EAS506 - Statistical Data Mining I

Homework 1 – Question 1

Aniket Kamlender Maheshwari

09/16/21

Abstract

This report summarizes the process undertaken to clean, transform and perform Exploratory Data Analysis (EDA) on cereal dataset to get a clean data for fitting a Linear Regression Model.

Initial Data set was:
“cereal.csv”
77 rows * 16 columns

The final cleaned data set was saved as:
“cleaned_cereal_data.RData”
73 rows * 14 columns

Content

1 Introduction

The Cereal data frame has 77 rows and 16 columns. The data come from the 1993 ASA Statistical Graphics Exposition and are taken from the mandatory F&DA food label. The data have been normalized here to a portion of one American cup.

This report summarizes the process undertaken to clean, transform and perform Exploratory Data Analysis (EDA) on cereal dataset to get a clean data for fitting a Linear Regression Model.

2 Method

2.1 Initialization Steps

- Clear the memory
- Install and load all required libraries.
- Briefly examine the data with functions like dim, str , summary.

2.2 List of all Feature Information.

- Examining and create a list of all features details.

2.2.1 Feature Details.

1. mfr : categorical feature, has 7 categories. One thing to note about this feature is that 'A' has only one value. This may be a possible outlier.
2. type : categorical feature, has 2 categories but 'H' has only 3 records, we might as well remove these records because they won't be helpful.
3. calories : integer feature. will plot histogram and box plot to look for potential outliers.
4. protein : Discrete variable. Most proportion of values are at protein level below 4. '5' and '6' protein level has only 3 points.
5. Fat : Discrete variable.
6. sodium : continuous integer variable.
7. fiber : Discrete variable.
8. Carbo : continuous numeric variable.
9. sugars : Discrete variable. One interesting thing to note in this is one data point has sugar value '-1'
10. potass : continuous integer variable.
11. vitamins : Discrete variable. has three output '0' , '25' and '100'.
12 - 14 : shelf , weight , cups : i don't really think these feature impact our target feature ('rating') much but I will plot a correlation plot first before removing these features.
15 : rating : our target feature. will plot a histogram to see if there is any potential outlier in this feature.

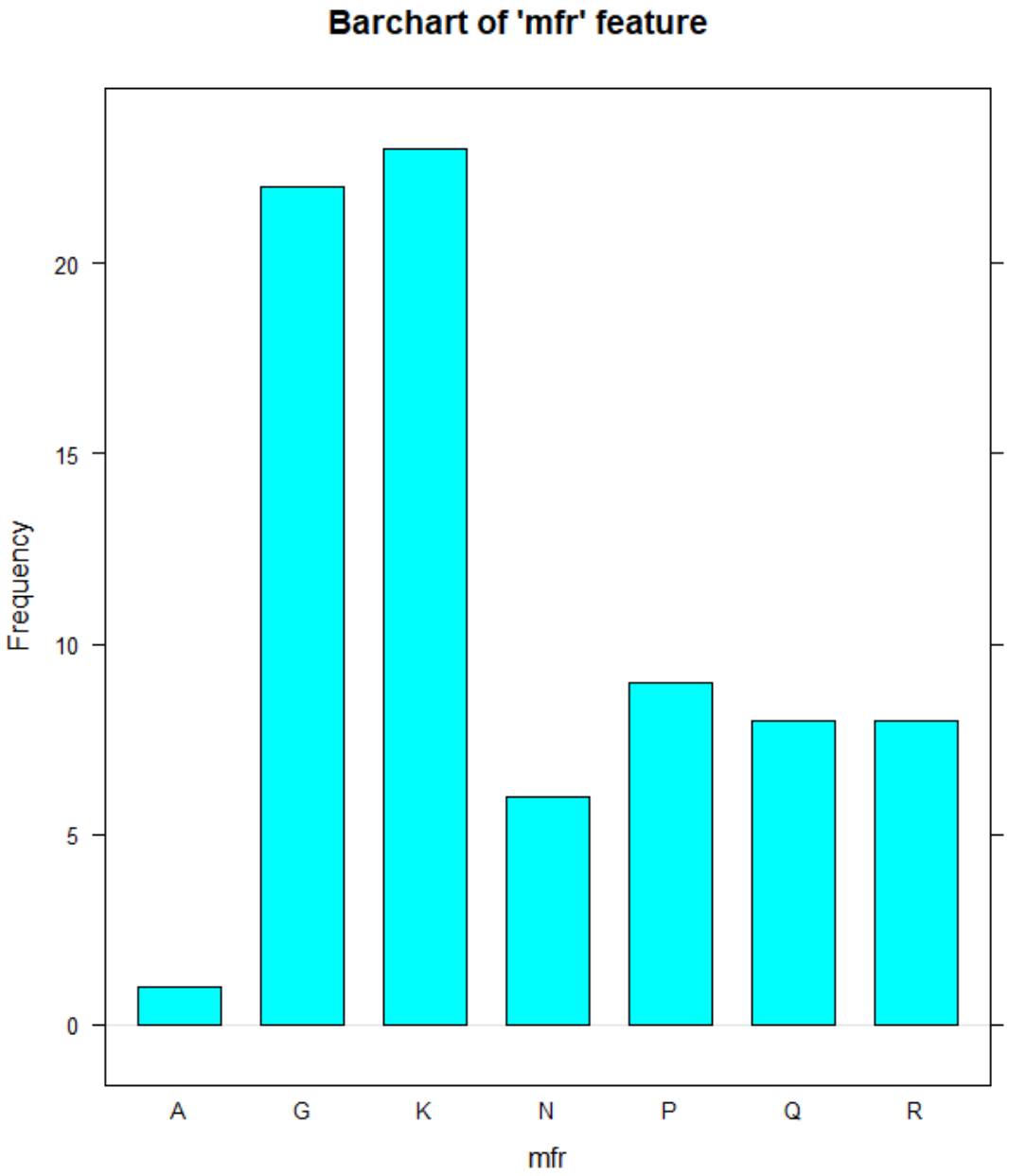
2.3 Finding Missing Values in the Dataset

- Examined if there are any missing values in the dataset.

2.4 Univariate Analysis on each Feature.

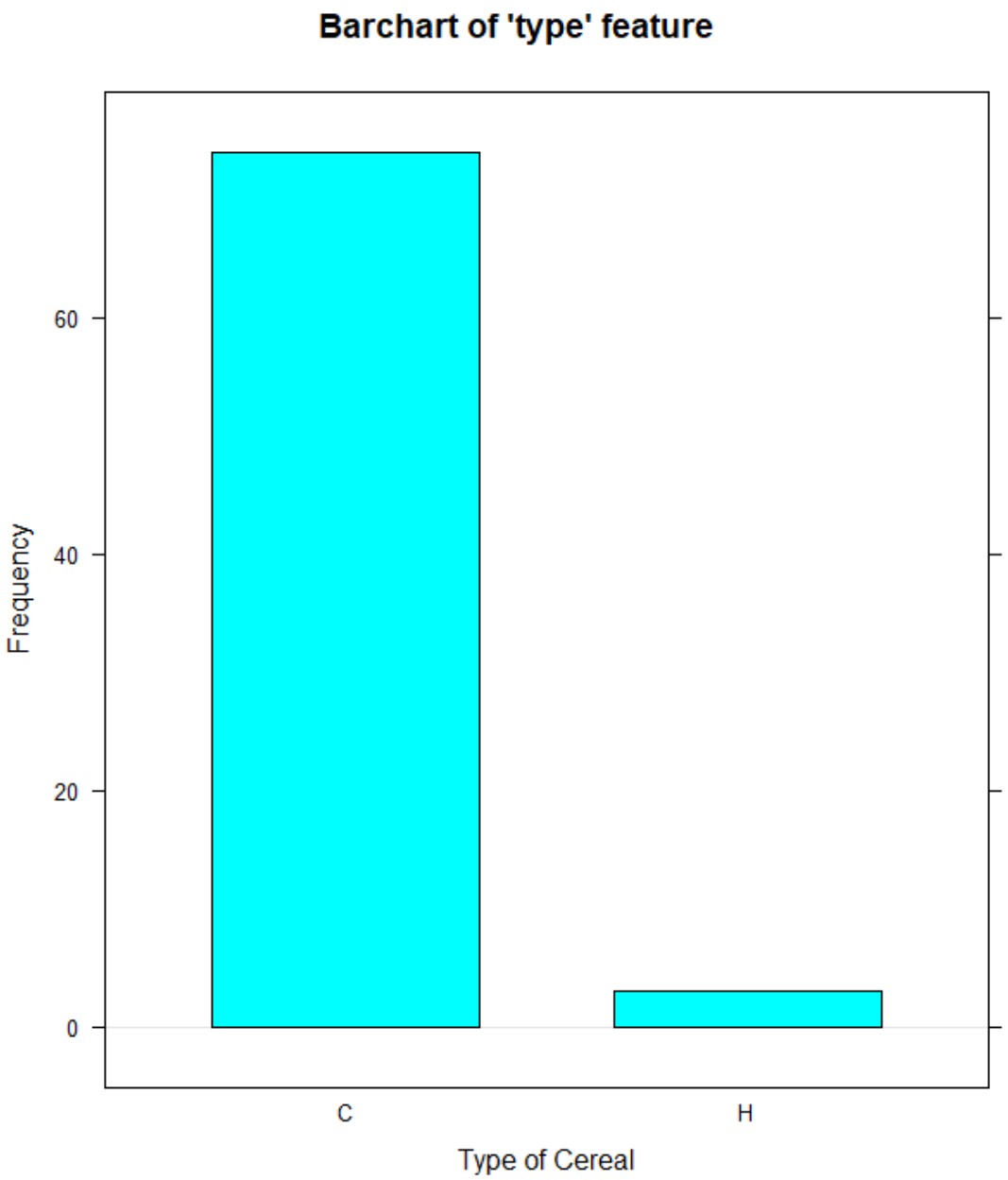
- Plotted Bar-chart for categorical Features.
- Histograms were plotted for continuous variable features and discrete variable.
- Boxplot were plotted for continuous variable feature and possible outliers were noted.
- Noted outliers of some features and saved their indexes. Dropped Indexes that were common outliers in two or more feature

