# Free Analytics Environment R

## Assignment 2: Report for part 1

### Student Number: 001010879

# Introduction:

We have been assigned a task of building a **regression model** that will indicate the **Murder Arrest** per **100,000** residents in Europe. For this we were given a data sets which included crimes throughout Europe such as **assault**, **murder**, **drugs** related crimes, **traffic violations**, **cyber crimes**, **domestic violence**, **alcohol** related crimes, **kidnappings** and relevant information such as **urban population** in that area and the number of recorded **car accidents**.
Our regression analysis will figure out the factors which maybe linked to murder arrests and relevant information.

## The data frame contained following explanatory variables:

| | Explanatory Variables |
|---|---|
| 1 | Assault |
| 2 | Murder |
| 3 | Drug |
| 4 | Traffic |
| 5 | Cyber |
| 6 | Kidnapping |
| 7 | Domestic |
| 8 | Alcohol |
| 9 | Car Accidents |
| 10 | Urban Population |

## Installed and loaded Packages:

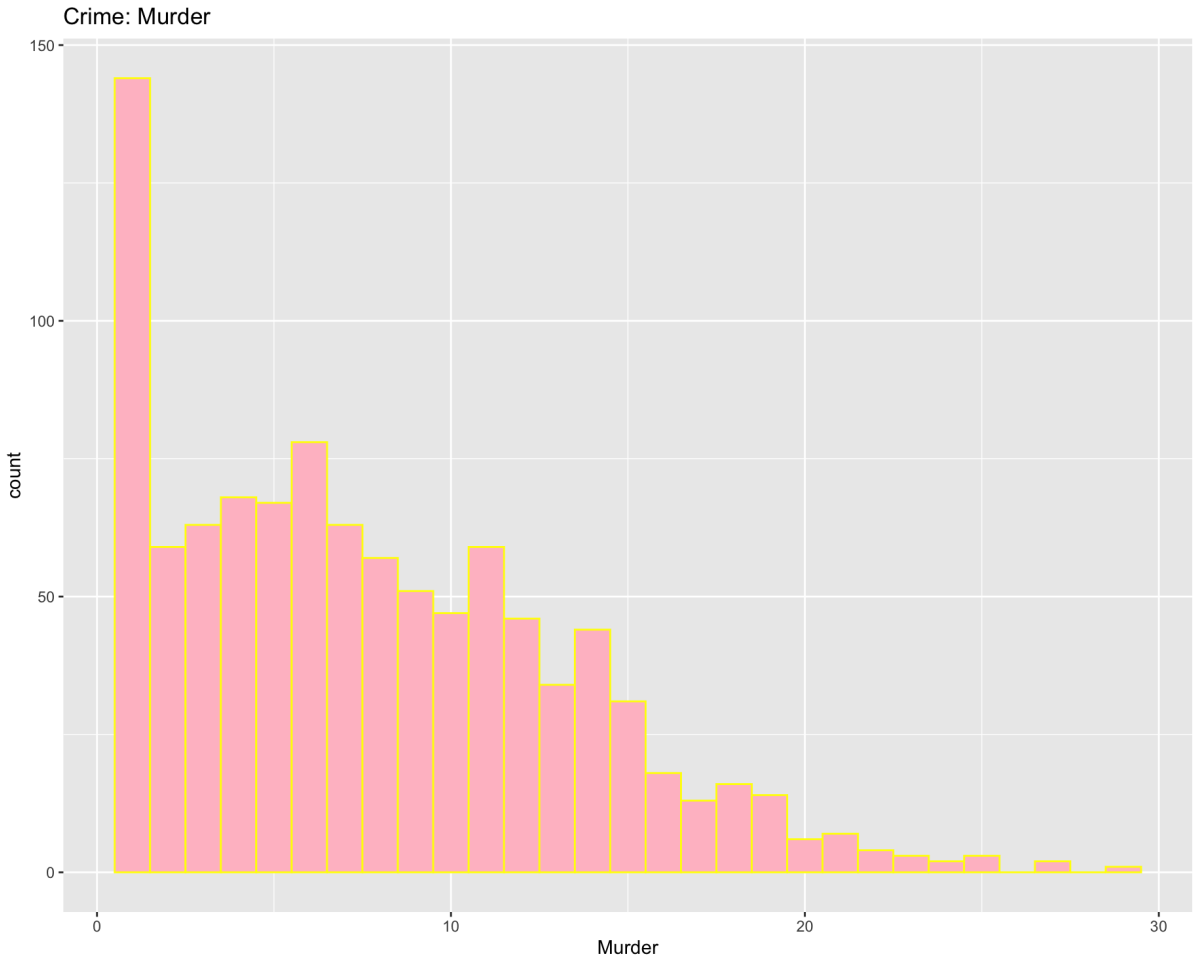
- Ggplot2
- Corrplot
- Tsoutliers
- Psych
- Dplyr
- Datasets

## 1- Loading data sets to R script to analyse data:

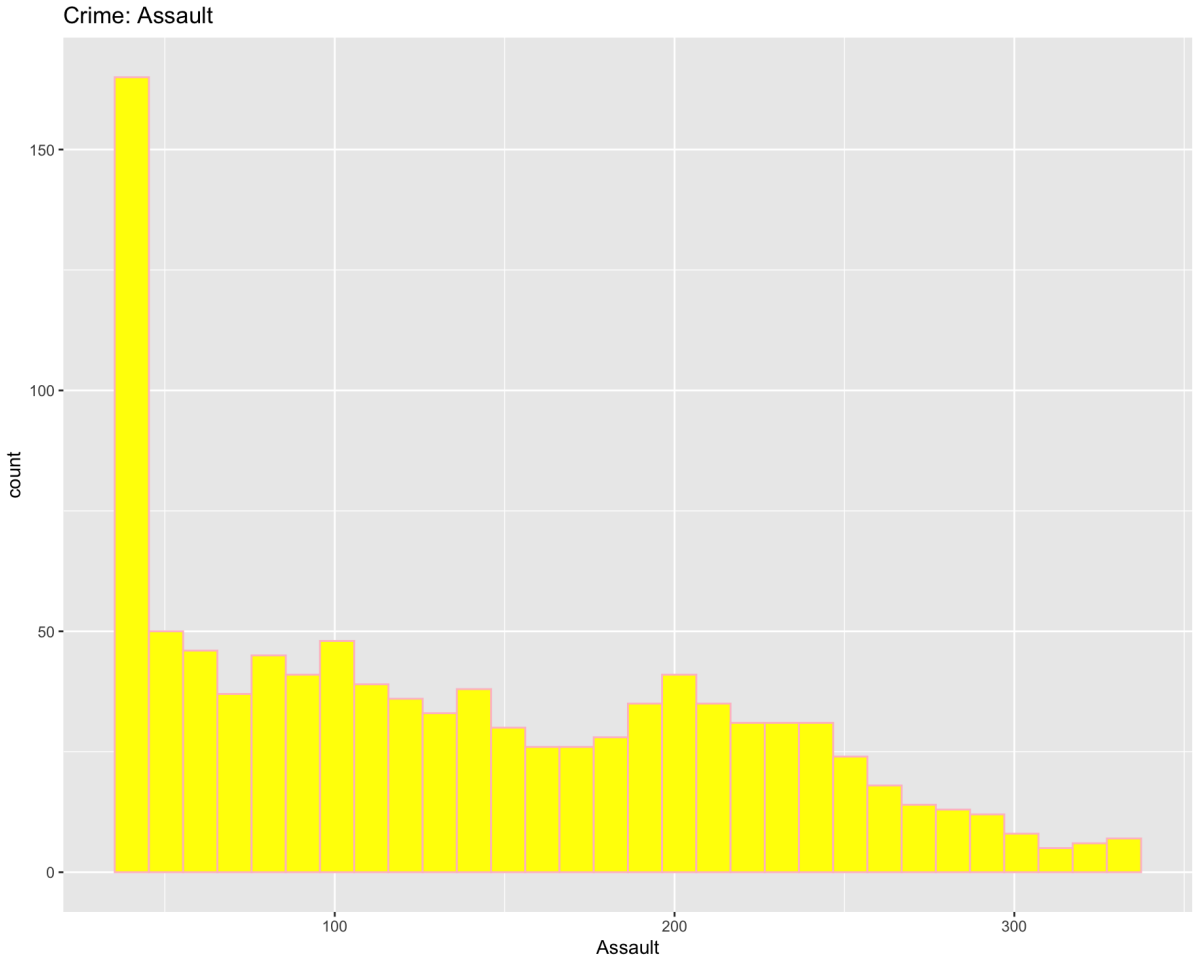
Used **read** command to load the data into our R environment so we can analyse the data sets. The data set was available in the "**dataArrests_Mac.csv file**" and we saved it into variable **my_data**.
After successfully loading the data, we used the structure command to observe that the data set consists of **995 observations** and **10 explanatory variables**. The explanatory variables had **Integer** and **numeric** classes.
On exploring, we figured out that the dataset had some **missing values** too hence those missing values were removed through **exploratory preprocessing techniques**.

