PGPBABI Oct. 2019 BATCH

Group 2

Advance Stats- Data Analysis

Preparers

|  |  |  |
| --- | --- | --- |
| Date of Preparation | Brief Description of Assignment | Prepared By |
| 10thFeb, 2020 | Statistical analysis on the Cereal Data set, Leslie Salt Data set and All Greens Franchise Data set. | * Avni Tandon * Bhavyadeep Garg * Madhuchandan * Nishitha Ramesh * Rohit Taunk |

Contents

[**Introduction** 3](#_Toc32265085)

[**Case 1: Cereal Data Set** 3](#_Toc32265086)

[Functionality & Solution 3](#_Toc32265087)

[Data Summary 4](#_Toc32265088)

[Univariate & Bivariate Analysis 5](#_Toc32265089)

[Factor Analysis- Eigen Values 6](#_Toc32265090)

[Factor Analysis- Scree Plot 6](#_Toc32265091)

[Factor Analysis - Output 7](#_Toc32265092)

[5Factors 7](#_Toc32265093)

[4 Factors 8](#_Toc32265094)

[Naming the Factors 9](#_Toc32265095)

[Case 2: Leslie Salt Data Set 9](#_Toc32265096)

[Functionality & Solution 10](#_Toc32265097)

[Data Structure 10](#_Toc32265098)

[Data Summary 11](#_Toc32265099)

[Boxplot 12](#_Toc32265100)

[Scatter Plot 12](#_Toc32265101)

[Correlation Matrix 13](#_Toc32265102)

[Multiple Linear Regression - Full Model 14](#_Toc32265103)

[Multiple Linear Regression – Full Model Summary 15](#_Toc32265104)

[Multiple Linear Regression – Model 2 Summary 15](#_Toc32265105)

[Multiple Linear Regression – Model 3 Summary 16](#_Toc32265106)

[Residual Vs Fitted 16](#_Toc32265107)

[Case 3: All Greens Franchise 18](#_Toc32265108)

[Functionality & Solution 19](#_Toc32265109)

[Data Structure 19](#_Toc32265110)

[Data Summary 19](#_Toc32265111)

[Univariate and Bivariate Analysis 19](#_Toc32265112)

[Box Plots 20](#_Toc32265113)

[Scatter Plots 20](#_Toc32265114)

[Correlation Matrix 22](#_Toc32265115)

[Multiple Linear Regression 23](#_Toc32265116)

[Multiple Linear Regression – Model Summary 23](#_Toc32265117)

# **Introduction**

This document gives a brief description and solution of the problems based on the Exploratory Data Analysis, Factor Analysis, Missing Value Treatment, Correlation Matrix, Linear Regression Analysis to the given data sets of Cereal Data, Leslie Salt Data and All Greens Franchise Data.

Each Case consists of unique techniques considered to derive at the conclusion.

# **Case 1: Cereal Data Set**

As part of a study of consumer consideration of ready-to-eat cereals sponsored by Kellogg Australia, Roberts and Lattin (1991) surveyed consumers regarding their perceptions of their favorite brands of cereals. Each respondent was asked to evaluate three preferred brands on each of 25 different attributes. Respondents used a five-point Likert scale to indicate the extent to which each brand possessed the given attribute.

For the purpose of this assignment, a subset of the data collected by Roberts and Lattin, reflecting the evaluations of the 12 most frequently cited cereal brands in the sample (in the original study, a total of 40 different brands were evaluated by 121 respondents, but the majority of brands were rated by only a small number of consumers).

Requirement to analyze:

1. EDA a) Basic data summary, Univariate, Bivariate analysis, graphs
2. PCA/FA a) Perform PCA/FA and Interpret the Eigen Values (apply Kaiser Normalization Rule)
3. PCA/FA b) Output Interpretation Tell which all factors need to be shortlisted along with their importance and which ones needs to ignored. Name the factors with correct explanations.

## Functionality & Solution

In order to derive and analyze the data set, we have used R and imported the data using the following function:

*cerealdata = read.csv(file.choose())*

* The file used for the data analysis consists of certain variables where the rating for certain variables was given as 6 while the highest ranking as per five-point Likert scale should not exceed 5.
* We identified a total of 7 such entries and to normalize the data, we have replaced the same with pointer 5 before analyzing the data.

**Problem 1.a) Basic data summary, Univariate, Bivariate analysis, graphs**

## Data Summary

The following function was used to derive the summary of all the attributes of Cereal rated by the customers. The summary function briefs the data as per below R screenshot.

*summary(cereal)*

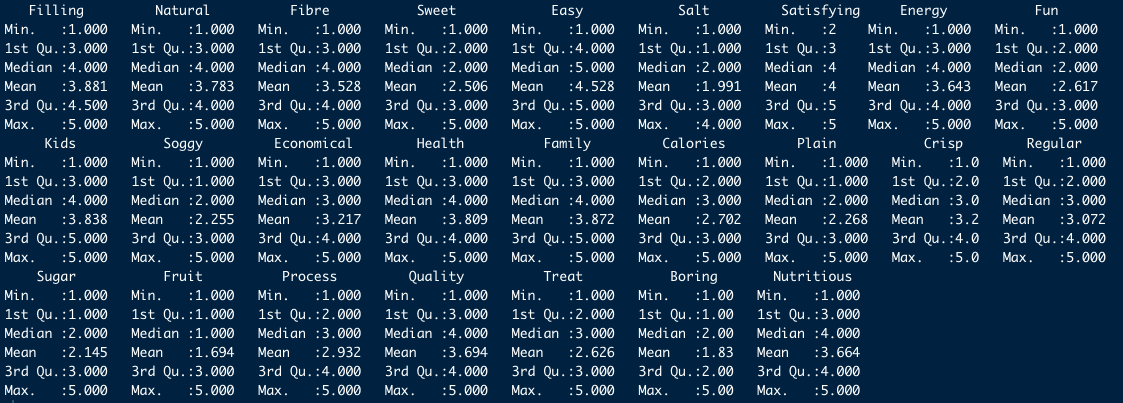


Figure 1: Cereal Data Summary

*str(cereal)*

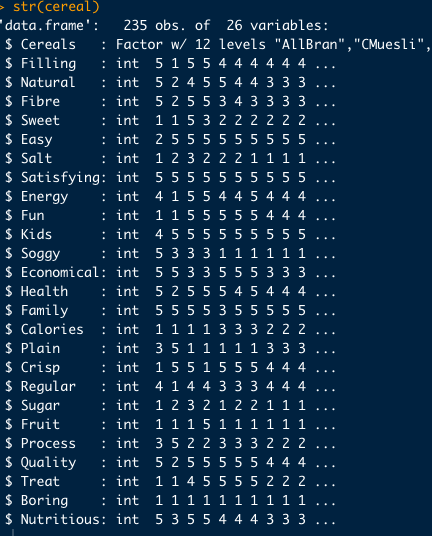


Figure 2: Cereal Data Structure

## Univariate & Bivariate Analysis

* We have plotted box plot to perform univariate analysis as all of the variables are on the same scale, we have plotted together.
* With the help of Box Plot, we could get a clarity on the association of variables. Such as ***Energy, Health, Quality & Nutrition*** shows the same pattern of responses from all the customers which emphasizes that these are the main point of preferences by the customer while choosing a cereal brand.
* Other similar patterns observed are: a) Crisp, Regular & Process b) Soggy, Plain & Sugar.

