Universiteit Leiden
The Netherlands

Opleiding Informatica

Accessibility of subgroup discovery on housing data through bar visualization

Benjamin Koen Sijpesteijn

Supervisors:
Matthijs van Leeuwen & Hugo Manuel Proença

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

July 27, 2020

Abstract

In this thesis we investigate whether visualizing the results of subgroup discovery on a housing data set makes them more accessible, i.e. more easily understandable for laymen.

Subgroup discovery is a data mining technique that can find subsets in a data set with similar and interesting behavior towards a target variable. The results of such an experiment can give relevant insights into the data, but the results are often displayed in a text format, which might be hard to interpret for non-experts.

Using the bar visualization technique on subgroup discovery results, results might become more accessible for people with no data mining affinity. Two surveys, one showing regular text results and the other showing the visualized results, ask respondents questions about the results. This yields an impression of their understanding of the results, and thus of the accessibility of the results.

The surveys found that the bar visualization did not make the subgroup discovery results significantly more accessible for non-experts. In general, respondents with experience in computer science appeared to be more confident interpreting the results than those who do not have this experience. This was found to be true regardless of how the results were presented. For the age groups that were large enough to give an impression of differences, no substantial differences between the understanding of the results were found.

A suggestion for further research is to use more data sets, in order to get a better understanding of differences in the accessibility of subgroup discovery results by using the bar visualization. Having more responses to the surveys could possibly find that age makes a difference. Using different subgroup discovery result sets and/or other visualization methods may also provide interesting insights in this area of research.

Contents

1 Introduction

Over the last decades, the volume of data has increased rapidly. From 2005 to 2020, the digital universe was predicted to grow from 130 exabytes to 40000 exabytes (Gantz and Reinsel, 2012). With this mass increase of data, automated data analysis will become necessary, because the volume of data will simply be too big to be analysed by humans. Humans are, however, very quick at visual processing (Thorpe et al., 1996), which means visualizations might help to understand results of analysis on big data sets. As the volume of all sorts of data increases, more data about the sales of houses can be found and analysed as well. This thesis looks into whether visualization can help make results of data mining on large housing data sets become more accessible.

There exist several data sets about house sales in a certain area or city. In these data sets each entry describes a house. Columns are used to describe variables which are characteristics of the house, for example the number of bedrooms above ground, the style of the house and whether or not the house has a pool. Each row, representing a house, takes values for each of these variables. The values for the example attributes could be "3", "Two-story" and "No".

For investors, real estate agents or people buying a house, it is valuable to be able to determine whether the listing price of a house is realistic. Subgroup discovery is a technique that can help with this process. It is a data mining technique that can find subsets of similar entries in a data set that show interesting behavior towards a certain target variable (Herrera et al., 2011). Entries are considered similar if they take the same, or more or less the same, value for one or more of the variables. These sets of similar entries, taking only certain variables into account, will also show similar and possibly atypical values towards the target variable.

In a housing context, subgroup discovery can be relevant to determine whether the listing price of a house is fair. For example: a two-story house with three bedrooms and no pool is for sale for €200,000. If houses with those characteristics usually sell for €300,000, then the current listing price is an atypical value towards the target variable. This means that this house can be a great investment. Conversely, if houses with those characteristics normally sell for €150,000, then this house is a bad investment, because it is overpriced.

While subgroup discovery is able to show interesting results, they are not always clear when you do not have experience within this field. The example from the previous paragraph seems quite clear, but using a subgroup discovery tool on an actual housing data set could, for example, give the following subgroup description as a result:

`D_Gr Liv Area` $= (765.00, 987.00]$ `D_Total Bsmt SF` $= (771.00, 1025.00]$ `Paved Drive = Y Garage Qual = TA Land Slope = Gtl Street = Pave -> SalePrice` $=$ `100K-150K`.

Note that this text-based example might not be very easy to interpret, especially for laymen. In fact, the above discription is one of the subgroups that will be used in this research. As seen in the result, a binary target is used: either the house falls in the `100K-150K` region or it does not. The text-based results being hard to interpret is the problem that we will try to solve in this research.

Taking the phrase "a picture is worth a thousand words" in mind, a visualization technique will be used on the results of subgroup discovery to display the results in a non-text-based format for non-experts. In particular, this thesis will consider the bar visualization technique by Novak et al. (2009) to show the results of a subgroup discovery experiment. The following research question is therefore proposed:

Can bar visualization of results of subgroup discovery on a housing data set make the results more accessible to non-experts?

The aim of this research is to use this visualization technique to a broad audience in order to gain insight into whether it enhances the comprehensibility of results of subgroup discovery on a housing data set. If that is the case, it means that everyone should be able to take advantage of the results, since this might not be the case at the moment. Surveys will be used to observe the difference in people's understanding of results displayed as text compared to visualized results.

1.1 Thesis overview

This thesis starts by analysing and discussing previous work done with subgroup discovery, housing data and subgroup discovery on housing data in Section 2. In Section 3, the data set used in this research will be introduced. The aim of this section is to get an impression of the data. Following this, the methods used to run our experiments and to obtain results will be explained in Section 4. The conducted experiment using the visualization to try to find an answer to the research question is described in Section 5. Subsequently, Section 6 will analyse the results we found using the surveys and Section 7 will present the discussion. Lastly, the conclusion is found in Section 8.

2 Related Work

This section will discuss work that has been done in the fields relevant to our research. We will look at previous research done on subgroup discovery, studies concerning housing data and, lastly, work dealing with the combination of these two: subgroup discovery on housing data. The aim of this section is put the present thesis into some scientific context. Previous work relevant to the visualization aspect of this thesis is omitted here, since it requires some more context to be properly understood. Instead, this work will be discussed in some detail in Section 4.2.

2.1 Applications of subgroup discovery

Subgroup discovery is a data mining technique that can find subsets in a data set with similar and interesting behavior towards the target variable. It has been used in numerous fields of research (Herrera et al., 2011). For example, it has been used in the medical domain, where it was used in order to find subgroups of patients who have a relatively higher risk for coronary heart disease (Gamberger et al., 2003). Besides finding that the risks for coronary heart disease for women start about ten to fifteen years later than for men, subgroup discovery found risk groups, which might help in the prevention process. This is helpful because people themselves can determine if they are in a risk group. With their research, they found that subgroup discovery can be useful, even if the data sets used contained bias and/or were not very big.

The fact that subgroup discovery can lead to new knowledge, as well as help in the decision making progress, indicates that it might be valuable in the housing sector as well. Getting a better understanding of the factors that influence the price of houses, could help investors, real estate agents or private buyers in the decision making process of buying a house. It helps them to better understand whether the listing price of a house is fair or not.

Another example of the application of subgroup discovery is for the analysis of points in tennis (de Leeuw et al., 2019). The research discovered that for the player they analysed, it was beneficial if service points lasted two strokes or less. Subgroup discovery here also found that the analysed player profited from not hitting a backhand.

Subgroup discovery was also used to gain an insight into political instability (Lambach and Gamberger, 2008). In this research, subgroup discovery was used to find conditions that might indicate vulnerability of countries to political instability. Subgroup discovery here managed to find important variables for the description of subgroups, that experts or other analyses could have overlooked. Some variables, such as GDP per capita, were an indicator for political instability when they decrease, but other times they were used as an indicator for political instability when they increase. Not all of the results were completely new insights, but they were valuable nonetheless, because the correct identification of some known predictors may indicate that other models or theories are useful. Lastly, the researchers note subgroup discovery is a technique that can also be used on data set with a lot of variables and not many entries.

Some advantages of subgroup discovery described in Lambach and Gamberger (2008): finding important variables that might otherwise be overlooked; not needing experts and it being usable

on small but very descriptive data sets. Our data set contains a lot of variables compared to the number of entries. With other techniques, some of these variables might otherwise be overlooked. Therefore, subgroup discovery is a technique that might be valuable to use in research on this data set.

2.2 Analysing housing data

Machine learning can be used on a housing data set to try to predict future house prices. An example of this is the study by Park and Bae (2015), where different algorithms are tested on a data set containing 5,359 houses in Fairfax County in Virginia. Several algorithms are tested on their ability to predict whether houses will be sold above or below the price that they are listed for. The different algorithms are compared by their classification accuracy score. The researchers found that machine learning algorithms, with the RIPPER algorithm scoring best, are useful to determine whether the price a house is listed for is realistic. Next to this, the study states it is a cheap way to analyse the houses and it could also be a quicker way for, for example a bank, to decide whether a loan can be given.

Originally, the Boston housing data set (Harrison and Rubinfeld, 1978) is a data set that was used for teaching, but it is also used to rate algorithms in order to compare the strength of different algorithms (Lim et al., 2000). The Boston housing data set has fourteen variables for 506 entries. This data set has become well-known, that it is even included in some `Python` (Van Rossum and Drake, 2009) packages, such as `scikit-learn` (Pedregosa et al., 2011). Furthermore, the data set has also been used for teaching regression modules (De Cock, 2011). One possible task in such courses is the construction of a model to predict house prices, using a house's values for the variables in the data set. The models give a weight to each variable in order to calculate the predicted sale price. Lim et al. (2000) use the Boston housing data set next to thirty-one other data sets to test, among many other things, the classification accuracy of twenty-two decision trees, nine statistical algorithms, and two neural network algorithms.

Because the Boston housing data set does not have many entries and because the paper dates from 1978 − making the data not representative − De Cock (2011) proposed the Ames housing data set as a new housing data set to use for teaching. Besides its use for teaching, the Ames housing data set can be found on the competition website Kaggle. The competition here is to use regression for prediction of the Sale Price of a house based on its values for the variables. The Ames housing data set has eighty attributes, making it a very descriptive set. Furthermore, it is newer than the Boston housing data set and has far more entries as well. Therefore, the Ames housing data set will be used in this research.

An important fact stated in Park and Bae (2015) is that analysing the house prices can be beneficial for people looking to buy a house as well as real estate companies, because it can give them a better understanding of the house prices (in a region). For this thesis, this means that the analysis performed may prove valuable, and using a visualization technique to make it more accessible might support this even more.

2.3 Subgroup discovery on a housing data set

The two preceding subsections dealt with previous work on subgroup discovery and housing data, respectively. This thesis looks to combine these two areas. This subsection lists some of the previous work conducted in this (as of yet) relatively small field of research.

Leman et al. (2008) conducted experiments with subgroup discovery on a housing data set. The data set used here is called the Windsor housing data set, introduced by Anglin and Gençay (1996). It contains 546 houses described by twelve attributes. In this paper, Leman et al. introduce exceptional model mining as an addition to subgroup discovery, which allows analysis towards a more complex target, i.e. multiple target attributes at the same time. Exceptional model mining finds a subgroup where houses with a driveway, a recreation room and a minimum of two bathrooms have a correlation of -0.090 between `sales_price` and `lot_size` instead of the 0.549 of the whole data set. This shows that although the `sales_price` and `lot_size` in the whole data set seem to have an influence on one another, this does not have to be the case within each subgroup.

3 Data

This section concerns the Anes housing data set. Firstly, the reader is provided with an introduction to the data set with a description of its origin, its set-up and its contents. Secondly, we will explore the data set to get a feel of the data it includes.

After this section, when the data set is more properly understood, Section 4 describes the methods used on this data set to conduct the experiments in Section 5.

3.1 Data description

As stated in 2.2, the Ames housing data set is a data set proposed by De Cock (2011) to be used for teaching. It can also be used as an alternative for the Boston housing data set, introduced in the same subsection.

The data in the proposed data set are house sales between $2006 - 2010$ in Ames. Ames is the seventh biggest city in the state of Iowa. The data set contains $2,930$ entries with 80 descriptive variables:

- 23 nominal, for example "Neighborhood": physical locations within Ames city limits;
- 23 ordinal, for example "Kitchen Qual": kitchen quality;
- 14 discrete, for example "Overall Qual": rates the overall material and finish of the house;
- 20 continuous, for example "Gr Liv Area": above grade (ground) living area square feet.

