End Term - I2R

Chiranjeevraja - D21010, Kirankumar A M - D21018

24/02/2022

1. Problem Statement:

In order to obtain the real compressive strength of concrete (target labels in the dataset), an engineer needs to cast sample(s) of concrete produced in the batching plant before sending it to the construction site for pouring (no. of samples vary based on the grade and the quantity of concrete to be used in site). **For the concrete to achieve its desired strength, one has to cure the sample for 28 days in order to find the final compressive strength of concrete as per IS standards (IS: 516 - 1959).** The cured samples will then be broken with the help of compression-testing machine. The failure load is divided by the cross-sectional area of the sample to obtain the compressive strength.

All resources and time spent in the above processes can be minimized if we can set up a machine learning model to predict the compressive strength of concrete with the help of features such as cement, slag, flyash, water, plasticizer, aggregate contents and the age of the concrete. Also construction contractors based on their design requirement can adequately plan for the concrete raw materials, well in advance, with the help of this model to produce concrete of required strength.

2. Objective:

The task is to predict the compressive strength of concrete (in MPa) using features which has the quantity (in Kg/cu.m) of individual raw materials used in the production of concrete. **The overall objective is to provide results swiftly and accurately using minimal resources and not make the end user wait till the last day of the curing period.**

3. Label and Features:

Details of the predictors and target variable are as follows,

Name – Data Type – Measurement – Description

Cement (component 1) – -quantitative – kg in a m3 mixture - Input Variable

Blast Furnace Slag (component 2) - quantitative – kg in a m3 mixture – Input Variable

Fly Ash (component 3) – quantitative - kg in a m3 mixture – Input Variable

Water (component 4) – quantitative – kg in a m3 mixture – Input Variable

Superplasticizer (component 5) – quantitative – kg in a m3 mixture – Input Variable

Coarse Aggregate (component 6) – quantitative – kg in a m3 mixture – Input Variable

Fine Aggregate (component 7) – quantitative – kg in a m3 mixture – Input Variable

Age (component 8) – quantitative – Days (1-365) – Input Variable

Concrete compressive strength – quantitative – CsMPa – Output Variable 1 csMPa = 10ˆ6 Newton/sq.m

4. Importing Dataset & Train-Test Split:

Dataset is to be imported and the train-test split (70%-30%) is made. Necessary type conversions are to be done and the dataset is checked for null values

```
concrete <-  read.csv('Concrete.csv')

train <- concrete[1:700,]

test <- concrete[701:nrow(concrete),]


#structure and summary of training dataset
head(train)
```

```
##   cement  slag flyash water superplasticizer coarseaggregate fineaggregate age
## 1  540.0   0.0      0   162              2.5          1040.0         676.0  28
## 2  540.0   0.0      0   162              2.5          1055.0         676.0  28
## 3  332.5 142.5      0   228              0.0           932.0         594.0 270
## 4  332.5 142.5      0   228              0.0           932.0         594.0 365
## 5  198.6 132.4      0   192              0.0           978.4         825.5 360
## 6  266.0 114.0      0   228              0.0           932.0         670.0  90
##   csMPa
## 1 79.99
## 2 61.89
## 3 40.27
## 4 41.05
## 5 44.30
## 6 47.03
```

```
summary(train)
```

```
##      cement           slag            flyash           water      
##  Min.   :102.0   Min.   :  0.00   Min.   :  0.00   Min.   :121.8  
##  1st Qu.:212.0   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:159.5  
##  Median :254.0   Median : 24.00   Median :  0.00   Median :178.5  
##  Mean   :286.9   Mean   : 71.99   Mean   : 54.64   Mean   :178.1  
##  3rd Qu.:374.0   3rd Qu.:130.53   3rd Qu.:118.30   3rd Qu.:192.0  
##  Max.   :540.0   Max.   :359.40   Max.   :174.70   Max.   :228.0  
##  superplasticizer coarseaggregate  fineaggregate        age        
##  Min.   : 0.000   Min.   : 801.0   Min.   :594.0   Min.   :  3.00  
##  1st Qu.: 0.000   1st Qu.: 936.0   1st Qu.:746.6   1st Qu.:  7.00  
##  Median : 6.500   Median : 968.0   Median :780.7   Median : 28.00  
##  Mean   : 6.507   Mean   : 979.9   Mean   :781.1   Mean   : 49.01  
##  3rd Qu.:10.900   3rd Qu.:1040.6   3rd Qu.:845.0   3rd Qu.: 56.00  
##  Max.   :32.200   Max.   :1145.0   Max.   :992.6   Max.   :365.00  
##      csMPa
##  Min.   : 2.33
##  1st Qu.:24.28
##  Median :36.62
##  Mean   :37.64
##  3rd Qu.:50.12
##  Max.   :82.60
```

```
str(train) #checking datatypes of features
```

```
## 'data.frame':    700 obs. of  9 variables:
##  $ cement          : num  540 540 332 332 199 ...
##  $ slag            : num  0 0 142 142 132 ...
##  $ flyash          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ water           : num  162 162 228 228 192 228 228 228 228 228 ...
##  $ superplasticizer: num  2.5 2.5 0 0 0 0 0 0 0 0 ...
##  $ coarseaggregate : num  1040 1055 932 932 978 ...
##  $ fineaggregate   : num  676 676 594 594 826 ...
##  $ age             : int  28 28 270 365 360 90 365 28 28 28 ...
##  $ csMPa           : num  80 61.9 40.3 41 44.3 ...
```

```
train$age <- as.numeric(train$age) #convert int to numeric data type

dim(train) #dimension of dataset
```

```
## [1] 700   9
```

```
is.null(train) #returns TRUE if there is any null value and FALSE if no null values
```

```
## [1] FALSE
```