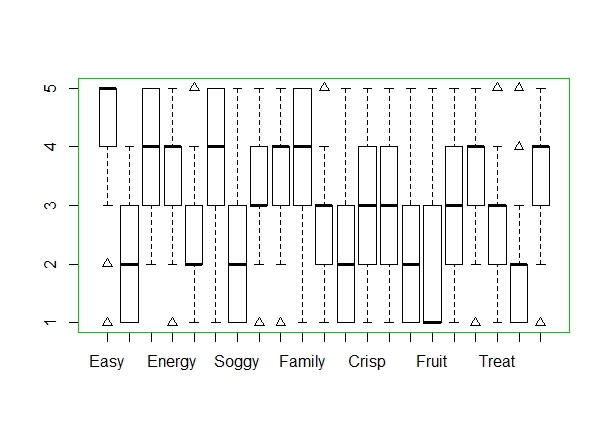


Figure 3: Box Plot- Cereal Attributes

* We have drawn correlation diagram to perform Bivariate analysis which visualizes relation among variables. Positive correlations are displayed in Red and negatives are in Blue. Color intensity is proportional to correlation coefficients.

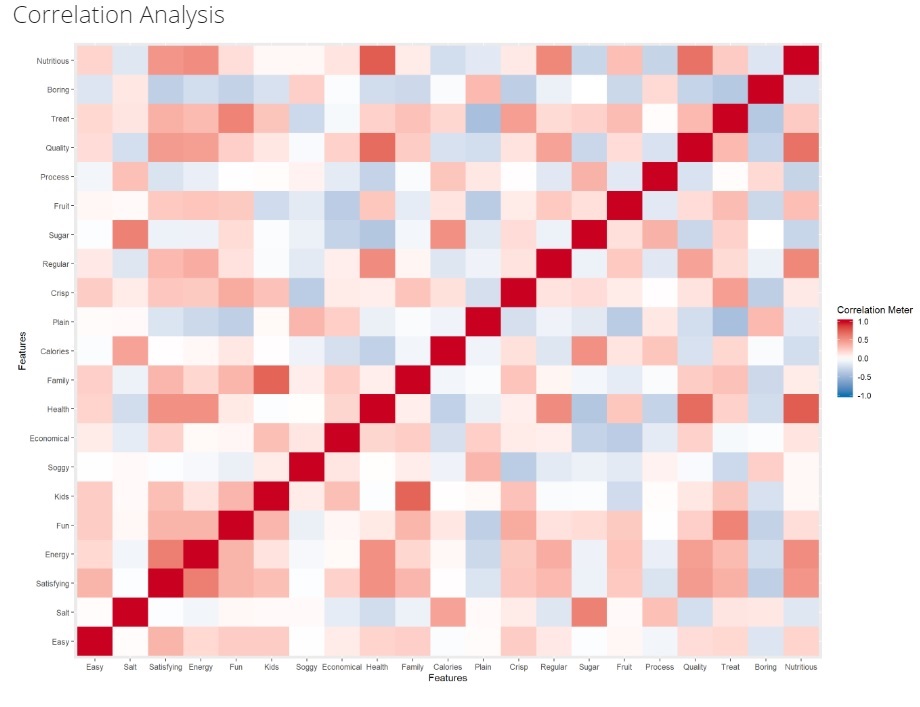


Figure 4 : Correlation Diagram

In above correlation diagram, we can see variable **Health** and **Nutritious** are strongly correlated, **Quality** and **Health** are strongly correlated, **Kids** and **Family** are strongly correlated.

**Problem 1.b) Perform PCA/FA and Interpret the Eigen Values (apply Kaiser Normalization Rule)**

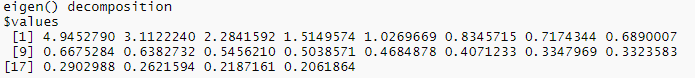
## Factor Analysis- Eigen Values

* Factor analysis is a technique used for dimension reduction where large number of variables are reduced to fewer number of factors. This method extracts maximum common variance from all the variables and generate a common score.
* Factor analysis is mostly used where there are hidden latent variables in dataset which cannot be observed directly but are reflected in the answers or variables of the data.
* We have loaded required packages to carry out further analysis. These packages are `psych` and `GPArotation`.
* Next, we will find out the number of factors that we’ll be selecting for factor analysis. This can be evaluated via methods such as `Parallel Analysis` and `eigenvalue`, etc. As the problem asks for the eigen value, we have plotted scree plot on eigen values.
* First & foremost we will calculate the eigenvalues for the matrix to get a better understanding of the linear transformation. Eigenvalues calculate the variance between and within the variables.
* The following function to be used to arrive at eigenvalues:

*ev\_cereal <- eigen(cor(cereal[,-1]))*

*ev\_cereal*

* The following values are derived from the eigenvalues function:



## Factor Analysis- Scree Plot

* Scree plot is plotted to show the eigenvalues on the y-axis and the number of factors on the x-axis. It displays a downward curve.
* The point where the slope of the curve is clearly leveling off (the elbow) indicates the number of factors that should be generated by the analysis.

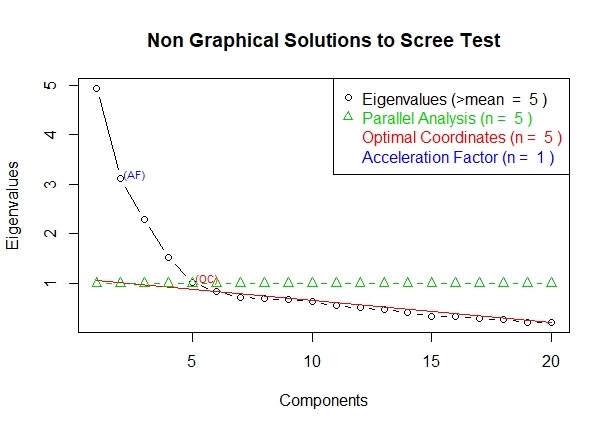


Figure 5: Cereal Data- Scree Plot

* Based on the above scree plot we have observed that it is forming an elbow at 5th factor.
* Therefore, we have considered to check the factor loading results with 5 factors onwards to identify which factor to be considered.

**1.c) Output Interpretation Tell which all factors need to be shortlisted along with their importance and which ones needs to ignored. Name the factors with correct explanations.**

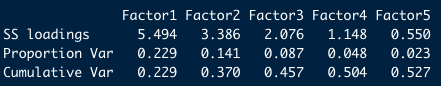
## Factor Analysis - Output

Now that we’ve arrived at probable number of factors, let’s start off with 5 as the number of factors. In order to perform factor analysis, we are using **‘factanal’** function. Given below are the arguments we’ll supply:

* x  – data frame
* nfactors – Number of factors to extract
* rotate – Although there are various types rotations, we are doing it without rotation, so rotation is ‘none’.