The right-most attribute of each states the SalePrice of the house, which is the target variable in this research. For a full overview of all variables and their description see Table 6 in A.1 (Kaggle). These definitions will be used throughout this thesis. The average house price in this data set is $\$180,796$. The aim of this research is to find subgroups (explained in Section 4.1) of similar houses in the data that show an interesting behavior towards the sale price of the houses. This is done by using subgroup discovery in Cortana and Orange introduced in Section A.2 and Section 4.3 respectively.

3.2 Data exploration

In order to get an impression of the data set, Figure 1 and Figure 3 were constructed using the `seaborn`-package (Waskom et al., 2017) in `Python` (Van Rossum and Drake, 2009). Figure 1 is a correlation matrix that uses all variables that are expressed in numbers. A correlation matrix is a grid where each row and each column represents a variable. Row i represents the same variable as column i. Each box in the grid shows the correlation between the row-variable and the column-variable. Correlation between two variables means that if one value increases the other increases as well. It is therefore evident that the boxes along the diagonal from the top left to bottom left all show a 1.00, since these show the correlation between a variable and itself.

From Figure 1 we see that `Overall Qual` (overall material and finish quality) is the highest correlating variable with our target variable `SalePrice` (0.80). `Gr Liv Area` (Above grade (ground)

living area square feet) with 0.71 is the second highest. It is intuitive that the sale price will increase as the overall quality of a house or the ground living area increase (or both). The third strongest correlating variable with `SalePrice` is `Total Bsmt SF`, the total area of the basement in square feet, with a correlation of 0.63. The least correlating variable from the correlation matrix is `BsmtFin SF 2`, which is the square feet of a basement finished in type 2, with 0.01.

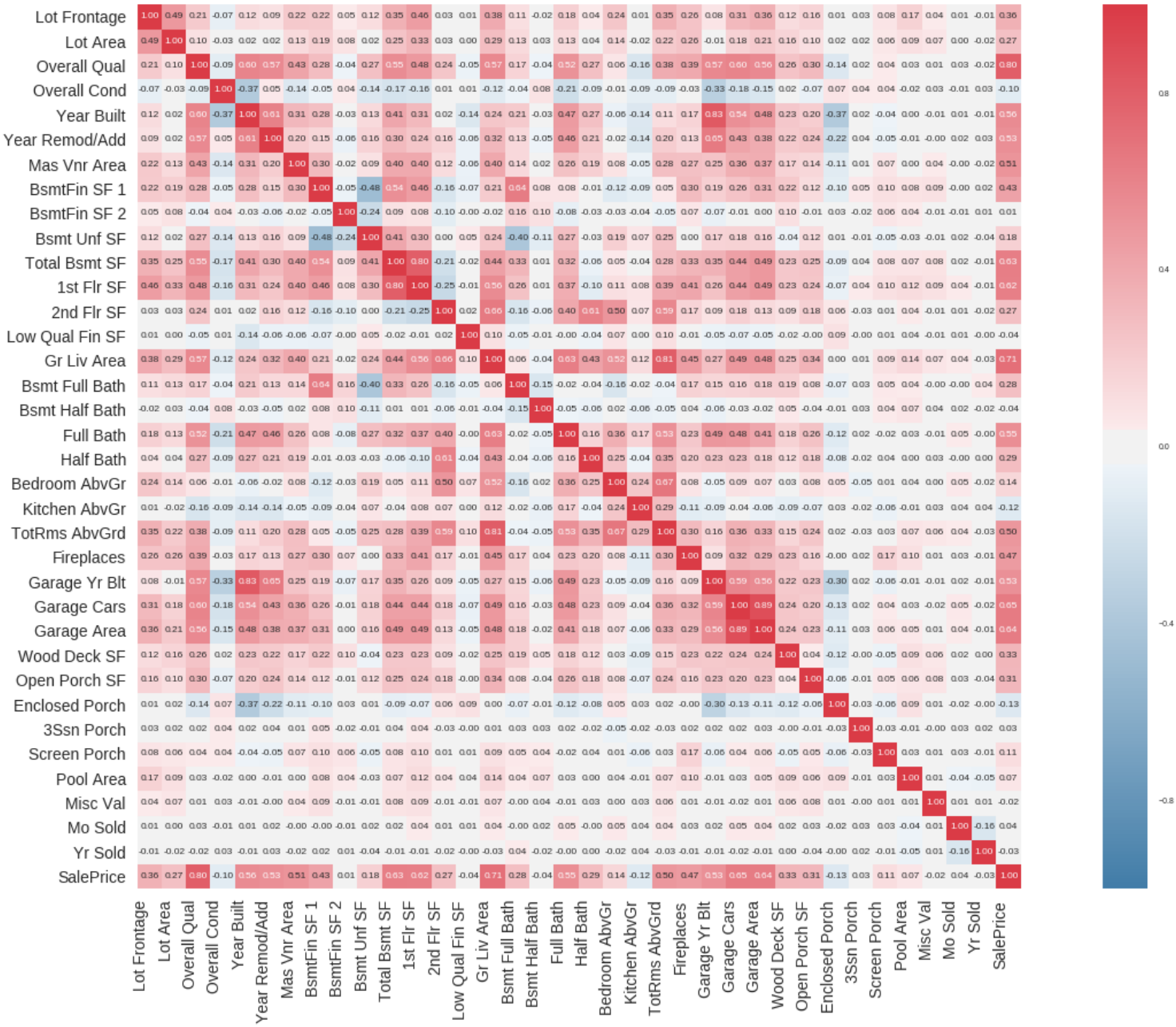
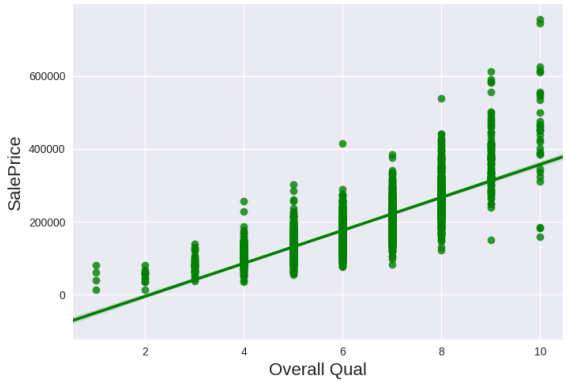


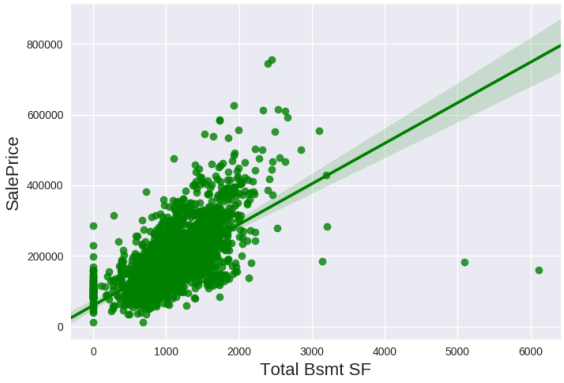
Figure 1: Correlation matrix of variables expressed in numbers

To get a feel of what these different values for the correlation looks like, Figure 2 shows plots for the three variables correlating the strongest with `SalePrice`, as well as for variable correlating the least with `SalePrice`. A scatterplot makes this very visible: for each house in the data set it plots the values for the two variables. It is now easy to see that `Overall Qual`, `Gr Liv Area` and `Total Bsmt SF`, seen in Figures 2a, 2b and 2c respectively, correlate strongly and that `BsmtFin SF 2` seen in Figure 2d does not. For the three plots with strong correlation, there seems to be a predictable reaction of the `SalePrice` when the value for the variable on the $x-$axis changes. For the fourth plot, with very weak correlation, the value for `SalePrice` moves seemingly more random as the value for `BsmtFin SF 2` changes.

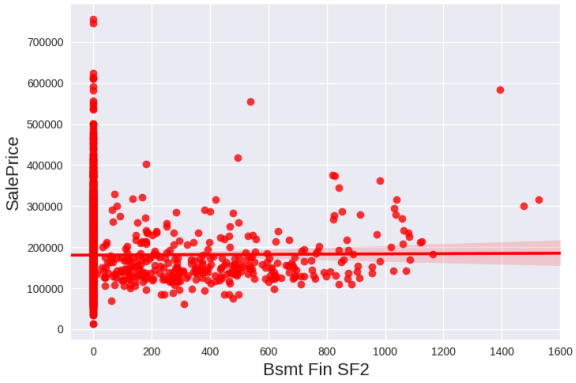


(a) `Overall Qual` plotted against `SalePrice`: correlation of 0.80

(b) `Gr Liv Area` plotted against `SalePrice`: correlation of 0.71



(c) `Total Bsmt SF` plotted against `SalePrice`: correlation of 0.63



(d) `BsmtFin SF 2` plotted against `SalePrice`: correlation of 0.01

Figure 2: Plots of the strongest three and weakest one correlating variables against `SalePrice`

Finally, the plot in Figure 3 shows the distribution of the sale price in dollars of the houses in this data set. It allows us to see what interesting behavior of our target variable could be: the plot shows what typical values can be and therefore also what atypical values for `SalePrice` could be. We see that a lot of the houses fall in the $\$150,000 - \$250,000$ region, with most outliers towards the

higher prices. Subgroups outside of this region might therefore be interesting to find with subgroup discovery. We can also see that the average house price of $\$180,796$, as stated in Section 3.1, seems to fit with this picture.

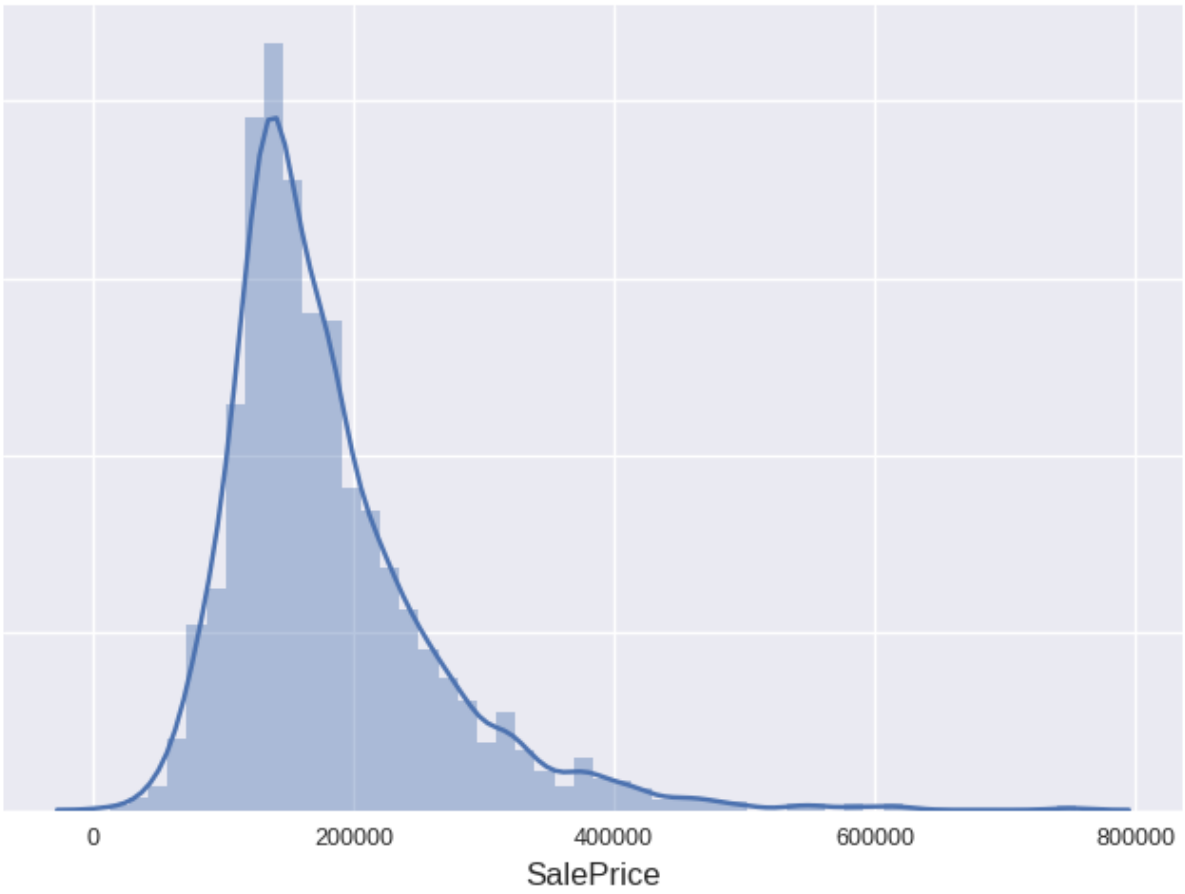


Figure 3: Histogram showing the distribution of `SalePrice`

4 Methods

This section describes the methods used in this research. These methods are the techniques and programs used to work towards finding an answer to the research question in Sections 5 and 6. Subsection 4.1 explains what exactly subgroup discovery is in more detail. Previous work done on the visualization of subgroup discovery results is discussed in Subsection 4.2. Finally, in 4.3, Orange will be introduced: a program for subgroup discovery and visualization of the subgroup discovery results.