2.4.1 Feature ~ mfr (Manufacturer)



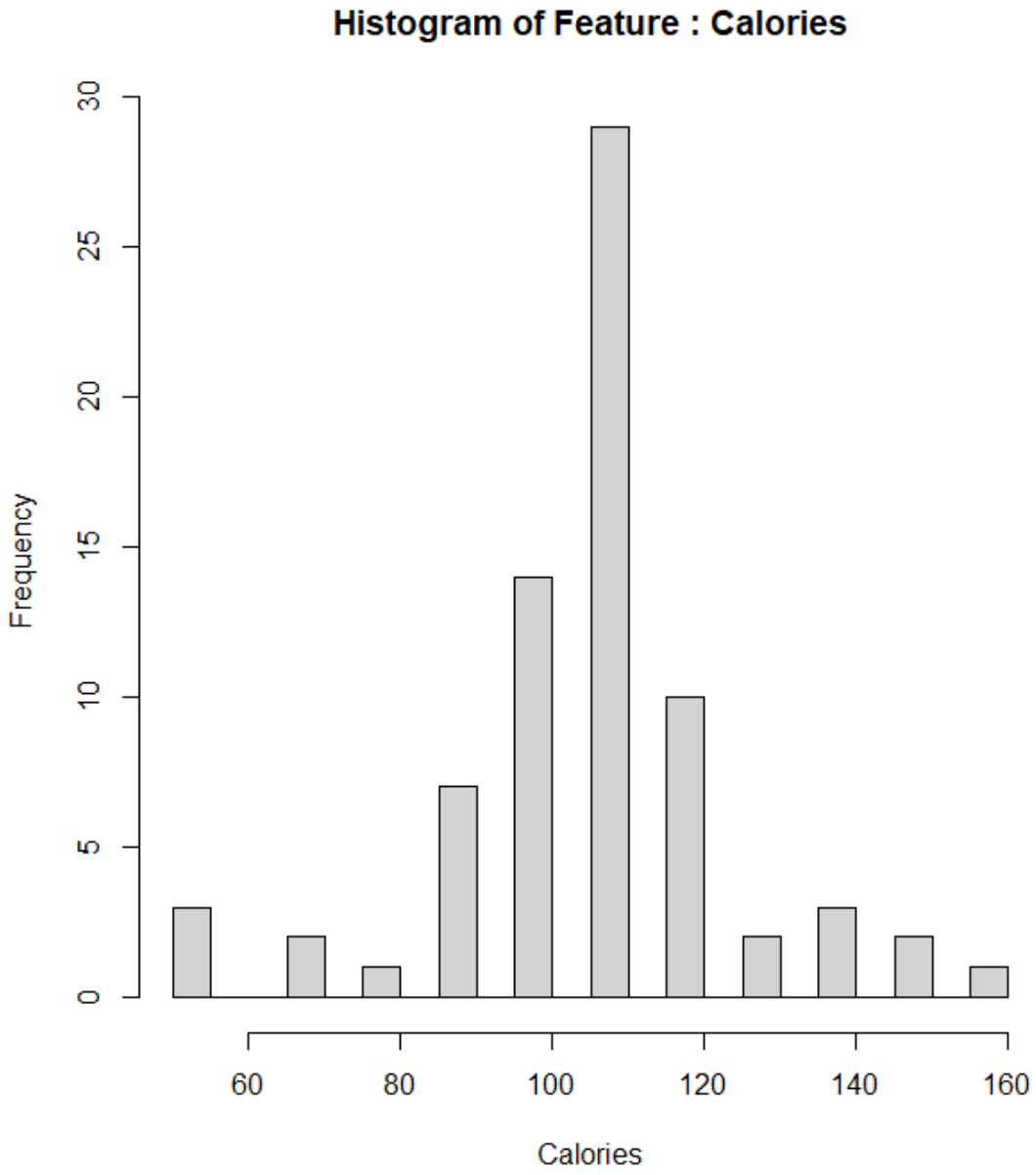
So, manufacturer 'A' has only 1 product to it. For now, I'll just save the index of that data point. Index was ‘44’.

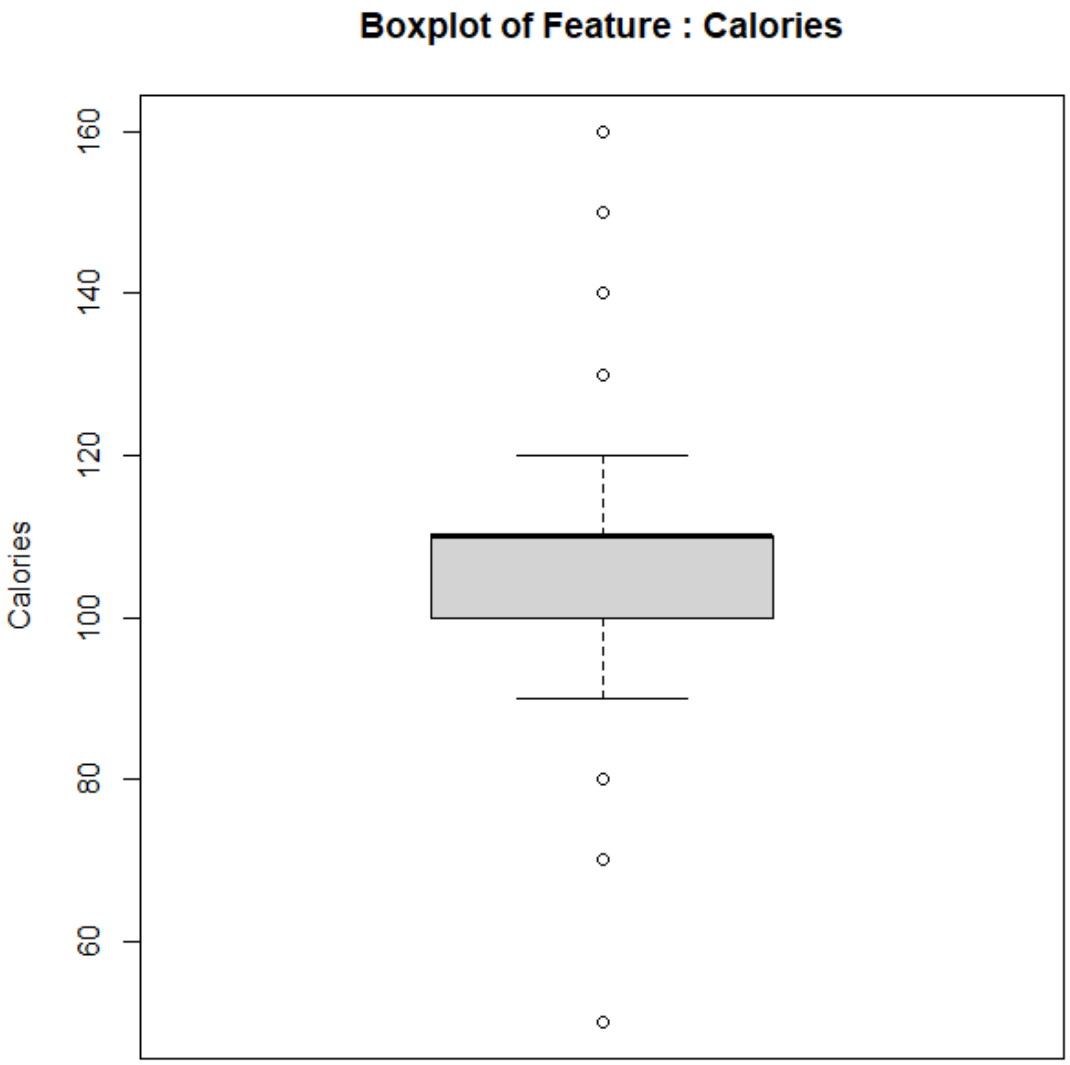
2.4.2 Feature ~ type (Type of Cereal)



So, here type ‘H’ is significantly less than ‘C’. So I found out the indexes and surprisingly index ‘44’ {outlier in ‘mfr’ feature} also came up here. So, I created a new data frame and dropped these 3 records from it.

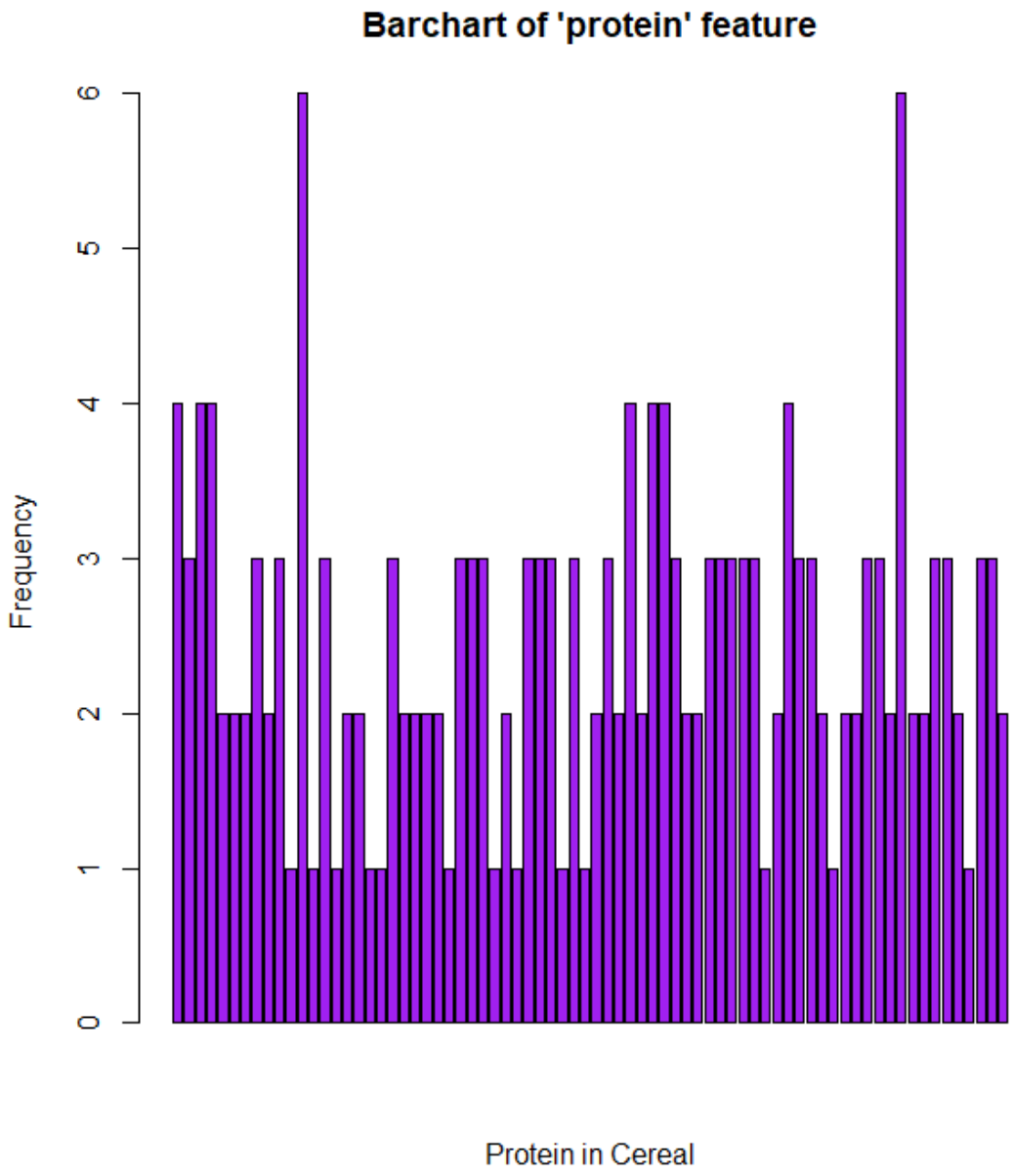
2.4.3 Feature ~ calories (Number of Calories in one portion)