### 5Factors

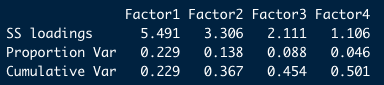
* Below is the section of output which shows SS loadings of factors and percentage of variance covered:



* To find a smaller number of interpretable factors that explain the maximum amount variability in the data we will classify it further into the most appropriate factor where the cut off eigenvalue ≥ 1. Using Kaiser Normalization Rule, we are reconsidering 4 factors.

### 4 Factors

* The output after using number of factors as 4 is as follows:



Also, we need to consider the loadings more than 0.4 and not loading on more than one factor. We have used cutoff = 0.4 for loadings to improve the visibility.

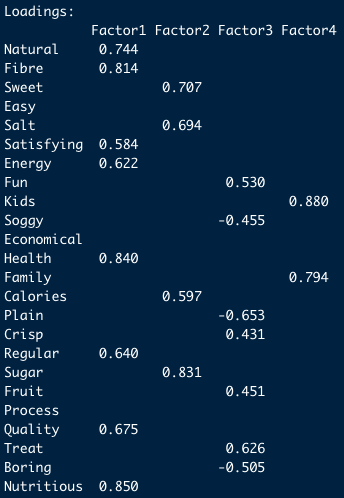


Figure 6: Factor loadings after cut off=0.4

We can see that it results in only single-loading. This is known as **simple structure**.

Now that we’ve achieved simple structure it’s time for us to validate our model. Now let’s check at the adequacy of the factors.

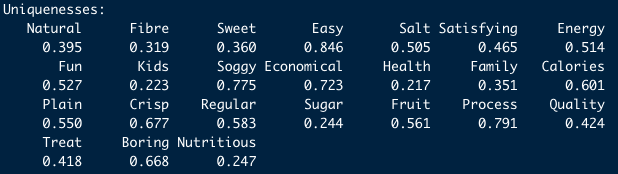


Figure 7: Uniqueness

This section of output provides the **uniqueness,** which ranges from 0 to 1. The uniqueness, sometimes referred to as noise, corresponds to the proportion of variability, which cannot be explained by a linear combination of the factors. A high uniqueness for a variable indicates that the factors do not account well for its variance.

In above screenshot, we can see the variables ‘**Easy’**, ‘**Soggy’**, ‘**Process’** has high uniqueness values in compared to other variables, that means the factors do not account well for its variance.

**Communality** is 1- uniqueness that describe how the larger amount of the variance in the variable extracted by the factor solution. Variable ‘**Kids’, ‘Health’, ‘Sugar’, ‘Nutritious’** have high communality.

## Naming the Factors

After performing the factor analysis with simple structure and looking at the factor loadings, it’s time for us to name the factors. We are forming the factors depending on the variable loadings. In this case, here is how the factors can be created:

|  |  |  |  |
| --- | --- | --- | --- |
| HEALTHY CEREAL | SWEET CEREAL | TREAT YOURSELF CEREAL | FAMILY PACK |
| Fibre | Sweet | Plain | Kids |
| Health | Sugar | Treat | Family |
| Nutritious |  |  |  |

Table 1: Factor Naming

# **Case 2: Leslie Salt Data Set**

In 1968, the city of Mountain View, California, began the necessary legal proceedings to acquire a parcel of land owned by the Leslie Salt Company. The Leslie property contained 246.8 acres and was located right on the San Francisco Bay. The land had been used for salt evaporation and had an elevation of exactly sea level. However, the property was dyked so that the waters from the bay park were kept out. The city of Mountain View intended to fill the property and use it for a city park.

Ultimately, it fell into the courts to determine a fair market value for the property. Appraisers were hired, but what made the processes difficult was that there were few sales of byland property and none of them corresponded exactly to the characteristics of the Leslie property. The experts involved decided to build a regression model to better understand the factors that might influence market valuation. They collected data on 31 byland properties that were sold during the previous 10 years. In addition to the transaction price for each property, they collected data on a large number of other factors, including size, time of sale, elevation, location, and access to sewers. A listing of these data, including only those variables, deemed relevant for this exercise.

Requirement to analyze:

1. Basic data summary, Univariate, Bivariate analysis, graphs
2. Missing Value treatment, Correlation check
3. Perform Multiple Linear Regression with Price as the Dependent Variable and the rest of them as Independent Variables
4. MLR summary interpretation and significance (R, R2, Adjusted R2, Degrees of Freedom, f-statistic, coefficients along with p-values)
5. Output Interpretation <making it meaningful for everybody.

## Functionality & Solution

In order to derive and analyze the data set, we have used R and imported the data using the following function:

*Lesliesalt <- read\_excel (file.choose())*

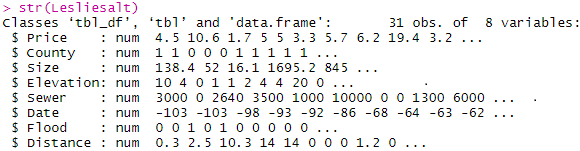
* Price variable is considered as Dependent Variable and rest of the variables are Independent variables.

**Problem 2.a) Basic data summary, Univariate, Bivariate analysis, graphs**

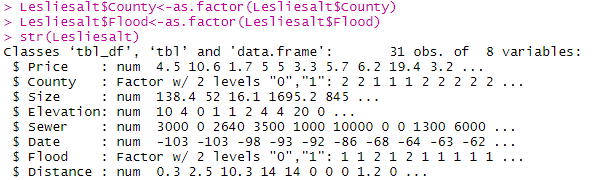
## Data Structure

The following function was used to analyze the structure of the Data Set. Independent variables such as Flood and County are found to be Numeric.

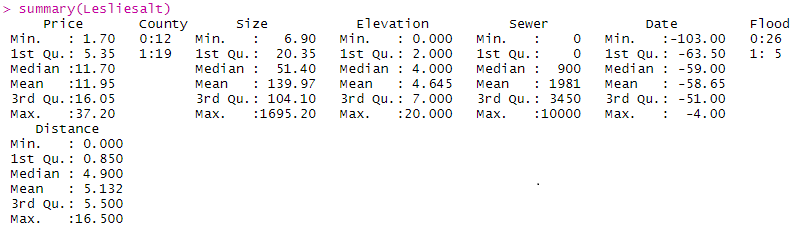
Str(Leslie salt)

**

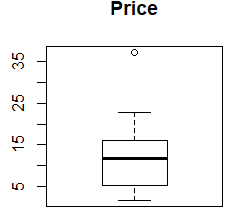
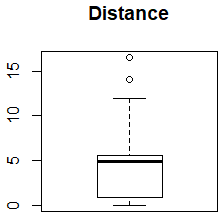
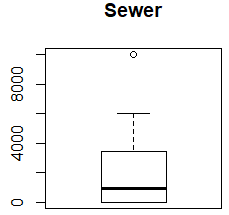
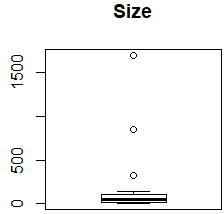
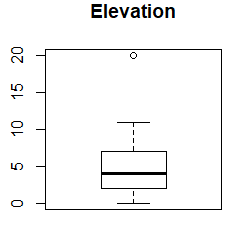
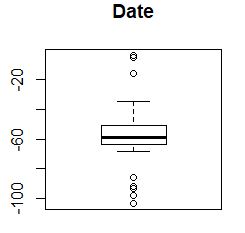
We have transformed the Independent Variables Flood and County from Numeric to Factor Variable before preparing the Model.

