ASSIGNMENT_2_000299549

1 Front Page

A220A0010 Free Analytics Environment R

Second Assignment – Autumn 2022

A VIVID REPORT ON A CRIME RESEARCH GROUP AND A WHOLESALE DISTRIBUTION

NOVEMBER 3, 2022

STUDENT NUMBER: 000299549

Contents

List of Figures

ASSIGNMENT 2

2 Assignment 2: PART 1

2.1 Answer for Question 1

On a first glance, the data has 1000 observations and has 10 variables.

Thus, it has 1000 rows and 10 columns. From the 10 variables, Murder, Drug, Cyber and Alcohol are numeric values, meaning they contain values that are either whole numbers or decimals in nature whiles Assault, UrbanPop, Traffic, Kidnapping, Domestic and Car Accidents are integer values implying that they are whole numbers in nature.

An important aspect of data observation is to check for any missing value and hence remove it accordingly with the aim of making the data viable for forecasting purposes. A numerous reason can result for that, these can be due to misunderstanding of data entry and equipment malfunction. Therefore, I proceed to check for missing values in the data.

After the removal of such NA's values, our data observation is now 995 and our variables remains 10, which is not far from our initial data set hence a good data to be analyzed because it contains complete observations.

2.2 Answer for Question 2

Now exploring four variables and the dependent variable with the histogram charts to give a data analysis on them. We start with Assault,

2.2.1 Assault

i. Assault's Visual's

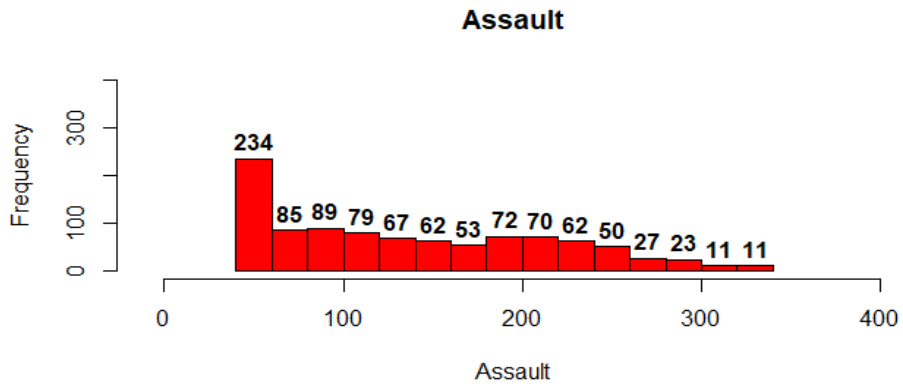


Figure 2.1: A graphical view of the Assault variable values (per 100'000)

Figure 2.1 depicts the histogram of Assault, from the histogram, Assaults values ranging between 45 and 100 per (100'000) had the highest count of over 400. Thus, the number of people recorded with these values contributed to over 40% from our observations in which, 234 of these counts was from the values between 45 and 60 (per 100'000).

Furthermore, the rest of the assault values (per 100'000), was relatively volatile with counts of 79 in the 101 and above (per 100'00) then decreasing to a count of 53. There was a

raise to 72 then falling to a count of 23 within the values of Assault 101 and 300 (per 100'000). The decline took a nose dive from the 300 value onwards (per 100'000) with a count of over 10.

The chart showed a right/positive skewness thus most of the distribution was to the right.

ii. Assaults Numeric

The Assault data recorded its lowest value (per 100'000) as 45.0 and its maximum as 337.0. The average or the mean of the Assault value is 137.9(per 100'000) and its median value is 124.0(per 100'000).

The first quartile implying that when arranged from an increasing order, 25% of our values (per 100'000) are less than 64(per 100'000) and 75% of per data for Assault are less than 202.0(per 100'000).

2.2.2 UrbanPop

i. UrbanPop Visual

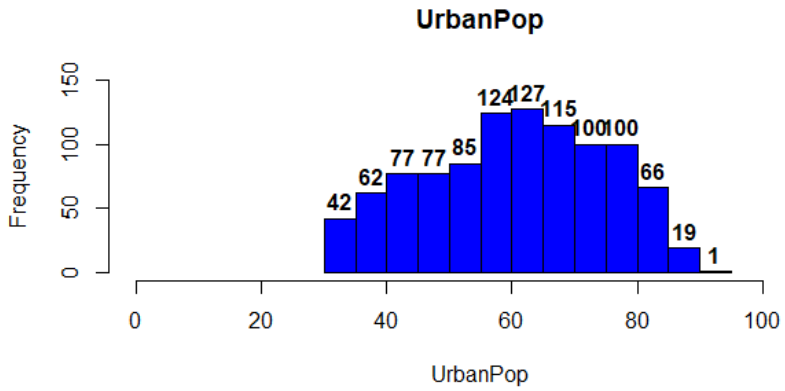


Figure 2.2: A graphical view of the UrbanPop variable values (per 100'000

Figure 2.2 shows a presentation of UrbanPop, it counts was increasing at a constant rate from the values ranging from 32 to 60 where it had a count of 124. The count started to diminish at a constant rate, then at an increasing rate till the count dropped as low as 1 count on the maximum value.

The histogram can be said to be symmetric as the distribution was approximately even on both the right and left sides.

ii. UrbanPop Numeral

The UrbanPop recorded its lowest value as 32.0 and its highest value as91.0 Its Mean value is 60.88 and Median is 62.0. The 1st Quartile is 50.00 and 3rd Quartile is 72.00

2.2.3 Traffic

i. Traffic Visual

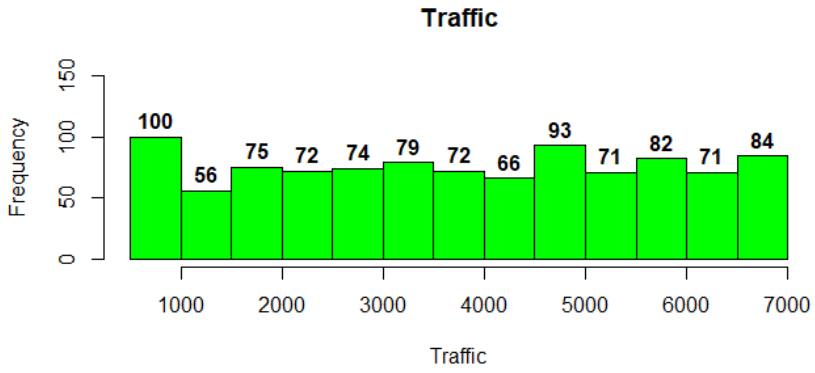


Figure 2.3: A graphical view of the Traffic variable values (per 100'000)

The values from traffic (per 100'000) were relatively the same counts throughout the values attained especially values from the ranges of 2000 and 7000 (per 100'000). The most counts were values ranges from approximately 500 to 999(per 100'000) which recorded a count of 100. The rest of the values had a count of 75,72,79,6 and 56 recording the lose count for values within the 1000 and 1500(per 100'000).

ii. Traffic Numerals

The minimum value for Traffic was 503(per 100'000) and its maximum is 6991(per 100'000). The mean and median for the traffic was 3767 and 3781(per 100'000). Its 1st Quartile is 2137 and 3rd Quartile is 5390

2.2.4 CarAccidents

i. CarAccidents Visuals

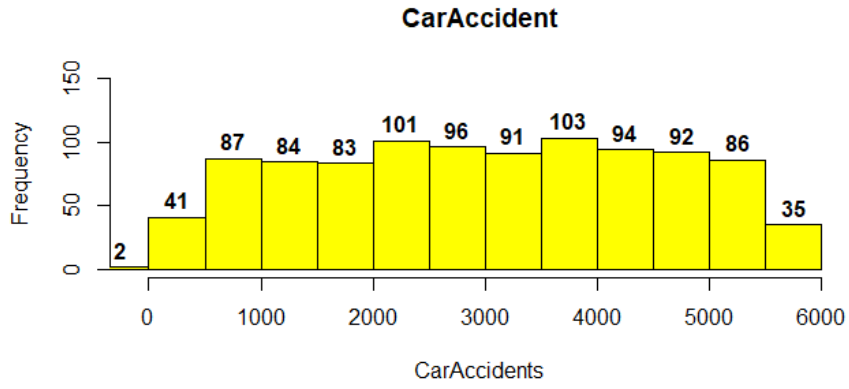


Figure 2.4: A graphical view of the CarAccidents's variable values (per 100'000)

CarAccidents negative values had a count of 2 in total. From 0 to 500 values (per 100'000) had a count of 41 which then increased relatively along the line. Values of 3500 to

4000(per 100’000) had the majority count with a score of 103. This count decreased till it reached 35. This corresponded to values in the ranges of 5500 to 6000(per 100’000).

ii. CarAccidents Numerals

The minimum and maximum value (per 100’000) was negative 66 and 5991 respectively. The mean and median value for CarAccidents (per 100’000) was 3004 and 3025 respectively. Its 1st Quartile was 1731(per 100’000) and 3rd Quartile was 4364(per 100’000).