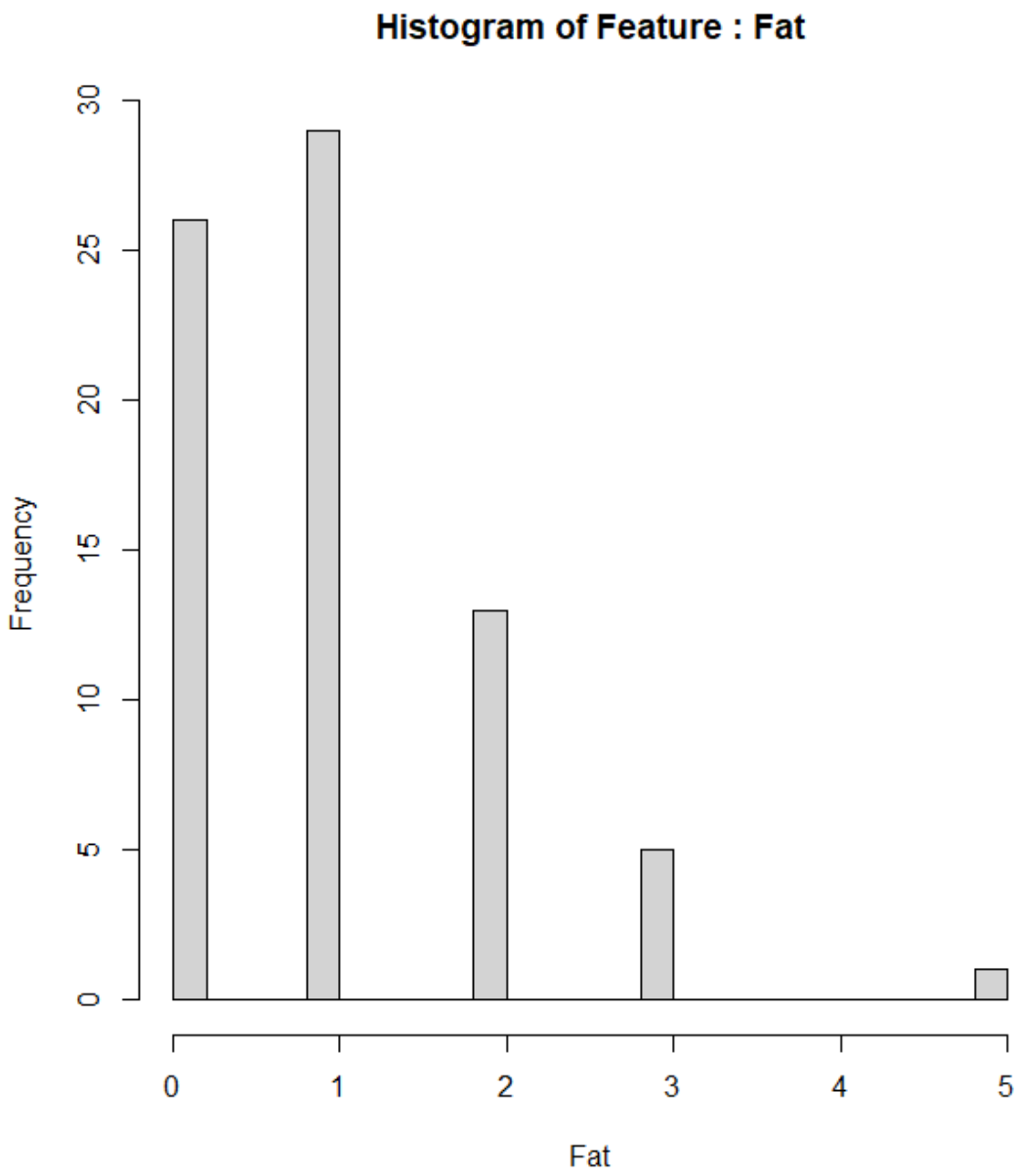


So, box plot of calories shows some outliers, but I won't remove them as of yet because we don't have much data. Instead, I'll again save those points index. Indexes were: 1 3 4 8 39 43 44 45 48 50 53 54 61 68

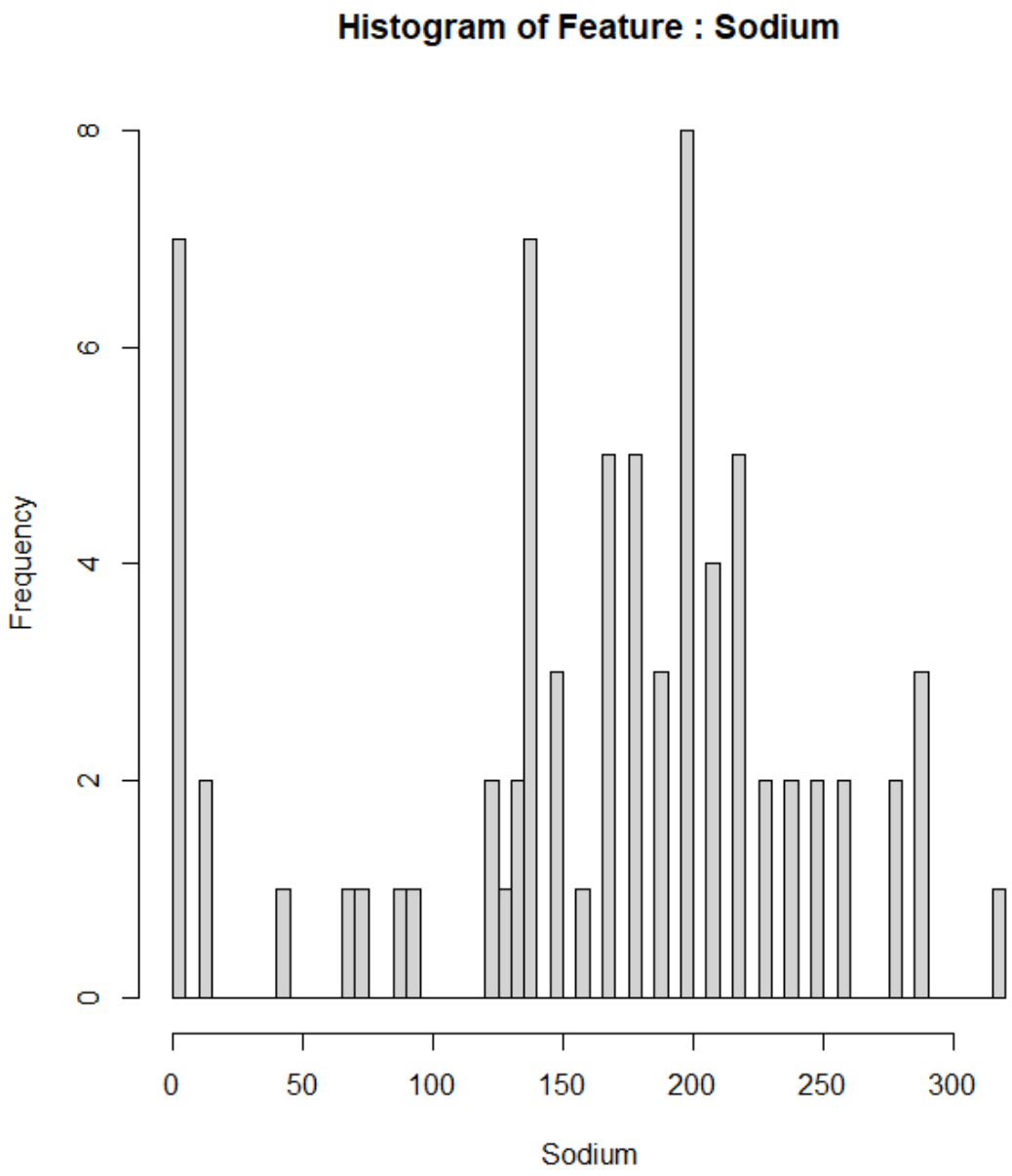
2.4.4 Feature ~ protein (Grams of Protein in one portion)