#========================== PART - 1 ==========================#

5. Univariate Analysis:

Each features should be analysed individually using box plots and histograms for numerical variables and bar plots for categorical variables. This will help us in visualizing the distribution and skeweness of the numerical features and the frequency distribution of categorical features. Presence of outliers and anamolies can also be seen from the above plots. Apart from the plots, descriptive statistics of all the columns should be analysed too.

5.1. Plot graphs for all columns

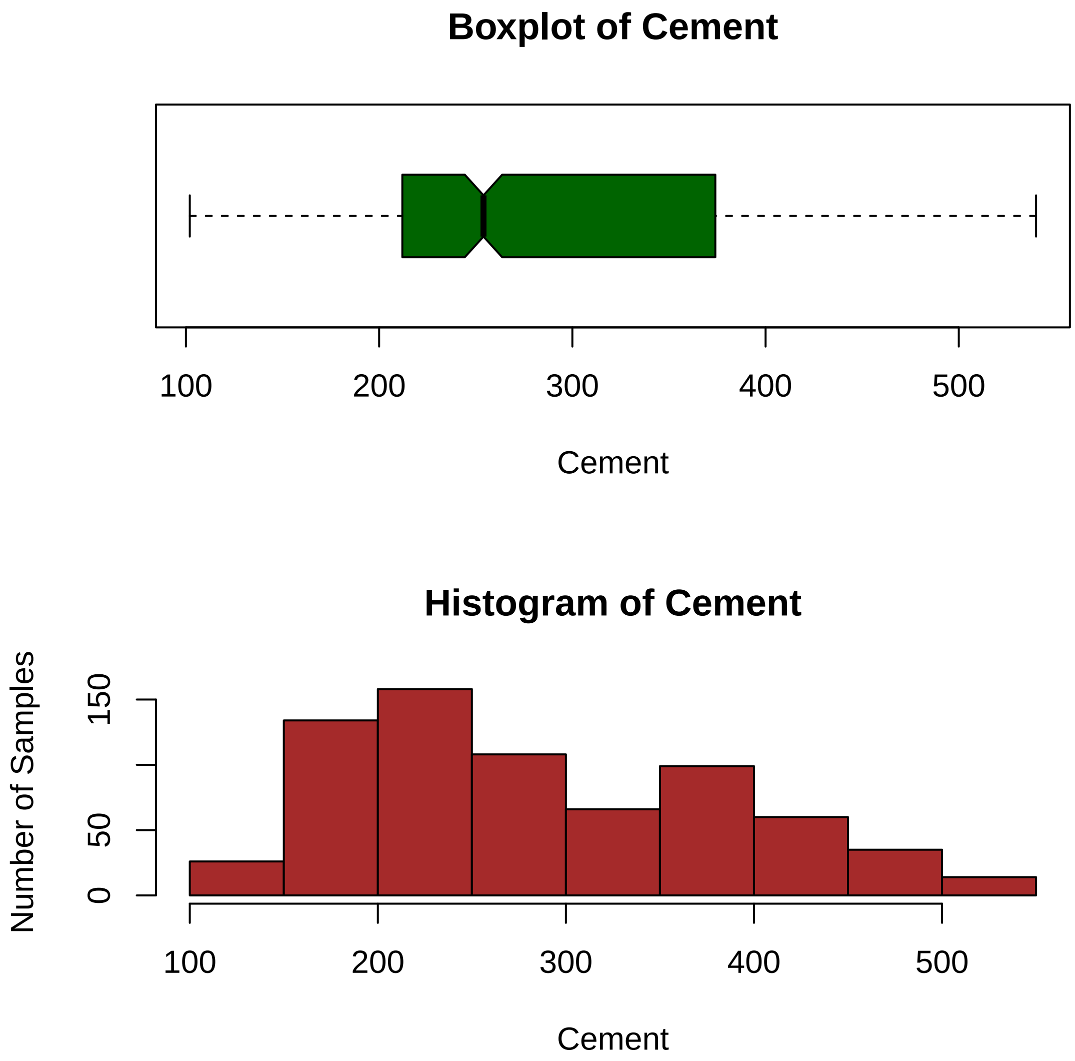
Box plot and Histogram for all numerical variables are plotted below

5.1.1 Feature - Cement

```r
par(mfrow = c(2,1))

boxplot(train$cement,horizontal = T, xlab = 'Cement',
        main = 'Boxplot of Cement',notch=T,
        col = 'dark green', border = 'black')

hist(train$cement,xlab = 'Cement', ylab = 'Number of Samples',
     main = 'Histogram of Cement',
     col = 'brown', border = 'black')
```



```r
summary(train$cement)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   102.0   212.0   254.0   286.9   374.0   540.0
```

```
sd(train$cement) #standard deviation
```

```
## [1] 101.3793
```

```
sd(train$cement)/mean(train$cement) * 100 #coefficient of variation
```

```
## [1] 35.33399
```

```
#install.packages('moments')
library(moments)
skewness(train$cement) #coefficient of skewness
```

```
## [1] 0.4962435
```

```
kurtosis(train$cement) #kurtosis
```

```
## [1] 2.279545
```