2.2.5 Murder

i. Murder Visuals

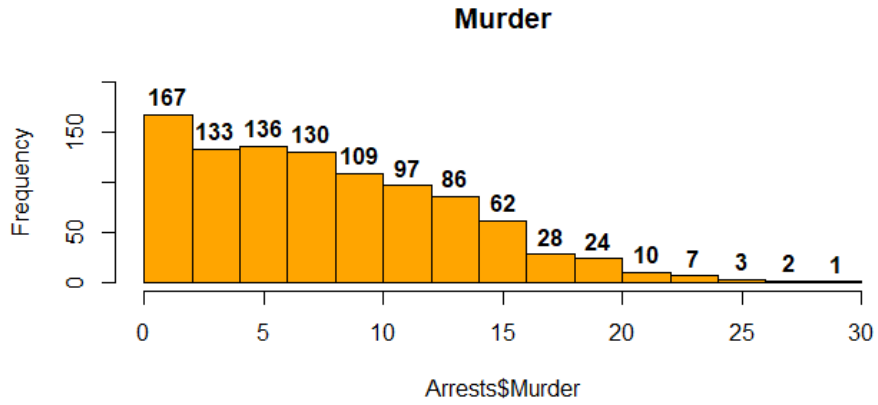


Figure 2.5: A graphical view of Murder variable values (per 100’000)

Murder data clearly showed a right skewedness; thus, its value was distributed to the right-hand side of the tail. The highest count of Murder was in the range of 0.500 to 2. The count for the values decreased at a decreasing rate for the range of values (per 100’000) 2 to 15. Values of 20 (per 100’000) to 29.5 (per 100’000) recorded a count below 10. Where 29.5 (per 100’000) had only one count.

ii. Murder Numerals

The maximum and minimum value (per 100’000) was 0.500 and 29.500 respectively. The mean and median value (per 100’000) was 6.900 and 7.747 respectively. The 1st Quartile and 3rd Quartile values (per 100’000) was 3.200 and 11.450 respectively.

2.3 Answer for Question 3

2.3.1 Answer for Question 3(a)

The correlation of our data is seen relatively against each other in Figure 1.6. On the first glance a common scene is that all values against itself is equal to 1 hence meaning against itself it is strongly correlated.

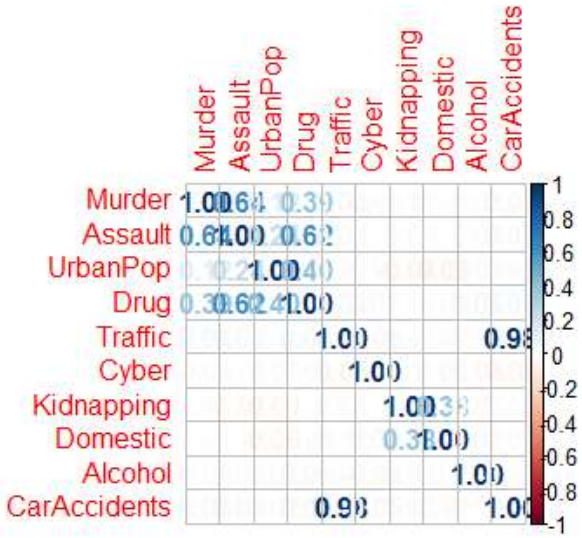


Figure 2.6: A graphical view of the correlation between variables.

2.3.2 Answer for question 3(b)

With absolute values taken into consideration, we look at those with the highest linear association with our independent variable in question (Murder arrest).

Assault and Drugs had the highest linear association with Murder arrest. They had a correlation of about 0.637 and 0.392 respectively. This implies that Assault and Drugs has a moderate correlation with Murder meaning both variables change in the same direction moderately.

A 1% change in Assault and Drug means Murder Arrest also move in the same change of direction of 0.637 and 0.392 respectively.

2.4 Answer for Question 4

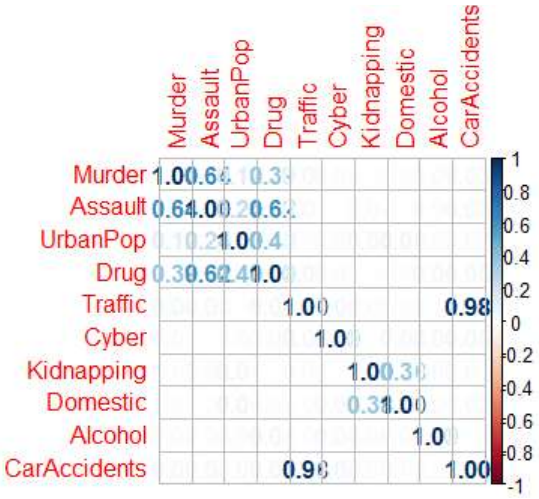


Figure 2.7: A display of the Correlation of our data in absolute values

Absolute values of our correlation imply ignoring the mathematical signs (+ and -) attached to coefficient values and describing it wholly hence looking or concentrating on the magnitude of the correlation.

But remember every variable correlates strongly to itself. We take the variables one after the other.

2.4.1 Murder

As Assault, Drug and UrbanPop variable changes with coefficients of 0.637,0.118 and 0.392, Murder also changes relatively stronger in that order. For the other variables due to its low absolute values, their correlation with Murder is very low thus 0.041, 0.029, 0.167,0.013,0.024 and 0.033 for Traffic, Cyber, Kidnapping, Domestic, Alcohol and CarAccidents respectively.

2.4.2 Assault

We can see here that Murder, UrbanPop and Drug have a relative strong correlation with Assault with coefficient values of 0.67,0.243 and 0.617 respectively. Thus, Assault has a moderately strong correlation with Murder. Traffic, Cyber, Kidnapping, Domestic and CarAccidents has a weak correlation of 0.033,0.009,0.022,0.004 and 0.034 with Assault, their correlation with assault is very low.

2.4.3 UrbanPop

It has a strong correlation with Murder, Assault and Drug. This implies that as these three changes in a certain direction, UrbanPop also follows suit, With the other variables in our data, UrbanPop has a low correlation with them.

2.4.4 Drug

Form our correlation display, Drug has a strong correlation with Murder, Assault and UrbanPop, with a correlation coefficient of 0.39,0.61and 0.40 implying that these have a strong magnitude compared to the other variables with a coefficient of 0.04 for Traffic, 0.02 for Cyber, 0.004 for Kidnapping,0.01 for Domestic, 0.047 for Alcohol and CarAccidents respectively.

2.4.5 Traffic

The only variable that highly strongly correlates with Traffic is CarAccidents. Thus, both variables always go in the same direction with a correlation of almost 1. But with the rest of the variables, it had a correlation coefficient of less than 0.1.

2.4.6 Cyber

From our graphical correlation, Cyber has no strong correlation with any of the variables but it has a relatively weak correlation with Traffic and CarAccidents with coefficient values of 0.048 and 0.049 respectively.

2.4.7 Kidnapping

Kidnapping has a strong correlation with domestic meaning both move in the same direction with a score of 0.38. But with the other variables in question, it has a relatively weak correlation existing between them.

2.4.8 Alcohol

Alcohol doesn't correlate strongly with any of the variables. It has a relatively low correlation with all of the variables in our data.

2.4.9 CarAccidents

Drug has a correlation of 1 with CarAccidents. Implying a very strong correlation with Drug. But with the other variables CarAccidents has a low correlation in the data.

2.5 Answer to Question 5

2.5.1 Answer to Question 5(a)

The variable "exavar" was created and all the explanatory variables were grouped into it. Then there was a removal of all the explanatory variables by making them first absolute and also taking all the diagonals which is the variables correlation values gained against itself. A loop was made to remove all the absolute values higher than 0.8.

From our exavar groupings, the CarAccidents variable have been removed.

2.5.2 Answer to Question 5(b)

The reason for the removal is to meet the first property of the OLS which states that our explanatory variables should not be linearly dependent.

This is because when they are dependent it indicates that changes in one will also associate in the shift of the other hence our model becomes difficult to examine the relationship between each explanatory variable.

In addition, when they are not removed, it makes it hard to explain ones co-efficient from the model and hence trust the p-values of the model because the coefficient becomes very sensitive to any small change in the model due to the high correlation in our explanatory variables, but upon removal we can understand the role of each explanatory variable in relation to our model.