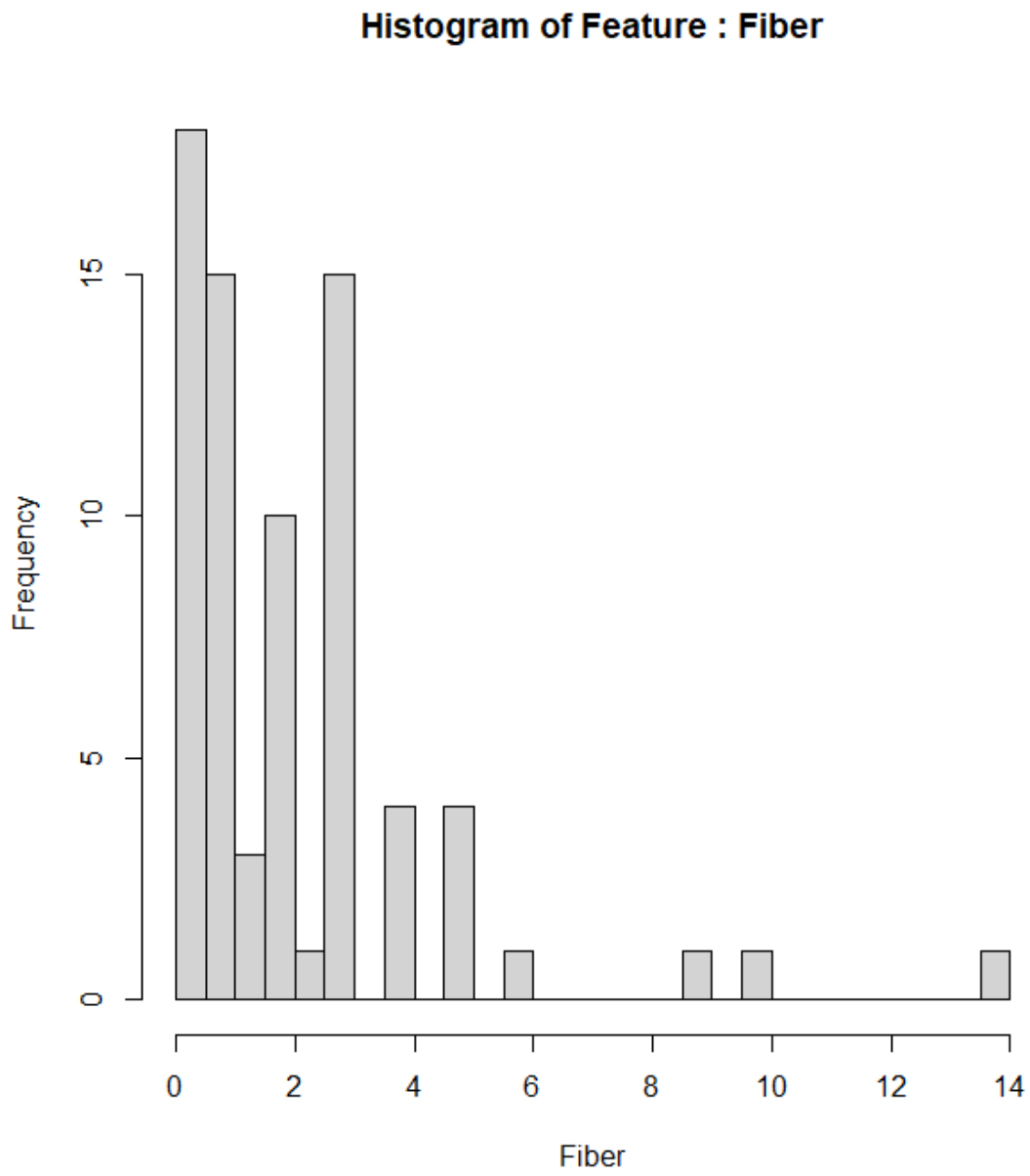
2.4.5 Feature ~ Fat (Grams of Fat in one portion)



2.4.6 Feature ~ Sodium (Milligrams of Sodium in one portion)

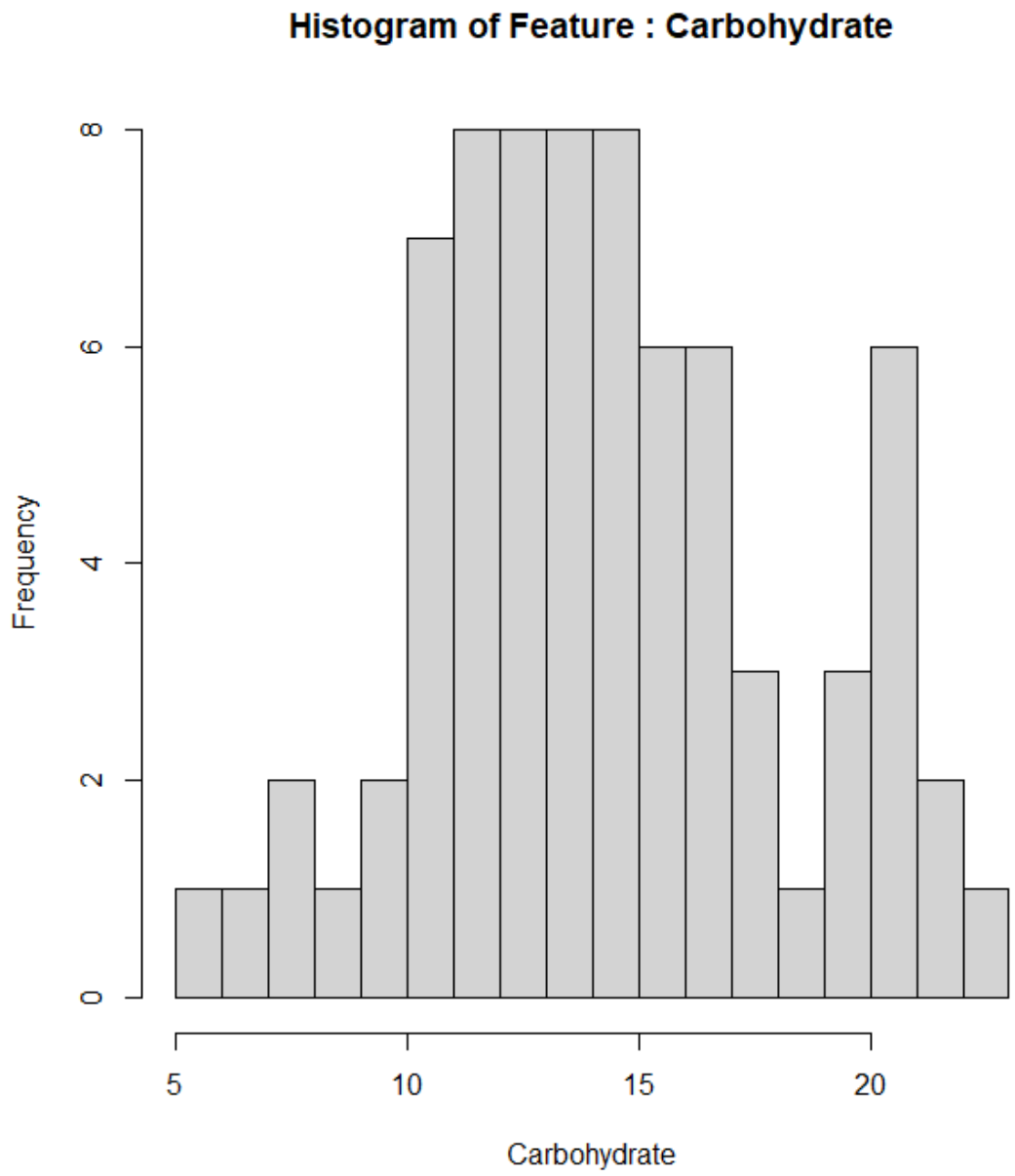


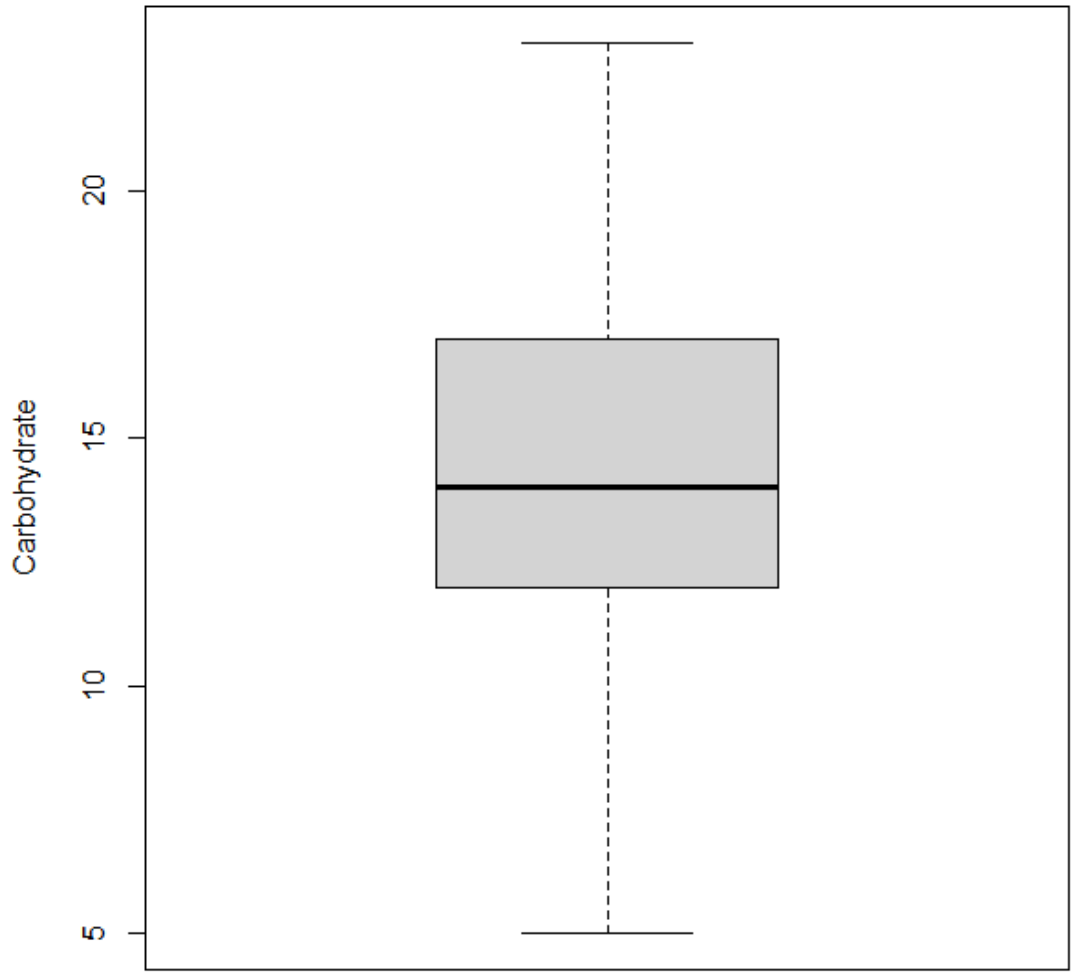
2.4.7 Feature ~ Fiber (Grams of dietary fiber in one portion)