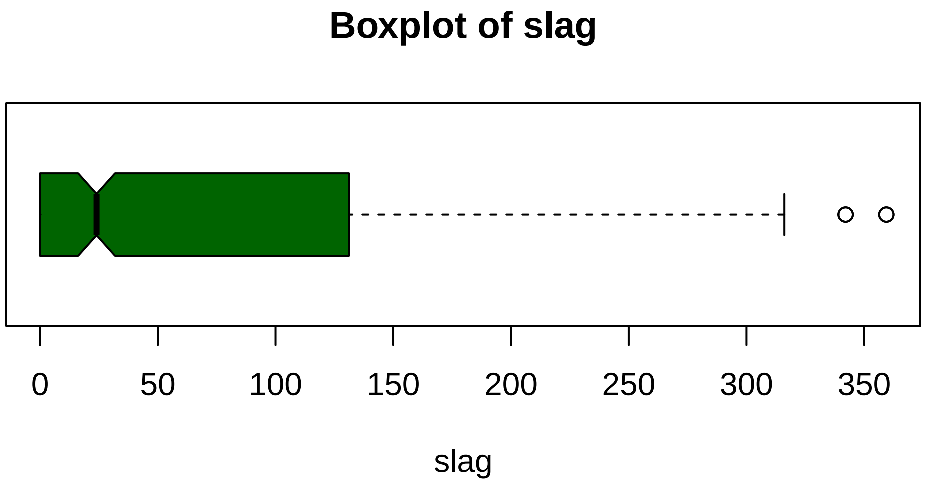
From the above graphs, it is seen that the values in cement columns are a little skewed to the right(positively skewed) without the presence of any outliers. The coefficient of variation is 35.33%. Kurtosis is less than 3 which means the distribution is platykurtic.It is a spread out distribution.

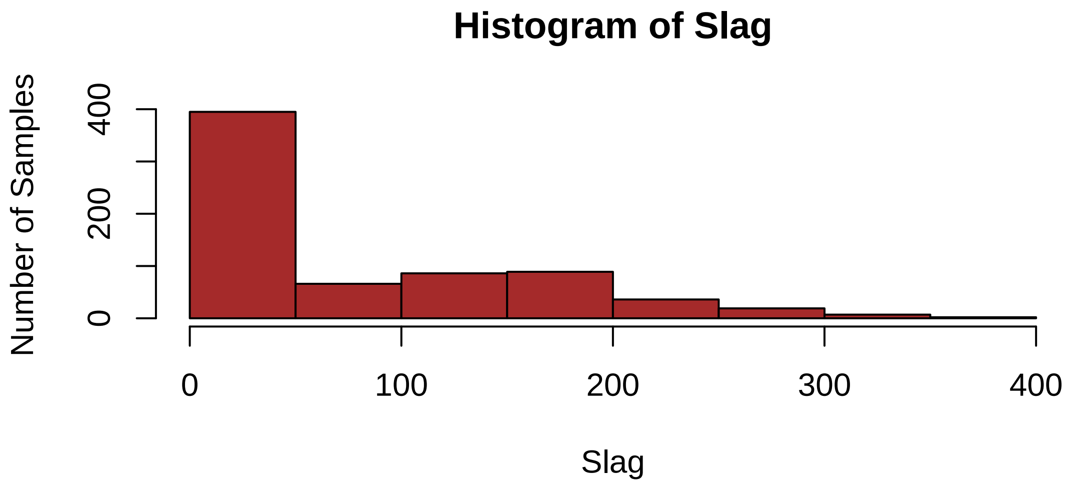
5.1.2 Feature - Slag

```
par(mfrow = c(2,1))

boxplot(train$slag,horizontal = T, xlab = 'slag',
        main = 'Boxplot of slag',notch=T,
        col = 'dark green', border = 'black')

hist(train$slag,xlab = 'Slag', ylab = 'Number of Samples',
     main = 'Histogram of Slag',
     col = 'brown', border = 'black')
```





```
summary(train$slag)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    0.00   24.00   71.99  130.53  359.40
```

```
sd(train$slag) #standard deviation
```

```
## [1] 86.17537
```

```
sd(train$slag)/mean(train$slag) * 100 #coefficient of variation
```

```
## [1] 119.712
```

```
skewness(train$slag) #coefficient of skewness
```

```
## [1] 0.9594471
```

```
kurtosis(train$slag) #kurtosis
```

```
## [1] 2.887892
```

