LUT University

A220A0010 Free Analytics Environment R

Christoph Lohrmann

Assignment 2 – Data-analysis

3.11.2022

0568396

Table of contents

1. Part 1 (Regression Analysis)

1.1 Purpose

The purpose of this analysis is to study what kind of factors may be linked to murder arrests, and to give information on what other kinds of crimes may be related to it. Particularly this analysis will try to create a regression model which attempts to explain the murder arrests (dependent variable), and to use the other available data as the explanatory variables.

1.2 Data and exploratory data-analysis

Firstly, when we take a look at our data-available, we see that we have different arrest types per 100000 residents and that we have 1000 rows of observations. All of the data we have available is numeric, and we can see that there are some missing values in the observations, so we will omit those observations from the analysis so they will not cause any. This will leave us with 995 observations.

If we look at the min and max values, means and standard deviations (STD) for the dependent variable Murder, and as well as the explanatory variables Assault on the Table 1., UrbanPop, Traffic and CarAccidents, we can see that the values differ a lot between the variables. Murder variable has the lowest Max, Min, Mean and STD for all of the variables. The Assault and UrbanPop variables seem to have low standard deviation and mean values, and the Traffic and CarAccidents seem to have larger values for them. We also see that CarAccidents has a minimum value of negative 66, which could indicate of a faulty value on the data. When we check the whole data for negative values, there seems to be another negative value also in the CarAccidents variable. We will not remove these observations from the data for this analysis, but we would have to confirm for further analysis that can these values really be negative.

	Murder	Assault	UrbanPop	Traffic	CarAccidents
Min	0.5	45	32	503	-66
Max	29.5	337	91	6991	5991
STD	5.53	78.76	14.45	1910.16	1551.81
Mean	7.75	137.94	60.88	3766.55	3004.22

Table 1. Exploratory data

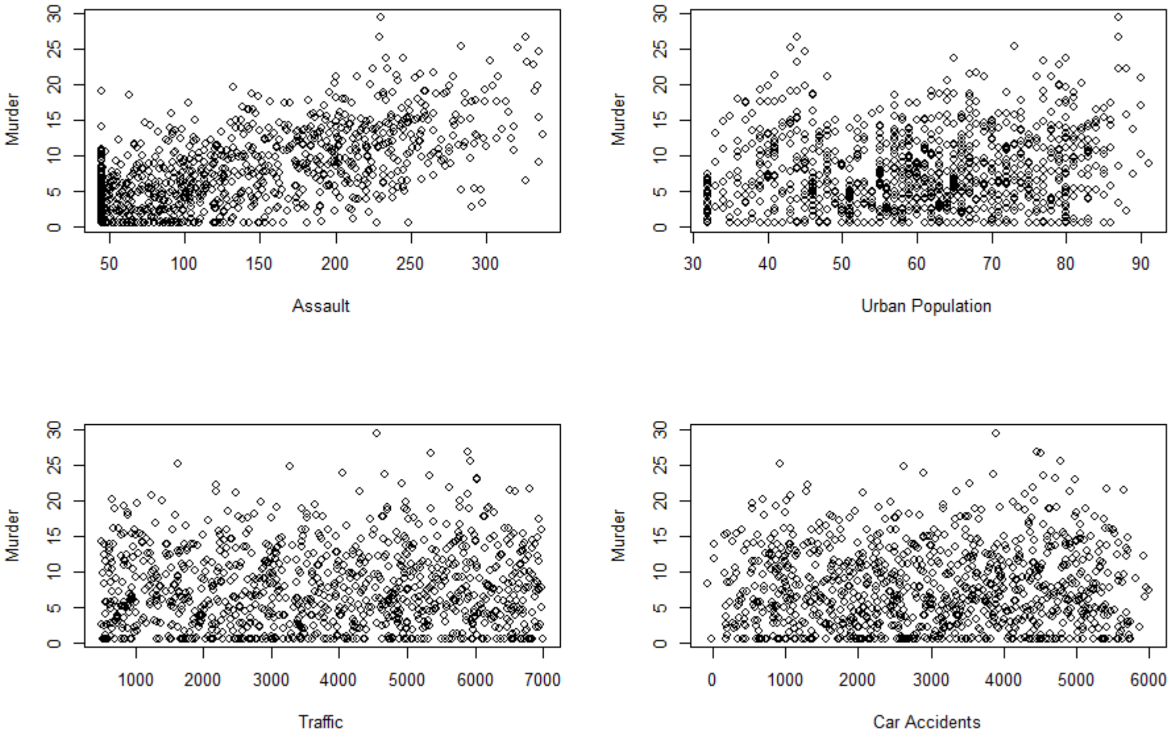


Figure 1. Exploratory plots

When we look at the plots of the explanatory variables as x-values and Murder as the y-values on the Figure 1, only the plot of Assault and Murder seems to have correlation between the variables, and it is positive correlation. There does not seem to be any other patterns that we would have to consider in further analysis.