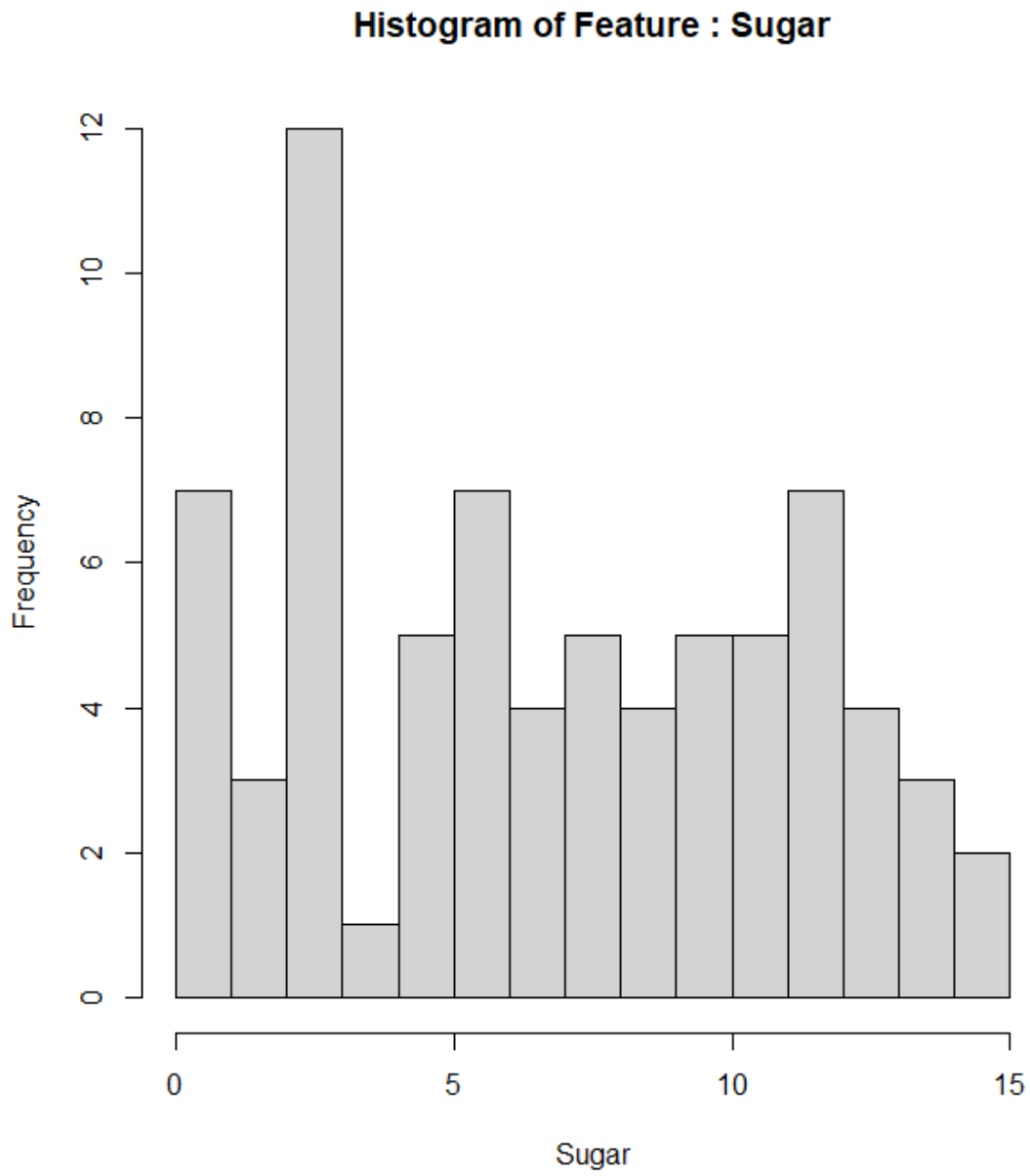
So, here one data point is way too far from other data point. Again, I saved the index of that point. Index was: ‘4’