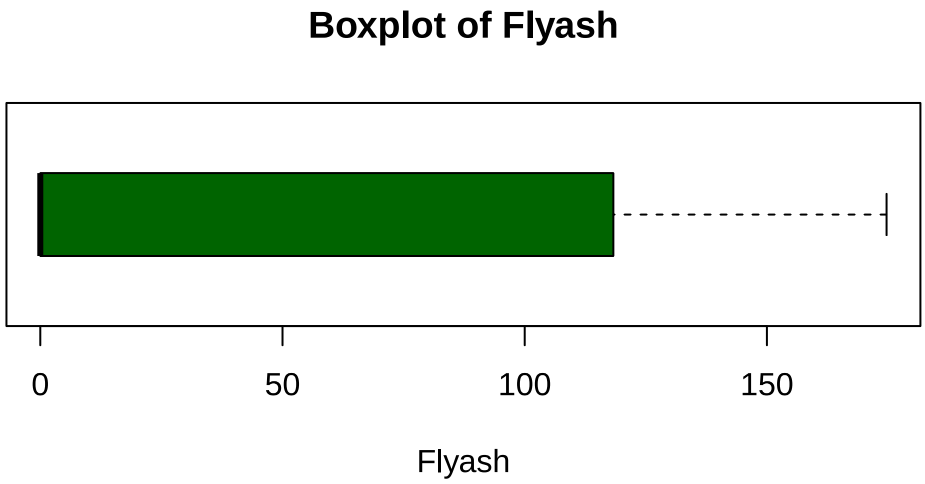
From the above graphs, it is seen that the values in slag columns are more skewed to the right(positively skewed) with presence of outliers. The coefficient of variation is 119.712%. Kurtosis is less than 3 which means the distribution is platykurtic.It is a spread out distribution.

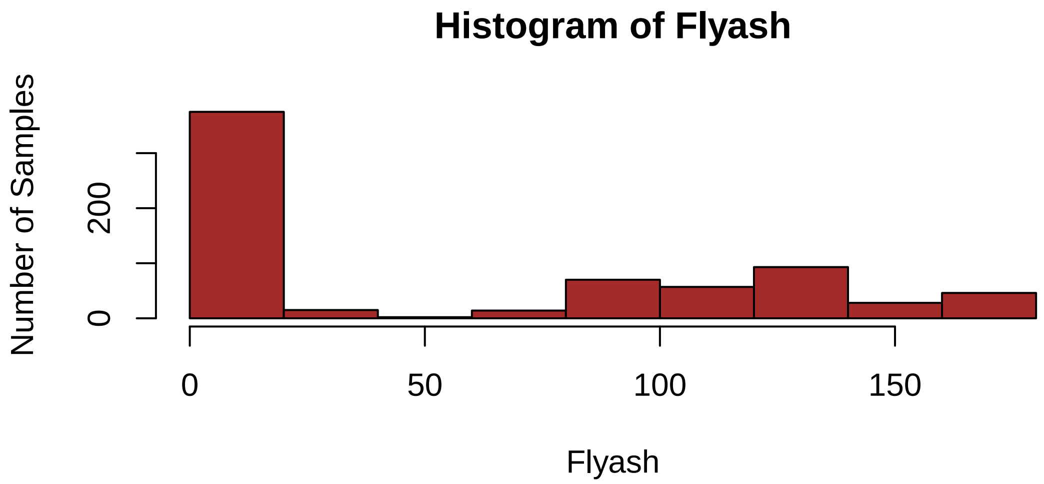
5.1.3 Feature - Flyash

```
par(mfrow = c(2,1))

boxplot(train$flyash,horizontal = T, xlab = 'Flyash',
        main = 'Boxplot of Flyash',
        col = 'dark green', border = 'black')

hist(train$flyash,xlab = 'Flyash', ylab = 'Number of Samples',
     main = 'Histogram of Flyash',
     col = 'brown', border = 'black')
```





```
summary(train$flyash)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    0.00    0.00   54.64  118.30  174.70
```

```
sd(train$flyash) #standard deviation
```

```
## [1] 62.88954
```

```
sd(train$flyash)/mean(train$flyash) * 100 #coefficient of variation
```

```
## [1] 115.095
```

```
skewness(train$flyash) #coefficient of skewness
```

```
## [1] 0.4630034
```

```
kurtosis(train$flyash) #kurtosis
```

```
## [1] 1.540368
```

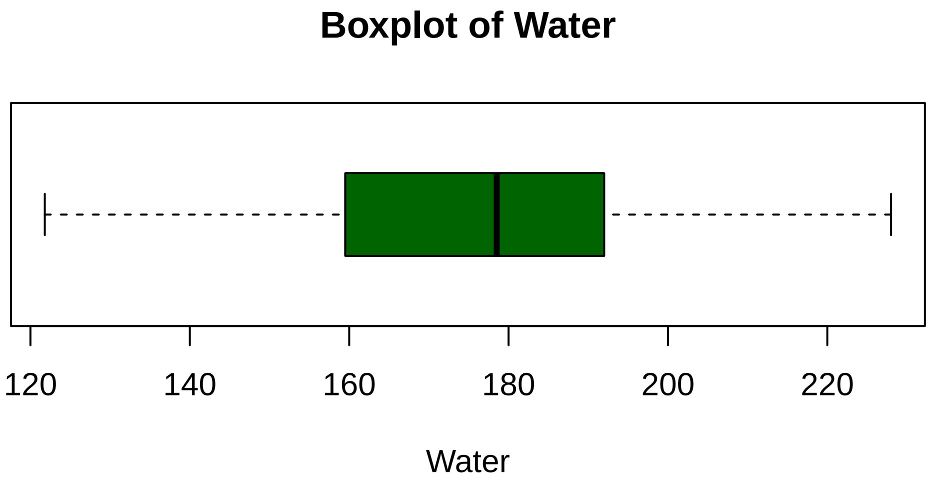
From the above graphs, it is seen that the values in flyash columns are skewed to the right(positively skewed) without the presence of outliers. The coefficient of variation is 115.095%. Kurtosis is less than 3 which means the distribution is platykurtic.It is a spread out and a roughly flat distribution.