2.6 Answer to Question 6

After implementing our linear regression, the coefficients section shows that,

The y intercept of our linear regression has an estimated value of 3.009e+00

The estimated effect of Assault on Murder is 4.476e-02 (almost 0) this is positive and means that for every 1% increase in Assault there is a correlated 4.476e-02 increase in Murder.

For UrbanPop, its estimated effect is (-1.725e-02) this implies that for every 1% increase in UrbanPop, there is a correlated 1.725e-02 decrease in Murder

Drug recorded an estimate of 1.709e-02. This signifies that for every 1% increase in drug, Murder increases correlatedly by 1.709e-02.

Traffic's estimate coefficient was 5.413e-05 telling us that for every 1% increase in traffic, Murder increases by 5.413e-05.

Cyber had an estimated coefficient of -6.418e-02 implying that for every 1% increase in Cyber, Murder decreases proportionally by 6.418e-02

Kidnapping had an estimated coefficient of 1.084e-04. The value shows that for every 1% increase in Kidnapping there is a correlation of 1.084e-04 increase in Murder

Domestic had an estimated coefficient of -3.809e-03. This implies that for every 1% increase in Domestic, Murder decreases by 3.809e-03

Alcohol had an estimated coefficient of 2.530e-03. Thus, for every 1% increase in Alcohol, there is a correlated 2.530e-03 increase in Murder.

The p-value column shows that according to the rule of significant, Assault is highly significant (three stars) to our dependent variable because its p-value is less than our usual significant level of 0.05 which implies that the results of Assaults are random from the dependent variables and is statistically significant therefore must be included in our model so as to increase the model precision.

The other explanatory variables all had a p-value higher than the threshold of 0.05 indicating that these variables aren't statistically significant with our dependent variable. Even though UrbanPop has a dot(.) it still doesn't meet the significant threshold of been less than or equal to 0.05)

These results of the p-value can change as we adjust the explanatory variables by removing those that have a higher p-value.

From our data, our R-squared measures how well our data are fitted to our regression line. Thus, how much variation of a dependent variable is explained by the independent variables.

The Adjusted R-square takes into account the observation as our independent variables. It is 40.41%. this value can increase only if the adjustment made to the observations improve our model if not it reduces. Generally, a lower R-square means our model isn't a good fit. it shows how distinctly far they are to our mean.

The R-square for our model is 40.89% this value always decreases or remains the same when variables are taken away from the model.

Our overall F-Statistic (82.25) value is greater and bigger than our overall p-value of 2.2e-16, implying that our variables have a joint effect together and none of them happen by chance.

Thus, it is significantly better than a mean only model.

2.7 Answer to Question 7

One must note that in selecting which regression model is better, there are three significant points to note

1. The Value of the Adjusted R-squared
2. The Value of the Multiple R-squared
3. The nearness of the variable's p-values to the significant levels.

This far with these in mind, from our explanatory discussion from our model in Question (6), we can see that we had an Adjusted R-squared of 0.4041, Multiple R-squared of 0.4089. Six out of nine of our independent variable's p-values are not significant with some having a bigger p-value.

We can't just select this model, but various adjustments must be made to the model and the one with roughly a higher Adjusted R-squared value and quite significant p-values will be selected as our prime model.

The steps involve in the adjustment is by removing independent variables that have too high p-values and hence checking if this doesn't reduce our Adjusted R-squared score.

2.7.1 STEP 1 – Removal of Kidnapping and Effect upon Removal

Mine removal of kidnapping is because it has a p-value of 0.99912 and also a standard error of 9.8 implying that its removal won't have any effect on our dependent variable.

Upon removal, we see an increase in our Adjusted R-squared value which is now 0.407 as against 0.4041. We see other variables p-values adjusting nearer to the significant level threshold. But we still can't accept take sides so we then run another regression model.

2.7.2 STEP 2- Removal of Alcohol and Effect Upon Removal

Alcohol's p-value is the highest with a score of 0.8297, therefore must be removed because it's impact on Murder is insignificant.

Again, we see an increase in the R-squared value implying that the removal of Alcohol is in a right direction. But yet again we must further adjust because we can see high values of p in our model regression.

2.7.3 STEP 3-Removal of Domestic and Effect After Removal

Domestic had a p-value of 0.616 and upon removal, our regression had an Adjusted R-squared value of 0.04057, thus an increase again. And our Multiple R-squared has reach its highest score of 0.487.

But we see a slight change in the p-value of UrbanPop which is relatively increasing and moving away from the significant level.

Which upon further iteration its p-value won't be significant.

2.7.4 Selected Model

After various adjustment and iterations, step 3 will be the final step, and the model will be accepted as our optimal model because upon removal of Kidnapping, Alcohol and Domestic variable, we see our model having an Adjusted R-squared value of 0.457 whiles Multiple R-squared is 0.4087. Because our Adjusted R-squared is not zero it means that our model is not worse than a mean only model.

Again, we see Assault and our intercept been highly significant thus having a significant value above 5%. UrbanPop also is significant with a significant value of 9.7% which makes it have an impact on our dependent variable. Cyber, Traffic and Drug been the other independent variables are not significant but with them having one of the lowest p-values we add them to make our model have 5 observations of independent variables.

Lastly our overall p-value of (2.2e-16) still remained zero as compared to our F-statistic indicating our model with five (5) variables plus the intercept is highly significantly better at modelling the relationship to the response variable than an 'intercept only Model'.

2.8 Answer to Question 8

2.8.1 The Final model for our regression is

Murder = 0.2993 + 0.04474*Assault – 0.01607*UrbanPop + 0.01101*Drug + 0.05434*Traffic – 0.06592*Cyber

This implies that for every 1% change in Assault, Murder increases by 0.04474, and also for every 1% change in UrbanPop, Murder reduces by 0.01607, for every 1% change in Drug, Murder increases by 0.01101, for every 1% change in Traffic, Murder increases by 0.05434 for every 1% change in Cyber, Murder reduces by 0.06592. The intercept of our model is 2.993 which is positive meaning that our model moves in the same direction as our dependent variable Murder.

One must also bear in mind that from the model I am confident that Assault and UrbanPop has an impact whiles Drug, Traffic and Cyber may or may not have an impact on Murder.

2.9 Answer to Question 9

To check the OLS properties, look at the following

2.9.1 Mean of the Residual

The mean of the residual is checked to meet the assumption that the mean of the residuals of our model should be zero.

And from our mean residual calculation of our model, we have 1.461391e-16, a value almost closer to zero. Hence, we have a good fit model

2.9.2 Homoskedasticity

The variance of the errors should be consistent for all observations. In other words, the variance does not change for each observation or for a range of observations. This preferred

condition is known as homoscedasticity (same scatter). This can be done visually from the residuals plotting.

The plot made and the model shows that the ranges of observation in our data do not change over the period thus no clear pattern hence meeting the Homoskedasticity assumption

2.9.3 Residuals are linear

To check this assumption, we plot the residuals of our model giving the colors of our residuals blue with a 'pch' of 16. Our y-axis named 'Residuals' and the title 'Residuals over time'.

Then we add some lines using 'abline' and hence we add red lines indicating a plus and a minus 3 standard deviations to assume that the residuals are normal and hence most of the residual's distribution shouldn't be above or below the threshold. And also, with a slope of zero. Lastly, we have another line within the residuals data to represent the mean. The Lines have a width of 2.

From our visualization even though a few residuals where outside the boundary, our model meets the assumption that the residuals are linearly independent of one another because their movement depicts that they aren't dependent on each other.

2.9.4 Correlation Between Residuals and Independent

We making sure that are residual values aren't correlating with the explanatory variables. Because if they do it implies, we can use the independent variables to predict our error term which therefore violate the notion that the error term represents unpredicted variables.

Therefore, we check for the correlation our explanatory variables

- Assault
  It has a value of 0.017 which implies it isn't correlated to the residuals data.
- UrbanPop
  It has a value of 0.0325 which implies it isn't correlated to the residuals data.
- Drug
  It has a value of 0.0397which implies it isn't correlated to the residuals data.
- Cyber
  It has a value of 0.03970 which implies it isn't correlated to the residuals data.
- Traffic
  It has a value of 0.02124 which implies it isn't correlated to the residuals data.

2.9.5 Check the Normal distribution of residuals

We use the Jarque-Bera test. A normal distribution is defined by its tail and skewedness. and that our Skewness should be zero (0), meaning that our data is symmetric to the mean and kurtosis should be three hence telling us the peak of our distribution tail.

From our Jarque – Bera Test run I had a skewness of 0.4065 which is closer to zero and a Kurtosis of 3.40 which is also closer to 3. Therefore, the model has a normal distribution.

3 Assignment 2: PART 2

3.1 Answer to Question 1

The Dataset of Wholesale contains 440 observations from 8 variables. The eight variables are Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper and Delicassen. All our variables have integer values.