4.1 Subgroup discovery

Subgroup discovery is a descriptive and exploratory data mining technique used to find subgroups in a data set. A subgroup is a set of different entries that show similar behavior with regards to some − or sometimes more than one − target variable. Each subgroup has one or more descriptive rules that restrict the values an entry can take for a number of attributes. An entry in the data set is a member of the subgroup if it follows these rules. The target variable is the variable that we try to predict, which is the `SalePrice` in this thesis.

Instead of taking the continuous variable `SalePrice` as the target variable, in this thesis a new binary target variable is set up using `SalePrice`. This binary variable is defined by a price range, evaluating to `True` if a house price falls within this range and `False` otherwise. This is an intuitive choice, because potential property buyers often have a price range set. Furthermore, having a binary target variable is useful computationally, as the software used is unable to process continuous variables. With the use of a binary target, subgroups contain positive and negative examples. Positive examples are houses in a subgroup that follow the description and have the chosen value for the target variable, whilst negative examples are houses that follow the description but do not have the desired value for the target variable.

A subgroup has a certain coverage, which is the number of entries in the data set that are described by the subgroup's rules. If the coverage is too high (closer to the total number of elements in the data set), the subgroup is too general as it describes a very big part of the data set. On the other hand, a very small coverage means the subgroup is very specific. This is also not optimal, because it is only relevant for very few entries.

Subgroups are ranked by a score on a quality measure. A subgroup that scores high on a quality measure is usually an interesting subgroup. Quality measures use a number of parameters to calculate a subgroup's score. An example of such a parameter is the coverage. Another example, for non-binary targets, is the use of the standard deviation of the target variable within the subgroup. A small standard deviation for a subgroup makes the subgroup more specific, which is good if the coverage is not too small, because a small coverage makes the subgroup too specific.

There are a number of different algorithms which can be used for subgroup discovery. Differences in the algorithms for example is whether the search strategy is exhaustive or heuristic (Helal, 2016). Exhaustive search methods take all data into account whereas heuristic search only covers a part of the data. Exhaustive search is not realistic for large data sets, but will generally result in

higher scoring subgroups. One such algorithm is CN2SD, a heuristic algorithm. A classification rule learner CN2 already existed. A classification rule learner is an algorithm that finds rules to use for predicting or classifying the data. The CN2-SD algorithm is a slightly altered version of CN2 to be used for subgroup discovery (Lavrač et al., 2004b).

As an example, consider the data set of cars seen in Table 1 below. As the target variable we take the price of the car. A subgroup could for example be based on the amount of horsepower a car has. Note that the BMW and the Audi have a similar amount of horsepower and also a similar price. Therefore, the Audi and the BMW form a possible subgroup. Similarly, the VW and the Ford may form a separate subgroup, with a similar descriptive rule. Now consider the variables #Doors and Color. A possible subgroup contains the BMW, Ford and Rolls-Royce. However, these do not show similar behavior toward the target variable and are therefore not a subgroup that would score high on a quality measure in this data set.

Brand	**Horsepower**	**#Doors**	**Color**	**Price (€)**
BMW	350	5	Black	50,000
Audi	300	3	Grey	40,000
VW	120	5	Blue	20,000
Ford	150	5	Black	25,000
Rolls-Royce	500	5	Black	250,000

Table 1: An example data set about features of cars and their price

4.2 Subgroup visualization

Visualization of subgroups might be useful to get a better understanding of the results. The four purposes of visualization described by Kralj et al. (2005) are:

- to better illustrate the model to the end user;
- enable comparison of models;
- increase model acceptance;
- enable support for "what-if questions".

Furthermore, Kralj et al. argue that since subgroup discovery is a descriptive pattern mining technique, which are techniques that tries to show patterns in the used data using conditions that the data must follow, it is important that the results are also understandable for the end user. In the paper, four visualization techniques are discussed:

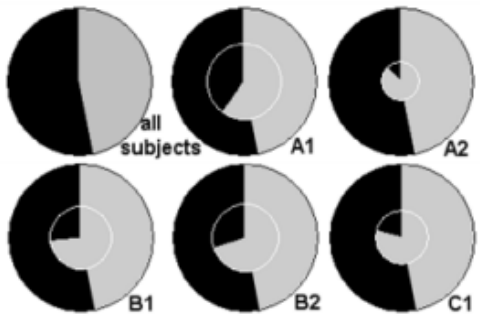
1. **Pie charts**: shown in Figure 4a. Each pie, except for the pie describing all subjects (the complete data set), represents a subgroup and is made up of two pies. The outer pie shows how many of the entries of the whole data set are described by the chosen value for the target variable. The inner pie indicates how many of the examples described by the subgroup are described by the chosen value for the target variable. The size of a subgroup is shown by the

size of the inner pie. When displaying the risk for atherosclerotic coronary heart disease to non-experts, it is found that pie charts are appealing because they have seen them before, but pie charts make it hard to determine and compare the sizes of the pies (Gamberger et al., 2002).

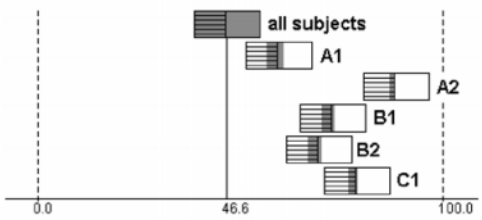
2. **Box plots**: shown in Figure 4b, this visualization also shows a visualization for all subjects. Each box describing a subgroup also visualizes every entry. The striped part of the boxes represents examples described by the subgroup that follow the value for the target variable, i.e. positive examples. In contrast, the white part of each box represents the negative examples, meaning examples described by the subgroup that do not have the chosen value for the target variable. The grey square in each box shows the distribution of positive and negative examples within the subgroup. A grey box more to the left indicates more positive examples. The horizontal location of a box shows the percentage of examples represented by the box that follow the chosen target variable's value. Box plots are harder to understand than pie charts, but they are able to properly show the coverage and make comparison of subgroups easier as well (Gamberger et al., 2002). In this paper, the authors show the box plots visualization method as an example of visualizing patients with a higher risk for atherosclerotic coronary heart disease.

3. **Distributions of a continuous attribute**: shown in Figure 4c. For this visualization we need at least one numeric variable, which is plotted on the $x-$axis. On the $y-$axis you place the target variable. The lines in the graph show the number of positive examples on the top half and the number of negative examples on the bottom half for each subgroup. In Figure 4c, we see all subjects, as well as a subgroup A1 and a subgroup B2. This subgroup discovery visualization method has been used to display the found subgroups of patients with a higher risk for atherosclerotic coronary heart disease (Gamberger et al., 2002). They use different attributes on the $x-$axis to compare the found subgroups for each of these attributes. This allows them to see the influence of the attributes used on the risk for atherosclerotic coronary heart disease per subgroup.

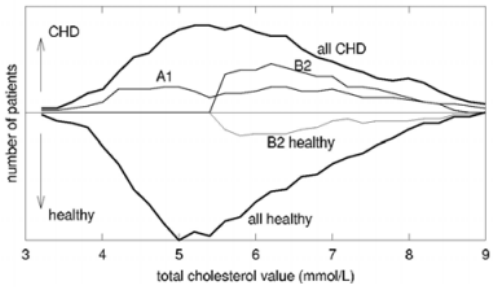
4. **ROC space**: shown in Figure 4d, displays the false positive rate on the $x-$axis and the true positive rate on the $y-$axis. False positives are examples that are described as positive when this is not the case. True positives are examples that are predicted to be positive and are positive. The closer a subgroup in this visualization is to the diagonal, the less interesting it is. The ROC space was used to compare how well the CN2 algorithm performed compared to the CN2-SD algorithm (Lavrač et al., 2004a). In the paper, they used both algorithms on the Australian UCI data set, results of both were visualized in the same ROC space. For both ways they visualize the two algorithms in the ROC space, the CN2-SD scores better.



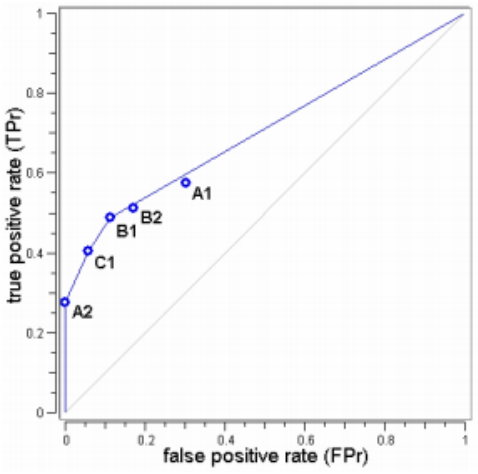
(a) Pie chart visualization



(b) Box plots visualization



(c) Distributions of a continuous attribute visualization



(d) ROC space visualization

Figure 4: An example of each of the subgroup discovery visualization techniques from (Kralj et al., 2005)

Next to these, a new visualization technique is proposed in Kralj et al. (2005): bar chart. An example of this visualization, from the same paper, is shown in Figure 5. The top bar represents the whole data set, the other bars show results of individual subgroups. The green area denotes positive examples and the red area indicates the negative examples. The subgroups are sorted based on the proportion of positive examples compared to negative examples. The paper finally summarizes the different visualization methods based on five evaluation criteria: the intuitiveness, the attractiveness, the correctness, the usefulness and whether or not the contents are displayed. The scores of each visualization method on the criteria can be found in Table 2. From this table, it can be seen that the newly proposed method scores high, and that the distributions of a continuous attribute does too.

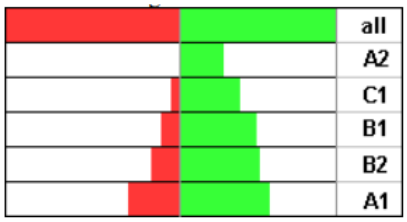


Figure 5: Example of the bar chart subgroup discovery visualization technique proposed in Kralj et al. (2005)

	Intuitive	Attractive	Correct	Useful	Contents
Pie charts	+	+	0	-	-
Box plots	-	0	+	+	-
Distr. of a cont. attr.	+	+	-	+	+
ROC space	+, -	0	+	+	-
Bar chart	+	+	+	+	-

Table 2: A table showing the scores of each technique for five evaluation points, from Kralj et al. (2005)

Novak et al. (2009) make a small difference in the evaluation criteria used by Kralj et al. (2005), but the distributions of a continuous attribute and the bar chart still score best. In this paper these two techniques are combined to get a good visualization that also shows the contents. An example of this visualization can be found in Figure 6 below. This is a great way to show the results of a subgroup discovery experiment: it will score a "+" on all evaluation criteria seen in Table 2. Based on these findings, we decide to use the bar chart visualization in this research to visualize subgroup discovery results of a housing data set.