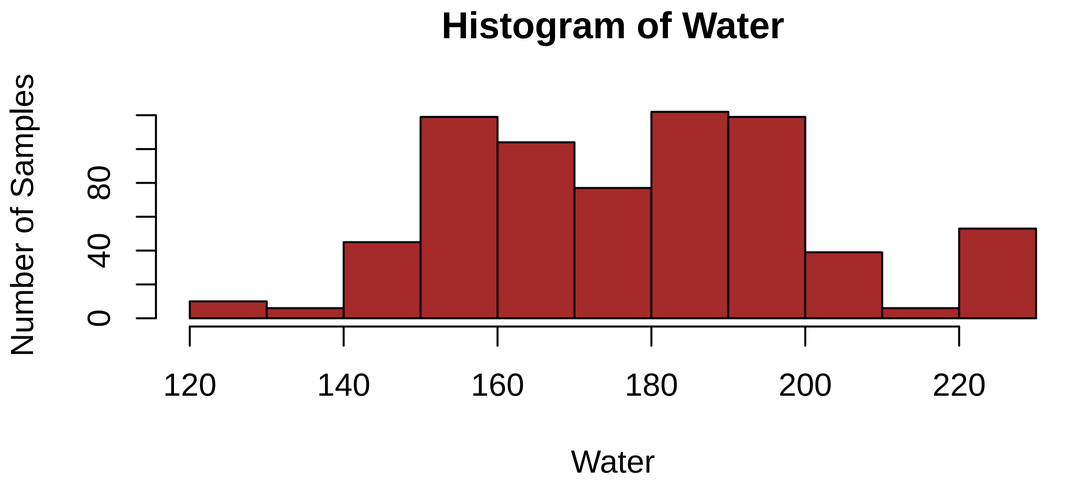
5.1.4 Feature - Water

```
par(mfrow = c(2,1))

boxplot(train$water,horizontal = T, xlab = 'Water',
        main = 'Boxplot of Water',
        col = 'dark green', border = 'black')

hist(train$water,xlab = 'Water', ylab = 'Number of Samples',
     main = 'Histogram of Water',
     col = 'brown', border = 'black')
```





```
summary(train$water)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   121.8   159.5   178.5   178.1   192.0   228.0
```

```
sd(train$water) #standard deviation
```

```
## [1] 23.03016
```

```
sd(train$water)/mean(train$water) * 100 #coefficient of variation
```

```
## [1] 12.93394
```

```
skewness(train$water) #coefficient of skewness
```

```
## [1] 0.3383557
```

```
kurtosis(train$water) #kurtosis
```

```
## [1] 2.84978
```

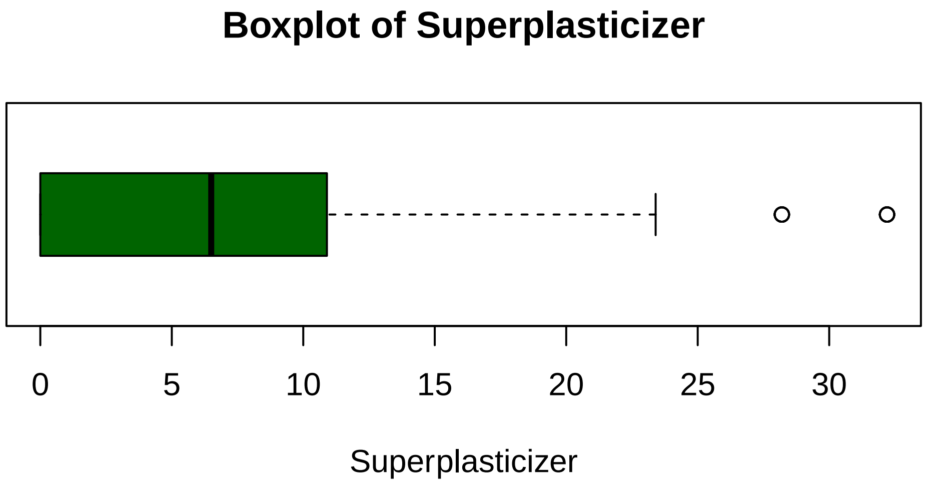
From the above graphs, it is seen that the values in water columns have very little skeweness without the presence of outliers. The coefficient of variation is 12.933%. Kurtosis is nearer to 3 which means the distribution is mesokurtic.It is resembling a normal distribution.