Channel and Region have discrete values. 1 & 2 for Channel and 1,2 & 3 for Region. The rest of the variables have continuous integer values.

3.2 Answer to Question 2

A detailed description of all variables is seen below.

3.2.1 CHANNEL

Channels Visual and Numerical description

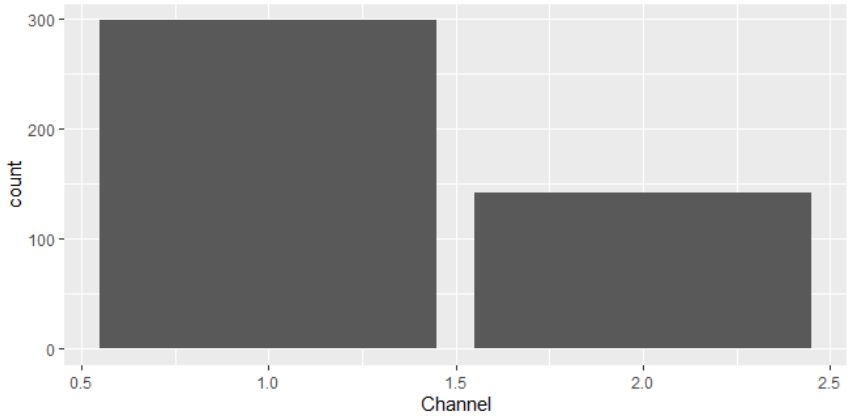


Figure 3.1: A graphical view of Channel values.

Figure 3.1 is a bar chart which was deployed using the ggplot to help give a visual description of the variable.

From the chart, the discrete value of 1 had a total count of 300 and 2 had a total count of 140.

The maximum value of Channel is 1 and that of its maximum is 2. The mean value for the variable is 1.323 and a median of 1. Its 1st quartile is 1 and 3rd quartile is 2.

3.2.2 Region

Region Visuals and Description

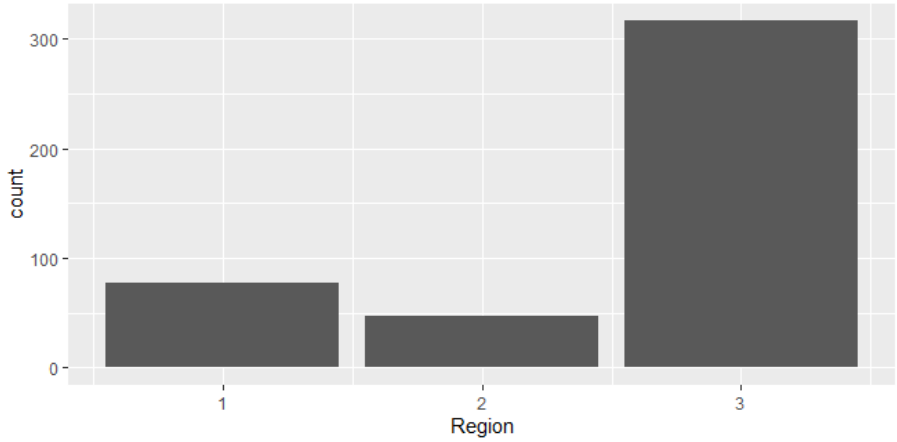


Figure 3.2: A graphical view of Region values.

A bar chart was deployed using the ggplot to help give a visual description of the Region variable. From the Chart, 1 had a count of 77 out of the total 440, whiles 2 had a count of 47

and 3 had a total frequency of 316 hence having the highest frequency. The mean value of the Region data is 2.543 and its median is 3.

3.2.3 FRESH

Fresh Visuals and Numeric Description

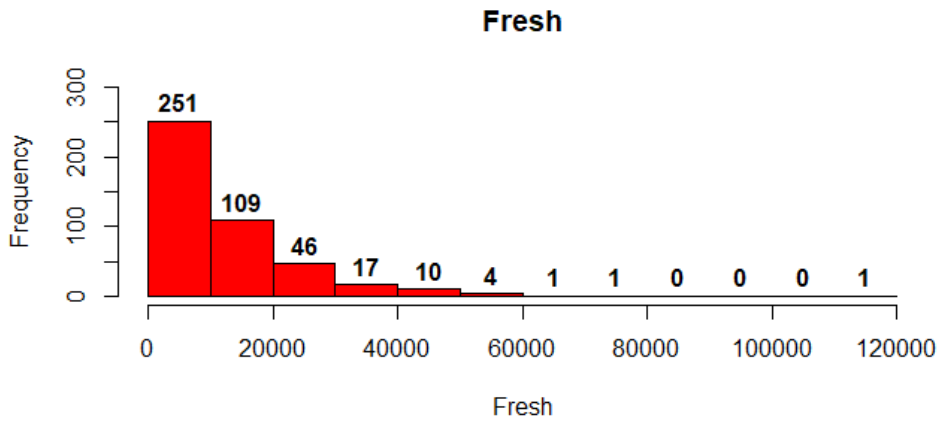


Figure 3.3: A graphical view of Fresh values.

Figure 3.3 is a histogram that shows the details of the variable Fresh. We have a right tiled diagram thus as the annual spending of Fresh product increases, the number of counts in that regard decreases.

The 0 – 1000 range had the highest count of 251 followed by 1001 to the 2000 range which had a value of 109 count. The 2001 – 6000 range had a total count of 77. We had one (1) count for the 100000 to 120000 range.

The maximum value of Fresh is 112151 and its minimum is 3. The mean score for the values of Fres is 12000 and its median value is 8504 The 1st Quartile is 3128 and 3rd Quartile is 16934.

3.2.4 MILK

Milk Visuals and Numeric

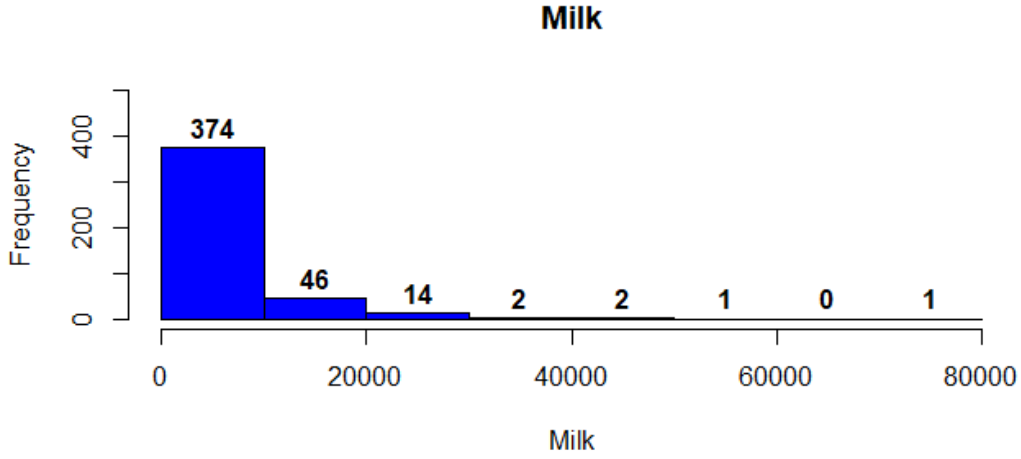


Figure 3.4: A graphical view of Milk values.

Figure 3.4 above is a histogram describing the number of counts of the annual spending on Milk products. The annual spending Interval is 10000.

The range of 0 to 10000 annual spending had a count of 374, which was the highest, followed by the second spending interval having a frequency of 46. The range of 20001 to 30000 had a count of 14 people annuals spending. the total people with an annual spending on the Milk product from 30001 to 80000 was 6.

The Annual spending od Milk had a minimum value of 55 and a maximum value of 73498. The mean score for the spending was 5796 with a median mark of 3627. The 1st Quartile was 1533 and the 3rd Quartile was 7190.

3.2.5 GROCERY

Figure 3.5: A graphical view of Grocery values.

From Figure 3.5, annual spending ranges from 0 to 1000 had the highest count of 323 followed by the annual spending range of 1001 to 2000 which had a count of 79. The count decreased to 27 and 7 for the ranges of 2001 to 3000 and 3001 to 4000 respectively. The annual spending from 4001 to 10000 had a total count of 5.

The data of Grocery has its lowest annual spending of 3 and highest annual spending of 92780. The mean for the grocery is 7951 and its median is 4756. The 1st Quartile was 2153 and 3rd Quartile is 10656.