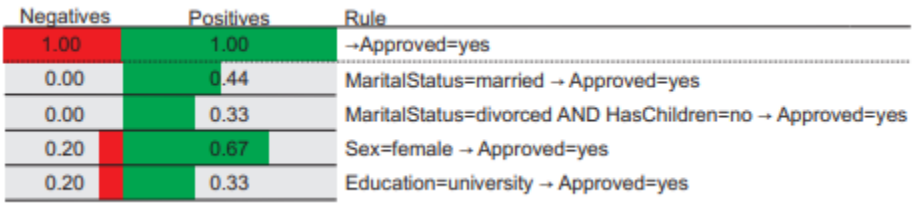


Figure 6: Example of the combination of the continuous attribute and bar chart subgroup discovery visualization techniques

4.3 Orange

Orange is machine learning and data visualization software. In this research we will use Orange for subgroup discovery and to visualize subgroup discovery results. It has a drag-and-drop interface

which makes machine learning and data visualization very accessible for people without a computer science background. We can see this in Figure 7. Suppose we want to make a plot of `Gr Liv Area` on the $x-$axis against `SalePrice` on the $y-$axis. As the figure shows, a simple sequence of clicks on the File icon and Scatterplot icon suffices. The resulting plot is displayed in Figure 8.

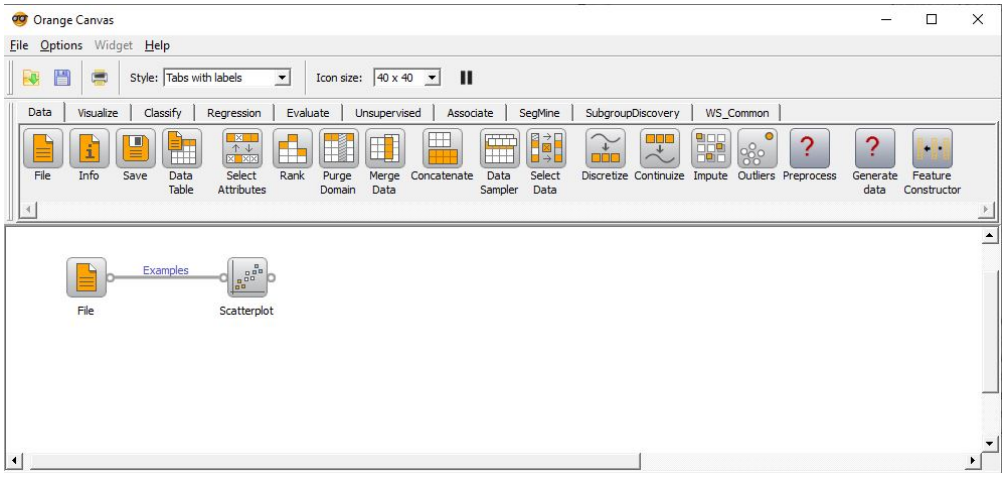


Figure 7: The simple drag-and-drop interface of `Orange`, here used to make a plot

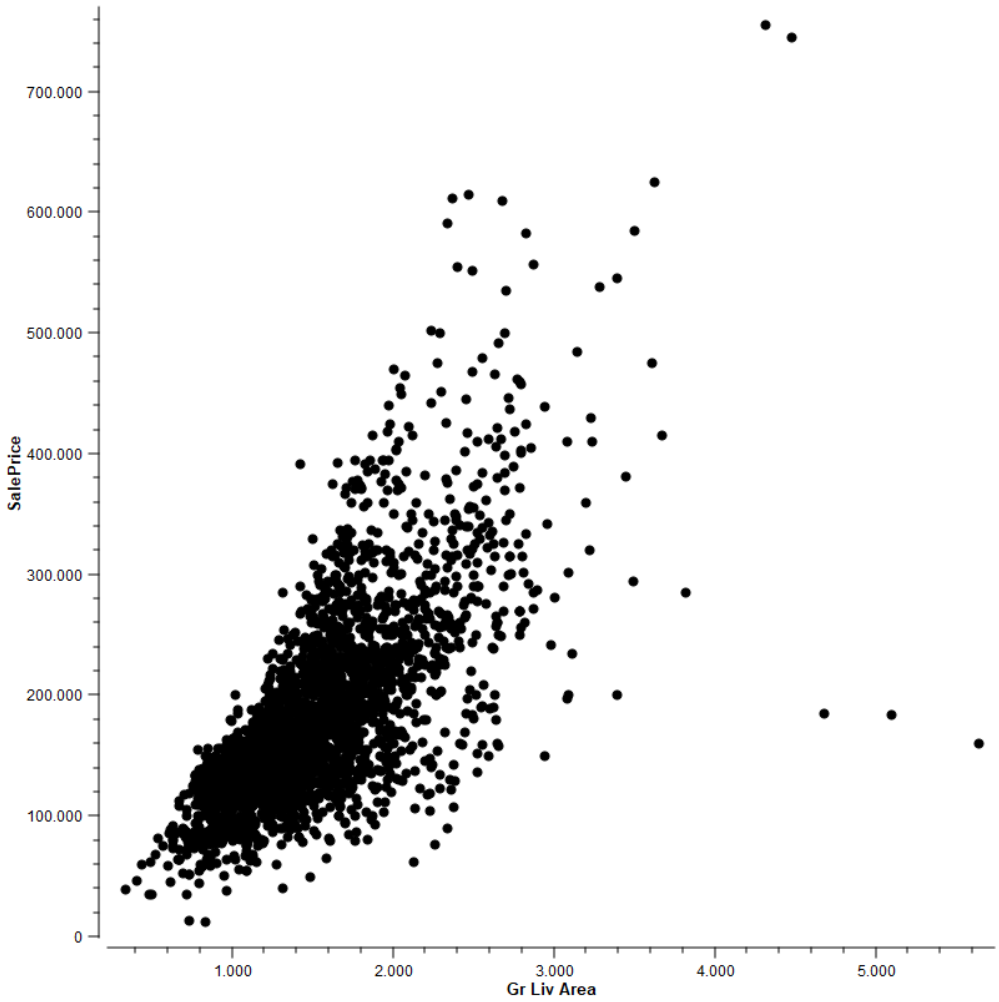


Figure 8: A plot of Gr Liv Area against SalePrice made in `Orange`

`Orange` allows users to install a subgroup discovery widget that enables options for subgroup discovery and subgroup discovery visualization. As mentioned in 4.1, the subgroup discovery tool in `Orange` does not allow for continuous variables. For this reason the discretize option in `Orange` is used to transform the continuous variables into intervals. However, the Discretize option only discretizes non-target variables, meaning the `SalePrice` has to be manually discretized in order to be able to perform subgroup discovery in `Orange`.

A long if-statement in Google Sheets determined in what interval of $\$50,000$ the `SalePrice` of a house fell. Intervals of $\$50,000$ were chosen in order to find subgroups with a sufficient coverage. Smaller intervals lead to smaller, too specific subgroups.

The drag-and-drop interface makes it easy to get subgroup discovery results as well as subgroup discovery visualizations: see Figure 9. Double-clicking 'Build Subgroups' allows us to set a number of parameters:

– The subgroup discovery algoorithm was set to CN2-SD, because it gave the best results to use

for the surveys discussed in the next section. One other algorithm did not work and results of the others were more redundant.

– The target variable was set to the discretized `SalePrice` variable.

– The value of the target variable was set to $\$100,000$ (1.10 standard deviation from the average) $-$ $\$150,000$ (0.39 standard deviation from the average). This interval of sale prices was chosen, because it covers 1016 houses of the whole data set; other intervals contained far fewer. More subgroups as well as smaller coverages make the bars in the visualization barely visible, for this research we want a visualization that helps to show the results.

– The desired number of subgroups was set to five. This allowed for diverse yet manageable results.

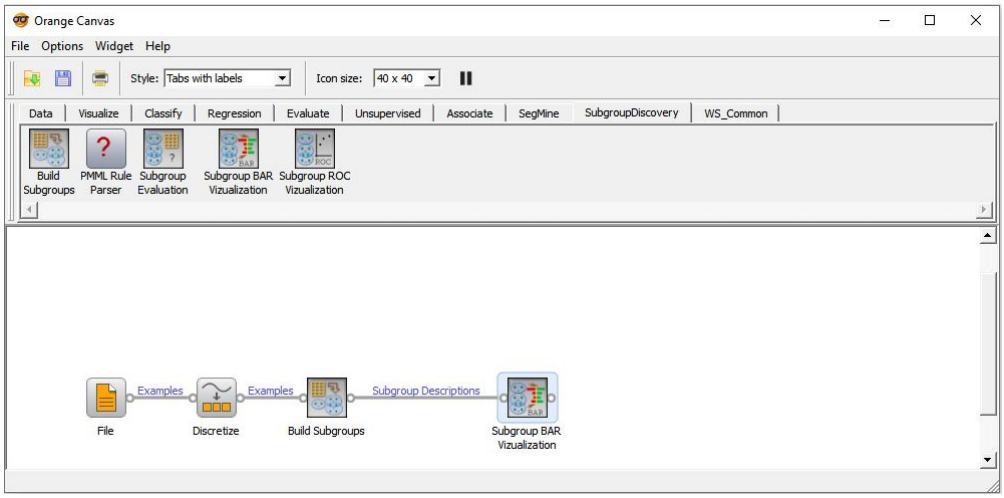


Figure 9: The flow of icons in Orange used to get subgroup discovery results and to visualize those results

5 Experiment

Section 4 explained the methods that will help to work towards answering the research question of this thesis. This section describes the experiment conducted: it explains how the results needed to answer the research question will be obtained. The way results will be acquired will be described in Section 5.1. The contents and how exactly the surveys were structured is discussed in Section 5.2. Lastly, the way these results will be analysed is explained in Section 5.3.

5.1 Surveys

In order to answer the research question, non-experts will have to be shown text-based results of subgroup discovery as well as visualized results. Each respondent will randomly be shown either the text-based results (survey A) or the visualized results (survey B). The results are shown in a survey, where some questions about the results are asked which allow the respondent to answer their understanding of the presented results.

By comparing the answers given to survey A and survey B, respondents' understanding of the results can be analysed. If, for example, respondents to survey B answer more positively on their comprehension of the results, this might indicate that visualizing the results helps to understand them. Other options would be that visualized results make the results harder to understand or that there is no significant difference.

It is important that the amount of people responding to the different surveys should be more or less equal. To achieve this, a website that redirects people fifty-fifty to the two surveys was made and used.

The surveys were set up in English, to be able to reach many people more easily. The text in the subgroup discovery results and visualization is also in English, this motivates respondents to answer in English as well. Having all answers in English, makes comparison easier.

Since surveys do not allow for follow-up questions or explanations of questions, the surveys need to be unambiguous and clear. Before sending the surveys out to as many people as possible, they were therefore sent to members of the Explanatory Data Analysis group of the Leiden Institute of Advanced Computer Science for feedback. After having received and processed feedback twice, the surveys were ready to be distributed.

5.2 Survey set up

The survey starts by asking some general demographic questions. These may be used to see whether, for example, age plays a role in the accessibility of either type of results. After that, respondents are asked for permission to use their answers for research purposes. Following this, the same set of questions about the results are asked in survey A and survey B. Below you will find the two content questions asked; the correct answers are also given here. These questions helped to give the respondents an idea of the information you can get from the results. Lastly, four questions aer

asked about how understandable and therefore accessible the results are. For each of these four questions, respondents choose one answer from the Likert scale. The Likert scale offers five possible answers: strongly agree, agree, neutral, disagree and strongly disagree. This scale was used to be able to compare easily: the answers were transformed to 1, 2, 3, 4 and 5 respectively. The questions using this scale will, from now on, be referred to as the "likert questions".