**

## Data Summary

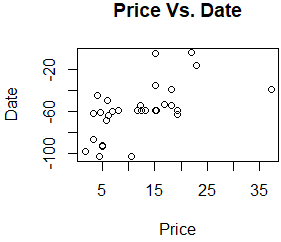
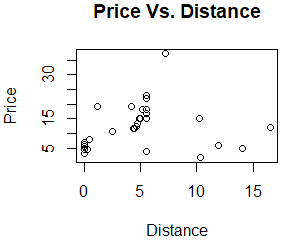
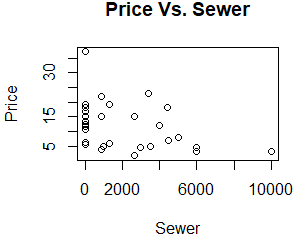
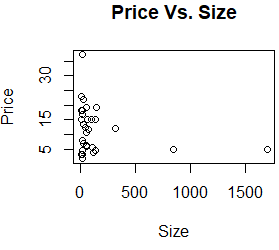
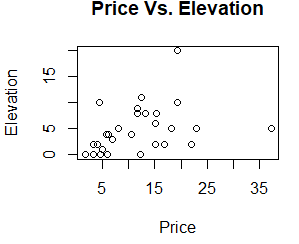


## Boxplot



## Scatter Plot

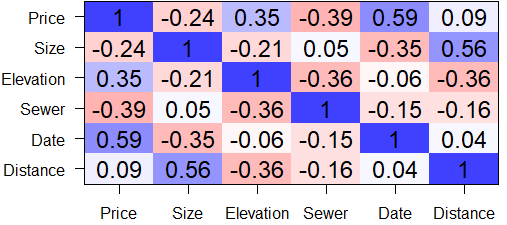
* From the Scatter Plot of “Size Vs. Price” we could observe that as the Size increases, the Price decreases which practically seems to be strange. Therefore, we have considered not to include this variable in the model.
* Scatter Plot of “Distance Vs. Price” doesn't show strong correlation.



**Problem 2.b) Correlation check**

## Correlation Matrix

* Price has positive correlation with Elevation and Date.
* Price has negative correlation with Size and strongly negative correlation with Sewer.
* Price has weak positive correlation with distance.



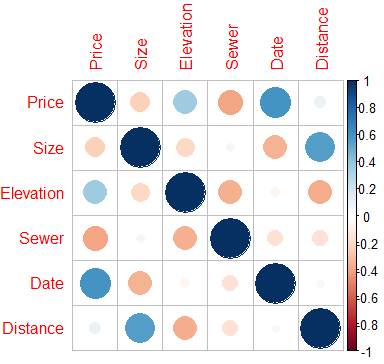


Figure 8: Correlation matrix

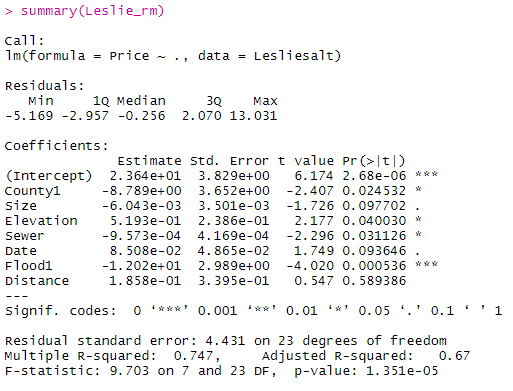
**Problem 2.c) Perform Multiple Linear Regression with Price as the Dependent Variable and the rest of them as Independent Variable**

## Multiple Linear Regression - Full Model

We have created a Multiple Linear Regression Model Considering Price as Dependent Variable and all other variables as Independent using following function.

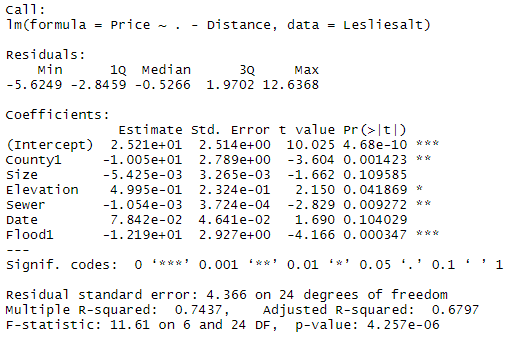
Leslie\_rm <- lm(Price~., data = Lesliesalt)

### Multiple Linear Regression – Full Model Summary



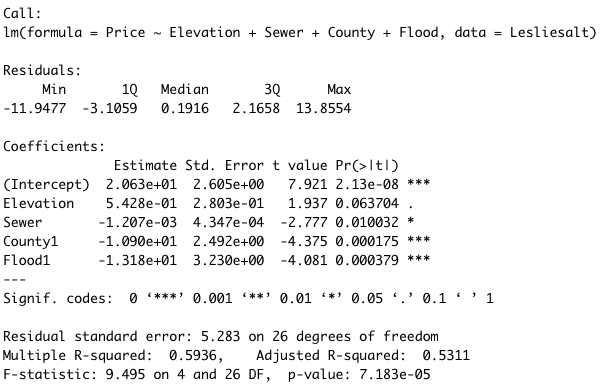
* As p-value of the model is very less, we can conclude that the model is valid.
* P-value obtained for the variable ‘Distance’ is very high. Therefore, we can consider it as insignificant. Hence, we will generate another model with other variables.
* Multiple R- squared Value indicates the current model is able to explain 74.7% of the total variance.

### Multiple Linear Regression – Model 2 Summary



* P-value of the model is less hence the model is valid.
* P-value obtained for variables Size & Date is very high, hence, these variables are not significant. Therefore, we will generate another model excluding Size & Date variables.
* Multiple R-squared value of this model is 74.3%.

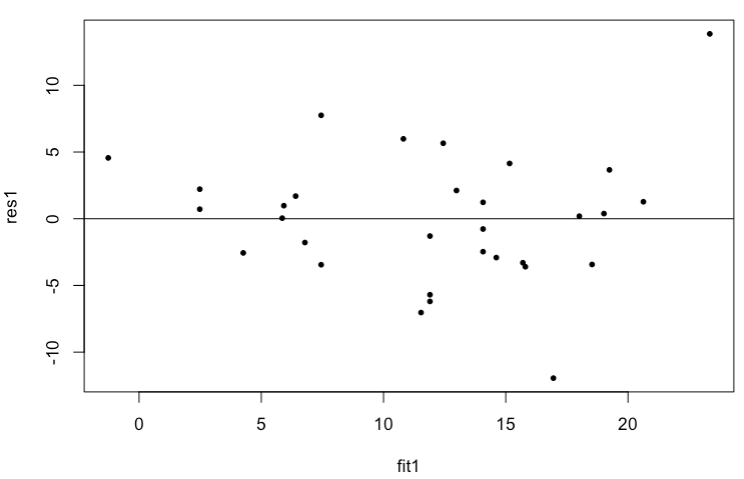
### Multiple Linear Regression – Model 3 Summary



* P-value of the model is less hence the model is valid.
* Multiple R-squared value of this model is 59.36% which is lesser than Model 1 but all the variables are significant.
* The P-value obtained for Model 3 is less than 0.5 which says overall model is significant. Therefore, we will go ahead with Model 3 for Prediction.

