LUT School of Business and Management

A220A0010 Free Analytics Environment R

Christoph Lohrmann

ASSIGNMENT 2

3.11.2022

0565878

Table of contents

Part 1: Regression analysis

1.1 – 1.2 Studying the data set and exploratory data analysis

For this part of the assignment, we have a dataset containing statistics of different types of arrests, number of car accidents and urban population in that area. The aim is to find out if it's possible to predict murder arrests with these given features/variables. This is carried out with linear regression analysis.

The dataset has 1000 observations, each containing ten variables. Every variable is non-categorical variable. Five of these observations were found to include missing value(s) and they were removed.

The purpose of exploratory data analysis (EDA) is to explore dataset and make possible interesting and beneficial observations from it. Here we investigate dependent variable Murder and four explanatory variables Assault, UrbanPop, Traffic and CarAccident both visually and numerically. These results are presented in the figures 1 and 2, and table 1.

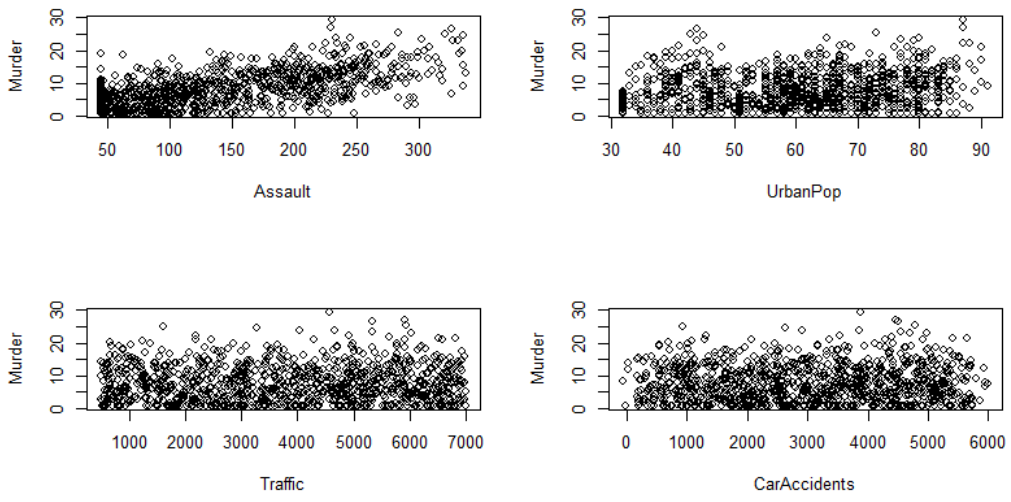

Figure 1: Four variables explaining Murder.

In the scatter plot figures it's demonstrated if the four explanatory variables could explain the dependent variable Murder. Assault and Murder have some positive linear correlation, since when the other variable's value is increasing, so does the other. That's why assault could be a good one to choose for explanatory variable to explain Murder. Traffic and CarAccidents don't show any dependency to Murder, so they can probably be excluded from the model. Instead UrbanPop is not that clear. Based on the figure it might have a little correlation with Murder, in other words the Murder values are very slightly increasing when UrbanPop is increasing, but that should be investigated more to decide whether it's suitable parameter to our model or not.

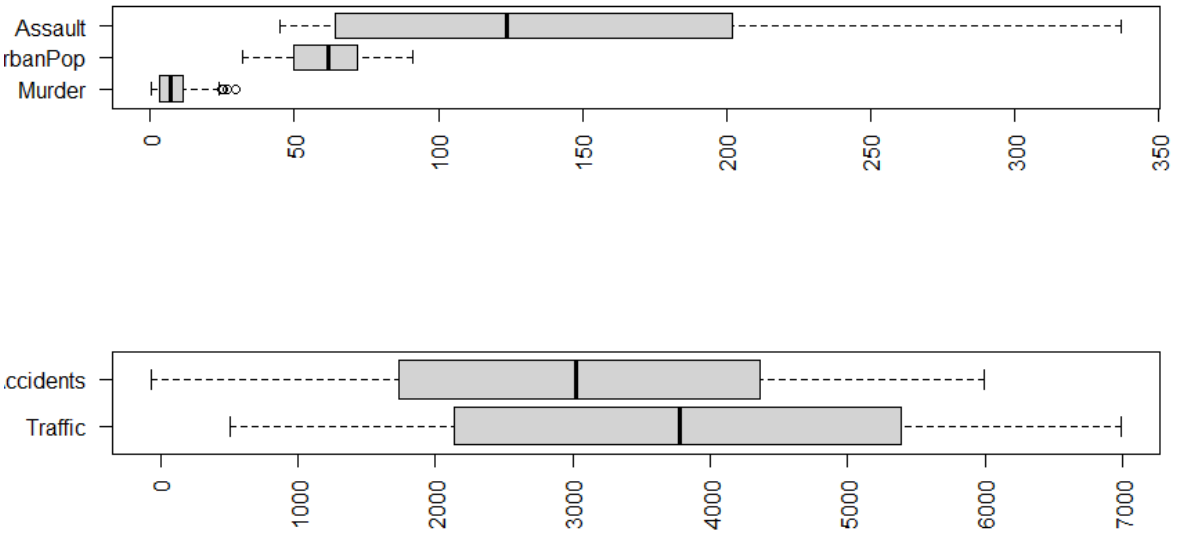

Figure 2: Boxplots of five chosen variables.