3.2.6 Frozen

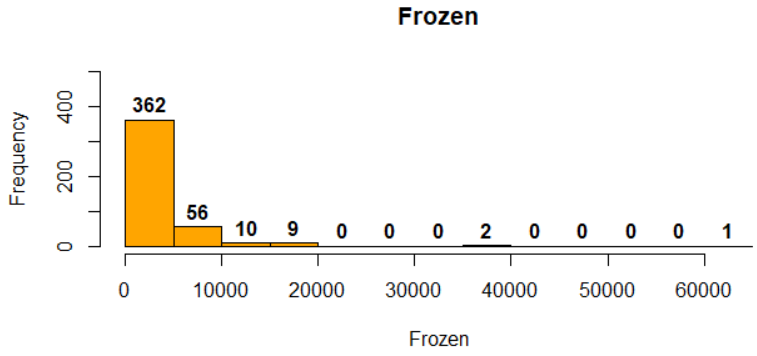


Figure 3.6: A graphical view of Frozen values.

Figure 3.6 is a histogram showing the values of Frozen. The highest annual spending count was 362 which was in the 0 to 5000 range. The next range of 5000 to 10000 had a total count of 56. The range of 10001 to the highest annual spending value had a total count of 22.

The mean value of Frozen was 3071.9 and its median score was 1526.0 The maximum value of Frozen is 40827 and its minimum value as 3. The 1st Quartile of Frozen from our data is 256.8 and 3rd Quartile is 3922.

3.2.7 Detergents_Paper

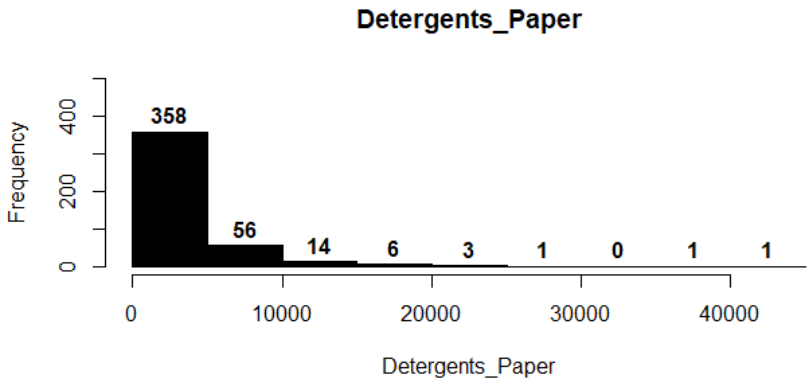


Figure 3.7: A graphical view of Detergents_Paper values.

The histogram above shows that, the range of 0 to 5000 had the highest count form the Detergents and Paper data with a score of 358 out of the 440 observations. The second highest count was a value of 56 which corresponds to the range of 50001 to 10000. The range of 10001 to 20000 had a total count of 20 whiles the ranges from 20001 to 41000 had a total count of 12.

The maximum value from the Detergent and Paper data is 40827 and its minimum annual spending was 3. The mean score from our data is 2881.5 and median score is 816.5. The 1st Quartile of the Detergent and Paper data is 256.8 and its 3rd Quartile is 3922

3.2.8 DELICASSEN

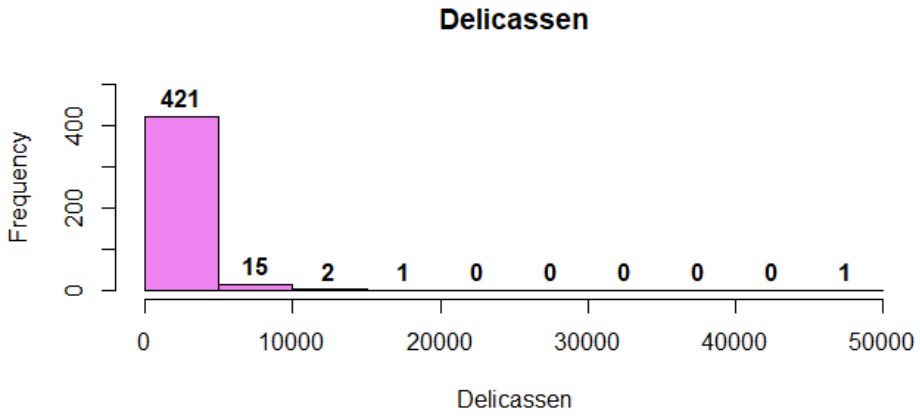


Figure 3.8: A graphical view of Delicassen values.

The histogram in Figure 3.8 is a graphical image of Delicassen observations form our data.

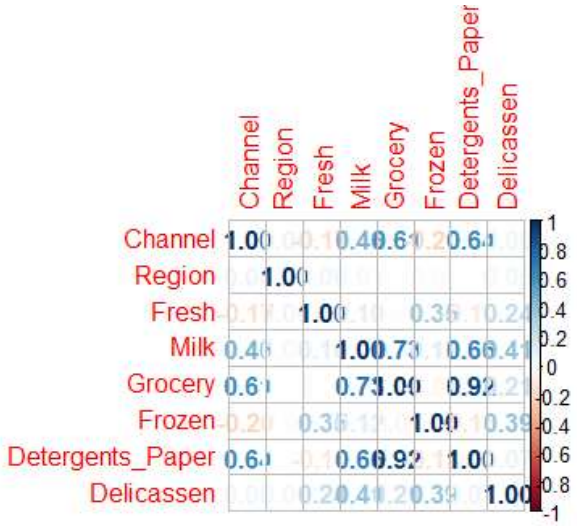
The highest annual spending of Delicassen was between 0 to 5000, which had a count of 421 out of 440 representing a percentage of approximately 95%. The remainder of the percentage was shared between the ranges from 5001 to 50000.

This shows that a huge chuck of our clients annual spending on Delicassen is within the first 10000.

3.3 Answer to Question 3

The values of the correlation of our variables are assigned to the WScor. The diagram below is a pictorial view of the correlation that exist between each variable. The corrplot was used in giving the pictorial view.

Figure 3.9: A graphical representation of the Coefficient of the Whole Dataset.



From Figure 3.9, the correlation that existed between our variables are explained below. Remember every variable correlates with itself highly hence giving a value of 1.

3.3.1 CHANNEL

Channel had a positive correlation with the variables Region, Milk, Grocery, Detergents and Delicassen with a correlation co-efficient of 0.062, 0.467,0.687,0.636 and 0.056 respectively. These correlations aren't strongly correlated but it shows that Channel and each of these variables moves in the same direction thus as channel increases, they also increase and vice versa.

Furthermore, Channel had a negative correlation with Fresh and Frozen with a coefficient of 0.169 and 0.202. these correlations are weak but it implies that Channel and each of these variables move in an opposite direction.

3.3.2 Region

The variable Region had a weak positive correlation with Channel, Fresh, Milk, Grocery and Delicassen. The coefficient was 0.062, 0.052, 0.032, 0.007, 0.045. this implies Region and these variables move in the same direction, thus if Region decrease these also decrease.

Additionally, Region has a weak negative correlation with Frozen and Detergents and Paper with a coefficient of 00.021 and 0.001. These variables move in opposite direction with Region.

3.3.3 Fresh

Fresh has a weak correlation with all the variables in our data. It had a positive coefficient of 0.055,0.100,0.345 and 0.244 with Region, Milk, Frozen and Delicassen respectively. Its negative coefficient was with the variables Channel, Grocery and Detergents and Paper with a value of 0.169,0.011 and 0.101.

3.3.4 Milk

From our Correlation diagram, Milk had a strong positive correlation with Grocery and Detergents and Paper. With a coefficient value of 0.728 and 0.66.

It had a weak positive correlation with Channel, Region, Fresh, Frozen and Delicassen. The coefficient value is 0.460,0.032,0.100, 0.12 and 0.406 respectively.

In conclusion, it had no negative correlation with any of the variables.

3.3.5 Grocery

For the variable Grocery, it had a very strong positive correlation with Detergent and Papers with a correlation coefficient of 0.924.

It had a positive correlation with Channel, Milk, Delicassen and Region with a coefficient of 0.608, 0.728, 0.205 and a 0.007. Thus, these variables also move in the same direction with Grocery.

3.3.6 Frozen

Form the chart, Frozen has a positive correlation with Fresh, Milk and Delicassen. The coefficient between these were 0.345, 0.123 and 0.390.

It had a negative correlation with Channel with a value of 0.202, Region with a value of 0.021, with grocery with a value of 0.040 and with Detergents and Paper with a value of 0.131. Implying that Frozen and these variables move in the opposite direction.

3.3.7 Detergents and Papers

Detergents and Papers had a very strong correlation with Grocery with a coefficient value of 0.924. It also had a positive correlation with Channel, Milk and Delicassen with a coefficient value of 0.636, 0.661 and 0.069.

Additional it had a negative correlation with Region, Fresh and Frozen, with a coefficient value of 0.001, 0.101 and 0.131 respectively.

3.3.8 Delicassen

Delicassen had a positive correlation with all the variables in the data. With Channel it had a coefficient value of 0.056, and with Region 0.045, Fresh it had a value of 0.244, Milk it had a value of 0.406, with Grocery it had a coefficient value of 0.205, Frozen it had a value of 0.399 and with Detergents and Paper it had a coefficient value of 0.069. this means that as the Delicassen and all the variables in the data moves in the same direction.

3.4 Answer to Question 4

Normalization of our data implies giving a set of weight or boundary to our variables so they can easily be measured equally. Thus, we want each column to give the same impact on the measurement. In other words, irrespective of how big a value is it is graded in a specified range so for easy comparing.

It is needed when variables are of incomparable values and hence common scaler must be given to variables. Either than that most values of variables will become redundant because only high values of our data will influence the measurement.

Therefore, we use the min-max method where we scale our data to the limit between zero (0) and one (1).

We assign this new data frame of Normalization as 'NorWS'.

3.5 Answer to Question 5

To determine the optimal number of clusters for K- means, these various methods will be used in ascertaining the cluster.

3.5.1 K-means Algorithm

One can use the K – means algorithm on randomly. K-means function works in a way that whatever you tell it to produce is what you get. This will be done by selecting a random normalization of a point thus a center of 2 for the clusters and another perimeter to use in the function is the 'nstart' which helps to randomly configurate the best possible cluster. Hence, we take 25.

The K-means cluster tells us the cluster membership of each of the membership of our variables.

The K-means center tells as the point at which all the variables point are been centered or compared against. It shows as the range for which our clusters should be in.

K-size tell us how many of the points falls within a particular cluster.

3.5.2 Using the Elbow Method

It is a visual method where we plot the within sum of squares distance for the entire clustering and their assigned clusters centers.

The rule is that we pick the number of cluster (K) for which we see a significant decrease in the sum of squared distance hence the point forming an elbow.

With the plotting, the number of clusters is on the x-axis and the total sum of means/distance is on the y-axis.

In this method there is a limit set for which the cluster centers should fall, it implies that our cluster centers should take these values and at the point where the total sum of squares begins to level up is selected as our point of improvement.

The only problem with the Elbow method is that the more the clusters, the smaller the average distance of points to the cluster center. (Decrease in the within sum of squares)

Figure 3.10: A graphical representation of the Elbow Method.

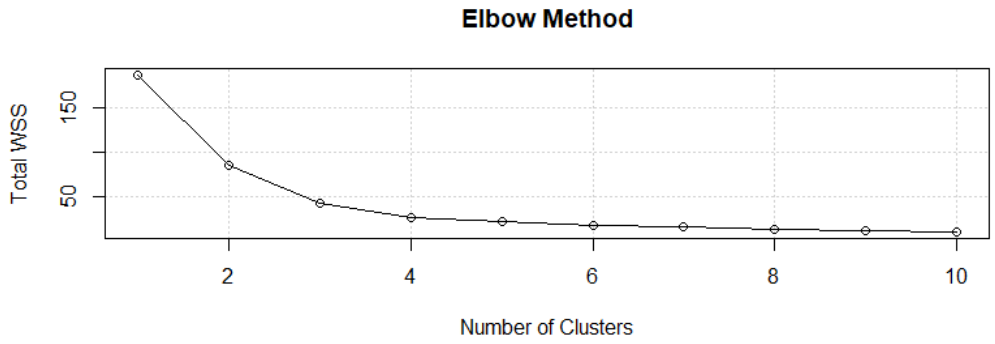


Figure 3.10 is the Elbow Method diagram. From the visualization the number of clusters bend at 2 thus at this point our line flattens and after wards there is very low marginal improvement after and hence the optimal cluster center is 2.

3.5.3 Using the Silhouette Method

It can be used to determine the degree of separation between clusters.

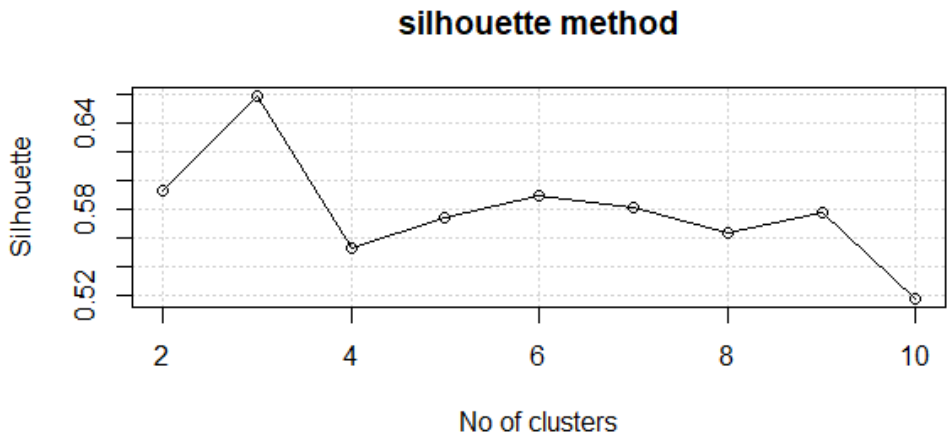


Figure 3.11: A graphical representation of the Silhuotte Method.

The Figure 3.11, is a visualization of the Silhouette method and at the highest points 2 and 3 we see that our data is well matched or similar to its own cluster. Because the closer it is to 1 the best it is, hence 2 is accepted in the silhoutte method.

3.5.4 Gap Method

Here our data is compared to the reference data with a random uniform distribution. And as they are compared, the point at which there is a big change is accepted as our optimal cluster center.

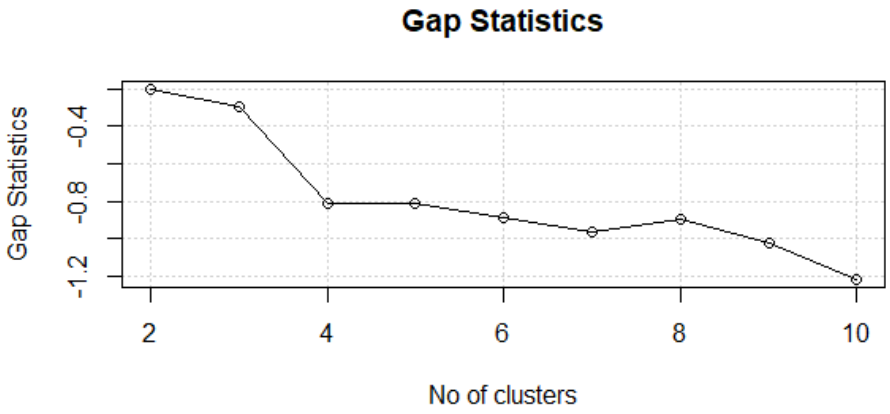


Figure 3.12: A graphical representation of the Gap Statistics Method

From Figure 3.12, at the cluster point 2 and 3 we see the difference in the comparison of the reference data and the random uniform distribution been the highest points. Therefore, I take point 2 as the optimal K cluster center.

3.5.5 Calinski-Harabasz Method.

This method takes into effect the intra class similarity and the inter class dissimilarity by using the Between Group Sum of Squares (BGSS) and the With-in Group Sum of Squares (WSGG).

The WSGG is the distance of observation to the cluster centers but BGSS is the difference of the cluster center and the mean value of the data. The higher the score the better because it means that our data has a higher intra class similarity and a lower inter class dissimilarity.

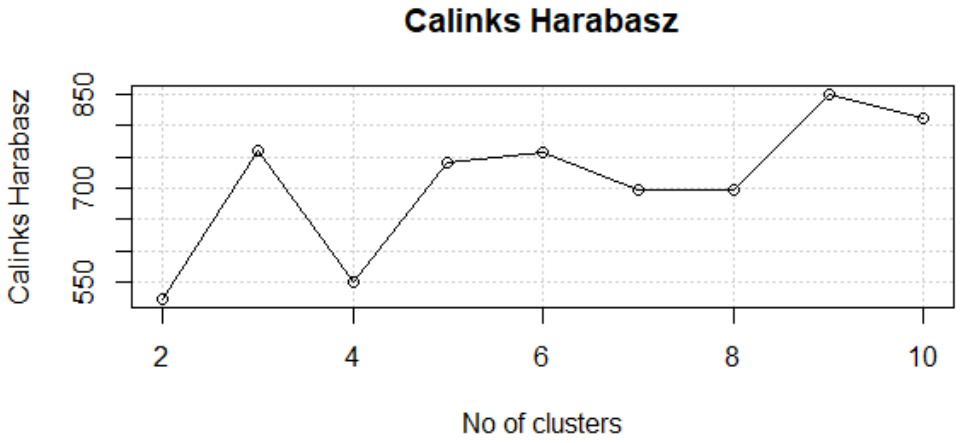


Figure 3.13: A graphical representation of the Calinks Harabasz.

Figure 3.13 is a graphical representation of the Calinks Harabasz method. From the graph, the highest points for the Clusters are 9,8 and 3.

3.5.6 CHOOSING CLUSTER NUMBER

From our Methods used we could see differences in the cluster center for each method. But among all this difference there is a common cluster number that repeats itself in the Gap Statistics, Silhuotte method and the Elbow method which is the cluster center been 2.

Hence I go with a cluster center of 2 because the elbow method has it as its optimal cluster number and also because the average distance between the points are close.

3.6 Answer to Question 6

Now focusing on the unnormalized data set with a cluster number of 2 I run the K-means algorithm with nstart of 25.

The nstart is used as a configurator in the K-means algorithm it serves to run the algorithm for a number of times and chooses the best cluster grouping to be used. So, meaning setting nstart to 25 will make the k-means algorithm to have a multiple configuration of 25 and hence select the best one that minimizes the cost.

The cluster generated by the k-means is mutated with the original data set to get the membership of variables values. The mean scores of the variables are derived to know which of the variables contributed greatly to the membership or grouping of the clusters.

Fresh had an average of 35401 for its contribution in cluster 1 and 7944 for its contribution in cluster 2. Milk also had 9514 as an average contribution in cluster 1 and 5152 for cluster 2. Grocery also had 10346 for cluster 1 and 7536 for cluster 2. Frozen also had 6463 for cluster 1 and 2484 for cluster 2. Detergents had a contribution of 2933 and 2873 for cluster 1 and cluster 2 respectively Lastly, Delicassen had a mean score of 3312 in cluster 1 and 1214 for cluster 2.

So, we can see that all the variables except for Channel and Region are driving the clustering.

But mainly Fresh and Grocery had a high contribution in defining the clusters.

3.6.1 Visualization of cluster result in grouping

Now we visualize and see how these two variables define our clustering by using a ggplot.

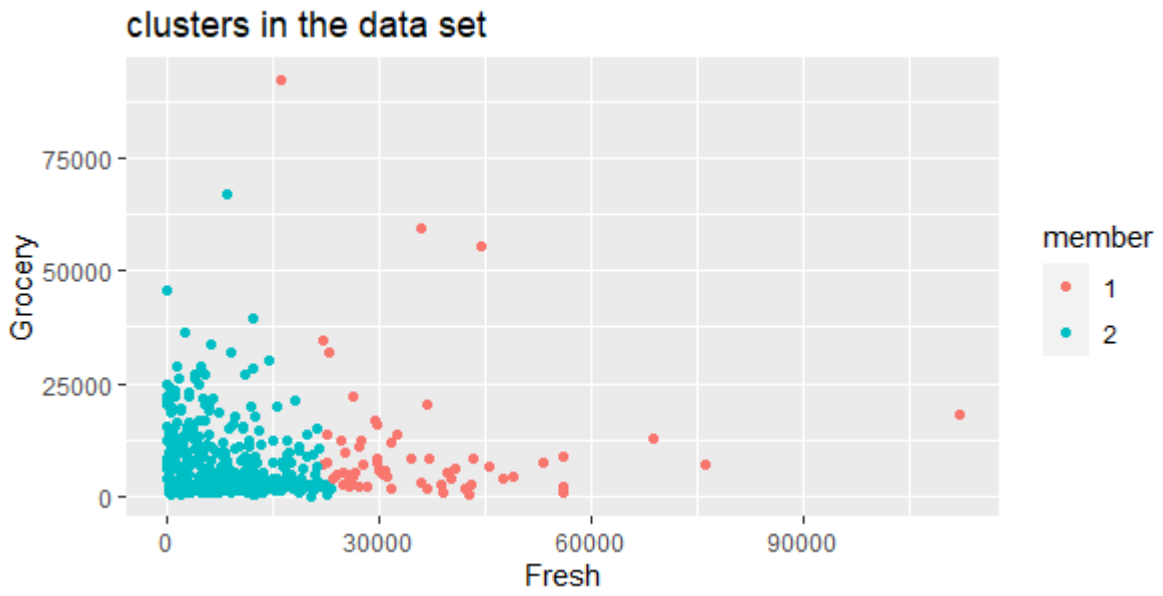


Figure 3.14:A plot of cluster for Fresh and Grocery.

Visualization of each variable with respect to the cluster group.

3.6.2 Fresh

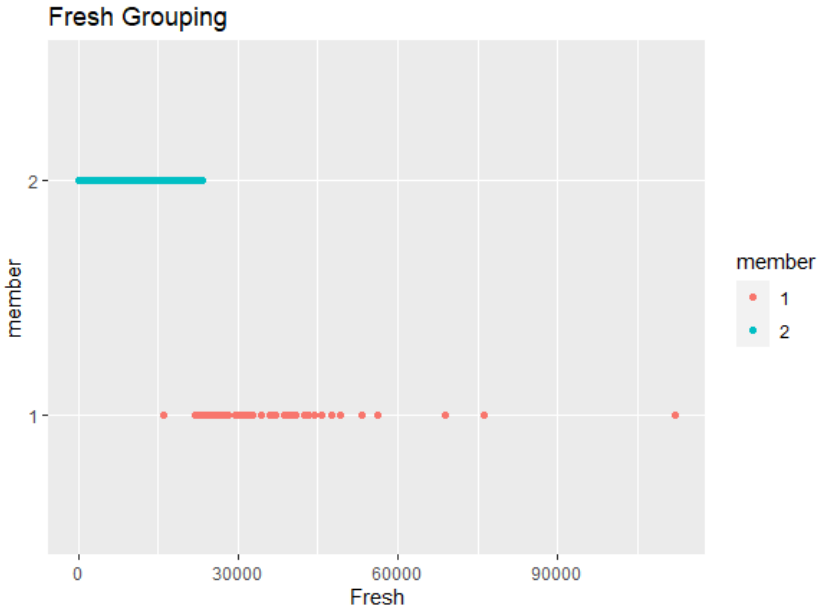


Figure 3.15: A cluster of Grouping for Fresh

From Figure 3.15 the plot shows majority of the values were found in the 2nd cluster group. And these were ranging from approximately 1800 spending and above. The Intra similarity in the cluster is quite wide.

3.6.3 Frozen

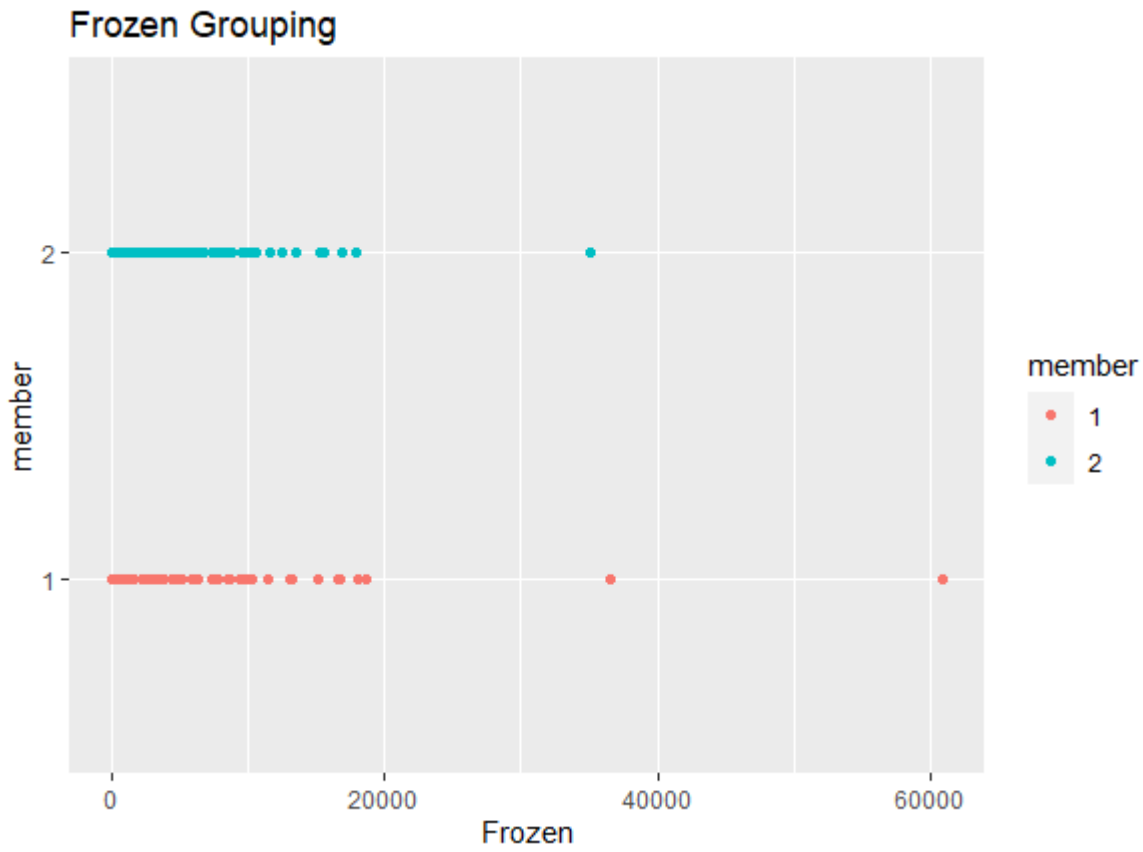


Figure 3.16: A cluster of Grouping for Frozen

The chart above shows that for the Frozen variable, the inter dissimilarity among the cluster grouping is almost the same. Implying that the range of values that are in class one are almost same for the cluster group 2

3.6.4 Milk

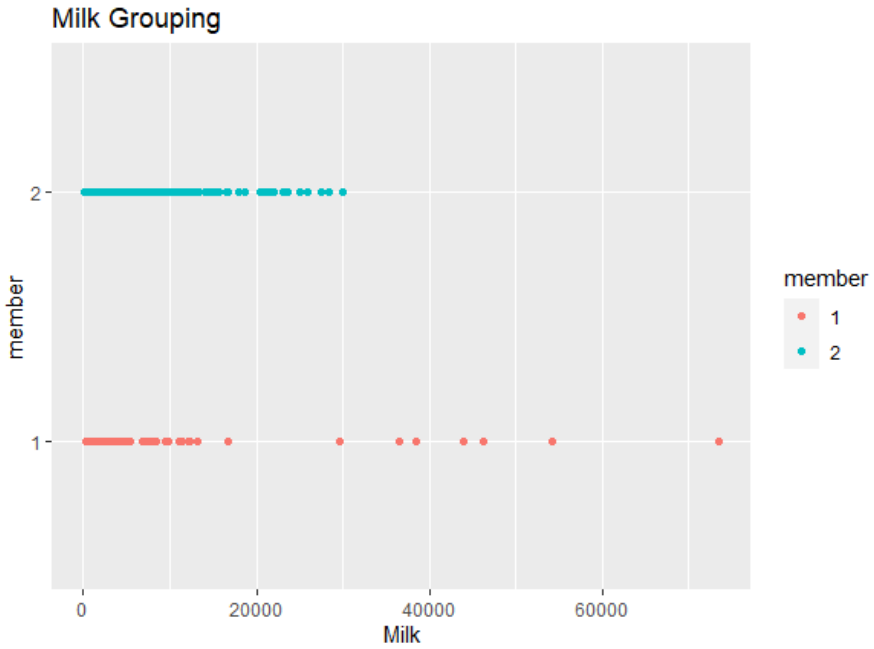


Figure 3.17: A cluster of Grouping for Milk

From Figure 3.17 the chart shows Milk had two cluster grouping which are really inter similar in nature. The spending ranges from 0 to approximately 1500 are inter similar to each other whiles the range of approximal 1600 spending and above were intra similar to each other in the cluster member 2.

3.6.5 Grocery

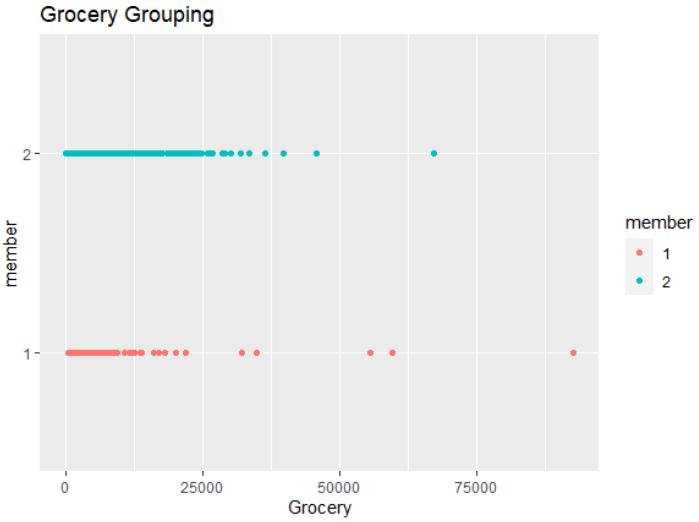


Figure 3.18: A cluster of Grouping for Grocery

Grocery spending range in the cluster inter dissimilarity is low, thus the grouping of this variable values is almost the same to each other.

3.6.6 Detergents_Paper

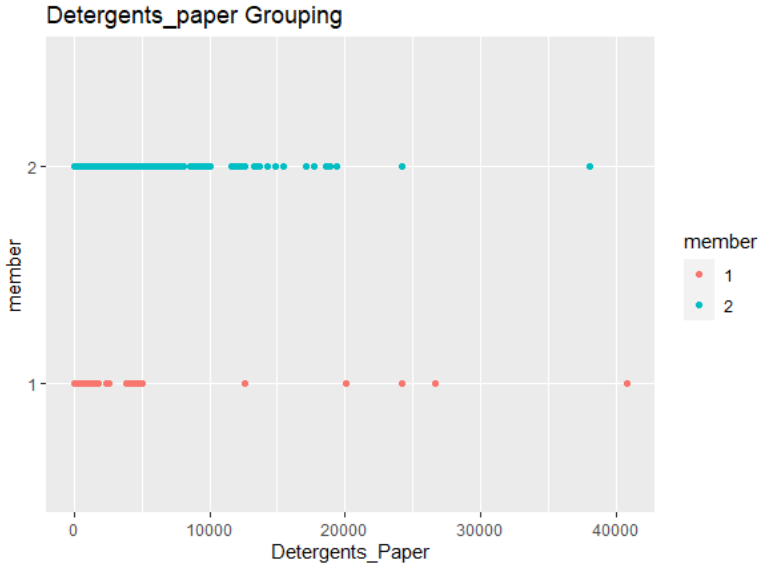


Figure 3.19: A cluster of Grouping for Detergents_Paper

Figure 3.19, shows that for the cluster groups spending of Detergents_Paper had a low inter dissimilarity at the beginning stage but after the 5000-spending mark almost all of the values where within the Cluster 2 hence having a high intra similarity.

3.6.7 Delicassen

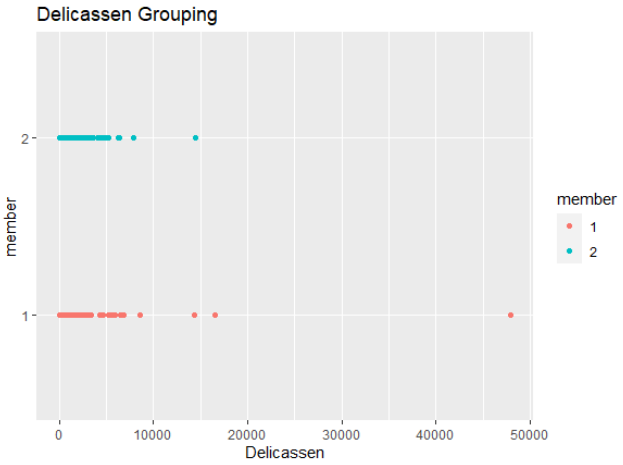


Figure 3.20: A cluster of Grouping for Delicassen

Delicassen cluster grouping spending range had a low inter dissimilarity among itself.

3.7 Question 7.

From the overall look at the clustering of the wholesale dataset, mine advise will be as follows the marketing department.

3.7.1 Channel

They should frequently use Channel 1 because via that channel, sales for the firm increases greatly as compared to Channel 2. For Channel 2, they should make huge improvements if they want to utilize that platform

3.7.2 Region

The third 3rd Region contributed a lot to the sales amount of the firm hence most products should be sent there for frequent increase in the sales figure.

3.7.3 Fresh

Most of our clients were in the cluster group 2 implying that the spending range of our customers for Fresh products is high. Thus, annually our clients spend 5000 and above individually meaning they consume a lot of Fresh products hence we should focus on producing more Fresh products for an increase in our sale revenue.

3.7.4 Grocery

Among all our variables the Grocery products fetching the firm a small amount of sales revenue. So, my advice is to advertise more on our Grocery products so as to encourage our clients to purchase more of the Grocery products hence increasing sales.

The rest of the products are averagely purchased by our clients hence there is no huge marketing strategy to do. The normal advertisement of these products should be done and as the year goes by another review will be made will see their impact on sales.