Assignment 2

Part 1 Classification

Data

The data set contains 156 observations of 25 variables. These variables are: age, blood pressure, gravity, albumin, sugar, red blood, pus_cell, pus_clumps, bacteria, glucose, urea, serum, sodium, potassium, hemoglobin, white blood, red blood count, cell volume, hypertension, diabeter, coronary atery, appetite, pedal edema, anemia, and class. Out of these we are only interested in age, blood pressure, glucose, coronary atery, and class. All the variables that will be used for the classification are deffined as int meaning numeric values. For the Class variable will have to be changed into a factor. For each patient, the dataset contains information on the subject's blood pressure, age, glucose and coronary artery disease, as well as information on whether the person has chronic kidney disease. The averages of each explanatory values were: Age=49.4, Blood Pressure= 73.8, and Glucose=131. The smallest values were: Age= 6, Blood pressure=50, and Glucose = 70. And the maximum values were: Age=83, Blood Pressure=110, and Glucose=490. From figure 1, we can see that these findings are consistent with the data. Age seems to be randomly distributed, with a larger amount of the patients being 40 or over. Most of the patients had blood pressure of 80 or under and glucose levels under 150. From figures 2 and 3 it can be noticed, that most of the patients belonged to the group without the coronary artery disease or chronic kidney disease.

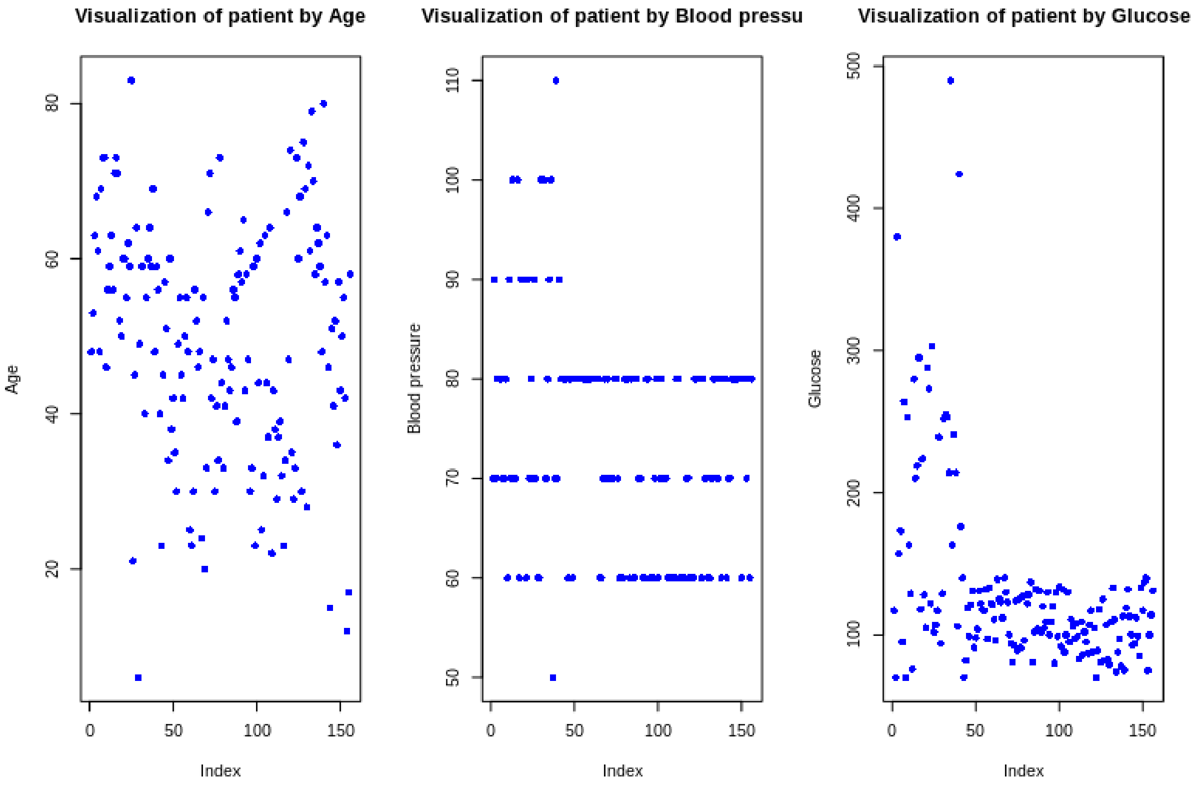


Figure 1 Basic plot of all explanatory variables

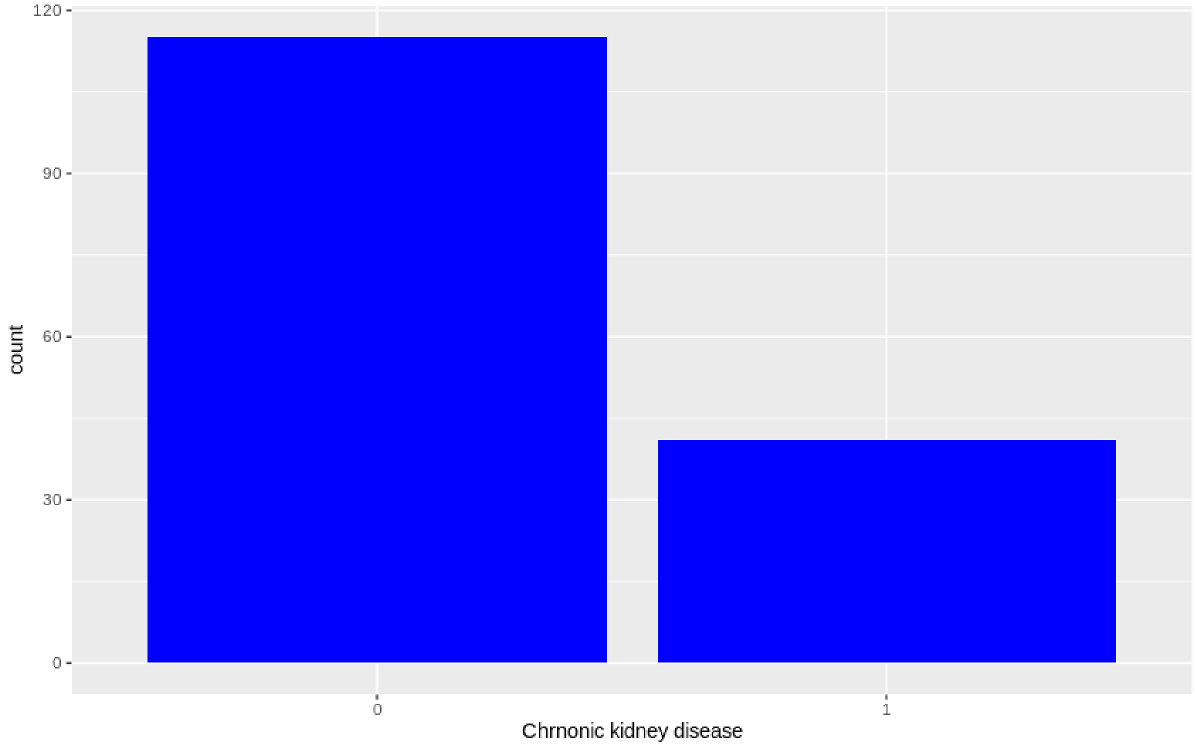


Figure 2 Distribution of chronic kidney disease in data set

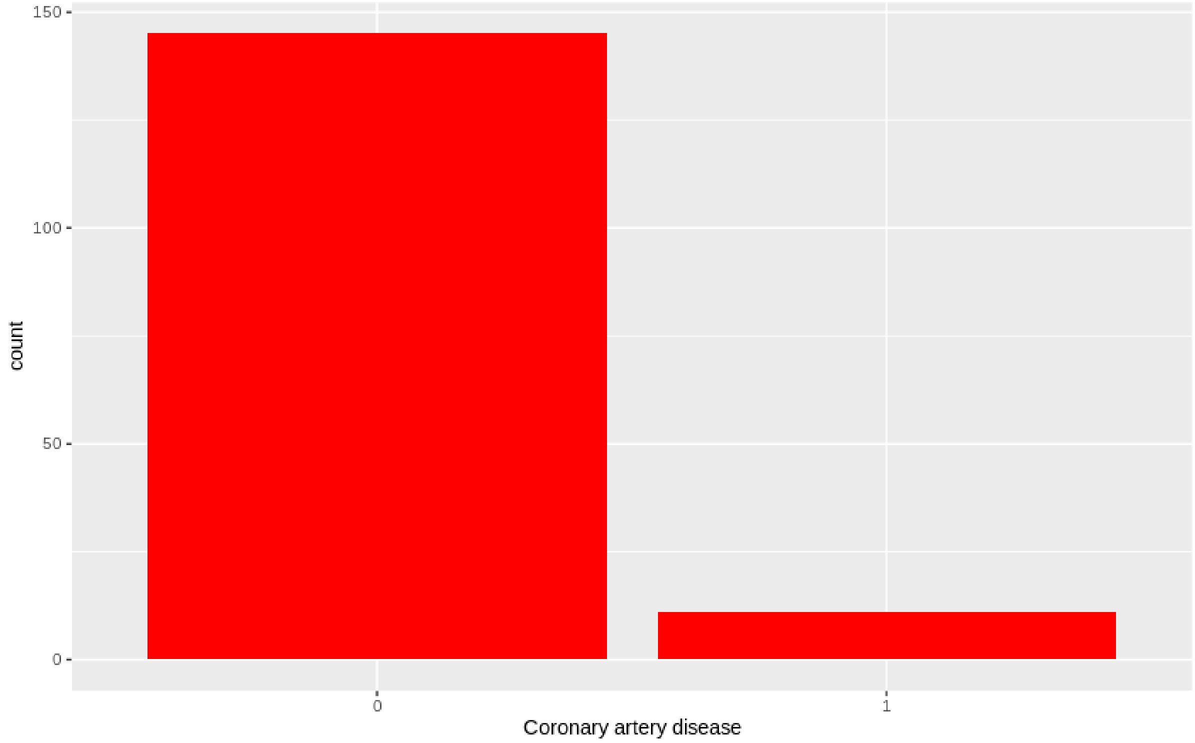


Figure 3 Distribution of coronary artery disease in data set

Exploratory data analysis

The main goal of the exploratory data analysis was to look at how different variables interact with each other.