## 2- Plotting

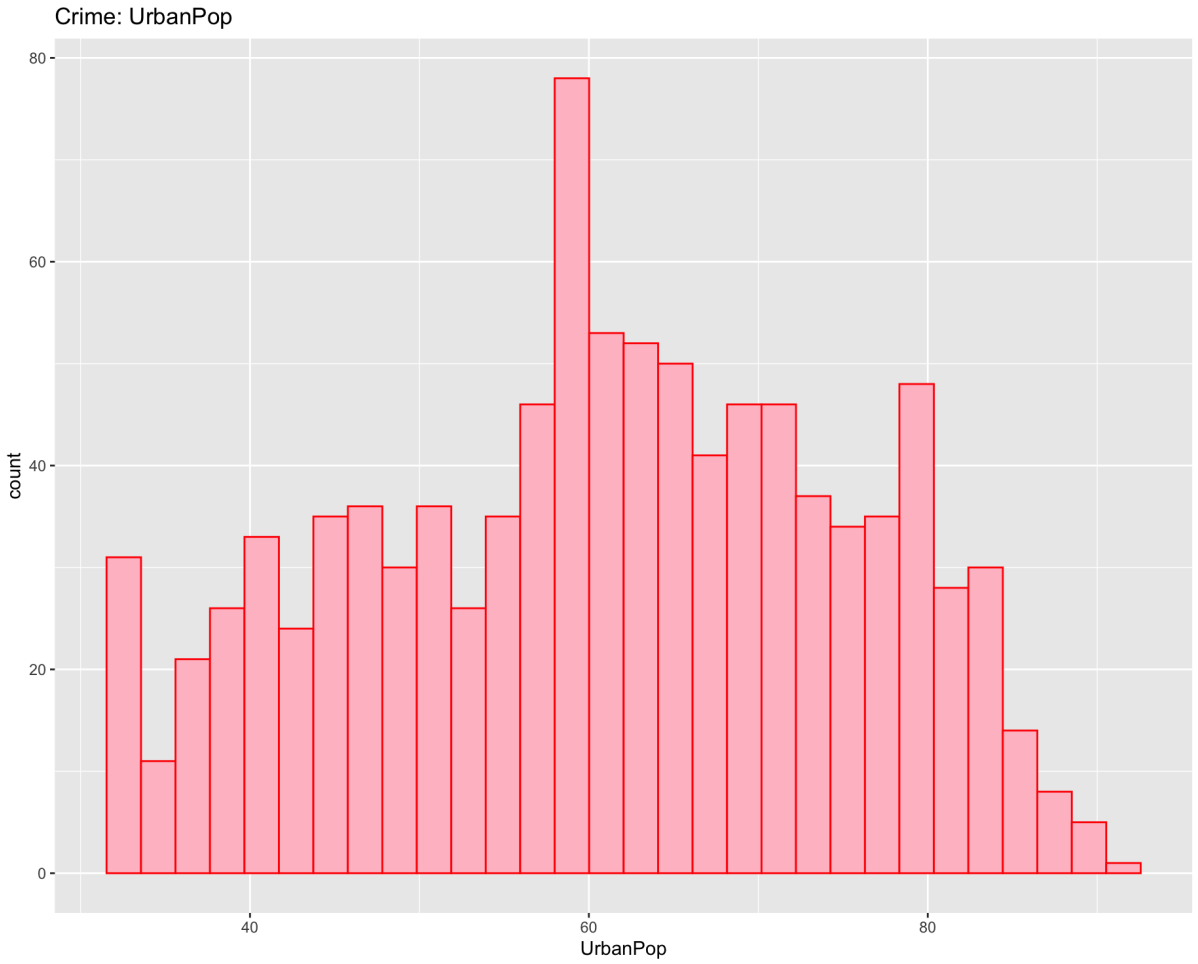
- Dependent variable Murder : We plotted dependent variable Murder in a histogram. By analysing histogram we found out that Murder was decreasing by count