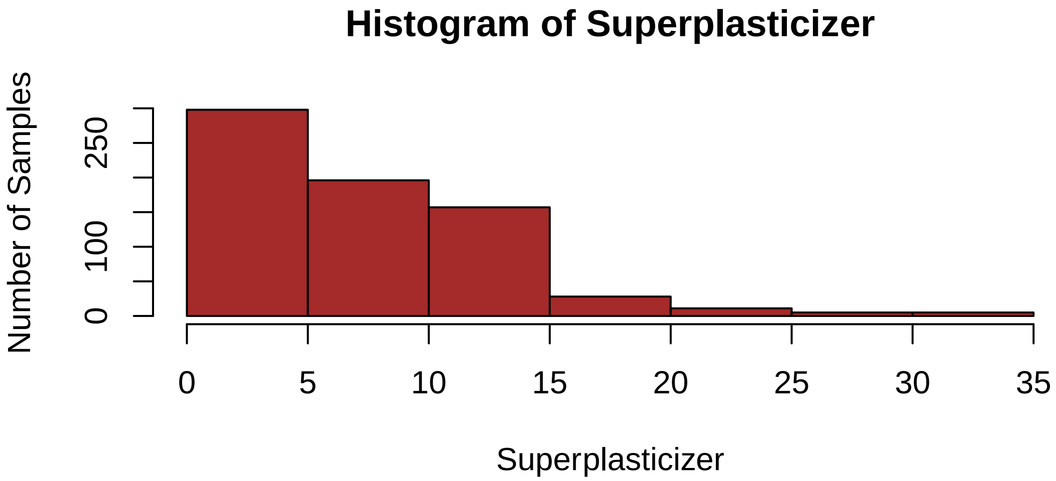
5.1.5 Feature - Super Plasticizer

```
par(mfrow = c(2,1))

boxplot(train$superplasticizer,horizontal = T, xlab = 'Superplasticizer',
        main = 'Boxplot of Superplasticizer', col = 'dark green', border = 'black')

hist(train$superplasticizer,xlab = 'Superplasticizer', ylab = 'Number of Samples',
     main = 'Histogram of Superplasticizer', col = 'brown', border = 'black')
```





```
summary(train$superplasticizer)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   6.500   6.507  10.900  32.200
```

```
sd(train$superplasticizer) #standard deviation
```

```
## [1] 6.289468
```

```
sd(train$superplasticizer)/mean(train$superplasticizer) * 100 #coefficient of variation
```

```
## [1] 96.65907
```

```
skewness(train$superplasticizer) #coefficient of skewness
```

```
## [1] 0.983874
```

```
kurtosis(train$superplasticizer) #kurtosis
```

```
## [1] 4.562749
```

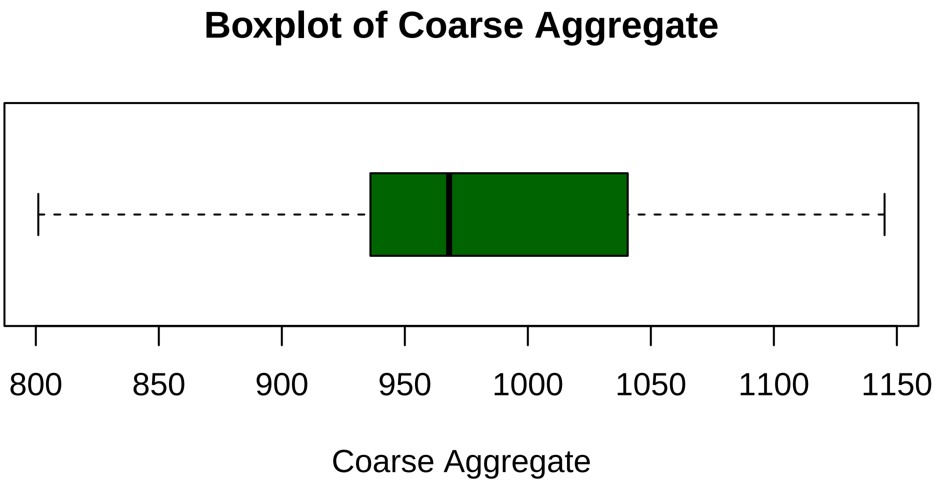
From the above graphs, it is seen that the values in superplasticizer columns are skewed to the right with presence of outliers. The coefficient of variation is 96.65%. Kurtosis is more than 3 which means the distribution is leptokurtic.The distribution has a sharp peak to its left.

