Assignment 2

Free Analytics with R

Lappeenranta University of Technology

Master's Degree in Business Analytics

2022

Table of contents

1 Linear Regression Model

This section includes the regression analysis of the "dataArrests" dataset. Firstly, this section gives the information about the dataset. Moreover, this section depicts the univariate and bivariate exploratory data analysis for explanatory variables i.e., Assault, Urban Prop, Traffic, Car Accidents, and dependent variables i.e., Murder. Furthermore, the dataset will be pre-processed before creating the linear regression model which can forecast number of murders using various explanatory variables. Similarly, this section will check if the model satisfies the 5 properties for linear regression using OLS. Lastly, this section will conclude with findings and results that will help the police allocate the resources in future.

1.1 Structure and Characteristics of the dataset

The dataset i.e., 'dataArrests' contain 1000 observation and 10 variables before the elimination of the missing values. All the variables belong to either numeric class or integer class. The dependent variable is 'Murder', and the explanatory variables are Assault, Urban Prop, Drug, Traffic, Cyber, Kidnapping, Domestic, Alcohol, and Car Accident. In pre-processing step all the rows with one or more missing values are eliminated. It is very important to handle the missing values in a dataset because it drastically impacts the quality of regression model. In this case, the researcher has eliminated all the rows instead of imputing the values because there are few missing values in the dataset. After the elimination of the missing values, there are 995 observation and 10 variables.

1.2 Exploratory data analysis

The explanatory variables i.e., Assault, Urban Prop, Traffic and Car Accident as well as dependent variable i.e., Murder are used for explanatory data analysis to get better understanding of the variables in the dataset. Both the numerical and graphical approach is considered for the analysis.

1.2.1 Univariate data analysis

In univariate data analysis, each variable will be examined separately. The measure of central tendency i.e., mean, and median and measure of spread i.e., standard deviation, range, skewness, kurtosis, and interquartile range will be use in the univariate data analysis. Table 1 shows the measure of central tendency and measure of spread for all the selected variables of the dataset.

Table 1 Measure of central Tendency and Measure of Spread

	Murder	**Assault**	**Urban Pop**	**Traffic**	**Car Accidents**
Mean	7.747437	137.9447	60.8794	3766.554	3004.219
Median	6.9	124	62	3781	3025
Min	0.5	45	32	503	-66
Max	29.5	337	91	6991	5991
SD	5.534178	78.75573	14.44575	1910.159	1551.81
Variance	30.62712	6202.465	208.6796	3648709	2408115
IQR	8.25	138	22	3253	2633.5

From table 1 we can see that on average 60.87 % live in the urban European cities. The percentage that separates the high half of population to the lower half of the population in urban area is 62%. The minimum percentage of urban population in European cities is 32% whereas the maximum number of urban populations in European cities is 91%. The standard deviation suggests that on average the urban population in the cities vary by 14.45% from the mean. Moreover, the interquartile range of 22% suggest the difference between the highest percentage of urban population and lowest percentage of urban population in the middle half of the distribution is 22%.

Figure 1 shows the histogram and box plot of the urban population. From the figure we can see that the distribution can be considered as symmetric and normally distributed. From the boxplot we can see that there are no potential outliers.

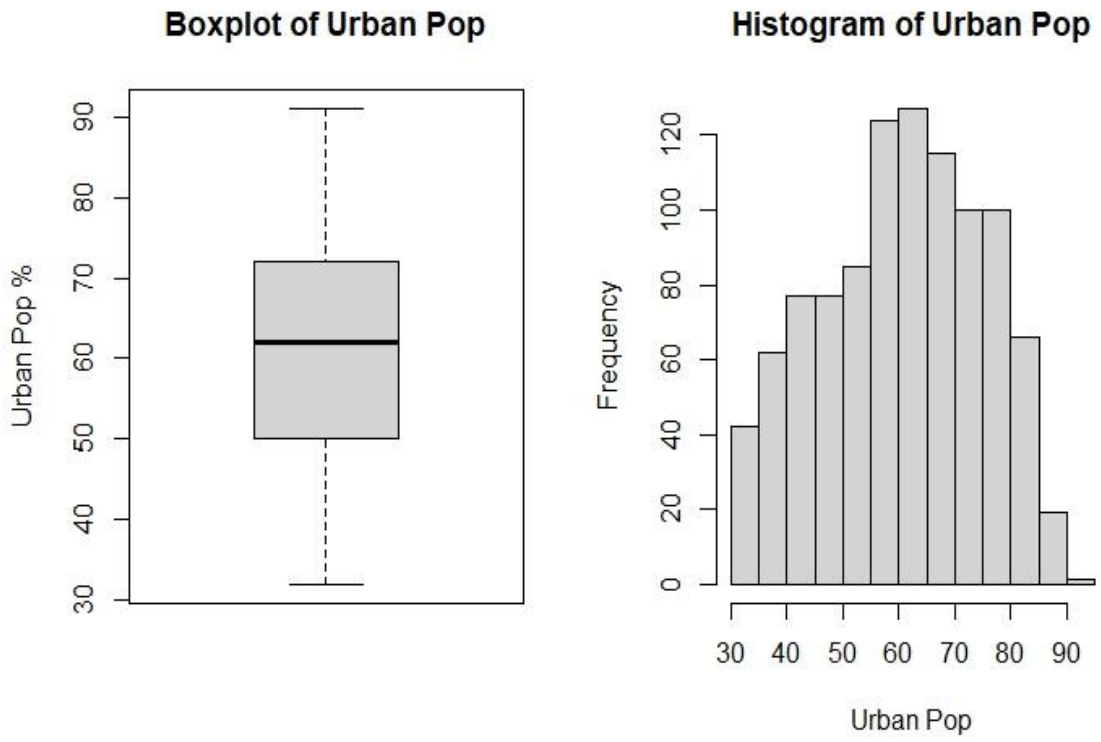

Figure 1 Boxplot and Histogram of Urban Pop

Table 1 depicts that the average number of murder arrest is 1.74 in the European cities. The minimum and maximum number of murder arrest are 0.5 and 29.5 in the European cities. Similarly, the number of murders that separate high half and low half of murder arrests in European cities is 6.9. Moreover, the Standard deviation of the murder arrests is 5.54. It suggests that on average the number of murder arrest differ the mean of murder arrests by 5.54. Furthermore, the difference between the highest number of murders and the lowest number of murders in the middle half of the distribution is 8.25.

Figure 2 shows the histogram and box plot of the murder arrest. From the figure we can see that the distribution can be considered as positively skewed. Moreover, the data is not distributed normally. From the boxplot we can see that there are potential outliers in the data. We can remove the outlier from the data to improve the quality of the model.

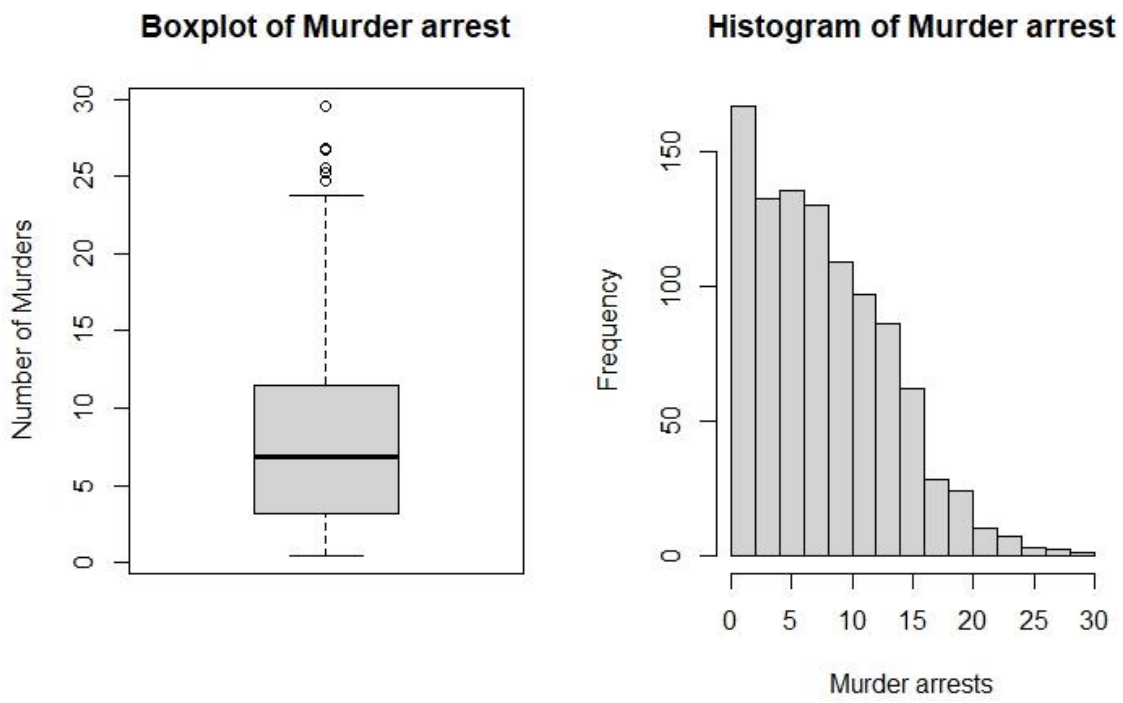

Figure 2 Boxplot and Histogram of Murder Arrest

Table 1 shows that the average number of assaults is 137.95 in the European cities. The minimum and maximum number of assaults are 45 and 337 in the European cities. Similarly, the number of assaults that separate high half and low half of assaults in European cities is 124. Moreover, the standard deviation of the assaults is 78.75. It suggests that on average the number of assaults differ the mean of murder arrests by 78.75. Furthermore, the difference between the highest number of assaults and the lowest number of assaults in the middle half of the distribution is 138.