Boxplots shows measures of location and tells information about symmetry of data and possible outliers. From the figure 2 it can be said that all the variables are pretty symmetrical. However, the interquartile range, and difference between minimum and maximum values, are quite huge for CarAccidents, Traffic and Assault. This can be seen from the table 1, too, when looking at standard deviation, minimum and maximum values. In addition, it seems like that Murder-variable may include a couple of outliers.

In addition to the figures, from the table it can be seen one interesting value at CarAccidents: minimum value for car accidents is negative. There's not known reason

for that, it can for example be due to some wrong previous inserted number that is fixed later, or it can be just a typo.

Table 1: Numerical summary of chosen variables.

Variable	MIN	MAX	MEAN	STD
Murder	0.5	29.5	7.747437	5.534178
Assault	45.0	337.0	137.944724	78.755728
UrbanPop	32.0	91.0	60.879397	14.445747
Traffic	503.0	6991.0	3766.553769	1910.159299
CarAccidents	-66.0	5991.0	3004.219095	1551.810140

(In addition, we should investigate if there's relationships between explanatory variables, but the correlation analysis is carried out on the next step, so it's not discussed here yet.)

In conclusion to these observed parameters, the Murder could be modelled with linear regression using variable Assault and perhaps UrbanPop. For a more specific model, negative values and possible outliers should be investigated more and some further action for them is recommended.

1.3 – 1.5 Correlation analysis

Correlations between all variables are presented as a matrix format in the figures 3 and 4. Some dependencies to our dependent variable Murder are found (which is a good thing), but additionally there're some correlation between explanatory variables (which is a bad thing).

Starting with the correlations with the variable Murder. Based on the correlation values, Murder has highest linear associations with Assault (0.64), Drug (0.39) and UrbanPop (0.12). Other variables, Traffic, Cyber, Kidnapping, Domestic, Alcohol and CarAccidents don't have even that much effect on Murder. Noteworthy is that none of these correlating values is relatively high, meaning that even though there's found some correlation, the model with these parameters might not be very accurate. From

these variables probably Assault is only suitable explanatory variable for the linear model.

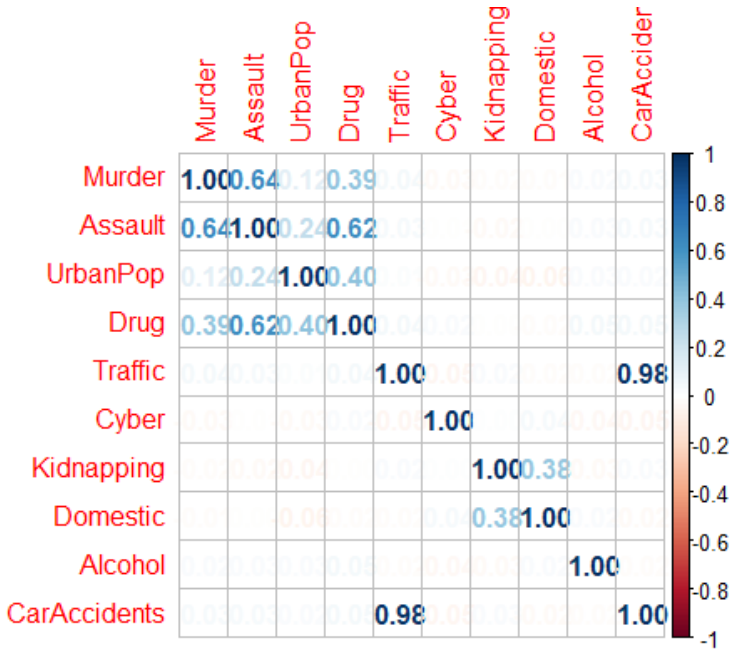

Figure 3: Correlation matrix of all variables.

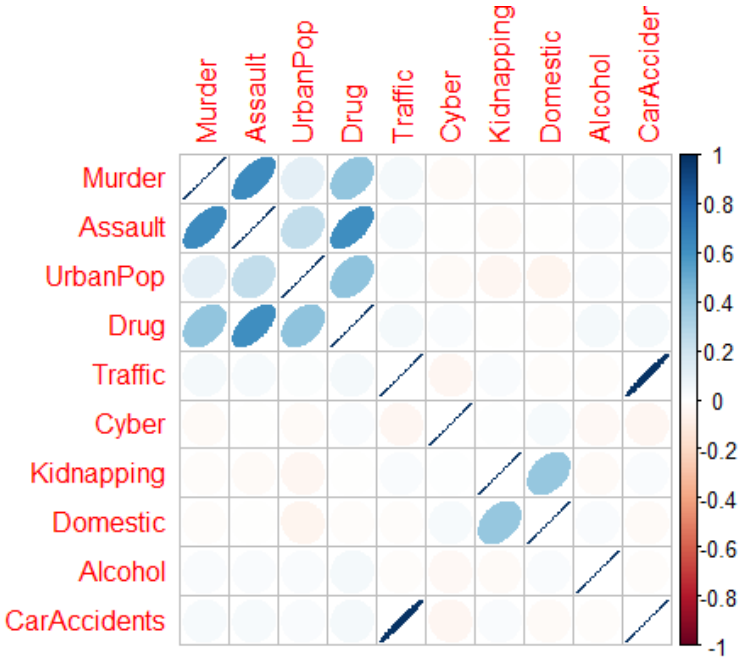

Figure 4: Visualizing correlations between all variables.

Then correlations between other variables. There's found a couple of correlations between explanatory variables, but the strongest one (0.98) is between CarAccidents

and Traffic. The other mention worthy correlating variable pair is Drug and Assault (0.62).