1. What houses within the data set do these results describe exactly?
   - Correct answer: Houses with a sale price between 100.000 and 150.000
2. What seem to be some important values of variables for these results?
   - Correct answers are the variables that show up more than once in the results:
     * Street = Pave
     * D_Enclosed Porch $\leq$ 16.00
     * D_Overall Qual = (4.00, 5.00]
     * Land Slope = Gtl
     * Full Bath = 1
3. I understand what the results shown at the top of this page mean.
   - No correct or incorrect answer, this question is asked to get an idea of how people themselves think they understood the results
4. The results are easy to understand.
   - No correct or incorrect answer
5. The way the results are shown in the picture is a good method to communicate the results.
   - No correct or incorrect answer
6. I would feel comfortable explaining the results to a friend.
   - No correct or incorrect answer

Figure 10 shows the text-based results that respondents to survey A will see and the visualized results used in survey B can be seen in Figure 11. As we can see from Figure 10 and Figure 11, five subgroups are displayed for houses with a sale price from \$100,000-\$150,000. The complete surveys can be found in Sections A.3 and A.4.

1: D_PID = (531475232.00, 532378240.00] D_Enclosed Porch = <=16.00 House Style = 1Story -> SalePrice = 100K-150K

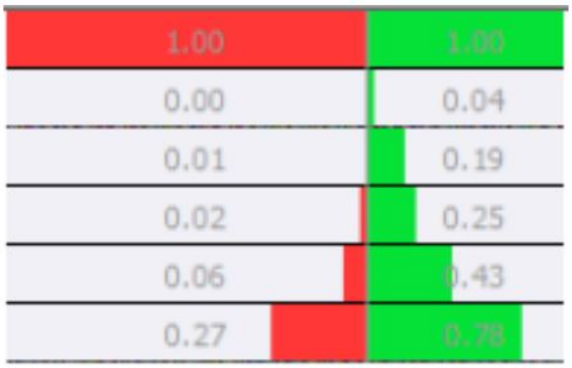
2: D_Gr Liv Area = (765.00, 987.00] D_Total Bsmt SF = (771.00, 1025.00] Paved Drive = Y Garage Qual = TA Land Slope = Gtl Street = Pave -> SalePrice = 100K-150K

3: D_Overall Qual = (4.00, 5.00] D_1st Flr SF = (811.00, 1081.00] Bsmt Cond = TA D_Enclosed Porch = <=16.00 Street = Pave -> SalePrice = 100K-150K

4: D_Overall Qual = (4.00, 5.00] Full Bath = 1 Central Air = Y Land Slope = Gtl -> SalePrice = 100K-150K

5: Full Bath = 1 Street = Pave -> SalePrice = 100K-150K

Figure 10: The regular subgroup discovery results presented in survey A



-> SalePrice = 100K-150K

1: D_PID = (531475232.00, 532378240.00] D_Enclosed Porch = <=16.00 House Style = 1Story -> SalePrice = 100K-150K

2: D_Gr Liv Area = (765.00, 987.00] D_Total Bsmt SF = (771.00, 1025.00] Paved Drive = Y Garage Qual = TA Land Slope = Gtl Street = Pave -> SalePrice = 100K-150K

3: D_Overall Qual = (4.00, 5.00] D_1st Flr SF = (811.00, 1081.00] Bsmt Cond = TA D_Enclosed Porch = <=16.00 Street = Pave -> SalePrice = 100K-150K

4: D_Overall Qual = (4.00, 5.00] Full Bath = 1 Central Air = Y Land Slope = Gtl -> SalePrice = 100K-150K

5: Full Bath = 1 Street = Pave -> SalePrice = 100K-150K

Figure 11: The visualized subgroup discovery results presented in survey B

The surveys were spread with the intention of reaching a diverse crowd: people with different fields of interest and educational amd social backgrounds. In order to get as many responses as possible, respondents were asked to forward the survey and accompanying message. Forwarding the survey was mainly done using WhatsApp, Facebook, Twitter and LinkedIn. The messages were sent multiple times, in order to remind people of the survey.

5.3 Analysis of results

The responses to the surveys were exported to a CSV-file (comma-separated-values). Selecting the same likert question from both surveys allows us to use the student t-test, a statistical test used to compare the means of two samples (McDonald, 2009), to find out the values for the following statistics:

- t-statistic: used to check whether `mean1` (the average of the answers of one of the questions of the first survey), significantly differs from `mean2` (the average of the answers of the same question of the second survey), where a large value means it does and a small value means it does not.
- Degrees of freedom: the total elements in the data − 2
- Critical value: if the absolute value of the t-statistic is bigger than the critical value (i.e. falls in the region described by the critical value), we reject the null hypothesis.
- p-value: assuming null hypothesis is correct, this value denotes the probability of getting these results. A low value is good because this means the results did not happen incidental. If the p-value is smaller or equal to alpha, we reject the null hypothesis.
- 95% confidence interval `data1` (all answers of one of the questions of the first survey): range that with a certainty of 95% contains the population average of `data1`
- 95% confidence interval `data2` (all answers of the same question of the second survey): range that with a certainty of 95% contains the population average of `data2`

All of these statistics are computed with an *alpha* of 0.05, a commonly used value for alpha (Samuels and Gilchrist, 2014).

For visualizing the results of the surveys, box plots will be used. An example of a box plot can be found in Figure 12. The box in a box plot indicates the data from the 25^{th} to the 75^{th} percentile. The horizontal line in the box represents the median. The areas from the horizontal lines outside of the box to the box indicate the minimum to the 25^{th} percentile and the 75^{th} to the maximum (Williamson et al., 1989). Showing two box plots next to each other, for example one for survey A and one for survey B, makes comparison easy. Seeing a difference in the vertical placement of the box allows for quick interpretation of the difference between the results of the two surveys. This will be useful in this research, because the vertical axis will represent the answers of the Likert scale. The placement of the box therefore indicates how confident respondents, of for example survey A compared to survey B, are with the results.

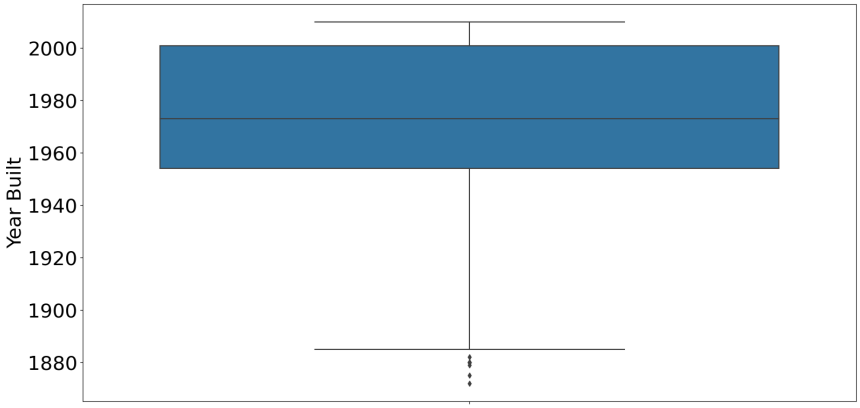


Figure 12: Box plot of the Year Built attribute in our data set

6 Results

This section will analyse the results acquired through the surveys. In this section we hope to find results that will help us to give an answer to the research question in Section 8. Recall from Subsection 5.3 that a large t-statistic score means that `mean1` significantly differs from `mean2` and that a p-value smaller or equal to the value of alpha means we reject the null hypothesis.

6.1 Content questions

This subsection will analyse the respondents' answers to the content questions. It should give an insight into whether respondents managed to understand the results presented, as these questions had correct answers, unlike the likert questions. These questions are listed in Subsection 5.2.

The first content question asked which houses within the data set are described by the results shown to the respondent. In survey A, 52 of the 63 respondents answered this question correct and in survey B, 46 of the 58 respondents did. Combined, this means that 98 of the 121 respondents answered the first content question correct, which equals 81%.

Table 3 shows the five correct answers for the second content question in the first column. The second column details in how many subgroup descriptions the relevant attribute occurs. The remaining four columns show the number of times each correct answer was chosen by the respondents.

Note that respondents identified `Street` = `Pave` as important most times; it is also the answer that showed up most in the subgroup descriptions. All other answers showed up twice in the descriptions, but surprisingly, `Land Slope` = `Gtl` was chosen at least 13% less than the other answers. A third interesting find, that is not displayed in this table, is that nine respondents to survey A and eight respondents to survey B gave the complete correct answer, i.e. correctly identified all five important characteristics. This is equivalent to 14%. Not only computer science experienced people managed to give the full correct answer: in survey A, three of the nine respondents who gave the full correct answer have no computer science experience; of the eight respondents to survey B who gave the complete correct answer, three have no computer science experience.

Correct answers	Occurrences	Survey A	Survey B	Total	Percentage
Street = Pave	3	35	42	77	64%
D_Enclosed Porch $\leq$ 16.00	2	34	18	52	43%
D_Overall Qual = (4.00, 5.00]	2	43	24	67	55%
Land Slope = Gtl	2	18	18	36	30%
Full Bath = 1	2	38	38	76	63%

Table 3: Table showing how many times each correct answer was answered per survey and in total

6.2 The understandability of the results shown in both surveys

Box plots of the respondents' answers to the likert questions for both surveys are shown in Figure 13 are displayed. Below the box plots, in Table 4, the results of the student t-test for each likert question are shown.

Here, the following null hypothesis and alternative hypothesis will be used to test against:

H_0: `mean1` does not significantly differ from `mean2`
H_a: `mean1` significantly differs from `mean2`.

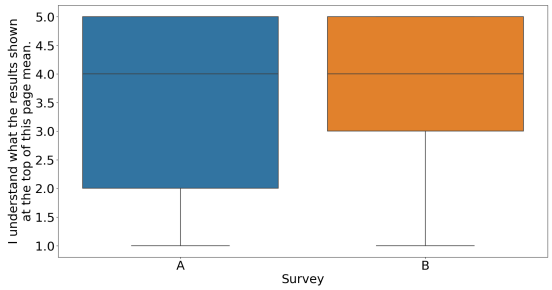
In these hypotheses, `mean1` describes the average value of a likert question of survey A, as the Likert scale answer was transformed to a number. The same but for survey B is what `mean2` indicates. Also, recall the four likert questions asked in both surveys:

- `Likert question 1`: I understand what the results shown at the top of this page mean.
- `Likert question 2`: The results are easy to understand.
- `Likert question 3`: The way the results are shown in the picture is a good method to communicate the results
- `Likert question 4`: I would feel comfortable explaining the results to a friend.

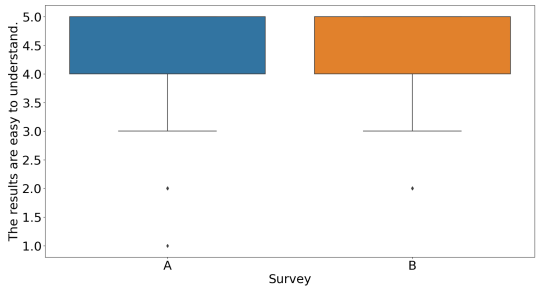
Intuitively, the null hypothesis states that there is no noticeable difference between the accessibility of survey A and survey B, i.e. bar visualization does not make subgroup discovery results more accessible. The alternative hypothesis then indicates that there is a difference.

Taking a look at the box plots first, the results between the two surveys do not seem to be very different; the box plots of survey A and survey B for each question are on a similar height. The height in these plots indicate the answers on the Likert scale to the question. As we know, a 5 means a respondent answered "Strongly disagree" and a 1 means a respondent answered "Strongly agree". The integers in between represent the other answers from the Likert scale. Note that the plots also indicate that respondents generally disagreed with the likert questions.