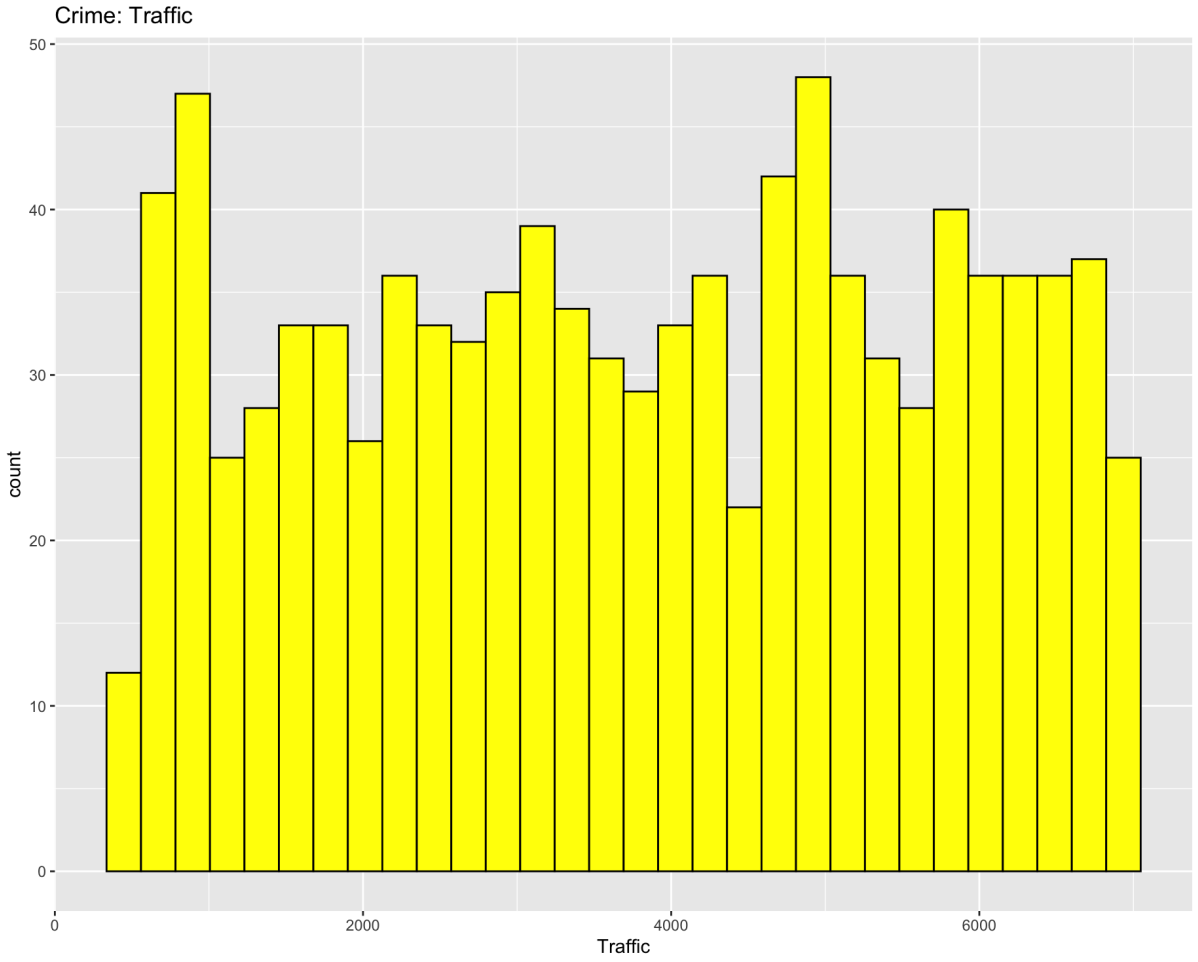
- **Assault Variable:** We plotted Assault Variable to found by analysing, found out that assault was decreasing by count



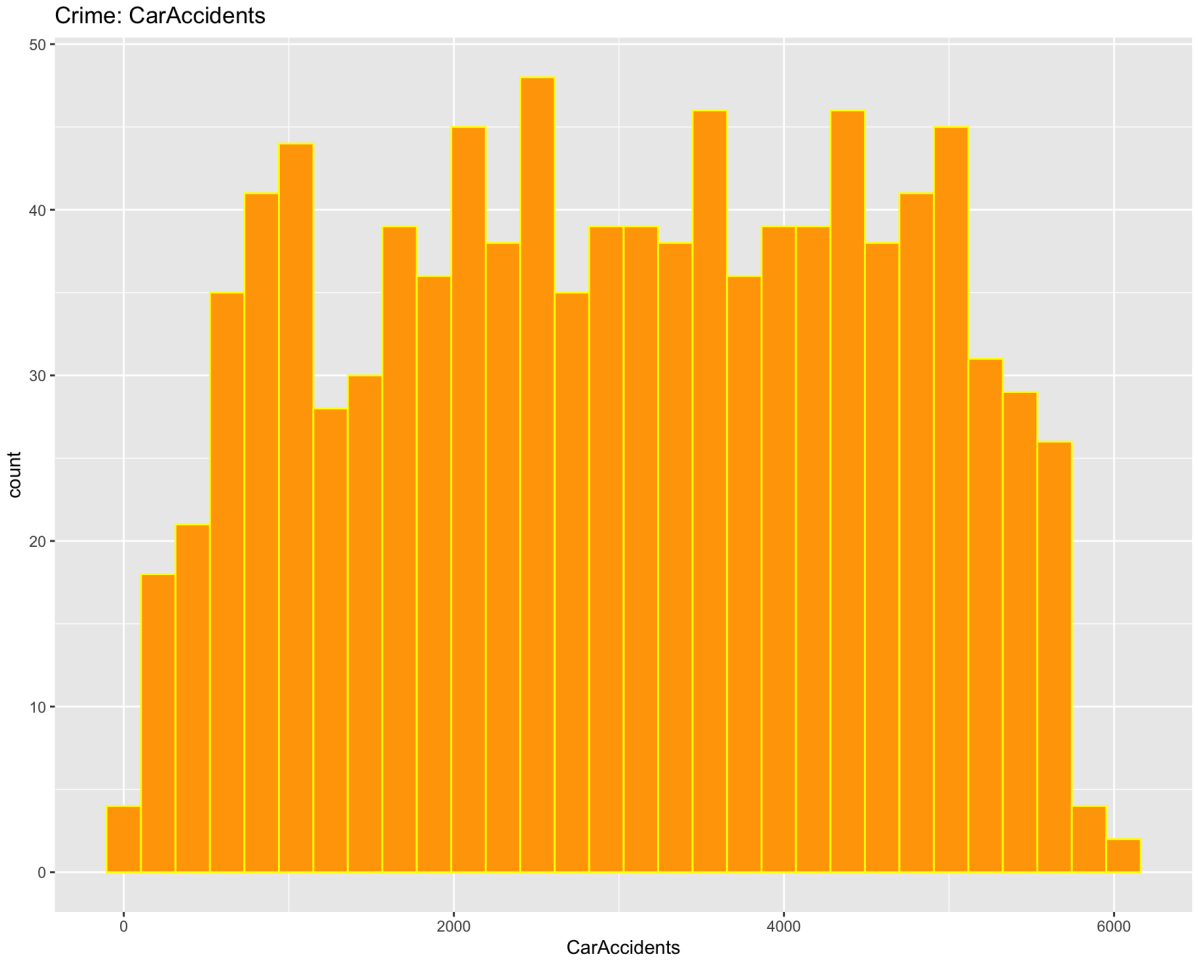
- **Urbanpop Variable:** After plotting urban pop in histogram we figured out that there was an increase till the middle then it started to decrease