Only explanatory variable pair having high absolute correlation (>0.8) is Traffic and CarAccidents. One of these variables should be removed from the dataset before starting to calculate linear model. One way to determine that is to calculate their average correlation to all other variables and then remove the one with higher average correlation. In this case CarAccidents has a little higher average correlation to other variables, so that is removed from the dataset.

It's important to exclude highly correlated explanatory variables before fitting a linear regression model, because without that the estimated model wouldn't be trustworthy. The basic idea behind the regression analysis is to find out how the dependent variable would change, if one explanatory variable would be changed when others are staying constant (=effect of coefficients). If explanatory variables are correlated with each other, changes at one variable would significantly change the other one, too. That would make coefficient estimates become very sensitive even to small changes in the model, and p-values of the variables couldn't be trusted anymore because the accuracy of the model may vary a lot (Frost, 2022).

1.6 – 1.8 Linear regression

After removing Car Accidents, the next step is to carry out linear regression with all the other variables. The model and its results are calculated with OLS. From the results one can see calculated coefficients for the used variables, their standard errors, t-values, significances, and in addition model's R-squared values and p-value. With this information we can determine if the model is good or if it can be improved.

As a result the first calculated model has three (3) significant coefficients (Pr(>|t|) < 0.1): intercept, Assault and UrbanPop. R-squared is 0.4089 and adjusted R-squared 0.4041. The other variables are at this case non-significant, and they can be removed from the model because they don't affect the prediction much. A good manner is to do removal one-by-one and starting from the most non-significant variable, since the

coefficients and their significances changes when the model parameters are changed (in other words, when removing variables, previous significant variables may become non-significant and vice versa). The variables removed are in this order: Kidnapping, Alcohol, Domestic, Drug, Traffic, Cyber and UrbanPop (which was originally significant). Noteworthy is that UrbanPop became non-significant variable due to removals, other variables' significance levels didn't change that much.

After the improvements, thus non-significant variable removals, the final model for estimating the number of murder arrests per 100 000 is:

$$x_{Murder} = 0.04x_{Assault} + 1.57,$$

where x_{Murder} is the estimated number of murder arrests and $x_{Assault}$ is the number of assault arrests. The fitted model for the datapoints of Assault and Murder is presented in the figure 5.

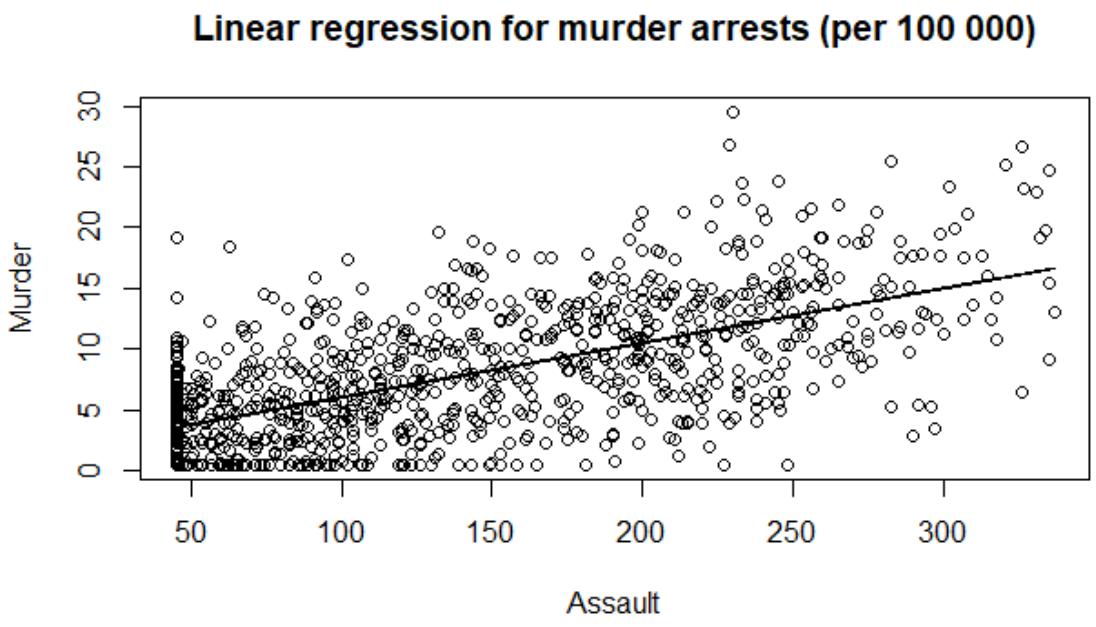

Figure 5: Fitted linear model to predict Murder arrests based on Assault arrests.

The R-squared value for this is 0.4062 and adjusted R-squared 0.4056. It's noteworthy that even if we improved the model, these R-values didn't change a lot (actually adjusted R-value became a little higher). Because R-value is closer to 0

than 1, our model is not that good predictor for the murder arrests, even though the plot looks okay.

1.9 Residual analysis

When conducting linear regression, there're 6 properties related to variables and residuals which should be considered. The one concerning the variables was already checked, meaning that the used explanatory variables are not linearly dependent. The remaining five can be investigated with different test related to residuals' mean, variance, covariance, and distribution. The residuals are presented in the figure 6, and already from that can be seen that something's wrong.