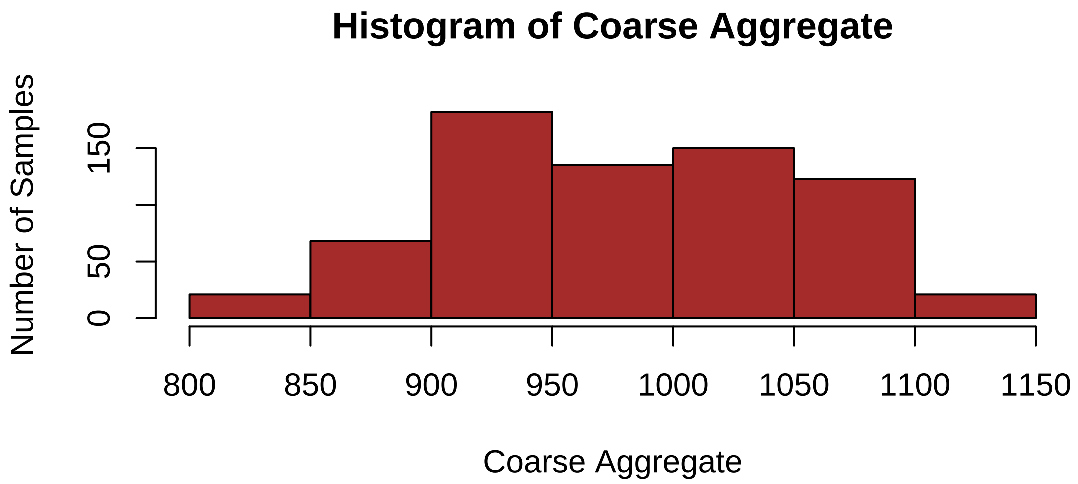
5.1.6 Feature - Coarse aggregate

```
par(mfrow = c(2,1))

boxplot(train$coarseaggregate,horizontal = T, xlab = 'Coarse Aggregate',
        main = 'Boxplot of Coarse Aggregate', col = 'dark green', border = 'black')

hist(train$coarseaggregate,xlab = 'Coarse Aggregate', ylab = 'Number of Samples',
     main = 'Histogram of Coarse Aggregate', col = 'brown', border = 'black')
```





```
summary(train$coarseaggregate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   801.0   936.0   968.0   979.9  1040.6  1145.0
```

```
sd(train$coarseaggregate) #standard deviation
```

```
## [1] 71.45157
```

```
sd(train$coarseaggregate)/mean(train$coarseaggregate) * 100 #coefficient of variation
```

```
## [1] 7.291968
```

```
skewness(train$coarseaggregate) #coefficient of skewness
```

```
## [1] -0.1449645
```

```
kurtosis(train$coarseaggregate) #kurtosis
```

```
## [1] 2.549439
```

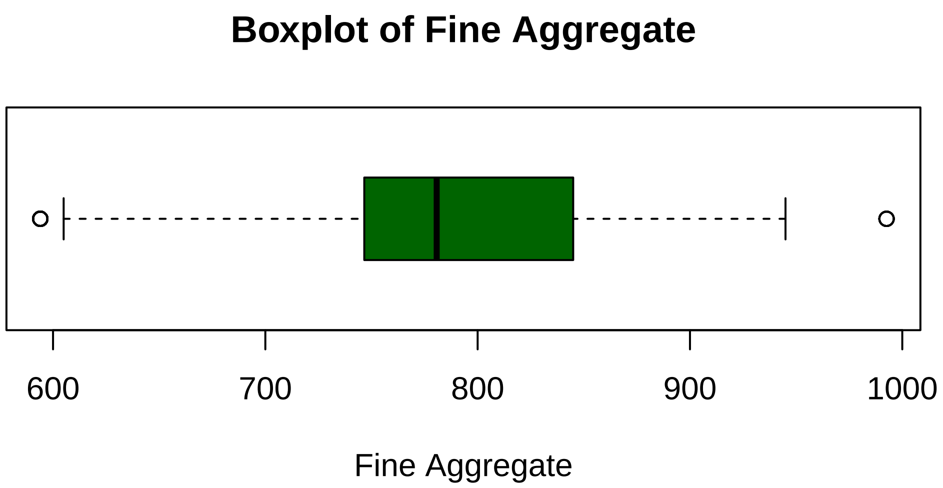
From the above graphs, it is seen that the values in coarseaggregate columns are slightly skewed to the left(negatively skewed) without the presence of outliers. The coefficient of variation is 7.29%. Kurtosis is less than 3 which means the distribution is platykurtic.It is a spread out and a roughly flat distribution.

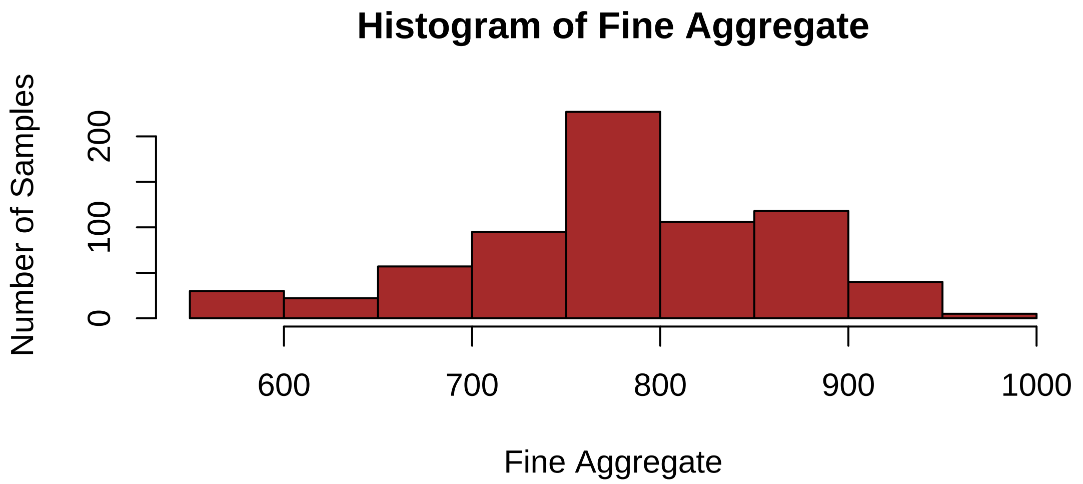
5.1.7 Feature - Fine aggregate

```
par(mfrow = c(2,1))

boxplot(train$fineaggregate,horizontal = T, xlab = 'Fine Aggregate',
        main = 'Boxplot of Fine Aggregate', col = 'dark green', border = 'black')

hist(train$fineaggregate,xlab = 'Fine Aggregate', ylab = 'Number of Samples',
     main = 'Histogram of Fine Aggregate', col = 'brown', border = 'black')
```





```
summary(train$fineaggregate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   594.0   746.6   780.7   781.1   845.0   992.6
```

```
sd(train$fineaggregate) #standard deviation
```

```
## [1] 82.72016
```

```
sd(train$fineaggregate)/mean(train$fineaggregate) * 100 #coefficient of variation
```

```
## [1] 10.59071
```

```
skewness(train$fineaggregate) #coefficient of skewness
```

```
## [1] -0.2861906
```

```
kurtosis(train$fineaggregate) #kurtosis
```

```
## [1] 2.944999
```