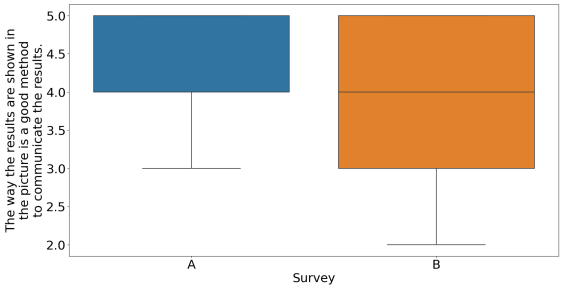
This first impression given by the box plots, is supported by the results of the student t-test shown in Table 4. We see that the 95% confidence interval for each likert question does not lie very far apart when comparing the two surveys. Adding on to that, for each likert question the p-value exceeds the alpha of 0.05. This means that we do not reject the null hypothesis, which suggests that a visualization does not significantly help to understand the subgroup discovery results.



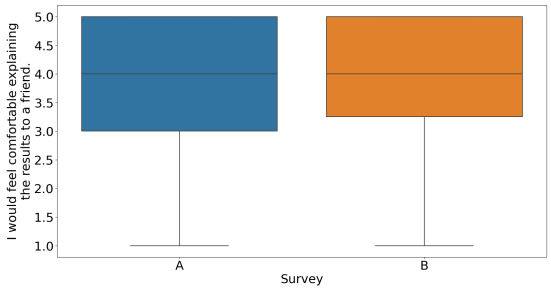
(a) First likert question: I understand what the results shown at the top of this page mean.



(b) Second likert question: The results are easy to understand.



(c) Third likert question: The way the results are shown in the picture is a good method to communicate the results.



(d) Fourth likert question: I would feel comfortable explaining the results to a friend.

Figure 13: Box plots of each likert question per survey, where a 1 indicates Strongly agree as the answer

Likert question	**p-value**	**95% confidence interval A**	**95% confidence interval B**
1	0.327	3.22 - 3.86	3.45 - 4.07
2	0.543	3.83 - 4.33	3.93 - 4.45
3	0.304	4.07 - 4.44	3.82 - 4.35
4	0.747	3.70 - 4.24	3.73 - 4.34

Table 4: Results of the student t-test on survey A and survey B for each likert question

6.3 The influence of computer science experience on the understandability of the results

In Section 6.2 we found that a visualization does not significantly help to understand subgroup discovery results. It might be interesting to see if there are other variables that do have an influence on the understandability of the results, i.e. what does help to understand the results. In this section, analysis is done on whether the respondents' experience in the computer science field makes a difference.

Again, we use two hypotheses to test against:

H_0: `mean1` does not significantly differ from `mean2`
H_a: `mean1` significantly differs from `mean2`.

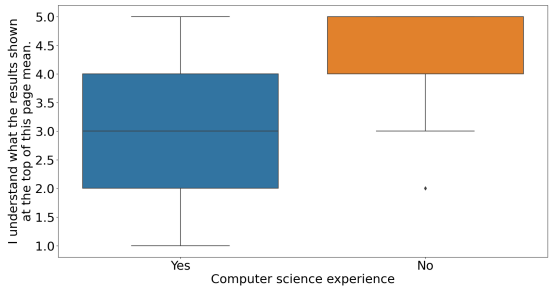
Here, `mean1` describes the average value of a likert question of respondents for respondents without computer science experience. The same but for respondents with computer science experience is what `mean2` indicates. The same four likert questions are used for comparison:

- `Likert question 1`: I understand what the results shown at the top of this page mean.
- `Likert question 2`: The results are easy to understand.
- `Likert question 3`: The way the results are shown in the picture is a good method to communicate the results
- `Likert question 4`: I would feel comfortable explaining the results to a friend.

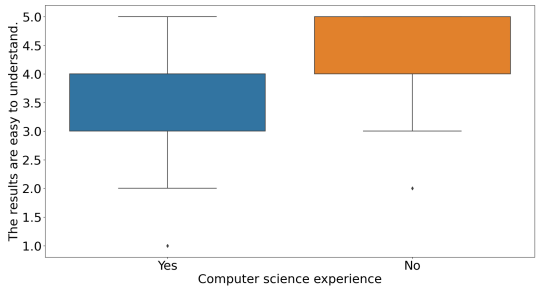
Once again, the null hypothesis states that there is no noticeable, this time between the accessibility for computer science experienced respondents and non-experts. The alternative hypothesis indicates that there is a difference.

Taking a look at the box plots in Figure 14, the first impression is that respondents with computer science experience tend to answer more towards "Strongly agree". The blue boxes, representing the answers of respondents with computer science experience, all seem to be closer to the 1.0−value on the $y-$axis.

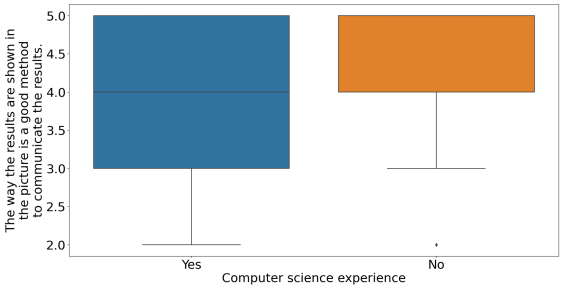
From Table 5, it is clear that $p - value < alpha$ holds for each likert question. This means that the null hypothesis is rejected. In turn, this implies that `mean1` is significantly different from `mean2`, indicating that experience in the computer science field does help to understand the results.



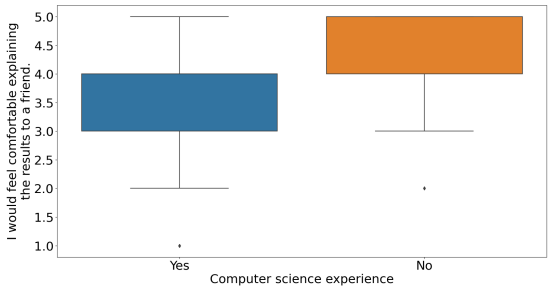
(a) First likert question: I understand what the results shown at the top of this page mean.



(b) Second likert question: The results are easy to understand.



(c) Third likert question: The way the results are shown in the picture is a good method to communicate the results.



(d) Fourth likert question: I would feel comfortable explaining the results to a friend.

Figure 14: Box plots of each likert question split on respondents' computer science experience

Likert question	**p-value**	**95% c.i. CS-experience**	**95% c.i. no CS-experience**
1	0.000	2.56 - 3.14	4.05 - 4.52
2	0.000	3.32 - 3.90	4.39 - 4.72
3	0.000	3.60 - 4.10	4.25 - 4.62
4	0.000	3.13 - 3.80	4.24 - 4.63

Table 5: Results of the student t-test on for each likert question with a split on computer science experience

7 Discussion

This section describes some limitations of the research as well as possible suggestions for further research. This thesis has provided interesting new insights, but there is much more to find out in this area of research.

7.1 The possible downsides of using surveys

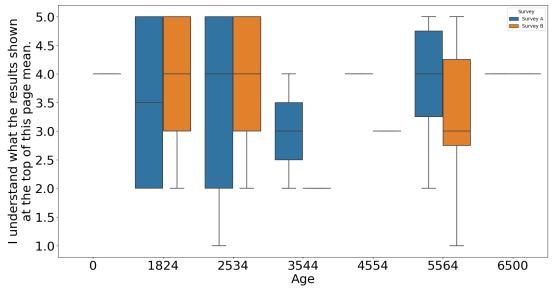
One of the difficulties with using surveys is that respondents might prematurely quit the surveys. This can have several causes. If the survey takes too long − or longer than the respondent expected − respondents might stop filling out the survey. The survey used in this research was said to take about fifteen minutes. This might already be too much time for some respondents.

Secondly, difficult questions could possibly scare respondents away. The two content questions that were used in this survey may have led to this. Unfortunately, it will remain unclear how many people may have prematurely quit the surveys. Having more respondents could lead to different and possibly more grounded results. Another possible complication of the use of surveys is the fact that communication takes place only through text. Therefore, there is no way for respondents to ask questions to clear up potential confusion. This might result in questions being interpreted differently than they are supposed to.

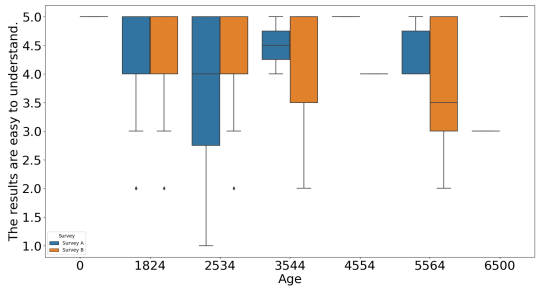
7.2 The influence of age

The age groups work as follows: age group "1824" represents the respondents who are 18-24 years old. It is important to note that there is an age group "0": this is the group of people who would prefer not to say their age. Box plots for all age groups per survey are displayed in Figure 15.

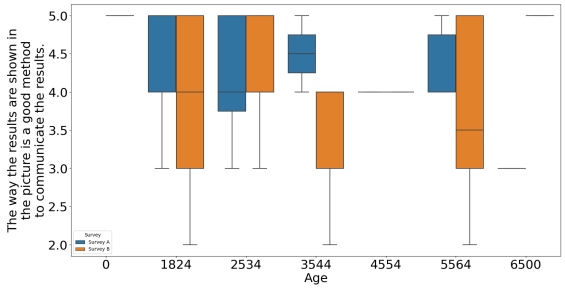
Age group "0" only has one respondent, which is not enough to draw conclusions from. For age group "1824" − with sixty-five respondents − there is not a clear difference between the answers given in survey A and survey B. In age group "2534" (twenty-seven respondents) we see that survey A seems to score slightly better, the difference is, however, minimal and probably does not say much. For age group "3544", survey B scores a little better, again this difference is minimal and it only contains five respondents, making it too small of a sample to be able to make well-grounded remarks on. The same applies to the age group "4554" with only three respondents. A small difference in favor of survey B can be seen in the age group "5564", which consists of eighteen respondents. With only two respondents, age group "6500" is again too small to tell us anything. While age may have an influence, the number of respondents per age group in this survey is too small to be able to confidently conclude anything.



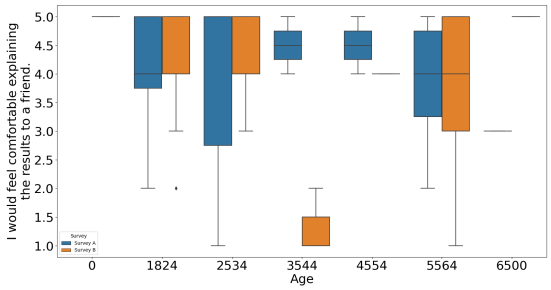
(a) First likert question: I understand what the results shown at the top of this page mean.



(b) Second likert question: The results are easy to understand.



(c) Third likert question: The way the results are shown in the picture is a good method to communicate the results.



(d) Fourth likert question: I would feel comfortable explaining the results to a friend.

Figure 15: Box plots of respondents' age against each likert question for each survey

7.3 Further research

The research can be improved and expanded in several ways. Doing this might lead to more and better insights into the research question. Below, improvements to the research will be described. Improvements can possibly make the research more complete and grounded. Following that, ways to expand the research are discussed. Expansions to the research are ways to build forward from this thesis.

By conducting this type of research on more data sets, one could possibly eliminate inconsistencies or parts that are unclear in the one data set used in this research. Additionally, a larger number of participants in the surveys may improve accuracy of the study: with the analysis of the responses to our surveys, it was found that some age groups had very few respondents. This made it impossible to say anything relevant about the influence of age, which may actually have an influence on the research question.