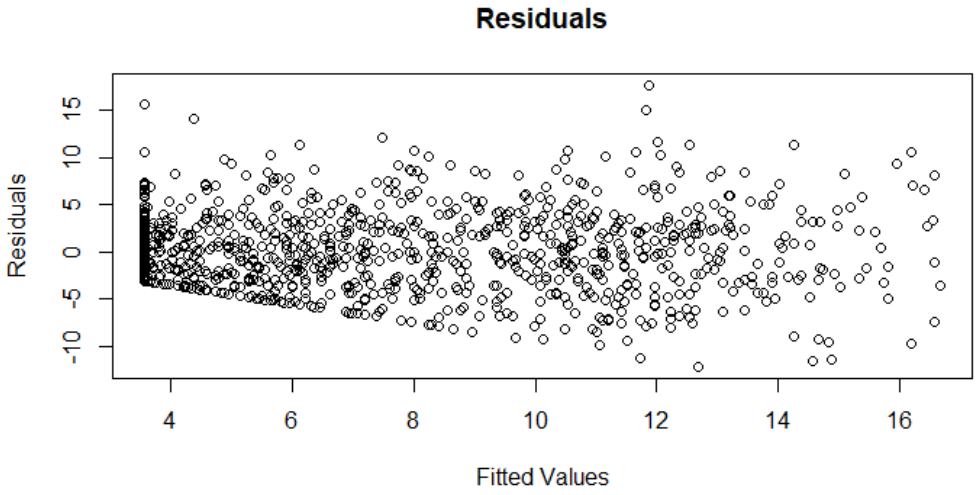

Figure 6: Residuals.

1. The residuals have zero mean

It should be tested that the residuals have zero mean. The calculated mean for the residuals is -2*e-17, which can be taken as a zero. This test passes.

2. The variance of the residuals is constant and finite

The variance being constant and finite can be tested with Breusch-Pagan test. The calculated test value is 45.66 with p-value 1.4e-11. Because p-value is small (<0.01), the null hypothesis (=homoscedasticity, error variances are equals) is rejected. This means that heteroskedastic is assumed, and this OLS linear regression analysis is not usable (Glen, 2022).

3. The residuals are linearly independent of one another

The independence of residuals can be tested with Durbin Watson test. The calculated D-W Statistic value is 2.02 with p-value 0.79 and autocorrelation -0.01. Because test statistic is a bit over 2, there's a little negative autocorrelation. However, a small range around the statistic value can be considered as normal (Kenton,2022), and the p-value is high (>0.05), so the null hypothesis can be accepted and there's no autocorrelation.

4. There's no relationship between the residuals and explanatory variables

The possible correlation between residuals and explanatory variables can be examined likewise the explanatory variables previously. The correlation value is -4.13e-17, so there's no relationship between residuals and explanatory variables and this test passes.

5. The residuals are normally distributed

The residuals' normality of distribution can be tested with Jarque-Bera test. The skewness statistics is 0.36 with p-value 3.18e-06, kurtosis statistic is 3.36 with p-value 0.02, and x-squared is 26.96 with p-value 1.40e-06. Because the final test score is far from zero and p-value is < 0.01, the null hypothesis (=normal distribution) is rejected. This indicates that linear model is not suitable fit to this data.

1.10 Conclusion

At this first part of the assignment, we aimed to find a suitable linear model to predict number of murder arrests with some of the given variables. First it was analyzed what kind of data we had and pre-processed it so that there're no missing values. Based on EDA the Murder could be linearly modelled with Assault.

Then correlation analysis was performed. The dependent variable murder had some explanatory variables (mainly Assault) correlating with it, and some of the explanatory variables correlated with each other (CarAccidents and Traffic). CarAccidents were removed from the data so that there isn't multicollinearity.

Linear regression was implemented and after reviewing coefficients and their significances, the model was improved by removing non-significant variables one-by-one starting from the most non-significant variable. As a result model to predict murder arrests, we got: $x_{Murder} = 0.04x_{Assault} + 1.57$. However, the R-squared and adjusted R-squared were such low, both about 0.4, indicating that our linear model may not be suitable.

The five properties of OLS related to residuals were tested with the help Breusch-Pagan, Durbin-Watson and Jarque-Bera tests. Mean of residuals was zero, residuals were linearly independent of one another and there's no relationship between residuals and explanatory variables. However, the variance of residuals wasn't constant, and the residuals weren't normally distributed, so only three out of five properties were met. Because the Durbin-Watson and Jarque-Bera tests failed, they indicate that this OLS regression isn't suitable for modelling murder arrests.

For the future, it's recommended to test other types of models (other than linear) for predicting Murder for example with weighted-sum-of-squares or logarithmic/polynomic/exponential model to get more accurate predictions. However, this analysis was not necessarily useless, since there were found positive correlations between Murders and Assaults, Murders and Drugs, Assaults and Drugs, and Traffic and CarAccidents. This information can be used when allocating police resources. For example, when knowing regions where is happening lots of drug dealing, there's higher probability to happen murders and assaults than in the regions without drug problems, and more polices there would be a good choice.

Part 2: Clustering

2.1 – 2.2 Studying the data set and exploratory data analysis

For this part of the assignment, we have a dataset containing statistics of annual spending of clients (supermarket chains) in different categories of goods. The aim is to find out if it's possible to find and group similar types of clients and try to understand characteristics of these groups. This is carried out with clustering.

The dataset has 440 observations, each containing eight variables. None of these observations were found to include missing value(s). Two of the variables, Channel and Region, are categorical variables and rest six non-categorical variables. This may cause problems later because categorical variables don't work with k-means algorithm (=discrete distribution -> constant Euclidean distances).

The data is visualized with scatter plots so that relationships between every variable are shown in the figure 7. Based on these plots, number of clusters for different types of clients can't be determined visually because any datapoint groups are not separating clearly from each other. However, there seems to be correlations between some variables, for example between Grocery and Detergents_Paper.