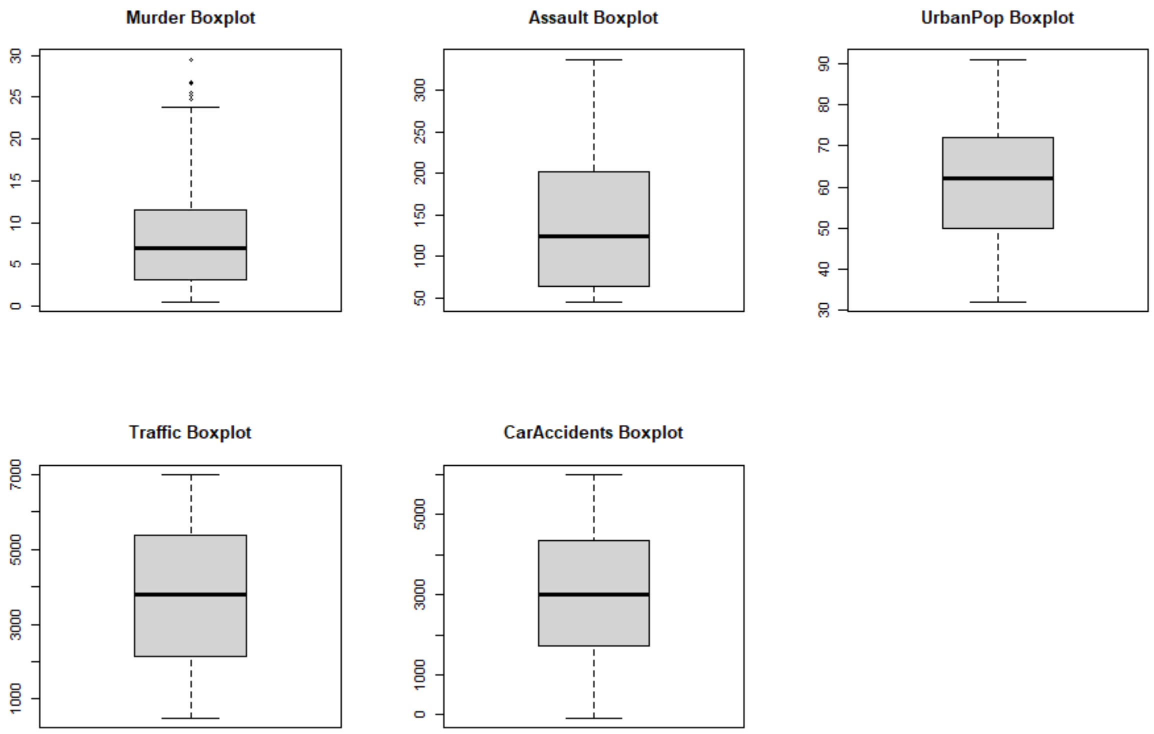


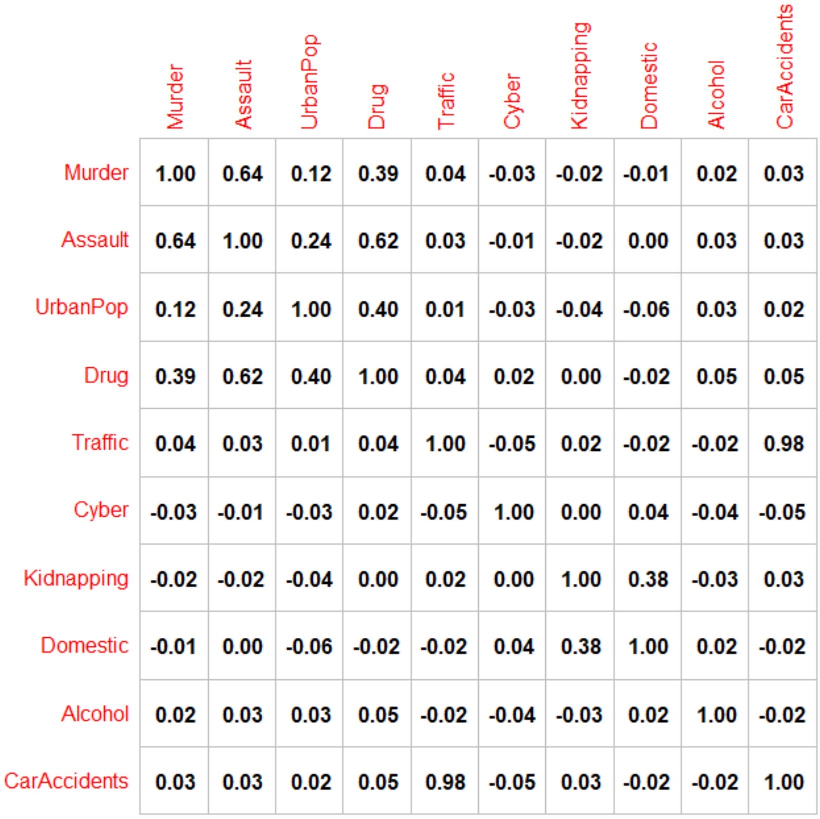
Figure 2: Box plots of the Murder and explanatory variables

The most notable mentions of the boxplots on the Figure 2. are that they seem to be mostly symmetric, and that the Murder variable seems to have some outliers in the data.However without further outlier analysis we will not do any actions for the outliers in the data, but it is worth keeping in mind for any further analysis.

1.3 Correlation

When we take a look at our plot of the correlation matrix on Figure 3., we see that the highest correlation for Murder is with Assault (0.64) and Drug (0.39) variables. Explanatory variables also have some correlations between themselves. Most notable ones are Assaults correlation with Drug (0.62) and CarAccidents correlation with Traffic (0.98).

When multiple independent explanatory variables have correlation between themselves, it is said that the regression model has multicollinearity. This is bad for the model, since it lowers the statistical significance of those independent variables, and it might cause wrong results for the model. Multicollinearity increases the standard error which those variables will have which lowers the chance of the coefficient being statistically significant. (A. Michael, 1997).



	Murder	Assault	UrbanPop	Drug	Traffic	Cyber	Kidnapping	Domestic	Alcohol	CarAccidents
Murder	1.00	0.64	0.12	0.39	0.04	-0.03	-0.02	-0.01	0.02	0.03
Assault	0.64	1.00	0.24	0.62	0.03	-0.01	-0.02	0.00	0.03	0.03
UrbanPop	0.12	0.24	1.00	0.40	0.01	-0.03	-0.04	-0.06	0.03	0.02
Drug	0.39	0.62	0.40	1.00	0.04	0.02	0.00	-0.02	0.05	0.05
Traffic	0.04	0.03	0.01	0.04	1.00	-0.05	0.02	-0.02	-0.02	0.98
Cyber	-0.03	-0.01	-0.03	0.02	-0.05	1.00	0.00	0.04	-0.04	-0.05
Kidnapping	-0.02	-0.02	-0.04	0.00	0.02	0.00	1.00	0.38	-0.03	0.03
Domestic	-0.01	0.00	-0.06	-0.02	-0.02	0.04	0.38	1.00	0.02	-0.02
Alcohol	0.02	0.03	0.03	0.05	-0.02	-0.04	-0.03	0.02	1.00	-0.02
CarAccidents	0.03	0.03	0.02	0.05	0.98	-0.05	0.03	-0.02	-0.02	1.00

Figure 3: Correlation matrix for all variables

This means that we need remove explanatory variables where the correlation is very high. For this analysis, let us consider the limit for high correlation to be 0.8, which means that the only variables for where the correlation is above that is between CarAccidents and Traffic. However, we do not want to remove both of the explanatory variables, since one of them could still be significant for the regression model. We will remove the variable which has the mean value of correlations between the explanatory variables. The Traffic has mean value of 0.242 and CarAccidents have mean value of 0.244, which means that we will remove the CarAccidents variable from the Model.

1.4 Linear regression

We used Ordinary least squared method to create our linear regression model. Our first iteration of the linear regression had very high p-values for all of the other variables than Intercept, Assault and UrbanPop (over 1). When we chose a significance level of 0.05 for the p-values, and always removed the highest p-value which was higher than our significance level, our final regression model has the explanatory variable Assault and the intercept. Our final model has the formula:

$y(x) = 0.044784 * x + 1.569684$, where the y depicts the Murder variable (per 100000) and x depicts the Assault variable.

Even though the p-values were very low for this iteration of the model, our Goodness-of-Fit (R-squared) value was 0.4062 which could indicate that the model might not have the best fit for the data.

.