### Residual Vs Fitted

As next action, We have extracted the fitted values and residual values from model 2 output, combined with data and plotted a scatterplot.



We can observe there is an outlier which on analysis we found that due to high Price variable which is a dependent variable, hence, we decided to keep in the data.

**Problem 2.d) MLR summary interpretation and significance (R, R2, Adjusted R2, Degrees of Freedom, f-statistic, coefficients along with p-values)**

We have considered **Model 2 for Summary** Interpretation and significance.

**R**

* Residuals are essentially the difference between the actual observed response values and the response values that the model predicted. In Model 3, the distribution of Residuals appears to be fairly symmetrical.

**R2**

* The R-squared statistic provides a measure of how well the model is fitting the actual data.
* In full model, the value of R2 is 0.7437 which indicates that 74.37% of the variance in Response Variable (Price) is explained by Predictor Variables.

**Adjusted R2**

* The Adjusted R2 adjusts for the number of variables in the model. The Adjusted R2 value actually decreases when variables don’t improve the model fit.
* In **Model 2** Adjusted R2 value is obtained as 0.6797 where as for Model 3 Adjusted R2 value is 0.5311, which indicates variables in Model 2 has improved model fit when compared to Model 3.

**Degrees of Freedom**

* Degree of freedom define the amount of information available relative to the number of properties to be estimated.
* **Model 2** has 24 error degrees of freedom, which indicates that 24 independent pieces of information are available for estimating the coefficient, while Model 3 has 26 degrees of freedom.

**F-statistic**

* F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. Higher the value gives more reasons to reject null hypothesis which says overall model is significant. As Model 2 has f-stats value of 11.61 but Model 3 has 9.49 f-stats value.
* In Model 1 we have 24 Degrees of freedom**.**

**Coefficients along with P-values**

* The coefficients describe the mathematical relationship between each [independent variable](https://statisticsbyjim.com/glossary/predictor-variables/) and the [dependent variable](https://statisticsbyjim.com/glossary/response-variables/). The p-values for the coefficients indicate whether these relationships are statistically significant. Lower the p-value allows us to reject null hypothesis and also the significance of the particular variable in that model. In model 3, we could see, all the variables are less than or closer to 0.05.

**Problem 2.e) Output Interpretation**

* P-value of the model is less hence the model is valid.
* P-value obtained for variables all the variables is less than 0.05, hence, all the variables used for this model are significant.
* R-square Value indicates the current model is able to explain 59.36% of the total variance.
* Adjusted R square value for model 2 is obtained as 67.97% indicates model 2 has improved fit.
* High F-stats value shows that the overall model is significant.

# Case 3: All Greens Franchise

Explain the importance of X2, X3, X4, X5, X6 on Annual Net Sales, X1. The data (X1, X2, X3, X4, X5, X6) are for each franchise store.

X1 = annual net sales/$1000 X2 = number sq. ft./1000 X3 = inventory/$1000

X4 = amount spent on advertising/$1000 X5 = size of sales district/1000 families

X6 = number of competing stores in district

The final deliverable should include:

1. Detailed Exploratory Data Analysis I.e. Basic data summary, Univariate, Bivariate analysis, graphs
2. Correlation check, explanations of the relationships discovered, checking for linear relationship using Regression

## Functionality & Solution

In order to derive and analyze the data set, we have used R and imported the data using the following function:

*Franchise <- read\_excel (file.choose())*

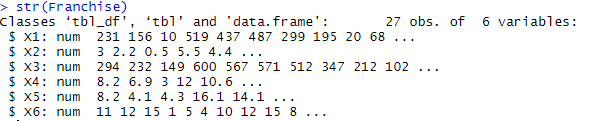
* The file used for the data analysis consists of variables where the variable X1 is the net sales generated by a franchise store and all the other variables are the factors which are affecting the net sales.

Problem 3.a) Detailed Exploratory Data Analysis I.e. Basic data summary, Univariate, Bivariate analysis, graphs

## Data Structure

The following function was used to analyse the structure of the Data Set. All the Independent variables are found to be Numeric.

Str(Franchise)

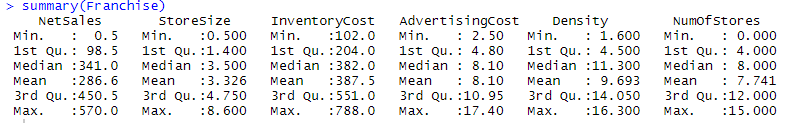


## Data Summary

For ease of understanding, we renamed the variables.

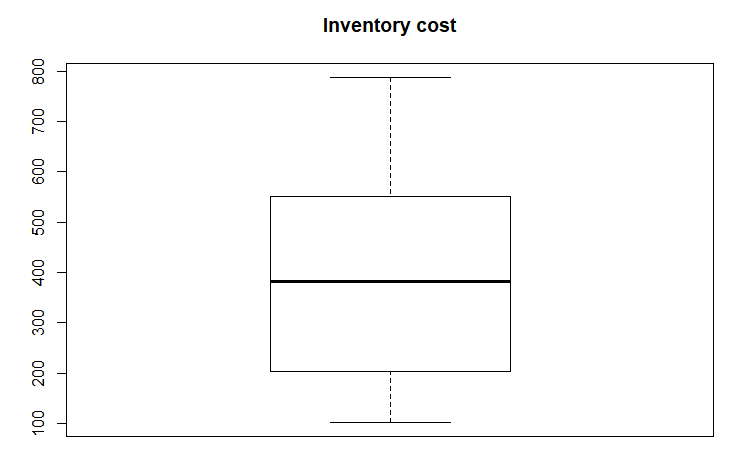
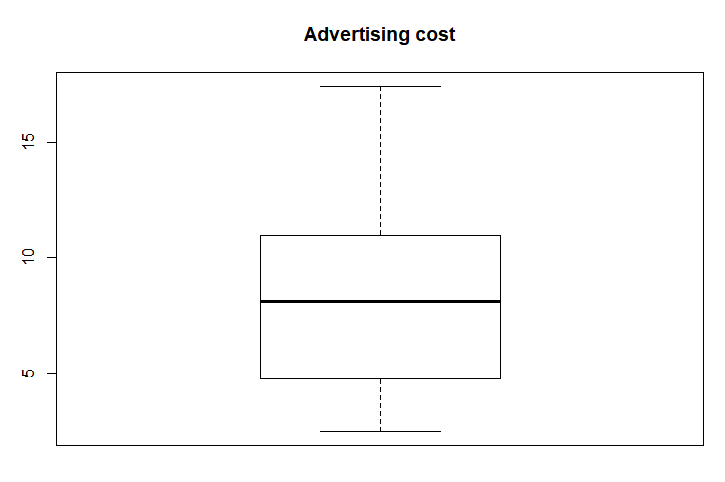
X1 – Net Sales, X2 – Store Size, X3-Inventory Cost, X4 – Advertising Cost, X5 – Density, X6 – No. of Stores

