SDM Assignment 1

Sri Balaji Muruganandam

17/09/2021

Setting Working Directory

```
rm(list = ls())
setwd("G:\\SDM_Sem01\\Assignment1")
```

Importing necessary libraries

```
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 4.1.1
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Importing Cereal data from local path

```
c_data <- read.delim("cereal.csv",sep = ",")
dim(c_data)
```

```
## [1] 77 16
```

Viewing the Sample data

```
head(c_data, 5)
```

```
##                        name mfr type calories protein fat sodium fiber carbo
## 1                 100% Bran   N    C       70       4   1    130    10     5
## 2         100% Natural Bran   Q    C      120       3   5     15     2     8
## 3                  All-Bran   K    C       70       4   1    260     9     7
## 4 All-Bran with Extra Fiber   K    C       50       4   0    140    14     8
## 5            Almond Delight   R    C      110       2   2    200     1    14
##   sugars potass vitamins shelf weight cups   rating
## 1      6    280       25     3      1 0.33 68.40297
## 2      8    135        0     3      1 1.00 33.98368
## 3      5    320       25     3      1 0.33 59.42551
## 4      0    330       25     3      1 0.50 93.70491
## 5      8     -1       25     3      1 0.75 34.38484
```

```
tail(c_data, 5)
```

```
##                   name mfr type calories protein fat sodium fiber carbo sugars
## 73             Triples   G    C      110       2   1    250     0    21      3
## 74                Trix   G    C      110       1   1    140     0    13     12
## 75          Wheat Chex   R    C      100       3   1    230     3    17      3
## 76            Wheaties   G    C      100       3   1    200     3    17      3
## 77 Wheaties Honey Gold   G    C      110       2   1    200     1    16      8
##    potass vitamins shelf weight cups   rating
## 73     60       25     3      1 0.75 39.10617
## 74     25       25     2      1 1.00 27.75330
## 75    115       25     1      1 0.67 49.78744
## 76    110       25     1      1 1.00 51.59219
## 77     60       25     1      1 0.75 36.18756
```

Getting High Level Overview of the data

```
str(c_data)
```

```
## 'data.frame':    77 obs. of  16 variables:
##  $ name    : chr  "100% Bran" "100% Natural Bran" "All-Bran" "All-Bran with Extra Fiber"
...
##  $ mfr     : chr  "N" "Q" "K" "K" ...
##  $ type    : chr  "C" "C" "C" "C" ...
##  $ calories: int  70 120 70 50 110 110 110 130 90 90 ...
##  $ protein : int  4 3 4 4 2 2 2 3 2 3 ...
##  $ fat     : int  1 5 1 0 2 2 0 2 1 0 ...
##  $ sodium  : int  130 15 260 140 200 180 125 210 200 210 ...
##  $ fiber   : num  10 2 9 14 1 1.5 1 2 4 5 ...
##  $ carbo   : num  5 8 7 8 14 10.5 11 18 15 13 ...
##  $ sugars  : int  6 8 5 0 8 10 14 8 6 5 ...
##  $ potass  : int  280 135 320 330 -1 70 30 100 125 190 ...
##  $ vitamins: int  25 0 25 25 25 25 25 25 25 25 ...
##  $ shelf   : int  3 3 3 3 3 1 2 3 1 3 ...
##  $ weight  : num  1 1 1 1 1 1 1 1.33 1 1 ...
##  $ cups    : num  0.33 1 0.33 0.5 0.75 0.75 1 0.75 0.67 0.67 ...
##  $ rating  : num  68.4 34 59.4 93.7 34.4 ...
```

```
summary(c_data)
```

```
##      name               mfr                type             calories    
##  Length:77          Length:77          Length:77          Min.   : 50.0  
##  Class :character   Class :character   Class :character   1st Qu.:100.0  
##  Mode  :character   Mode  :character   Mode  :character   Median :110.0  
##                                                           Mean   :106.9  
##                                                           3rd Qu.:110.0  
##                                                           Max.   :160.0  
##     protein           fat            sodium          fiber       
##  Min.   :1.000   Min.   :0.000   Min.   :  0.0   Min.   : 0.000  
##  1st Qu.:2.000   1st Qu.:0.000   1st Qu.:130.0   1st Qu.: 1.000  
##  Median :3.000   Median :1.000   Median :180.0   Median : 2.000  
##  Mean   :2.545   Mean   :1.013   Mean   :159.7   Mean   : 2.152  
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:210.0   3rd Qu.: 3.000  
##  Max.   :6.000   Max.   :5.000   Max.   :320.0   Max.   :14.000  
##      carbo          sugars           potass          vitamins     
##  Min.   :-1.0   Min.   :-1.000   Min.   : -1.00   Min.   :  0.00  
##  1st Qu.:12.0   1st Qu.: 3.000   1st Qu.: 40.00   1st Qu.: 25.00  
##  Median :14.0   Median : 7.000   Median : 90.00   Median : 25.00  
##  Mean   :14.6   Mean   : 6.922   Mean   : 96.08   Mean   : 28.25  
##  3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00  
##  Max.   :23.0   Max.   :15.000   Max.   :330.00   Max.   :100.00  
##      shelf           weight          cups           rating     
##  Min.   :1.000   Min.   :0.50   Min.   :0.250   Min.   :18.04  
##  1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.670   1st Qu.:33.17  
##  Median :2.000   Median :1.00   Median :0.750   Median :40.40  
##  Mean   :2.208   Mean   :1.03   Mean   :0.821   Mean   :42.67  
##  3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:50.83  
##  Max.   :3.000   Max.   :1.50   Max.   :1.500   Max.   :93.70  
```

Checking if there are any missing values

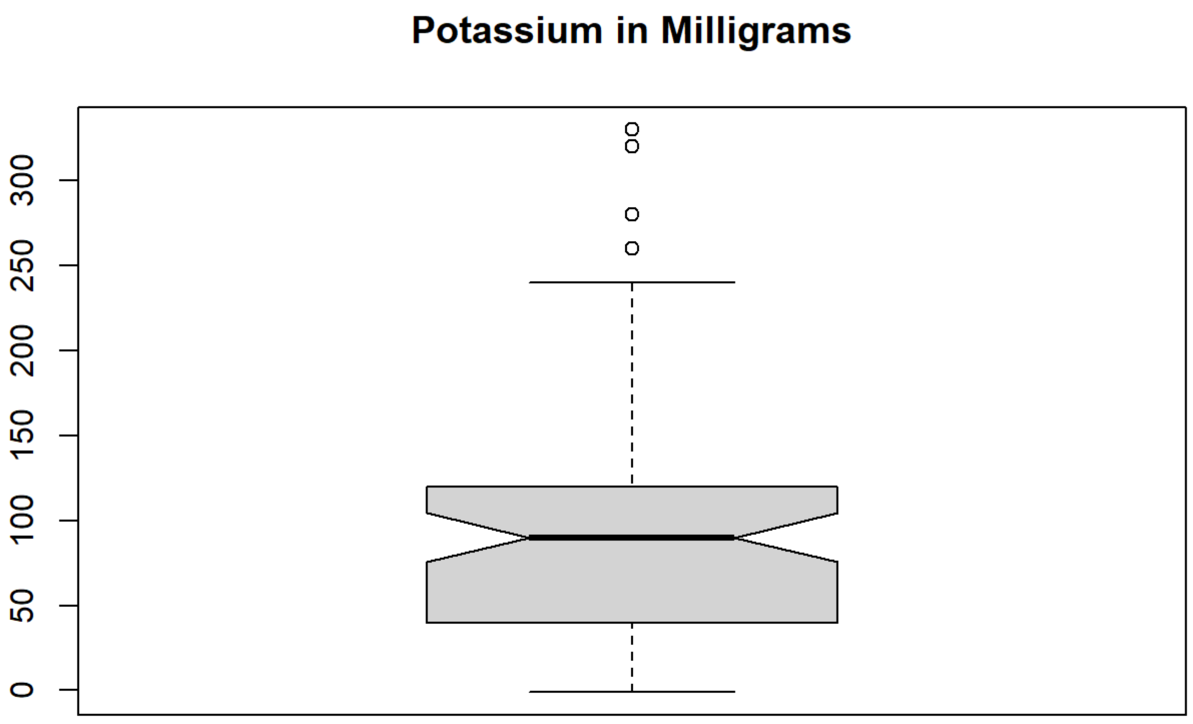
```
sum(is.na(c_data))
```

```
## [1] 0
```

There is no missing data in the dataset

Outliers and Graphs

```
boxplot(c_data$potass, notch = TRUE, main = "Potassium in Milligrams")
```



##### There are outliers in the potass column

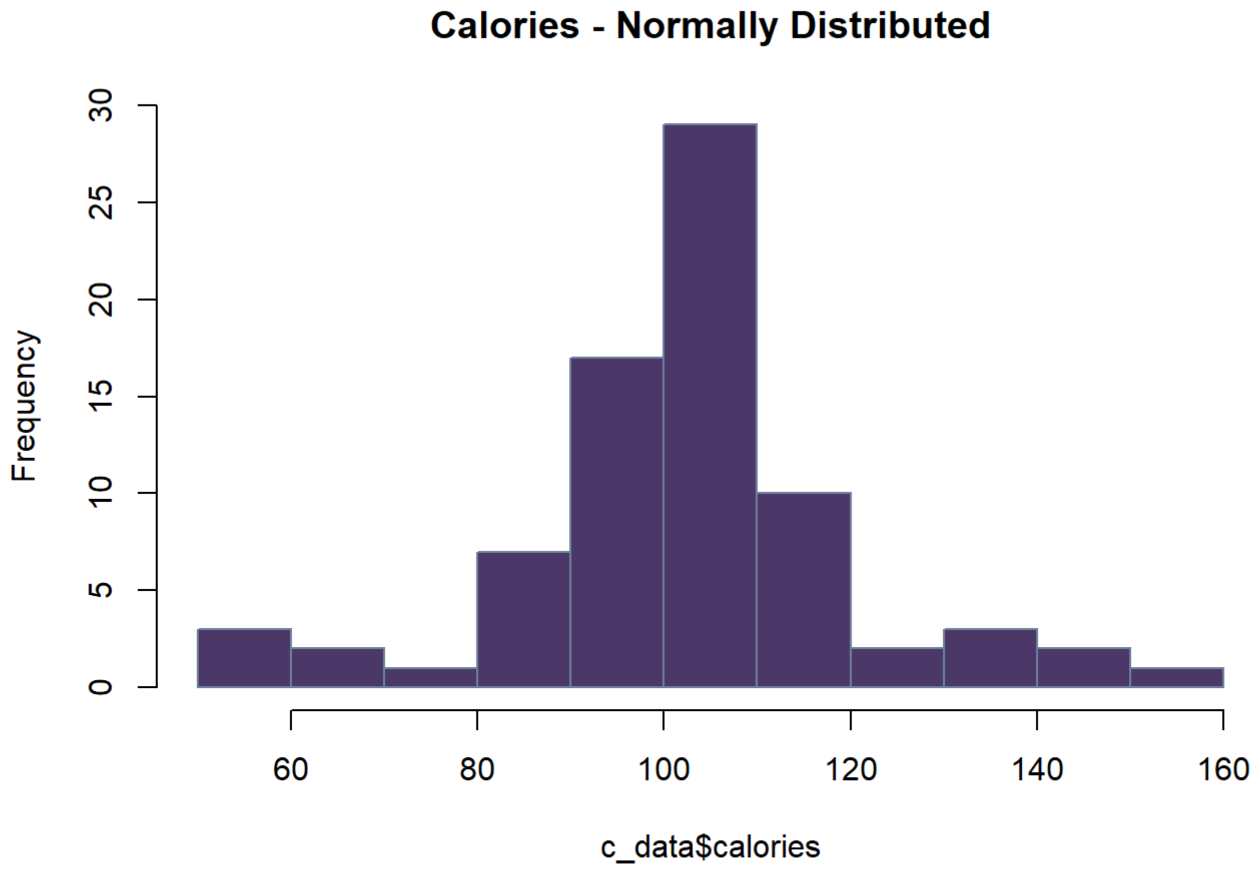
Removing the Outliers

As the data is very less. Keeping all of the data for Regression model.

```
#boxplot.stats(c_data$potass)$out
#out_potass <- boxplot.stats(c_data$potass)$out
#c_data <- filter(c_data, potass != out_potass)
#dim(c_data)
#head(c_data, 10)
```

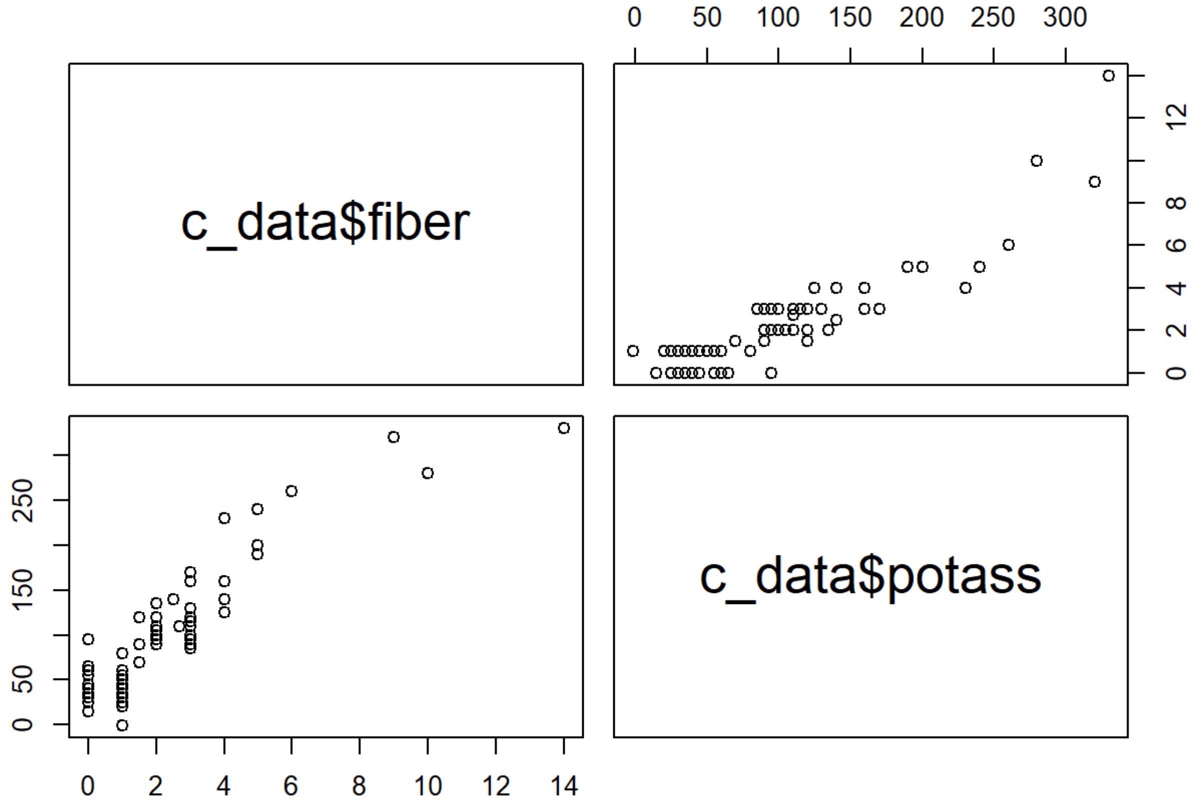
Calories is normally distributed

```
hist(c_data$calories,breaks = 10, col="#4B3869", border = "#6D8299", main = "Calories - Norma
lly Distributed")
```



There is a positive correlation between fiber and potass

```
pairs(c_data$fiber~c_data$potass)
```



Checking the unique values of mfr and type

```
table(c_data$mfr)
```

```
##
##  A  G  K  N  P  Q  R
##  1 22 23  6  9  8  8
```

```
table(c_data$type)
```

```
##
##  C  H
## 74  3
```

Converting categorical variables into numerical values

```
mfr_fact <- as.factor(c_data$mfr)
c_data$mfr <- as.numeric(mfr_fact)

c_data$type<-ifelse(c_data$type=="C",1,0)
```

Data Transformation

Converting vitamins(percentage) to float value

```
table(c_data$vitamins)
```

```
## 
##   0  25 100 
##   8  63   6
```

```
c_data$vitamins <- c_data$vitamins / 100

#c_data$vitamins_num <- log(c_data$vitamins, base = 10)/2
#c_data$vitamins_num<-ifelse(c_data$vitamins_num =="-Inf",0,c_data$vitamins_num)
```

Converting rating to float value

```
c_data$rating <- c_data$rating / 100
```

Taking log to Potass Column

Values of the Potass varying so much.

```
table(c_data$potass)
```

```
## 
##  -1  15  20  25  30  35  40  45  50  55  60  65  70  80  85  90  95 100 105 110 
##   2   1   1   4   4   5   4   4   1   3   3   1   1   1   1   5   4   3   2   5 
## 115 120 125 130 135 140 160 170 190 200 230 240 260 280 320 330 
##   1   3   1   1   1   2   2   2   2   1   1   1   1   1   1   1
```

```
c_data$potass <- log(c_data$potass, base = 10)
```

```
## Warning: NaNs produced
```

```
c_data <- na.omit(c_data)

head(c_data, 10)
```

```
##                          name mfr type calories protein fat sodium fiber carbo
## 1                  100% Bran   4    1       70       4   1    130  10.0   5.0
## 2          100% Natural Bran   6    1      120       3   5     15   2.0   8.0
## 3                   All-Bran   3    1       70       4   1    260   9.0   7.0
## 4  All-Bran with Extra Fiber   3    1       50       4   0    140  14.0   8.0
## 6    Apple Cinnamon Cheerios   2    1      110       2   2    180   1.5  10.5
## 7                Apple Jacks   3    1      110       2   0    125   1.0  11.0
## 8                    Basic 4   2    1      130       3   2    210   2.0  18.0
## 9                  Bran Chex   7    1       90       2   1    200   4.0  15.0
## 10               Bran Flakes   5    1       90       3   0    210   5.0  13.0
## 11              Cap'n'Crunch   6    1      120       1   2    220   0.0  12.0
##    sugars   potass vitamins shelf weight cups    rating
## 1       6 2.447158     0.25     3   1.00 0.33 0.6840297
## 2       8 2.130334     0.00     3   1.00 1.00 0.3398368
## 3       5 2.505150     0.25     3   1.00 0.33 0.5942551
## 4       0 2.518514     0.25     3   1.00 0.50 0.9370491
## 6      10 1.845098     0.25     1   1.00 0.75 0.2950954
## 7      14 1.477121     0.25     2   1.00 1.00 0.3317409
## 8       8 2.000000     0.25     3   1.33 0.75 0.3703856
## 9       6 2.096910     0.25     1   1.00 0.67 0.4912025
## 10      5 2.278754     0.25     3   1.00 0.67 0.5331381
## 11     12 1.544068     0.25     2   1.00 0.75 0.1804285
```

```
sum(is.na(c_data))
```

```
## [1] 0
```

Saving the preprocessed data in Rdata file

```
save(c_data, file = "cereal_clean_data.RData")
```