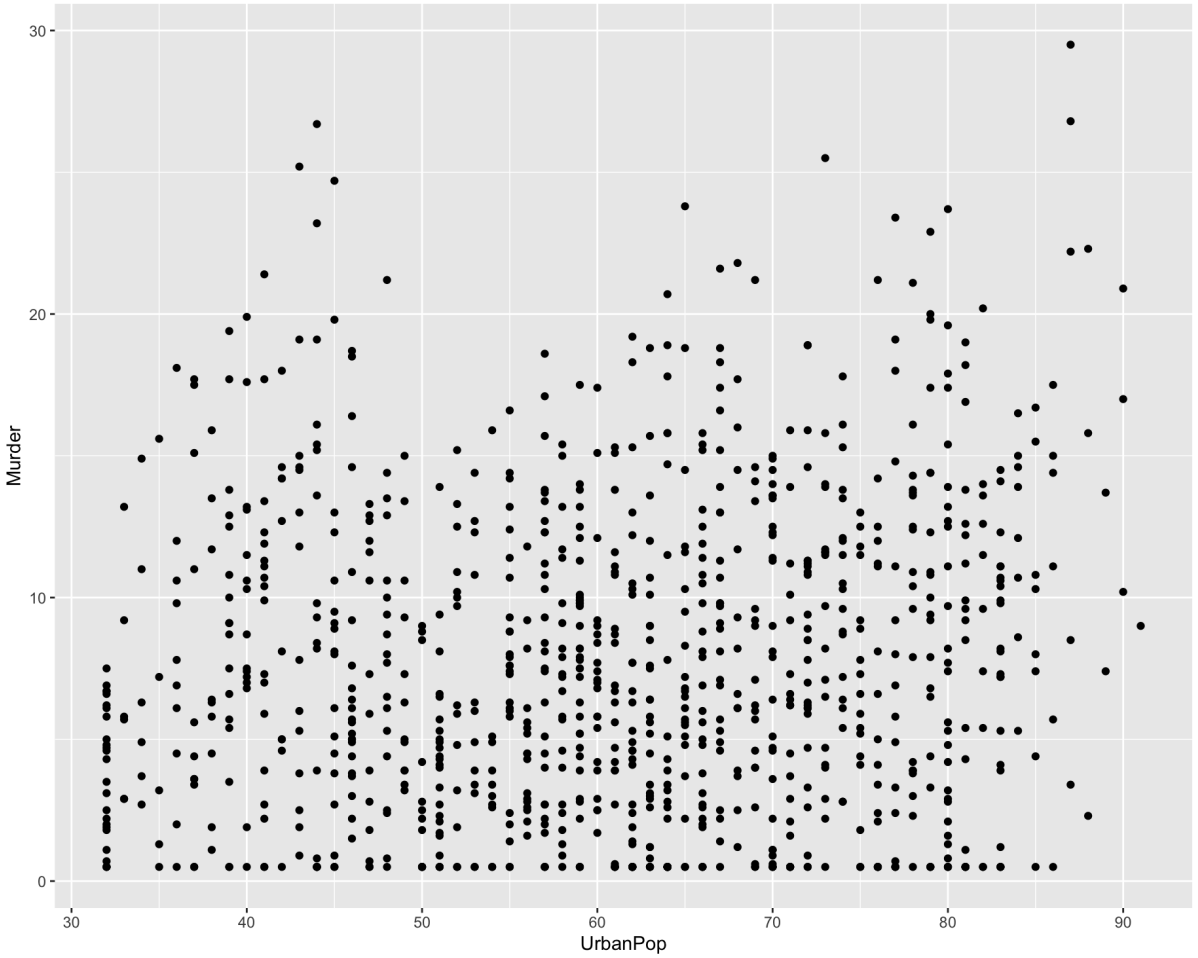
- **Traffic Variable:** After plotting traffic variable in histogram, we figured out that traffic variable increases till the middle and then it starts to decrease irregularly



- **Car Accidents Variable:** After plotting car accident variable in histogram, we analysed that car accidents increased till the middle and then started to decrease irregularly

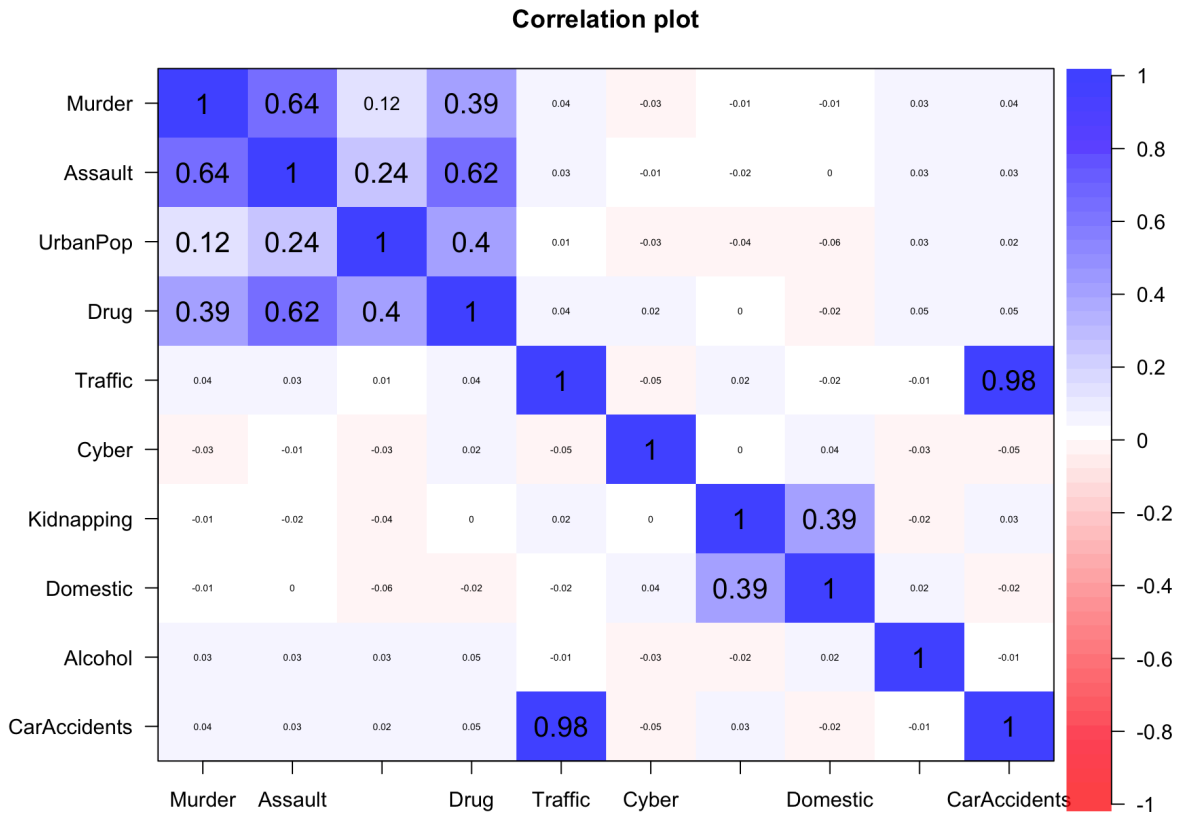



- Then we used ggplot to plot, Urban pop against Murder


30
20
10
0
Murder
30
40
50
60
70
80
90
UrbanPop

## 3- Correlation Plotting:

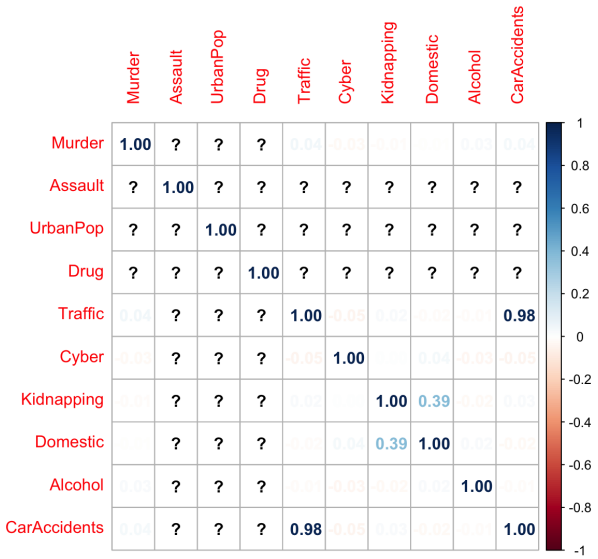
We used correlation plotting from corrplot package which provides a visual exploratory tool on correlation matrix that supports automatic variable re-arranging to help detect hidden patterns in a given data set.
By using corPlot, we figured out that **car accidents** Variable and **traffic variable** are strongly correlated with each other with a value of **0.98** and there is **positive linear correlation** between the same variables.