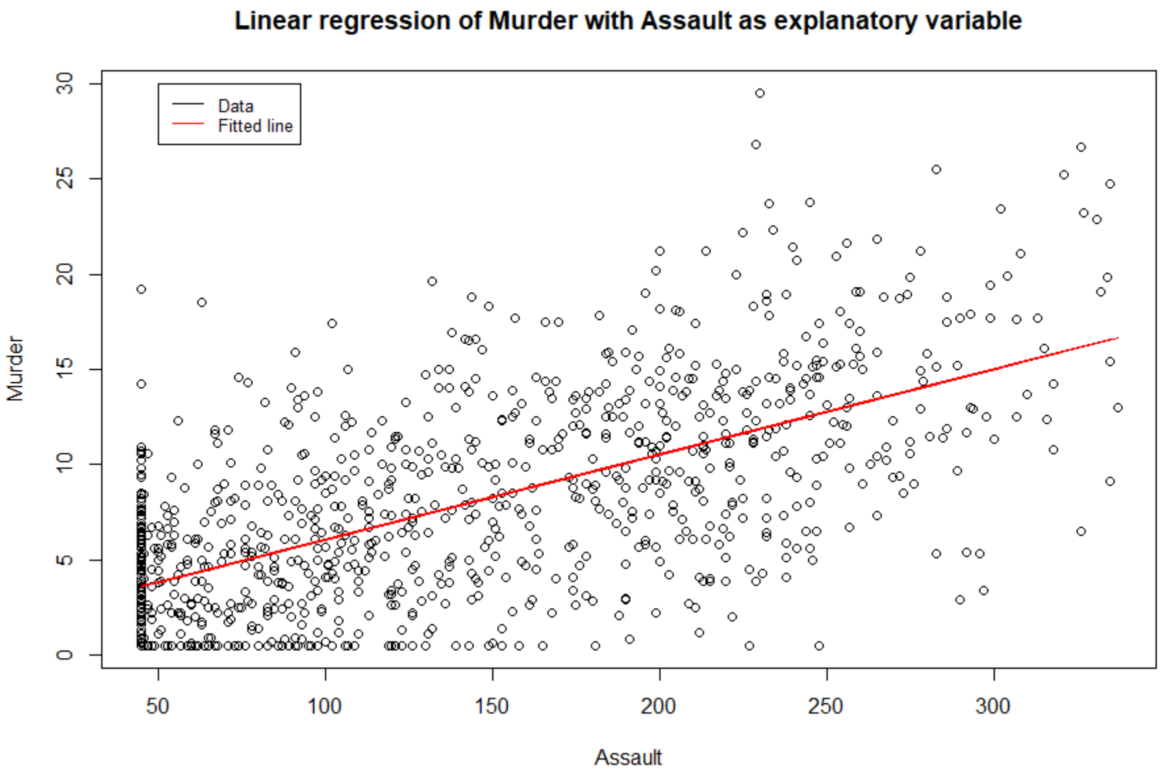


Figure 4. Fitting a line on the Murder data

When we use the model to fit a line to our data on Figure 4, we see that it fits between the data points, but there is a lot variance between the datapoints and the line

1.5 Model validation

The linear regression model OLS (ordinary least squares) model has few requirements for the variables, and the models' residuals which must be met so we can validate the model as valid. The variables requirement is that they are not linearly dependent (the correlation between them is 1 or -1), which we confirmed on the correlation section.

The first requirement for residuals in OLS is that the models' residuals has a mean value of zero. Our model has a mean value -2.386*10^-17, which is close enough to zero and this means that we can say that this criterion is met.

The second criteria is that the variance of the residuals must be constant and finite meaning that the model is homoscedastic. We used Breusch-Pagan test for heteroskedasticity to test this, which has the null hypothesis that the regression model has heteroscedasticity present. We received a p-value of 1.4*10^-11, and even with significance level of 0.01 we reject the null hypothesis, meaning that we have sufficient evidence to say that there is heteroscedasticity present in the model, which would indicate that we can not use the OLS model.

Third requirement is that the residuals must not be autocorrelated, meaning that they are linearly independent of one another. We used Durbin-Watson test to validate this, which has the null hypothesis of that there is no correlation among the residuals and the p-value we received was 0.786. Since our p-value is significantly higher than 0.05 (or even 0.1), we can conclude that there is enough evidence to accept the null hypothesis and that residuals in this regression model are not autocorrelated.

The fourth requirement for the residuals is that there must not be correlation between the residuals and the explanatory variable Assault. We received a very low correlation of -4.132*10^-17 which indicates that there is none (or close to none) correlation between the residual and Assault variable.

The fifth and last criteria is that the residuals of the model must be normally distributed. We used the Jarque-Bera test to test this, which has the null hypothesis that the model's skewness = zero, and excess kurtosis = 0. From the test we receive high test scores (Skewness test-statistic = 0.3618 and Kurtosis test-statistic = 3.3558), and very low p-values (Skewness p-value = 3.175*10^-6 and Kurtosis p-value = 0.02198). Both the high test score and low p-value indicates that we must reject our null hypothesis, and that the residuals are not normally distributed. This essentially means that we can not reliably use this OLS model to model our data.

1.6 Conclusions

The dependent variable Murder that we focused our analysation on, has notable positive correlation with the explanatory variables Assault and Drug, meaning that when the values of Assault or Drugs rise, also the Murders values increase.

The model which we created to predict Murder variables values was otherwise good, but it did not meet one of the requirements for the OLS model, which were that the residuals of the models must be normally distributed and that they are homoscedastic. This means that we can not say that you can reliably use this model to predict values for the Murder with the Assault variables values. The models R-squared value is also not very high (0.4062), which indicates that it does not give that good predictions that fits the given data.

The dependent variable Murder also seems to have some outliers on its data, which is a good thing to pay mind to in further data-analysis. Further research should consider that are the outliers in the boxplots real outliers, and if they are, should they for example just remove the observations which contain outliers.

Even if the police may not use this model to predict the Murder values in Europe, it is still useful to know that there is positive correlation between Murder and Assault or Drugs. This could also indicate that there could be made some modifications for the model so it could be used, for example transform the data in some way so the residuals meet the assumption of normality.

The models predictions could however be used to give indications on how the Murder arrests could look like, even if the predictions would not be that reliable. The police should however consider allocatin resources for other kinds of prediction methods, for example if you could cluster the given data, or to see if some other kind of non-linear model would fit well.

2. Part 2 (Clustering)

2.1 Purpose

The purpose of this analysis to find out if there are similar types of clients and group them with clustering using the data that we are given. When we have clustered the data together, we want to see if there are any relations between the different clusters and how the variables are distributed among them.

2.2 Data and exploratory data-analysis

The data that we have contains 8 different variables. Two of the different variables are categorical data: Channel with values between 1–2 and Region with values between 1–3. Rest of the variables contain only numerical values. On the Table 2. we have each of the variables, minimum, maximum, mean and standard deviation (STD) values.

	Channel	**Region**	**Fresh**	**Milk**	**Grocery**	**Frozen**	**Detergents_Paper**	**Delicassen**
Min	1	1	3	55	3	25	3	3
Max	2	3	112151	73498	92780	60869	40827	47943
STD	0.47	0.77	12647.33	7380.38	9503.16	4854.67	4767.85	2820.11
Mean	1.32	2.54	12000.3	5796.27	7951.28	3071.93	2881.49	1524.87

Table 2. Exploratory data for clustering

We see that the Fresh variable has the largest Max, STD and Mean value of the variables. Generally, all of the non-categorical variables seem to have enough variation between the values, so that we need to consider normalization for the data before we do any clustering. We can also see that there are no negative values on the dataset that we would have to worry about. The data also does not contain any missing values, so we do not have to remove any observations or consider imputation.

On the Figure 5. we can see the plots of all of the different variable combinations. There are no clear clusters visible on the plots, except on the Region and Channel which is caused by the fact that they are categorical data, meaning that all of the data gets packed on the categories. The Fresh variable seems to have the data most packed with only few outliers on the plots of the other non-categorical variables.