As an expansion, it would be interesting to use different result sets in the research: this thesis only used the results of one subgroup discovery experiment. One might find that visualizations could help for certain types of results, but not for others. Different result sets could, for example, be result sets with a different value for the target variable; with a different number of conditions describing the

subgroups or by having more subgroups in the result set. The same could be interesting for different types of data sets. Lastly, from the results it seems the bar visualization is not a final solution to making subgroup discovery accessible for everyone. Asking respondents for feedback about the results might make it possible to suggest some improvements or expansions to the bar visualization. With this information, it might be possible to enhance the accessibility of the visualization. Adding on to this, it would also be interesting to research how respondents would experience the other visualization methods discussed in Subsection 4.2.

8 Conclusion

This research aimed to get an insight into the accessibility of subgroup discovery results. The results of a subgroup discovery experiment can be very valuable to all people, as they can provide interesting information about all kinds of data. The results can, however, be quite hard to understand. The objective was to find out whether the bar visualization method for visualizing subgroup discovery results made the results more accessible to non-experts. The subgroups used in this research were results from performing subgroup discovery on the Ames housing data set.

Overall, this research found that regardless the way the subgroups discovery results were presented, text-based or through bar visualization, they are hard to understand for the respondents. The answers to the two surveys were not significantly different; this means that for this research, visualizing the subgroups did not make them much more clear. The accessibility has therefore not been improved through bar visualization in this thesis.

This could have several causes. Possibly, the concept of the subgroup descriptions is too difficult to grasp if you have no computer science experience, and therefore a visualization also makes no significant improvement. Another cause could be that the visualization does not provide a deeper explanation of the results, but only gives extra information. It was found that, in general, respondents with computer science experience actually were more comfortable interpreting the results. In order to make results more accessible for non-experts, more than just changing the method used to communicate the results needs to happen. Possibly, the improvements and expansions suggested in Subsection 7.3 could help with this.

References

Paul M. Anglin and Ramazan Gençay. Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, 11(6):633–648, 1996. ISSN 08837252, 10991255. URL http://www.jstor.org/stable/2285156.

Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19, 11 2011. doi: 10.1080/10691898.2011.11889627.

Arie-Willem de Leeuw, Aldo Hoekstra, Laurentius Meerhoff, and Arno Knobbe. Tactical analyses in professional tennis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 258–269. Springer, 2019.

Dragan Gamberger, Nada Lavrac, and Dietrich Wettschereck. Subgroup visualization: A method and application in population screening. In *Proceedings of the 7th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-02)*, pages 31–35, 2002.

Dragan Gamberger, Nada Lavrač, and Goran Krstačić. Active subgroup mining: A case study in coronary heart disease risk group detection. *Artif. Intell. Med.*, 28(1):27–57, May 2003. ISSN 0933-3657. doi: 10.1016/S0933-3657(03)00034-4. URL https://doi.org/10.1016/S0933-3657(03)00034-4.

John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007(2012):1–16, 2012.

David Harrison and Daniel Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 03 1978. doi: 10.1016/0095-0696(78)90006-2.

Sumyea Helal. Subgroup discovery algorithms: a survey and empirical evaluation. *Journal of Computer Science and Technology*, 31(3):561–576, 2016.

Franciso Herrera, Cristóbal José Carmona, Pedro González, and María José del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, 29:495–525, 12 2011. doi: 10.1007/s10115-010-0356-2.

Kaggle. House prices: Advanced regression techniques | kaggle. https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview. (Accessed on 07/18/2020).

Petra Kralj, Nada Lavrač, and Blaz Zupan. Subgroup visualization. In *8th International Multiconference Information Society (IS-05)*, pages 228–231. Citeseer, 2005.

Daniel Lambach and Dragan Gamberger. Temporal analysis of political instability through descriptive subgroup discovery. *Conflict Management and Peace Science - CONFLICT MANAG PEACE SCI*, 25:19–32, 03 2008. doi: 10.1080/07388940701860359.

Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupčo Todorovski. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, 5(Feb):153–188, 2004a.

Nada Lavrač, Branko Kavšek, Peter Flach, and Ljupco Todorovski. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, 5:153–188, 02 2004b.

Dennis Leman, Ad Feelders, and Arno Knobbe. Exceptional model mining. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 1–16. Springer, 2008.

T. Lim, W. Loh, and Y. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40:203–228, 09 2000. doi: https://doi.org/10.1023/A:1007608224229.

John H McDonald. *Handbook of biological statistics*, volume 2. sparky house publishing Baltimore, MD, 2009.

Petra Kralj Novak, Nada Lavrač;, and Geoffrey I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10(14):377–403, 2009. URL http://jmlr.org/papers/v10/kralj-novak09a.html.

Byeonghwa Park and Jae Kwon Bae. Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert Systems with Applications*, 42(6):2928–2934, 2015.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Peter Samuels and Mollie Gilchrist. Statistical hypothesis testing, 04 2014.

Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381:520–2, 07 1996. doi: 10.1038/381520a0.

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017. URL https://doi.org/10.5281/zenodo.883859.

David F Williamson, Robert A Parker, and Juliette S Kendrick. The box plot: a simple visual method to interpret data. *Annals of internal medicine*, 110(11):916–921, 1989.

A Appendix

A.1 Ames housing variable description

See Table 6 below for a full description of all variables found in the Ames housing data set (Kaggle).

Table 6: A table introducing each variable from the Ames housing data set with their corresponding description

Data description table	
Variable	**Description**
MSSubClass	The building class
MSZoning	The general zoning classification
LotFrontage	Linear feet of street connected to property
LotArea	Lot size in square feet
Street	Type of road access
Alley	Type of alley access
LotShape	General shape of property
LandContour	Flatness of the property
Utilities	Type of utilities available
LotConfig	Lot configuration
LandSlope	Slope of property
Neighborhood	Physical locations within Ames city limits
Condition1	Proximity to main road or railroad
Condition2	Proximity to main road or railroad (if a second is present)
BldgType	Type of dwelling
HouseStyle	Style of dwelling
OverallQual	Overall material and finish quality
OverallCond	Overall condition rating
YearBuilt	Original construction date
YearRemodAdd	Remodel date
RoofStyle	Type of roof
RoofMatl	Roof material
Exterior1st	Exterior covering on house
Exterior2nd	Exterior covering on house (if more than one material)
MasVnrType	Masonry veneer type
MasVnrArea	Masonry veneer area in square feet
ExterQual	Exterior material quality
ExterCond	Present condition of the material on the exterior
Foundation	Type of foundation
BsmtQual	Height of the basement
BsmtCond	General condition of the basement
BsmtExposure	Walkout or garden level basement walls
BsmtFinType1	Quality of basement finished area
BsmtFinSF1	Type 1 finished square feet
BsmtFinType2	Quality of second finished area (if present)

Continuation of Table 6	
Variable	**Description**
BsmtFinSF2	Type 2 finished square feet
BsmtUnfSF	Unfinished square feet of basement area
TotalBsmtSF	Total square feet of basement area
Heating	Type of heating
HeatingQC	Heating quality and condition
CentralAir	Central air conditioning
Electrical	Electrical system
1stFlrSF	First Floor square feet
2ndFlrSF	Second floor square feet
LowQualFinSF	Low quality finished square feet (all floors)
GrLivArea	Above grade (ground) living area square feet
BsmtFullBath	Basement full bathrooms
BsmtHalfBath	Basement half bathrooms
FullBath	Full bathrooms above grade
HalfBath	Half baths above grade
Bedroom	Number of bedrooms above basement level
Kitchen	Number of kitchens
KitchenQual	Kitchen quality
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)
Functional	Home functionality rating
Fireplaces	Number of fireplaces
FireplaceQu	Fireplace quality
GarageType	Garage location
GarageYrBlt	Year garage was built
GarageFinish	Interior finish of the garage
GarageCars	Size of garage in car capacity
GarageArea	Size of garage in square feet
GarageQual	Garage quality
GarageCond	Garage condition
PavedDrive	Paved driveway
WoodDeckSF	Wood deck area in square feet
OpenPorchSF	Open porch area in square feet
EnclosedPorch	Enclosed porch area in square feet
3SsnPorch	Three season porch area in square feet
ScreenPorch	Screen porch area in square feet
PoolArea	Pool area in square feet
PoolQC	Pool quality
Fence	Fence quality
MiscFeature	Miscellaneous feature not covered in other categories
MiscVal	$Value of miscellaneous feature
MoSold	Month Sold
YrSold	Year Sold
SaleType	Type of sale

Continuation of Table 6	
Variable	**Description**
SaleCondition	Condition of sale
SalePrice	The property's sale price in dollars.
End of Table	

A.2 Cortana

Developed by Leiden University, Cortana is free to use software that can, among other functionalities, be used for subgroup discovery. For this research, Cortana has been used extensively to get familiar with subgroup discovery and the results it produces. It helped to know what results look like and what parameters have big effects on the results. Before starting experiments, it is important to have sufficient knowledge, Cortana has helped with this part of the research.

The first thing to do when opening Cortana is to choose a data set. Then, the screen shown in Figure 16 is presented. This interface shows four tabs, each of these tabs may be used to change some settings for the experiment to try to get the best results. The most influential settings that can be adjusted will be explained below.

- **Dataset:** this tab shows some information about the chosen data set. The browse button is used to (de)select variables to be used for the experiment.
- **Target concept:** this tab allows to set the target variable. The target type specifies what kind of variable the target variable is, in the case of Figure 16: single numeric. Another option could, for example, be single nominal which are variables that can only take certain values from a category. For example: the style of a house (one, two or three stories). You also choose the quality measure you wish to use. The quality measure is the base for the ranking of the subgroups in the results. Not all subgroups score high on every quality measure, making this an influential parameter. Finally, the target variable has to be chosen in the primary target option.
- **Search Conditions:** refinement depth states the maximum amount of conditions a subgroup can have. The first example from Table 1 above has only one condition: the amount of horsepower. The second example uses two conditions to describe the subgroup: the number of doors and the color. As refinement depth increases, more variables can be used for the description of a subgroup. This means that more specific subgroups can be described. This results in an increase in the score of the quality measure scores, but subgroups being more specific also leads to a decrease in their size.
- **Search Strategy:** the settings in this tab are used to choose how to analyse the data set. The strategy type specifies the way Cortana searches in the data: depth first is an exhaustive search method, an example of a heuristic search method is beam search, which is another option for this setting. The choice of the numeric strategy decides how numeric variables are dealt with. This strategy can for example be set to a chosen number of bins, this will split the numeric space of the variable in a chosen amount of groups.

The subgroup discovery button will run the experiment with the chosen settings. To give an idea of the way this looks, the top 15 results from the experiment with the settings as seen in Figure 16 are shown in Figure 17.