- **Murder** Variable had the highest correlation with **0.64**, then drug had **0.39**, then Urban population had correlation of **0.12**.
- Explanatory variables **cyber** and **alcohol** had **no correlation** or **linear association**.
- **Assault** had a **linear association** with **Urban population**, with correlation value of **0.24**.

## 4- After correlation plotting:

We weren't able to obtain all the values as these values will be negligible or near to be negligible.




## 5- Removal of high correlation explanatory variables

Then we removed high correlation explanatory variables from the set of all explanatory variables and saved the variables in explanatory/independent variables and dependent variables.

## 6- Figuring out correlation

Then we were able to figure out the correlation

## 7-Assigning '0' to the diagonal

After that, we assigned 0 to the diagonal of correlation of explanatory variables.

## 8- High correlation between explanatory variables

Due to high correlation between explanatory variables, it was required to remove the explanatory variables which had high correlation with each other so we used a while loop to search for max explanatory variables having values greater than or equal to **0.8**. Then, we were able to find the variables with highest absolute correlation and highest average correlation between the variables.

## 9- Removal of High absolute correlation variables

Then we were able to remove the explanatory variables having the highest absolute correlation.

## 10- Developing our Linear Regression Model

At this stage our variables were stable, and we were ready to develop our linear regression model. Hence we used built-in functions  to develop our linear regression model for whole variables.
In addition, we also developed:

- Linear Regression Model without Drugs
- Linear Regression Model without Domestic
- Linear Regression Model without Traffic
- Linear Regression Model without Cyber

Residual standard error: 4.272 on 986 degrees of freedom

Our model's **Multiple R-squared value: 0.4089** and **Adjusted R-squared value: 0.4041** values were between **0 and 1**, and **p-value 2.2e-16 which is ≤ 0.05** means the p-value is **significant**.

## 11- Conducting homoscedasticity test

We also conducted a **homoscedasticity test** for our model to check for dissimilarities in a population and as it invalidates statistical tests of significance that assume that the modelling errors all have the same variance. We also checked whether our **residuals were linearly independent or not**.

## 12- Fulfilment of properties of Linear Regression OLS

Now, at the final stage, we will see if our model fulfils all the properties of Linear Regression OLS, which are as follow:

- The explanatory variables are not linearly dependent
- The residuals have zero mean
  Mean of Residuals: **1.271313e-16**
- The variation of the residual value is constant
- Residuals are normally distributed
- There is no relationship between residuals and each of explanatory variables
- The explanatory values are non-linearly distributed
  To check whether explanatory values are non-linearly distributed or not, we conducted a JarqueBera test to find out whether skewness is equal to **0** and kurtosis is equal to **3**.

So as our **Jarque.Bera test** shows that the value of **skewness** is **nearly equal to 0** and value of **kurtosis** is **nearly equal to 3**, having **p-values 1.641e-07 and 0.009757**, it indicates that our sample data is not normally distributed.

# Assignment 2: Report for part 2

## 1- Loading the file in our R environment

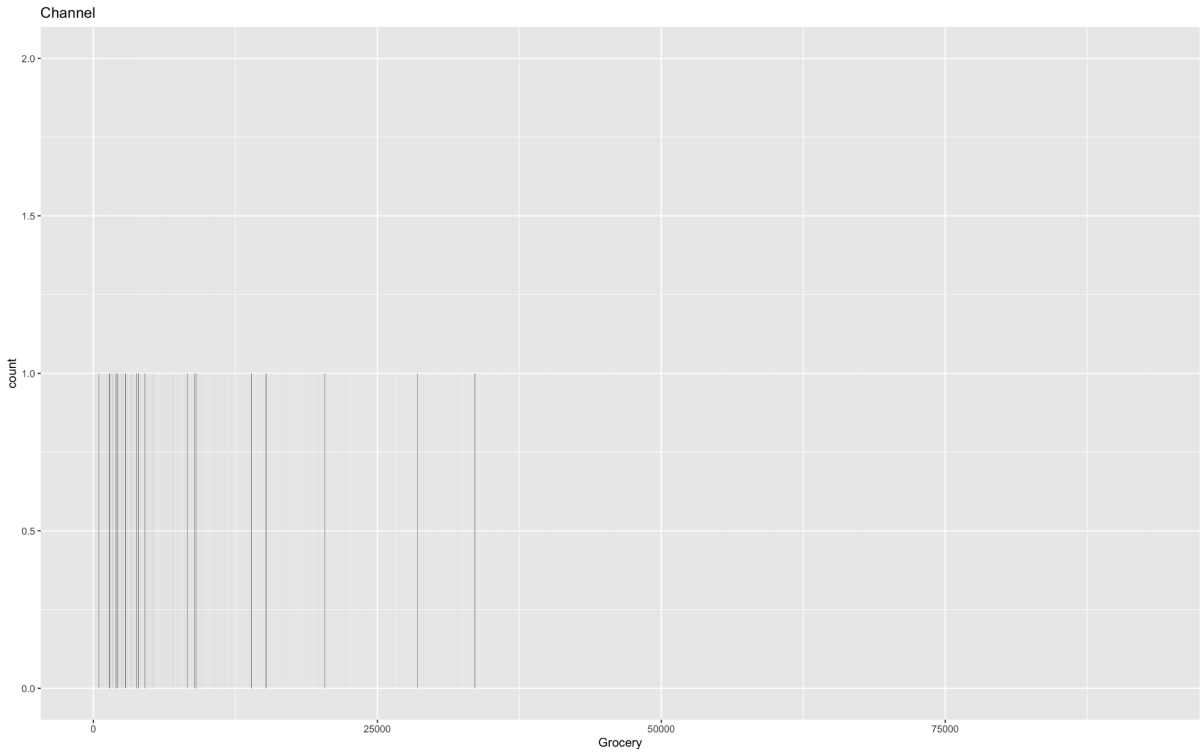
We loaded the file “Wholesale_Mac.csv” in our R environment and saved the data set in my_data, analysed the structure of the given dataset and determined that there are not any null values present in the data.

## 2- Determining whether they are categorical variable

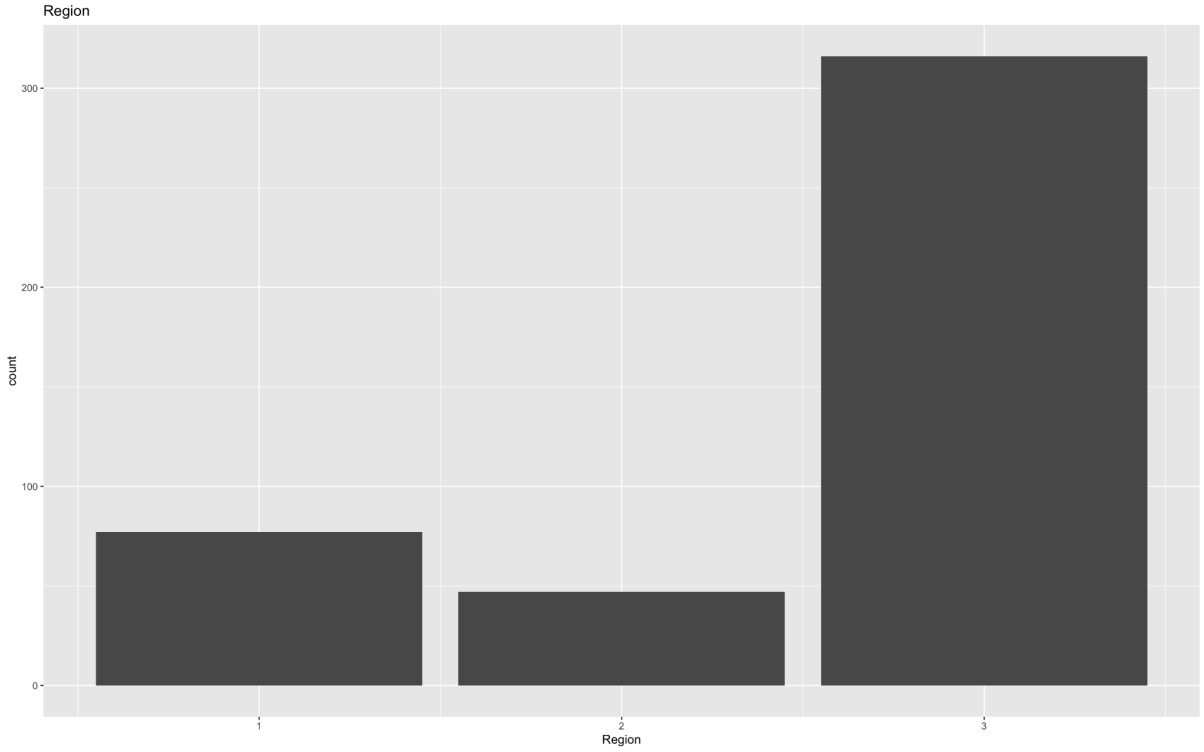
- Channel were categorical variables with class integer 1 and 2
- Region were also categorical variables with class integer 1, 2 and 3

## 3- Plotting

Plot for Channel/Grocery



Plot for Region Count

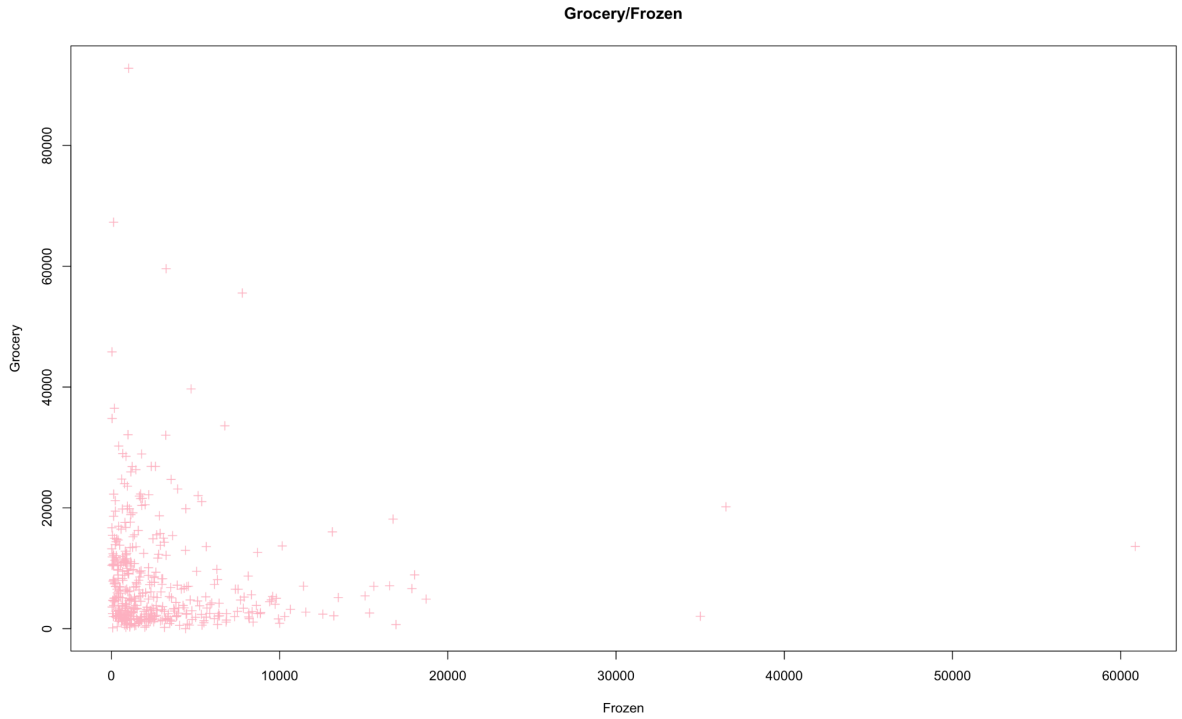



## 4- Creating Matrix

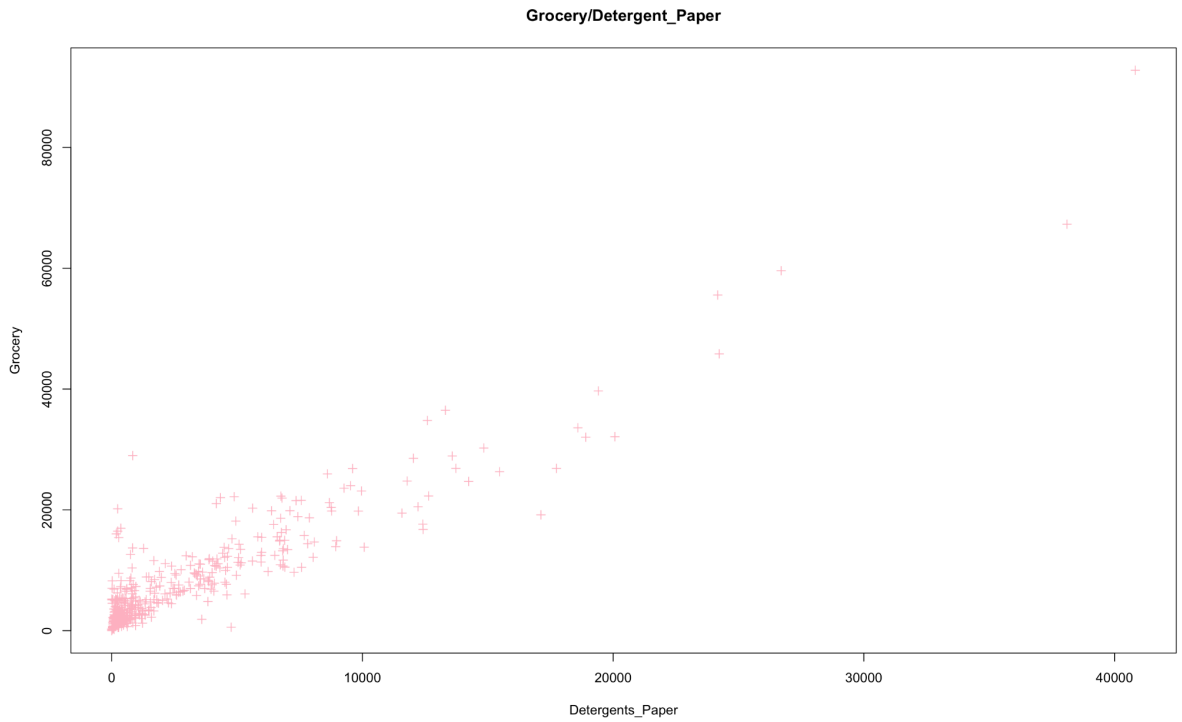
Then we created a 2x3 matrix for Grocery/Milk, Grocery/Detergent_Paper, Grocery_Frozen, Grocery/Delicassen and Grocery/Fresh

## 5- Plotting Grocery against different variables

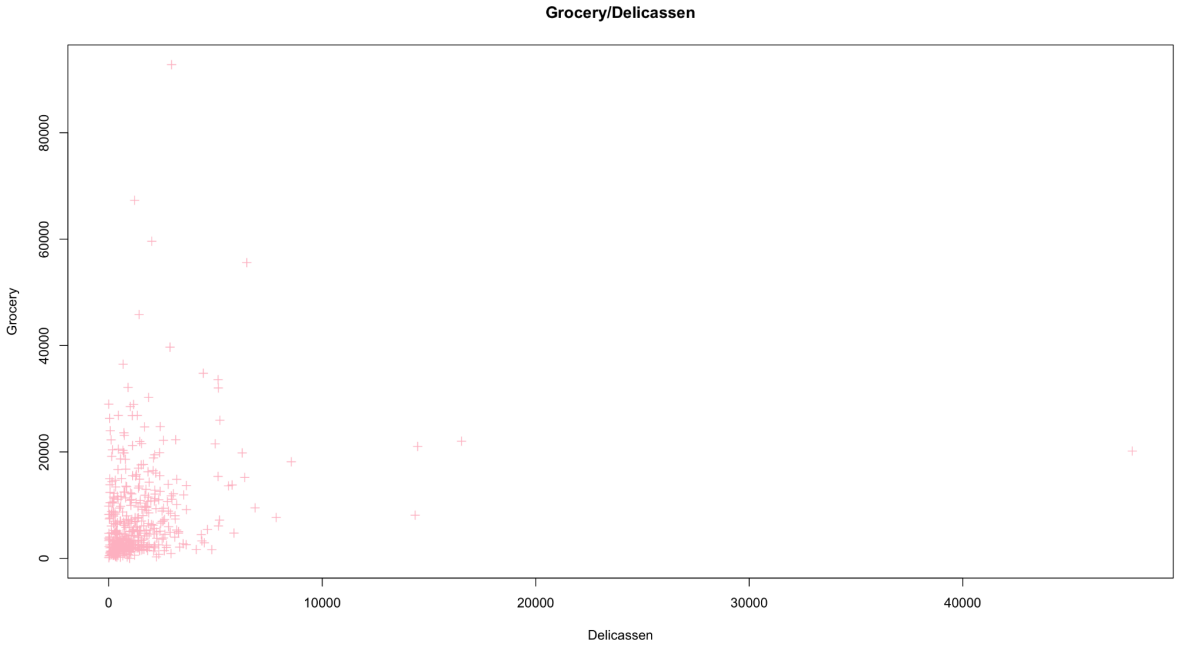
- Grocery against Frozen



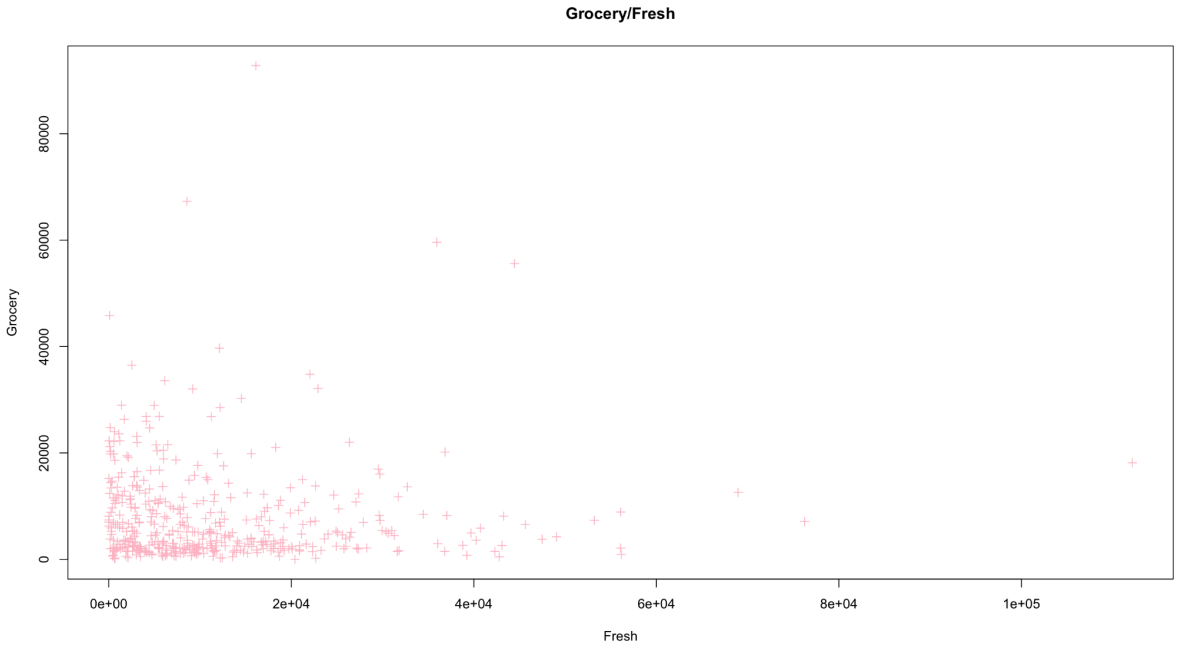

- Grocery against Detergents_Paper



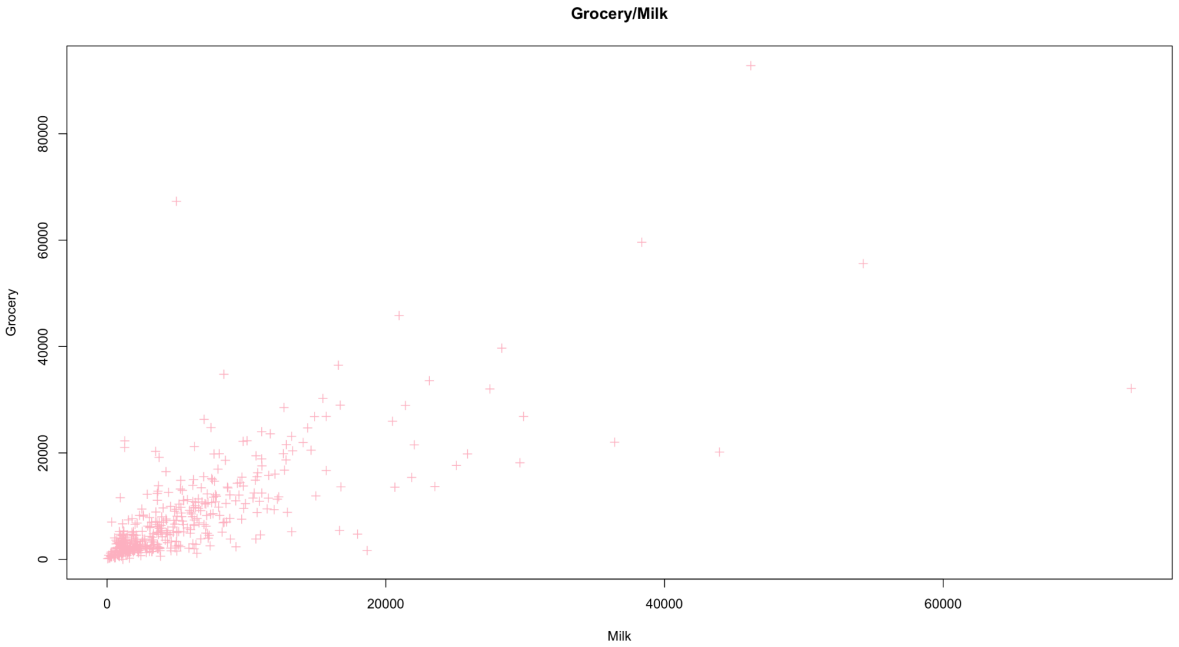
- Grocery against Delicassen



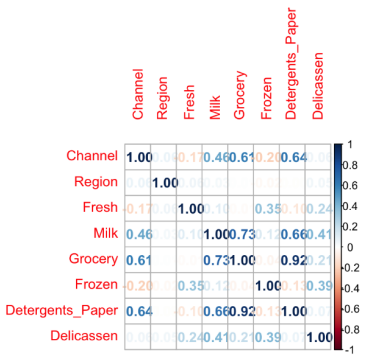

- Grocery against Fresh



- Grocery against Milk



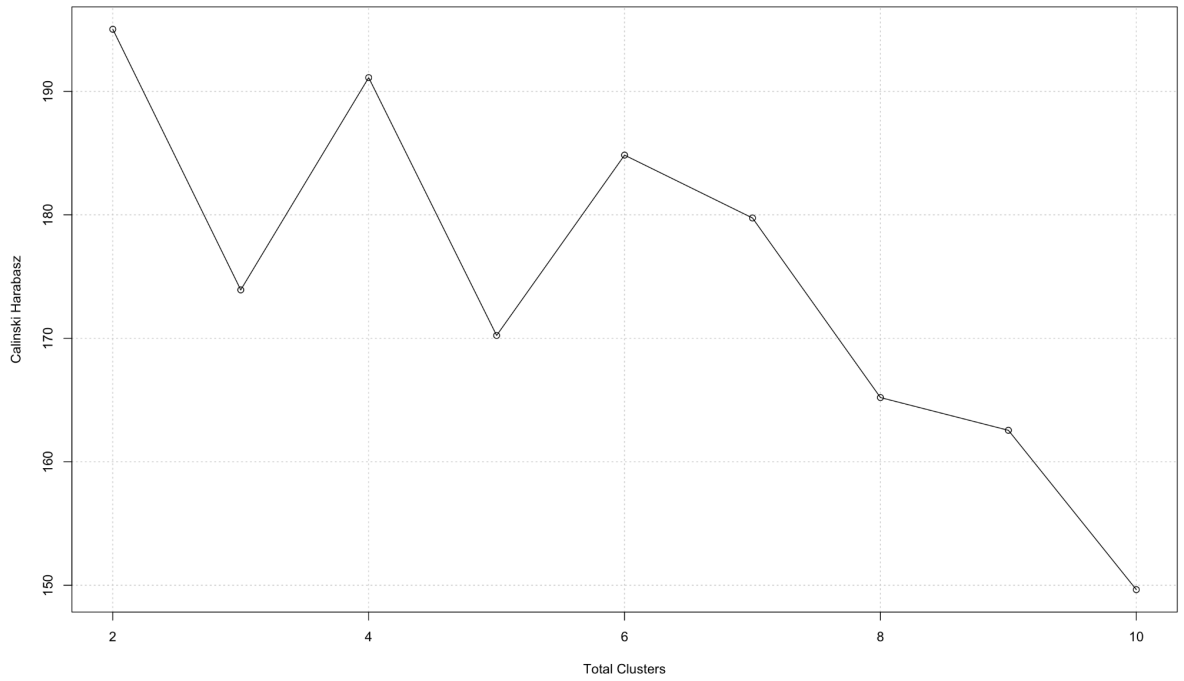

## 6- Determined the correlation of my_data and plotted correlation



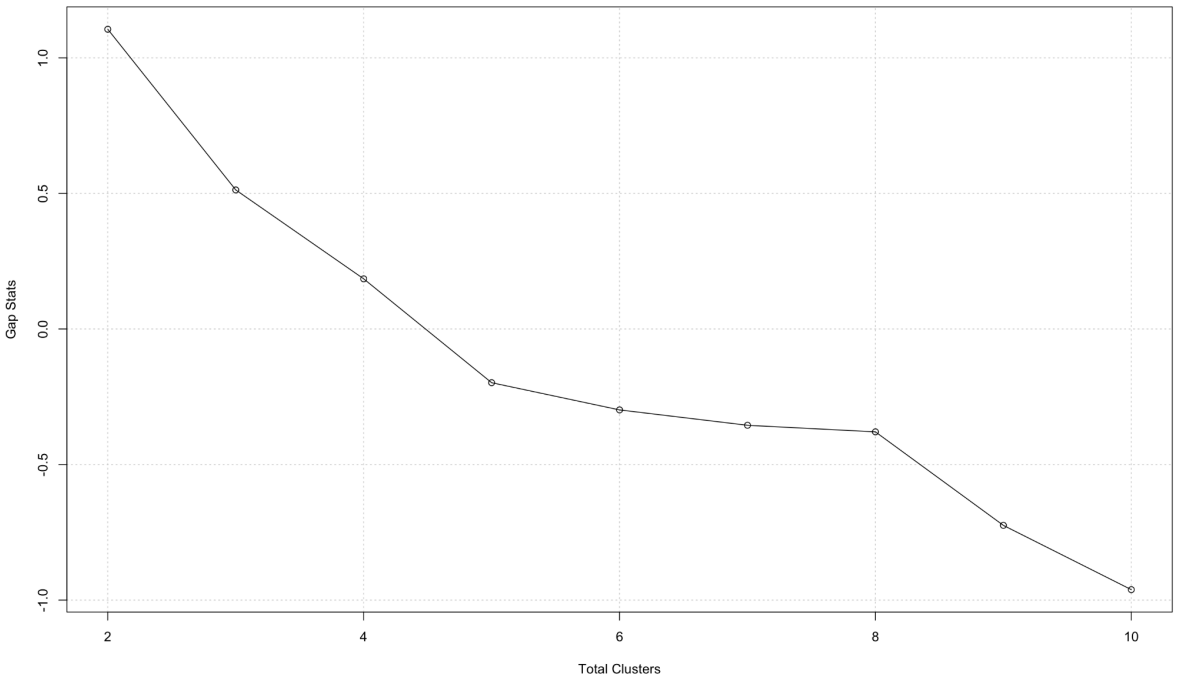
## 7- Determining total clusters

Then we were able to determine the total number of clusters. We used multiple methods to determine the total number of clusters such as the Elbow method, Silhouette method, Gap method, Calinski Harabasz method.

Plotting Calinski Harabasz Method



Plotting Gap Method




## 8- Clustering and Cluster Membership

After all the plottings, we started clustering the data sets, hence we added cluster membership

## 9- Determining total observations and Average mean

We were able to determine the total number of observations in each cluster and figured out avg mean of every variable cluster

## 10- Creating new variable

Then we added cluster membership to all un-normalised datasets
Variable my_new_data has values 1 2 and 394 46

## 11- Plotting

Then we plotted, Cluster against Fresh, Cluster against Milk, Cluster against Grocery, Cluster against Detergents_Paper, Cluster against Frozen and Cluster against Delicassen. We also plotted Income/Saving of detergent/frozen and detergent/milk