The analysis began by looking at the way the Age variable interacted with the Blood Pressure variable, in regard to the chronic kidney disease factor. It can be noticed in figure 4 that the probability of kidney disease increases when the patient’s blood pressure was over 80 mm/Hg, as every patient with blood pressure over 80 mm/Hg in this dataset had kidney disease. Kidney disease was still present in patients with blood pressure under 80, thus it is not the only factor that affects the presence of kidney disease. In the fourth figure it can also be noticed that kidney disease is more likely in patients over the age of 40, with only 2 outliers to this statement. In the fifth figure the Age variable was compared with the Glucose variable. It appears that patients with Glucose levels over 150mgs/dl were more likely to have kidney disease. Figure 6 seems to corroborate this as well, as most patient that had blood pressure under 80 and glucose under 150 didn’t have chronic kidney disease, with only 2 outliers to this statement.

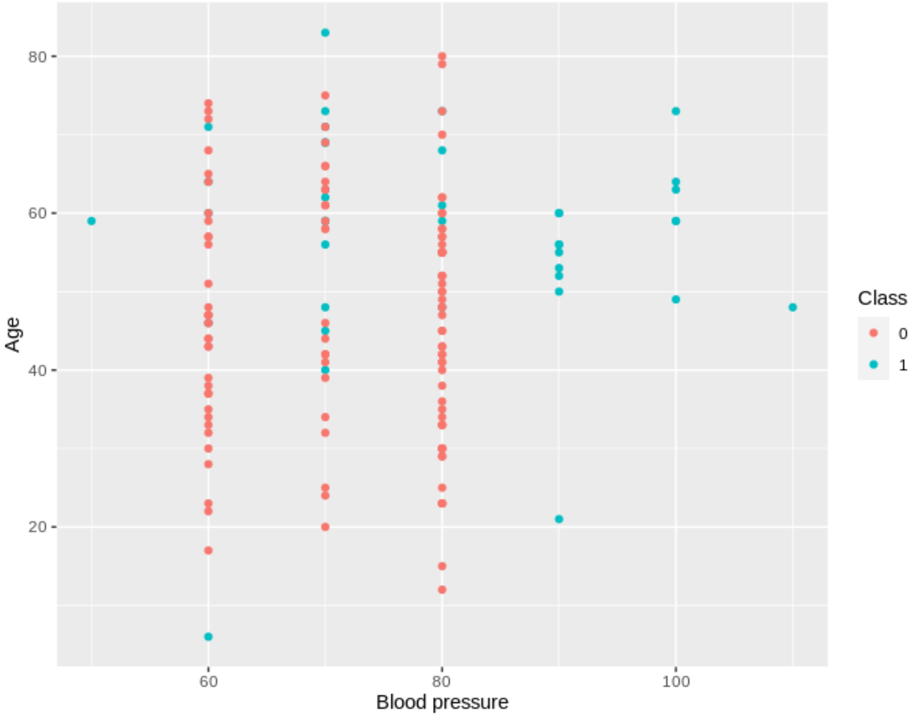


Figure 4 Age and Blood Pressure by Class

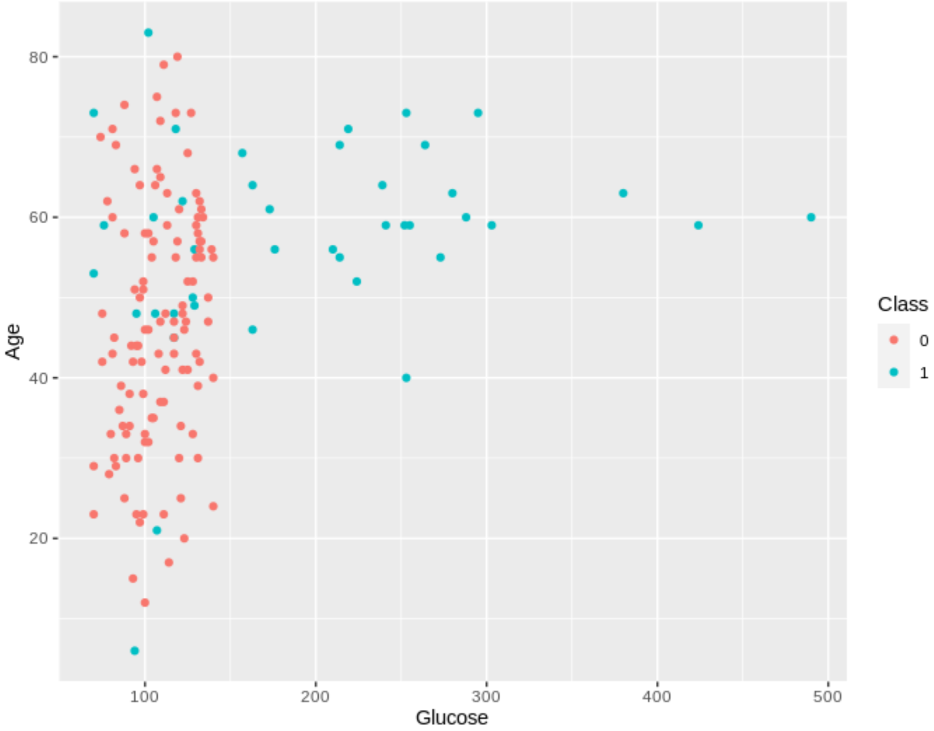


Figure 5 Age and Glucose by Class

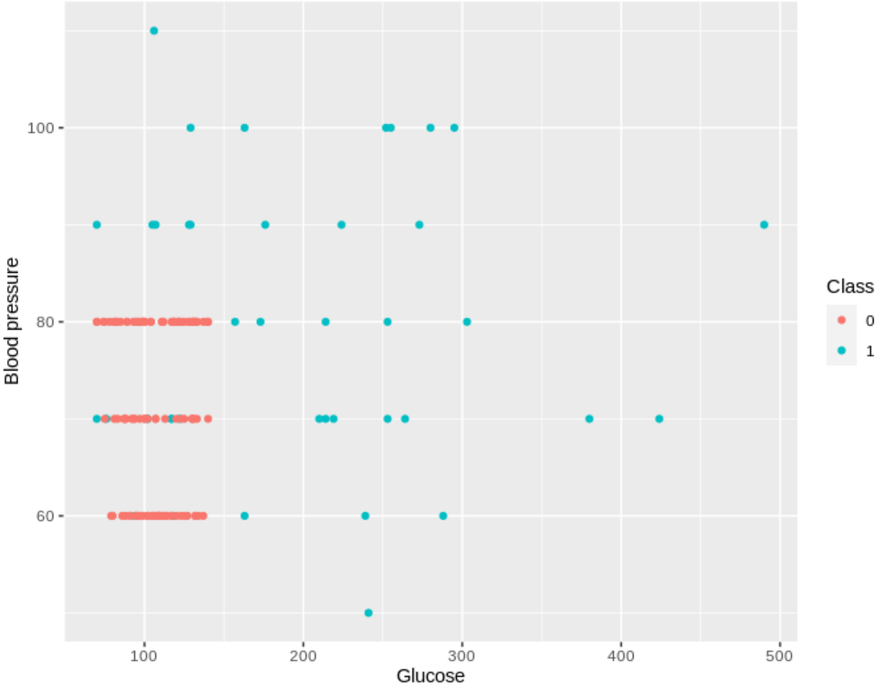


Figure 6 Glucose and Blood pressure by class

Next it was important to analyse how the presence of coronary artery disease affects the presence of chronic kidney disease. It appears that the presence of coronary artery disease automatically meant presence of kidney disease, as all patients that had coronary artery disease also had kidney disease. Visualization in figure 7.

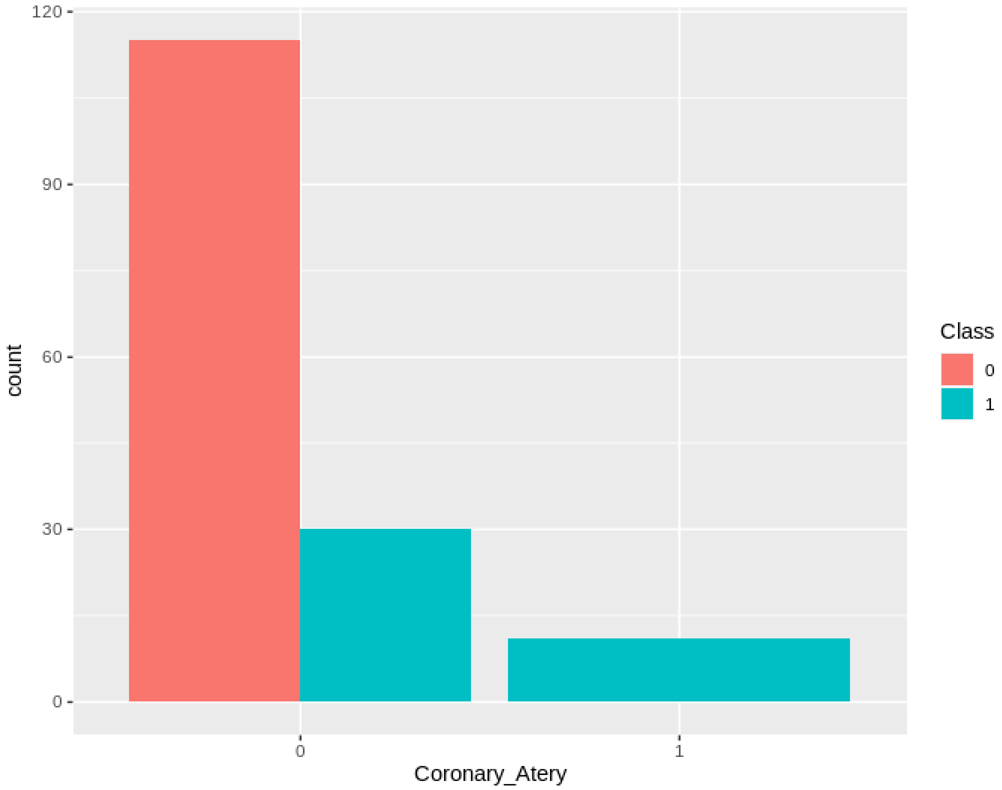


Figure 7 Kidney disease when Coronary Artery disease

Further analysis was needed on how coronary artery disease is present with the other variables: Age, Blood Pressure, and Glucose.

On analysis it appears that coronary artery disease is more common in ages over 40, as well as for patients with blood pressure over 80 (figure 8). It is important to note that coronary artery disease wasn’t present in our dataset in patients under the age of 40 and with blood pressure under 80 mm/Hg. It also appeared that coronary artery disease was more common in patients with glucose levels over 150 mgs/dl, with only 1 patient with lower glucose levels that had coronary artery disease (figure 9). These results were consistent with the results of the analysis of presence of chronic kidney disease, when compared to the same variables.

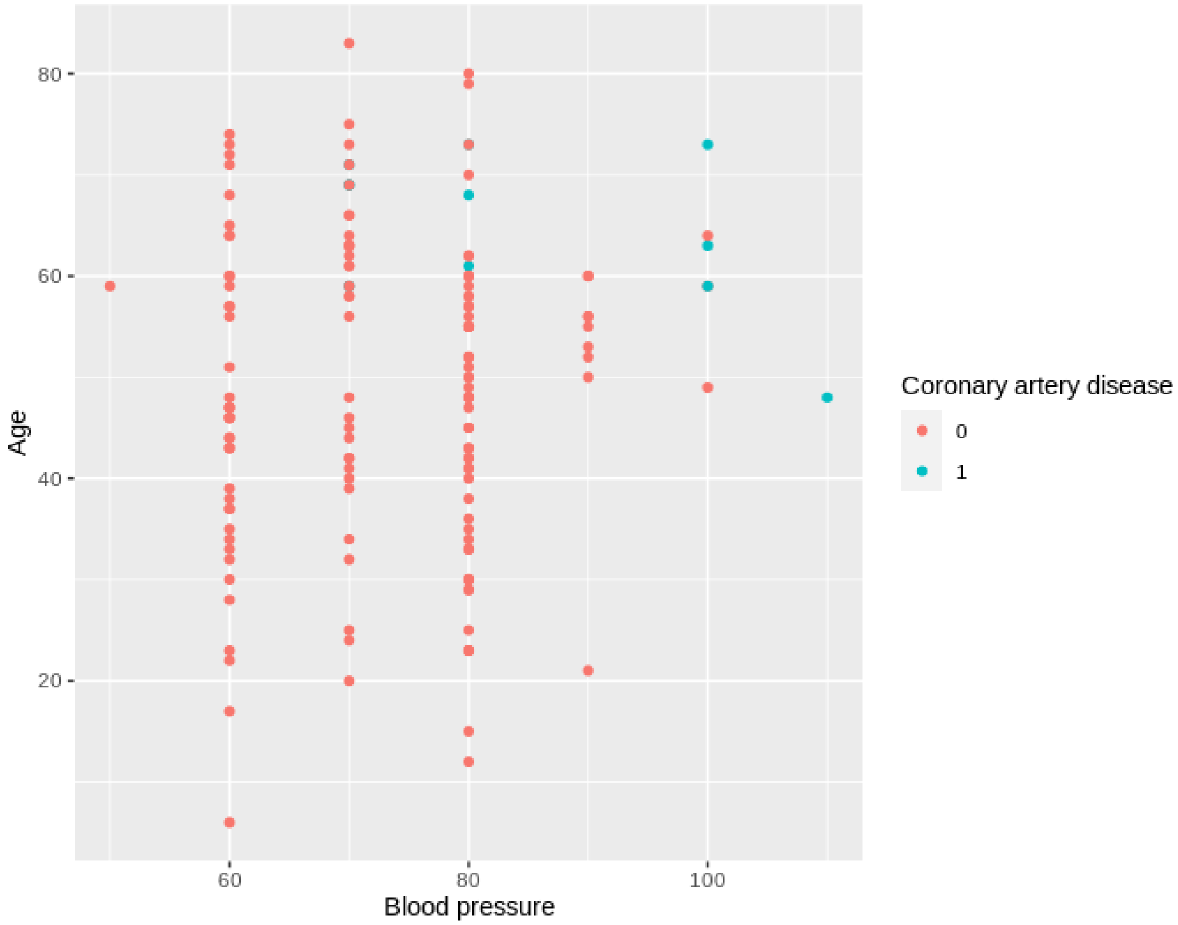


Figure 8 Blood pressure and age by coronary artery disease

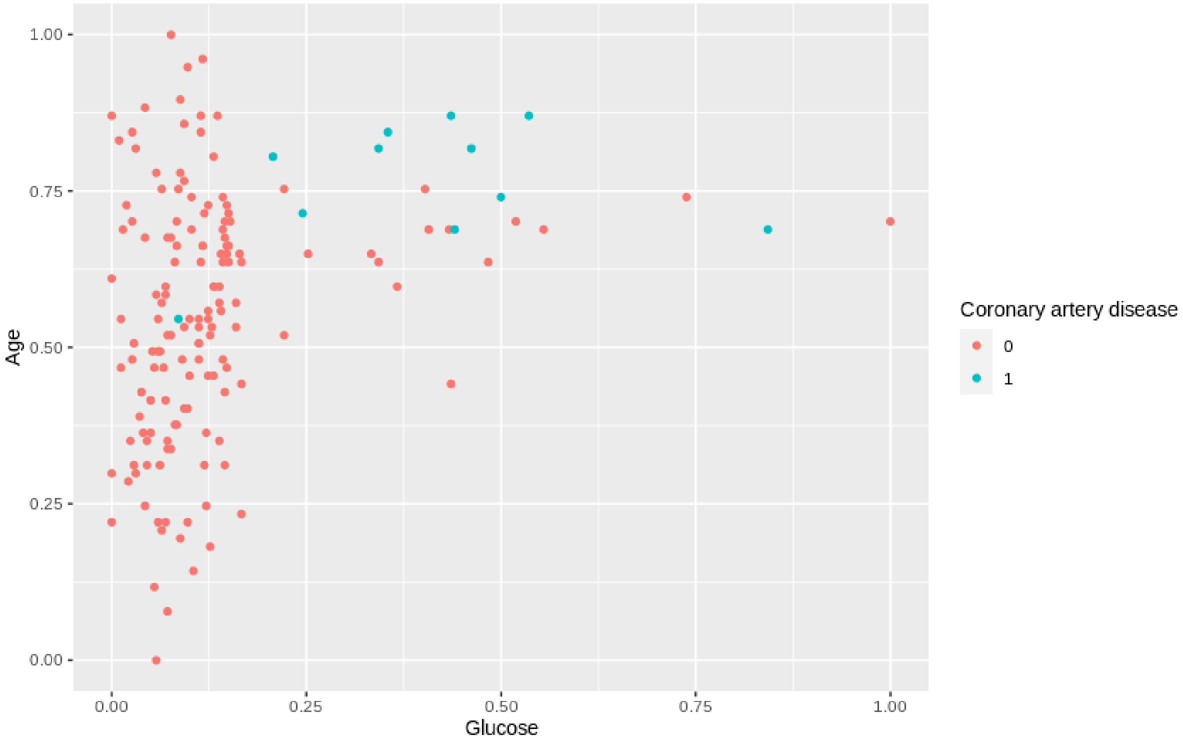
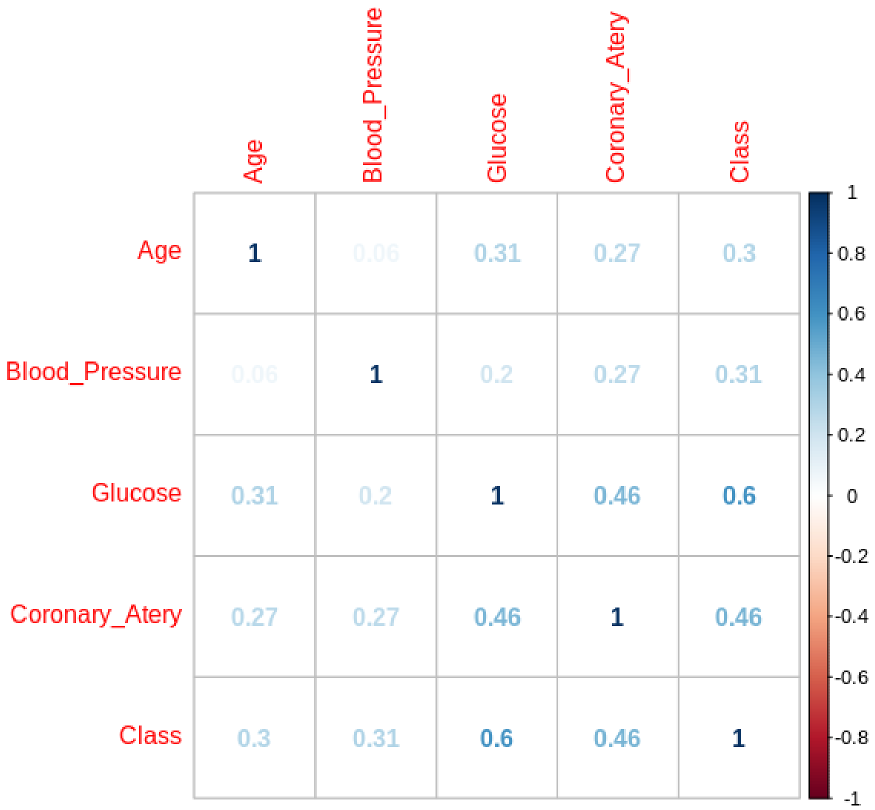


Figure 9 Age and glucose by coronary artery disease

To confirm these findings analysis with the dplyr-package was done. The idea was to look whether these statements were true, when only analysing patients that had chronic kidney disease. First the average values were calculated for each explanatory variable, when grouped by Class. The averages of patients with chronic kidney disease were: Age=57.2 , Blood Pressure=79.5, and Glucose=197. The averages indicate that for the age, blood pressure, and the glucose variables the values are higher than the averages of each variable, without the grouping by Class. Additionally, the smallest value from each variable was found, when grouped by Class.

Next it was important to verify our initial analysis of threshold values of blood pressure over 80, age 40 or over, and glucose over 150. It appeared that the blood pressure and glucose level threshold values were correct, as there were no patients without kidney disease when smaller values than the threshold values were filtered away. Age was not as strong of a threshold value with only 39/160 patients 40 or over having chronic kidney disease.

After this correlation analysis was performed on the given variables. The strongest correlation was positive correlation between the variables Class and Glucose. This seems to be consistent with the previous findings, as every patient with glucose levels over 150 mgs/dl has chronic kidney disease. Additionally, there appears to be some correlation between Coronary Artery and Glucose, as well as Coronary Artery and Class. These findings are also consistent with the previous results. Lastly there is slight correlation between Class with Age and Blood Pressure, as well as Coronary Artery disease with Age and Blood pressure, as well as Glucose and Age. These findings are also consistent with the previous exploratory data analysis. All the correlations were positive. Figure 10 and Figure 11 are visualizations of the correlation between variables.



	Age	Blood_Pressure	Glucose	Coronary_Atery	Class
Age	1	0.06	0.31	0.27	0.3
Blood_Pressure	0.06	1	0.2	0.27	0.31
Glucose	0.31	0.2	1	0.46	0.6
Coronary_Atery	0.27	0.27	0.46	1	0.46
Class	0.3	0.31	0.6	0.46	1

Figure 10 Numeric correlation

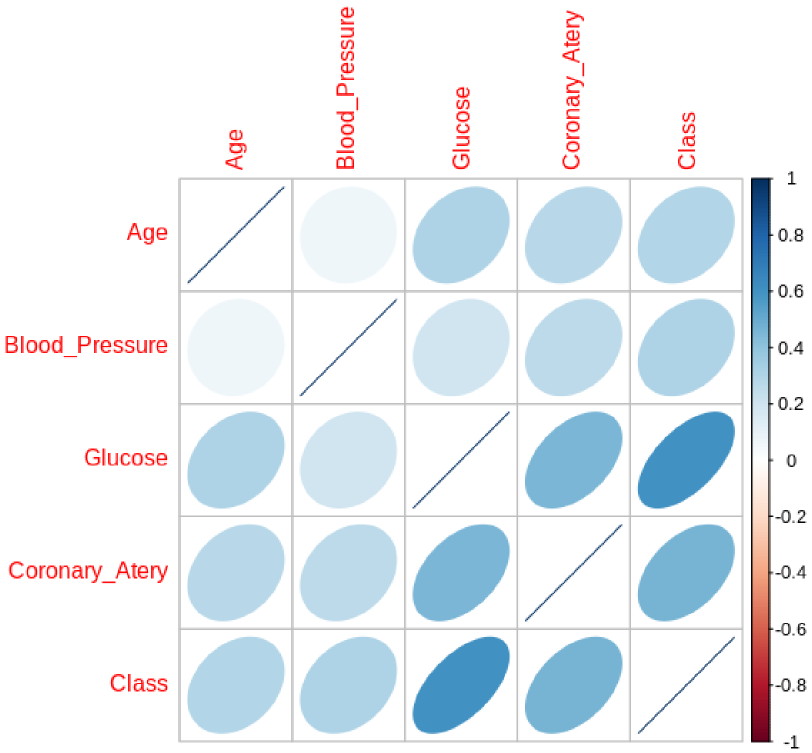


Figure 11 Visual correlation

Initial decision tree model

Before building the initial tree model, the balancing of the data was checked. It appeared that the data was not too balanced, as 74% of the data belonged to Class false. This means that we are more interested in how our model performs on the smaller Class of true, and we are looking for an accuracy that is better than 74%.

The data was split into training and testing data, with a ratio of 70:30 respectively. Then the initial tree model was trained, with minimum leaf size of 1. The resulting tree can be seen in figure 12. The tree is making decisions on similar threshold values as in the initial exploratory data analysis, with for example if the glucose level is over 149 then the patient has chronic kidney disease. Also, blood pressure over 85 is also an indication of chronic kidney disease. From the figure it might be able to see that the model is a bit overfit right now, as the criteria for deciding whether someone has kidney disease is very detailed. This was checked by making predictions with the training data and the testing data separately and then comparing the results. On the training data the model delivered perfect results of 100% for accuracy, recall, and precision. On the testing data the model produced an accuracy of 85%, recall of 83%, and precision of 67%. While the results are not horrible, it is still a drastic drop from the results on the training data, meaning that the model is overfit and will not generalize well. The drop in precision indicates that the model predicts true positives 67%, which means that 33% are false positives. Additionally, recall indicates that the model correctly finds positive cases 83% of the time, which means that 17% of the patients with chronic kidney disease would get a false negative. For medical data, these values need to be higher.

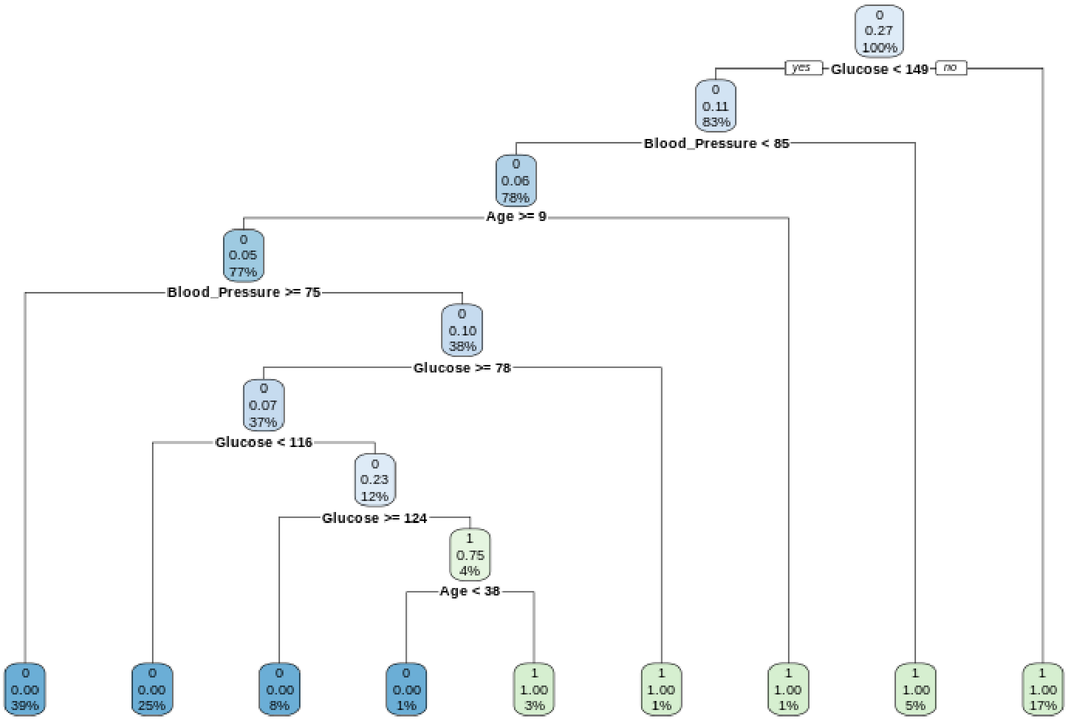


Figure 12 Initial decision tree model

Decision tree model with model selection

The model can be improved, by finding the best generalizing value for the minimum tree leaf size. For the model selection the data set was divided into 3 smaller sets: training, validation, and testing sets at a ratio of 60:20:20. The validation data is important, as it enables the testing data to be completely separate from the model selection process, and thus the testing data will be a good way to evaluate the model generalization. If the model was selected only using training and testing data, then we will select the parameters that perform best on the testing data, and not the parameters that generalize best, thus ending up with a model that gives good predictions on the testing data, but not good ones in general. The model selection was ran 100 times, with minimum leaf sizes of 1 to 50. The best parameter (minimum tree leaf size) is decided by the performance of the validation data. The leaf size that produces the best validation result will generalize best. In figure 13 the performance of the models with different minimum leaf size values are visualized, with the blue line being performance on the training data, and the red line being on the validation data. From the graph it can be seen, that the model overfits with leaf sizes of 1-3, and underfits for sizes after 3. The best validation value was with minimum leaf size 3, so this is the parameter chosen in the model selection.

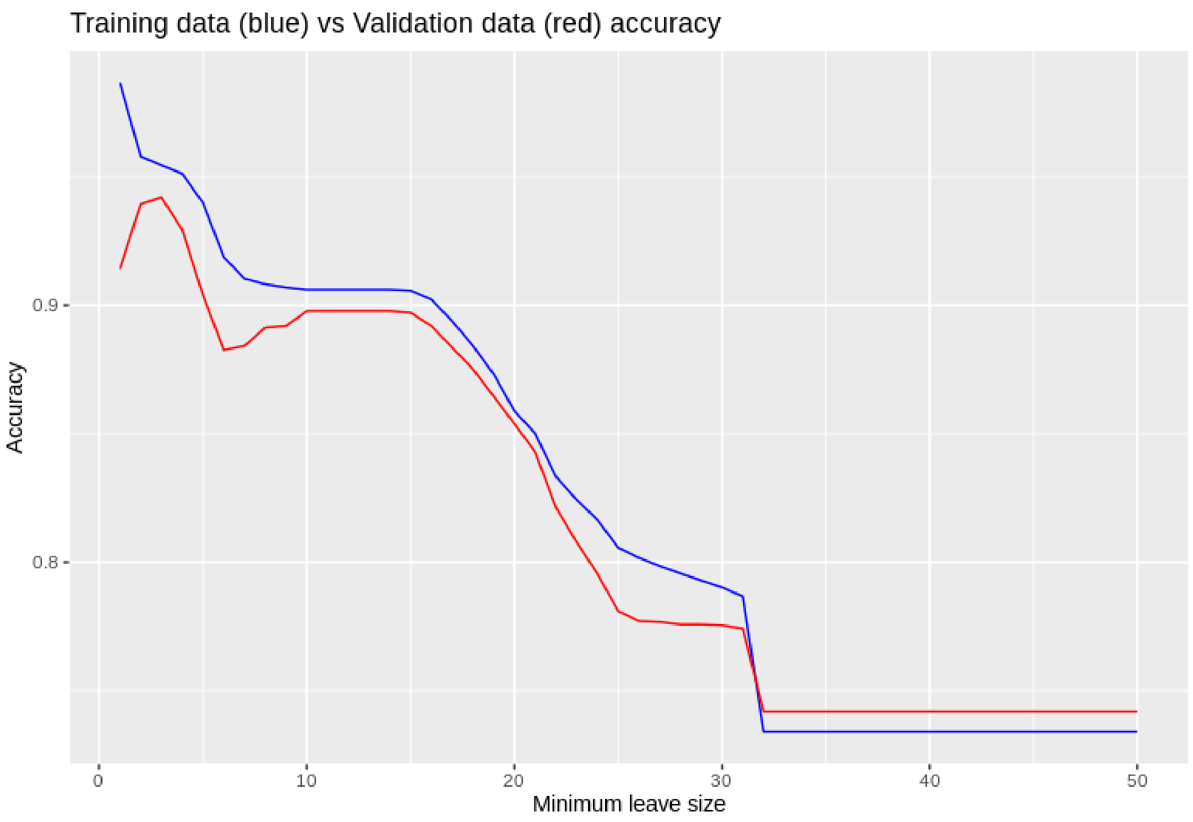


Figure 13 Training vs validation data accuracy

Next the model was trained with the validation and training data combined and tested separately on the testing data. On the training data the model got an accuracy of 95%, recall of 82%, and precision of 1. On the testing data the model got an accuracy of 90%, recall of 63%, and precision of 1. These results appear to be pretty good, with the exception of recall, as ideally recall would be as small as possible, as it is the value that indicates how many percent of patients that had chronic kidney disease got correctly diagnosed. The model would not appear to be overfit or underfit based on these results, meaning that the model should generalize well with new data.

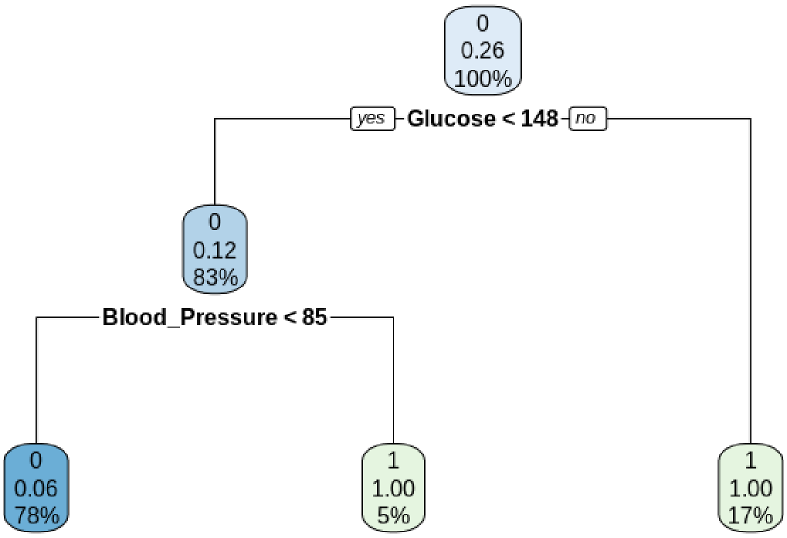


Figure 14 Final decision tree

K-nearest neighbour classifier

For the k-nearest neighbour classifier the data was normalized. Normalization helps but all the values of each variable in a similar scale, which is great for models that use distance between different data points to make their predictions. Then we repeated the same steps for model selection as in with the decision tree. The only difference was that instead of looking for the optimal value of the minimum leaf size, the optimal number of neighbours was searched for. This means how many closest neighbours are considered in the making of the prediction for each point. From the results the best number of neighbours would be 1, but I don't believe that the model would generalize well if it only considers the 1 closest point (Figure 14). I still tested out the model with parameter 1, and the outcome was an overfit model as I suspected. This is why I redid the model with the second-best parameter of 3. With the parameter 3, the model evaluation on training data resulted in accuracy of 94%, recall of 79%, and precision of 1. On the testing data the model produced an accuracy of 94%, recall of 75%, and precision of 1. It appears that overall the k-nearest neighbour model had better results than the decision tree.

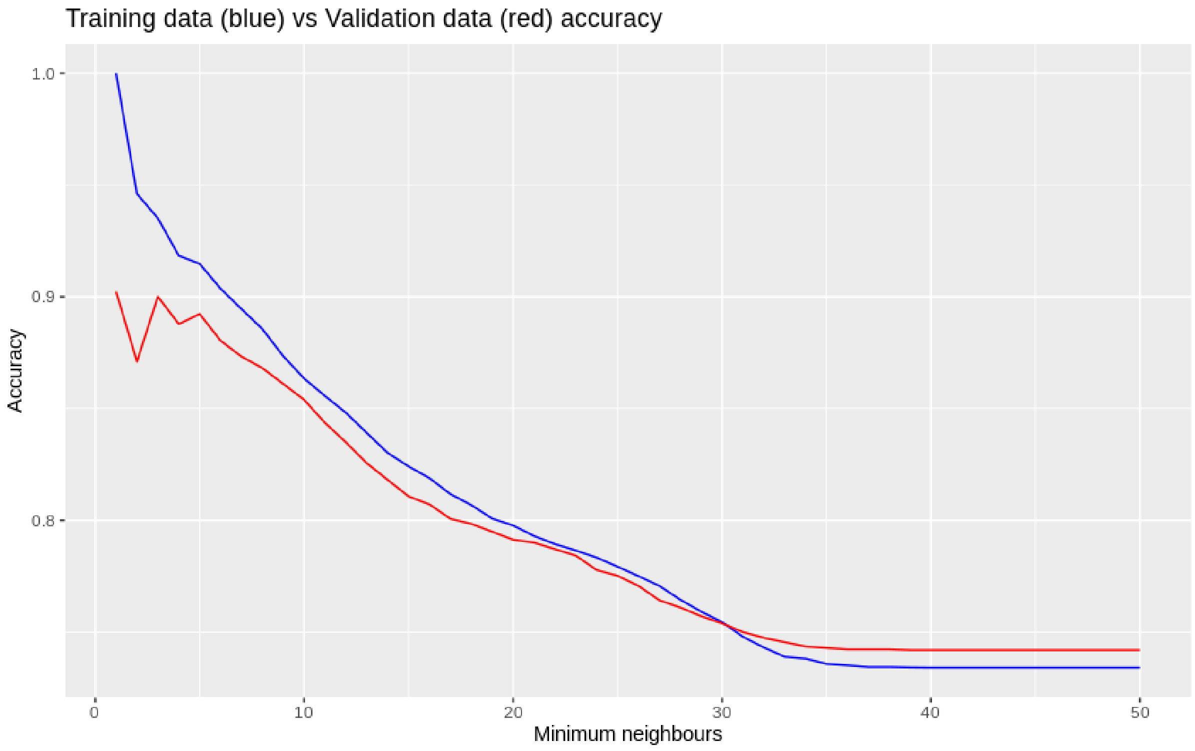


Figure 15 Training vs Validation data accuracy

Conclusion and discussion

The results on the testing data indicate that the k-nearest neighbour model performed overall better, when compared to the decision tree model. While the results of the model selection indicated that the best parameter for the knn model would be 1, it was a good call to choose parameter 3, as the model would not generalize well. Overall, both models appear to generalize well on new data after the model selection, and both gave respectable results. The most important part for medical data classification is the recall value, as this shows how many patients from all patients that have chronic kidney disease are classified correctly. The knn model produced the best recall value, which is why in this case, this model is recommended. The knn model also produced better accuracy for the testing data, and the same precision as the decision three model with the testing data.

These models can help professionals in the medical field to be able to detect signs of chronic kidney disease in an early stage. This is because a machine learning model can find indications of chronic kidney disease from the patient information, especially the information of blood pressure and glucose levels. Additionally, a machine learning model is less prone to make human error and can process large amounts of data quickly and efficiently. The machine learning model can be improved by adding additionally information on other parameters from patients, as additional parameters might give the machine learning model more information on indicators of chronic kidney disease.

Part 2 Clustering

Data

The data set consisted of 440 observations of 8 variables. The variables were: Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicassen. The first two variables appear to be categorical data, and the other variables are explanatory data of annual spending on the products by variable name. All the variables contain numeric data. The data set did not contain any empty values. Plot 16 visualizes the explanatory variables. It can be noticed that most of the costumers spent in a similar range for all of the products, with a smaller group of costumers spending annually more on each product. Next we looked at the distribution of costumers between Region and Channel. From figure 17 and 18 it can be noticed that most of the consumers shopped in region 3, and the most popular sales channel was 1.

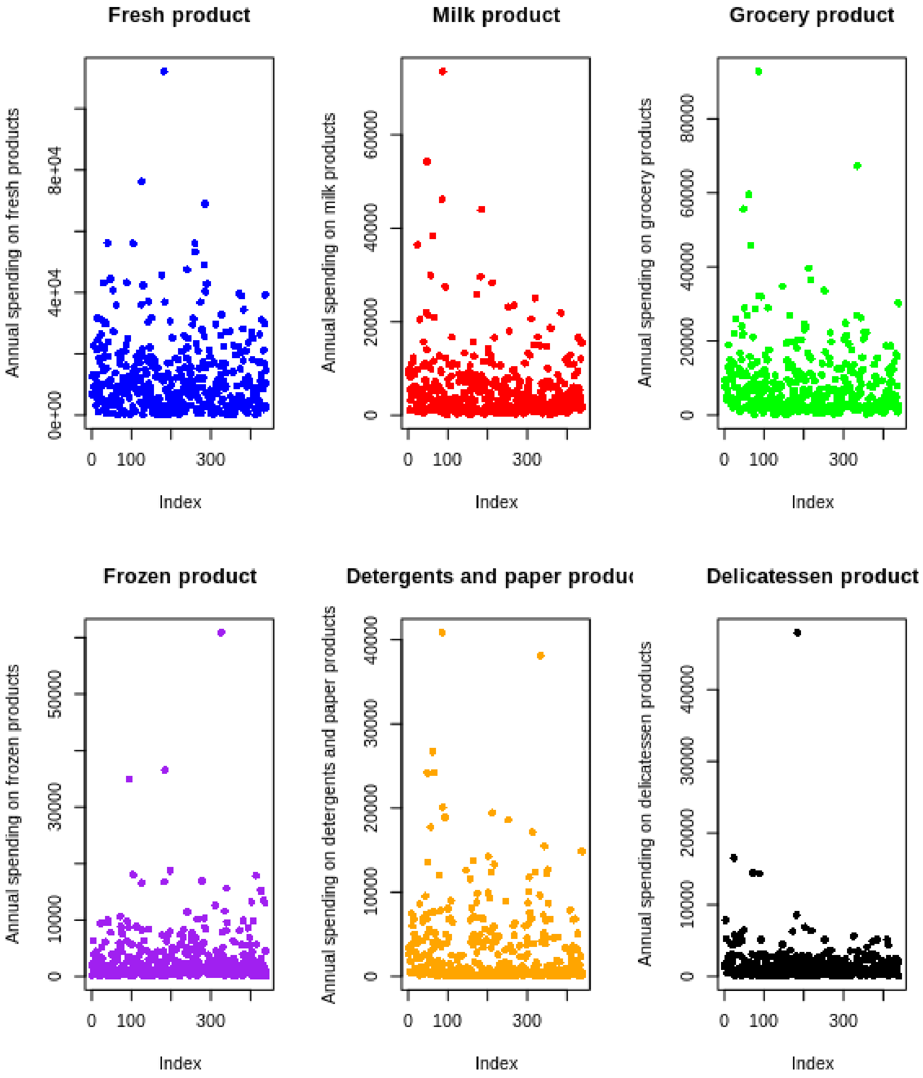


Figure 16 Analysis of exploratory variables

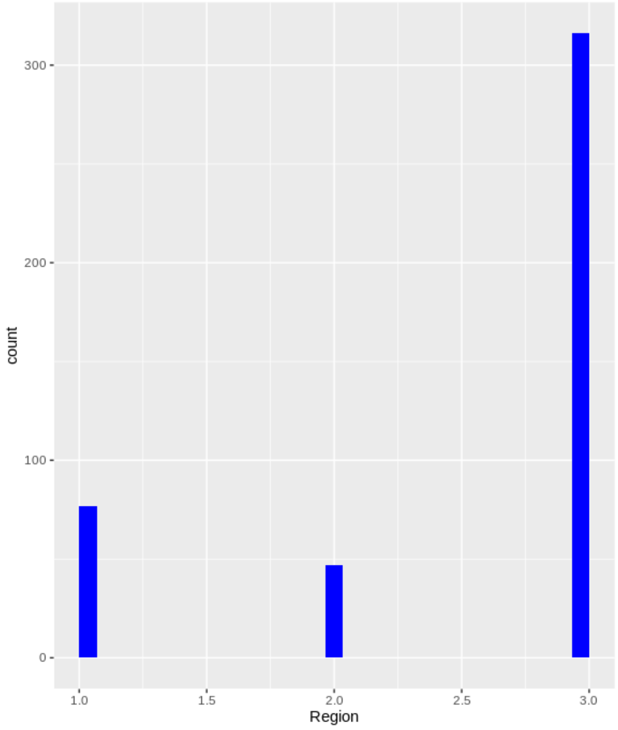


Figure 17 Distribution of costumers by region

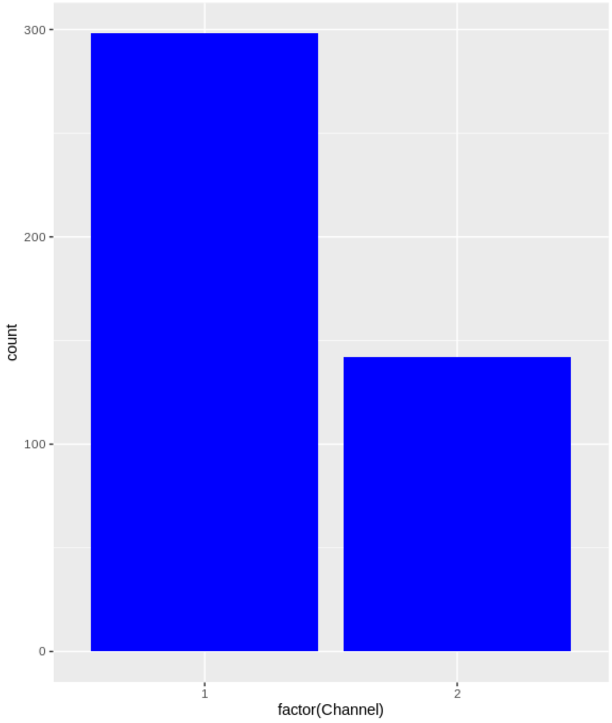


Figure 18 Distribution of costumers by channel

Exploratory data analysis

The explanatory data analysis began with the calculation of correlation between variables. Figures 19 and 20 are visualizations of the correlation. It can be noticed that the strongest positive correlation was between grocery and detergents_paper. There was also positive correlation between variables: grocery and milk, milk and detergents_paper, detergents_paper and channel, milk and channel, and grocery and channel. Additionally there was some positive correlation between: delicatessen and milk, delicatessen and frozen, as well as frozen and fresh. The assumption would be that these values might have some impact on the clustering of the data, especially the ones with strong correlation.

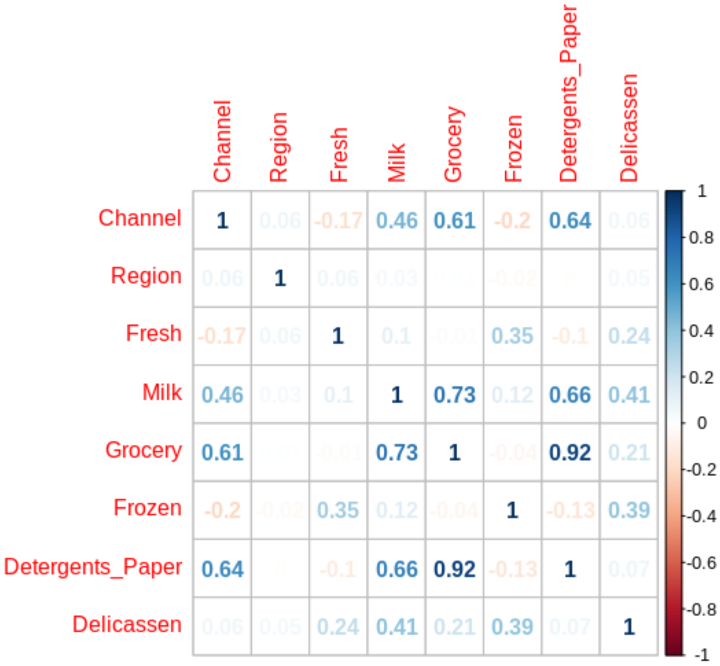


Figure 19 Numeric correlation analysis

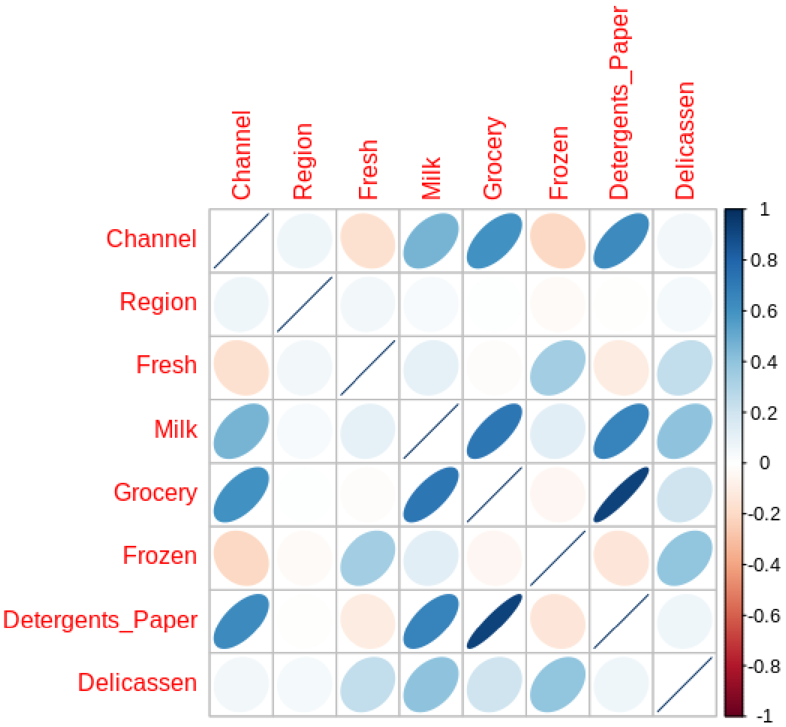


Figure 20 Visual correlation analysis

Next the ones with strong or normal correlation were visualized. Figure 21 represents the annual spending on grocery and detergents and papers by channel. There is clear positive correlation, as if the consumer used more money on grocery, then they also used more money on detergents and paper. Additionally, it appears that consumers that used sales channel 1 spent on average less money, than consumers that used channel 2.

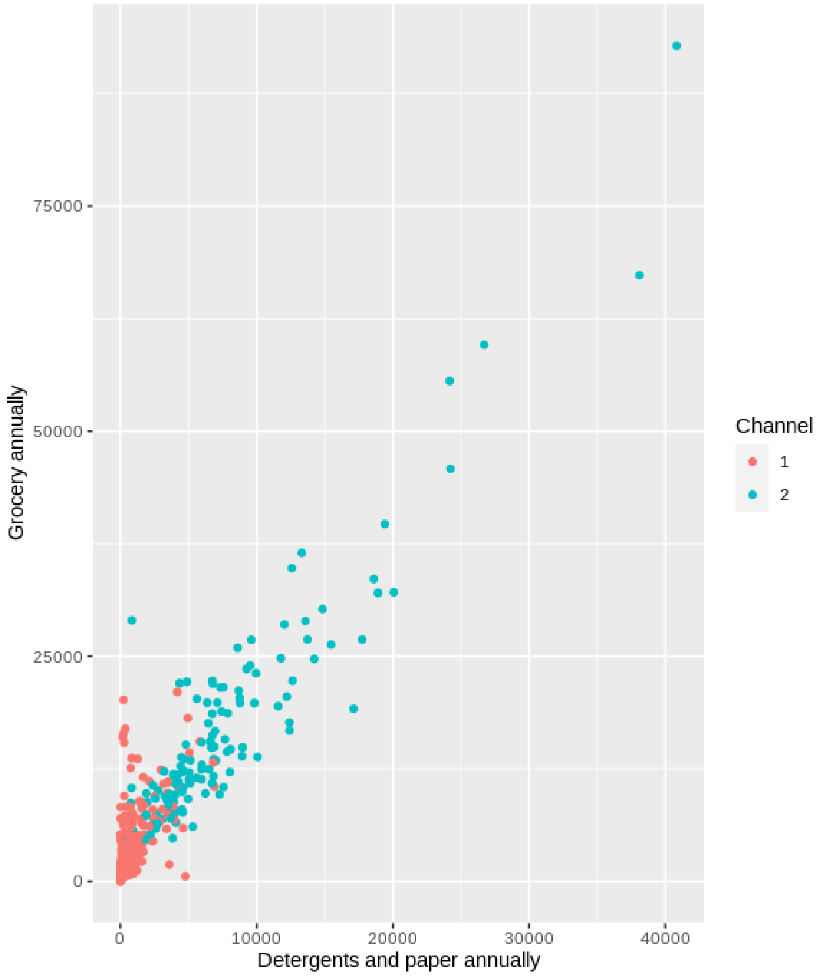


Figure 21 Grocery and Detergents by channel

Figure 22 represents the annual spending on milk and detergents and paper by channel. There seems to also be positive correlation between these two values, but not as strong as the previous ones. It would appear that if a costumer used more money on milk, they also spent more money on detergents and paper, although only slightly more as in comparison to figure 21, most of the costumers are under the spending of 15 000 on detergents and paper, while in figure 21 most of the costumers used around 5000 more on detergents and paper. There are also similar findings in regard to channel, as those that used channel 1 spent annually less than those that used channel 2, with a few outliers that spent more on milk products, but not on detergents and paper.

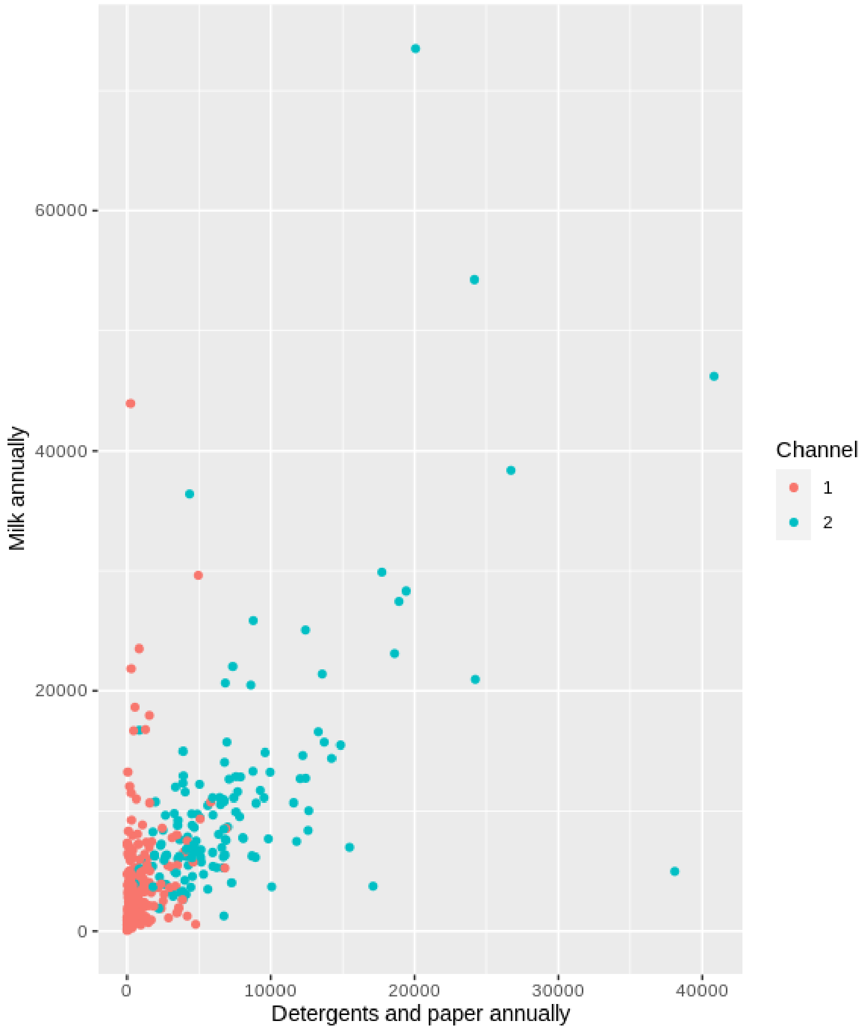


Figure 22 Milk and detergents by channel

Figure 23 represents the amount spent on milk and grocery annually by channel. The findings appear to be the same as above. There is some correlation between how much a costumer spent on milk, with how much they spent on grocery. The costumers that used sales channel 1 spent less money, than the costumers that used sales channel 2.

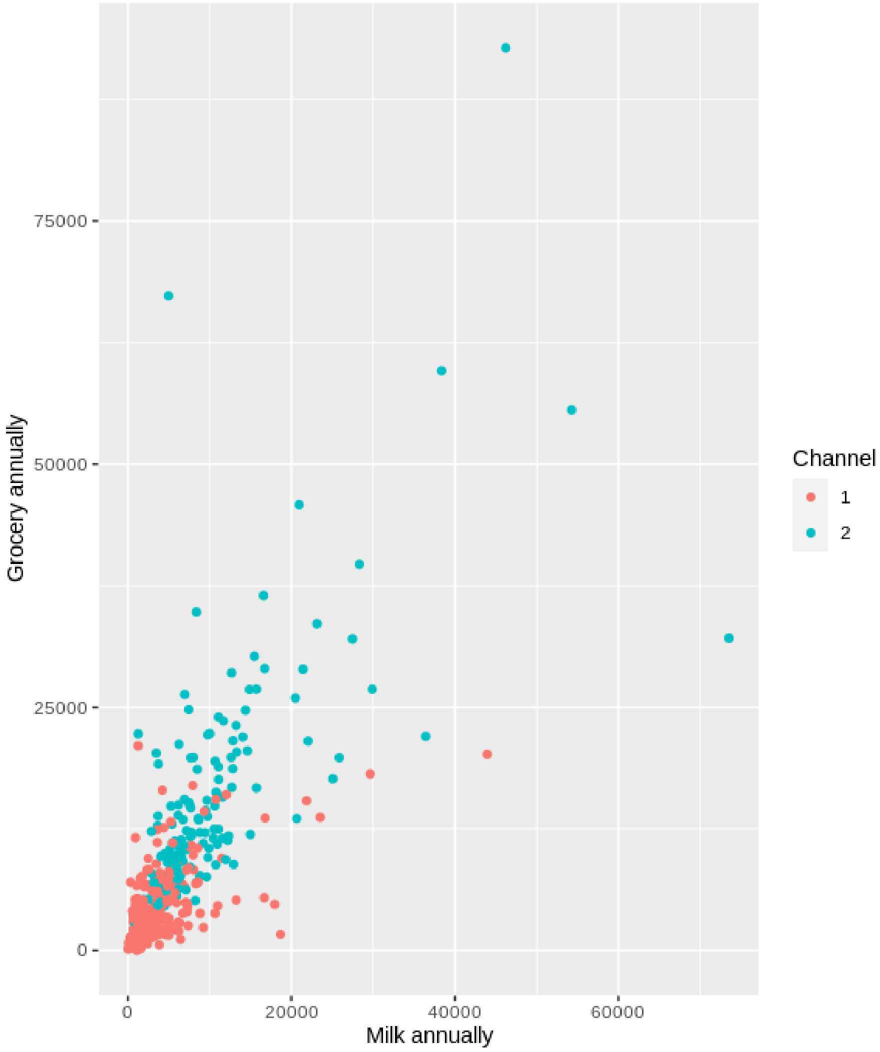


Figure 23 Grocery and Milk by channel

Next analysing was done on how channel affects the average, minimum, and maximum spending. It appears that those that used channel 1, spent more money on fresh and frozen products, than those that used channel 2. In the other variables those that used channel 2 spent more money than those that used channel 1. The minimum values in costumers that used channel 1, were smaller than those that used channel 2, except for delicatessen products, which was equal to those that used channel 2. The maximum values correspond with the previous findings, with consumer from channel 1 that used more money on frozen, fresh, and delicatessen products, but otherwise channel 2 had the higher annual spending in the other variables.

Optimal number of clusters

The data was normalized using min-max-normalization, which means that all of the variables were turned into values between 0 and 1. Normalization is especially important in clustering, as clustering works based on the distances between different data points. When we have data that has a lot of variety between different variables, as well as within each variable the results of the clustering can be worse. Normalization turns all the values into a similar range, which improves the efficiency and accuracy of the clustering model, as it reduces the distance between different data points.

Next the different models for determining the number of clusters were applied. Starting with the Elbow method. Figure 24 represents the elbow methods results. It would appear that 4 would be a good amount of clusters, as in the graph this appears to be the point where the graph starts to gradually decrease. The elbow method works by calculating the sum of square distance for each point to it’s assigned center. The elbow method calculates the distances based on how many clusters are wanted, usually range of 1-10. The output is the average sum of within square distance score for all the clusters. Basically the elbow method gives the total sum of the variation within each cluster.

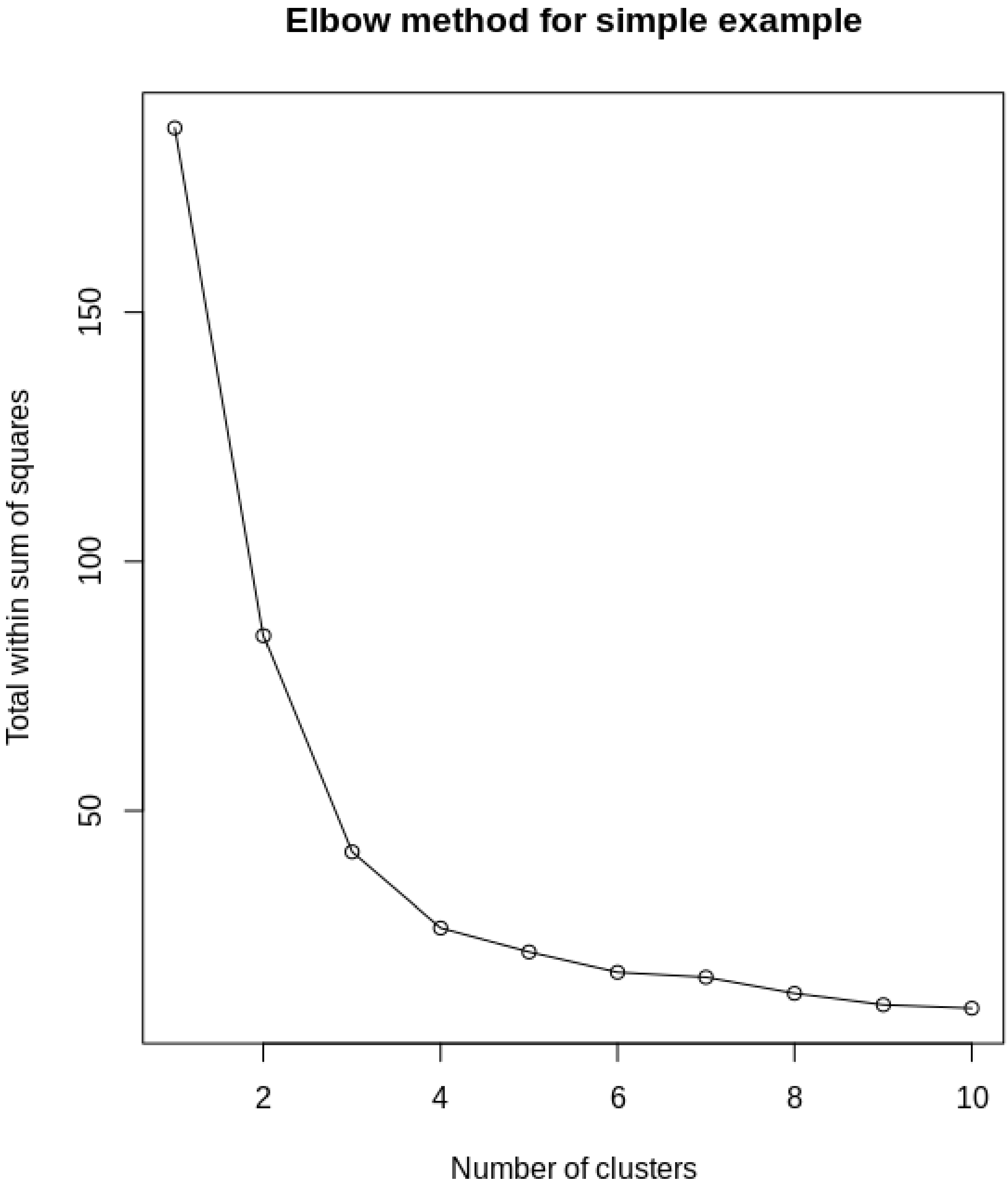


Figure 24 Elbow method results

Next all the other models were created: Silhouette method, Gap statistic method, and Calinski-Harabasz method (Figure 25). The silhouette method indicates that the best number of clusters would be 3, the gap method 2, and the CH method 9. The figures could not be more different, so the decision needs to be decided, by which number of clusters satisfies most of the methods, at least partially. For the elbow method a cluster amount of 3 would work fine, which is the ideal for the silhouette method. Additionally, a cluster number of 3 would also work for the Gap method and the CH method, as it is the second-best value for both of these, when cluster number is 3. A cluster number of 3 is chosen.

The CH method measures the ratio between the between group sum of squares and the within group sum of squares. This means that it looks at the variance between different clusters and then looks at the variance inside each cluster. The CH method then calculates how far are the cluster centers from each other and how dispersed each cluster is within. The ideal outcome would be the distance between clusters is very big, and the variance inside each cluster is very small. So the higher the result of CH is, the better the number of clusters will work.

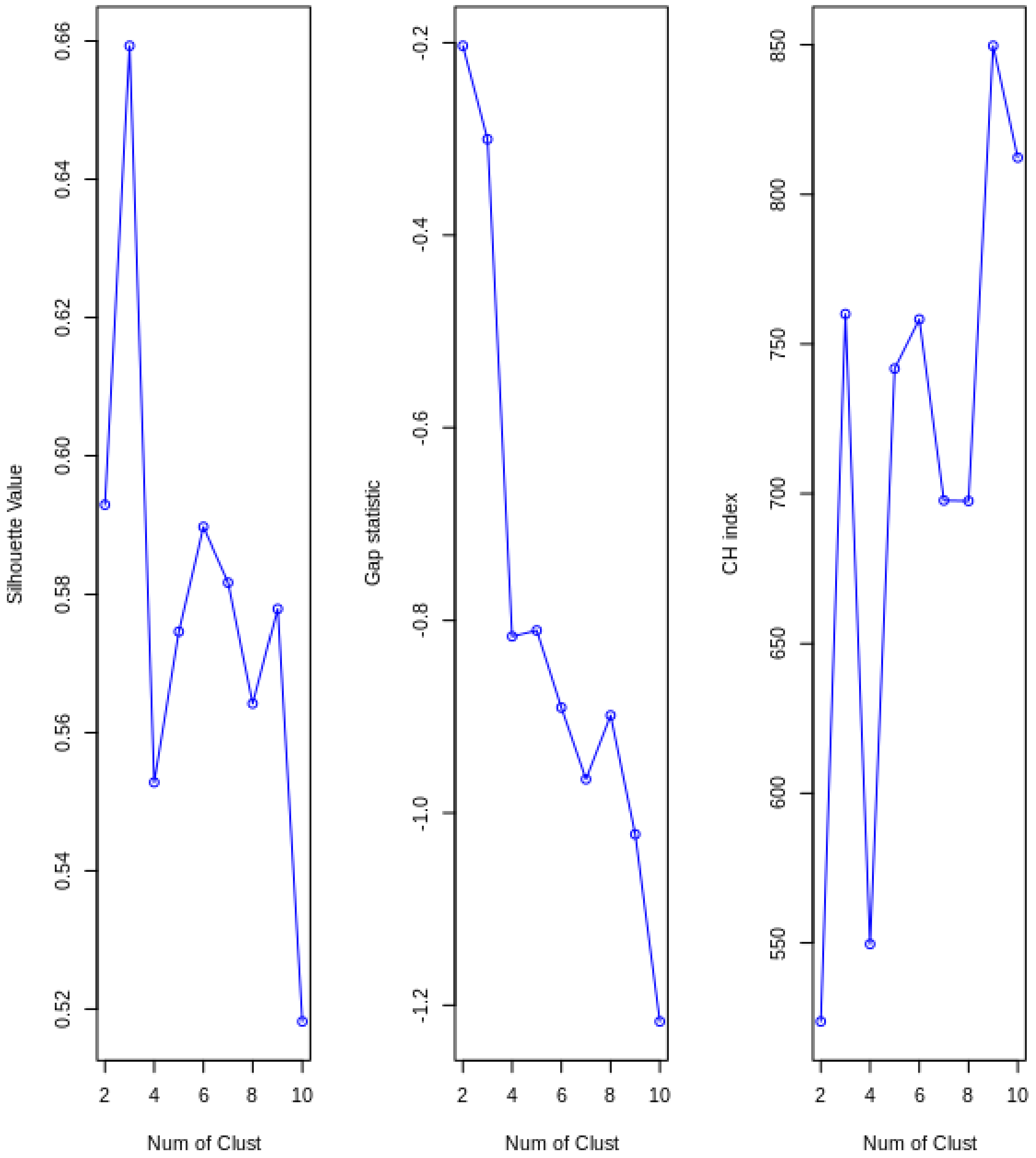


Figure 25 Other methods results

Clustering results

The model was then created with the ideal amount of centers (3). Upon analysing the clustering results, they seemed to not be very consistent. It appeared that in most of the variables average values by cluster, that there were always two clusters with similar averages and then the third cluster had a very different average. This would indicate that the cluster centers of two variables overlap, which is not desirable for a clustering model, but because the variables that have same averages vary depending on variable, it is not possible to merge two clusters into one. Standard deviation had more variation between clusters in variables, which would indicate that the variables shouldn’t overlap too much. Lastly we plotted the data with the model, but it appears that the result are inconclusive (Figure 26).

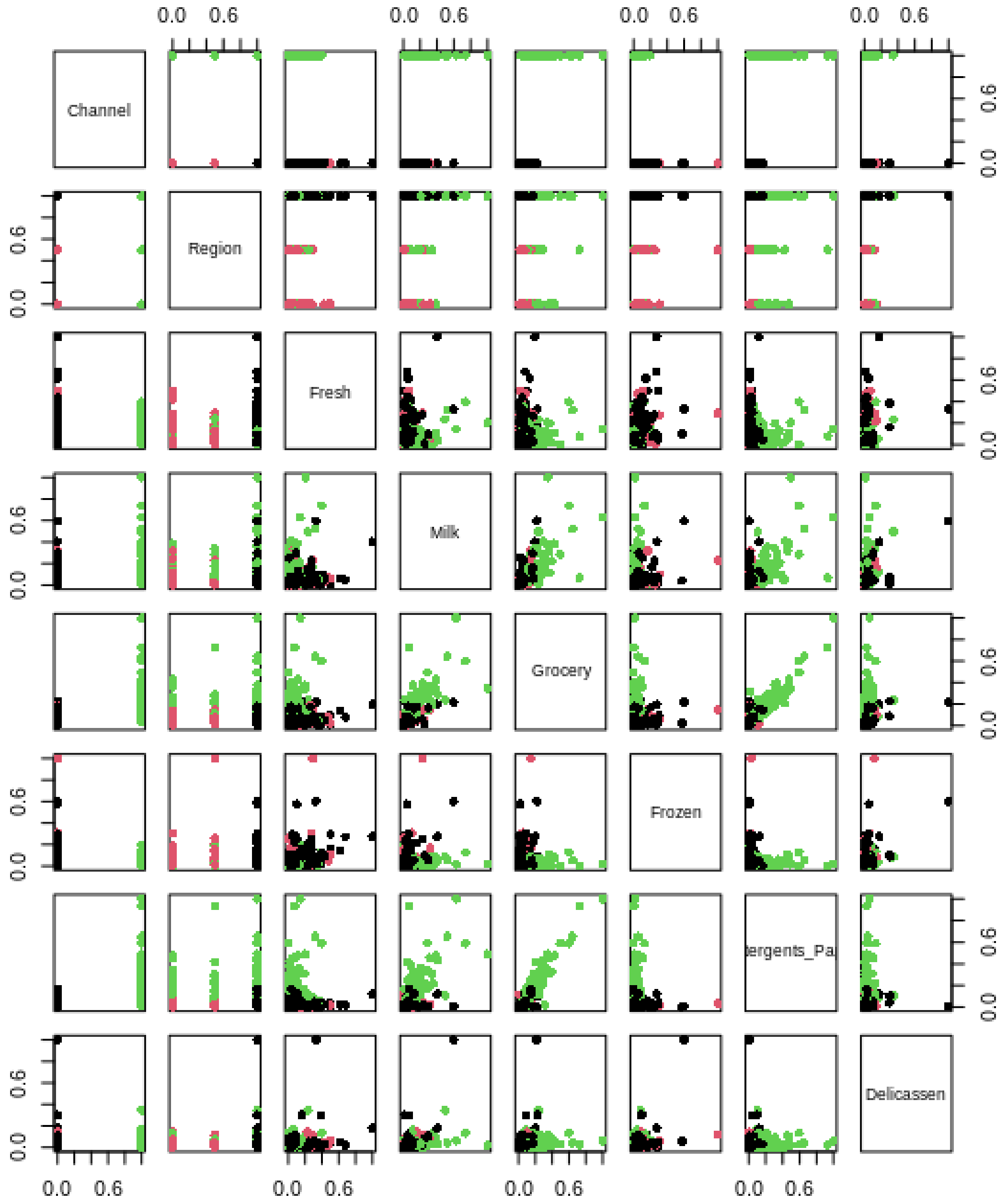


Figure 26 Clustering model results

Conclusion and discussion

Sadly it appears that the results were inconclusive and clusters were not found using the data. Additional research might be required and while the model was tested with other numbers of clusters as well, the results were still inconclusive. Perhaps there was some variable, that affected the clustering model negatively or maybe the data simply doesn’t lead to any clustering. If clusters could have been found the wholesale company could have used the findings of the model to for example market products to consumers who are likely to buy them based on their other purchases.