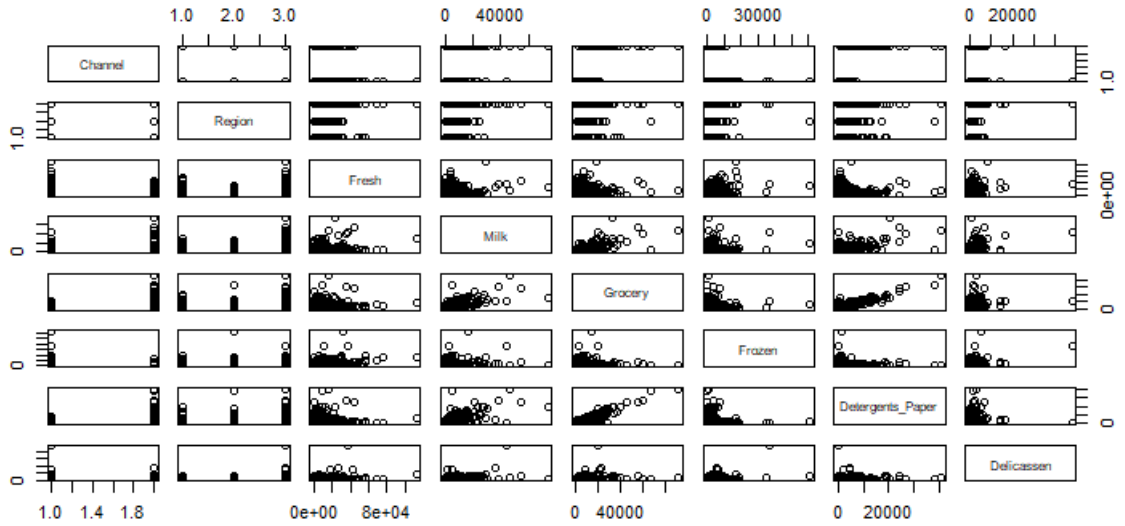

Figure 7: Scatter plots of all the data.

The deviation of the data is visualized, too. Categorical variables Channel and Region are visualized with barplots in the figure 8 and other variables with boxplots in the figure 9. Based on barplots, it seems that there're clear differences how frequently channels are used and where the sales have happened. Most used channel is 1 and most sales have been made at region 3. The boxplot of non-categorical variables looks kind of interesting. Based on that, every variable contains many outliers, and every variable's distribution seems to be more or less left skewed, meaning that the data is asymmetrical. This can cause problems with data analysis if they're not taken into account properly.

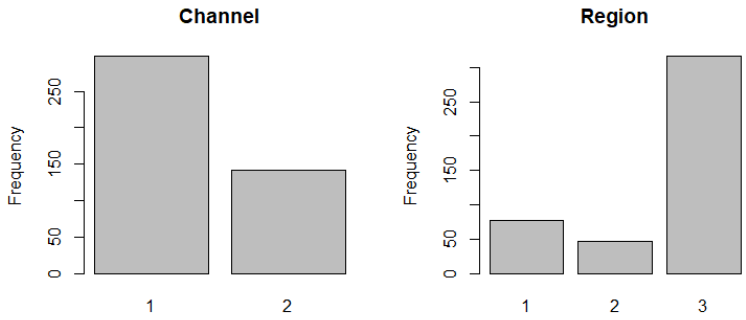

Figure 8: Barplot of the categorical data.

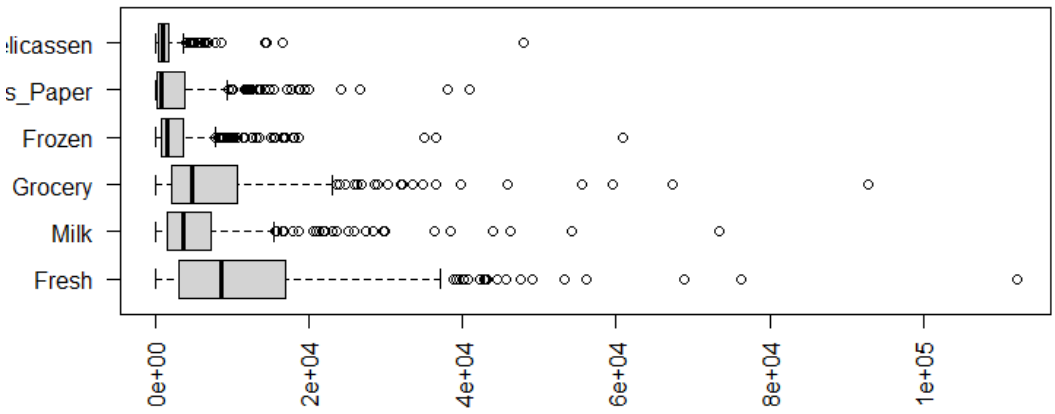

Figure 9: Boxplot of the non-categorical data.

In addition, some basic statistics of the data is calculated and presented at the table 2. Noteworthy from these values is that the variables are having different scales of values,

so for any data analysis and modelling it's recommended to normalize data so that they can be compared equally. Moreover, the standard deviation for non-categorical variables seems to be large, so the data items are locating in a wide range.

Table 2: Numerical summary of all variables.

Variable	MIN	MAX	MEAN	STD
Channel	1	2	1.322727	4.680516e-01
Region	1	3	2.543182	7.742724e-01
Fresh	3	112151	12000.297727	1.264733e+04
Milk	55	73498	5796.265909	7.380377e+03
Grocery	3	92780	7951.277273	9.503163e+03
Frozen	25	60869	3071.931818	4.854673e+03
Detergents_Paper	3	40827	2881.493182	4.767854e+03
Delicassen	3	47943	1524.870455	2.820106e+03

2.3 Correlation analysis

Compared to the part 1, this time there seems to be a lot more correlations between different variables. Correlation matrix and its visualization are presented in the figures 10 and 11.