2.4.8 Feature ~ Carbo (Grams of Complex Carbohydrates in one portion)

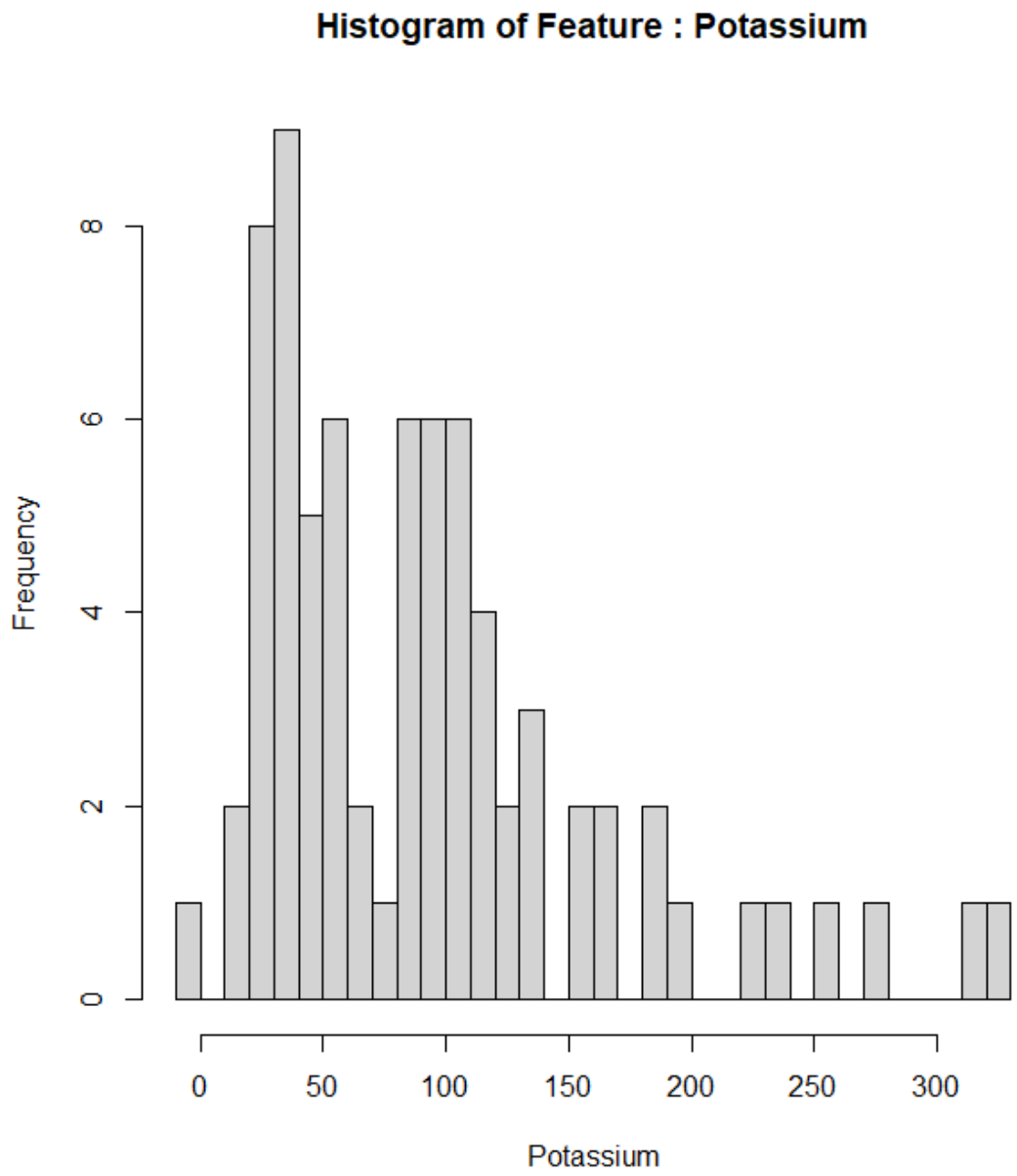


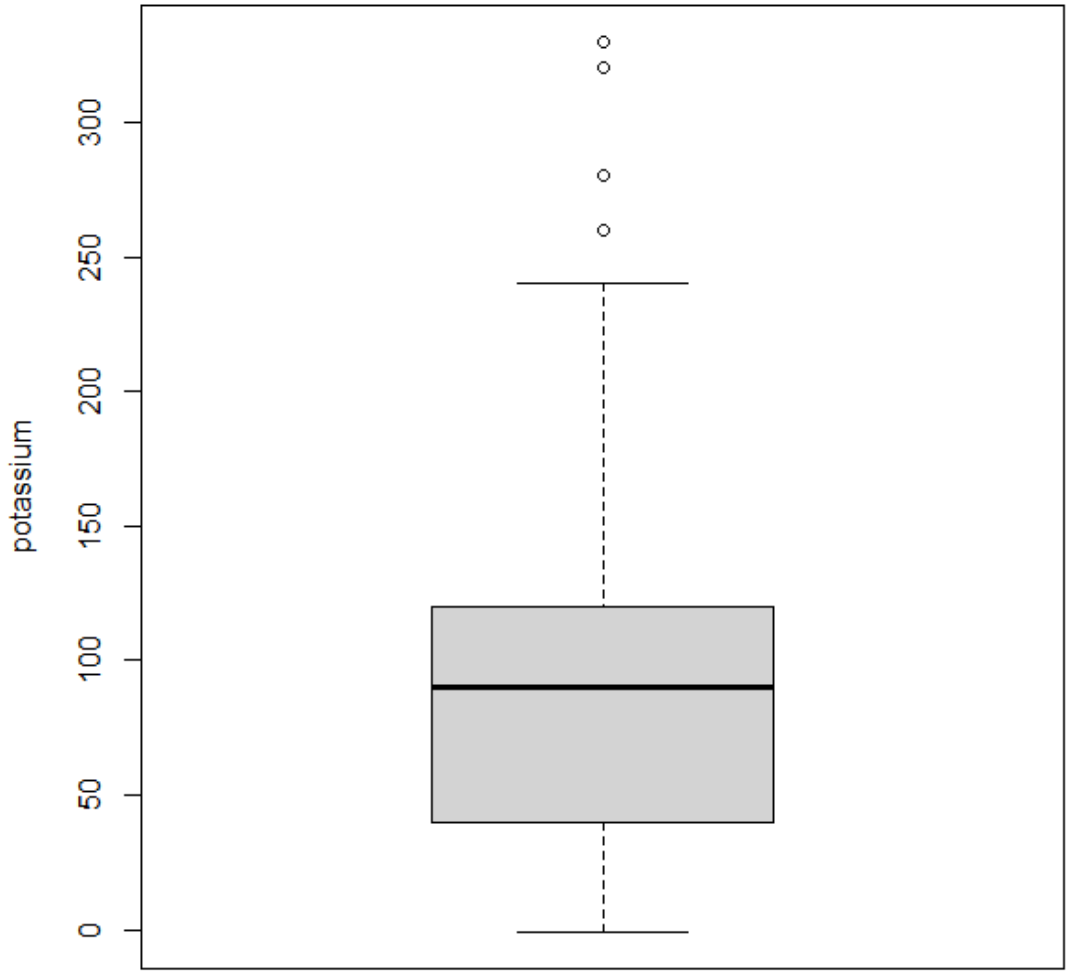

Carbohydrate
5
10
15
20

2.4.9 Feature ~ sugars (Grams of sugar in one portion)



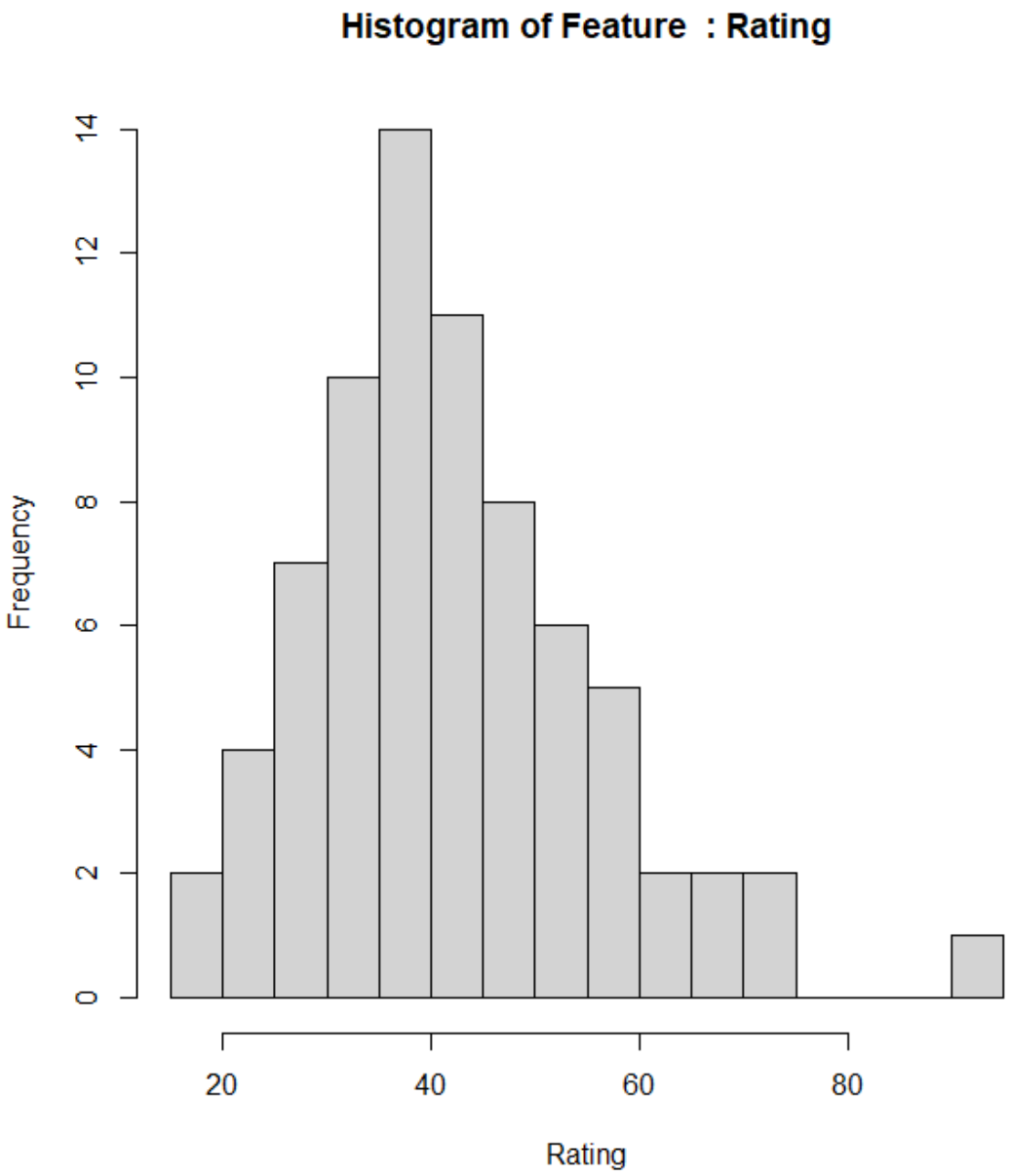
2.4.10 Feature ~ potass (Grams of potassium in one portion)

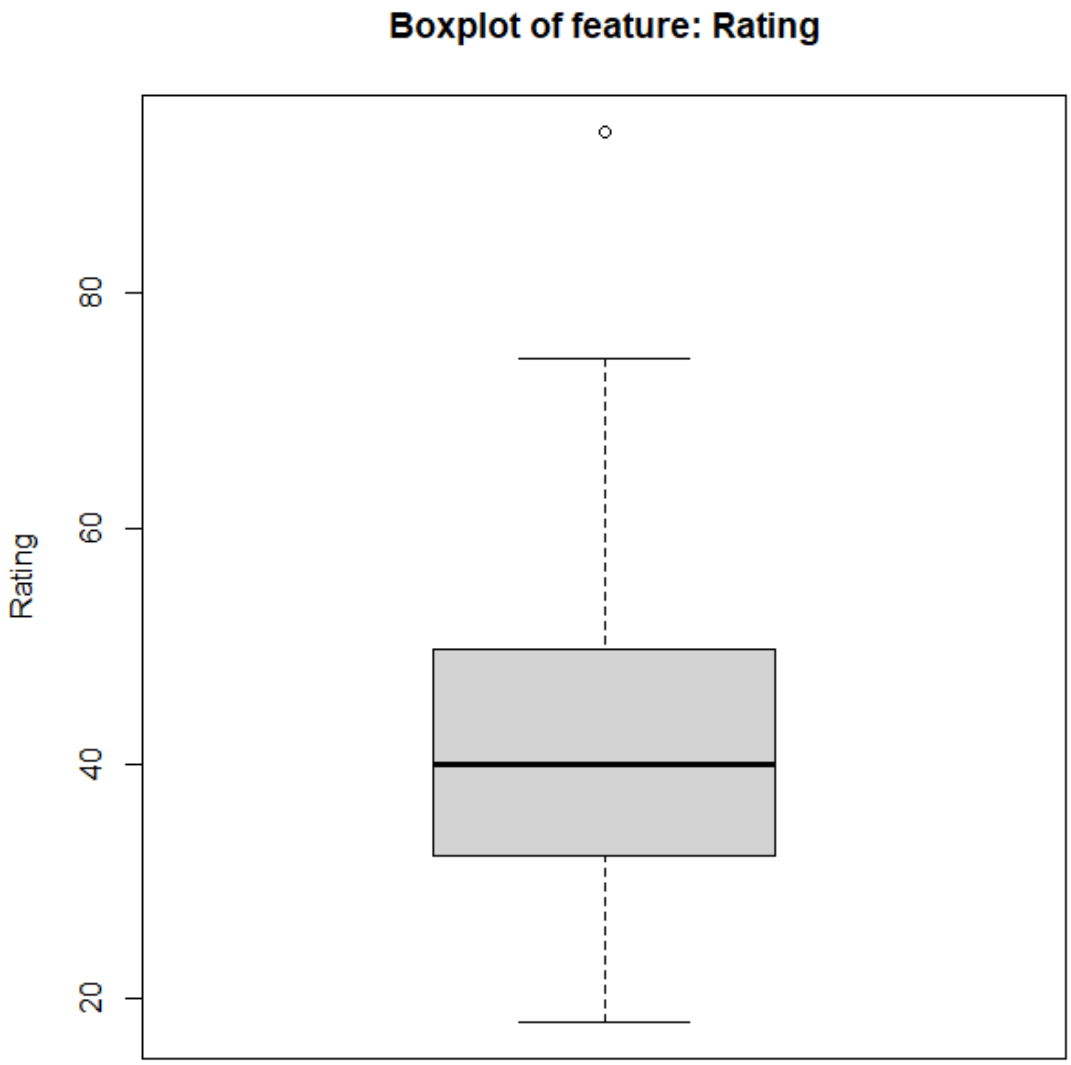




Here Boxplot shows some data point as outliers. I saved the indexes of all those points. Indexes were: 1 3 4 51