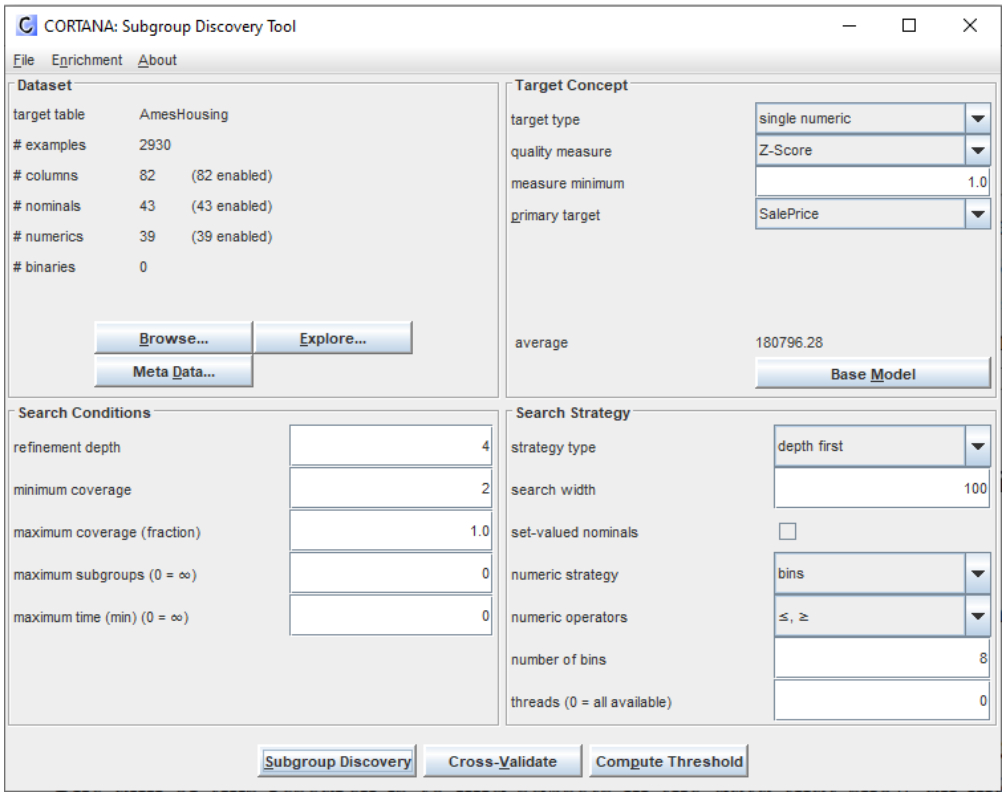


Figure 16: The interface of Cortana after having chosen the data set

15168 subgroups found; target table = AmesHousing; quality measure = Z-Score

Nr.	Depth	Coverage	Quality	Average	St. Dev.	p-Value	Conditions
1	4	328	35,683834	338.197,756098	88.723,88951	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1478.0 AND 1st Flr SF >= 1086.0 AND Gr Liv Area >= 1592.0
2	4	319	35,663586	340.312,125392	88.908,361531	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND Paved Drive = 'Y' AND 1st Flr SF >= 1082.0
3	4	320	35,659042	340.042,4	88.855,64642	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND 1st Flr SF >= 1080.0 AND Half Bath <= 1.0
4	4	320	35,659042	340.042,4	88.855,64642	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND Half Bath <= 1.0 AND 1st Flr SF >= 1080.0
5	4	319	35,655174	340.274,507837	88.948,272509	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND Central Air = 'Y' AND 1st Flr SF >= 1086.0
6	4	319	35,655174	340.274,507837	88.948,272509	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND Garage Cars >= 2.0 AND 1st Flr SF >= 1086.0
7	4	320	35,649246	339.998,65	88.945,734684	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND Heating = 'GasA' AND 1st Flr SF >= 1082.0
8	4	320	35,635349	339.936,6	89.061,43512	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND 1st Flr SF >= 1080.0 AND Misc Feature = 'NA'
9	4	320	35,635349	339.936,6	89.061,43512	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND 1st Flr SF >= 1080.0 AND Misc Val <= 0.0
10	4	320	35,635349	339.936,6	89.061,43512	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND Misc Feature = 'NA' AND 1st Flr SF >= 1080.0
11	4	320	35,635349	339.936,6	89.061,43512	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND Misc Val <= 0.0 AND 1st Flr SF >= 1080.0
12	4	357	35,632927	331.454,072829	88.320,298266	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND Alley = 'NA' AND Half Bath <= 1.0
13	4	357	35,632927	331.454,072829	88.320,298266	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1616.0 AND Half Bath <= 1.0 AND Alley = 'NA'
14	4	362	35,625145	330.377,325967	88.440,477041	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1478.0 AND Alley = 'NA' AND Gr Liv Area >= 1602.0
15	4	328	35,6161	337.898,97561	89.116,070711	-	Overall Qual >= 8.0 AND Gr Liv Area >= 1478.0 AND Gr Liv Area >= 1595.0 AND 1st Flr SF >= 1080.0

Figure 17: Results of the subgroup discovery experiment from Figure 16

A.3 Survey A

Survey introduction

Dear respondent,

After some introductory questions, this survey will ask you some questions about results of a subgroup discovery experiment. Subgroup discovery is a data mining technique that can find subsets in a data set with similar and interesting behavior towards the target variable. The data set used in this experiment is a data set that show the sales of houses. Each row describes one house and its attributes. Different columns describe different attributes of the houses, for example the number of cars that fit in the house's garage or the number of kitchens the house has.

An example result could be: houses with a four car garage and two kitchens have a price between €300.000 and €350.000, while the average house price is €150.000. In the format used in the picture in this survey this would look as follows: Garage Cars = 4 Kitchens = 2 -> SalePrice = 300000-350000.

The survey should take about fifteen minutes. Thanks a lot for your participation!

Please only fill out this survey once.
* Required

1. I give permission to use my answers for research. *

   Mark only one oval.

   - ◯ Yes
   - ◯ No

2. What is your age? *

   Mark only one oval.

   - ◯ 0-17 years old
   - ◯ 18-24 years old
   - ◯ 25-34 years old
   - ◯ 35-44 years old
   - ◯ 45-54 years old
   - ◯ 55-64 years old
   - ◯ 65+ years old
   - ◯ I would prefer not to say

3. What is your gender? *

Mark only one oval.

- ◯ Female
- ◯ Male
- ◯ Other
- ◯ I would prefer not to say

4. Do you have experience in the computer science field? This also includes (self) study as well as job(s). *

Mark only one oval.

- ◯ Yes
- ◯ No

Content questions

The following questions concern the five subgroups shown in the picture below, as obtained with subgroup discovery. Each subgroup first lists the features determining the houses in the subgroup, and ends with the sale price category of those houses. Most subgroups contain positive and negative examples. Positive examples are houses in a subgroup that follow the description AND have the chosen value for the target variable, while negative examples are houses that follow the description but fall in a different price category.

The results are found in the picture below

1: D_PID = (531475232.00, 532378240.00] D_Enclosed Porch = <=16.00 House Style = 1Story -> SalePrice = 100K-150K

2: D_Gr Liv Area = (765.00, 987.00] D_Total Bsmt SF = (771.00, 1025.00] Paved Drive = Y Garage Qual = TA Land Slope = Gtl Street = Pave -> SalePrice = 100K-150K

3: D_Overall Qual = (4.00, 5.00] D_1st Flr SF = (811.00, 1081.00] Bsmt Cond = TA D_Enclosed Porch = <=16.00 Street = Pave -> SalePrice = 100K-150K

4: D_Overall Qual = (4.00, 5.00] Full Bath = 1 Central Air = Y Land Slope = Gtl -> SalePrice = 100K-150K

5: Full Bath = 1 Street = Pave -> SalePrice = 100K-150K

5. What houses within the data set do these results describe exactly? *

Mark only one oval.

- ◯ All houses
- ◯ Houses with a sale price between 100.000 and 150.000
- ◯ Houses with an D_Overall Qual(ity) of a 4 or a 5
- ◯ Houses with one Full Bath

6. What seem to be some important values of variables for these results? *

More than one answer can be correct.

Check all that apply.

- ☐ Street = Pave
- ☐ D_Enclosed Porch <= 16.00
- ☐ Year Built >= 1980
- ☐ House Style = 2Story
- ☐ D_Overall Qual = (4.00, 5.00]
- ☐ Pool Area >= 0
- ☐ Land Slope = Gtl
- ☐ Bedroom AbvGr <= 3
- ☐ Full Bath = 1
- ☐ Kitchen Qual = Gd

7. I understand what the results shown at the top of this page mean. *

Mark only one oval.

- ◯ Strongly agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly disagree

8. The results are easy to understand. *

Mark only one oval.

- ◯ Strongly agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly disagree

9. The way the results are shown in the picture is a good method to communicate the results. *

Mark only one oval.

- ◯ Strongly agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly disagree

10. I would feel comfortable explaining the results to a friend. *

Mark only one oval.

- ◯ Strongly agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly disagree

A.4 Survey B

Survey introduction

Dear respondent,

After some introductory questions, this survey will ask you some questions about results of a subgroup discovery experiment. Subgroup discovery is a data mining technique that can find subsets in a data set with similar and interesting behavior towards the target variable. The data set used in this experiment is a data set that show the sales of houses. Each row describes one house and its attributes. Different columns describe different attributes of the houses, for example the number of cars that fit in the house's garage or the number of kitchens the house has.

An example result could be: houses with a four car garage and two kitchens have a price between €300.000 and €350.000, while the average house price is €150.000. In the format used in the picture in this survey this would look as follows: Garage Cars = 4 Kitchens = 2 -> SalePrice = 300000-350000.

The survey should take about fifteen minutes. Thanks a lot for your participation!

Please only fill out this survey once.
* Required

1. I give permission to use my answers for research. *

   Mark only one oval.

   - ◯ Yes
   - ◯ No

2. What is your age? *

   Mark only one oval.

   - ◯ 0-17 years old
   - ◯ 18-24 years old
   - ◯ 25-34 years old
   - ◯ 35-44 years old
   - ◯ 45-54 years old
   - ◯ 55-64 years old
   - ◯ 65+ years old
   - ◯ I would prefer not to say

3. What is your gender? *

Mark only one oval.

- ◯ Female
- ◯ Male
- ◯ Other
- ◯ I would prefer not to say

4. Do you have experience in the computer science field? This also includes (self) study as well as job(s). *

Mark only one oval.

- ◯ Yes
- ◯ No

Content questions

The following questions concern the five subgroups shown in the picture below, as obtained with subgroup discovery. Each subgroup first lists the features determining the houses in the subgroup, and ends with the sale price category of those houses. Most subgroups contain positive and negative examples. Positive examples are houses in a subgroup that follow the description AND have the chosen value for the target variable, while negative examples are houses that follow the description but fall in a different price category. Positive examples are displayed by the green bars, while the red bars represent negative examples; each of rows 2-6 corresponds to one of the five subgroups, in the same order (the top row corresponds to the complete dataset).

The results are found in the picture below

The numbers of the red bars from top to bottom: 1.00 - 0.00 - 0.01 - 0.02 - 0.06 - 0.27 and the numbers of the green bars from top to bottom: 1.00 - 0.04 - 0.19 - 0.25 - 0.43 - 0.78

-> SalePrice = 100K-150K

1: D_PID = (531475232.00, 532378240.00] D_Enclosed Porch = <=16.00 House Style = 1Story -> SalePrice = 100K-150K

2: D_Gr Liv Area = (765.00, 987.00] D_Total Bsmt SF = (771.00, 1025.00] Paved Drive = Y Garage Qual = TA Land Slope = Gtl Street = Pave -> SalePrice = 100K-150K

3: D_Overall Qual = (4.00, 5.00] D_1st Flr SF = (811.00, 1081.00] Bsmt Cond = TA D_Enclosed Porch = <=16.00 Street = Pave -> SalePrice = 100K-150K

4: D_Overall Qual = (4.00, 5.00] Full Bath = 1 Central Air = Y Land Slope = Gtl -> SalePrice = 100K-150K

5: Full Bath = 1 Street = Pave -> SalePrice = 100K-150K

5. What houses within the data set do these results describe exactly? *

Mark only one oval.

- ◯ All houses
- ◯ Houses with a sale price between 100.000 and 150.000
- ◯ Houses with an D_Overall Qual(ity) of a 4 or a 5
- ◯ Houses with one Full Bath

6. What seem to be some important values of variables for these results? *

More than one answer can be correct.

Check all that apply.

- [ ] Street = Pave
- [ ] D_Enclosed Porch <= 16.00
- [ ] Year Built >= 1980
- [ ] House Style = 2Story
- [ ] D_Overall Qual = (4.00, 5.00]
- [ ] Pool Area >= 0
- [ ] Land Slope = Gtl
- [ ] Bedroom AbvGr <= 3
- [ ] Full Bath = 1
- [ ] Kitchen Qual = Gd

7. I understand what the results shown at the top of this page mean. *

Mark only one oval.

- ◯ Strongly agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly disagree

8. The results are easy to understand. *

Mark only one oval.

- ◯ Strongly agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly disagree

9. The way the results are shown in the picture is a good method to communicate the results. *

Mark only one oval.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

10. I would feel comfortable explaining the results to a friend. *

Mark only one oval.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree