 3 | 10.0 | |
| 2013-01-09 00:01:00 | cycling | 14 | 47.0 | |
| 2013-01-09 00:01:00 | lyingDown | 9 | 30.0 | |
| 2013-01-09 00:01:00 | walking | 5 | 16.0 | $3_{rd}$ |
| 2013-01-09 00:01:00 | household | 2 | 7.0 | |
| 2013-01-09 00:01:30 | sitting | 29 | 96.0 | |
| 2013-01-09 00:01:30 | household | 1 | 3.0 | $4_{th}$ |

Figure 3.4: Data Example of 30s with 90% Threshold Aggregation

In figure 3.4, we can observe four different 30-seconds time window. In $1_{st}$ time window, we notice that *'walking'* accounts for 93% of the total activity in that window. Similarly, in the $4_{th}$ window, *'sitting'* accounts for 97%, so we will use these time windows for our experiments. In the $2_{nd}$ and $3_{rd}$ time window, there is no activity with threshold more than 90%. Therefore, we disregard those time intervals for our experiments.

| Time Window | Seconds | Days |
|---|---|---|
| **Baseline** | 84898008 | 982.6 |
| **10s: Majority Class** | 84681680 | 980.1 |
| **10s: 80% Threshold** | 77407970 | 895.9 |
| **30s: 90% Threshold** | 65755560 | 761.1 |

Table 3.4: **Data**$_{pre}$ (Amount of data)

| Time Window | Seconds | Days |
|---|---|---|
| **Baseline** | 85415629 | 988.6 |
| **10s: Majority Class** | 85348390 | 984.8 |
| **10s: 80% Threshold** | 77075620 | 892.1 |
| **30s: 90% Threshold** | 66067680 | 764.7 |

Table 3.5: **Data**$_{post}$ (Amount of data)

The amount of total data in seconds and days for **Data**$_{pre}$ and **Data**$_{post}$ is shown in tables 3.4 to 3.5. From the table 3.4 for **Data**$_{pre}$, we notice that without aggregation, there are about

982 days of data, whereas with 30 s time window having activity with 90% threshold, we lose approximately 222 days. Similarly, from table 3.5 for $\mathbf{Data}_{post}$, we can observe that the number of lost days is approximately 224 days. We decided to use the 30-seconds aggregation with a 90% threshold for our experiment purpose because of the following reasons:

- For active profiles, we are primarily interested in continuous prolonged activities.
- With this time window, we expect to filter out small misclassifications.
- Although we are losing some fraction of the time windows (roughly 22%), we are also gaining certainty in the data.

Therefore, we used the 30 s: 90% Threshold aggregation for constructing features for our dataset.

## 3.3 Feature Construction

To construct features from our processed data for every participants, we extracted frequency of 30 second aggregation for each of the activities per hour for all individual days. Then we normalised all the frequencies by computing the average over those days. Therefore, for each participants, we have extracted following features:

### 3.3.1 Average Per Hour

To identify minor changes in daily physical activities, hourly measurements are crucial. Accordingly, we decided to construct features by computing the average of each of the participants with hourly frequency of the activities i.e. for every activity we have 24 features. Remember our dataset has 6 activity labels. Therefore, in total we have 24 x 6 features i.e. 144 features.

### 3.3.2 Period of Hours

Reducing the number of features helps to remove multi-collinearity from our data, which is beneficial for our machine learning models. It also aids in tweaking the speed of computations for these models. In order to reduce the number of features, we merged the earlier extracted 144 features together with respect to different period of time i.e. we combined hourly averages into the following four six-hour quadrants:

- Night: 00:00 to 06:00
- Morning: 06:00 to 12:00
- Afternoon: 12:00 to 18:00
- Evening: 18:00 to 23:00

Therefore, in this representation, we have 4 x 6 features i.e. 24 features.

## 3.4 Z-Normalisation

The features we constructed from the data have different ranges. Therefore, in order to convert the values of numeric columns in our dataset to a common scale, we performed Z-normalization. Z-Normalization refers to the process of normalizing every single column in a dataset in such a way that the mean of the values is 0 and the standard deviation is 1.

**Z-Normalization Formula:**

$$x' = \frac{(x - \mu)}{\sigma}$$

where, $x$ is the original value, $\mu$ is the mean and $\sigma$ is the standard deviation of the data. We used the formula above to perform Z-Normalization on every column of our dataset.

# Chapter 4

# Methods

For our experiments, we made extensive use of different clustering techniques. This chapter describes in detail the clustering techniques used and their respective functioning.

## 4.1 Cluster Analysis

Cluster analysis or clustering [16] is a statistical method used to group similar objects, i.e. data points, into different categories based on the degree of similarity between them. The features of the data points define this similarity. It is a commonly used technique for performing exploratory data analysis. It is used in many fields such as image analysis, information retrieval, pattern recognition and machine learning. Clustering is an unsupervised learning method, meaning we do not have predefined/existing categorical labels for the data points that are supposed to be categorized into clusters. For our experiments, we used two unsupervised clustering methods:

1. $k$-means Clustering Algorithm
2. Hierarchical Clustering Algorithm

A detailed explanation of each method, the data points used in each respective method, and the features of these data points are described in the following sections.

## 4.2 $k$-means Clustering Algorithm

$k$-means clustering algorithm [26, 8] is one of the most popular unsupervised machine learning algorithms. The algorithm aims to find patterns in the data based on similarity and aggregate the data points into $k$ number of different clusters. Each cluster is defined by the "centroid" (cluster centre), which is nothing but the data point that possesses the mean feature vector values of all the data points in that cluster. For our experiments, we used the *scikit-learn* [38] package library for Python to perform $k$-means clustering. Clustering is performed on a set of 114 data points, with each having the following features: 144 features for Average per Hour; 24 features for Period of Hours.

The Algorithm 1 describes the pseudo-code for $k$-means clustering algorithm.

**Algorithm 1** $k$-means Algorithm Pseudo-code

**Input:**

$D = d_1, d_2..., d_n$ ▷ set of $n$ data items

$k$ ▷ number of desired clusters

**Output:**

A set of $k$ clusters

**Step:**

**1.** Arbitrarily choose $k$ data-items from $D$ as initial centroids.

**2. Repeat**

- Assign each item $d_i$ to the cluster which has the closest centroid based on Euclidean distance.
- Calculate new mean for each cluster.
- Assign the new centroid of each cluster based on the respective means.

**Until convergence criteria is met**

The convergence criteria for $k$-means is either based on the maximum number of iterations or on the stability of clusters formed, i.e. the iterations must stop once no further change in clusters (and cluster centres) is observed.

The number of clusters must be determined beforehand for the $k$-means algorithm. The algorithm itself cannot predict the number of clusters in the data. Deciding the number of clusters is a challenging task. In our work, we used two metrics that gave us some intuition about the optimal number of clusters, $k$:

1. Elbow Method
2. Silhouette Analysis

Each of these methods is described in detail in the further subsections.

### 4.2.1 Elbow Method

The Elbow method [5, 27, 47] is used to determine the optimal number of clusters to be formed in $k$-means clustering. As the number of clusters increases, the number of data points in each cluster decreases, decreasing the average dispersion of data points in each cluster. In other words, with the increasing number of clusters, the data points in each cluster come closer to the cluster centre. Figure 4.1 [43] depicts an example of the average dispersion versus number of clusters $k$ plot. The bend in the graph represents the "elbow point", and the corresponding $k$ value represents the optimal number of clusters. The elbow point or the bend in the graph represents the optimal number of clusters, as shown in Figure 4.1. Therefore, the optimal number of clusters for this example would be 3.

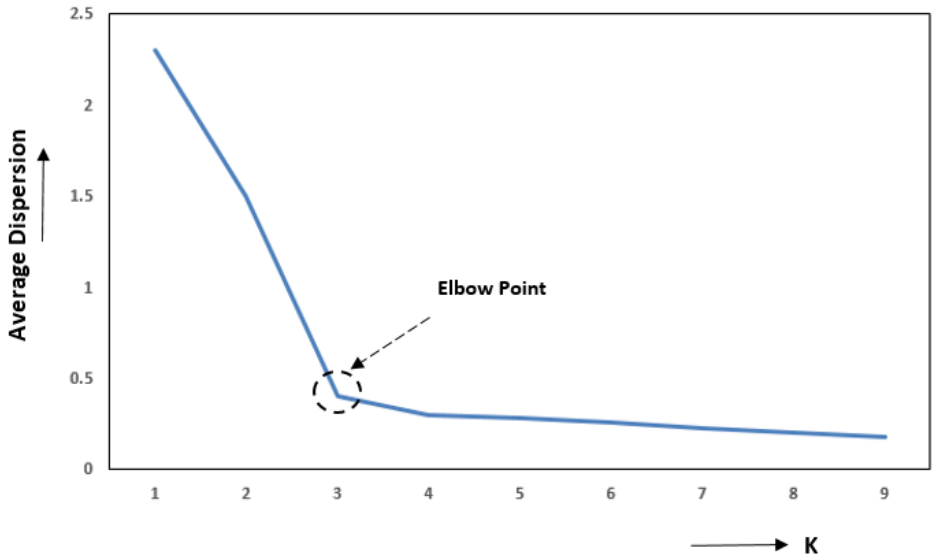



Figure 4.1: Elbow Method Example

The elbow method runs $k$-means clustering on a dataset for a range of values for $k$ and then computes an average score for all clusters for each value. By default, the scoring parameter metric is set to distortion, which computes the sum of squared distances from each point to its assigned centre.

### 4.2.2 Silhouette Coefficient

Silhouette Coefficient or Silhouette score [35, 39] is a metric used to calculate the "goodness" of a clustering technique. Any clustering technique aims to minimize the distance between a cluster centre and the data points in that cluster and maximize the distance between different cluster centres so that clusters are well distinguished. The Silhouette score is quite helpful in validating the working of the clustering algorithm when dealing with higher dimensions. It can attain three different values, i.e. 1, 0 and -1. The significance of each of these values is as follows:

- 1: Suggests that clusters are well-separated from each other
- 0: Suggests that distance between clusters is insignificant
- -1: Suggests that clusters are assigned incorrectly

The Silhouette score can be calculated as follows:

$$\textbf{Silhouette Score} = \frac{(b-a)}{max(a,b)}$$

where, $a$ is the average intra-cluster distance i.e. the average distance between data points within a cluster, and $b$ is the average inter-cluster distance i.e. the average distance between clusters/cluster centres.

## 4.3 Hierarchical Clustering Algorithm

Hierarchical clustering, also known as hierarchical cluster analysis [18], is another clustering algorithm that categorizes similar objects/data points into several groups. The result is a set of clusters, where each cluster is distinct from every other cluster, and the data points within each cluster are broadly similar. There are mainly two types of hierarchical clustering techniques available:

- Agglomerative Hierarchical Clustering [42]
- Divisive Hierarchical Clustering [36, 32]

For our experiments, we used the agglomerative hierarchical clustering technique, which is described in detail in the following subsection.

### 4.3.1 Agglomerative Hierarchical Clustering

In this type of hierarchical clustering [12, 48], we build a cluster tree to represent data, where each group of data points links to two or more successor groups. These groups are nested and organized as a tree, ideally as a meaningful classification scheme. For our experiments, we perform agglomerative clustering on a set of $114$ data points, with each data point representing a participant and each participant or data point having the following number of features: $144$ features for Average per Hour; 24 features for Period of Hours.

Each node in the cluster tree contains a group of similar data; nodes group on the graph next to other similar nodes. Clusters at one level join with clusters in the next level, using a degree of similarity. The process continues until all nodes are in the tree, providing a visual snapshot of the data in the whole set. The total number of clusters is not predetermined before the start of tree creation.

### 4.3.2 Dendrogram

A dendrogram [44, 37] is a type of tree diagram that depicts hierarchical clustering relationships between similar sets of data. We make use of dendrograms to determine the optimal number of clusters. They are frequently used in biology to show clustering between genes or samples, but they can represent any grouped data. To demonstrate the working of dendrogram, let us consider the figure 4.2 [6]. The figure shows the hierarchical clustering of six observations shown on the scatter plot to the left.

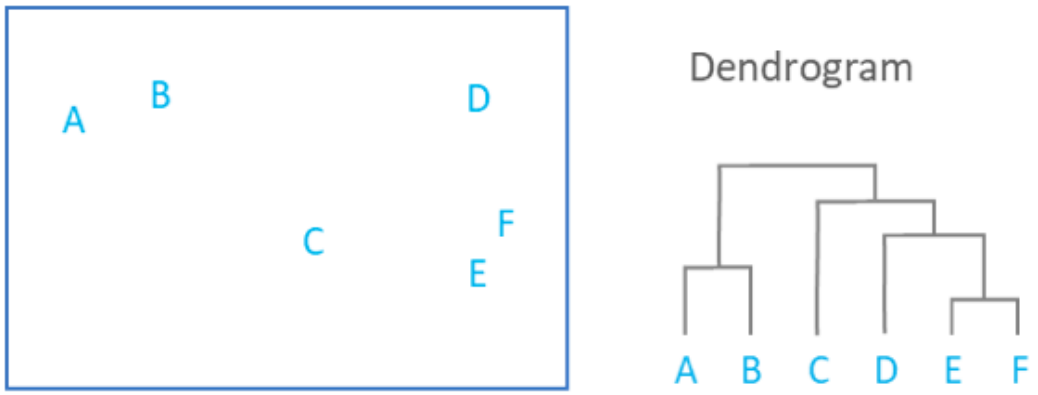



Figure 4.2: Dendrogram Example

In Figure 4.2, we notice that data points $E$ and $F$ are most similar, as the height of the link that joins them together is the smallest. The following two most similar objects are $A$ and $B$. In the dendrogram above, the height of the dendrogram indicates the order in which the clusters are joined. In this dendrogram, we can observe a significant difference between the clusters of $A$ and $B$ with that of $C$, $D$, $E$, and $F$.

## 4.4 Principal Component Analysis

Principal component analysis (PCA) [30, 17] is a dimensionality reduction technique. It aims to reduce the dimensions of a dataset such that for each data point, the trivial features are discarded, and a set of new features is created wherein only the important features, also referred to as principal components, are retained. The new features formed are uncorrelated and retain the highest amount of information and variation in the data. From the new features/principal components thus formed, the first component would have the highest amount of information, the second one would have the information that the previous component does not contain and so on.

We used the *sklearn.decomposition* [41] module to perform principal component analysis for dimensionality reduction for our generated features. In our experiments, we performed PCA on Average per Hour with $144$ features for each participant. For the features mentioned above, a covariance matrix is constructed, and its eigenvalues are calculated, based on which we obtain the principal components for our dataset.

# Chapter 5

# Results and Discussion

In this chapter, we will describe the results obtained from our experiments. We used the data collected from 114 participants with both ankle and wrist data for our experiments.

Objective for each sections are explained briefly as follows:

5.1 To remove data related problems, i.e. find anomalies in data using the $k$-means clustering algorithm.

5.2 To check the effect of principal component analysis on our high-dimensional feature dataset.

5.3 To profile the participants using $k$-means and hierarchical clustering algorithms.

5.4 To compare our methods of dimensionality reduction techniques, i.e. PCA with time-aggregated feature (Period of Hours)

5.5 To identify clearer activity behaviour in our participants after division by gender.

For all experiments described in the further subsections, we used the Silhouette coefficient as the evaluation metric to measure the quality of clusters obtained.

## 5.1 Initial Run

The objective of the initial run of experiments is to perform quality control, i.e. remove data-related problems and detect outliers. We aim to find anomalies in our data using the $k$-means clustering algorithm. To determine the optimal number of clusters, we used Silhouette score as the evaluation metric. We translated the 30-second aggregates to minutes for easier interpretation of the results and plots.
The following subsections discuss the demographic results and the normalized frequency of each activity of clusters obtained, using $k$-means clustering on our constructed feature-dataset, namely Average per Hour with 144 features and Period of Hours with 24 features.

### 5.1.1 Average per Hour (144 Features)

The possible number of clusters and their respective Silhouette scores for the Average per Hour feature-dataset are given in the table 5.1.

| No. of Clusters | Silhouette Coefficient |
|---|---|
| 2 | 0.145 |
| 3 | 0.128 |
| **4** | **0.152** |
| 5 | 0.044 |
| 6 | 0.041 |
| 7 | 0.030 |
| 8 | 0.050 |
| 9 | 0.042 |

Table 5.1: Silhouette Score (Average per Hour: 144 Features)

Using the Silhouette score, as seen in Table 5.1, we selected $k = 4$ as the optimal number of clusters for the $k$-means clustering algorithm. We also checked the scree plot for additional proof of our selection of $k$. Figure 5.1 depicts the scree plot for the same input data. We can observe the elbow point, marked with a red dot in the plot which also indicates four as the optimal choice for $k$. Afterwards, we used the $k$-means clustering algorithm to obtain clusters.

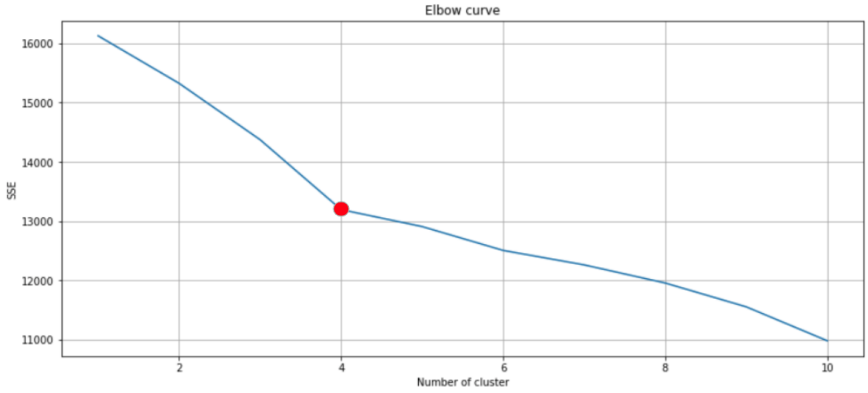



Figure 5.1: Scree Plot (Average per Hour) - 144 Features

In Table 5.2, we can observe demographic results of the obtained clusters. Clusters C-1 and C-2 contain the most participants, whereas clusters C-0 and C-3 are smaller, but these clusters contain a higher proportion of female participants than the other two clusters.

| Cluster | C-0 | C-1 | C-2 | C-3 |
|---|---|---|---|---|
| **Size** | 5 | 50 | 46 | 12 |
| **G: Female** | 4 | 21 | 21 | 8 |
| **G: Male** | 1 | 29 | 25 | 4 |
| **Age** | 66.22 ± 2.20 | 62.40 ± 5.98 | 62.66 ± 5.38 | 63.26 ± 5.27 |
| **BMI** | 26.43 ± 2.17 | 26.72 ± 2.57 | 27.07 ± 2.32 | 26.61 ± 2.91 |

Table 5.2: Demographic Results: Average per Hour (144 Features)

In Table 5.3 we show the normalized activity frequency results of the obtained clusters. Cluster C-0 has more household activities and walking frequency than other clusters. For cluster C-3, we can notice that the standing frequency is much higher than in the other clusters. Cluster C-2 seem to be participating more in sedentary activities such as lying down.

| Cluster | C-0 | C-1 | C-2 | C-3 |
|---|---|---|---|---|
| **Walking** | 25.61 | 20.76 | 18.63 | 18.15 |
| **Household** | 188.26 | 74.15 | 74.59 | 102.69 |
| **Cycling** | 5.05 | 16.90 | 13.76 | 14.02 |
| **Sitting** | 319.18 | 380.72 | 301.80 | 298.14 |
| **Lying Down** | 348.48 | 319.71 | 465.96 | 289.32 |
| **Standing** | 17.22 | 15.11 | 11.76 | 104.69 |

Table 5.3: Activity Frequency - Normalized: Average per Hour (144 Features)

We reviewed the participants of clusters C-0 and C-3 to remove the possible unwanted days that we might have missed while performing quality control for our dataset using area plots. We note the participants clustered together in each clusters below along with their ID, for future comparison.

[Cluster C-0: 60015, 60890, 61284, 62419, 70232]
[Cluster C-3: 60113, 60123, 60361, 60361, 60362, 60963, 62194, 62278, 70516, 70589, 70668, 70780, 70801]

### 5.1.2 Period of Hours (24 Features)

The possible number of clusters and their respective Silhouette scores for the Period of Hours feature-dataset are given in the Table 5.4.

| No. of Clusters | Silhouette Coefficient |
|---|---|
| 2 | 0.103 |
| 3 | 0.094 |
| **4** | **0.116** |
| 5 | 0.077 |
| 6 | 0.070 |
| 7 | 0.072 |
| 8 | 0.042 |
| 9 | 0.043 |

Table 5.4: Silhouette Score (Period of Hours: 24 Features)

We can infer from Table 5.4 that the highest Silhouette score corresponds to $k = 4$. Therefore, in this case too, we used four as our optimal choice for number of clusters.

| Cluster | C-0 | C-1 | C-2 | C-3 |
|---|---|---|---|---|
| **Size** | 7 | 12 | 82 | 12 |
| **G: Female** | 6 | 5 | 34 | 9 |
| **G: Male** | 1 | 7 | 48 | 3 |
| **Age** | 64.58 $\pm$ 4.29 | 59.82 $\pm$ 4.43 | 63.07 $\pm$ 5.71 | 62.62 $\pm$ 5.59 |
| **BMI** | 25.85 $\pm$ 2.13 | 26.27 $\pm$ 2.81 | 27.05 $\pm$ 2.38 | 26.50 $\pm$ 2.92 |

Table 5.5: Period of Hours (24 Features)

Table 5.5 displays demographic results for the obtained clusters for 114 participants using the features constructed concerning different periods of time, i.e. 24 features. We can observe that cluster C-2 contains the most participants. The mean age of cluster C-0 is the highest, whereas cluster C-1 has the minimum mean age among the four clusters. The mean BMI seems balanced in all groups with minor differences.

| Cluster | C-0 | C-1 | C-2 | C-3 |
|---|---|---|---|---|
| **Walking** | 25.01 | 18.01 | 19.87 | 18.38 |
| **Household** | 194.08 | 70.12 | 70.74 | 109.33 |
| **Cycling** | 9.04 | 43.12 | 10.8 | 17.10 |
| **Sitting** | 295.85 | 350.81 | 346.73 | 280.64 |
| **Lying Down** | 366.96 | 311.78 | 400.46 | 290.47 |
| **Standing** | 14.50 | 15.92 | 13.08 | 106.21 |

Table 5.6: Period of Hours (24 Features)

In Table 5.6, we can observe the normalized activity frequency of the obtained clusters. Cluster C-0 has more walking frequency and household activities than other clusters. Participants of cluster C-1 have more cycling and sitting frequency, whereas cluster C-2 took more part in other sedentary activities. Furthermore, for Cluster C-3, the standing frequency is higher than the other clusters.

We reviewed the participants of clusters C-0, C-1 and C-3 to find possible anomalies. We can observe the participants clustered together in each clusters below along with their ID.

[Cluster C-0: 60015, 60822, 60890, 60968, 61284, 62419, 70232]
[Cluster C-1: 60666, 60823, 60871, 61479, 61480, 62020, 62266, 62340, 62370, 70001, 70468, 70766]
[Cluster C-3: 60113, 60123, 60361, 60362, 60963, 62194, 62278, 70516, 70589, 70668, 70780, 70806]

After further manual investigation of participants of the clusters above, we discovered anomalies in the accelerometer data for almost all participants for some days.

### 5.1.3 Conclusion

We performed $k$-means clustering for both Average per Hour with 144 features and Period of Hours with 24 features, for $k = 4$. After looking into the clusters formed, we found that the participants of two clusters (C-0 and C-3) obtained after clustering for Average per Hour with 144 features dataset, were the same as the respective participants of the corresponding two clusters (C-0 and C-3) obtained after clustering for Period of Hours with 24 features dataset. After comparing and checking all the statistics of the obtained clusters, we found that the $k$-means algorithm was clustering together participants with some unusual reading, i.e. we could detect outliers with our initial run of $k$-means algorithm in the clusters obtained. We discovered anomalies in the accelerometer data for some days containing low physical activity involvement, possibly due to a lack of device-wearing time or incorrect activity classification. After identifying the anomalies, we removed those days from our data for the selected participants. We used the obtained clean data for further analysis.

## 5.2 Average per Hour (144 Features) Vs. PCA

On the obtained clean data, we performed the analysis again for the constructed feature-dataset with 144 features. We determined the optimal number of clusters using the Silhouette score. Later, we used a dimensionality reduction technique on our hourly feature to check if we could obtain similar results. We used principal component analysis to reduce the number of dimensions in our data. After performing PCA, we used $k$-means clustering algorithms to obtain clusters. The results obtained before PCA and after performing PCA on the given 144 features are described in the following subsections.

### 5.2.1 Average per Hour (144 Features)

The possible number of clusters and their respective Silhouette scores for the Average per Hour feature-dataset, after data cleaning, are given in the Table 5.7. We can notice that the

Silhouette score is highest for $k = 2$. Therefore, we selected 2 as the best choice for optimal number of clusters.

| No. of Clusters | Silhouette Coefficient |
|---|---|
| **2** | **0.110** |
| 3 | 0.044 |
| 4 | 0.035 |
| 5 | 0.033 |
| 6 | 0.029 |
| 7 | 0.025 |
| 8 | 0.024 |
| 9 | 0.024 |

Table 5.7: Silhouette Score (Average per Hour)

The demographic results of $k$-means clustering with $k = 2$ for the Average per Hour feature-dataset, after data cleaning, are summarized in the Table 5.8.

| Cluster | C-0 | C-1 |
|---|---|---|
| **Size** | 107 | 7 |
| **G: Male** | 57 | 3 |
| **G: Female** | 50 | 4 |
| **Age** | 63.01 $\pm$ 5.52 | 57.44 $\pm$ 5.20 |
| **BMI** | 26.78 $\pm$ 2.40 | 28.24 $\pm$ 3.48 |

Table 5.8: Demographics Results (Average per Hour)

In Table 5.8, cluster C-0 contains majority population with 107 participants, whereas cluster C-1 contains only 7 participants. The gender distribution appears to be balanced in both clusters. The difference between mean age of cluster C-0 and cluster C-1 is almost five years which could be one reason for them being clustered apart. The mean BMI for participants of cluster C-1 is slightly more than for those of cluster C-0.
The normalized activity frequencies for the clusters obtained are summarized in Table 5.9.

| Cluster | C-0 | C-1 | All |
|---|---|---|---|
| **Walking** | 21.36 | 23.69 | 21.51 |
| **Household** | 84.10 | 92.51 | 84.61 |
| **Cycling** | 13.99 | 54.81 | 16.50 |
| **Sitting** | 354.81 | 269.37 | 349.57 |
| **Lying Down** | 331.70 | 243.32 | 326.27 |
| **Standing** | 11.10 | 7.56 | 10.88 |

Table 5.9: Activity Frequency - Normalized (In Minutes) (Average per Hour)

There is a minor difference in the frequency of walking and household activities for active

activities between both the obtained clusters. Cluster C-1 is more involved in cycling than cluster C-0. For all sedentary level activities, cluster C-1 is less engaged than cluster C-0. Overall, cluster C-1, with 7 participants, is more active and performs less sedentary activities than the majority of the population.

### 5.2.2 Average per Hour - PCA (37 Features)

Using PCA, we were able to reduce the number of features to 37 while retaining 80% variance by the features, as we can observe in Figure 5.2.

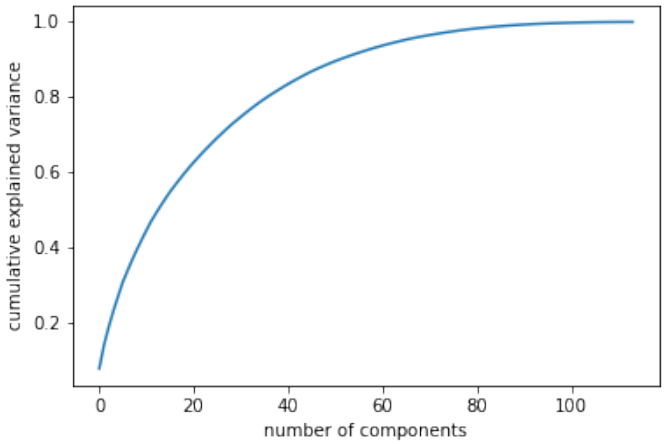



Figure 5.2: PCA Line Plot (Average per Hour)

In order to verify the optimal value for $k$ i.e. number of clusters, we used the Silhouette Coefficient.

| No. of Clusters | Silhouette Coefficient |
|---|---|
| **2** | **0.056** |
| 3 | 0.045 |
| 4 | 0.046 |
| 5 | 0.047 |
| 6 | 0.041 |
| 7 | 0.042 |
| 8 | 0.030 |
| 9 | 0.030 |

Table 5.10: Silhouette Score (Average per Hour - PCA)

In Table 5.10, we can see that for $k = 2$, the Silhouette Coefficient is closer to 1 as compared to those for other values of $k$. Therefore, we decided to use 2 as the optimal value of $k$ for $k$-means. The results that we obtained by performing $k$-means clustering with a value of $k = 2$, are shown in the tables 5.11 to 5.12.

| Cluster | C-0 | C-1 |
|---|---|---|
| **Size** | 72 | 42 |
| **G: Male** | 49 | 11 |
| **G: Female** | 23 | 31 |
| **Age** | 63.66 ± 5.11 | 60.96 ± 6.17 |
| **BMI** | 26.84 ± 2.42 | 26.90 ± 2.61 |

Table 5.11: Demographics Results (Average per Hour - PCA)

In Table 5.11 we have the demographic summary of the different clusters that we obtained. We can observe that the cluster C-0 has around 68% male participants, whereas cluster C-1 has around 73% female participants. This difference can also be observed in the mean age of clusters C-0 and C-1 with approximately 3 year gap in them as the female participants were on average a few years younger than the male participants. The mean BMI for both clusters are similar.

| Cluster | C-0 | C-1 | All |
|---|---|---|---|
| **Walking** | 23.31 | 18.41 | 21.51 |
| **Household** | 66.64 | 115.42 | 84.61 |
| **Cycling** | 12.22 | 23.82 | 16.50 |
| **Sitting** | 395.99 | 269.98 | 349.57 |
| **Lying Down** | 337.23 | 307.50 | 326.27 |
| **Standing** | 12.92 | 7.39 | 10.88 |

Table 5.12: Activity Frequency - Normalized (In Minutes) (Average per Hour - PCA)

The Table 5.12 shows the normalized frequency of the activities for both clusters i.e. C-0 and C-1. We can observe that the cluster C-0 is a bit more active in walking, in comparison to C-1. Whereas, in household activities, participants of cluster C-1 are more active. We can notice that the female participants are more active in household activities than the male participants, the same can be observed in the cluster C-1 as it has more female participants. The participants of cluster C-1 are almost twice more active in the house than the participants of cluster C-0. Whereas, for sedentary activities such as sitting, lying down and standing the frequency of cluster C-0 is more than cluster C-1. These statistics provide evidence that the participants of cluster C-1 are in general more active than participants of cluster C-0.

### 5.2.3 Conclusion

Even though we lost 20% variance by the features when we performed principal component analysis, we observed that the clusters obtained after PCA are clearer than those obtained by the original feature dataset with 144 features. From the clustering results after PCA, we could observe that the two obtained clusters have one dominant male group and another female dominant. Whereas, in the case of clustering before PCA, we were getting one cluster that contained more than 90% of the population size and the other with the remaining participants.

The difference between the results obtained by clustering with PCA and without PCA can be that the $k$-means clustering algorithm is sensitive to minor changes in the input data. Therefore, the number of clusters obtained can vary even with a slight difference in input data. Since we are getting more refined and clear clusters after performing PCA, for our further analysis, we decided to use the reduced feature dataset, i.e. with 37 features, which retains 80% variance. This would also increase the efficiency of the algorithms we are using for our experiments.

## 5.3 Profiling the participants using Average Per Hour

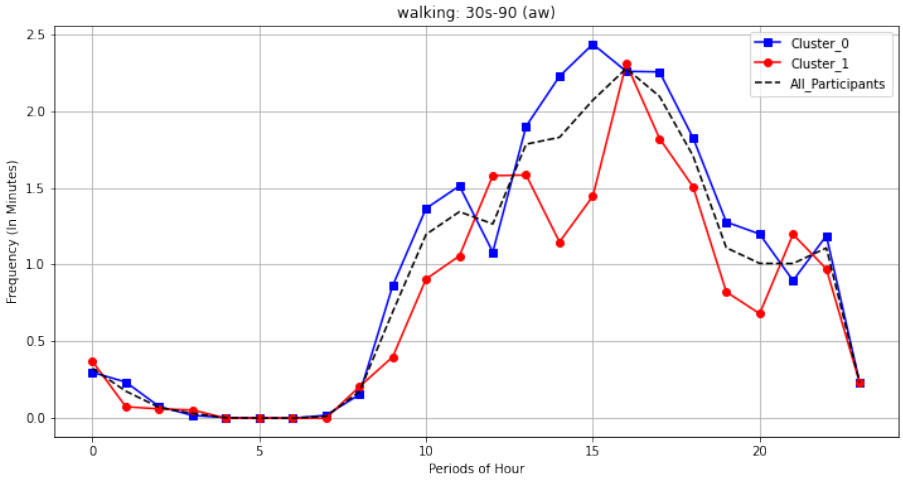
After cleaning the data of anomalies that we found in our initial run, we decided to use the 37 features we obtained by performing PCA on hourly data, i.e. average per hour with 144 features for the efficiency of experiments. We aim to profile the participants using two unsupervised learning algorithms, namely $k$-means and hierarchical clustering. We explain the acquired results using both clustering algorithms in the following subsections.

### 5.3.1 $k$-means Clustering Results

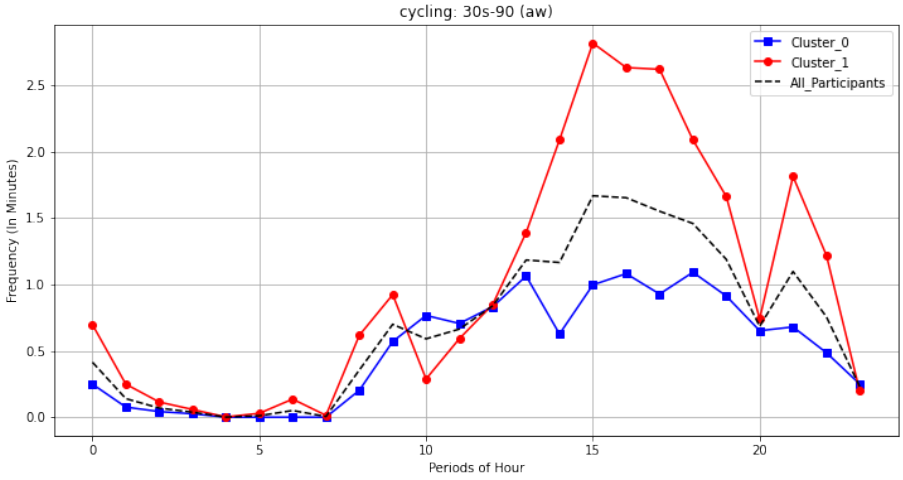
For the two obtained clusters using $k$-means clustering, we visually investigated each activity's normalized frequency per hour for 24 hours in figs. 5.3 to 5.4. In Figure 5.3a for walking, we can observe that cluster C-0 is more active than cluster C-1, with specific peaks from hours 13:00 to 18:00. In contrast, we can see that the participants of cluster C-1 have a sudden drop from 13:00 to 14:00. In Figure 5.3b for cycling, we can observe that cluster C-1 is more active overall than cluster C-0. For cluster C-1, we can observe peaks from 13:00 to 19:00 and then another sudden rise at 21:00. The reason for this is that some participants might prefer cycling at night. In contrast, most of them usually cycle around noon to evening. In the household activities 5.3c, we can see a significant difference in the frequency of both the clusters from 9:00 to 23:00, with cluster C-1 having more frequency. The reason could be that cluster C-1 consists of more female participants than cluster C-0.

In Figure 5.4b for sitting, cluster C-0 has more frequency overall, i.e. cluster C-0 is more active in sedentary activities than cluster C-1. The same can be observed in Figure 5.4a for lying down from 3:00 to 7:00, while for the rest of the day, both clusters show almost the same pattern of this activity overall. In Figure 5.4c for standing, we can observe that cluster C-0 spends more time overall than cluster C-1.

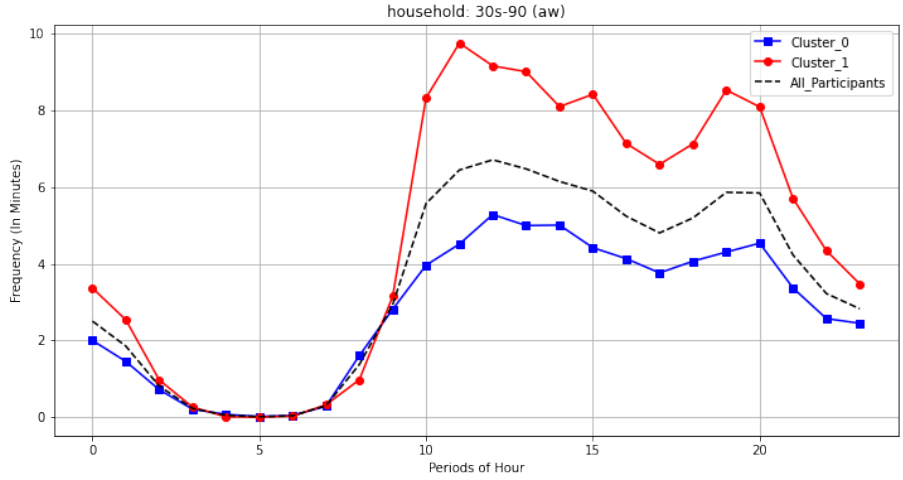
With all the evidence gathered from the investigation of the clusters obtained by the $k$-means clustering, we observed that the participants of cluster C-1 are more active than those in cluster C-0 in case of active activities whereas, for sedentary activities such as lying down, sitting and standing, participants of cluster C-0 have more frequency than the participants of cluster C-1.




(a) Walking

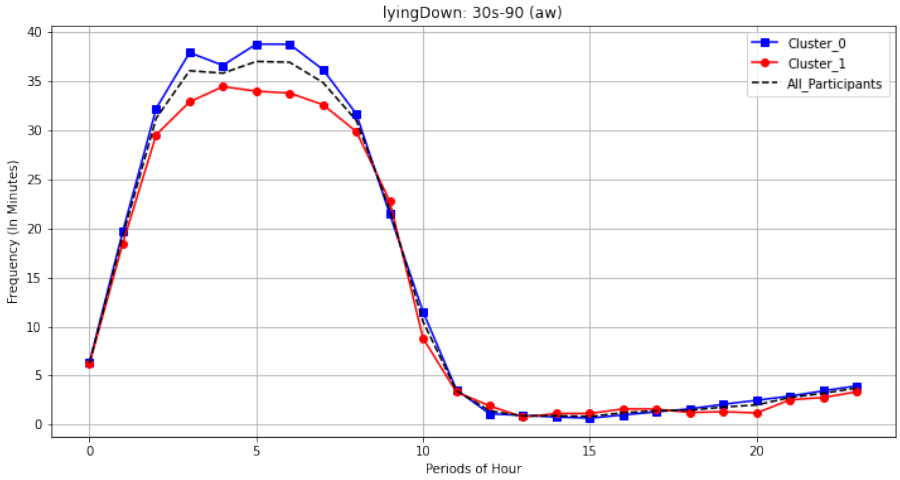



(b) Cycling

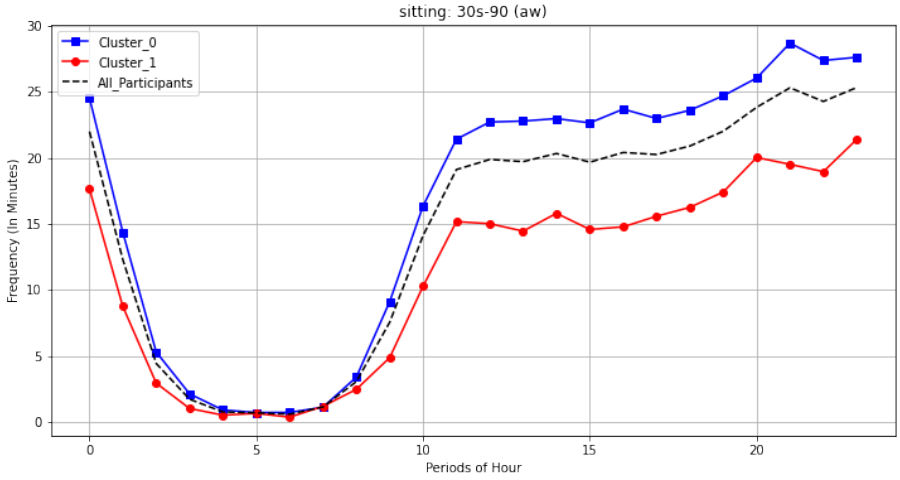



(c) Household

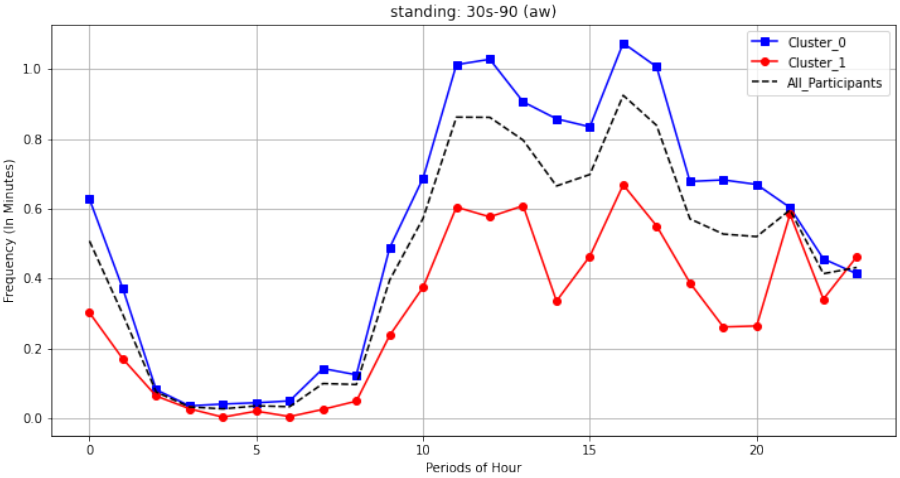
Figure 5.3: Normalized Frequency per Hour - Active Activities (K=2)




(a) Lying Down




(b) Sitting




(c) Standing

Figure 5.4: Normalized Frequency per Hour - Inactive Activities (K=2)

### 5.3.2 Hierarchical Clustering Results

After visually checking the dendrogram shown in Figure 5.5, we selected the total number of clusters as two, as we can see from the line that cuts the dendrogram at a distance of 35.

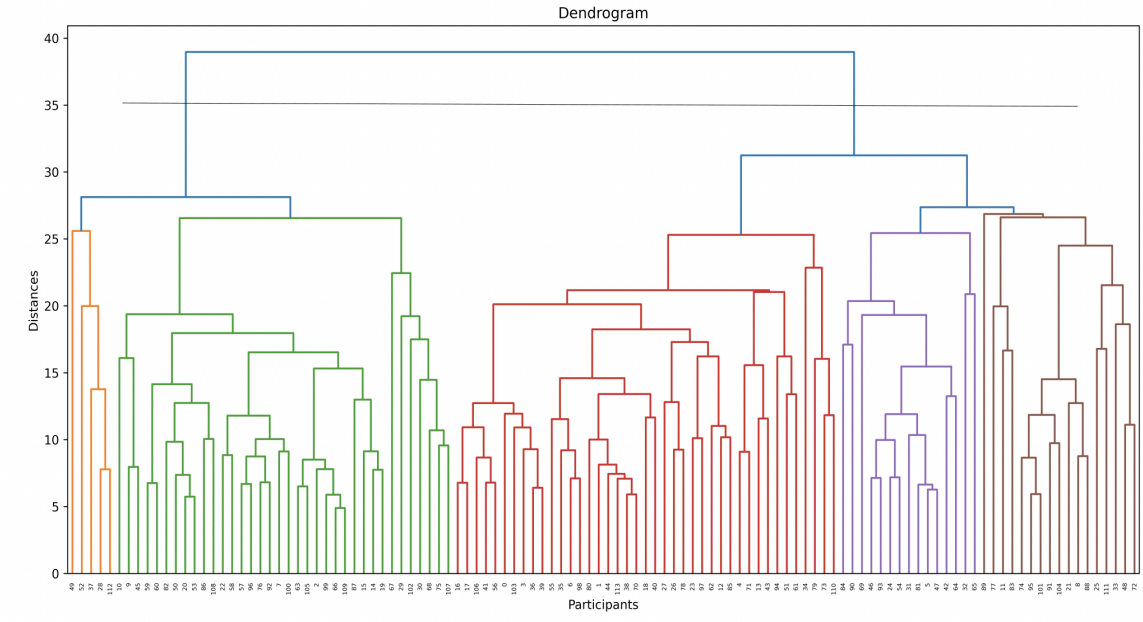



Figure 5.5: Dendrogram (Average per Hour - PCA)

After applying hierarchical clustering for two clusters, we get the following demographic results and normalized activity frequency.

| Cluster | C-0 | C-1 |
|---|---|---|
| **Size** | 73 | 41 |
| **G: Male** | 50 | 10 |
| **G: Female** | 23 | 31 |
| **Age** | 63.95 $\pm$ 5.16 | 60.35 $\pm$ 5.79 |
| **BMI** | 26.77 $\pm$ 2.26 | 27.04 $\pm$ 2.85 |

Table 5.13: Demographics Results (Average per Hour - PCA) (Hierarchical Clustering)

| Cluster | C-0 | C-1 | All |
|---|---|---|---|
| **Walking** | 23.80 | 17.43 | 21.51 |
| **Household** | 68.47 | 113.37 | 84.61 |
| **Cycling** | 12.43 | 23.75 | 16.50 |
| **Sitting** | 389.72 | 278.07 | 349.57 |
| **Lying Down** | 343.05 | 296.40 | 326.27 |
| **Standing** | 13.30 | 6.56 | 10.88 |

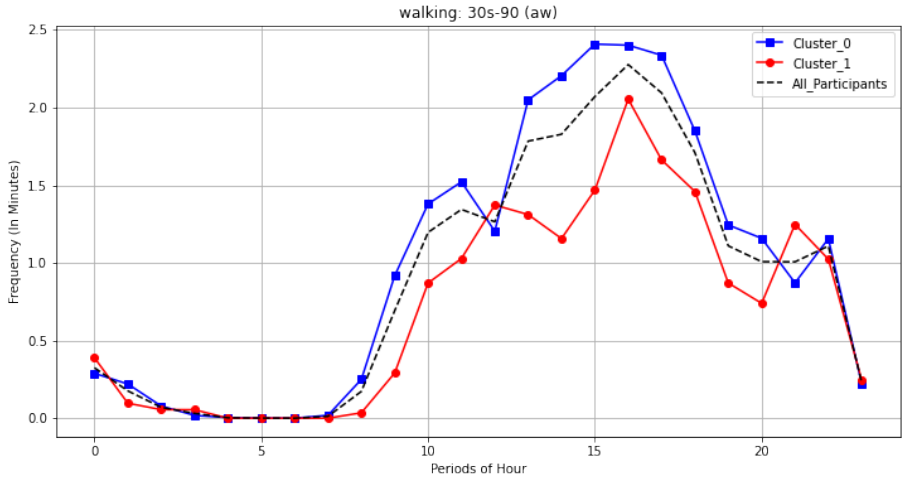
Table 5.14: Activity Frequency - Normalized (Average per Hour - PCA) (Hierarchical Clustering)

After surveying the results in Table 5.13 and 5.14, we can notice that the clusters obtained using hierarchical clustering has similar demographic and activity frequency results as those of $k$-means. The normalized frequency per hour for each activity in 24 hours, for hierarchical clustering for the two obtained clusters, is shown in figs. 5.6 to 5.7.

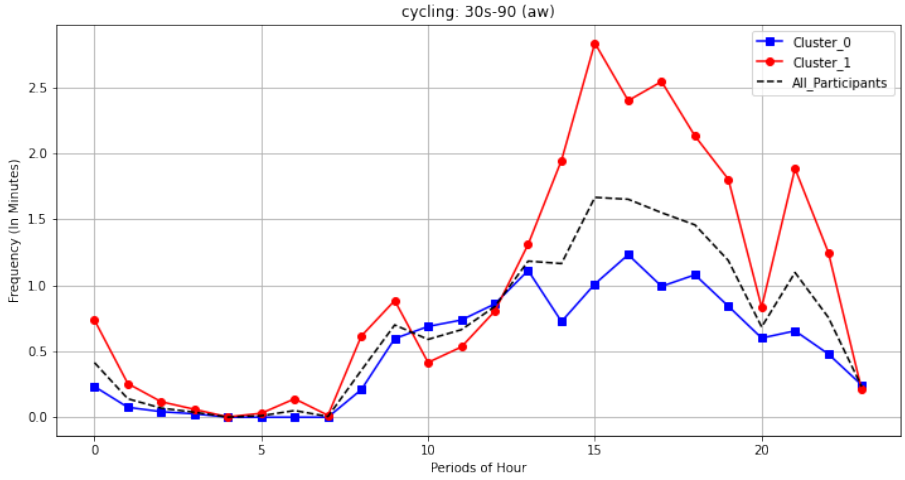
Furthermore, after comparing the produced plots for both unsupervised learning algorithms, it is clear that the respective normalized frequencies per hour, of active activities, in Figure 5.3 and Figure 5.6 are similar. Similarly, the respective normalized frequencies per hour, of sedentary activities, in Figure 5.4 and Figure 5.7 are similar. Therefore, we can infer that the two clusters obtained after $k$-means clustering and their respective participants' activity behaviours are similar to those obtained after hierarchical clustering.

### 5.3.3 Conclusion

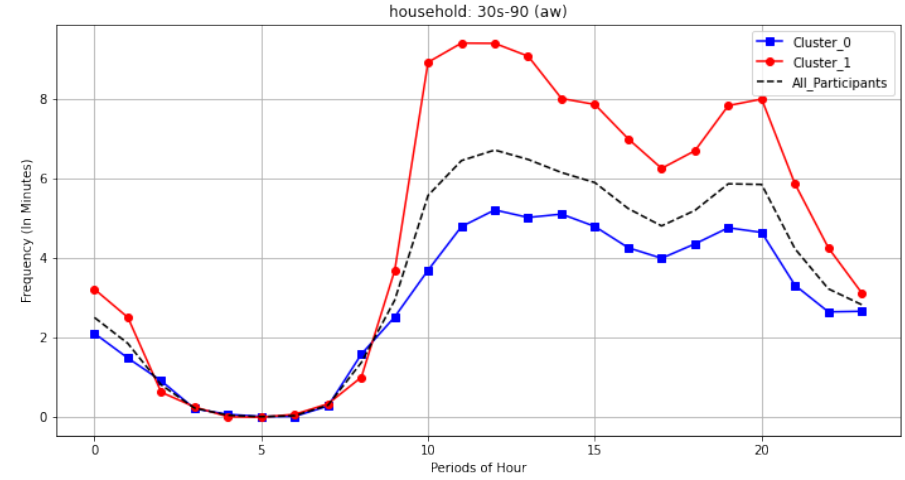
It is observed that, using the $k$-means clustering algorithm with $k = 2$, we obtained two clusters, out of which one group was male-dominant, and another was female-dominant. In this case, the clusters with a higher female percentage were generally more active than the cluster with a higher proportion of male participants, with an exception in the physical activity 'cycling'. From Table 5.10, we can see that at $k = 2$ we get slightly higher Silhouette scores than other values of $k$. From all the results we gathered, i.e. demographic results, activity data and their normalized frequency plots, we could identify the participants' active and non-active behaviours. The fact that the results obtained after $k$-means clustering and hierarchical clustering are similar with respect to their respective clusters can serve as an extra shred of evidence for our findings.




(a) Walking

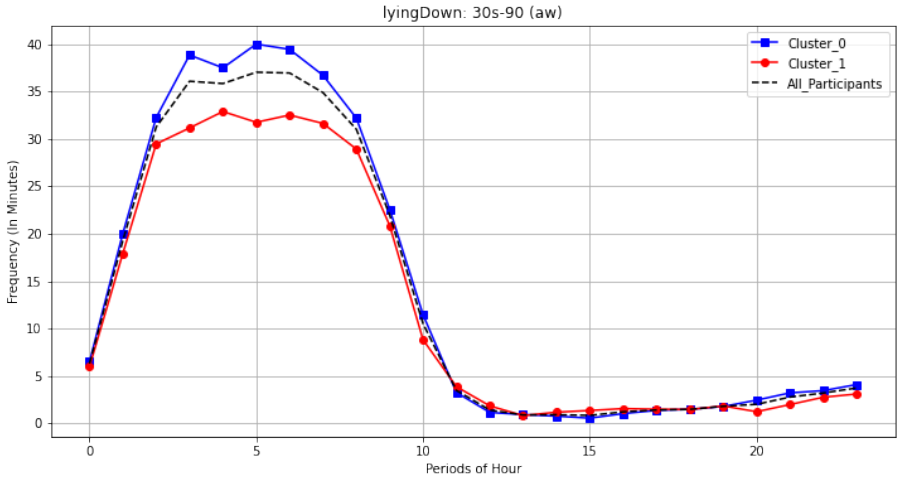



(b) Cycling

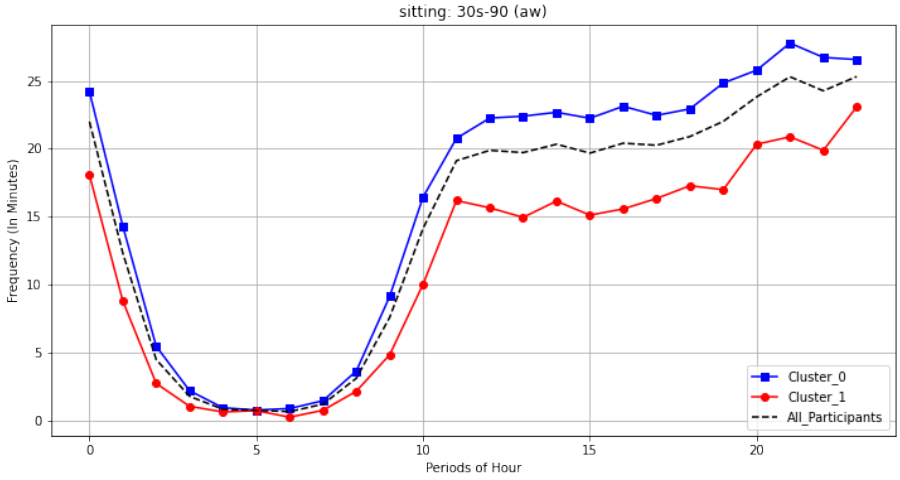



(c) Household

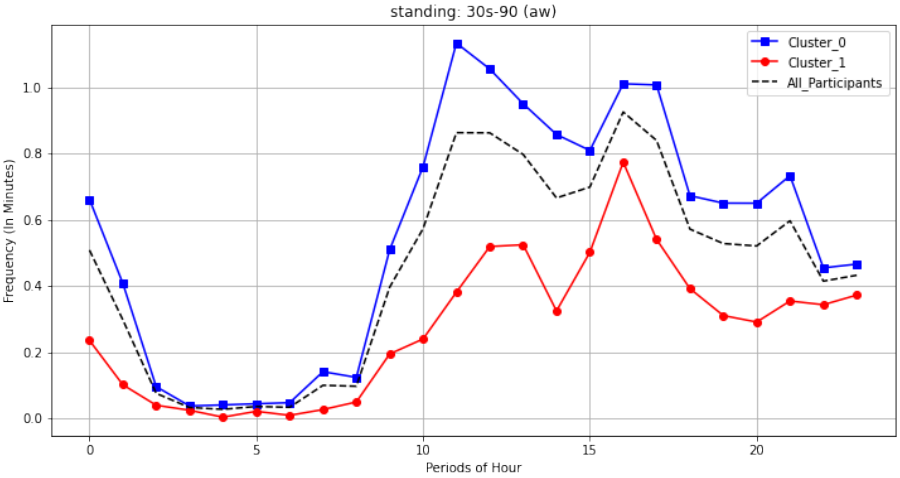
Figure 5.6: Normalized Frequency per Hour - Active Activities (Hierarchical Clustering)




(a) Lying Down




(b) Sitting




(c) Standing

Figure 5.7: Normalized Frequency per Hour - Inactive Activities (Hierarchical Clustering)

## 5.4 Comparing Average per Hour to Periods of Hours

To investigate the effect of dimensionality reduction techniques on our original feature dataset, i.e. Average per Hour with 144 features, we performed PCA on it, followed by $k$-means clustering algorithm, and then compared the cluster results to those of time-aggregated features, i.e. Periods of Hours with 24 features. The PCA results were already described in section 5.2.2. We explain the obtained results for the Period of Hours in the following sections.

### 5.4.1 $k$-means Clustering Results

We used the Silhouette Coefficient to check cluster quality before executing the $k$-means algorithm. In Table 5.15, we can see that for the value of 2, the Silhouette Coefficient is closer to 1 than those for other $k$ values. Therefore, we decided to use 2 as the optimal value of $k$ for the Periods of Hours dataset also.

| No. of Clusters | Silhouette Coefficient |
|---|---|
| **2** | **0.096** |
| 3 | 0.078 |
| 4 | 0.080 |
| 5 | 0.066 |
| 6 | 0.080 |
| 7 | 0.090 |
| 8 | 0.052 |
| 9 | 0.053 |

Table 5.15: Silhouette Score (Period of Hours)

The results that we obtained by performing $k$-means clustering with the value of $k$ as 2, are shown in tables 5.16 to 5.17.

| Cluster | C-0 | C-1 |
|---|---|---|
| **Size** | 63 | 51 |
| **G: Male** | 45 | 15 |
| **G: Female** | 18 | 36 |
| **Age** | $63.76 \pm 5.24$ | $61.30 \pm 5.88$ |
| **BMI** | $27.01 \pm 2.42$ | $26.69 \pm 2.56$ |

Table 5.16: Demographic Results (k $= 2$) (Period of Hours)

Table 5.16 displays the demographic results for the cluster we obtained. Cluster C-0 contains more participant, whereas cluster C-1 contain more female participants. The average age and BMI for cluster C-0 is slightly more than cluster C-1.

| Cluster | C-0 | C-1 | All |
|---|---|---|---|
| **Walking** | 25.51 | 16.56 | 21.51 |
| **Household** | 60.10 | 114.90 | 84.61 |
| **Cycling** | 10.93 | 23.37 | 16.50 |
| **Sitting** | 398.89 | 288.64 | 349.57 |
| **Lying Down** | 330.16 | 321.48 | 326.27 |
| **Standing** | 13.73 | 7.36 | 10.88 |

Table 5.17: Normalized Activity Frequency in Minutes (k = 2) (Period of Hours)

Table 5.17 shows the normalized frequency of the activities for both clusters i.e. C-0 and C-1. Cluster C-0 has more frequency of sedentary physical activities and also for walking than cluster C-1, whereas cluster C-1 which has more percentage of female participants is more active in household activities and cycling.

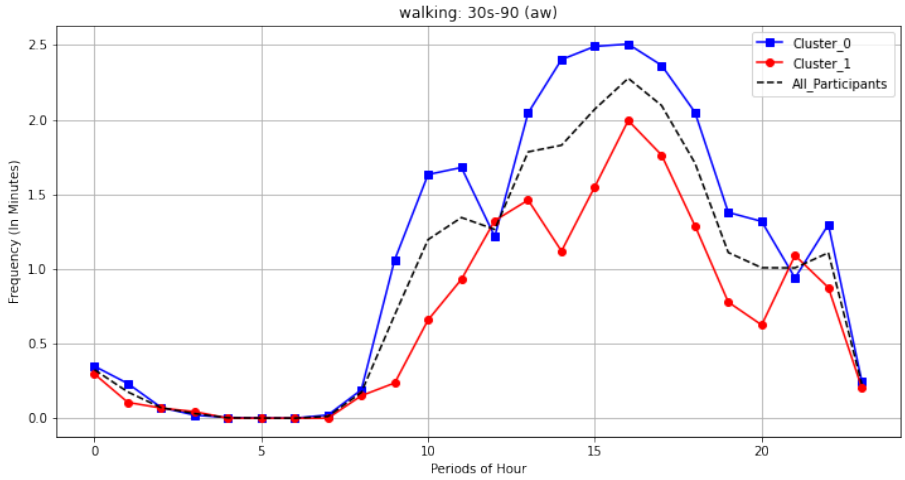
We further investigated visually the normalized frequency per hour for each of the activities in 24 hrs for the two obtained clusters in figs. 5.8 to 5.9. In Figure 5.8a for walking, we can notice that the participants of cluster C-0 are overall more active than those in cluster C-1. But in Figure 5.8b for cycling, cluster C-1 is more active with peaks in the evening. We can observe a similar pattern in household activities in Figure 5.8c for cluster C-1 with some peaks from afternoon to evening.
Participants of both clusters C-0 and C-1 seem to have similar patterns for lying down as shown in Figure 5.9a. In Figures 5.9b and 5.9c, the participants of cluster C-0 are seen to be spending more time sitting and standing than in those in cluster C-1.

### 5.4.2 Common Participants

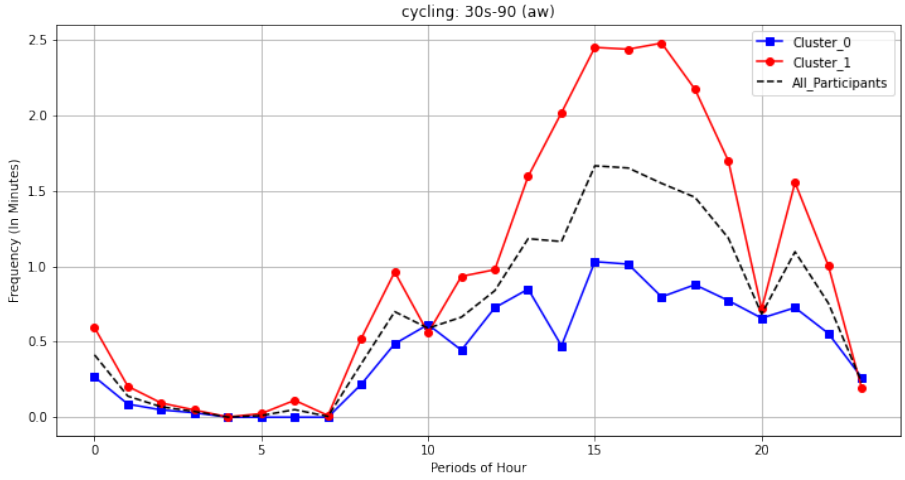
Table 5.18 depicts the common participants in respective clusters obtained using $k$-means clustering algorithm for Average per Hour after performing PCA and Periods of Hours feature-set. From this table, we can infer that the respective clusters obtained from both feature-sets are similar, having 62 common participants for cluster C-0 and 41 common participants for cluster C-1.

| | | **Period of Hours** | |
|---|---|---|---|
| | | **C-0 [63]** | **C-1 [51]** |
| **Average per Hour** | **C-0 [72]** | 62 | 10 |
| | **C-1 [42]** | 1 | 41 |

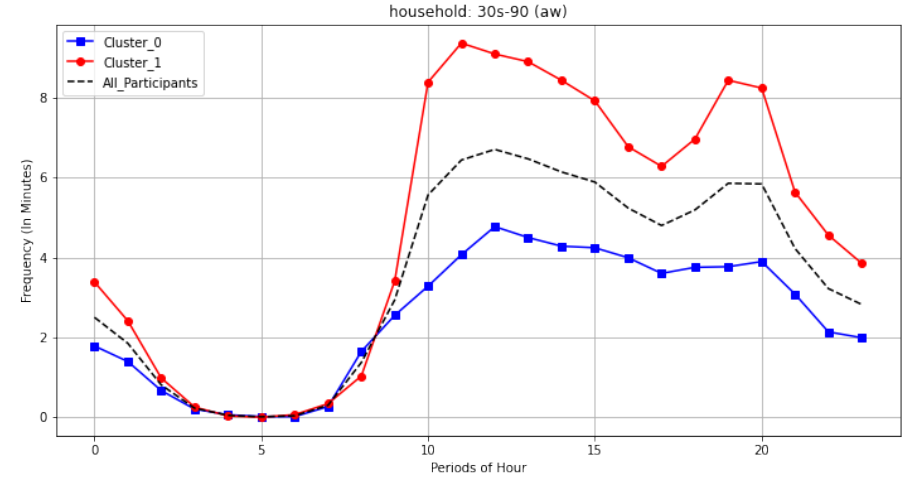
Table 5.18: Common Participants: Average per Hour and Periods of Hours




(a) Walking

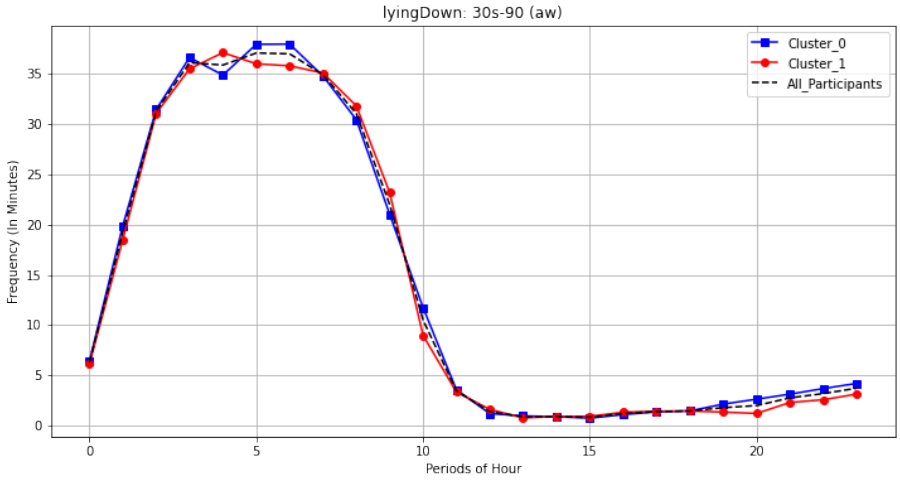



(b) Cycling

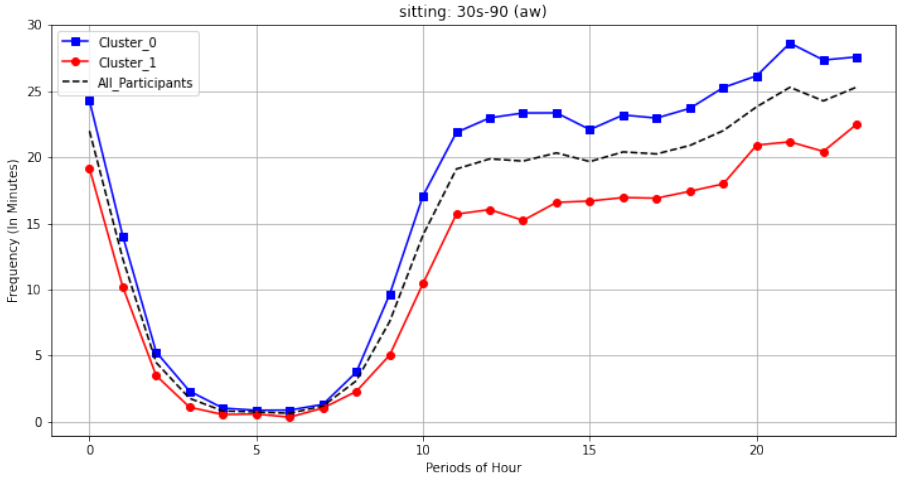



(c) Household

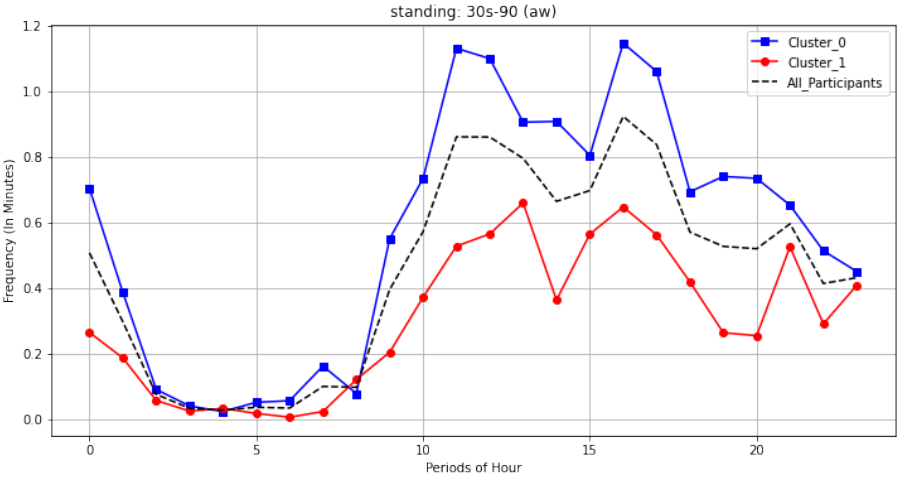
Figure 5.8: Normalized Frequency per Hour - Active Activities (k=2) (Period of Hours)




(a) Lying Down




(b) Sitting




(c) Standing

Figure 5.9: Normalized Frequency per Hour - Inactive Activities (k=2) (Period of Hours)

### 5.4.3 Conclusion

After using the $k$-means clustering algorithm with $k = 2$ for our constructed features, we obtained two clusters, where one cluster is male dominant and the other is female dominant. The cluster with more female participants was in general more active than the cluster having more male participants. We observed similar activity behaviour when we analysed the plots and demographic results of the obtained clusters using both dimensionality reduction technique. The common participants we found also suggests the same. Therefore, we can conclude that our time-aggregated feature-set also shows promise for generating precise results with respect to the popular dimensionality reduction technique, i.e. principal component analysis.
From the results and plots obtained, we can infer that female participants are likely more active than male participants, therefore to find better activity behavior among participants, we decided to check for clusters for both male and female participants separately. We further performed this analysis only for our one feature-set i.e. Average Per Hour with $144$ features for each participant.

## 5.5 Clustering after Division by Gender

Typically, females tend to perform more household activities than males. The results that we obtained for our sample population also suggest the same. Therefore, to identify the explicit activity behaviour of our participants, we chose to perform the $k$-means clustering algorithm after separating the population based on gender.

### 5.5.1 Male Participants

After extracting the male participants ($N = 60$) from our sample population, we check for the clusters using $k$-means clustering. Next, we verify the optimal value of $k$ by using the Silhouette coefficient for $k$-means clustering. We also perform Principal Component Analysis before checking for clusters. We can observe the variance explained by components in the line plot shown in Figure 5.10. To reduce the number of features from 144, we performed PCA that gave us 26 features while retaining 80% variance by the features.

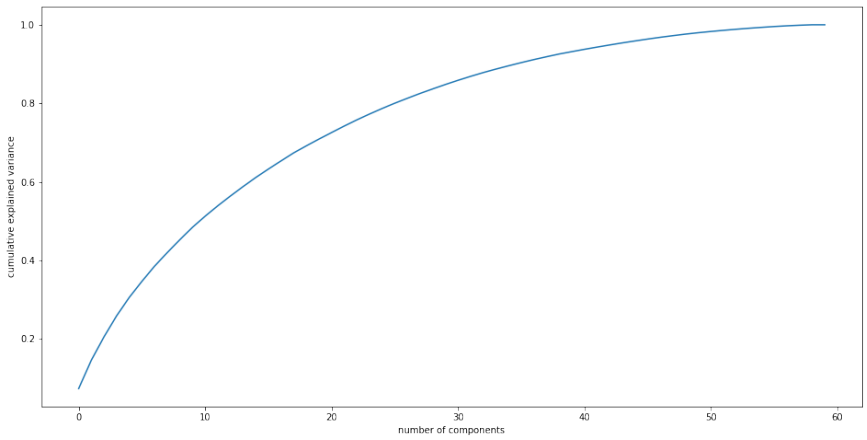



Figure 5.10: PCA Line Plot (Male Participants)

### $k$-means Clustering Results

We used Silhouette score to find the ideal value of $k$. The Silhouette coefficients we estimated can be seen in Table 5.19. We can notice than for $k = 2$, the Silhouette score is significantly higher than that for other values of $k$.

| No. of Clusters | Silhouette Coefficient |
|---|---|
| **2** | **0.180** |
| 3 | 0.041 |
| 4 | 0.045 |
| 5 | 0.049 |
| 6 | 0.048 |
| 7 | 0.032 |
| 8 | 0.012 |
| 9 | 0.040 |

Table 5.19: Silhouette Score (Male Participants)

We verified the optimal value of $k$ using the Silhouette score, and selected $k$ as 2. With $k = 2$, we got two clusters with sizes 1 and 59 respectively. We explored the cluster with size 1 and investigated the participant present in it i.e. Participant 62343. We can observe that this male participant has more frequency of walking than other male participants. For example, the walking frequency of this participant is 71 minutes, whereas the mean frequency of this activity for male participants is 23 minutes. We removed participant 62343 from our input data and performed the analysis again.

| No. of Clusters | Silhouette Coefficient |
|---|---|
| **2** | **0.050** |
| 3 | 0.040 |
| 4 | 0.037 |
| 5 | 0.028 |
| 6 | 0.038 |
| 7 | 0.030 |
| 8 | 0.029 |
| 9 | 0.024 |

Table 5.20: Silhouette Score (Male Participants) - After removing participant 62343

Again for this case, from Table 5.20, the best choice for the number of clusters is two. The demographic results and activity frequency for this setup is displayed in Table 5.21 and Table 5.22.

| Cluster | C-0 | C-1 |
|---|---|---|
| **Size** | 31 | 28 |
| **Age** | $64.78 \pm 5.15$ | $62.18 \pm 5.72$ |
| **BMI** | $26.62 \pm 2.33$ | $26.70 \pm 2.09$ |

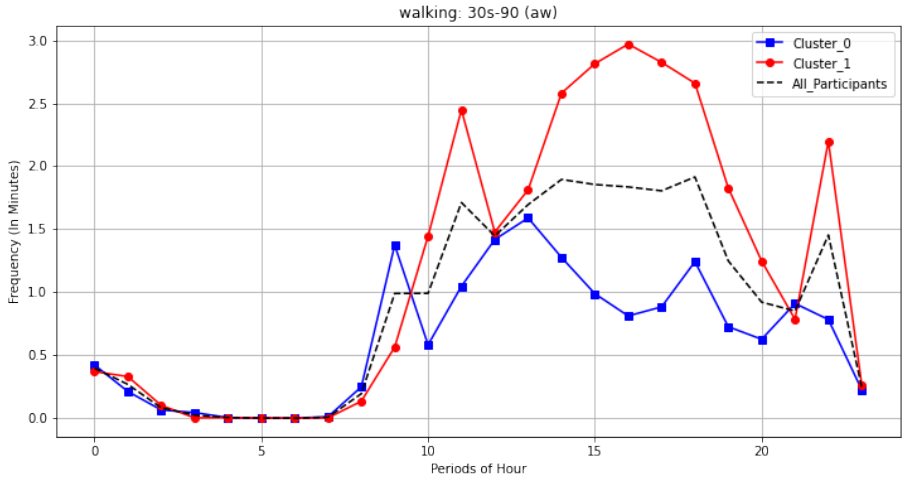
Table 5.21: Demographic Results (k = 2) (Male - After removing participant 62343)

In Table 5.21, we can notice that the clusters are almost equal in size. Cluster C-0 contains male participants having a higher mean age with approximately two years of a gap with cluster C-1. The mean BMI is also alike in both obtained clusters.

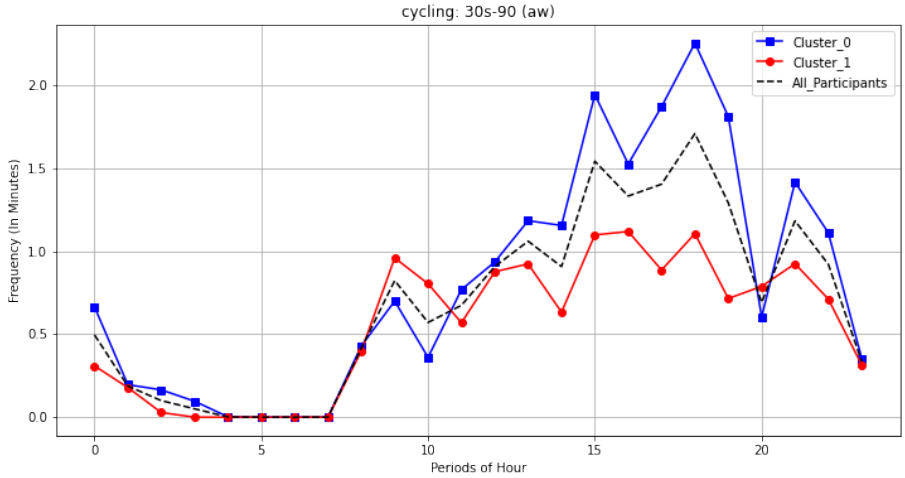
| Cluster | C-0 | C-1 | All |
|---|---|---|---|
| **Walking** | 15.49 | 28.86 | 21.84 |
| **Household** | 77.51 | 44.41 | 61.80 |
| **Cycling** | 19.54 | 13.34 | 16.60 |
| **Sitting** | 355.94 | 417.06 | 349.57 |
| **Lying Down** | 351.30 | 320.50 | 336.69 |
| **Standing** | 9.32 | 14.31 | 11.69 |

Table 5.22: Activity Frequency in Minutes (k = 2) (Male - After removing participant 62343)

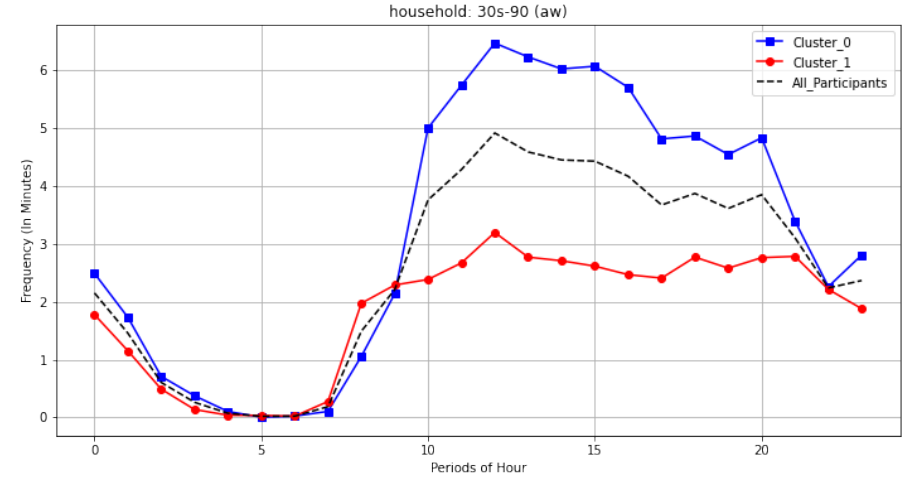
Table 5.22 displays the activity frequency in minutes for the obtained clusters. For active-level physical activities, cluster C-0 is more involved in cycling and household activities. In contrast, cluster C-1 is more active in walking. For sedentary-level activities, participants of cluster C-0 are lying down more, whereas cluster C-1 are sitting more. Therefore, we can infer that both clusters' non-active physical activities are balanced.
We visually investigated each activity's normalized frequency per hour for 24 hours, which can be seen in figs. 5.11 to 5.12.




(a) Walking

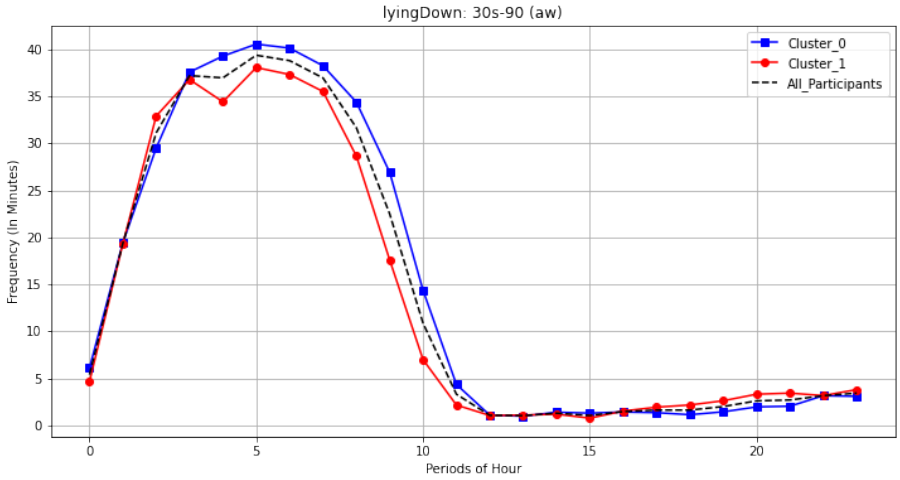



(b) Cycling

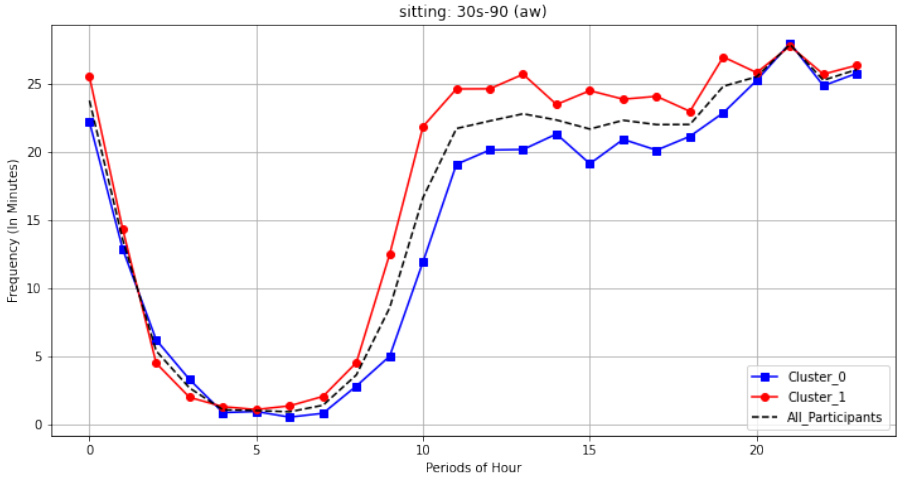



(c) Household

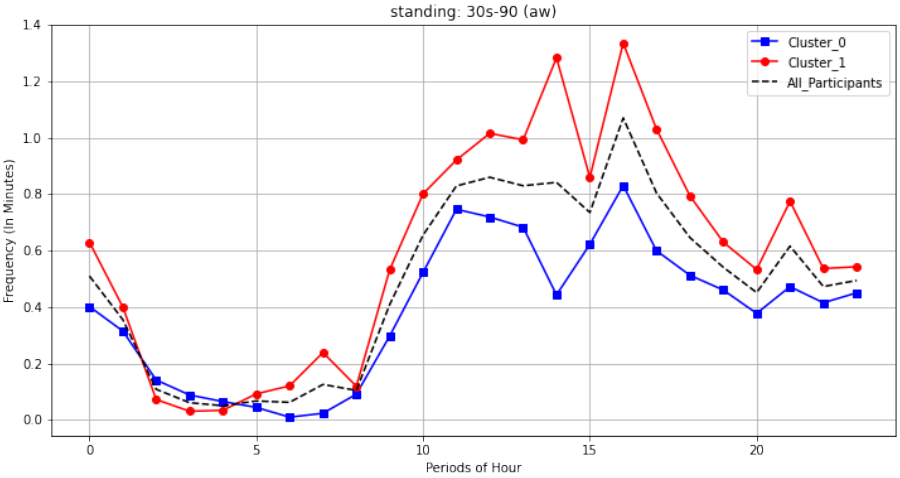
Figure 5.11: Normalized Frequency per Hour - Active Activities (k=2) (Male - After removing participant 62343)




(a) Lying Down




(b) Sitting




(c) Standing

Figure 5.12: Normalized Frequency per Hour - Inactive Activities (k=2) (Male - After removing participant 62343)

In the plot shown in Figure 5.11a for walking, cluster C-1 is more active from noon to night as compared to cluster C-0. For cycling, as shown in Figure 5.11b, cluster C-0 is more engaged, with the highest peak at 18:00. In plot 5.11c for household activities, cluster C-0 is showing more frequency of activity from 9:00 till night time. In sedentary activities, we can observe that clusters C-0 and C-1 are almost balanced, with cluster C-0 having more frequency for lying down 5.12a. In contrast, cluster C-1 is doing more sitting 5.12b throughout the day.

### 5.5.2 Female Participants

Similarly, after extracting the female participants ($N = 54$), we checked for clusters using the $k$-means clustering algorithm. Then, we verified the optimal value of $k$ using Silhouette Coefficient. We performed Principal Component Analysis and then checked for clusters. We can observe the variance explained by components in the line plot in Figure 5.10. To reduce the number of features from 144, we performed Principal Component Analysis which gave us 24 features after reduction while retaining 80% variance by the features in case of female participants.

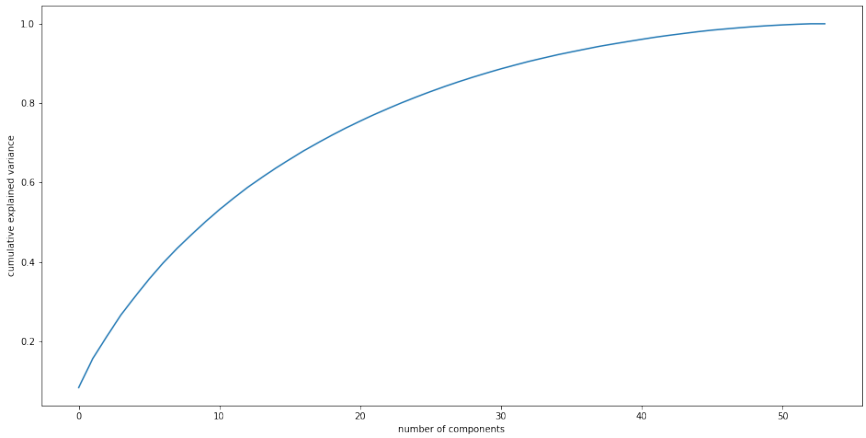



Figure 5.13: PCA Line Plot (Female Participants)

#### $k$-means Clustering Results

We also used Silhouette score in this case to find the ideal value of $k$ . The estimated Silhouette coefficients can be seen in Table 5.23. We can see that for $k = 2$, the Silhouette score is significantly higher than that for other values of $k$.

| No. of Clusters | Silhouette Coefficient |
|---|---|
| **2** | **0.240** |
| 3 | 0.050 |
| 4 | 0.055 |
| 5 | 0.037 |
| 6 | 0.035 |
| 7 | 0.032 |
| 8 | 0.037 |
| 9 | 0.042 |

Table 5.23: Silhouette Score (Female Participants)

We verified the optimal value of $k$ using the Silhouette Coefficient, and selected $k$ as 2. With $k = 2$, we got two clusters with sizes 53 and 1 respectively. We explored the cluster with size 1 and investigated the participant present in it i.e. Participant 70668. We can observe that this female participant has more frequency of active activities than other female participants and less sedentary activities. The walking frequency of this female participant is 47 minutes, whereas the mean frequency of this activity, for female participants, is 20 minutes. Also in case of cycling, this participant has 34 minutes of cycling frequency, whereas the mean frequency of this activity, for female participants, is 17 minutes. The household activity frequency for participant 70668 is also higher than the mean frequency of this activity for female participants. We removed participant 70668 from our input data and performed the analysis again.

| No. of Clusters | Silhouette Coefficient |
|---|---|
| **2** | **0.140** |
| 3 | 0.027 |
| 4 | 0.029 |
| 5 | 0.032 |
| 6 | 0.006 |
| 7 | 0.012 |
| 8 | 0.026 |
| 9 | 0.025 |

Table 5.24: Silhouette Score (Female Participants) - After removing participant 70668

Again, from Table 5.24, we can observe the best choice for the number of clusters is two. The demographic results and activity frequency for this setup is displayed in Table 5.25 and Table 5.26.

| Cluster | C-0 | C-1 |
|---|---|---|
| **Size** | 51 | 2 |
| **Age** | 61.80 $\pm$ 5.61 | 63.40 $\pm$ 7.63 |
| **BMI** | 27.05 $\pm$ 2.79 | 26.00 $\pm$ 0.71 |

Table 5.25: Demographic Results (k = 2) (Female - After removing participant 70668)

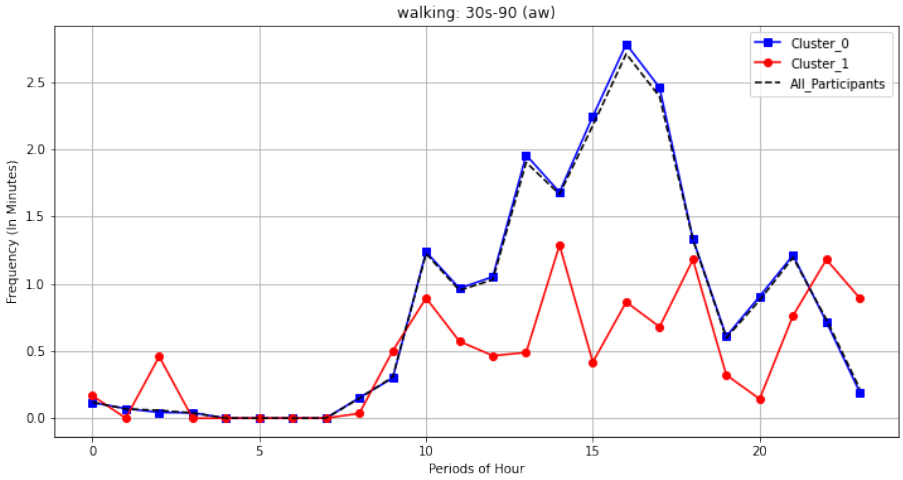
From Table 5.25, we notice that we are getting two clusters with sizes 51 and 2, respectively. Cluster C-0 has a lower mean age but slightly higher mean BMI than cluster C-1.

| Cluster | C-0 | C-1 | All |
|---|---|---|---|
| **Walking** | 20.05 | 11.29 | 19.72 |
| **Household** | 110.70 | 85.80 | 109.76 |
| **Cycling** | 15.43 | 36.83 | 16.24 |
| **Sitting** | 310.07 | 438.79 | 314.93 |
| **Lying Down** | 319.99 | 270.47 | 318.12 |
| **Standing** | 9.71 | 9.64 | 9.70 |

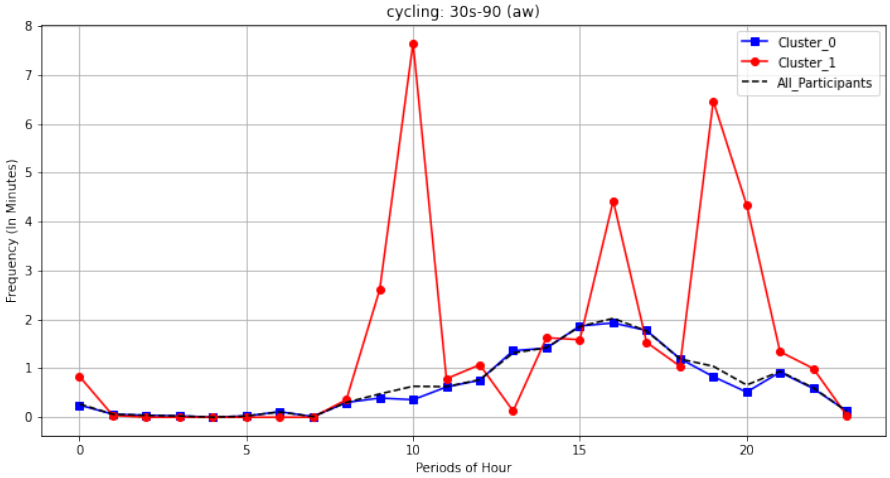
Table 5.26: Activity Frequency in Minutes (k = 2) (Female - After removing participant 70668)

Table 5.26 displays the activity frequency in minutes for the obtained clusters for female participants. Cluster C-0 is more active than cluster C-1 in the case of walking and household activities. Cluster C-1, which contains only two participants, is generally more involved in cycling as seen from the mean cycling frequency. For sedentary activities, participants of cluster C-1 lie down less but sit more as compared to those of cluster C-0. The frequency of standing activity remains almost similar for both clusters.

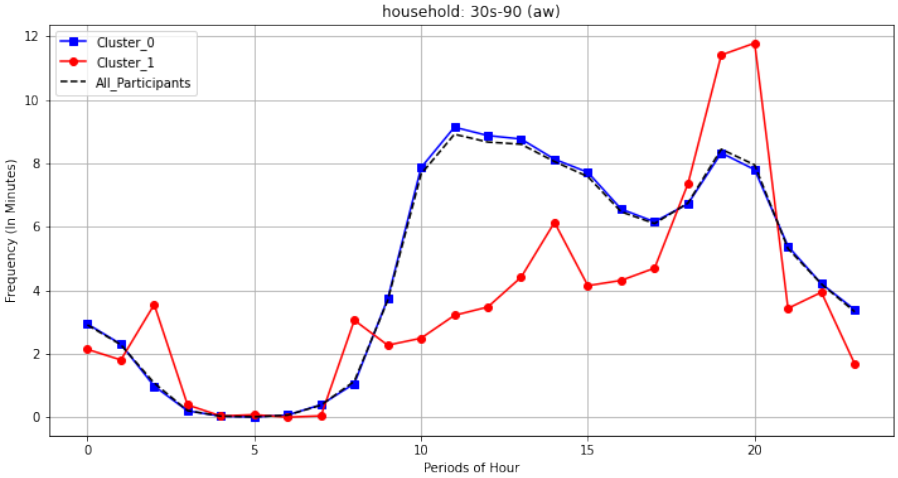
Again for female participants, we visually investigated each activity's normalized frequency per hour for 24 hours, which can be seen in figs. 5.14 to 5.15. In the case of walking as shown in Figure 5.14a, cluster C-0 is more active than cluster C-1 during the evening. In the plot for cycling as shown in Figure 5.14b, cluster C-1 is more engaged, with two high peaks one at 10:00 and one at 19:00. For household activities 5.14c, cluster C-0 is more engaged till late in the afternoon, while C-1 has a high peak, higher than C-0, of household activity in the evening (19-20:00). In sedentary activities, we can observe that cluster C-0 has a more smoothed frequency of lying down across the 24hr, as seen in Figure 5.15a. In contrast, participants in cluster C-1 are sitting more, as shown in Figure 5.15b, throughout the day. For standing, as shown in Figure 5.15c, both cluster seem to have a similar behaviour, though cluster C-1 has a high peak at around 20:00, which relate with to the household peaks.




(a) Walking

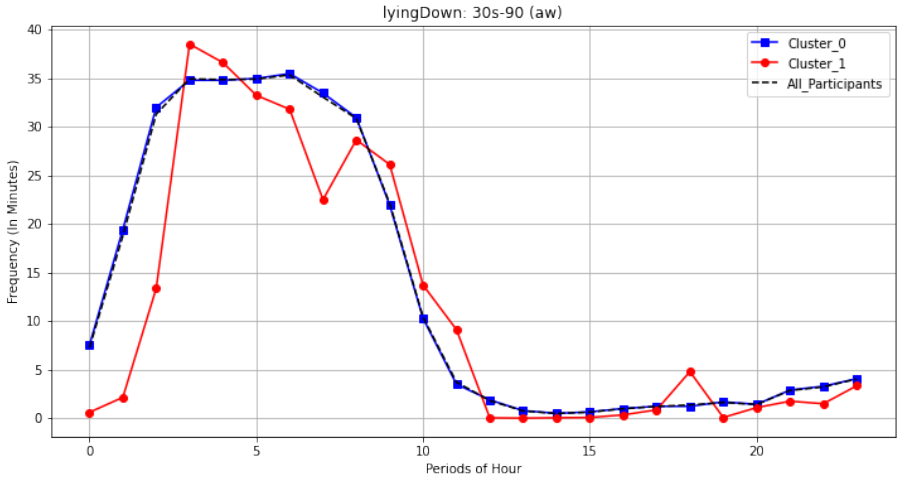



(b) Cycling

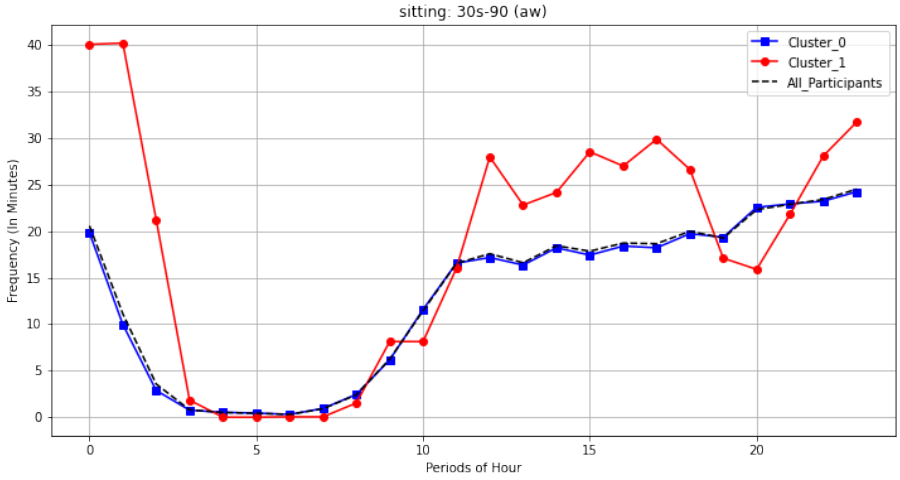



(c) Household

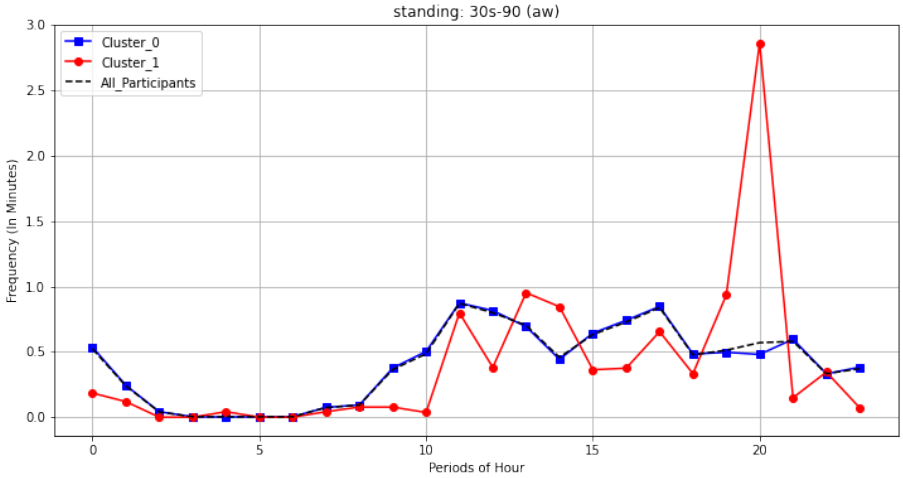
Figure 5.14: Normalized Frequency per Hour - Active Activities (k=2) (Female - After removing participant 70668)




(a) Lying Down




(b) Sitting




(c) Standing

Figure 5.15: Normalized Frequency per Hour - Inactive Activities (k=2) (Female - After removing participant 70668)

### 5.5.3 Conclusion

We separated our participants based on their gender, namely male and female. After division by gender we obtain 60 male participants. Afterwards, we performed PCA that gave us 26 features while retaining 80% variance by the features constructed. From the Silhouette scores of male participants, we chose $k = 2$ as our optimal choice of $k$ for $k$-means clustering algorithm. From the cluster results, we were able to find one male participant that was exhibiting significantly more physical activity behavior than other male participants in terms of 'walking'. Since this participant was not representative of the general population, we discarded it and then executed $k$-means clustering on the remaining male participants. From Table 5.21, we can infer that male participants can be classified into two clusters: C-0 with 31 participants and C-1 with 28 participants. Figure 5.11c depicts that C-0 represents male participants who are overall more active in household work as compared to those in C-1. Irrespective of the difference in the degree of activity of participants in both clusters, Figure 5.11c tells us that the respective highest peaks for both clusters are observed at 12:00. This implies that male participants are most likely engaged in household work at 12:00.

Similarly, we obtained 54 female participants. Then, we performed PCA that gave us 24 features while retaining 80% variance by the features constructed. From the Silhouette scores of female participants, we chose $k = 2$ as our optimal choice of $k$ for $k$-means clustering algorithm. From the cluster results, we were able to find one female participant that was showing significantly more physical activity behavior than other participants in terms of 'walking', 'cycling' and 'household activities'. Since this participant was not representative of the general population, we discarded it and then executed $k$-means clustering on the remaining female participants. From Table 5.25, we can infer that female participants can be classified into two clusters: C-0 with 51 participants and C-1 with just 2 participants. Figures 5.14a and 5.14c depicts that C-1 represents female participants who are overall less active in walking and household work as compared to those in C-0, which is also confirmed by the higher sitting time (see Figure 5.15a). However, it seems that C-1 cluster is really active in the evening, see Figures 5.14b, 5.14c and 5.15c.
From this we can infer that the respective highest peaks for both clusters are observed at 11:00 and 20:00. Since the number of participants in cluster C-0 is significantly higher than that in cluster C-1, we can safely imply that female participants are most likely engaged in household work at 11:00 i.e., highest peak for C-0.

From Figures 5.11c and 5.14c, it is clear that the highest peak of the cluster corresponding to active male participants is much lower even than the highest peak of the cluster corresponding to non-active female participants, in terms of household activity. This simply means that even non-active female participants are engaging in household activity significantly more that most male participants. These results are similar to those seen in Table 5.11 and Figure 5.3c, wherein the male-dominated cluster C-0 has highest peak significantly lower than the highest peak of

female-dominated cluster C-1. This confirms our previous conclusion that in terms of household activity, female participants are overall more active than male participants.

## 5.6 Discussion

We used the 30-seconds aggregation with a 90% threshold for our experiments. To verify the optimal choice for $k$ in $k$-means clustering algorithm, we used Silhouette score. With the initial run of our feature dataset, we wanted to remove data-related problems using the $k$-means clustering algorithm. We could detect various outliers for some participants in our dataset from the cluster results. We removed these anomalies to get clearer and distinct clusters.

After removing the anomalies, we used the clean data to check the outcome of PCA on our high-dimensional feature dataset. After comparing the cluster results of the original feature dataset with 144 features to the reduced feature dataset with 37 features performed by PCA, we observed that we were getting more reasonable clusters. After this analysis, we decided to use the reduced feature dataset, i.e. 37 features that retained 80% variance.

Afterwards, we profiled the participants from the clean data using $k$-means and hierarchical clustering algorithms. The results of both methods are similar, with one male dominant group and another female dominant. We can validate that the cluster results obtained by $k$-means clustering algorithms are appropriate with relatively identical cluster results obtained by hierarchical clustering.

Next, we compared dimensionality reduction techniques on our hourly generated feature dataset. We compared the hourly feature dataset with PCA, i.e. Average per Hour with 37 features that retain 80% variance of the dataset, to our time aggregated features, i.e. Period of Hours with 24 features per participant. Later, we checked for common participants for both techniques in their respective obtained clusters and found that the participants were clustered similarly. Therefore, we can validate that our reduced time aggregated features also provided similar statistics and results compared to PCA and can be used for dimensionality reduction.

Furthermore, to identify our participants' explicit activity behaviour, we chose to perform the $k$-means clustering algorithm after separating the population based on gender. After clustering for both male and female participants separately, we obtained clusters that exhibited clear separation of activity behaviour for certain activities such as walking and household work, both in males as well as females. Based on this, we could identify active and non-active groups within male and female clusters. We could also see how female clusters were overall more active than male clusters.

We were able to identify active and non-active profiles with few irregular profiles among the

114 older adults having both ankle and wrist accelerometer data. In future, it might also be beneficial to merge the Standing class with the Household activities class because standing activities alone did not provide adequate or acceptable information from the obtained results.

# Chapter 6

# Conclusion

This research aimed to identify different physical activity behaviours by exploring activity data of older individuals with the use of unsupervised methods. The accelerometer data was provided by GOTO study, which had in total of 164 older adults whose daily activities were measured using wearable sensors over the span of two weeks, one before and one after a lifestyle intervention. We aggregated the obtained data at various scales i.e. we combined different window lengths and aggregation methods to obtain the best possible combination. From our analysis performed, we decided to use the 30-seconds aggregation with a 90% threshold for our experiments. After performing data pre-processing such as comparing different accelerometer data i.e. ankle and wrist accelerometers and filtering out days using area plots, we removed some of the participants and thus obtained clean data for 114 participants to be used in our methods. We constructed features from this processed data for our algorithms. For every participants, we extracted frequency of 30 second aggregation for each of the activities per hour. We focus our work on physical activity profiling by using the $k$-means clustering algorithm.

From the experiments performed and the results obtained in our initial run, we were able to detect outliers using the $k$-means clustering algorithm. These outliers were mainly participant activity logs for certain days where the activities were classified incorrectly, possibly due to lack of device wearing time. For example, if the participant does not wear the device on a particular day and the device is kept stationary, the physical activity being performed would be classified as 'standing', which is incorrect. Similarly, for certain particular days the frequencies of some activities like 'sitting' and 'lying down' were also abnormal. Therefore, we chose to discard the data for these particular days. After filtering out the outliers, we performed the experiments again using $k$-means clustering. We primarily used the Silhouette score to check the quality of clusters for the $k$-means clustering algorithm. Considering the Silhouette score, the best obtained clusters are the one with 2 as the optimal value of $k$. We were able to determine clear activity behaviours from the results of clustering, and normalized frequency plots after performing the analysis for both male and female participants separately. We were able to successfully identify the active and non-active groups along with some irregular participants. After clustering for male and female participants separately, we were also able to infer that

female participants are overall more active than the male participants of this study, particularly in activities such as household work. This is not surprising since generally, females are commonly engaged in significantly higher amount of household work as compared to males. Hence, we can also conclude that the GOTO study data is representative of the general elder population. Since irregularities are precisely identified and active and non-active clusters obtained are well differentiated, we can conclude that wearable sensor data is successful and can be relied upon in fairly profiling participants based on their activity behaviours.

We explored different ways to deal with the high dimensionality and high correlations of our generated feature-dataset. We applied Principal Component Analysis (PCA) to reduce the number of features constructed. We were able to produce clearer and balanced cluster results using PCA with respect to the original feature-dataset (144 features). Since, we lost 20% variance after performing PCA, we were getting different cluster results but clearer. Another approach we used to tackle this problem was to merge the earlier extracted 144 features together with respect to different period of time into 24 features for every participant. The cluster results we obtained with this feature-set were similar to the cluster results obtained using 144 features after performing PCA to a certain extent. Since $k$-means clustering algorithm is highly sensitive to changes in data points; the respective similar clusters obtained for the 144 feature dataset after PCA and for the time-aggregated feature-set is indicative of the fact that feature construction has been done accurately in both cases. We can also conclude that dimensionality reduction techniques such as PCA can be precisely applied on wearable sensor data, and it is also possible to construct features manually using time-aggregation for wearable sensor data such that minimum variance is lost. Feature construction is a strenuous task but if executed properly, it can positively help in mapping wearable sensor data to actual physical activities performed by an individual.

## 6.1 Future Scope

In this project, we were dealing with the data in the form a sequence of data points indexed in time order i.e. time-series data. We used Euclidean distance as the similarity measure for our clustering algorithms to obtain similarity between two time-series. Therefore, it would be interesting to see the application and results of clustering algorithms that use similarity measures such as Dynamic Time Warping (DTW). Furthermore, investigating some intuitive and non-intuitive correlations between activity features and patterns along with health parameters, for example, blood-cholesterol levels, heart-rate etc. can help us understand more about healthy ageing among other things in older adults of our society. Since we also possess data of participants after a lifestyle intervention, we can use our current results to assess the effect of lifestyle intervention on individuals in terms of the above-mentioned health parameters.

Furthermore, it would also be worthwhile to carry out an in-depth investigation of the obtained clusters, particularly those of non-active participants, and analyse their health parameters in detail in order to produce suggestive measures to improve their health conditions if needed. We can also study the obtained clusters further and associate the physical activity behaviour to serious health-risks or diseases.

A major limitation of our work is that we were restricted in terms of the choice of clustering algorithms, due to the less number of participants in our data. A higher number of participants in the data would have enabled us to explore other complex clustering algorithms.

# Bibliography


[1] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. "Time-series clustering – A decade review". en. In: *Information Systems* 53 (Oct. 2015), pp. 16–38. DOI: 10.1016/j.is.2015.04.007.

[2] Farhad Ahamed, Seyed Shahrestani, and Hon Cheung. "Internet of Things and Machine Learning for Healthy Ageing: Identifying the Early Signs of Dementia". eng. In: *Sensors (Basel, Switzerland)* 20.21 (Oct. 2020), E6031. DOI: 10.3390/s20216031.

[3] Navid Amini et al. "Accelerometer-based on-body sensor localization for health and medical monitoring applications". en. In: *Pervasive and Mobile Computing* 7.6 (Dec. 2011), pp. 746–760. DOI: 10.1016/j.pmcj.2011.09.002.

[4] Soumya Banerjee. "A class-contrastive human-interpretable machine learning approach to predict mortality in severe mental illness". en. In: *npj Schizophrenia* (2021), p. 13.

[5] Purnima Bholowalia. "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN". en. In: *International Journal of Computer Applications* 105.9 (), p. 8.

[6] Tim Bock. *What is a Dendrogram?* en-US. Mar. 2018.

[7] Francisco Félix Caballero et al. "Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the ATHLOS project". en. In: *Scientific Reports* 7.1 (Apr. 2017), p. 43955. DOI: 10.1038/srep43955.

[8] Lucien Marie Le Cam and Jerzy Neyman. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification*. en. Google-Books-ID: IC4Ku_7dBFUC. University of California Press, 1967.

[9] Min Chen et al. "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities". In: *IEEE Access* 5 (2017), pp. 8869–8879. DOI: 10.1109/ACCESS.2017.2694446.

[10] Suresh Dara et al. "Machine Learning in Drug Discovery: A Review". eng. In: *Artificial Intelligence Review* 55.3 (2022), pp. 1947–1999. DOI: 10.1007/s10462-021-10058-4.

[11] Gautam Das et al. "Rule Discovery from Time Series". en. In: (), p. 7.

[12] Chris Ding and Xiaofeng He. "Cluster Merging and Splitting in Hierarchical Clustering Algorithms". In: *Proc. IEEE Int'l Conf. Data Mining*. 2002, pp. 139–146.

[13] Carolyn Domingo, Solomon See, and Roberto Legaspi. "Unsupervised Habitual Activity Detection in Accelerometer Data". en. In: *Mechatronics and Machine Vision in Practice 3*. Ed. by John Billingsley and Peter Brett. Cham: Springer International Publishing, 2018, pp. 253–272. DOI: 10.1007/978-3-319-76947-9_19.

[14] Christos Faloutsos, M Ranganathan, and Yannis Manolopoulos. "Fast Subsequence Matching in Time-Series Databases". en. In: (), p. 11.

[15] Sebastian Försch et al. "Artificial Intelligence in Pathology". In: *Deutsches Ärzteblatt International* 118.12 (Mar. 2021), pp. 199–204. DOI: 10.3238/arztebl.m2021.0011.

[16] Paul A. Gore. "11 - Cluster Analysis". In: *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. Ed. by Howard E. A. Tinsley and Steven D. Brown. San Diego: Academic Press, 2000, pp. 297–321. DOI: https://doi.org/10.1016/B978-012691360-6/50012-4.

[17] H. Hotelling. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441. DOI: 10.1037/h0071325.

[18] A. K. Jain, M. N. Murty, and P. J. Flynn. "Data clustering: a review". en. In: *ACM Computing Surveys* 31.3 (Sept. 1999), pp. 264–323. DOI: 10.1145/331499.331504.

[19] Minsu Jang et al. "Dynamic Time Warping-Based K-Means Clustering for Accelerometer-Based Handwriting Recognition". en. In: *Developing Concepts in Applied Intelligence*. Ed. by Kishan G. Mehrotra et al. Studies in Computational Intelligence. Berlin, Heidelberg: Springer, 2011, pp. 21–26. DOI: 10.1007/978-3-642-21332-8_3.

[20] Jiaxin Jin et al. "Activity Pattern Mining for Healthcare". In: *IEEE Access* 8 (2020), pp. 56730–56738. DOI: 10.1109/ACCESS.2020.2981670.

[21] Petra Jones et al. "Towards a Portable Model to Discriminate Activity Clusters from Accelerometer Data". en. In: *Sensors* 19.20 (Oct. 2019), p. 4504. DOI: 10.3390/s19204504.

[22] R. Jothi. "Clustering Time-Series Data Generated by Smart Devices for Human Activity Recognition". en. In: *Intelligent Systems Design and Applications*. Ed. by Ajith Abraham et al. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020, pp. 708–716. DOI: 10.1007/978-3-030-16657-1_66.

[23] Dafne van Kuppevelt et al. "Segmenting accelerometer data from daily life with unsupervised machine learning". en. In: *PLOS ONE* 14.1 (Jan. 2019). Ed. by Maciej S. Buchowski, e0208692. DOI: 10.1371/journal.pone.0208692.

[24] King-Ip Lin, Harpreet S Sawhney, and Kyuseok Shim. "Fast Similarity Search in the Presence of Noise, Scaling, and Translation in Time-Series Databases". en. In: (), p. 12.

[25] Qin Liu et al. "Machine learning models for predicting critical illness risk in hospitalized patients with COVID-19 pneumonia". In: *Journal of Thoracic Disease* 13.2 (Feb. 2021), pp. 1215–1229. DOI: 10.21037/jtd-20-2580.

[26] S. Lloyd. "Least squares quantization in PCM". en. In: *IEEE Transactions on Information Theory* 28.2 (Mar. 1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[27] Ezequiel López-Rubio, Esteban J. Palomo, and Francisco Ortega Zamorano. *Unsupervised learning by cluster quality optimization*. 2018. DOI: 10.1016/j.ins.2018.01.007.

[28] Ariel B. Neikrug et al. "Characterizing Behavioral Activity Rhythms in Older Adults Using Actigraphy". en. In: *Sensors* 20.2 (Jan. 2020), p. 549. DOI: 10.3390/s20020549.

[29] Stylianos Paraschiakos et al. "Activity recognition using wearable sensors for tracking the elderly". en. In: *User Modeling and User-Adapted Interaction* 30.3 (July 2020), pp. 567–605. DOI: 10.1007/s11257-020-09268-2.

[30] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (Nov. 1901), pp. 559–572. DOI: 10.1080/14786440109462720.

[31] Subhanik Purkayastha et al. "Machine Learning-Based Prediction of COVID-19 Severity and Progression to Critical Illness Using CT Imaging and Clinical Data". In: *Korean Journal of Radiology* 22.7 (July 2021), pp. 1213–1224. DOI: 10.3348/kjr.2020.1104.

[32] M Reddy, Vivekananda Makara, and Satish R U V N. "Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering". In: 5 (Oct. 2017).

[33] Ondine van de Rest et al. "Metabolic effects of a 13-weeks lifestyle intervention in older adults: The Growing Old Together Study". en. In: *Aging* 8.1 (Jan. 2016), pp. 111–124. DOI: 10.18632/aging.100877.

[34] Neguine Rezaii, Elaine Walker, and Phillip Wolff. "A machine learning approach to predicting psychosis using semantic density and latent content analysis". en. In: *npj Schizophrenia* 5.1 (Dec. 2019), p. 9. DOI: 10.1038/s41537-019-0077-9.

[35] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. DOI: https://doi.org/10.1016/0377-0427(87)90125-7.

[36] Maurice Roux. "A comparative study of divisive hierarchical clustering algorithms". In: *Journal of Classification* 35 (June 2015). DOI: 10.1007/s00357-018-9259-9.

[37] Matthias Schonlau. "Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams". en. In: *Computational Statistics* 19.1 (Feb. 2004), pp. 95–111. DOI: 10.1007/BF02915278.

[38] *scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentation*.

[39] Ketan Rajshekhar Shahapure and Charles Nicholas. "Cluster Quality Analysis Using Silhouette Score". In: *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. Oct. 2020, pp. 747–748. DOI: 10.1109/DSAA49011.2020.00096.

[40] Santosh Shinde and P.R. Rajeswari. "Intelligent health risk prediction systems using machine learning: A review". In: *International Journal of Engineering and Technology(UAE)* 7 (June 2018), pp. 1019–1023. DOI: 10.14419/ijet.v7i3.12654.

[41] *sklearn.decomposition.PCA*. en.

[42] P. H. A. Sneath and R. R. Sokal. "Numerical taxonomy. The principles and practice of numerical classification." English. In: *Numerical taxonomy. The principles and practice of numerical classification.* (1973).

[43] *The elbow method - Statistics for Machine Learning [Book]*. en. ISBN: 9781788295758.

[44] "Multivariate Regression Models". en. In: *Applied Multivariate Analysis*. Ed. by Neil H. Timm. Springer Texts in Statistics. New York, NY: Springer, 2002, pp. 185–309. DOI: 10.1007/978-0-387-22771-9_4.

[45] Rieke Trumpf et al. "Quantifying Habitual Physical Activity and Sedentariness in Older Adults—Different Outcomes of Two Simultaneously Body-Worn Motion Sensor Approaches and a Self-Estimation". en. In: *Sensors* 20.7 (Mar. 2020), p. 1877. DOI: 10.3390/s20071877.

[46] Matthew Willetts et al. "Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants". en. In: *Scientific Reports* 8.1 (Dec. 2018), p. 7961. DOI: 10.1038/s41598-018-26174-1.

[47] Shuo Xu et al. "Reviews on Determining the Number of Clusters". en. In: *Applied Mathematics & Information Sciences* 10.4 (July 2016), pp. 1493–1512. DOI: 10.18576/amis/100428.

[48] Marie Lisandra Zepeda-Mendoza and Osbaldo Resendis-Antonio. "Hierarchical Agglomerative Clustering". en. In: *Encyclopedia of Systems Biology*. Ed. by Werner Dubitzky et al. New York, NY: Springer, 2013, pp. 886–887. DOI: 10.1007/978-1-4419-9863-7_1371.

# Appendices

# Appendix A

# Additional Analysis

## A.1 Initial Run

### A.1.1 Average per Hour - Principal Component Analysis (42 Features)

We performed PCA on the Average per Hour i.e. 144 extracted features. After applying PCA, we obtained 42 features that were retaining 80% variance. Like before, we applied $k$-means with $k = 4$. The demographic results and the normalized frequency of each activity of the clusters that we obtained can be seen in Table A.1 and Table A.2.

| Cluster | C-0 | C-1 | C-2 | C-3 |
|---|---|---|---|---|
| **Size** | 5 | 54 | 42 | 12 |
| **G: Female** | 4 | 22 | 20 | 8 |
| **G: Male** | 1 | 32 | 22 | 4 |
| **Age** | 66.22 $\pm$ 2.20 | 62.36 $\pm$ 5.76 | 62.74 $\pm$ 5.62 | 63.26 $\pm$ 5.27 |
| **BMI** | 26.43 $\pm$ 2.17 | 26.57 $\pm$ 2.52 | 27.30 $\pm$ 2.32 | 26.61 $\pm$ 2.91 |

Table A.1: Average per Hour - Principal Component Analysis (42 Features)

| Cluster | C-0 | C-1 | C-2 | C-3 |
|---|---|---|---|---|
| **Walking** | 25.61 | 20.29 | 19.03 | 18.15 |
| **Household** | 188.26 | 71.22 | 78.41 | 102.69 |
| **Cycling** | 5.05 | 20.08 | 9.37 | 14.02 |
| **Sitting** | 319.18 | 368.85 | 309.55 | 298.14 |
| **Lying Down** | 348.48 | 333.04 | 462.75 | 289.32 |
| **Standing** | 17.22 | 14.53 | 12.21 | 104.69 |

Table A.2: Average per Hour - Principal Component Analysis (42 Features)

From the observation of the results, we notice the we have obtained almost similar cluster results after performing PCA for dimensionality reduction on the 144 features constructed.

## A.2 Profiling the participants using Average Per Hour

### A.2.1 Hierarchical Clustering Results

Table 5.10 shows five as the second best choice for the optimal number of clusters based on silhouette score. From the Figure A.1 of dendrogram, we can observe the obtained five clusters in different colours using hierarchical clustering algorithm.

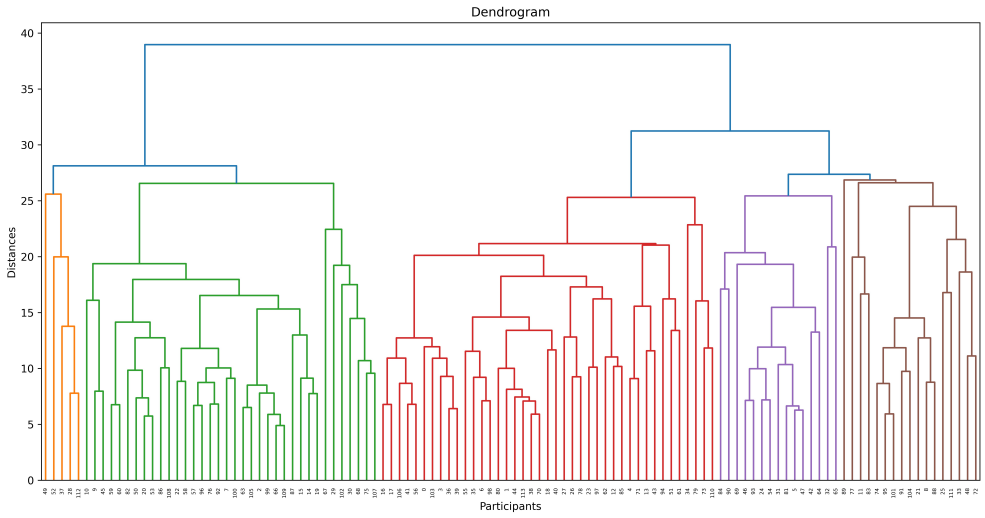



Figure A.1: Dendrogram (Average per Hour)

After applying hierarchical clustering for 5 clusters, we get the following demographic results and normalized activity frequency.

| Cluster | C-0 | C-1 | C-2 | C-3 | C-4 |
|---|---|---|---|---|---|
| **Size** | 17 | 36 | 41 | 5 | 15 |
| **G: Male** | 11 | 10 | 26 | - | 13 |
| **G: Female** | 6 | 26 | 15 | 5 | 2 |
| **Age** | 65.19 ± 1.77 | 60.17 ± 6.00 | 64.39 ± 5.25 | 61.66 ± 4.21 | 61.36 ± 6.74 |
| **BMI** | 26.61 ± 1.52 | 27.27 ± 2.87 | 26.97 ± 2.67 | 25.42 ± 2.25 | 26.39 ± 1.80 |

Table A.3: Demographic Results (k=5)

From the demographic summary given in the Table A.3, we can say that the number of participants in cluster C-0 and cluster C-4 are almost similar, cluster C-1 and cluster C-2 have most of the participants, leaving cluster C-3 with 5 participants. The cluster C-1 has the most number of female participants compared to all other clusters followed by clusters C-2, C-0, C-3 and finally C-4 with the least number of female participants. Cluster C-3 only has female participants. From the mean age of the clusters, we can see that cluster C-1 and cluster C-3 have a lower mean age than other clusters. The reason behind this is the number of female participants in the clusters. The clusters with higher percentage of male participants have mean age of around 65, but cluster C-4 is an exception with

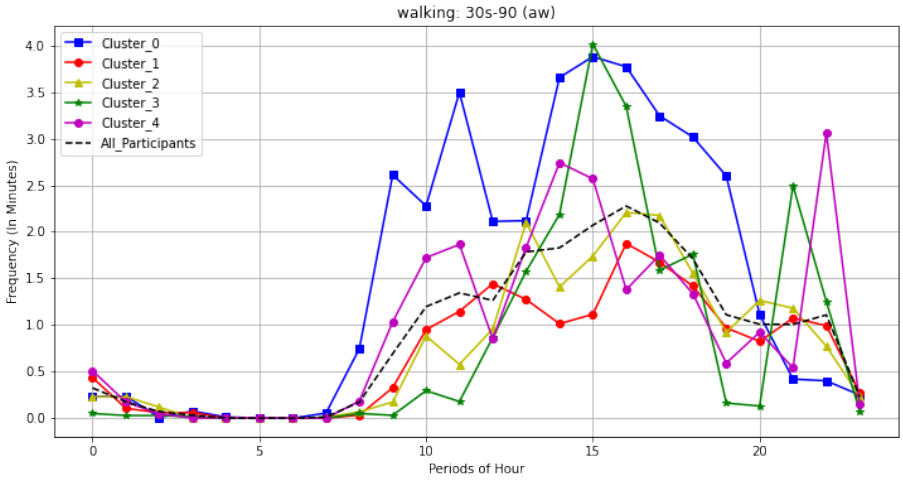
mean age of 61. The members of cluster C-0 have a mean age of 65 with least standard deviation among the clusters. The mean BMI of all the clusters are almost similar.

| Cluster | C-0 | C-1 | C-2 | C-3 | C-4 | All |
|---|---|---|---|---|---|---|
| **Walking** | 36.32 | 17.05 | 18.80 | 20.12 | 23.26 | 21.51 |
| **Household** | 66.22 | 101.12 | 76.23 | 201.50 | 49.81 | 84.61 |
| **Cycling** | 14.88 | 24.40 | 12.21 | 19.03 | 10.24 | 16.50 |
| **Sitting** | 343.98 | 275.55 | 381.33 | 296.22 | 464.50 | 349.57 |
| **Lying Down** | 355.00 | 296.55 | 354.85 | 295.31 | 297.28 | 326.27 |
| **Standing** | 11.65 | 6.47 | 13.63 | 7.23 | 14.29 | 10.88 |

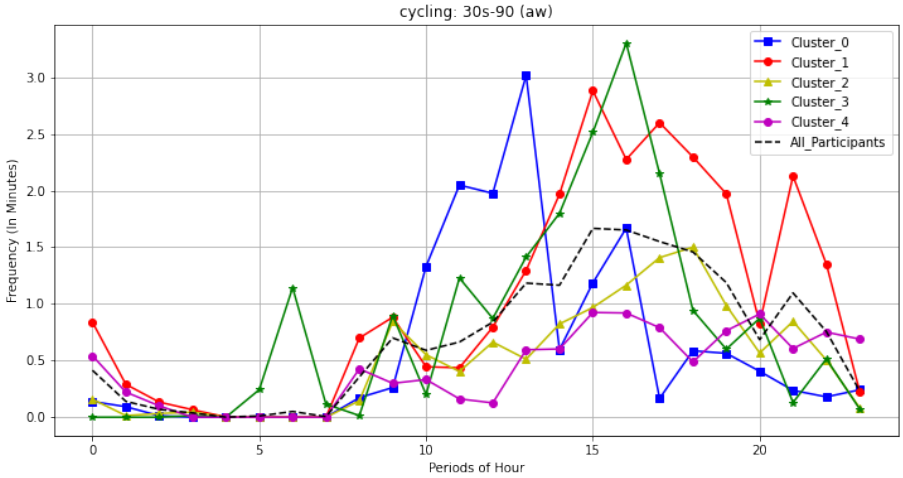
Table A.4: Activity Frequency - Normalized (In Minutes) (k=5)

From the normalized activity frequency table A.4, we can observe that cluster C-0 has significantly more frequency of walking than the other clusters. In the household activities, cluster C-3 which has only female participants is significantly larger than other clusters. The same can be seen in cluster C-1, whereas, in cluster C-4 with only 2 female participants, it has the least frequency among the clusters obtained. This suggests that the female participants are more active in household activities than the male participants. The same can be said for the cycling, with cluster C-1 dominating other clusters followed by cluster C-3. The results above suggest that clusters with more percentage of female participants are more active in cycling. For sedentary activities, i.e. in the case of sitting, cluster C-4 has significantly more frequency than other clusters. We can observe here that clusters with more percentage of male participants are performing sedentary activity more than compared to those with more percentage of female participants. We can notice similar behavior in the case of lying down too. But in cluster C-4, the participants were sitting more rather than lying down. For standing, the frequency is less in the clusters C-1 and C-3 compared to other three clusters i.e. C-0, C-2, C-4.

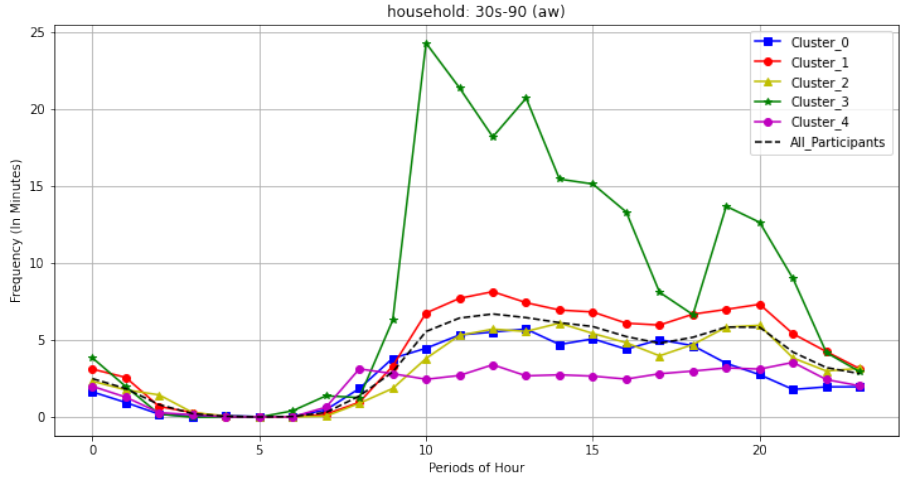
We further investigated visually the normalized frequency per hour for each of the activities in 24 hrs for all the five clusters obtained in figs. A.2 to A.3.




(a) Walking

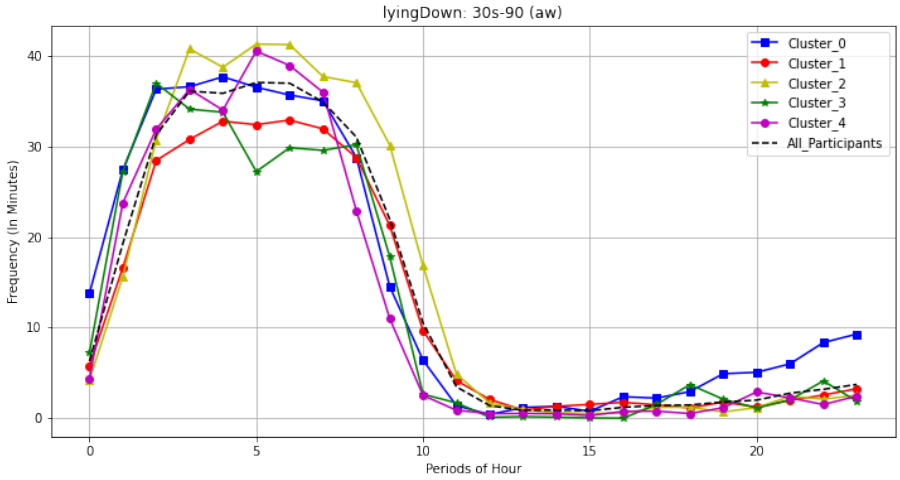



(b) Cycling

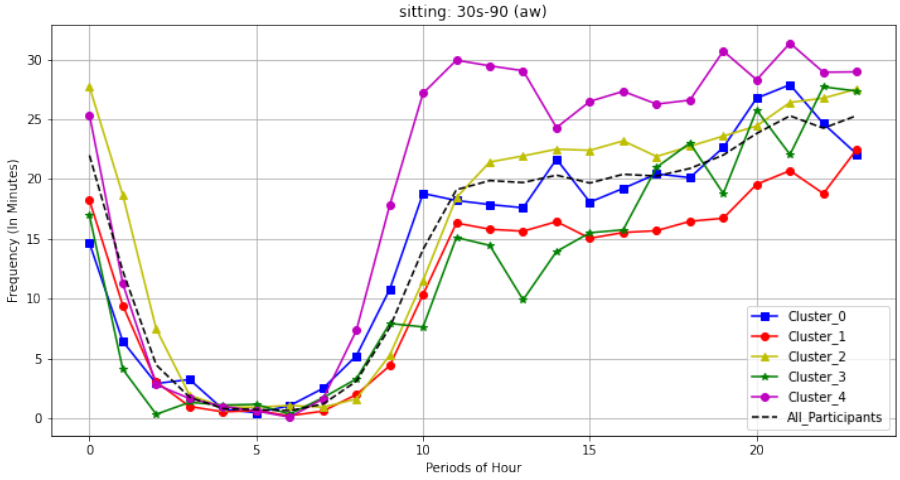



(c) Household

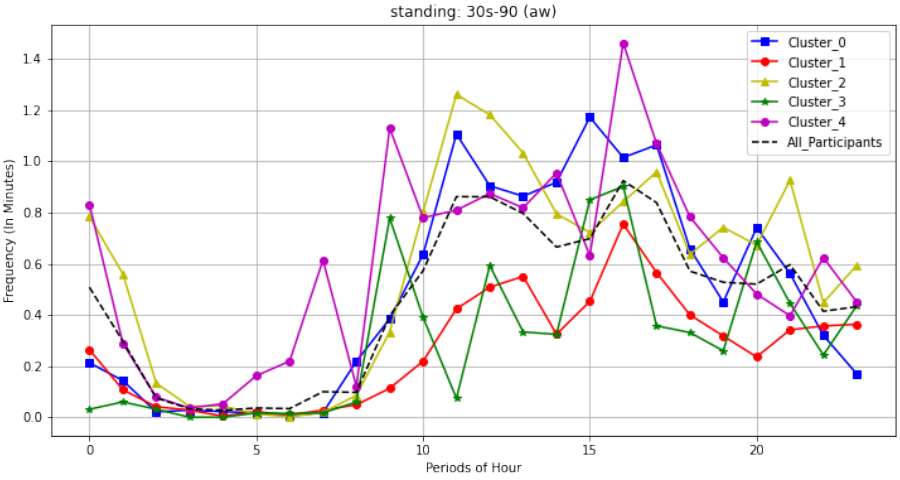
Figure A.2: Normalized Frequency per Hour - Active Activities (K=5)




(a) Lying Down




(b) Sitting




(c) Standing

Figure A.3: Normalized Frequency per Hour - Inactive Activities (K=5)

In the plot for walking A.2a, we can observe certain peaks by cluster C-0 at 9:00, 11:00 and 15:00. Overall participants of cluster C-0 show more frequency of walking activity throughout the day than other clusters. Cluster C-3 shows two peaks, one during noon and another at night time. Cluster C-4 shows a similar peak at night. This suggests that participants of clusters C-3 and C-4 might have a habit of taking a walk at night. From A.2b for cycling, we can clearly observe that clusters C-0, C-1 and C-3 are more active in cycling than other two clusters i.e. C-2 and C-4. Cluster C-3 shows a few small peaks at 6:00 and 9:00, with a big peak at 16:00. We can observe peaks by cluster C-0 from 11:00 to 13:00 hrs. Participants from cluster C-1 showed some increased cycling activity during night time. In A.2c for household activity, with cluster C-3 having only 5 female participants, we can observe that household activities by these participants is significantly more than other clusters during noon with small peaks during night time. The rest of the clusters follow the same pattern as the mean value of all the participants represented by a dashed black line. We can also see that cluster C-4 has least household activity among the clusters obtained.
In A.3a for lying down, we can observe that all the clusters follow almost the same sleeping pattern with cluster C-0 having some peak during night time, whereas cluster C-2 shows more frequency of lying down during day time. In A.3b for sitting, we can see that cluster C-4 has significantly more frequency of sedentary activity throughout the day, whereas cluster C-1 and C-3 has less sedentary activity all over on average. In A.3c for standing, we can observe that clusters C-0, C-2 and C-4 show more frequency of standing throughout the day than clusters C-1 and C-3 respectively. We can notice certain peaks by cluster C-4 during several periods i.e. at 7:00, 9:00 and 16:00 hrs. A common peak can be observed at 11:00 hr by both clusters C-0 and C-2.

## A.3 Profiling the participants using Periods of Hours

### A.3.1 Hierarchical Clustering Results

In order to find the optimal number of clusters for the hierarchical clustering method we used a dendrogram. Figure A.4 represents the dendrogram for generated features (Period of Hours) for every participant. From dendrogram, we can select the optimal number of clusters as 4 for our feature-dataset with 24 features.

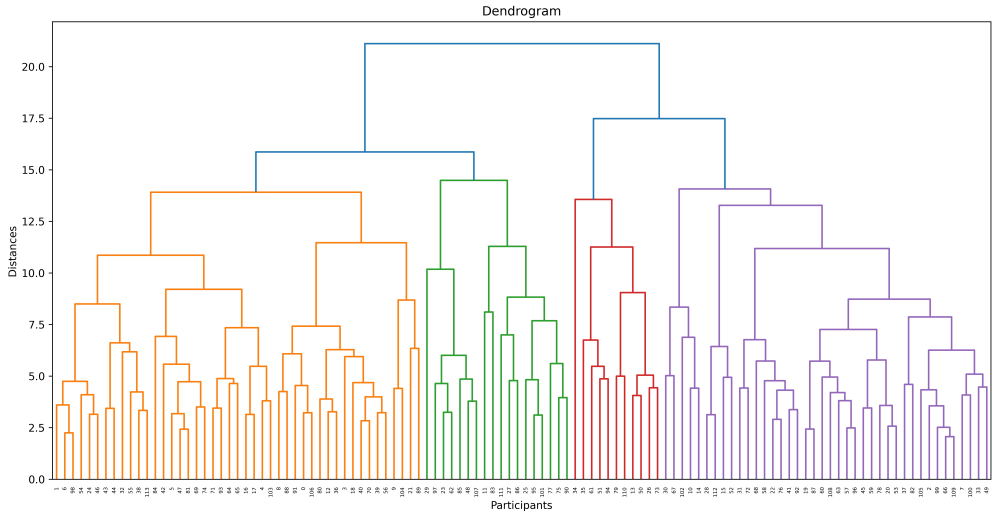



Figure A.4: Dendrogram (Period of Hours)

After applying hierarchical clustering for 4 clusters, we get the following demographic results A.5 and normalized activity frequency A.6.

| Cluster | C-0 | C-1 | C-2 | C-3 |
|---|---|---|---|---|
| **Size** | 18 | 40 | 11 | 45 |
| **G: Male** | 11 | 11 | 5 | 33 |
| **G: Female** | 7 | 29 | 6 | 12 |
| **Age** | 61.48 $\pm$ 5.23 | 60.91 $\pm$ 5.67 | 66.15 $\pm$ 5.32 | 63.83 $\pm$ 5.34 |
| **BMI** | 26.96 $\pm$ 2.80 | 26.49 $\pm$ 2.35 | 26.25 $\pm$ 2.54 | 27.33 $\pm$ 2.44 |

Table A.5: Demographic Results (k = 4) (Period of Hours)

From table A.5, we can see that clusters C-1 and C-3 contain most of the participants, with cluster C-1 having more female participants and cluster C-3 having more percentage of male participants. Cluster C-2 contains the least number of participants and has a mean age of 66, which is more than the overall mean age of participants. The BMI for all the clusters lies between 26 to 27.

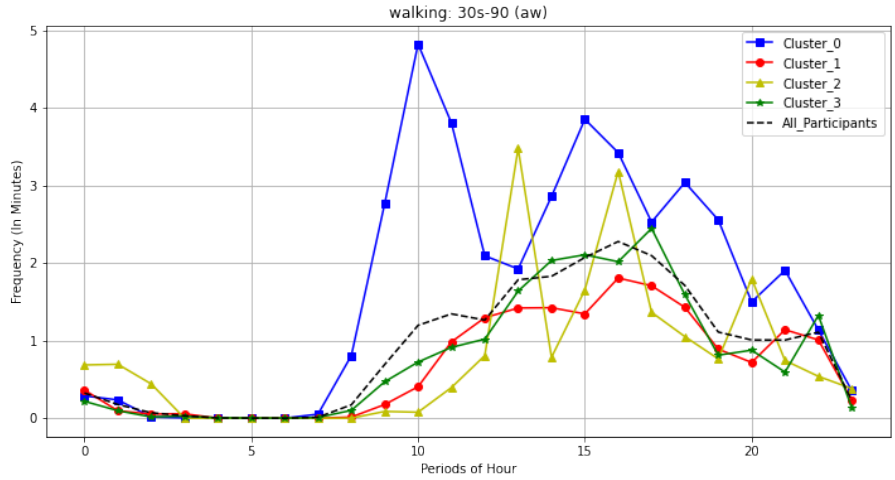
| Cluster | C-0 | C-1 | C-2 | C-3 | All |
|---|---|---|---|---|---|
| **Walking** | 39.96 | 16.56 | 18.91 | 19.16 | 21.51 |
| **Household** | 73.45 | 113.07 | 107.89 | 58.10 | 84.61 |
| **Cycling** | 24.83 | 23.90 | 8.73 | 8.48 | 16.50 |
| **Sitting** | 339.62 | 287.15 | 382.40 | 401.00 | 349.57 |
| **Lying Down** | 275.42 | 309.04 | 346.41 | 357.01 | 326.27 |
| **Standing** | 12.14 | 6.20 | 18.44 | 12.69 | 10.88 |

Table A.6: Normalized Activity Frequency in Minutes (k = 4) (Period of Hours)

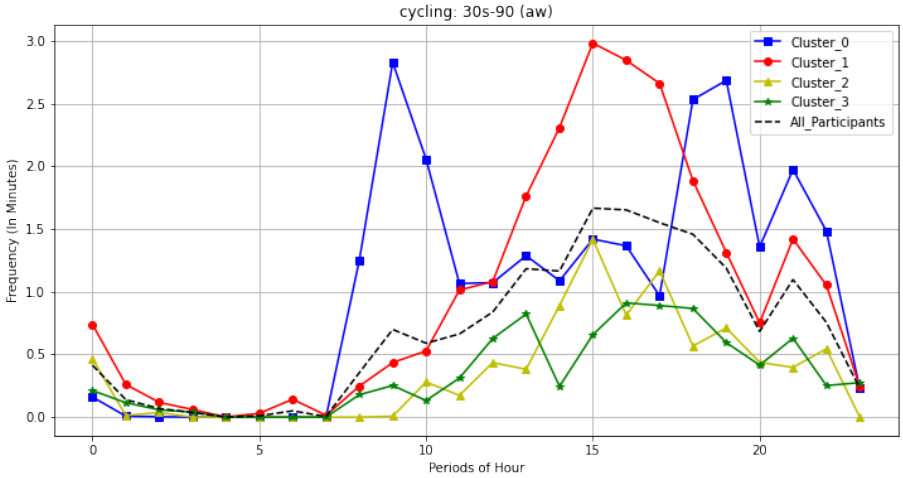
Table A.6 shows normalized activity frequency for all the activities for the obtained clusters. Cluster C-0 is more active in vigorous activities like walking and cycling. For household activities, cluster C-1

has more frequency followed by cluster C-2. For sedentary level activities such as sitting and lying down, cluster C-3 is spending more time than other clusters.

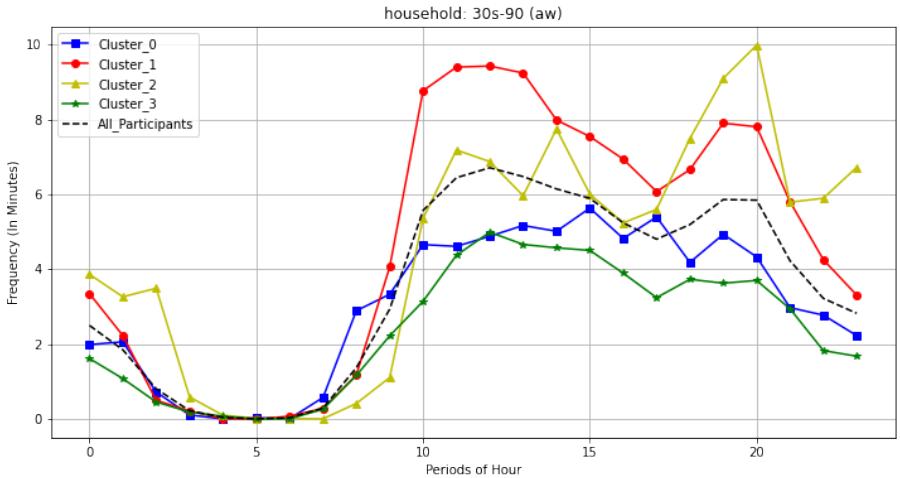
We used line plot to visualize the normalized frequency per hour for each of the activities in 24 hrs for all the four clusters obtained in figs. A.5 to A.6. In A.5a for walking, cluster C-0 shows a peak from 09:00 to 12:00 and at some time periods during the afternoon to evening. Some peaks for cluster C-2 can also be observed during the time period from noon to evening. In A.5b, cluster C-0 has certain peaks during the late evening, whereas in the afternoon, cluster C-1 has a sudden peak. From the pattern of walking and cycling activities, we can draw the inference that the participants of cluster C-0 might be the working group of individuals who need to walk or cycle for commuting to their respective places of work. For household activities A.5c, cluster C-1, which has the largest percentage of female participants seems to be more active from morning to the evening than other clusters.
For sedentary activities i.e. lying down A.6a and sitting A.6b, some peaks for cluster C-2 can be observed in the evening time. Participants of clusters C-2 and C-3 are spending more time lying down and sitting than participants of other clusters.




(a) Walking

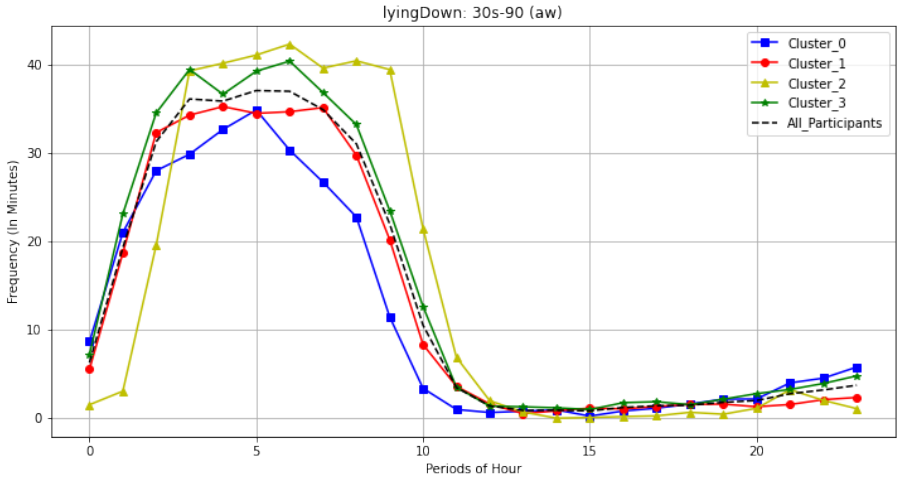



(b) Cycling

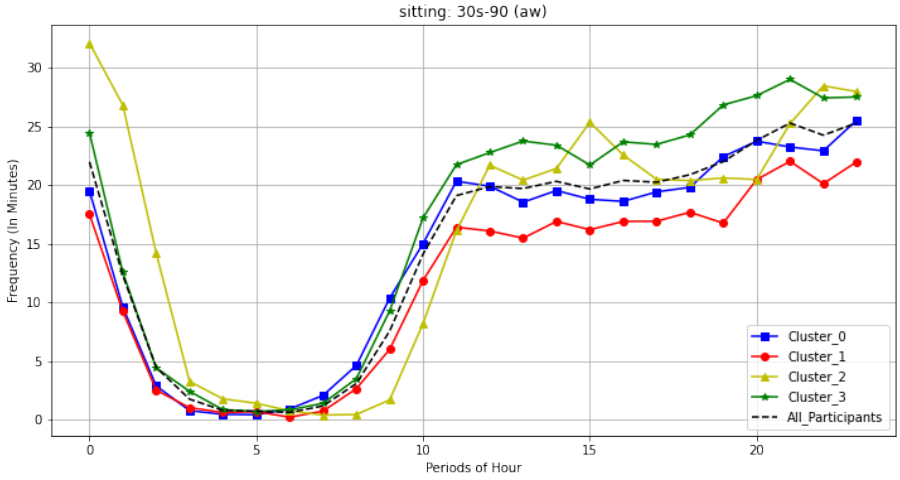



(c) Household

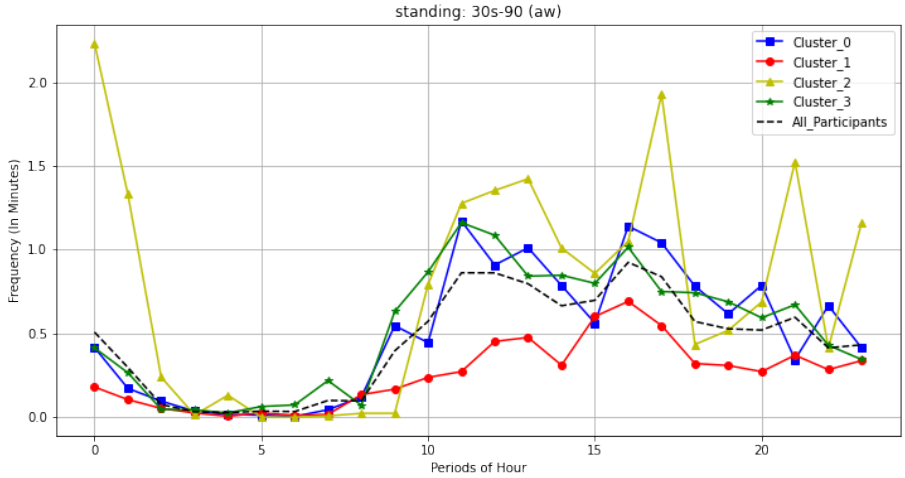
Figure A.5: Normalized Frequency per Hour - Active Activities (k=4) (Period of Hours)




(a) Lying Down




(b) Sitting




(c) Standing

Figure A.6: Normalized Frequency per Hour - Inactive Activities (k=4) (Period of Hours)

### A.3.2 Conclusion

Later, we used a dendrogram to visualize the optimal number of clusters for the hierarchical clustering algorithm. We found $4$ to be the optimal number of clusters using the dendrogram. Similarly, from the table 5.15, we can observe that for $k = 2$ and $k = 4$, we obtained a higher silhouette score than other values of $k$. This provides extra evidence supporting the choice of $k$ for our constructed features.

## A.4 Male Participants

### A.4.1 $k$-means Clustering Results

Afterwards, we checked for $k = 3$ to verify if we can find more clear clusters. The summary table A.7 shows the frequency of all activities with $k = 3$. Participant of cluster C-0 is the same as the clusters obtained with $k = 2$ i.e. participant 62343. While comparing only clusters C-1 and C-2, we can notice that cluster C-1 is more active in walking and household activities, whereas cluster C-2 is more active in cycling. For sedentary level activities, cluster C-1 is spending more time in lying down, whereas cluster C-2 is spending more time in sitting.

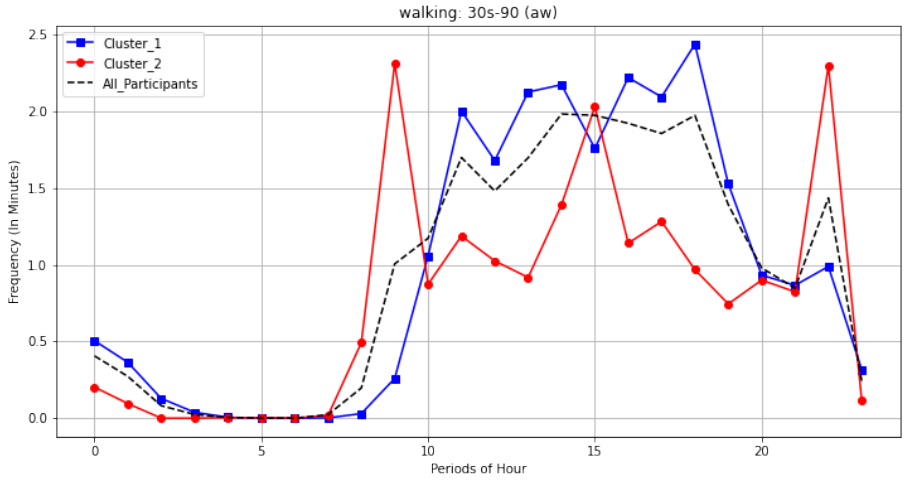
| Cluster | C-0 | C-1 | C-2 | All |
|---|---|---|---|---|
| **Walking** | 71.42 | 23.51 | 18.82 | 22.66 |
| **Household** | 51.00 | 69.02 | 48.74 | 61.62 |
| **Cycling** | 6.57 | 15.59 | 18.43 | 16.43 |
| **Sitting** | 276.72 | 358.47 | 432.86 | 383.15 |
| **Lying Down** | 300.28 | 358.33 | 297.52 | 336.08 |
| **Standing** | 30.50 | 11.44 | 12.13 | 12.00 |

Table A.7: Normalized Activity Frequency in Minutes (k = 3) (PCA)

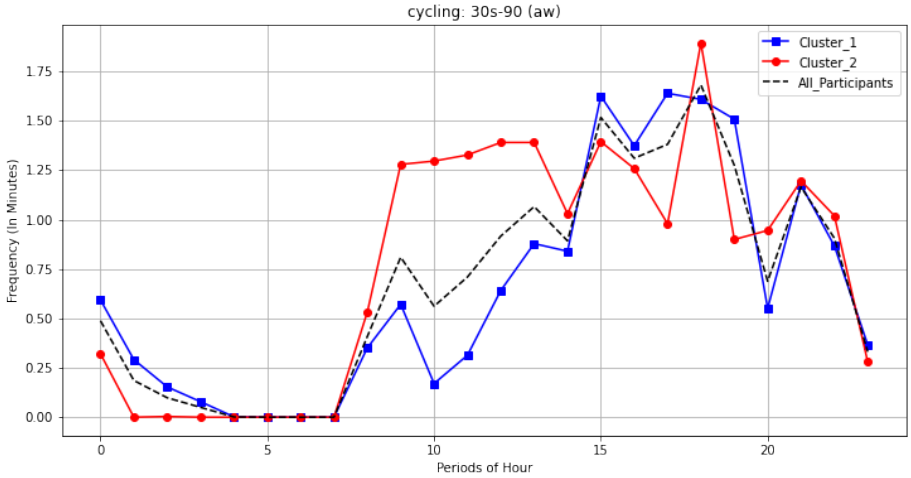
To understand the clusters further, we investigated visually the normalized frequency per hour for each of the activities in 24 hrs for two large clusters i.e. Clusters C-1 and C-2 for better visualization in figs. A.7 to A.8.

In plot A.7a for walking, cluster C-2 shows some interesting patterns. We can notice that the participants of this cluster has certain peak during late morning and late evening. Cluster C-1 is more active than cluster C-2 during noon. Similarly, in plot A.7b for cycling, cluster C-2 seems to be slightly more active from 9:00 to 13:00 and a sudden peak can also be noted at 18:00. For household activities A.7c, cluster C-1 is more active than cluster C-2 during the duration of late morning to early evening.

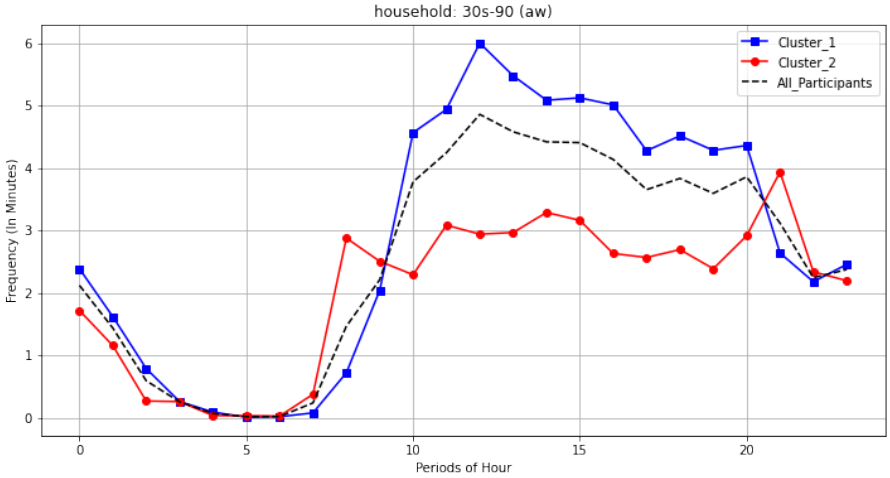
In plot A.8a for lying down, cluster C-1 seems to be spending more time throughout the day, whereas in plot A.8b for sitting, cluster C-2 is spending more time all day. In plot A.8c, for standing, it is hard to deduce meaningful pattern from line plots of the clusters.




(a) Walking

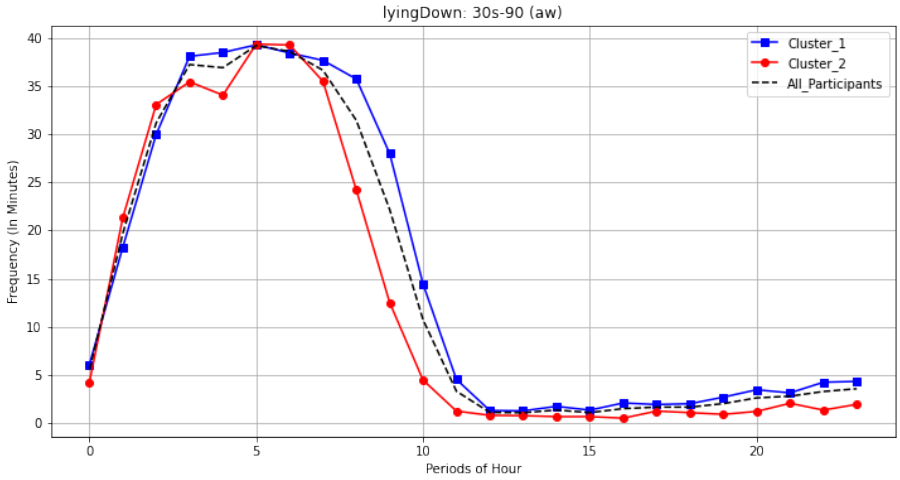



(b) Cycling

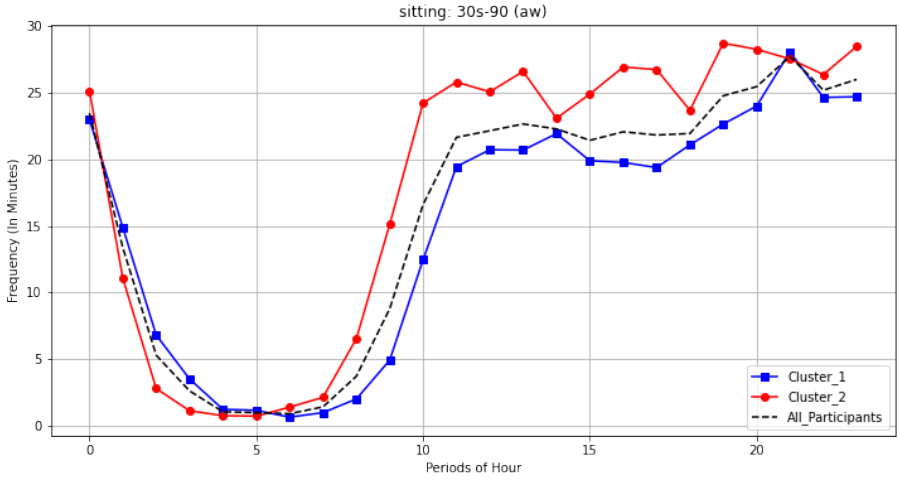



(c) Household

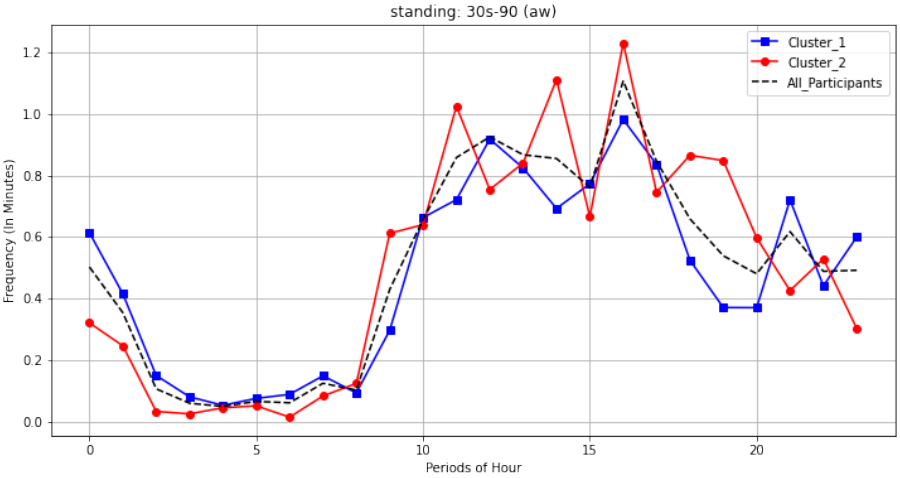
Figure A.7: Normalized Frequency per Hour - Active Activities (k = 3) ($k$-means)




(a) Lying Down




(b) Sitting




(c) Standing

Figure A.8: Normalized Frequency per Hour - Inactive Activities (k = 3) ($k$-means)

### A.4.2 Hierarchical Clustering Results

We used the Hierarchical Clustering method after verifying the optimal value for the number of clusters by using dendrogram. Figure A.9 illustrates the dendrogram created.

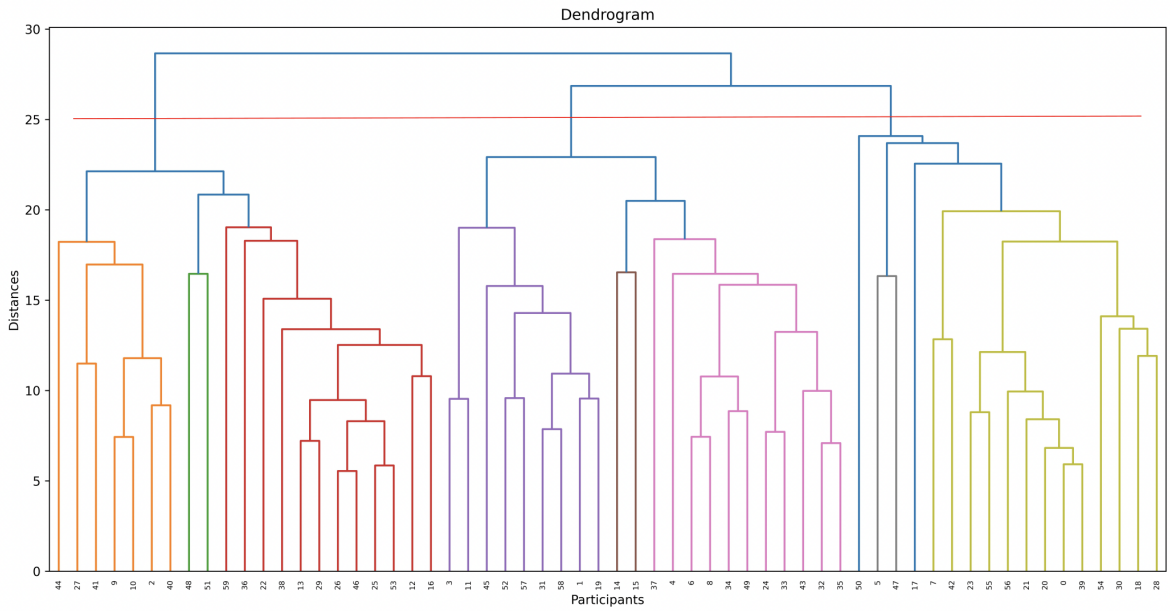



Figure A.9: Dendrogram (Male Participants)

After visually checking the dendrogram, we selected the total number of clusters as 3 as we can see from the red line cut at distance 25 of the dendrogram. The summary table A.8 shows the normalized frequency of all activities in minutes of all male participants.

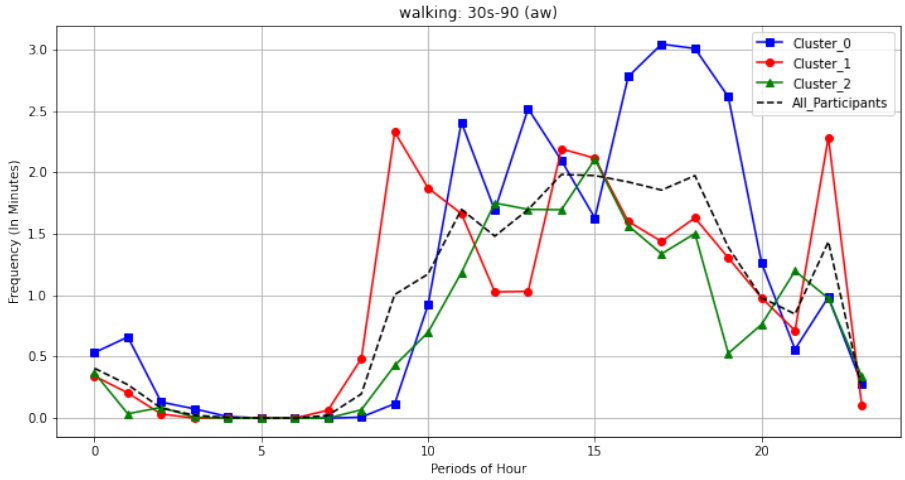
| Cluster | C-0 | C-1 | C-2 | All |
|---|---|---|---|---|
| **Age** | 66.85 ± 3.47 | 60.72 ± 5.71 | 63.68 ± 5.26 | 63.67 ± 5.32 |
| **Walking** | 27.33 | 23.42 | 18.34 | 22.66 |
| **Household** | 62.83 | 48.67 | 73.05 | 61.62 |
| **Cycling** | 4.91 | 17.47 | 24.34 | 16.43 |
| **Sitting** | 410.07 | 424.72 | 322.65 | 383.15 |
| **Lying Down** | 319.37 | 311.18 | 372.76 | 336.08 |
| **Standing** | 13.43 | 14.73 | 8.29 | 12.00 |

Table A.8: Normalized Activity Frequency in Minutes (k = 3) (Hierarchical Clustering)

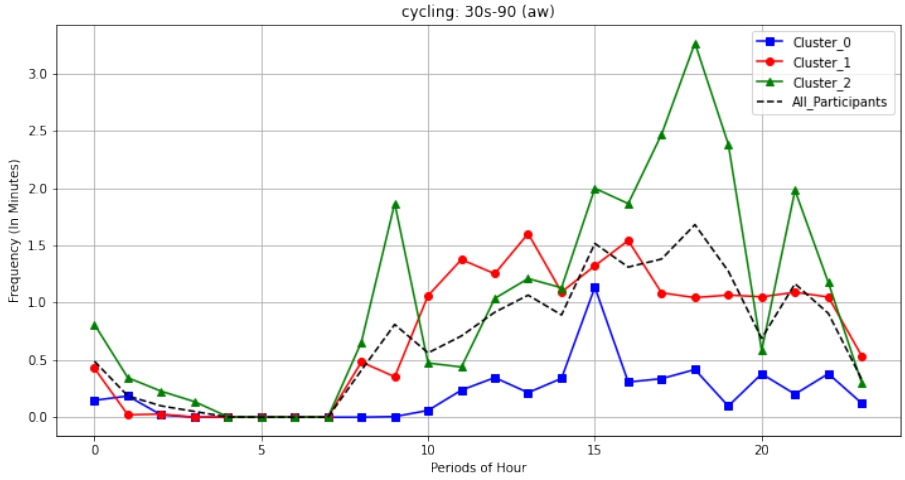
From the table A.8, we can see that male participant of cluster C-0 have more frequency of walking than other clusters, whereas cluster C-2 has more cycling and household activity. For sedentary activities, sitting together with lying down, seems to be approximately equal in all the clusters obtained. The figs. A.10 to A.11 shows the normalized frequency per hour for each of the activities in 24 hrs for $k = 3$.

In plot A.10a for walking, cluster C-0 seems to be more active during afternoon to early evening. For cluster C-1, peaks can be observed in early morning and late evening i.e. at 09:00 and 23:00. In

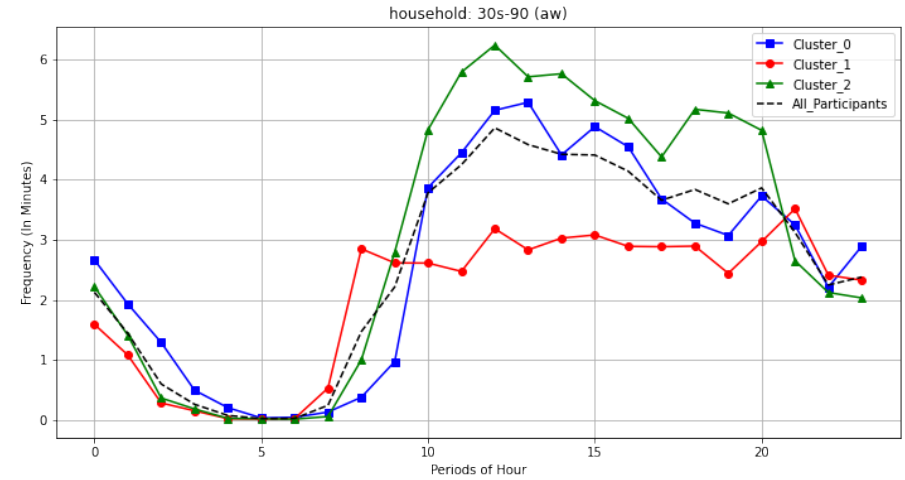
plot A.10b for cycling, cluster C-0 seems to be the least active throughout the day among the three clusters obtained. Cluster C-2 is more active in early morning around 09:00 and also in afternoon. For household activities A.10c, cluster C-2 is the most active cluster overall during the duration of afternoon to early evening.
In the plot A.11a, cluster C-0 is spending less time lying down during night time. While in plot A.11b, cluster C-2 is spending least time sitting among the three obtained clusters.




(a) Walking

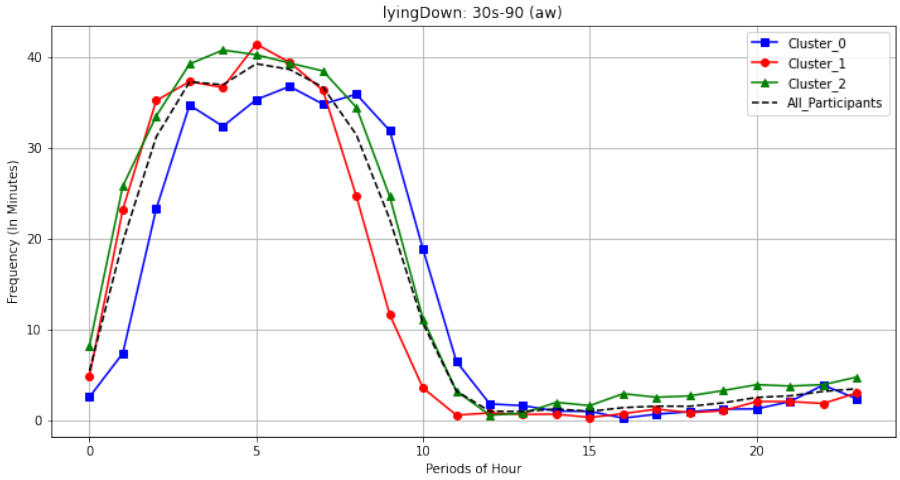



(b) Cycling

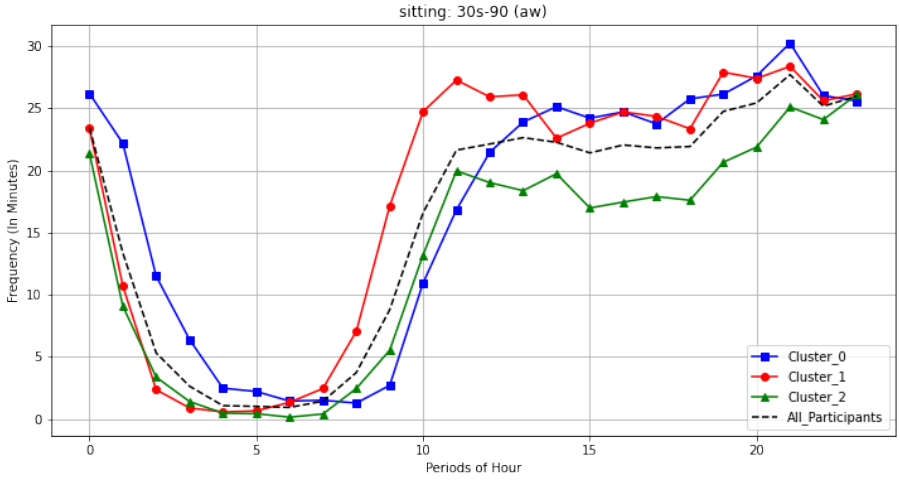



(c) Household

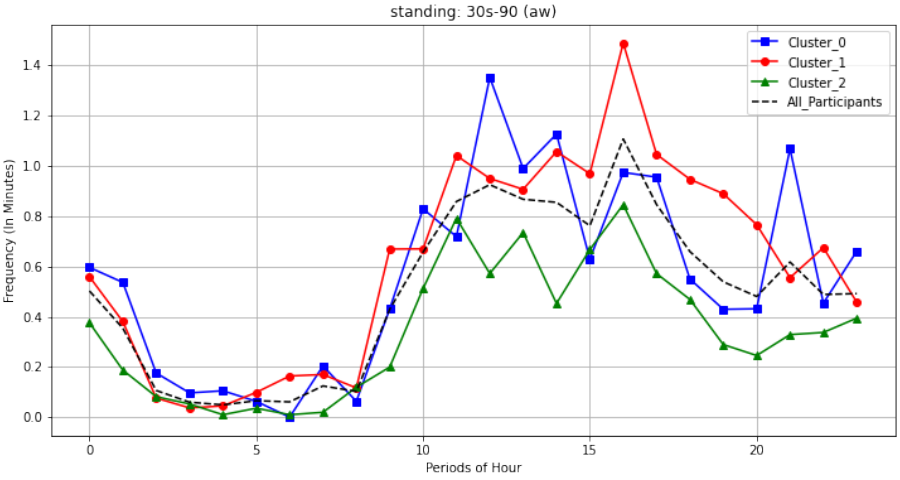
Figure A.10: Normalized Frequency per Hour - Active Activities (k = 3) (Hierarchical Clustering)




(a) Lying Down




(b) Sitting




(c) Standing

Figure A.11: Normalized Frequency per Hour - Inactive Activities (k = 3) (Hierarchical Clustering)

### A.4.3 Conclusion

For Hierarchical clustering algorithm, we used dendrogram and selected 3 as our choice for the number of clusters. Comparing results for both unsupervised clustering method, we can observe that we are obtaining different clusters in both methods. Therefore, from the cluster results of both the methods, it is hard to identify which method is better in terms of producing clusters.

## A.5 Female Participants

### A.5.1 $k$-means Clustering Results

We checked for $k = 3$ to verify if we can find more clear clusters. The summary table A.9 shows the frequency of all activities with $k = 3$.

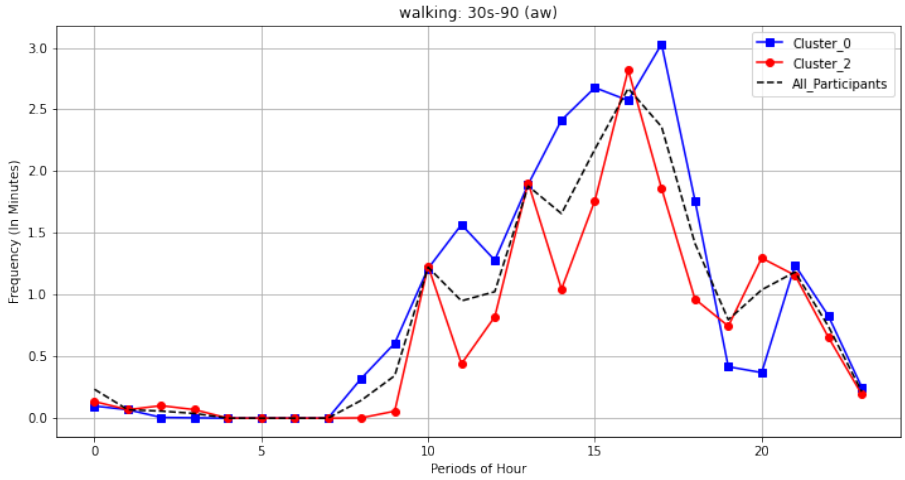
| Cluster | C-0 | C-1 | C-2 | All |
|---|---|---|---|---|
| **Size** | 24 | 1 | 29 | 54 |
| **Walking** | 22.61 | 46.64 | 17.33 | 20.22 |
| **Household** | 121.32 | 131.28 | 100.20 | 110.16 |
| **Cycling** | 20.09 | 34.07 | 13.06 | 16.57 |
| **Sitting** | 302.19 | 170.50 | 325.48 | 312.26 |
| **Lying Down** | 315.60 | 170.00 | 320.20 | 315.38 |
| **Standing** | 9.06 | 6.14 | 10.23 | 9.63 |

Table A.9: Normalized Activity Frequency in Minutes (k = 3) ($k$-means)

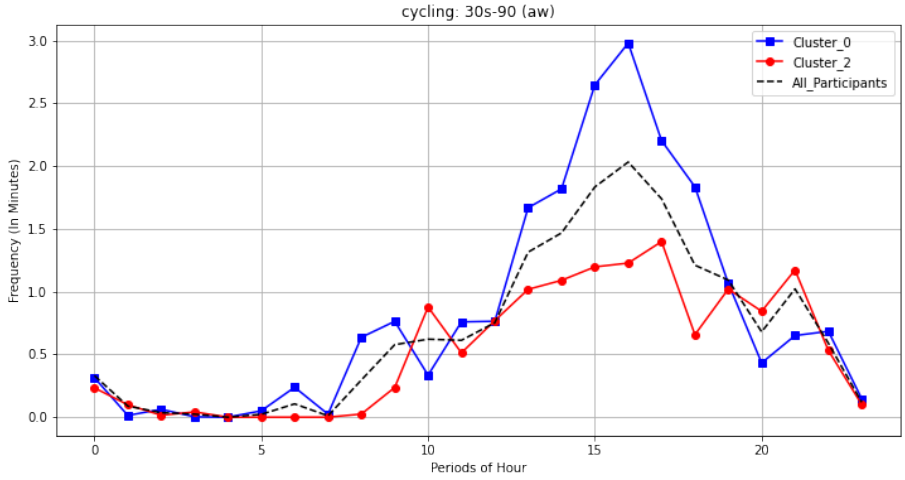
In this case, participant of cluster C-1 is the same as the clusters obtained with $k = 2$ i.e. participant 70668. From the table A.9, we can see that the female participants of cluster C-0 seem to be more active than female participants of cluster C-2. We can also observe the frequency of sedentary activities, which clearly shows that cluster C-2 spends more time in sedentary activities than cluster C-0. Afterwards, we checked visually the normalized frequency per hour for each of the activities in 24 hrs for $k = 3$ in figs. A.12 to A.13.

In plot A.12a for walking and plot A.12b for cycling, cluster C-0 clearly seems to be more active than cluster C-2. Female participants of cluster C-0 are cycling more in the afternoon than female participants of cluster C-2. Similarly, in plot A.12c for household activities cluster C-0 is more active with peaks at 11:00 to 13:00 and at 15:00 i.e. afternoon time period.

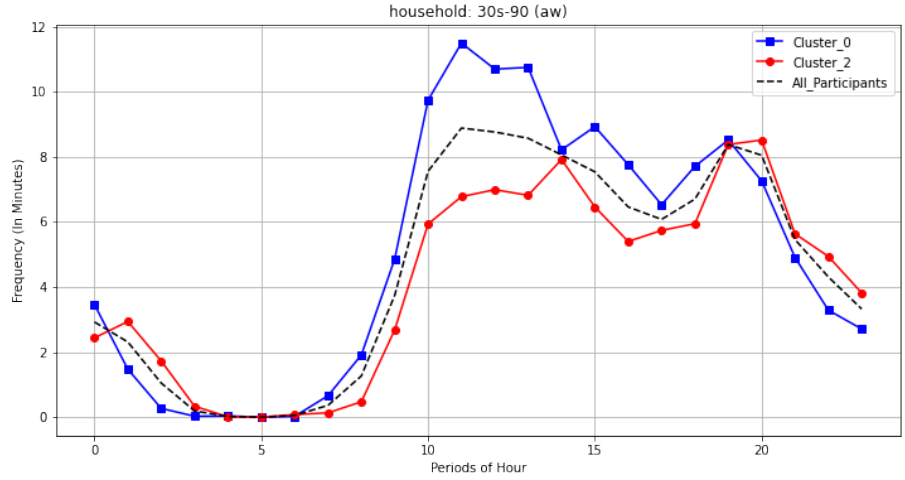
In figs. A.13a to A.13c for sedentary level physical activities, cluster C-0 is less active throughout the day than cluster C-2.




(a) Walking

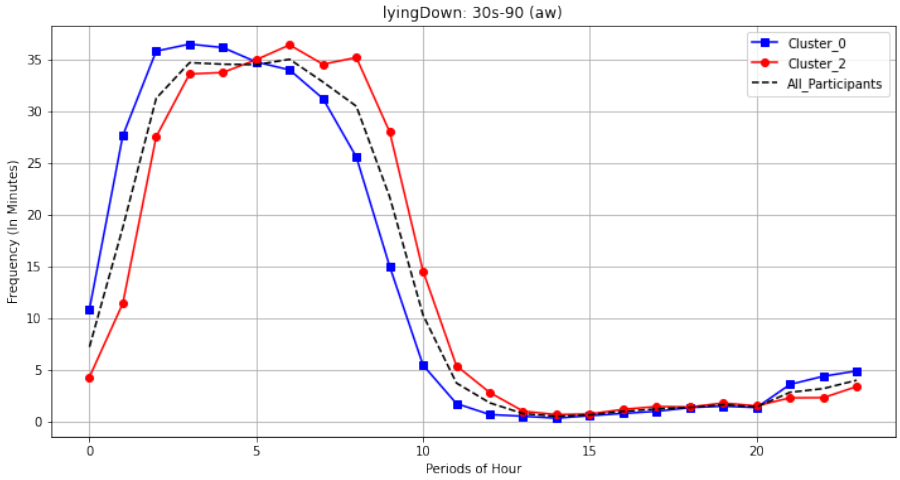



(b) Cycling

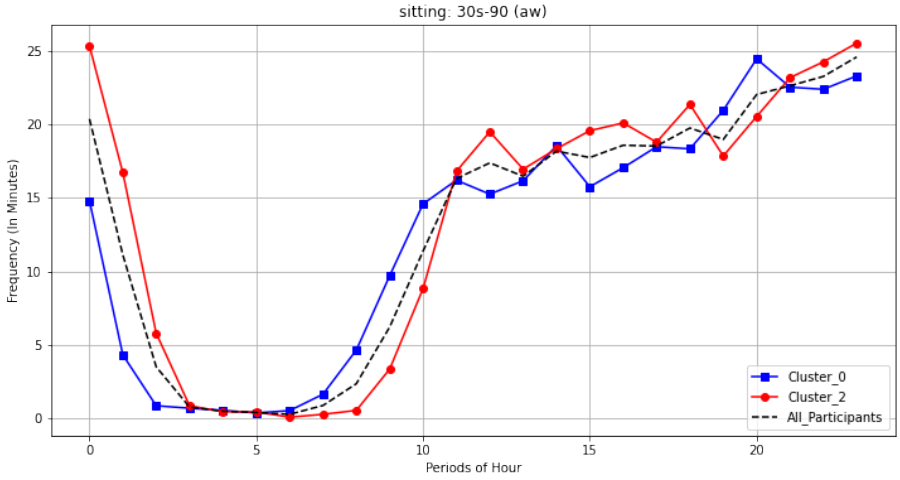



(c) Household

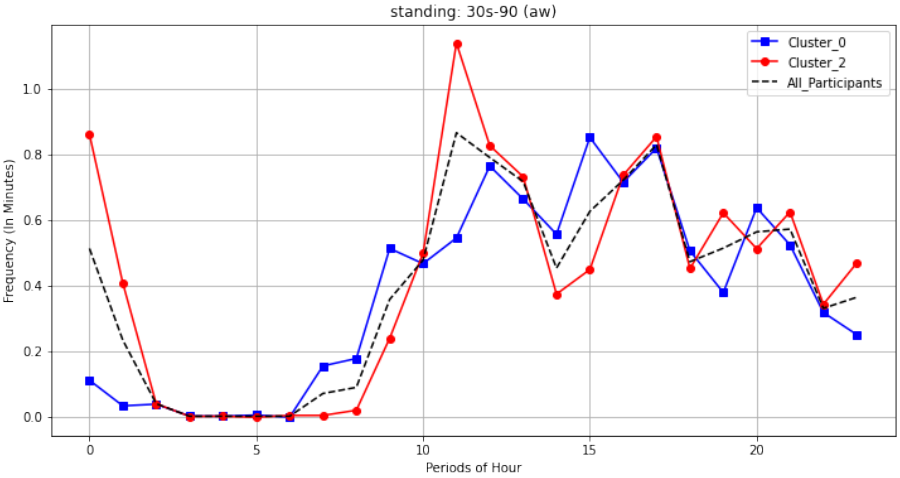
Figure A.12: Normalized Frequency per Hour - Active Activities (k = 3) ($k$-means Clustering)




(a) Lying Down




(b) Sitting




(c) Standing

Figure A.13: Normalized Frequency per Hour - Inactive Activities (k = 3) ($k$-means Clustering)

### A.5.2 Hierarchical Clustering Results

We used the Hierarchical Clustering method after verifying the optimal value for the number of clusters by using dendrogram. Figure A.14 represents the dendrogram formed using only female participants.

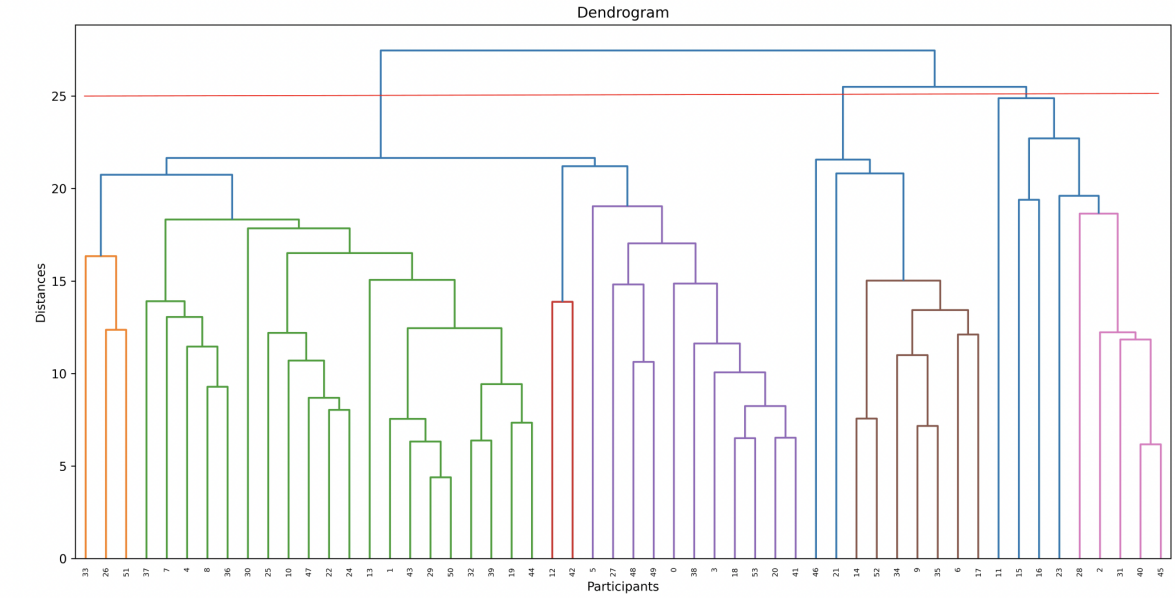



Figure A.14: Dendrogram (Female Participants)

We selected the total number of clusters as 3 as we can see from the red line cut at the distance 25 of the dendrogram. The summary table A.10 shows the normalized frequency of all activities in minutes for all female participants.

| Cluster | C-0 | C-1 | C-2 | All |
|---|---|---|---|---|
| **Size** | 9 | 36 | 9 | 54 |
| **Age** | 64.35 ± 3.11 | 61.07 ± 6.15 | 61.43 ± 5.55 | 61.70 ± 5.62 |
| **Walking** | 29.51 | 16.85 | 24.40 | 20.22 |
| **Household** | 98.53 | 97.31 | 173.21 | 110.16 |
| **Cycling** | 8.84 | 18.52 | 16.52 | 16.57 |
| **Sitting** | 362.44 | 311.36 | 265.68 | 312.26 |
| **Lying Down** | 316.67 | 324.90 | 275.98 | 315.38 |
| **Standing** | 13.23 | 9.12 | 8.12 | 9.63 |

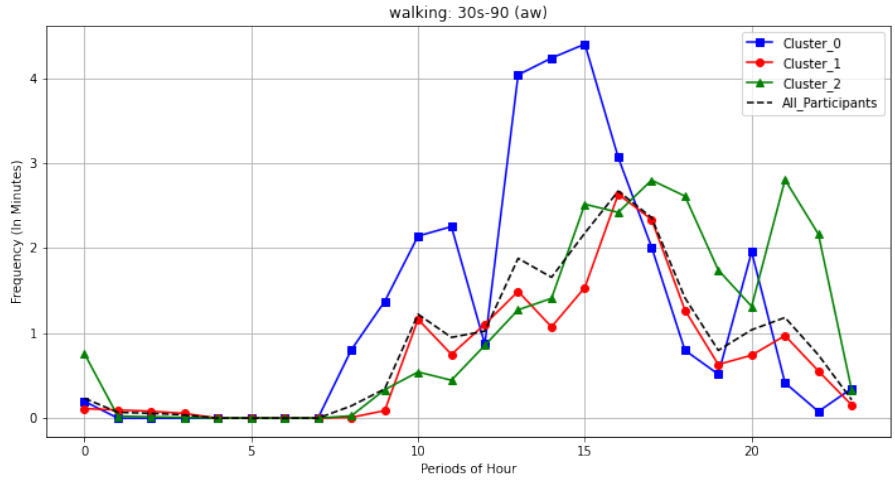
Table A.10: Normalized Activity Frequency in Minutes (k = 3) (Hierarchical Clustering)

From the table A.10, we can see that female participants of cluster C-2 seem to be the most active cluster in household activities than female participants of clusters C-1 and C-0. We can also observe the frequency of sedentary activities, which clearly shows that cluster C-2 spends less time in sedentary activities than clusters C-0 and C-1. Overall we can observe that female participants of cluster C-2 are the most active group among the 3 clusters obtained from the Hierarchical clustering method. Afterwards, we investigated visually the normalized frequency per hour for each of the activities in 24

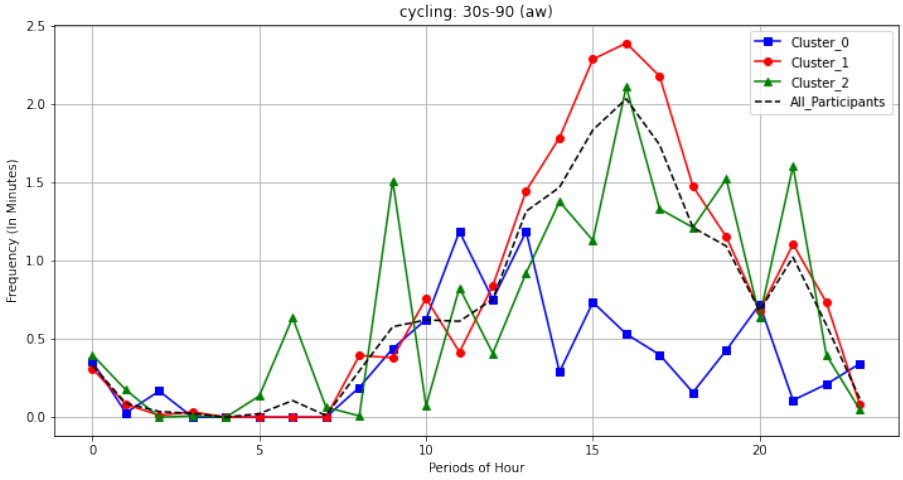
hrs for all 3 clusters in figs. A.15 to A.16.
In plot A.15a for walking, we can observe that female participants of cluster C-0 are more active in walking that other clusters during evening. Similarly, in case for cycling A.15b, cluster C-1 shows some peaks around 15:00 to 18:00. Female participants of cluster C-2 are the most active group in case of household activities as we can see from plot A.15c. For sedentary activities i.e. figs. A.16a to A.16c, cluster C-2 seems to have less frequency than other clusters obtained.

### A.5.3 Conclusion

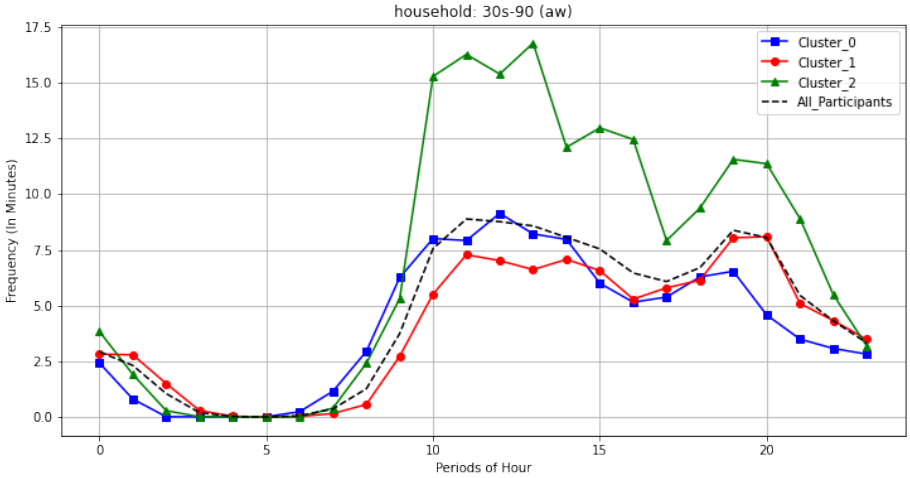
For Hierarchical clustering algorithm, we used dendrogram and selected 3 as our choice for the number of clusters. Similarly in this case too, the clusters that we obtained from both the methods were not identical.




(a) Walking

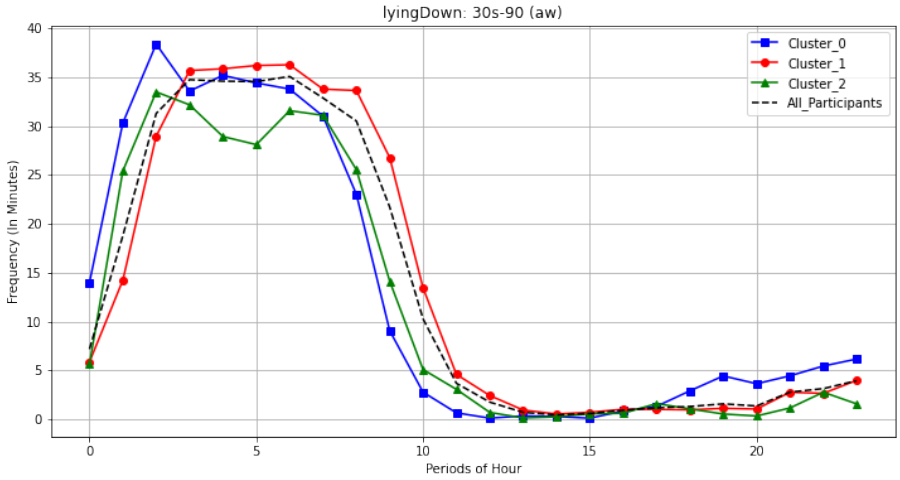



(b) Cycling

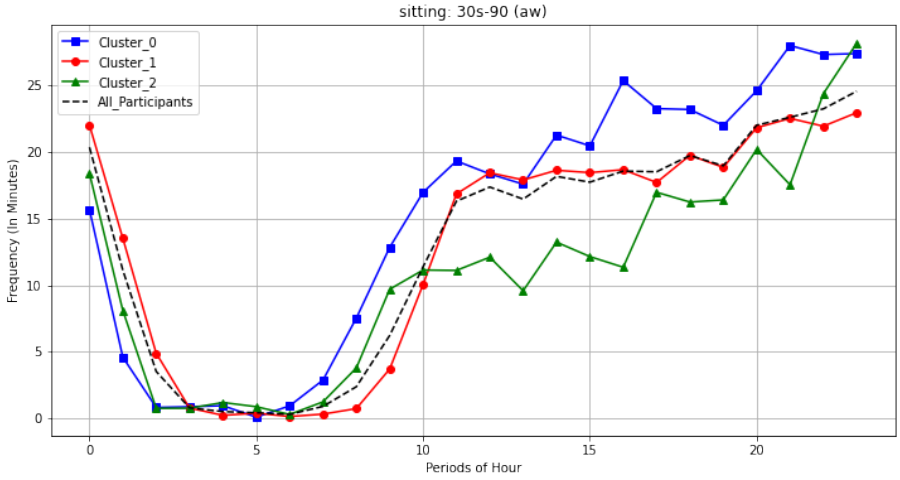



(c) Household

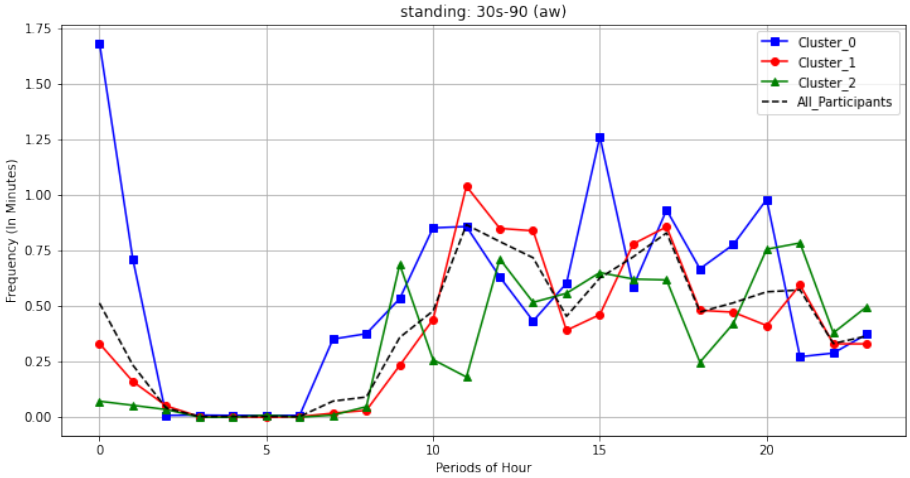
Figure A.15: Normalized Frequency per Hour - Active Activities (k = 3) (Hierarchical Clustering)




(a) Lying Down




(b) Sitting




(c) Standing

Figure A.16: Normalized Frequency per Hour - Inactive Activities (k = 3) (Hierarchical Clustering)