Highest linear association seems to appear between Detergents_Paper and Grocery (as high as 0.92). Other noteworthy relationships are between Grocery and Milk (0.73), Detergents_Paper and Milk (0.66), Detergents_Paper and Channel (0.64) and Grocery and Channel (0.61). Others are below 0.5, so there isn't much correlation noticed.

In addition, it's interesting to detect that not all correlations are positive. Negative correlations (however, small ones) appear for example between Channel and Fresh, Channel and Frozen, Fresh and Detergents_Paper, and Frozen and Detergents_Paper.

When with linear regression the collinearity between explanatory variables is a big problem, in clustering it's not, or at least it's different. In clustering if some variables are dependent on each other, it makes some variables have a higher weight than others. And if some variables are highly correlated, they basically represent the same

concept, and that concept is then represented twice and getting twice the weight of all the other variables. That can cause the solution to be skewed. (Sambandam, 2022).

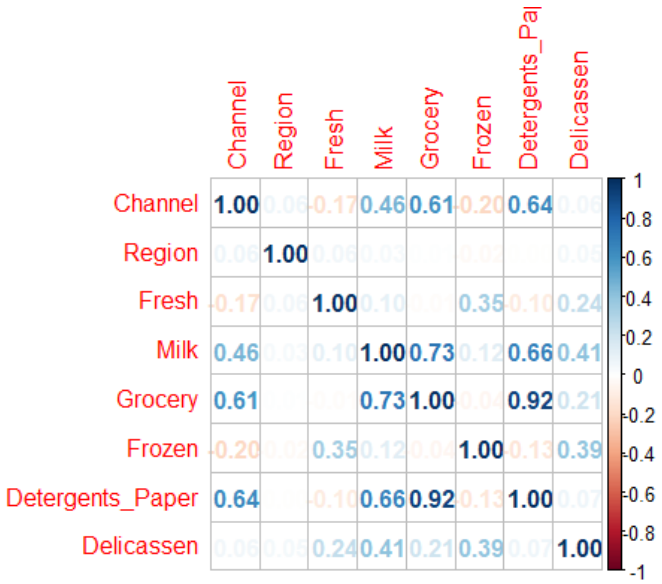

Figure 10: Correlation matrix between all variables.

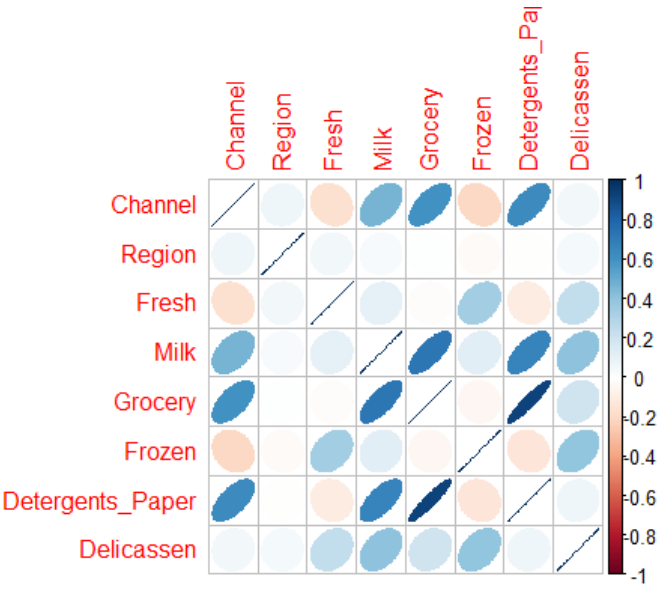

Figure 11: Visualizing correlation values between all variables.

2.4 Normalizing data

The given data is normalized by using min-max method. In practice this means that instead of variables' own scales, every observation value is now scaled to between 0 and 1.

When clustering (and often in other methods, too) it's helpful and even recommended to normalize data before calculations and data analysis. Often the dataset contains different kinds of data, for example items having different units, so different variables cannot be compared equally directly. This is important especially in clustering, because it does matter how far data items locates from each other (Euclidean distance), so if the normalization wasn't done, the variables with largest scales would have more weight than the variables with smaller scales. But with normalizations, different kinds of variables with different units and scales can be compared and modelled equally.

2.5 Finding optimal number of clusters

The optimal number of clusters is solved with k-means algorithm together with four methods: Elbow method, Silhouette method, Calinski-Harabasz Index and Gap statistic method. Determining the most optimal number of clusters can be done visually from the plots, where Total WSS / Silhouette value / CH Value / Gap Value is demonstrated in a relationship with the numbers of clusters. These are presented in the figure 12.

When using the Elbow method, the main idea is to calculate the within sum of squares per cluster and trying to minimize that sum. However, the optimal cluster number is not where the WSS is the smallest, but where the next cluster's WSS hasn't improved significantly anymore. This can be seen from the graph as a fold, "elbow". When using Calinski-Harabasz method, the main idea is to calculate how similar one item is to its own cluster and compared to other clusters. Now instead of minimizing that value, the higher value is better, because it means that clusters are dense and clearly separatable (Calinski-Harabasz Index, 2022). This can be seen from the graph as a high peak.

With Elbow method, the optimal number can be found where after that the total WSS improvement is comparably low and based on the graph that could be 3. For the other methods the optimal number can be found where the "peak" is highest. With silhouette method that would be 3, too, but with Calinski-Harabasz Index it's 9 and with Gap Statistic 2.

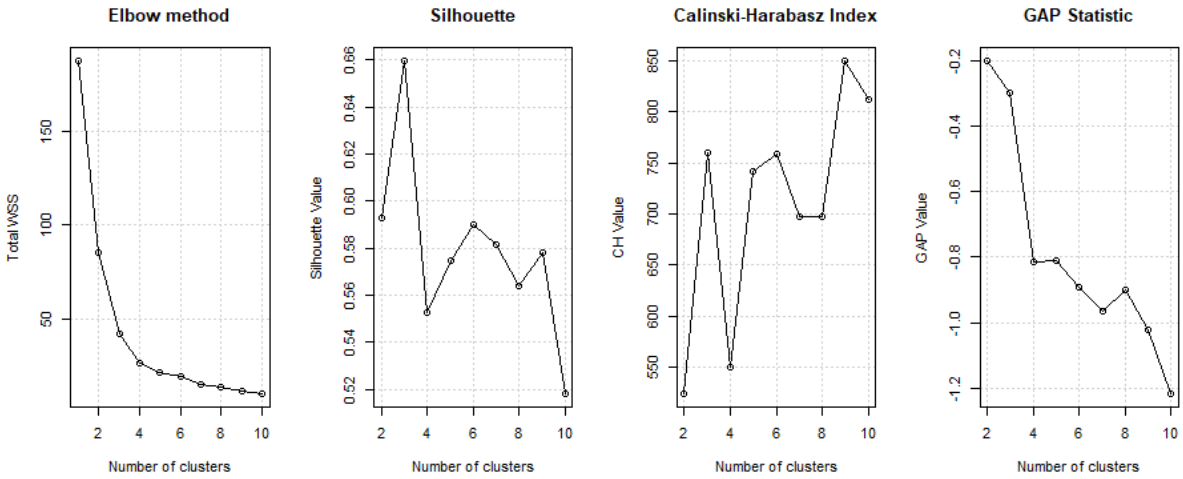

Figure 12: Determining the number of clusters with different methods.

So, it's not certainly clear which one to choose to be a good number of clusters, but I would suggest that it's approximately three. That's because 1) two out of four methods suggest that as an optimal number, and 2) when comparing the other two methods' graphs a cluster number three seems also quite a good option, too (with Gap statistic it's the second highest peak, and with Calinski-Harabasz index it's the third highest peak).

2.6 K-means algorithm and clustering results

Now when we know the optimal number of the cluster, we can present the same scatter plot as in the beginning but now with found three clusters after running k-means algorithm to unnormalized data. From the figure 13 can be seen how the wholesale company could divide their clients into three different segments. All these three groups can be identified when comparing different variables, but in most cases the segment borders are not visually clear. In addition, K-means algorithm has some problems with categorical data, so probably the client segment modelling could be done better than this.

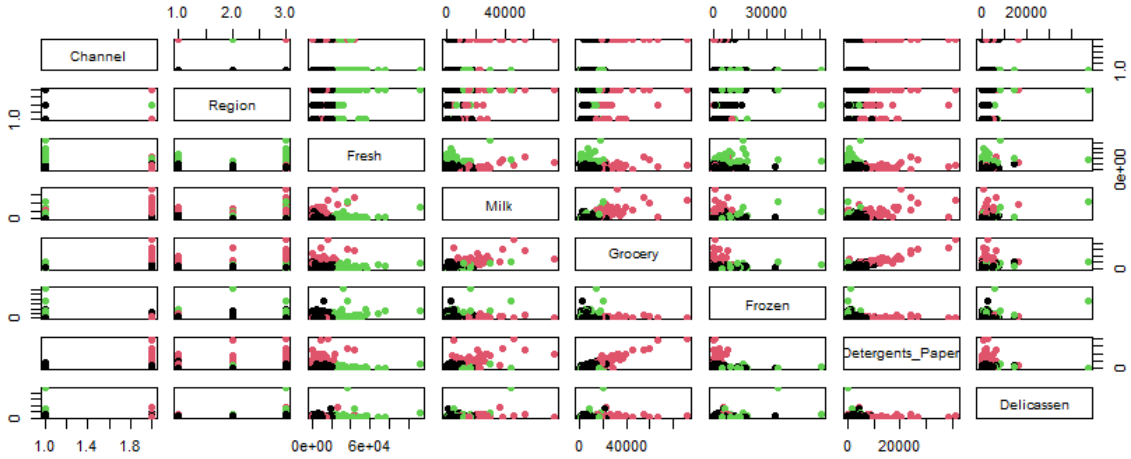

Figure 13: Data classified into three clusters.

The clearest and best separated clusters can probably be seen with Fresh and Milk, Fresh and Grocery, and Fresh and Detergents_Paper. They're presented as larger scale in the figure 14. In addition, between categorical variables Channel and Region there're clear clusters seen: channel 1 reaches one (black) client group at all three regions, but channel two reaches the other client groups at different regions (red group at regions 1 and 3, and green group only at region 2). This is presented in the figure 15.