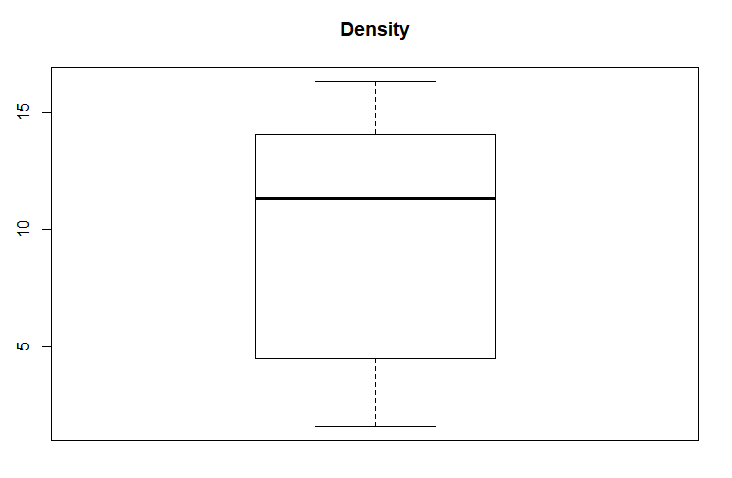
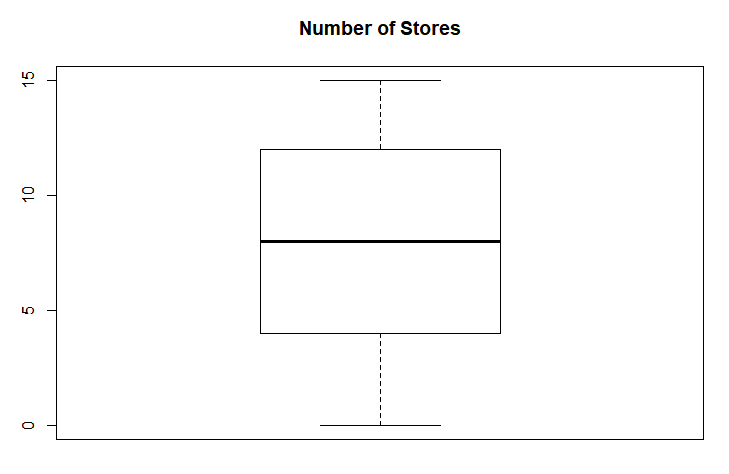
The following function was used to derive the summary of all the attributes of All Green Franchise attributes.

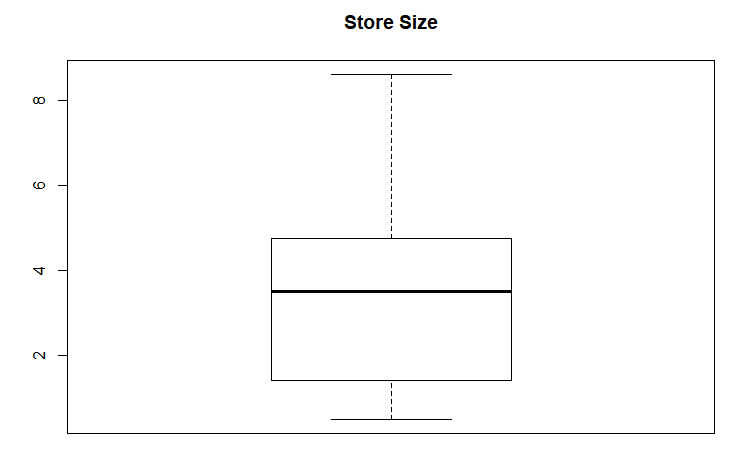
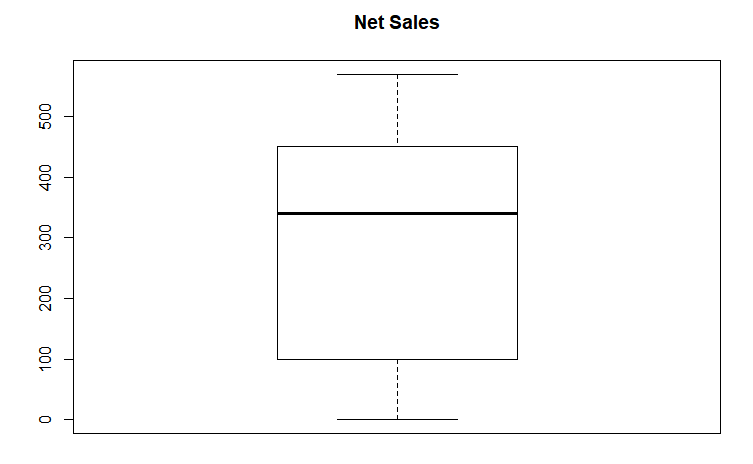


## Univariate and Bivariate Analysis

### Box Plots

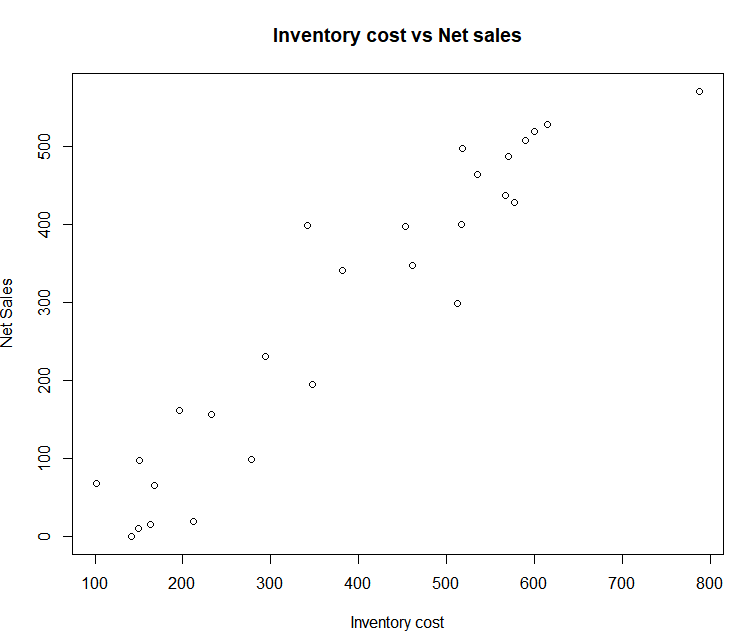
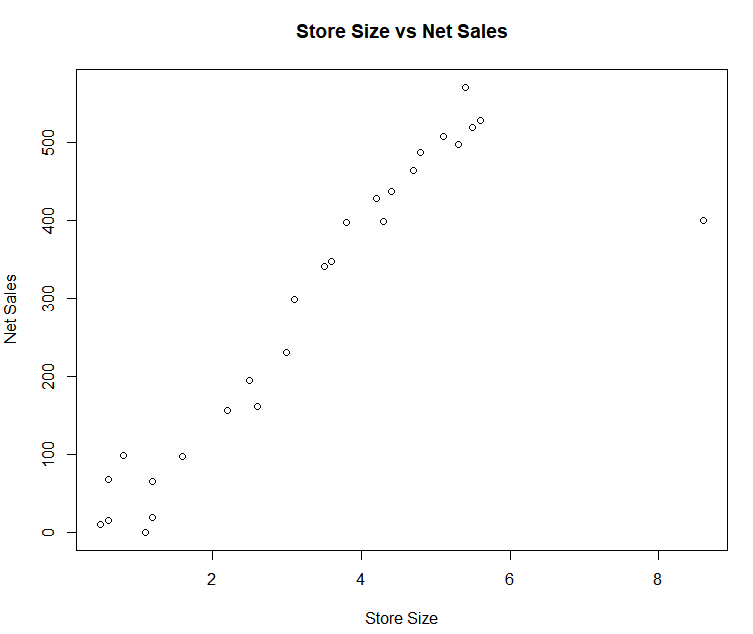
 