2.4.11 Feature ~ Rating (Rating of Cereal)



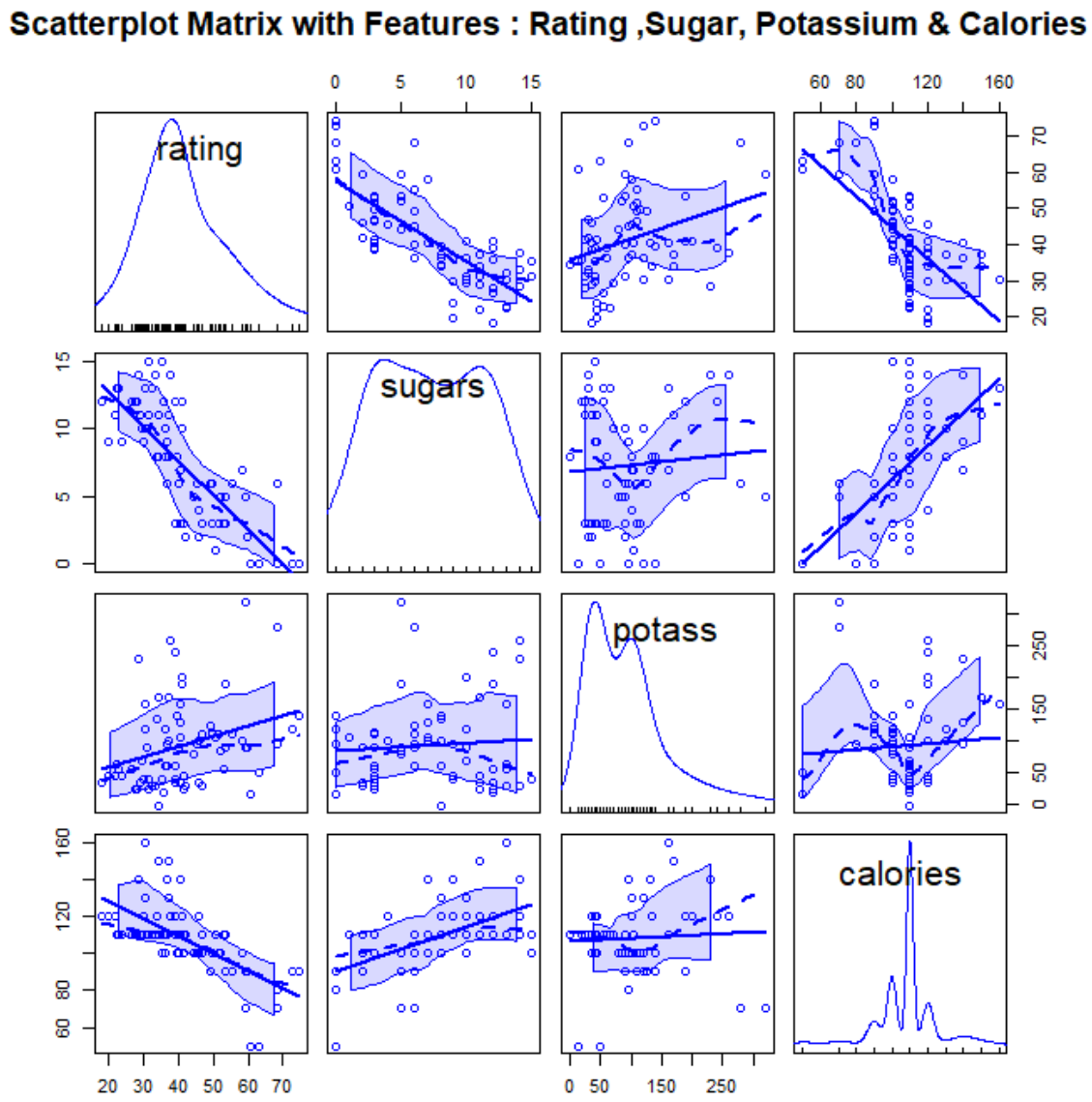


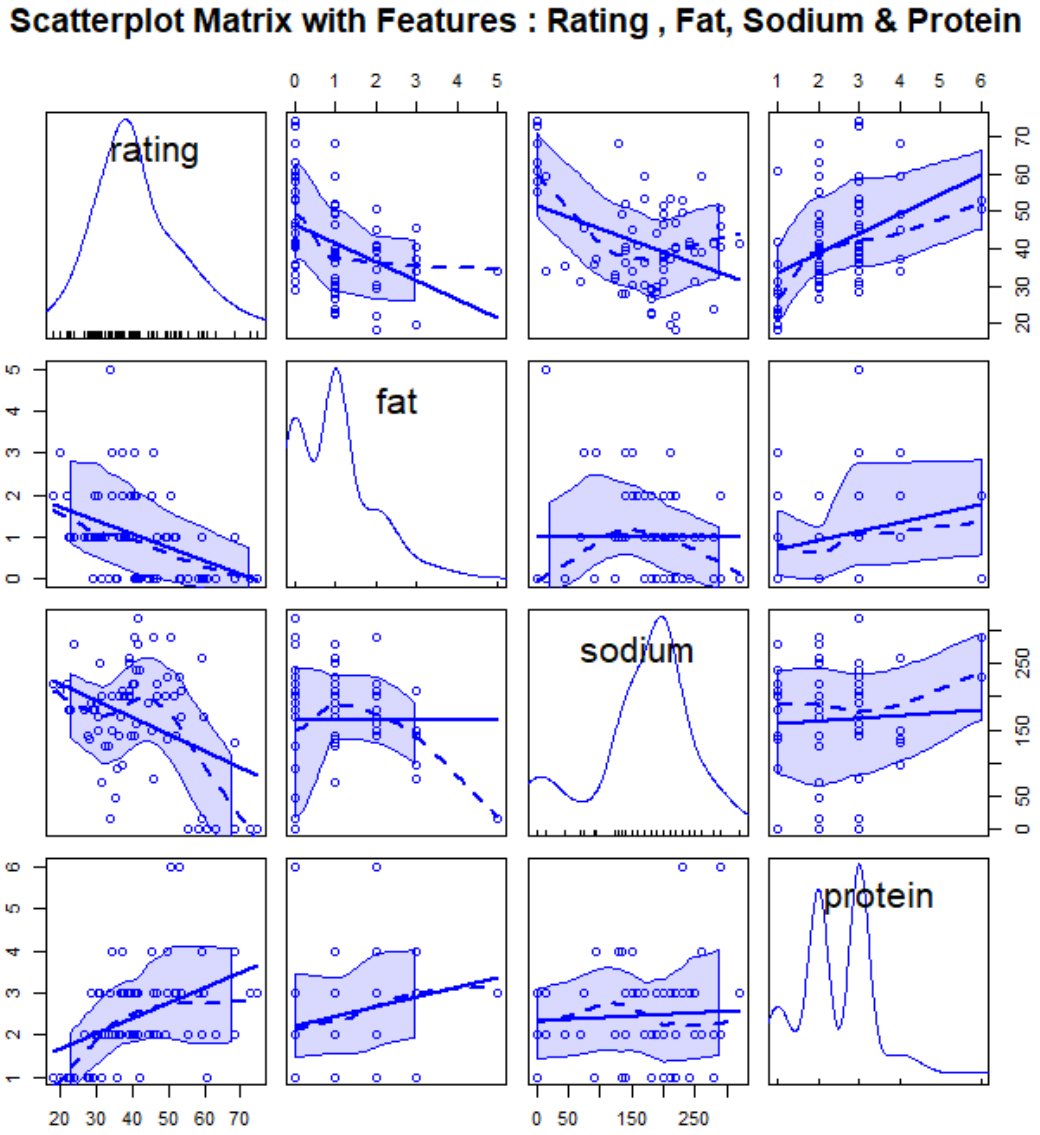
Here, both histogram and boxplot show a clear outlier. I saved its index value. Index was: ‘4’

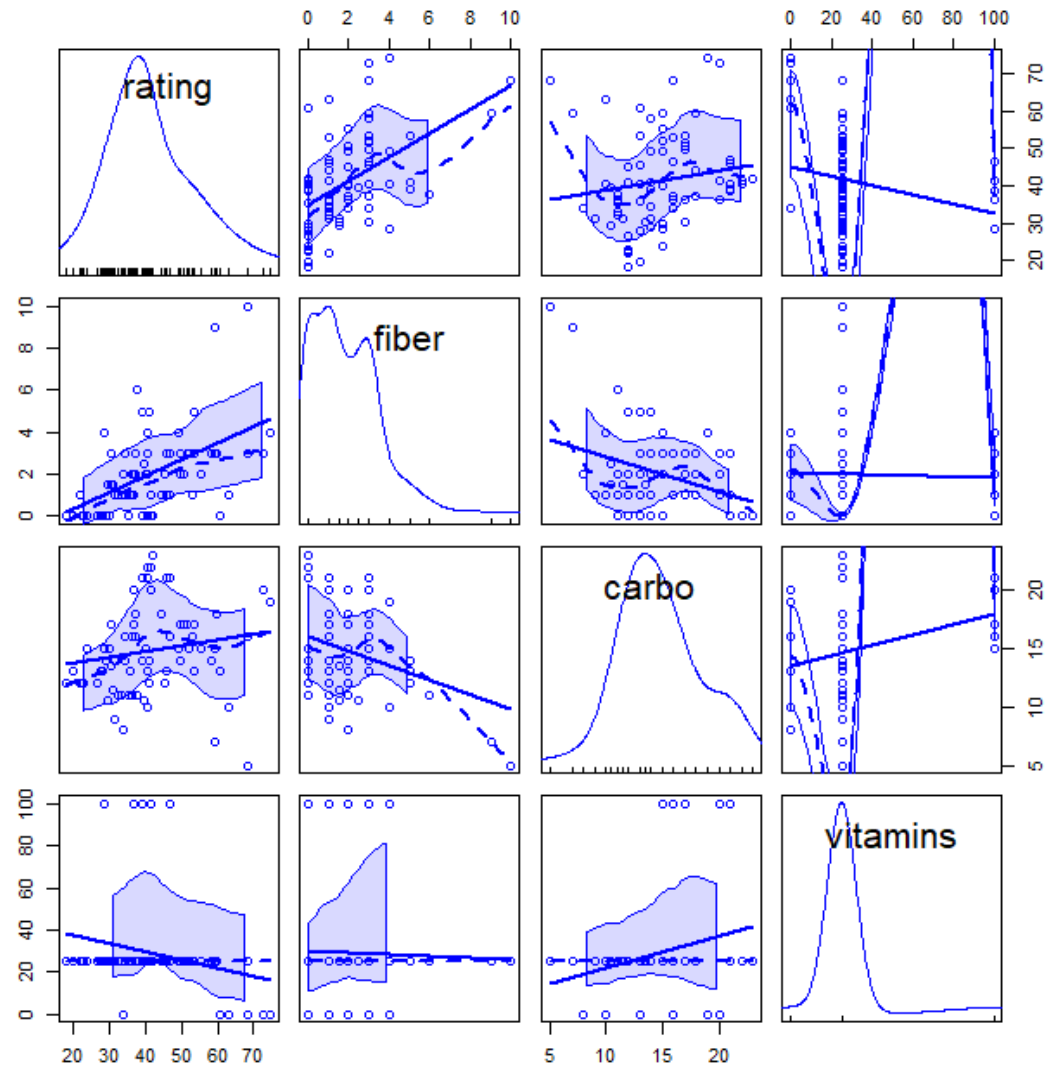
After saving indexes of all possible outcomes in each feature, I found index ‘4’ is possible outlier for four features (rating, potass, fiber, calories). So, I removed that row from our data frame.