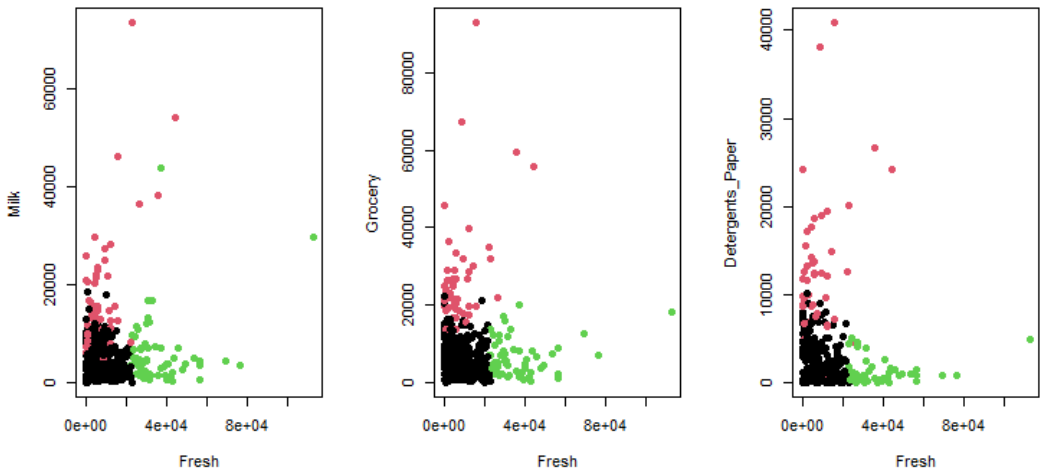

Figure 14: The clearest clusters.

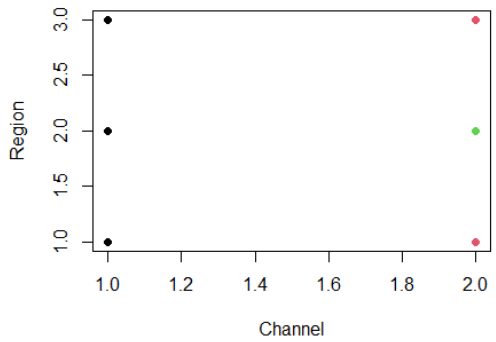

Figure 15: Clusters between Channel and Region.

When comparing all clustered data in figure 13, there can be seen similar patterns appearing for almost every variable. Black group is almost always in the left lower corner, meaning that this client segment spends money relatively same amount to every category of goods and altogether less than two other client segments. Then there're these two other client segments, that mostly buys more another type of goods and less the other type of goods. This is however not that clear for every variable pair.

It can be noticed that except the Grocery and Detergents_Paper variable pair, there seems to be no client group that would spend lot for any variable pairs, as the other three "corners" (spending less and less, more and less, less and more) are usually dominated by one client group.

When run the k-mean algorithm, the parameter nstart was used. This parameter makes n initial configurations and chooses the best one of them for the algorithm. With k-means, the algorithm starts with k randomly chosen centroids. Clustering results are often sensitive to the initial selection of centroids (meaning that if initial centroids vary a lot, the final solution can also vary a lot), so using high nstart-value gives some stability to the results when running the clustering algorithm multiple times. If it's not used, the clustering results can change when rerunning the algorithm.

2.7 Conclusion

At this first part of the assignment, we aimed to find some client segments with clustering based on the given data. Data was found it include both categorical and non-categorical variables in our data. Then the correlation analysis was carried out and multiple variables were found to correlate with each other, but with clustering it isn't that big problem as with classification and linear regression problems.

Then the dataset was normalized, so that the optimal number for cluster could be figured out. That was performed with k-means algorithm with four different methods (Elbow method, Silhouette method, Calinski-Harabasz Index and Gap statistic method), and the optimal cluster number was chosen to be three. After that k-means algorithm and clustering the data into three different client segments were run.

The clearest cluster borders could be seen with fresh products, and mostly based on them the groups could be defined as 1) Buys everything relatively equally and in moderation, and 2 and 3) Prefers other goods over the other goods and respectively spends more money to other goods than the other. The supermarket chains can maybe be divided into groups based on what they concentrate to sell: either a lot of different goods with less variety per product, or less variety of different goods but more options within one type of goods. However, with other variables the clusters' borders were not that clear to separate, so the characteristics described above may not be valid.

For the future, it's recommended to try to create new clustering model for example without categorical variables and additionally to investigate more possible outliers and if they should be removed. After that the clusters may have a different outcome and have perhaps clearer cluster borders. When thinking of the marketing perspective, it's interesting how the black group is reached in all regions through one channel, but those two other groups are only with channel two and in specific regions. This can be taken in account when thinking of how these findings can help the wholesale company. When the markets' purchasing behavior is known, the specific goods marketing can be targeted to the specific channels and regions. For example, it can be a good choice to advertise fresh products to the group green at region 2 with the channel 2, because it's known that they spend most money to those products, and they're reached best from these region and channel.

References

Calinski-Harabasz Index – Cluster Validity indices | Set 3. [Cited 2.11.2022]. https://www.geeksforgeeks.org/calinski-harabasz-index-cluster-validity-indices-set-3/

Jim Frost. “Multicollinearity in Regression Analysis: Problems, Detection, and Solutions” From Statistics By Jim: Making statistics intuitive. [Cited 2.11.2022]. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Rajan Sambandam. “Cluster Analysis Gets Complicated” From TRC Market Research. [Cited 2.11.2022]. https://www.greenbook.org/marketing-research/cluster-analysis

Stephanie Glen. "Breusch-Pagan-Godfrey Test: Definition" From StatisticsHowTo.com: Elementary Statistics for the rest of us! [Cited 2.11.2022]. https://www.statisticshowto.com/breusch-pagan-godfrey-test/

Will Kenton. “Durbin Watson Test: What It Is in Statistics, With Examples”. [Cited 2.11.2022]. https://www.investopedia.com/terms/d/durbin-watson-statistic.asp