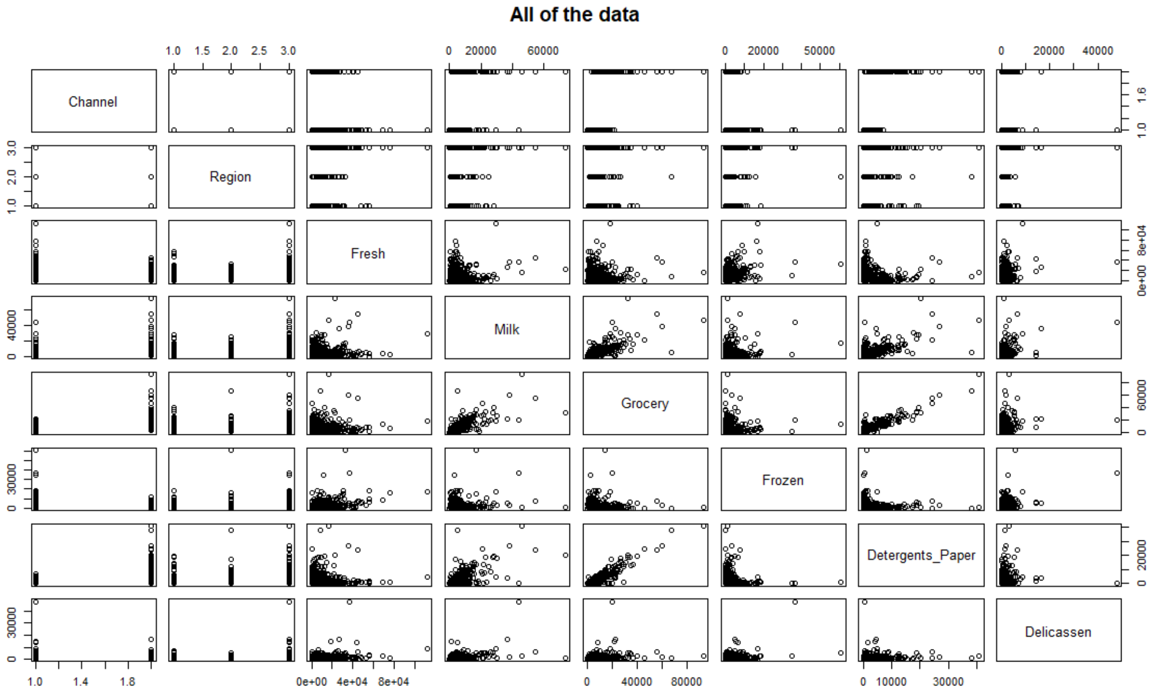


Figure 5. All of the variable combinations as plots

When we take a look at the Boxplots of the non-categorical variables on Figure 6., we see that all of the non-categorical variables seems to have outliers on the data, which might be related to the fact that they belong to a single cluster. The data also seems to not be very symmetric here and seems to be mostly left-skewed.

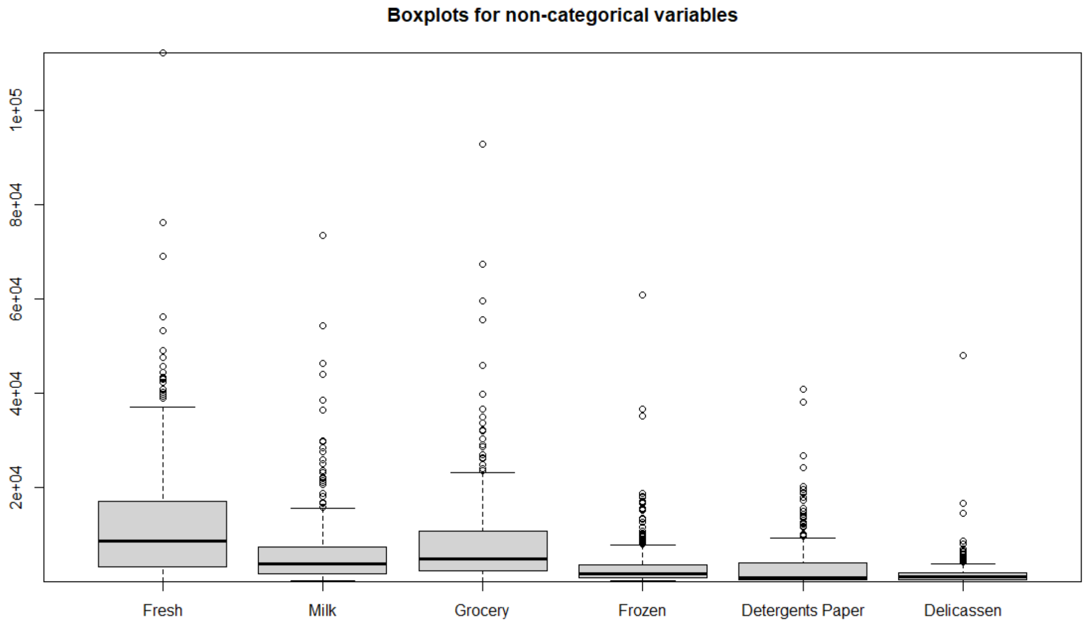


Figure 6. Box plots of the non-categorical variables

On the Figure 7. we see the bar plots for the categorical variables. We can see that the data is not distributed evenly on either of the categorical variables, and that there is clearly more datapoints on one of the categories. Channel variables category 1 has clearly more datapoints and Region variables category 3 has the most datapoints by a large margin.

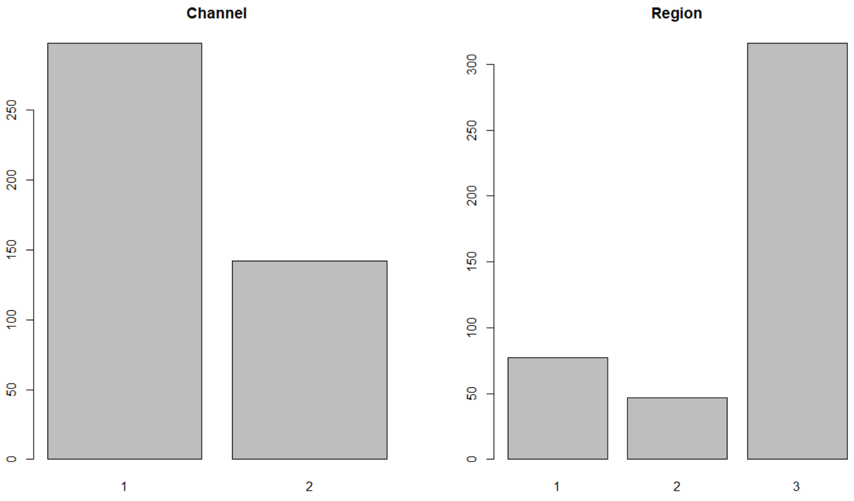
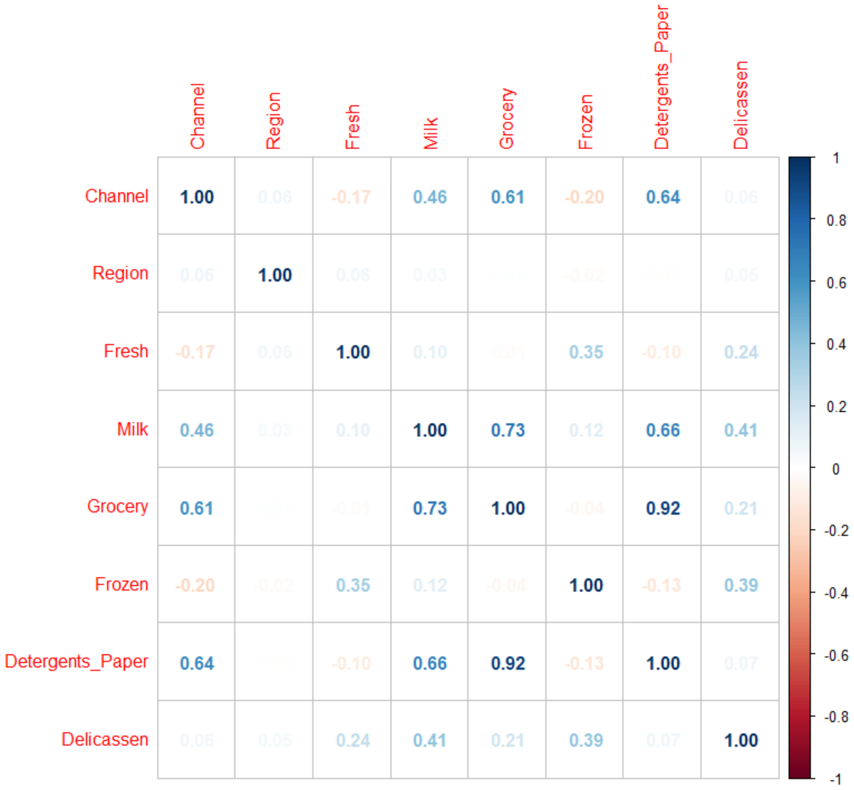


Figure 7. Bar plots for the categorical variables

2.3 Correlation

On the Figure 8. we see the correlation matrix with only the values closest to 1 and -1 visible. Most notable mentions are that the channel variable seems to have large correlation with Grocery and Detergents_Paper. The Milk variable has high correlation with Grocery and Detergents_Paper, and the Grocery variable has in addition to Milk a very high correlation with Detergents_Paper.



	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Channel	1.00	0.06	-0.17	0.46	0.61	-0.20	0.64	0.06
Region	0.06	1.00	0.06	0.03	0.01	-0.02	0.00	0.05
Fresh	-0.17	0.06	1.00	0.10	-0.01	0.35	-0.10	0.24
Milk	0.46	0.03	0.10	1.00	0.73	0.12	0.66	0.41
Grocery	0.61	0.01	-0.01	0.73	1.00	-0.04	0.92	0.21
Frozen	-0.20	-0.02	0.35	0.12	-0.04	1.00	-0.13	0.39
Detergents_Paper	0.64	0.00	-0.10	0.66	0.92	-0.13	1.00	0.07
Delicassen	0.06	0.05	0.24	0.41	0.21	0.39	0.07	1.00

Figure 8. Correlation matrix for all of the variables with the largest values visible

2.4 Normalization and finding optimal amount of clusters

Normalization is important in cases where the different variables have different size classes, for example when they have completely different units used (cm, kg, etc.). Normalization makes it so that each of the data points have the same weights when clustering, which helps when for exable one of the variables have values between 1 and 10, and other variable has values between 1000–10000. After normalization with min-max normalization for example, each of the variables would have values only between 0–1. This especially needed when we are using clustering which uses Euclidean distance as the distance measure.

We used the min-max normalization for our data, and after the data was normalized, we used Elbow method, Silhouette method, Calinski-Harabasz Index and Gap statistic to determine how many clusters we should use for this data and their results can be seen on the Figure 9.

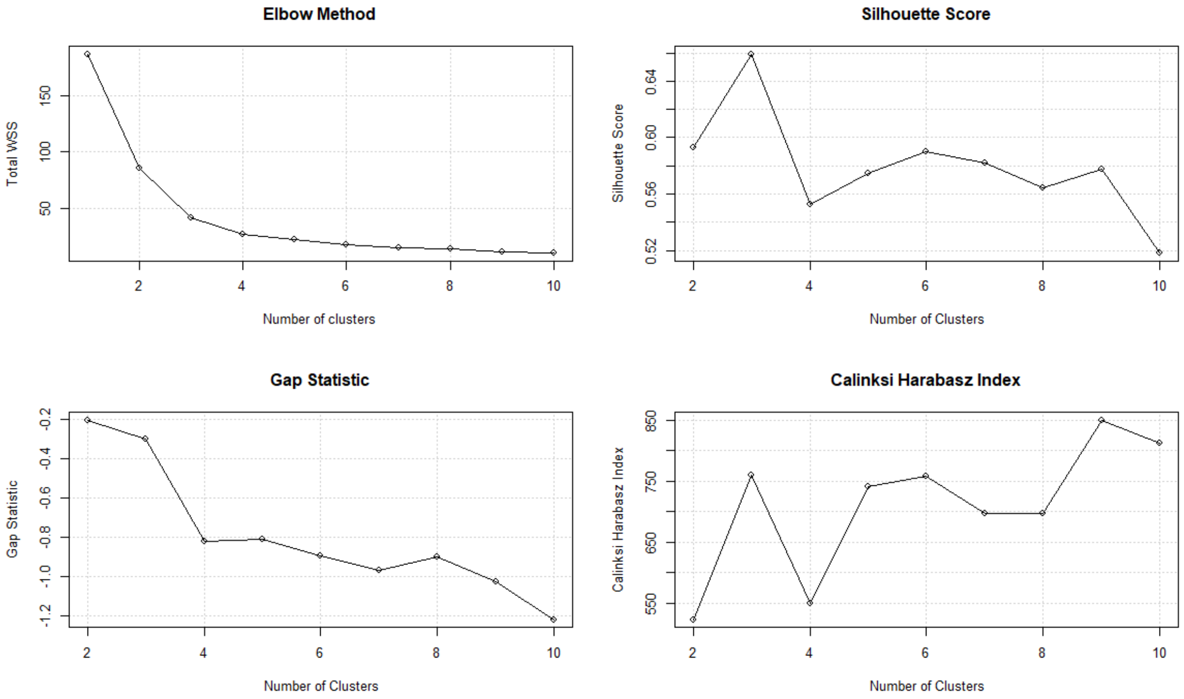


Figure 9. Different methods to define how many clusters are needed

The elbow method essentially compares the within sum of squares (WSS) value with different amounts of clusters. When we know that the more clusters you have, the smaller the WSS value will be and that the WSS’s value starts to decrease less with every new cluster we add, we can find the point where the improvement starts to clearly decrease compared to previous improvement in WSS value. This point is called the elbow point which looks similar to humans elbow from an 2D perspective. Either 2 clusters or 3 clusters would be good amount for our our data visually.

Calinski-Harabasz Index however measures the ratio between the Between Group Sum of Squares (BGSS) and Within Group Sum of Squares. The within group sum of squares means that we calculate the sum of squares for each of the clusters on their own, and use the clusters mean value for that. The Between Group Sum of Squares means that we calculate the the sum of squares using the mean of the whole data set and the means of each of the cluster to calculate that. A good cluster will have a large variance between different cluster, but small variance inside the cluster, and because of this the higher the index for Calinski Harabasz, the more likely it is that the best number of clusters is with that index. Looking at the Figure 9. the highest indexes are with 3 or 9 clusters.

Gap statistic and Silhouette score also are choosing the number of cluster with the highest score. When we compare the results between the four different methods:

- Elbow method: 2 or 3 clusters
- Gap Statistic: 2 or 3 clusters
- Silhouette Score: 3 clusters
- Calinksi Harabasz Index: 3 or 9 clusters

We can conclude that the optimal number of clusters for our data would be three clusters since it is an option for each of the different methods.

2.5 Clustering

For the clustering of our data, we use k-means algorithm with three clusters and 25 random initializations for the centroids. K-means might have problems with the categorical data we have on two variables, but in this analysis, we left both of the variables into the model.

The nstart parameter defines how many random initializations we will create for the k-means to clusterize, and then choose the one which creates best results for us. However, the nstart variables default value is 1, which essentially means that it will generate only one random centroid for the k-means algorithm to compute which might be a lot different than the optimal clusters could be. When we have more random initial configurations for the k-means algorithm, there is a higher chance of one of them being the optimal (or atleast closer to optimal) clusters. However, the larger the nstart parameters value is, the more computationally heavy running the k-means will be, which is why usually we use value of 25 for the nstart.

When we plot all of our data, we get the following Figure 10 as a result. There is no clear visual indication of which of the variables plots would have the best clusters, but it seems that the variable Fresh has the most definied clusteres with all of the different non-categorial variables.

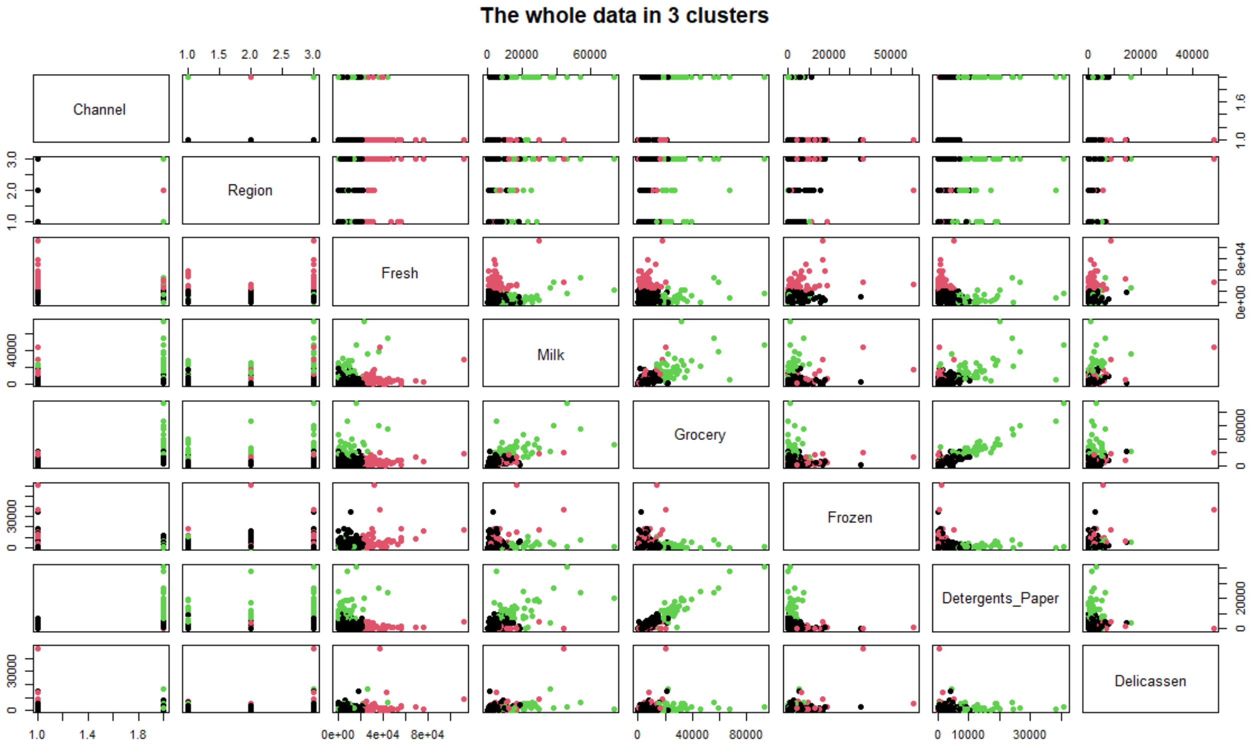


Figure 10. All of the data clustered with K-means algorithm

On the Figure 11. we can see a closer look at the clustering for the Fresh and Detergents_Paper plot. This clustering for the Figure 11. indicates that there would be 3 different types of people buying these:

- People who buy a lot of Detergents and Paper products, but low amounts of Fresh products (Blue colored)
- People who buy a lot of Fresh products but low amounts of Detergents and paper products (Red colored)
- People who buy moderately both of the products (Green colored).

There also does not seem to be a group of people who would buy a lot of both products. This same phenomenon can be observed on the Figure 12. where there is Fresh products on the x-axis and Grocery products on the y-axis. The result of these two being very similar is not however that surprising, since we know from the correlation analysis that these two variables has almost perfect correlation between themselves (0.92). This means that it is expected that they might behave similarly when compared to other variables.

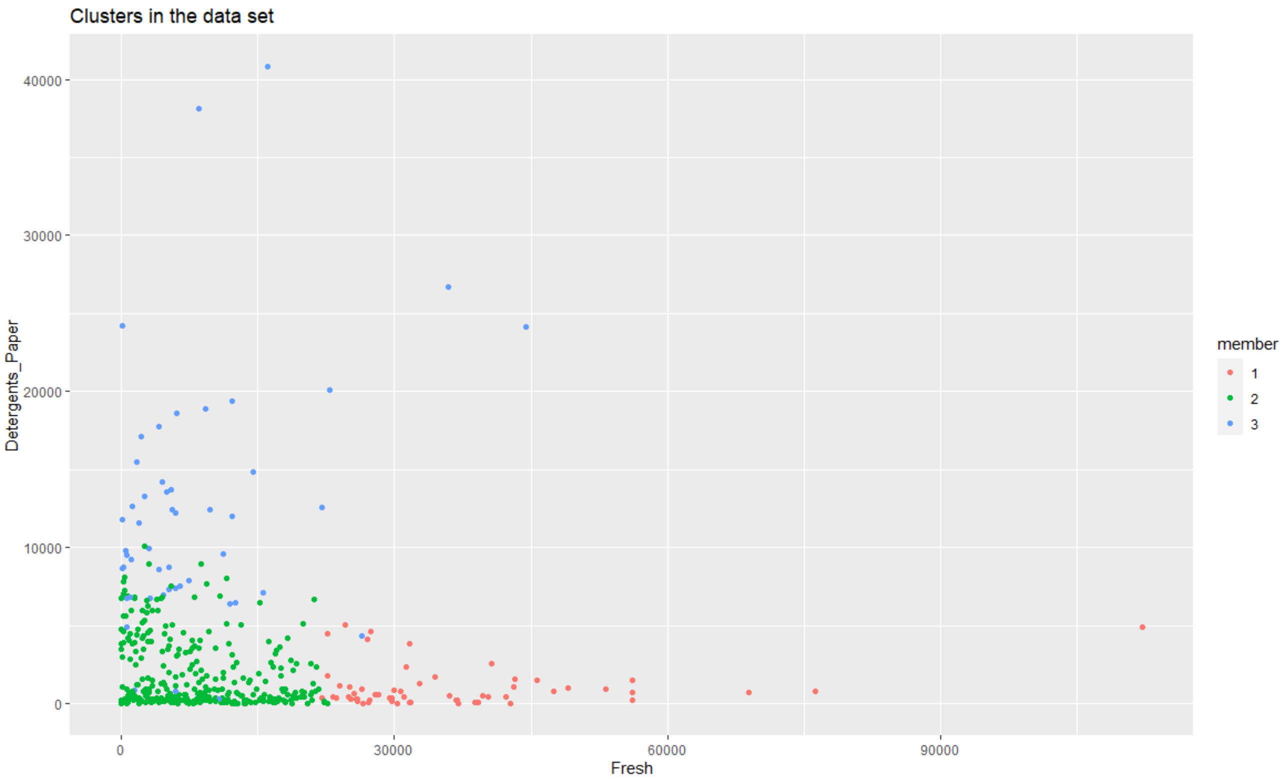


Figure 11. Fresh and Detergents_Paper plot with clusters colored

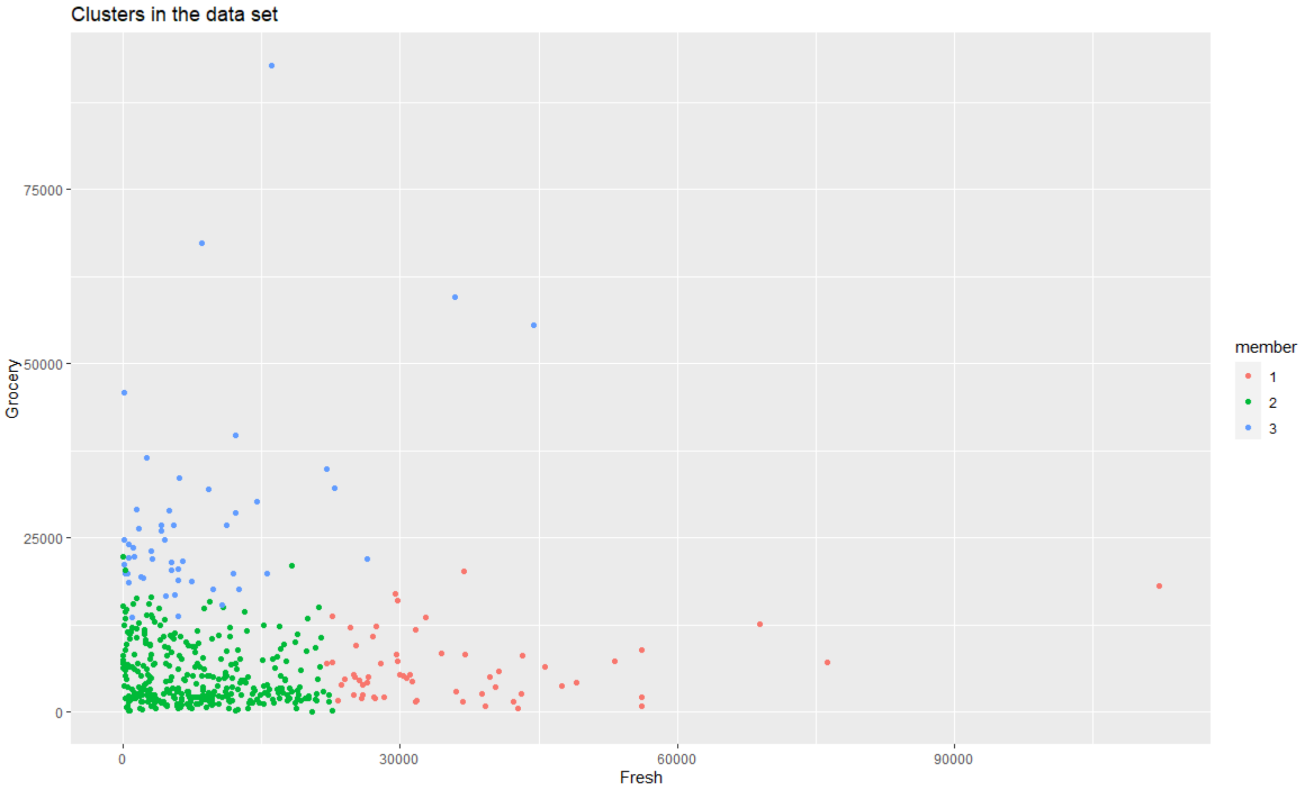


Figure 12. Fresh and Grocery plot with clusters colored

Another clear clusters can be seen on the Channel and Region variables plot on the Figure 13. We can conclude from this figure that the people using the sales channel 1 on any of the three regions belongs to the cluster 2, the people using sales channel 2 and are from region 3 or 1 belongs to cluster 3 and the people from region 2 and sales channel 2 belongs to cluster 1.

Combining this knowledge with the previous analysis of the Fresh variable, we can conclude that the people who moderately buys both Fresh and Grocery products belong to Sales channel 2 but can come from any of the regions. The people who buys a lot of fresh products but low amount of groceries uses the sales channel 2 and comes from the region 2, and the people who buy a lot of groceries but low amounts fresh products uses the sales channel 2 but belongs to either region 1 or 3.

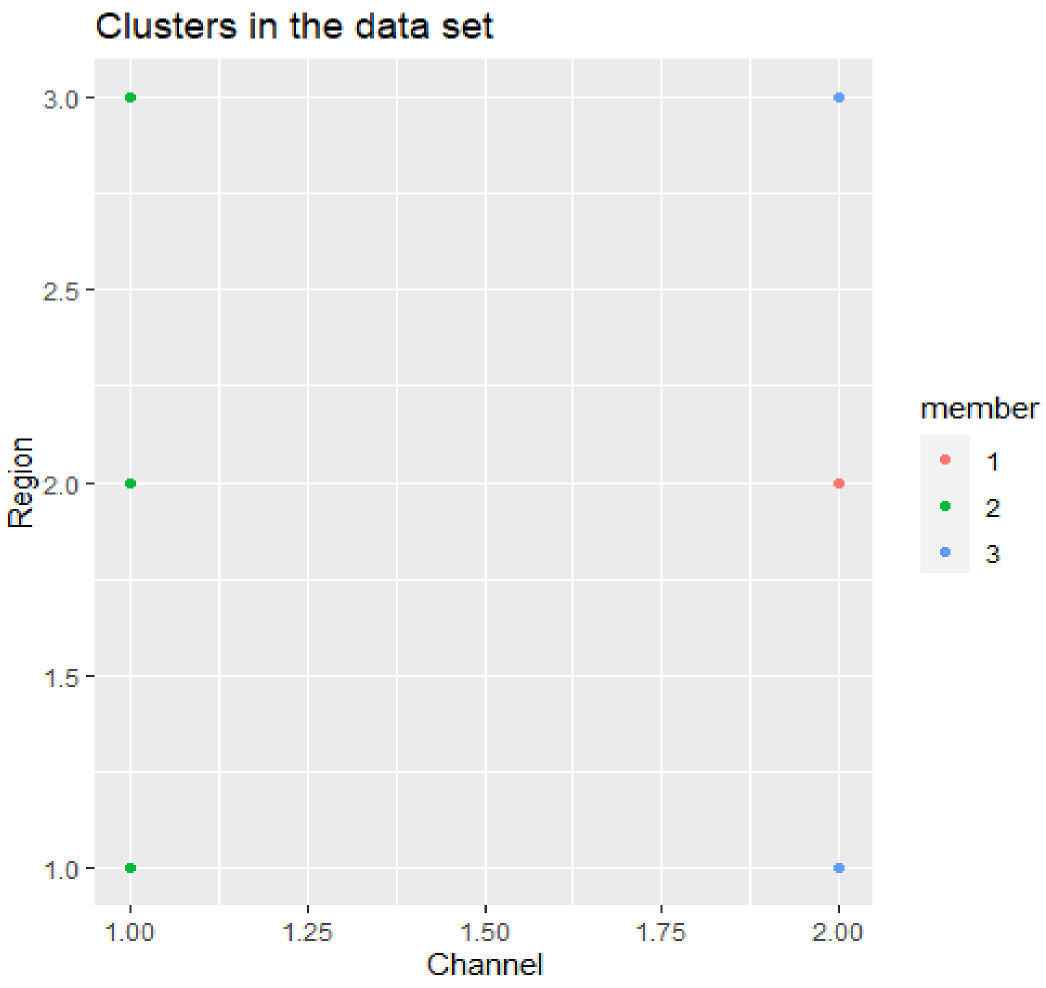


Figure 13. Channel and Region variables plot with clusters colored

If we consider how well the model distributes a single variable in the three cluster on Figure 14., we see that for the Fresh variable there seems to be some outliers on the clusters 1 and 3. The distributions seems to be slightly left skewed, and the cluster 1 clearly has the largest values on it, while the clusters 2 and 3 has the same lower values approximately. Overall, the box plots seems to contain similar values since they are comparatively short. We can say that the clustering performs decently for the variable Fresh, since there does not seem to be too many outliers and the median seems to be close to the middle of the box plot.

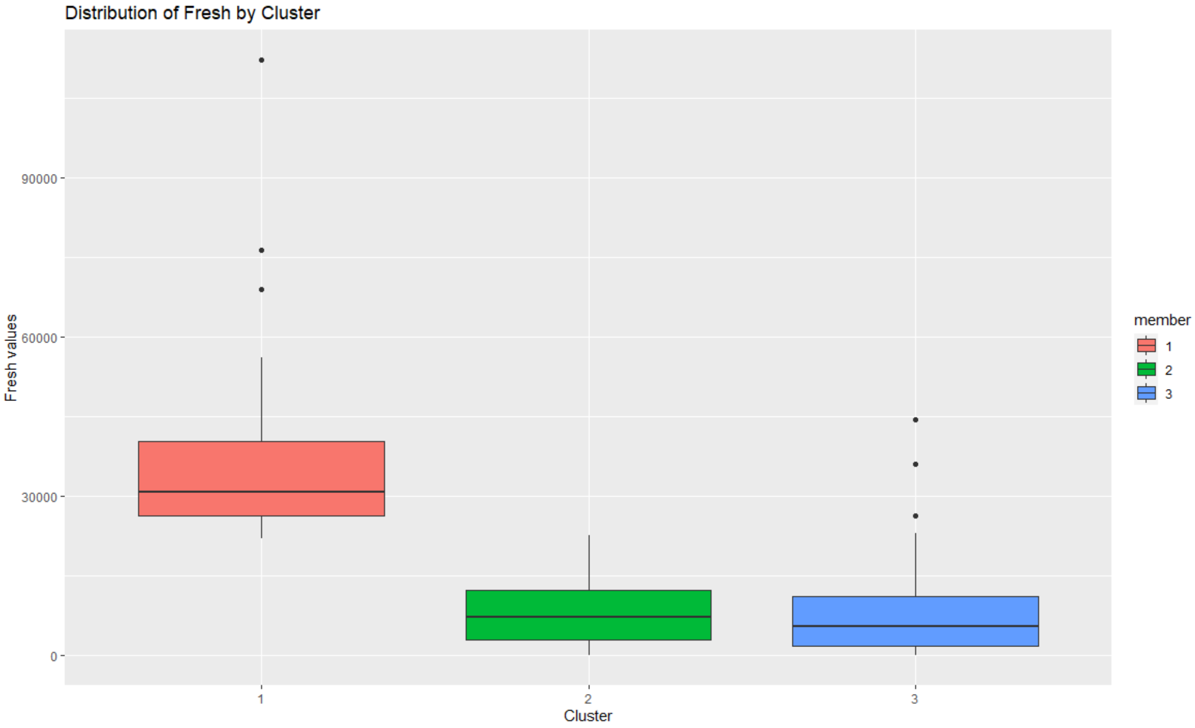


Figure 14. Boxplot of Fresh variable divided in three clusters

2.6 Conclusion

We managed to create a k-means clustering using three clusters for our data. Visually comparing the clusters in the plots of the variables, the clearest clusters can be seen on the Fresh variables plots with Milk, Grocery and Detergents_Paper. However there does not seem to be a clear distinctive clusters on the data, and most of the clusters are somehow overlapped as can be seen on the Figure 10.

We managed to cluster the Region and Channel categorical variables in a clusters where members of cluster 1 uses channel 2 and comes from region 2, members of cluster 2 uses channel 1 and comes from all of the regions and members of cluster 3 uses channel 2 but comes from either region 1 or 3.

The wholesale company can use especially the regional and channel clusters to help them decide on where they will distribute their products and which channels they can use to advertise to specific client groups. For example, they know that the channel 1 reaches best to the people who buys moderately fresh and grocery products, so they can focus their advertising on the channel 1 to focus on those people.

The wholesale company also know from the clustering that there are two groups who either buy a lot of Fresh products and low amount of Grocery/Detergent products, or lot of Grocery/Detergent products and low amount of Fresh products. They can use the knowledge that both of these groups use the sales channel 2 and comes from either the region 2 or from 1 and 3. For example now that they know that the group of people who buys the most Fresh products use the channel 2 and comes from the region 2, they know that they can use that channel to advertise Fresh products to that client group. They also know that they need to be ready to distribute more Fresh products to the region 2, since that region buys the most Fresh products. And vice versa they need to be ready to deliver more groceries, detergents and paper to regions 1 and 3, since the client group who buys most of these comes from these regions.

One example of how the wholesale company could use the clustering information for marketing could be that they could use the sales channel 1 to give paired product discounts for example for Fresh and Grocery products. Because the the group who uses the channel 1 already buys both products a lot, they can increase the sales of both products when they give out discounts on where if you for example you buy a Fresh product x, you get a Grocery product y cheaper.

One thing which should be considered in further analysis is that most of the non-categorical variables we use seems to have outliers in them. It is especially known that K-means algorithm is very sensitive to outliers and in any further clustering models we should check on what the effect of the outliers are for the clusters, and to confirm if the outliers are really outliers and to see if we need to remove them from the data.

Overall, the clustering created us some clusters that we can use to create distinctions between the client groups, but for some of the variables it did not seem to create any clear clusters that we could use for analysing. One of these variables is the Delicassen variable for example.

References

Michael Allen (1997). The problem of multicollinearity. In: Understanding Regression Analysis. Springer, Boston, MA. https://doi.org/10.1007/978-0-585-25657-3_37