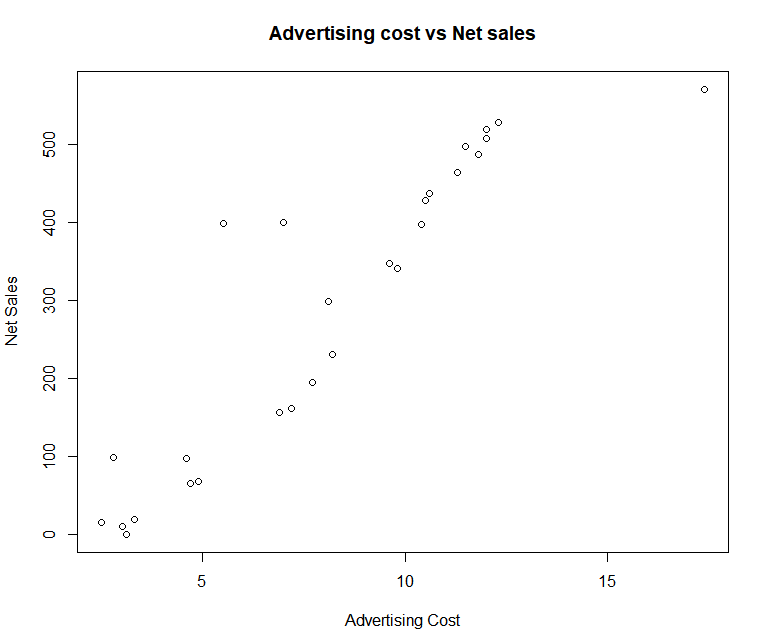
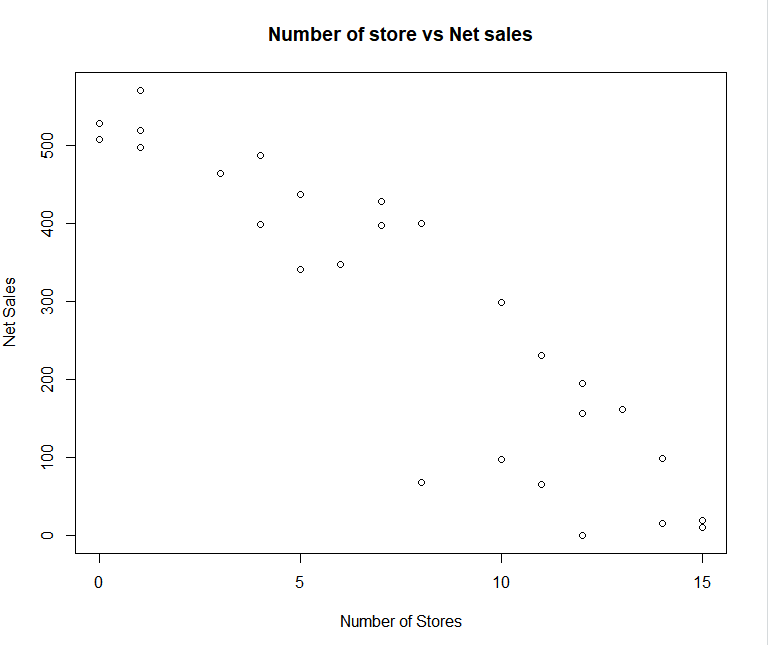


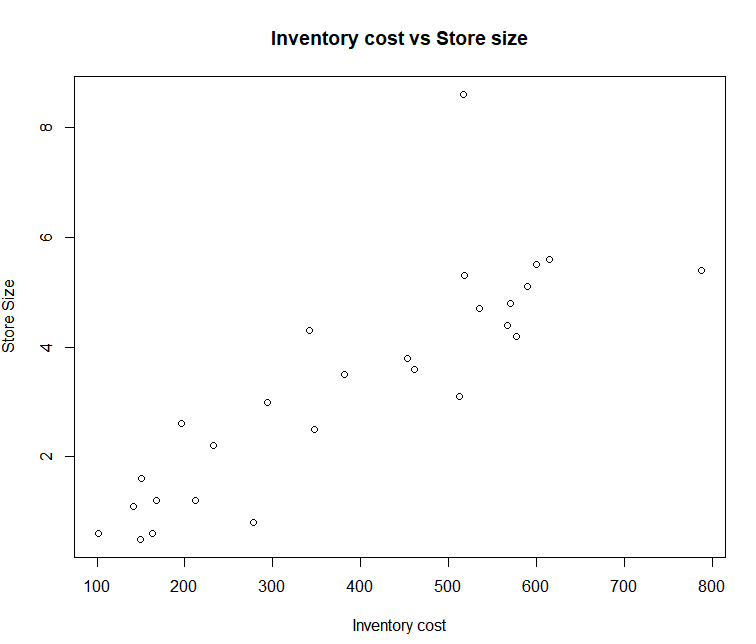
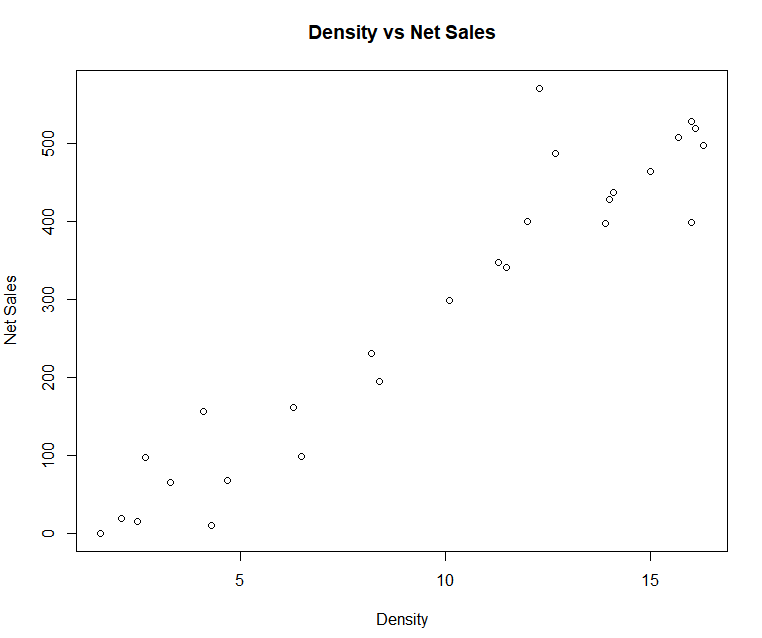
### Scatter Plots