Figure 3 shows the histogram and box plot of the assault arrest. From the figure we can see that the distribution can be considered as positively skewed. Moreover, the data is not distributed normally. From the boxplot we can see that there are no potential outliers in the data.

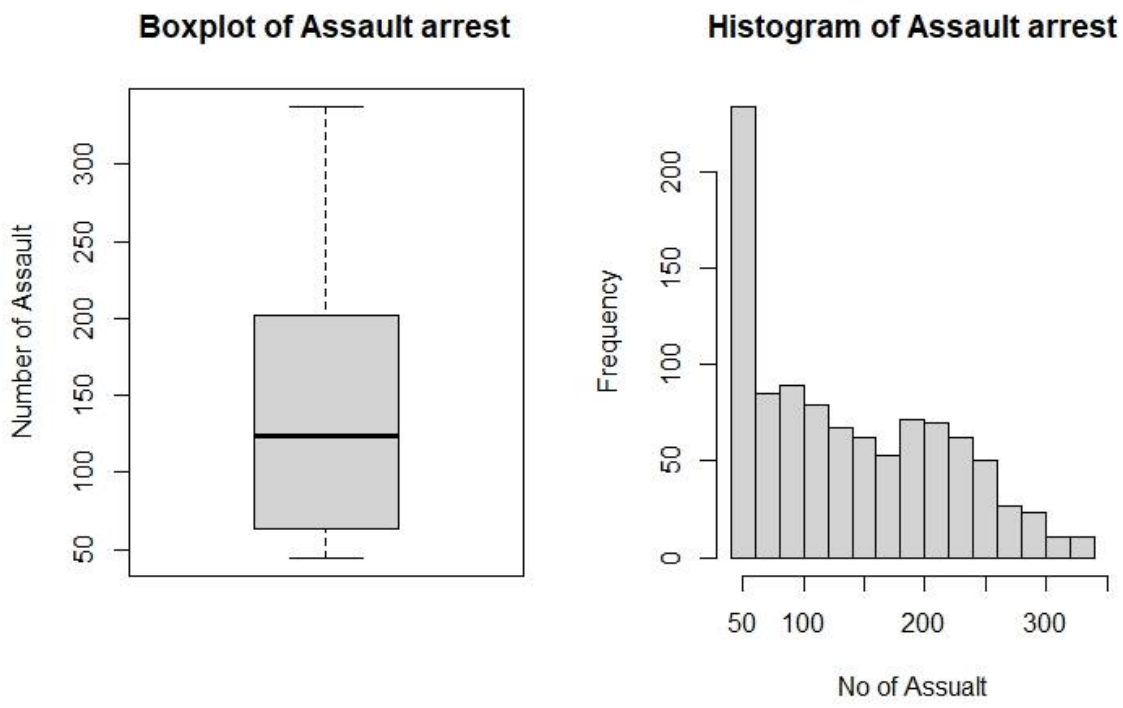

Figure 3 Box plot and Histogram of Assault arrest

Table 1 illustrates that the average number of traffic arrests is 3766.554 in the European cities. The minimum and maximum number of traffic arrests are 503 and 6991 in the European cities. Similarly, the number of traffic arrests that separate high half and low half of traffic arrests in European cities is 3781. Moreover, the standard deviation of the traffic arrests is 1910.159. It suggests that on average the number of traffic arrests differ the mean of traffic arrests by 1910.159. Furthermore, the dispersion between the highest number of traffic arrest and the lowest number of traffic arrests in the middle half of the distribution is 3253.

Figure 4 shows the histogram and box plot of the assault arrest. From the figure we can see that the distribution is not distributed normally. It can be considered as uniformly distributed. From the boxplot we can see that there are no potential outliers in the data.

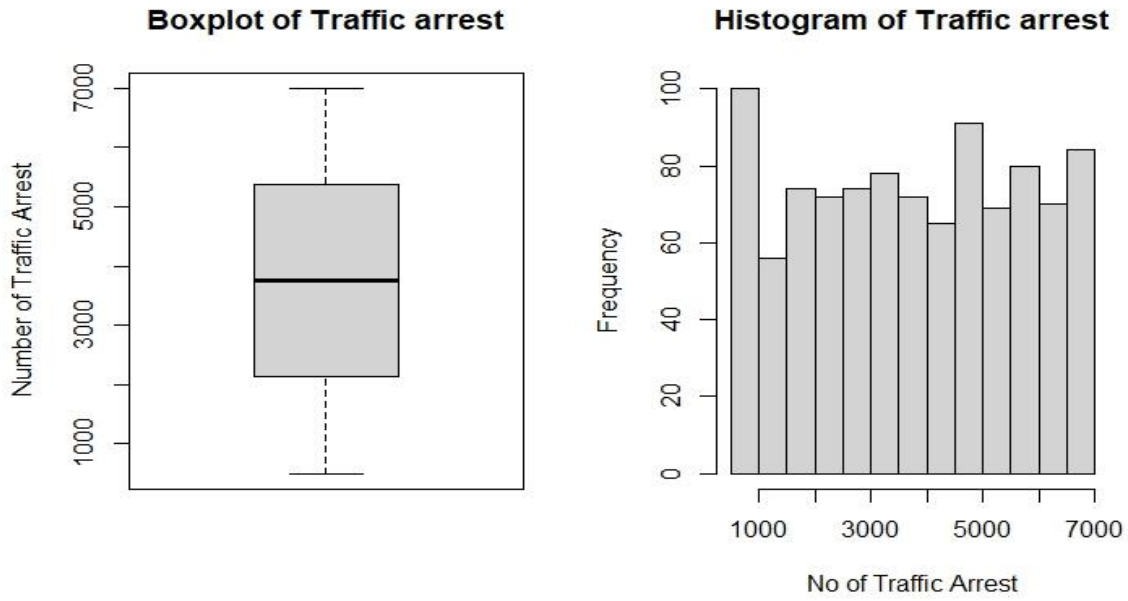

Figure 4 Boxplot and Histogram of Traffic arrest

Table 1 shows that the average number of car accidents is 3004.219 in the European cities. The minimum and maximum number of car accidents are -66 and 5991 in the European cities. Similarly, the number of car accidents that separate high half and low half of accidents in European cities is 3025. Moreover, the standard deviation of the car accidents is 1551.81. It suggests that on average the number of accidents differ the mean of car accidents by 1551.81. Furthermore, the difference between the highest number of car accidents and the lowest number of car accidents in the middle half of the distribution is 2633.5.

Figure 5 shows the histogram and box plot of the assault arrest. From the figure we can see that the data is distributed normally. From the boxplot we can see that there are no potential outliers in the data.

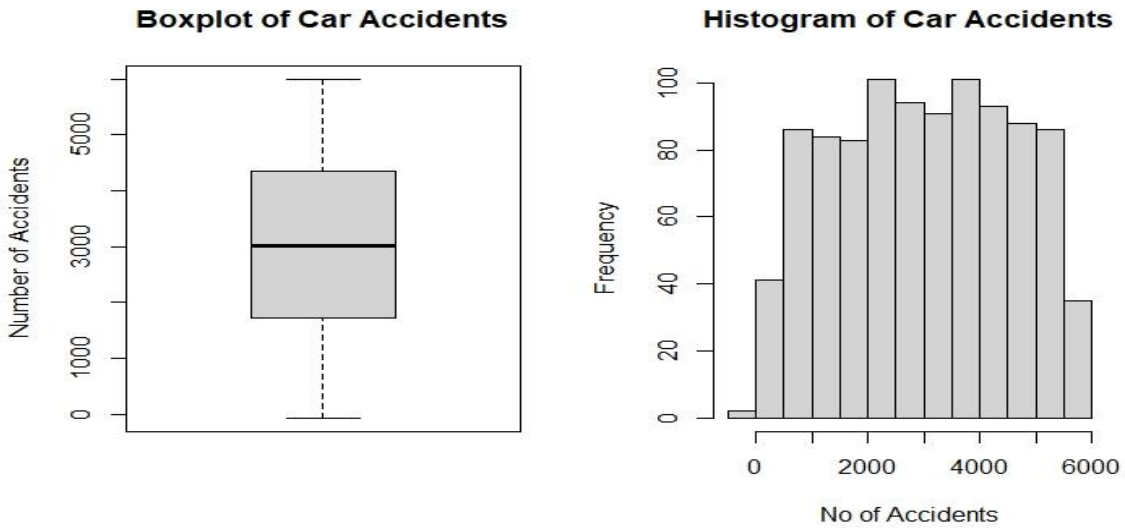

Figure 5 Boxplot and Histogram of Car Accidents

1.2.2 Multivariate data analysis

In multivariate data analysis, each variable will be analysed with the relation to other variable. The correlation matrix and correlation plot are used to show the relationship between the variables.

Figure 6 shows the correlation between the variables. The dependent variable i.e., Murder is moderately correlated with the variable Assault with the correlation of 0.64. Furthermore, the dependent variable has very weak positive correlation with Drug, Urban Population, Traffic, Alcohol and Car Accidents with the values 0.39, 0.12, 0.04, 0.02 and 0.03 respectively. Moreover, the dependent variable has very weak negative correlation with Cyber, Kidnapping and Domestic with the values -0.02, -0.016 and -0.013 respectively. Furthermore, the variable Drug arrest is moderately correlated with the variable Assault with the value of 0.62. Similarly, the variable Traffic and Car Accidents have a very strong positive correlation with the value of 0.98.

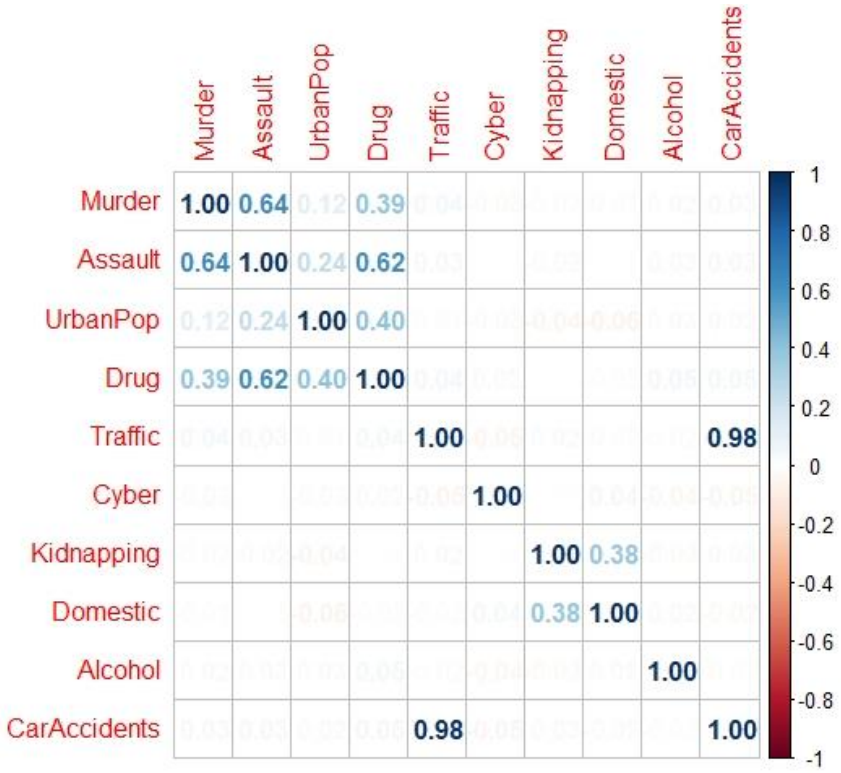

Figure 6 Correlation Matrix and Correlation plot

1.3 Elimination of explanatory variables

From the figure 6 we can see that there is a very strong positive correlation between the explanatory variable Car Accidents and Traffic. The correlation value is higher than 0.80. So, one of the two variables should be removed based on the average correlation with other variables. In this case, the average correlation of the Car Accidents was higher than the variable Traffic. Therefore, the explanatory variable was removed from the dataset to get the better fit for the linear model.

In the other hand, we saw some outliers in the dataset when we were analysing the boxplot of the dependent variable i.e., Murder. So, we are going to include the observation with murder arrest less than 23 to create the linear regression model. After this step the total number of observations is 985.

1.4 Implementation of the linear regression model

For the first linear model all the exploratory variables were used. The estimated intercept is 2.941e+00 and the p value is 0.0301. The p value is less than 0.05 which shows the significance of intercept at 5% significance level. The coefficient of the variable Assault, Urban Pop, Drug, Traffic, Cyber, Kidnapping, Domestic, Alcohol is 4.233e-02, -1.771e-02, 9.689e-03, 2.778e-05, -5.930e-02, 2.115e-03, -4.000e-03 and 7.265e-03 respectively. The variable which has p value less than 0.05 is the variable Assault with p value <2e-16. This shows the significance of Assault at 5% significance level. The R squared and Adjusted R squared are 0.3907 and 0.3857 respectively. This model is not optimal model, and this model can be improved.

1.5 Steps to obtain optimal linear regression model

For improvement of the linear model, I removed the exploratory variables based on highest p value one at a time and kept a close eye on the significance of the remaining exploratory variables along with R-squared and adjusted R squared. I was looking for increase or minimum decline in R-squared and increase in adjusted R squared along with increase in the number of significant exploratory variables.

Firstly, I created the linear model with all the variables. The R squared and Adjusted R squared are 0.3907 and 0.3857 respectively for the model. I looked at the significance of variables i.e., p value. The p value of the variable Kidnapping was the highest among all the variable with the value 0.8247. This shows the insignificance of the variable in the model.

Secondly, I created another linear model removing the explanatory variable Kidnapping. The R squared and Adjusted R squared are 0.3907 and 0.3863 respectively for the model. The R squared value remain the same and Adjusted R squared increased. This shows that this model is better than previous model. I looked at the significance of variables i.e., p value. The p value of the variable Traffic was the highest among all the variables in this model with the value 0.6820 and it is not significant at 5% significance level.

Thirdly, I created another linear model removing the explanatory variable Traffic from the previous model. The R squared and Adjusted R squared are 0.3906 and 0.3869 respectively for the model. The R squared value decrease slightly and Adjusted R squared increased. The

increase in adjusted R squared shows that this model is better than previous model. I looked at the significance of variables i.e., p value. The p value of the variable Domestic was the highest among all the variables in this model with the value 0.64177 and it is not significant at 5% significance level.

After this, I created another linear model removing the explanatory variable Domestic from the previous model. The R squared and Adjusted R squared are 0.3905 and 0.3873 respectively for the model. The R squared value decrease slightly and Adjusted R squared increased. The increase in adjusted R squared shows that this model is better than previous model. I looked at the significance of variables i.e., p value. The p value of the variable Drug was the highest among all the variables in this model with the value 0.63190 and it is not significant at 5% significance level.

After this, I created another linear model removing the explanatory variable Drug from the previous model. The R squared and Adjusted R squared are 0.3903 and 0.3878 respectively for the model. The R squared value decrease slightly and Adjusted R squared increased. The increase in adjusted R squared shows that this model is better than previous model. I looked at the significance of variables i.e., p value. The p value of the variable Alcohol was the highest among all the variables in this model with the value 0.53270 and it is not significant at 5% significance level.

After this, I created another linear model removing the explanatory variable Alcohol from the previous model. The R squared and Adjusted R squared are 0.3901 and 0.3882 respectively for the model. The R squared value decrease slightly and Adjusted R squared increased. The increase in adjusted R squared shows that this model is better than previous model. I looked at the significance of variables i.e., p value. The p value of the variable Cyber was the highest among all the variables in this model with the value 0.3467 and it is not significant at 5% significance level.

After this, I created another linear model removing the explanatory variable Cyber from the previous model. The R squared and Adjusted R squared are 0.3895 and 0.3883 respectively for the model. The R squared value decrease slightly and Adjusted R squared increased. The increase in adjusted R squared shows that this model is better than previous model. I looked at the significance of variables i.e., p value. The p value of intercept and the variable Assault were less than 0.05. It shows that the variables are significant at 5% significance level. The

p value of the variable Urban Pop was the highest among all the variables in this model with the value 0.0943 and it is not significant at 5% significance level. However, it is significant at 10% significance level.

Lastly, I created another linear model removing the explanatory variable Urban Pop from the previous model. The R squared and Adjusted R squared are 0.3878 and 0.3872 respectively for the model. The R squared value and Adjusted R squared both decreased. Therefore, in my opinion the optimal model is the model with the intercept, Assault and Urban Pop.

1.6 Final model for linear regression

The final linear model is the model with the intercept, Assault and Urban Pop. The equation from the final linear model is given below:

Murder = 2.664005 + (0.042982) * Assault + (-0.015780) Urban Pop

The equation shows the relationship between variable Assault, Urban Pop, and Murder. For instance, 1 unit increase in Assault arrest positively impact the murder arrest by 0.0429. Similarly, 1 unit increase in Urban Pop negatively impact the murder arrest by 0.015.

1.7 Properties of linear regression model using OLS

1.7.1 The residuals have zero mean

The mean of the residuals is 1.589108e-16 which is very close to zero. Thus, we can say the errors from the model is zero

1.7.2 The variance of the residuals is constant (Homoskedasticity)

From figure 7 we can see that only 3 points are outside the line i.e., 3*SD and -3*SD out of 985 points. Therefore, we can assume that the variance of the residuals is constant.

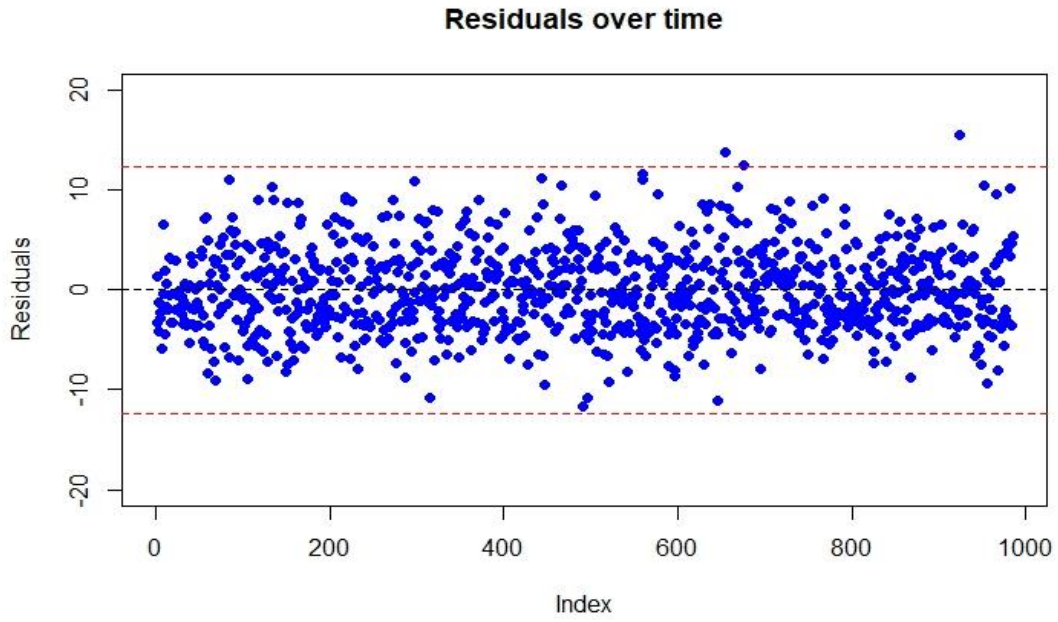

Figure 7 Residual Analysis for Homoskedasticity

1.7.3 The residuals are linearly independent of one another

From figure 7, we can see that residuals are completely random. There is no pattern in the residuals. Thus, the residuals of the model are linearly independent of one another.

1.7.4 There is no relationship between the residuals and each of the variables

The correlation of Assault and Urban Pop with the residual is 7.464943e-18 and 7.836895e-17. The correlation between the variables and the residual is very weak. Thus, there is no relationship between the residuals and the explanatory variable

1.7.5 The residuals are normally distributed

The X squared and the p value from the Jarque Bera Test is 17.252 and 0.0001794. The p value of the test is lower than 0.05. We can reject the null hypothesis that the data is from normal distribution. Thus, we can say the data is not from the normal distribution and the model violets the assumption of normal distribution.

1.8 Findings and results

The aim of the regression analysis was to find the model that could predict the murder arrest based on the variables. The final linear model that consists of assault and urban population as the explanatory variable of the model. The model suggests that 1 unit increase in Assault arrest positively impact the murder arrest by 0.0429. Similarly, 1 unit increase in Urban Pop negatively impact the murder arrest by 0.015. However, this model violets on of the 5 assumptions of linear regression i.e., test of normality. This suggest that the model that is created might be unreliable. This model is not able to generalize our findings from the sample data to the overall population.

Figure 6 shows that the correlation between the traffic arrests and car accidents are highly correlated with each other. This suggest that traffic violations arrests that the main cause of car accidents. Moreover, the correlation between the drug and the assault is moderately correlated. This shows the increase in the drug use can be the cause of the assaults. Furthermore, assault is moderately correlated with murder arrests. Therefore, we can also argue that the increase in drug use can increase the number of assaults and then in turn increase the number of murders in the city.

In conclusion, to decrease the murder arrests in the city, the police must focus their resources on decreasing the drug use in the European cities. Moreover, educating citizens about the traffic violation can decrease the car accidents.

2 Clustering

This section includes the clustering of the “wholesale” dataset. Firstly, this section gives the information about the dataset. This section also depicts exploratory data analysis for variables. Furthermore, the dataset is pre-processed and scaled before creating the clusters from the dataset. The objective of clustering is to group of observations which have similar characteristics to each other with in the group and different from the observations which belong to another group. To ensure the optimal number of clusters different type of methods is used before using the KMEAN clustering algorithm. Lastly, this section is concluded with findings and results that will help the wholesale company in the future.

2.1 Structure and characteristics of dataset

The dataset i.e., ‘wholesale’ contains 440 observation and 8 variables. All the variables belong to integer class. The variables used in the dataset are Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_paper and Delicassen. There are no missing values in the dataset. Moreover, Channel and Region are the categorical variable with distinct values of 1 and 2 and 1, 2 and 3 respectively.

2.2 Explanatory data analysis

The variables i.e., Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_paper and Delicassen are used for explanatory data analysis to get better understanding of the variables in the dataset. Both the numerical and graphical approach is considered for the analysis.

2.2.1 Univariate data analysis

In univariate data analysis, each variable will be examined separately. The measure of central tendency i.e., mean, and median and measure of spread i.e., standard deviation, range, skewness, and interquartile range will be use in the univariate data analysis. Table 2 shows the measure of central tendency and measure of spread for all the selected variables of the dataset.

Table 2 Measure of central tendency and spread

	Fresh	Milk	Grocery	Frozen	Detergent_ Paper	Delicassen
Mean	12000	5796	7951	3071.9	2881.5	1524.9
Median	8504	3627	4756	1526.0	816.5	965.5
Min	3	55	3	25	3	3
Max	112151	73498	92780	60869.0	40827.0	47943.0
SD	12647	7380	9503	4855	4768	2820
Variance	159954927	54469967	90310104	23567853	22732436	7952997
IQR	13806	5657	8503	2812	3665	1412

As there are few distinct values in variable channel and region. We can use a bar graph to explore the data and get useful insight. From figure 8 we can see the channel, region, and their frequency. The figure depicts that majority of the goods were sold via channel 1 and majority of the sales happened in region 3.

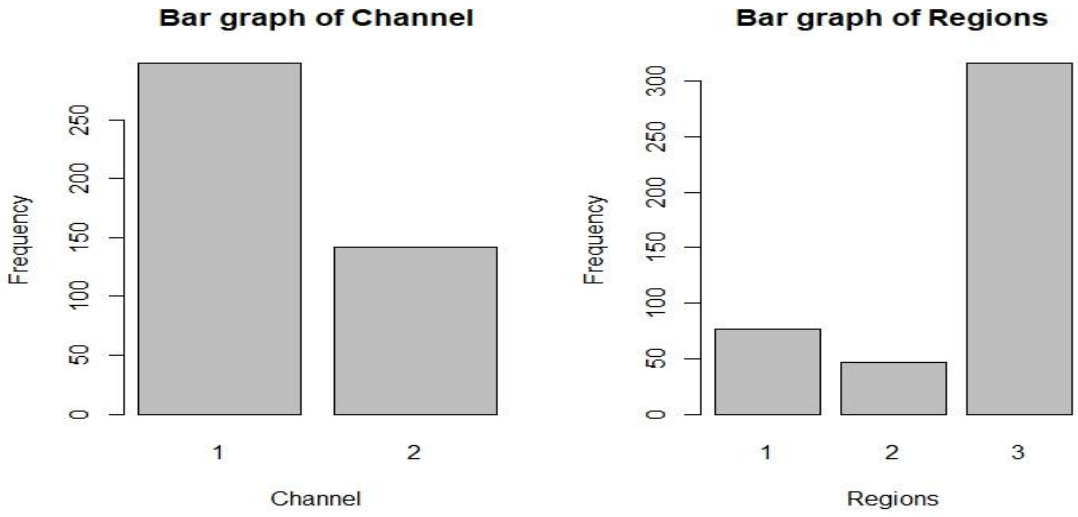

Figure 8 Bar graph of Channel and Region

Table 2 shows that the average annual sale of Fresh products is 12000. The minimum annual sale was 3 and the maximum annual sale was 112151. Similarly, the annual sale that separate the high half and low half of the annual sale is 8504. Moreover, the standard deviation of 12647 suggests that on average the annual sale deviate by 12647 from the mean.

Furthermore, the difference between the highest annual sale and lowest annual sale in the middle half of the distribution is 13806.

Figure 9 shows the box plot and histogram of variable Fresh. We can see that the data is not normally distributed. It is positively skewed. It shows that most of the annual sales of fresh products were below 20000.

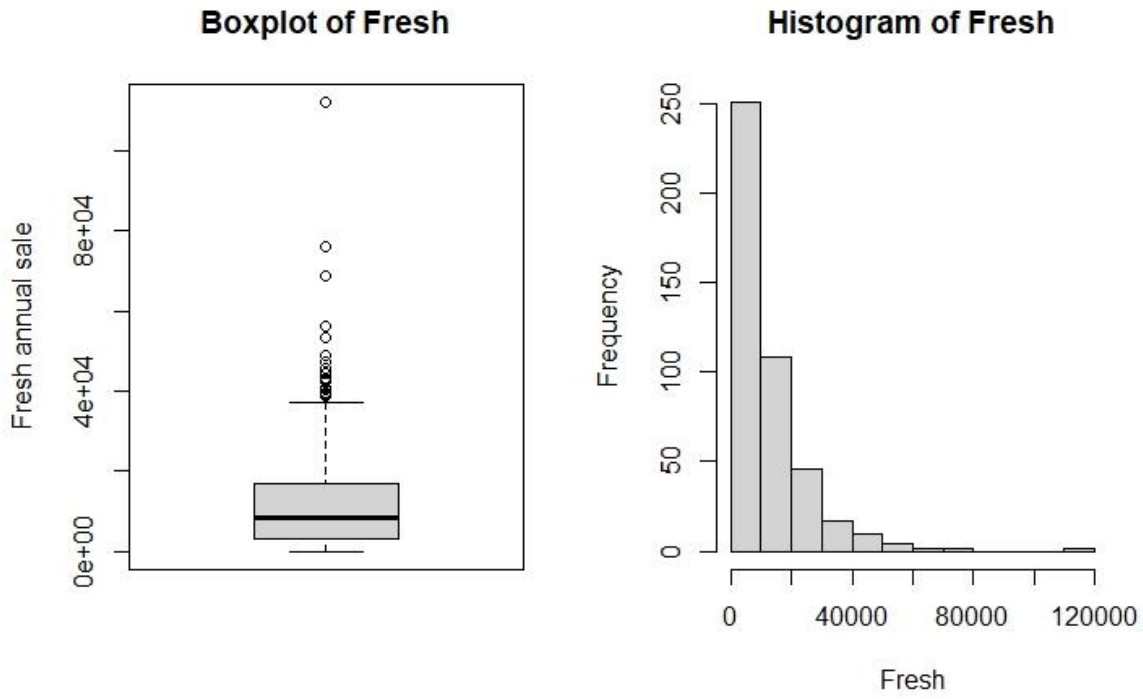

Figure 9 Box plot and histogram of Fresh product

Table 2 depicts that the average annual sale of Milk products is 5796. The minimum annual sale of Milk products was 55 and the maximum annual sale of Milk products was 73498. Similarly, the annual sale of Milk products that separate the high half and low half of the annual sale is 3627. Moreover, the standard deviation of 7380 suggests that on average the annual sale deviate by 7380 from the mean of Milk products. Furthermore, the difference between the highest annual sale and lowest annual sale of Milk products in the middle half of the distribution is 5657.

Figure 10 shows the box plot and histogram of variable Milk. We can see that the data is not normally distributed. It is positively skewed. It shows that most of the annual sales of milk products were below 10000.

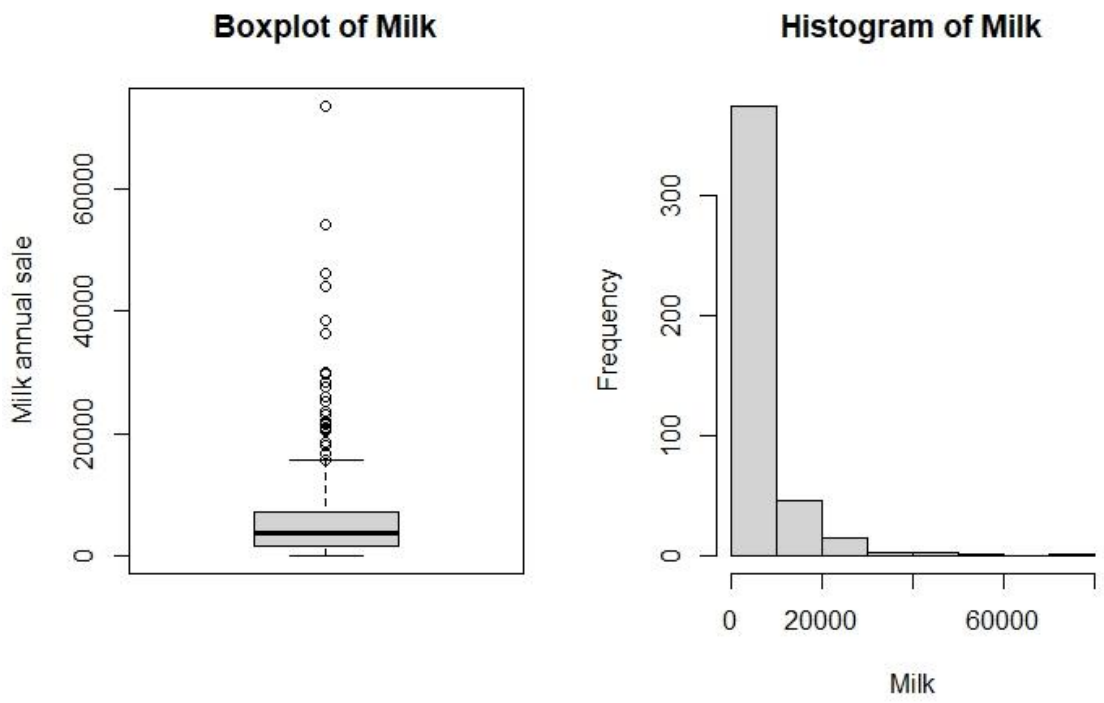

Figure 10 Box plot and histogram of Milk

Table 2 shows that the average annual sale of Grocery products is 7951. The minimum annual sale was 3 and the maximum annual sale was 92780. Similarly, the annual sale that separate the high half and low half of the annual sale is 4756. Moreover, the standard deviation of 9503 suggests that on average the annual sale deviate by 9503 from the mean. Furthermore, the difference between the highest annual sale and lowest annual sale in the middle half of the distribution is 8503.

Figure 11 shows the box plot and histogram of variable Grocery. We can see that the data is not normally distributed. It is positively skewed. It shows that most of the annual sales of grocery products were below 10000.

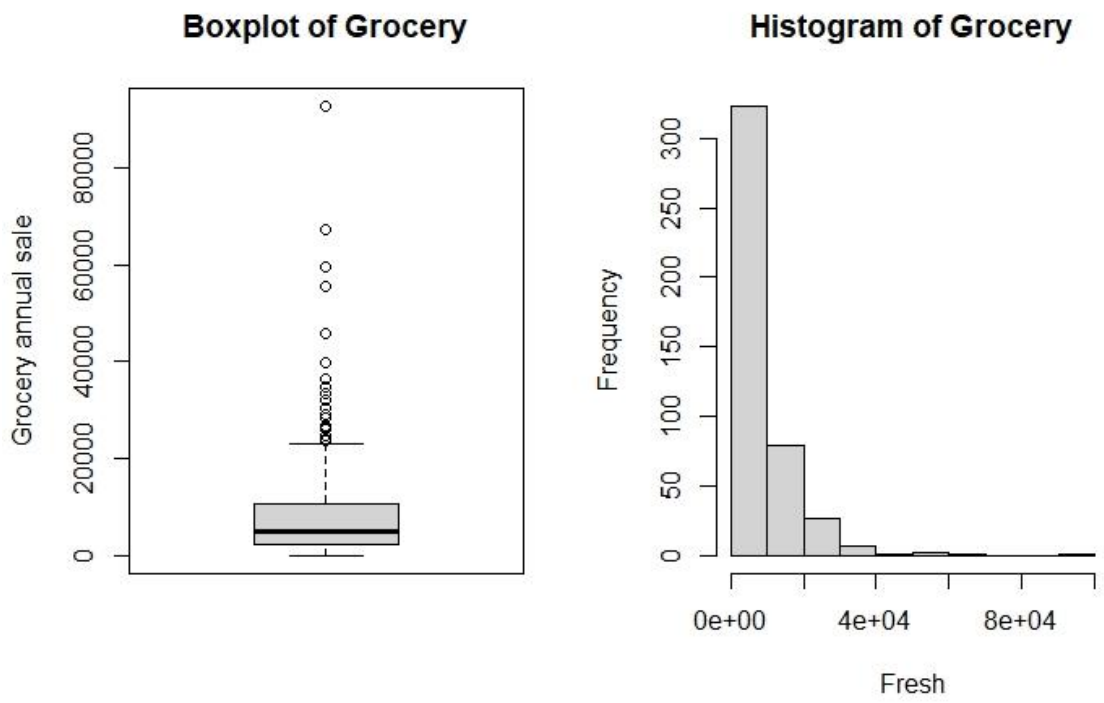

Figure 11 Box plot and histogram of Grocery

Table 2 shows that the average annual sale of Frozen products is 3071.9. The minimum annual sale was 25 and the maximum annual sale was 60869. Similarly, the annual sale that separate the high half and low half of the annual sale is 1526. Moreover, the standard deviation of 4855 suggests that on average the annual sale deviate by 4855 from the mean. Furthermore, the difference between the highest annual sale and lowest annual sale in the middle half of the distribution is 2812.

Figure 12 shows the box plot and histogram of variable Frozen. We can see that the data is not normally distributed. It is positively skewed. It shows that most of the annual sales of frozen products were below 5000.

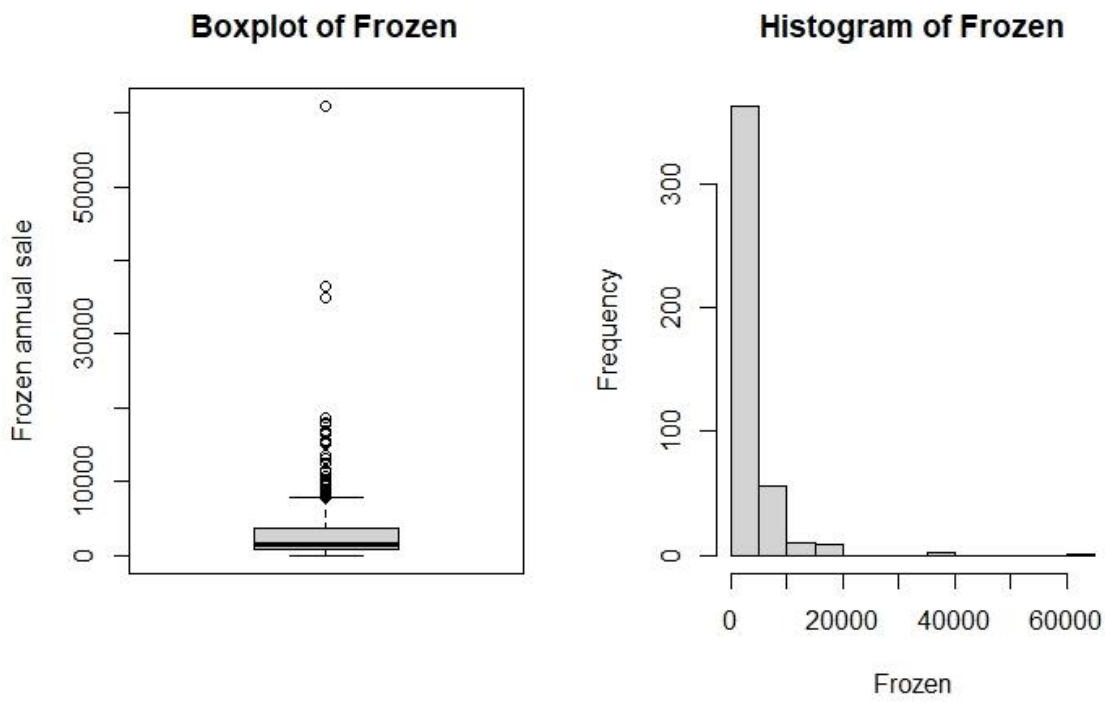

Figure 12 Boxplot and histogram of Frozen

Table 2 shows that the average annual sale of Detergents products is 2881.5. The minimum annual sale was 3 and the maximum annual sale was 40827. Similarly, the annual sale that separate the high half and low half of the annual sale is 816.5. Moreover, the standard deviation of 4768 suggests that on average the annual sale deviate by 4768 from the mean. Furthermore, the difference between the highest annual sale and lowest annual sale in the middle half of the distribution is 3665.

Figure 13 shows the box plot and histogram of variable Detergents Paper. We can see that the data is not normally distributed. It is positively skewed. It shows that most of the annual sales of detergents products were below 5000.

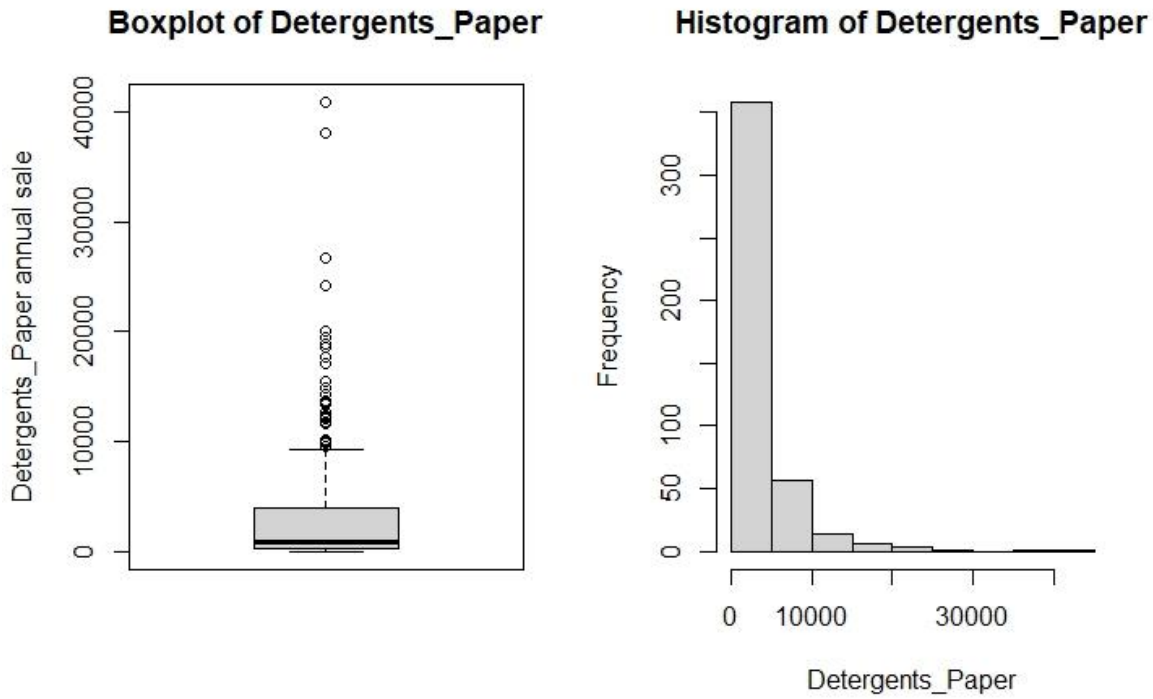

Figure 13 Boxplot and histogram of Detergent Paper

Table 2 shows that the average annual sale of Delicassen products is 1524.9. The minimum annual sale was 3 and the maximum annual sale was 47943. Similarly, the annual sale that separate the high half and low half of the annual sale is 8504. Moreover, the standard deviation of 2820 suggests that on average the annual sale deviate by 2820 from the mean. Furthermore, the difference between the highest annual sale and lowest annual sale in the middle half of the distribution is 1412.

Figure 14 shows the box plot and histogram of variable Delicassen. We can see that the data is not normally distributed. It is positively skewed. It shows that most of the annual sales of delicassen products below 5000.

Figure 14 Boxplot and histogram of Delicassen

2.2.2 Multivariate analysis

In multivariate data analysis, each variable will be analysed with the relation to other variable. The correlation matrix and correlation plot are used to show the relationship between the variables.

Figure 15 shows the correlation of each variable with respect to the other variables in the dataset. The correlation of channel with grocery and detergent paper is 0.61 and 0.64. The value indicates moderate correlation. This suggests that higher sales are done in higher channel i.e., 2. The variable grocery and the detergents paper have a high positive correlation with the correlation value of 0.92. This suggest that the increase in the grocery sale is related with the increase in detergents papers. Moreover, the correlation between the region and other variables is very weak. This suggests that the sales do not depend on the region. The variable milk is moderately positively correlated with the variable detergents paper and grocery with the value of 0.66 and 0.73. This might suggest that the increase in the sale of milk is related with the increase in the sale of grocery and increase in detergents paper.

Furthermore, channel is negatively correlated with the variable fresh and frozen with the weak correlation of -17 and -0.20. It suggests that there is a very weak possibility that the fresh and frozen are sold via lower channel. Similarly, there detergent paper is negatively correlated with the fresh and frozen products with a weak correlation of -0.10 and -0.13 respectively.

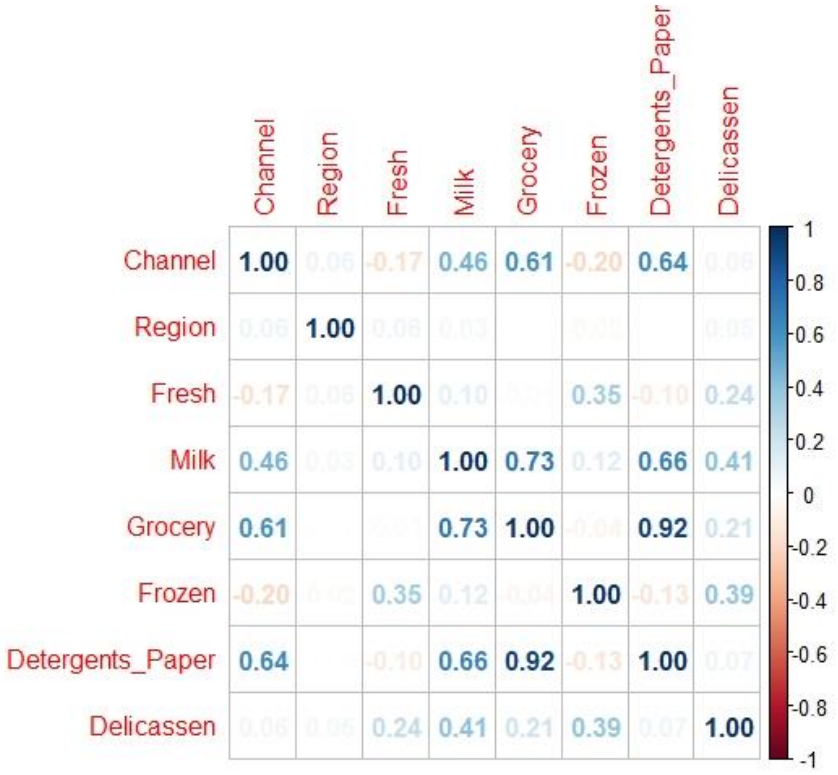

Figure 15 Correlation matrix

2.3 Normalization of the variables

Min-max normalization method is used to scale the data between 0 and 1. Generally, normalization is needed to improve the performance of the result. The k means algorithm uses distance like Euclidian distance to determine the clusters in the datasets. It is helpful in clustering when the variables in the datasets have different scale of measurement. It is useful because it helps to equalize the size and the magnitude as well as the variability of the variables. Therefore, it is very important to normalize the dataset for clustering.

2.4 Determination of optimal number of the clusters

It is very important to determine the optimal number of clusters while clustering the dataset. There are many ways to find the optimal number of clusters. However, 4 methods will be considered in this report i.e., Elbow method, Calinski Harabasz Index, silhouette method and Gap method. The maximum number of clusters that will be used in these methods is 8.

Elbow method is the graphical way of determining the clusters. In the method Total Within-cluster Sum of Squared i.e., TWSS is calculated. The WSS is plotted over the number of clusters. The number of clusters is chosen from the point when the improvement in WSS is decreasing. Figure 16 shows that after cluster 3 the imporvement in TWSS is decreasing significantly so the appropriate number of clusters according to elbow method is 3.

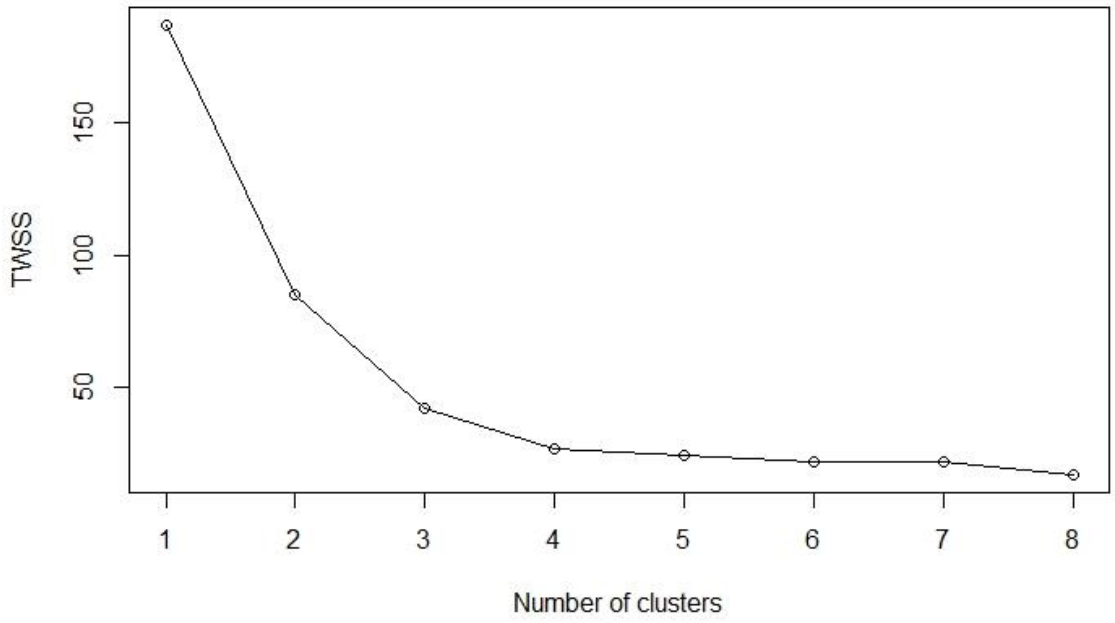

Figure 16 Elbow method

Calinski Harabasz Index is based on the variance ration criterion. It measures the ratio between the BGSS i.e., Between group sum of squares and WGSS i.e., Within group sum of squares. Figure 17 shows the Calinski Harabasz Index is maximum at cluster 3 so the optimal number of clusters according to Calinski Harabasz Index is 3

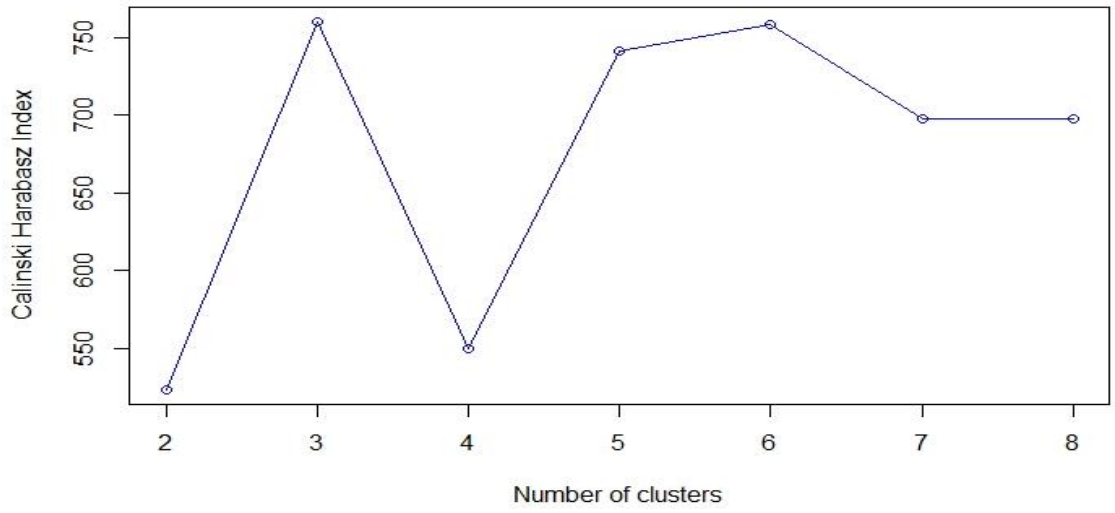

Figure 17 Calinski Harabasz Index

Silhouette method measures the consistency within the clusters. In other words, it measures how much a point is similar to its own cluster compared to other clusters. Figure18 shows that the maximum silhouette value is at cluster 3. Thus, the optimal number of clusters according to silhouette method is 3

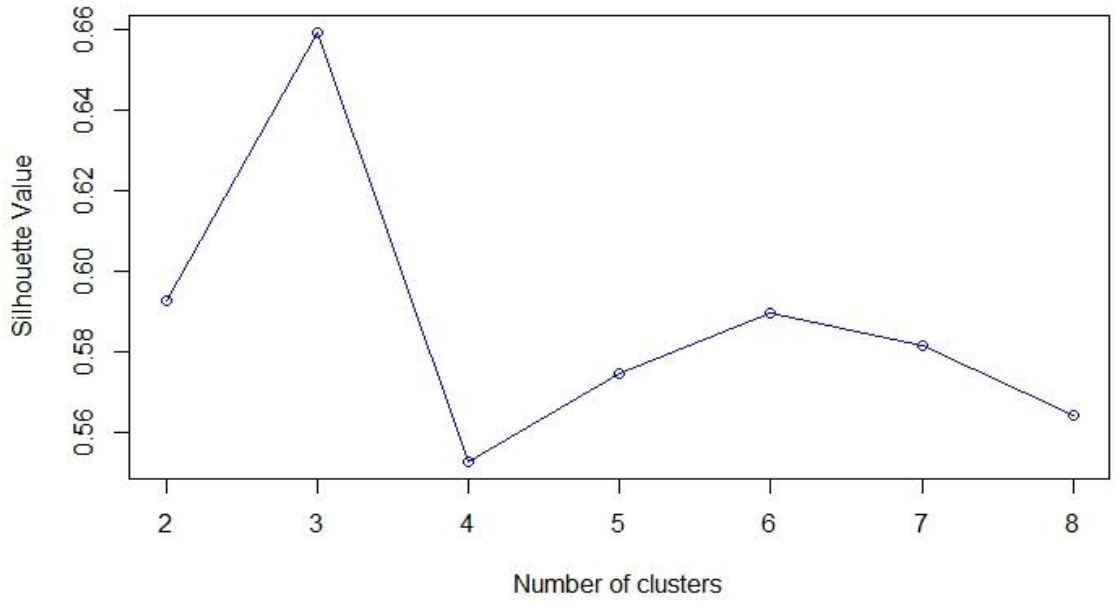

Figure 18 Silhouette method

Gap statistics measures the difference between the reference data and the actual data. The optimal number of clusters is determined by the largest distance between the reference data and actual data. Figure 19 suggests that the largest difference is at cluster 2. So the optimal cluster number according to gap statistic is 2

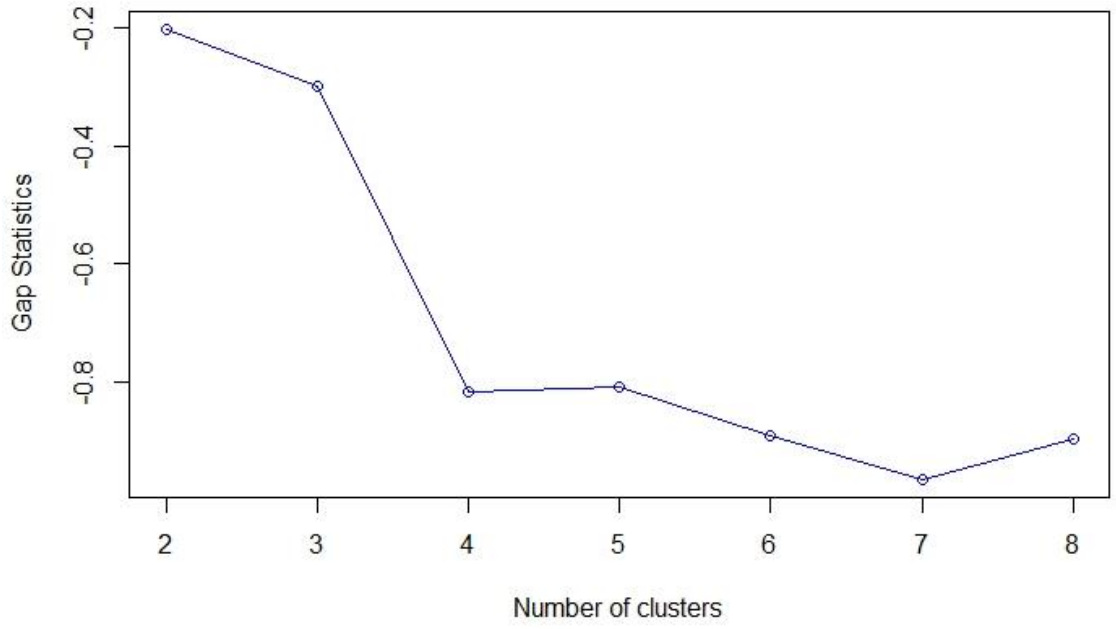

Figure 19 Gap Statistics

To conclude, having analysed all the methods the all the method suggests the cluster 3 except gap statistics. So, the optimal number of clusters for the model is 3.

2.5 Implementation of K-means algorithms

The k means algorithm is used to create the clusters for the scaled data. The number of clusters used in the algorithm is 3 and the “nstart” value is 25. Here the “nstart” parameter generates the 25 initial random centroids and choose the best one for the algorithm. If we do not use the parameters, then the clusters created will not be stable and changes the clusters every time the k means algorithm is run. In other words, if we do not set the “nstart” value there is higher possibility that the cluster will change every time the model is created. Figure 20 depicts 3 clusters in the data. The cluster does not overlap each other. This suggests that the clusters have distinct characteristics with each other.

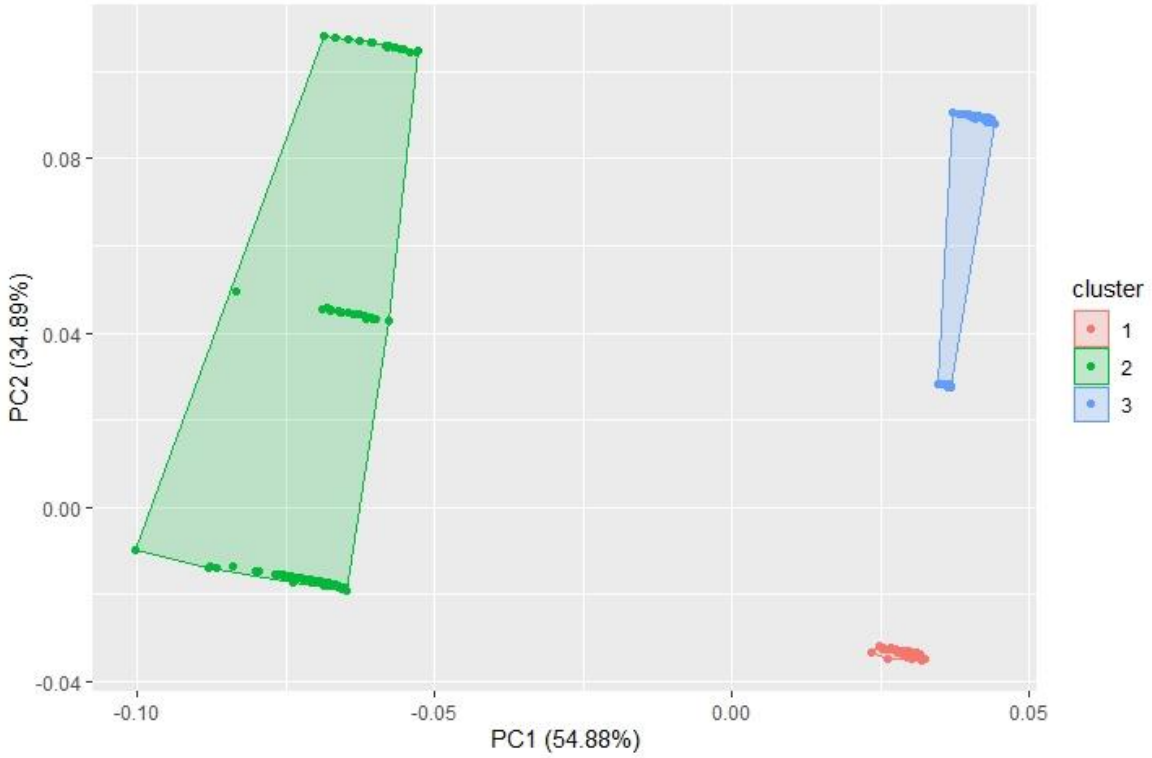

Figure 20 Clustering in scaled data

Table 3 Mean of the clusters

Cluster	Region	Channel	Fresh	Milk	Grocery	Frozen	Detergent	Delicassen
1	2.61	2	8904	10716	16323	1653	7270	1753
2	3	1	13878	3487	3887	3657	787	1518
3	1.32	1	12499	3366	4145	3970	800	1168

Table 4 SD of Channel and Region in clusters

Clusters	SD Channel	SD Region
1	0	0.703
2	0	0
3	0	0.470

Table 3 shows the mean of the clusters. Based on the mean we can derive some characteristics of the variables in the clusters. The table depicts that the cluster 1 the annual sales are done mostly by channel 2 and in region 2. Moreover, the sales in cluster 2 and 3 are done via channel 1. However, most of the sales for cluster 1 are one in the region 2 and 3 since its mean is 2.61 and the standard deviation is 0.703. Similarly, cluster 3 has the region mean of 3 and the standard deviation is 0 which suggests that the sale is done in only region

3. Furthermore, the mean and standard deviation for region in cluster 3 is 1.32 and 0.470 which suggests that the most of sale for cluster 3 are in region 1 and region 2.

Figure 21 and Table 2 depicts that cluster 1 has lower Fresh and Frozen average annual sale and higher average annual sale for the variable Milk, Grocery, Detergent and Delicassen. Furthermore, cluster 2 and cluster 3 have higher average annual sale of Fresh and Frozen and lower annual sales for the variable Milk, Grocery, Detergent and Delicassen.

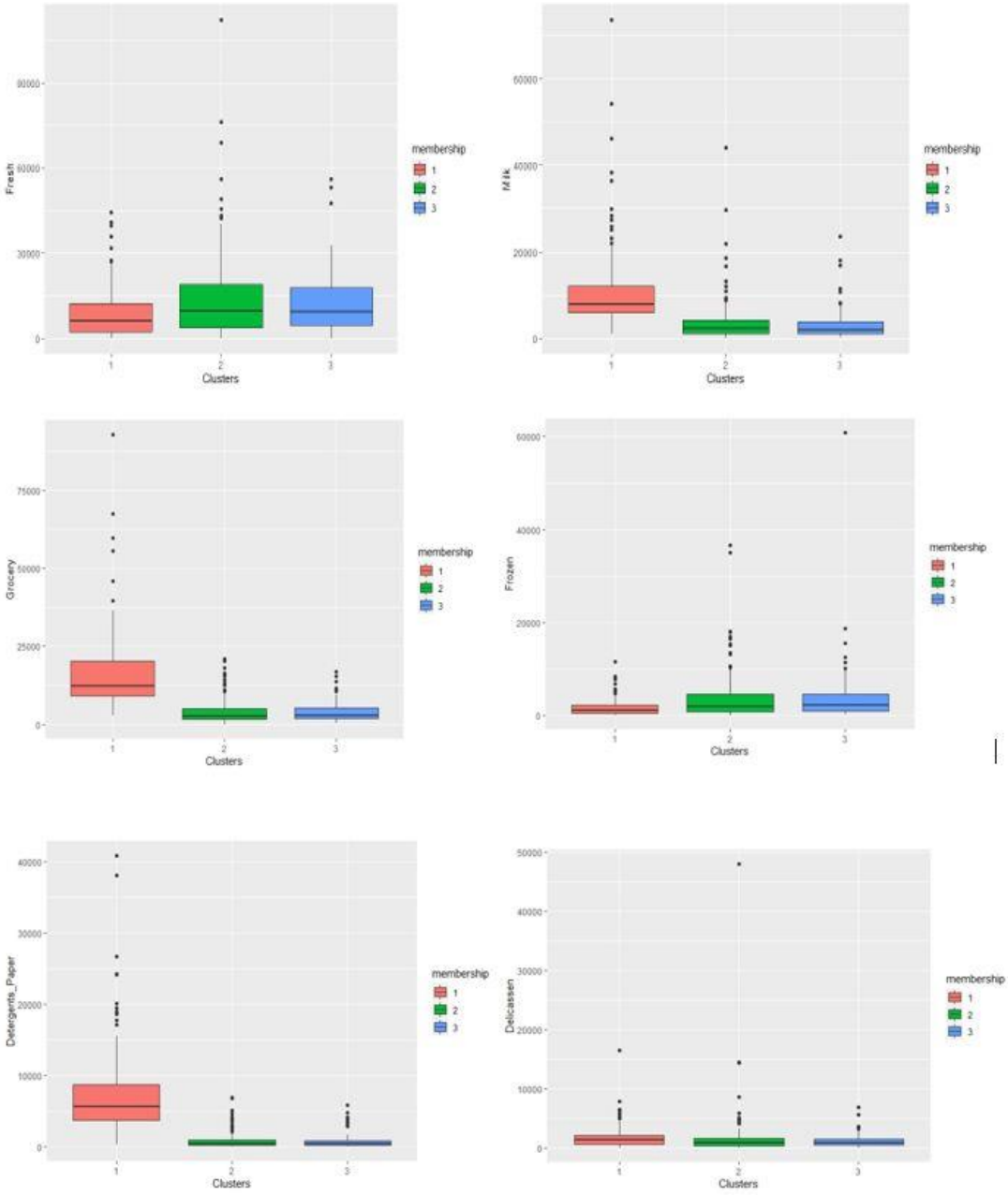

Figure 21 Box plot showing the distribution in each cluster

2.6 Findings and results

The KMEAN algorithm suggests 3 clusters with distinctive features. In my opinion, the wholesaler must consider these features when making the marking marketing decisions. In cluster 1, there is high annual sale of Milk, Grocery, detergents paper, and delicassen products in region 2 and 3. The cluster also suggests that the sales were made via channel 2. Moreover, in cluster 2 and cluster 3 there is high annual sale of fresh and frozen products via channel 1. However, the sales in cluster 2 was made in the region 2 and the sales in cluster 3 was made in region 1 and 2.

To recapitulate, having analysed all the clusters we can make specific plans to increase the sale of products in for each cluster. For instance, cluster 1 can be targeted with marketing of products like frozen and fresh products in region 2 and 3 via channel 2 to further increase the annual sale of the company. Moreover, cluster 2 can be targeted with the marketing of milk, Grocery, detergent and delicassen products in region 3 via channel 1. Furthermore, cluster 3 can be targeted with the marketing of milk, Grocery, detergent and delicassen products in region 1 and 2 via channel 1. Similarly, most of the sales were made in region 3 using channel 1 so region 3 and channel 1 must be the primary focus for the marketing strategies.