2.5 Multivariate Analysis ~ Scatterplot.

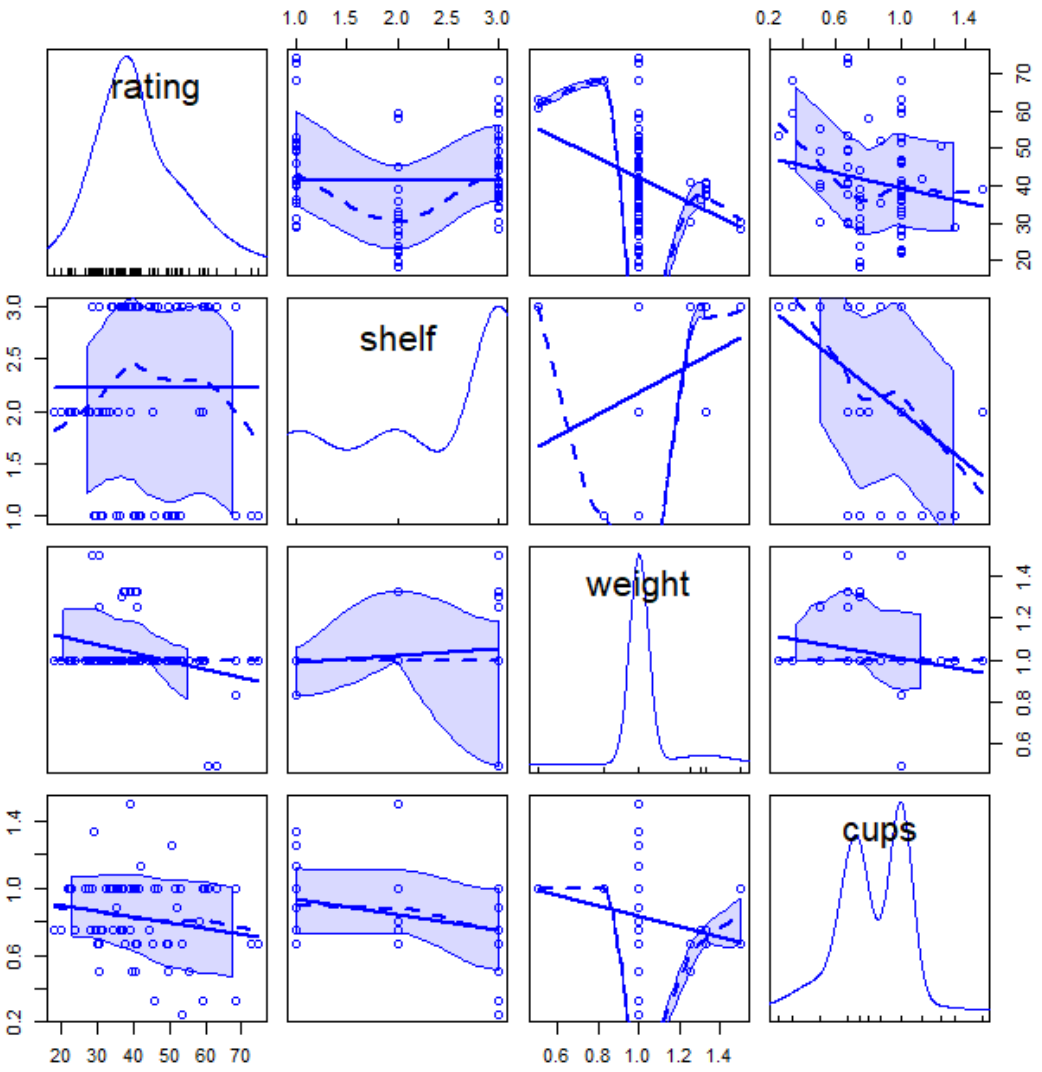
2.5.1 Plotting Multiple scatterplots and finding each features relation with our target feature 'rating'.




Scatterplot Matrix with Features : Rating , Fat, Sodium & Protein
rating
fat
sodium
protein


Scatterplot Matrix with Features : Rating , Fiber, Carbo & Vitamins
rating
fiber
carbo
vitamins

Scatterplot Matrix with Features : Rating , Shelf, Weight & Cups



2.5.2 Scatterplot Findings:

- Scatterplot 1: Features: Rating, Sugar, Potassium & Calories

  ~ From this scatterplot I can tell that, Sugar and calories has negative impact on the Rating Feature.
  ~ Potassium feature has a slight positive curve that mean it is slightly positively correlated to rating.

- Scatterplot 2: Features: Rating, Fat, Sodium & Protein

  ~ From this scatter plots I can tell that fat and sodium have negative impact on the Rating Feature.
  ~ protein feature is positively correlated to rating.

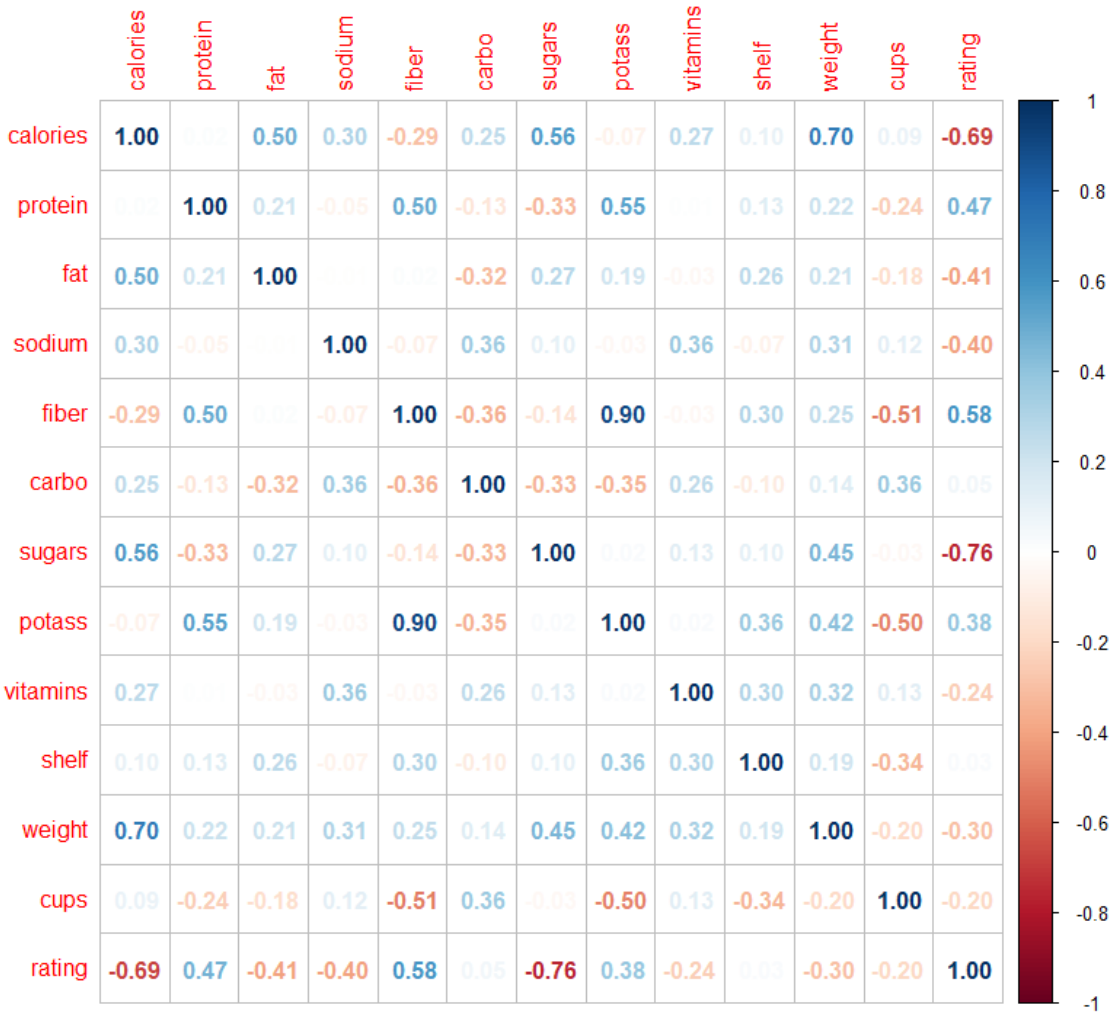
- Scatterplot 3: Features: Rating, Fiber, Carbo & Vitamins

  ~ Fiber feature is positively correlated to our target feature.
  ~ Vitamin feature is slightly negatively correlated to our target feature.

- Scatterplot 4: Features: Rating, Shelf, Weight & Cups

  ~ Features shelf, weight and cups has either a slight line or slightly negative curve. These features are not important to our target feature.
  ~ I will plot a co-relation plot and if they are not correlated then I'll remove them from our data frame.

2.6 Co-relation Plot.



2.6.1 Finding:

- Weight, Cups and Shelf is not highly co-related with rating, but weight is positively correlated to calories. So, I'll remove Cups and shelf from our data frame.
- Calories and Sugar is highly negatively correlated to rating feature.
- Fiber is positively correlated to rating feature.

3 Cleaned Data :

The final cleaned data set was saved as:
dataset named "Data4"
"cleaned_cereal_data.RData"
73 rows * 14 columns

4 Citations:

a) **https://datacornering.com/check-if-a-column-has-a-missing-values-na-in-r/**
b) **https://stackoverflow.com/questions/5863097/selecting-only-numeric-columns-from-a-data-frame**