The scatter plot shows the below results:

1. ***Store Size Vs Net Sales:*** Shows a positive co-relation such as larger the store size more is the sales
2. ***Inventory cost Vs. Net Sales:*** Shows a positive co-relation such as more the inventory cost involved more the sales.
3. ***Advertising Cost Vs. Net Sales:*** Shows a positive co-relation such as more the amount spent on advertising more it enhances the sales.
4. ***Number of Store Vs. Net Sales:*** Shows a negative co-relation such as more the number of stores less is the net sales. It may happen that there are a lot of stores opened in an area where there is less density.
5. ***Density Vs. Net Sales:*** Shows a positive co-relation such as more the population of the area more is the sales.
6. ***Inventory Cost Vs. Store Size:*** Shows a positive co-relation such as more the Inventory Cost involved the bigger is the store size required.
7. ***Number of Stores Vs. Density:*** Shows a negative co-relation such as larger number of stores in an area is not affected by the density of the area.



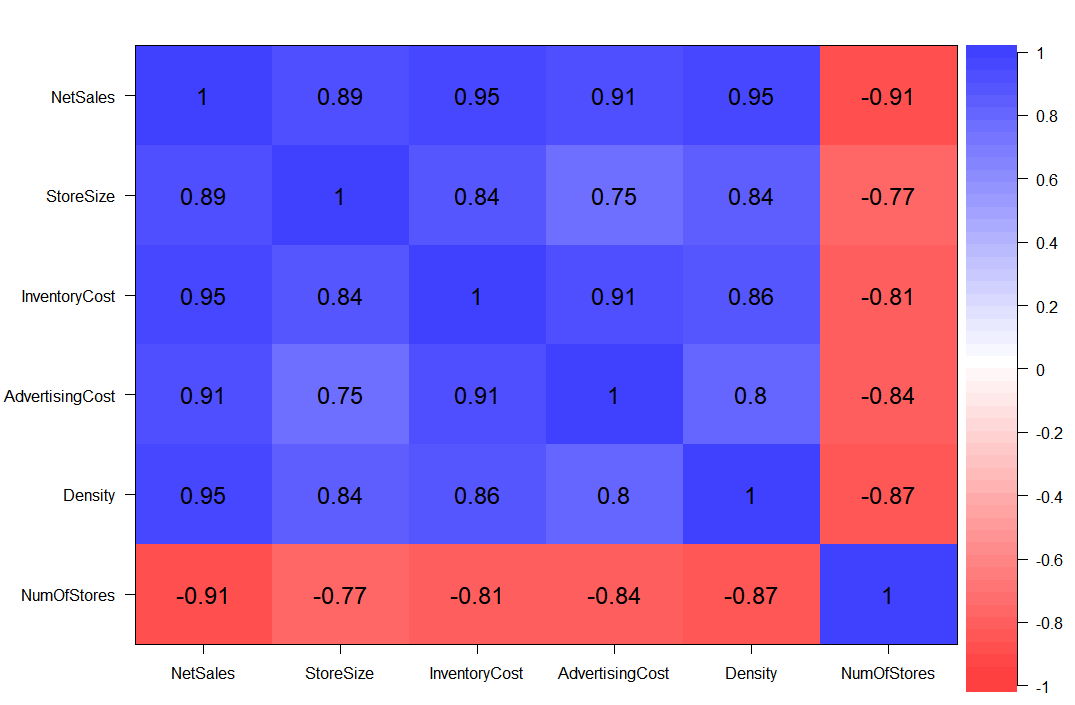




Problem 3.b) Correlation check, explanations of the relationships discovered, checking for linear relationship using Regression

## Correlation Matrix

* From the below Correlation plot we can observe that Net Sales is inversely co-related to Number of Stores.
* Net Sales has positive correlation with Store Size, Inventory Cost, Advertising Cost and Density.

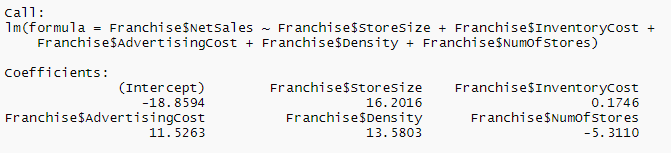


## Multiple Linear Regression

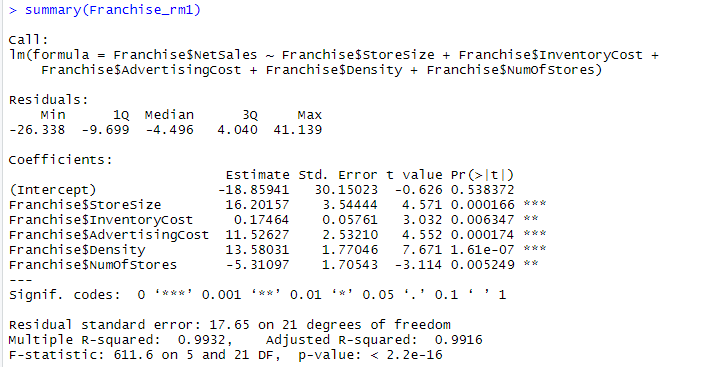
We have created a Multiple Linear Regression Model Considering Net Sales as Dependent Variable and Store Size, Inventory Cost, Advertising Cost, Density, Number of Stores as Independent Variables. The below function is used:

Franchise\_rm1<-lm(Franchise$NetSales~Franchise$StoreSize+Franchise$InventoryCost+Franchise$AdvertisingCost+Franchise$Density+Franchise$NumOfStores)

*Output:*



### Multiple Linear Regression – Model Summary



**R**

* In above model, the distribution of Residuals appears to be fairly symmetrical skewed toward positive.

**R2**

* The value of R2 is 0.99 which indicates that 99% of the variance in Response Variable (Net Sales) is explained by Predictor Variables.

**Adjusted R2**

* The Adjusted R2 value of 0.9916 is almost same as multiple R-square as all the variables are significant.

**Degrees of Freedom**

* Above Model has 21 error degrees of freedom, which indicates that 21 independent pieces of information are available for estimating the coefficient.

**F-statistic**

* Model has f-stats value of 611.6 which is high that shows the significance of overall model.

**Coefficients along with P-values**

* In above model, we could see, all the variables are less than 0.05, all the variables are significant to the model.
* Store Size, Advertising Cost & Density are highly significant variables since the PR value is close to 0.