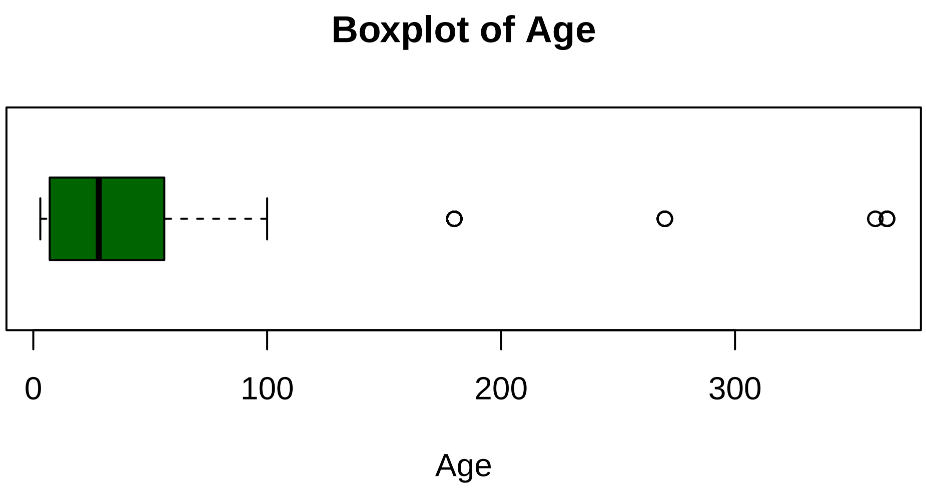
From the above graphs, it is seen that the values in fineaggregate columns are slightly skewed to the left(negatively skewed) with presence of outliers. The coefficient of variation is 10.59%. Kurtosis is nearer to 3 which means the distribution is mesokurtic.It is resembling a normal distribution.

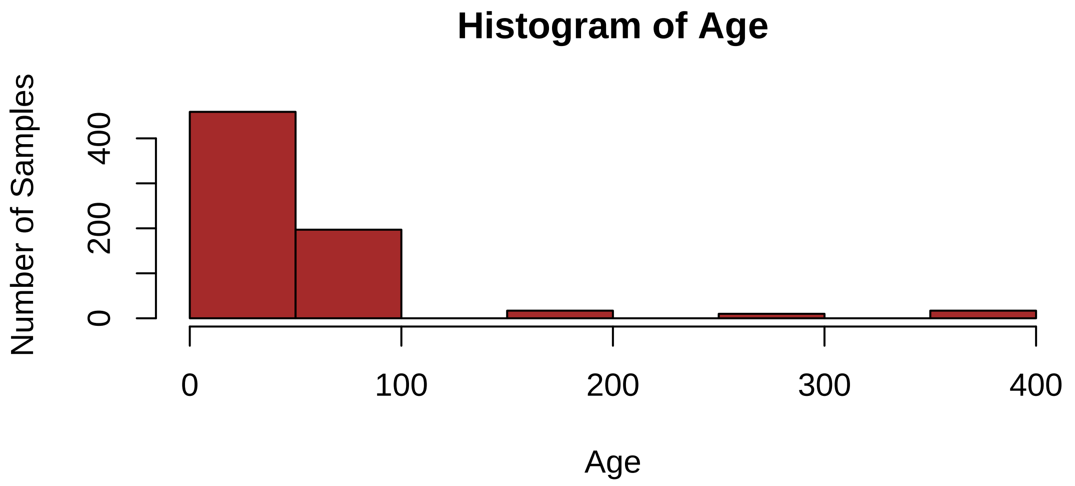
5.1.8 Feature - Age

```
par(mfrow = c(2,1))

boxplot(train$age,horizontal = T, xlab = 'Age',
        main = 'Boxplot of Age', col = 'dark green', border = 'black')

hist(train$age,xlab = 'Age', ylab = 'Number of Samples',
     main = 'Histogram of Age', col = 'brown', border = 'black')
```





```
summary(train$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3.00    7.00   28.00   49.01   56.00  365.00
```

```
sd(train$age) #standard deviation
```

```
## [1] 68.40098
```

```
sd(train$age)/mean(train$age) * 100 #coefficient of variation
```

```
## [1] 139.5653
```

```
skewness(train$age) #coefficient of skewness
```

```
## [1] 2.980642
```

```
kurtosis(train$age) #kurtosis
```

```
## [1] 12.95543
```

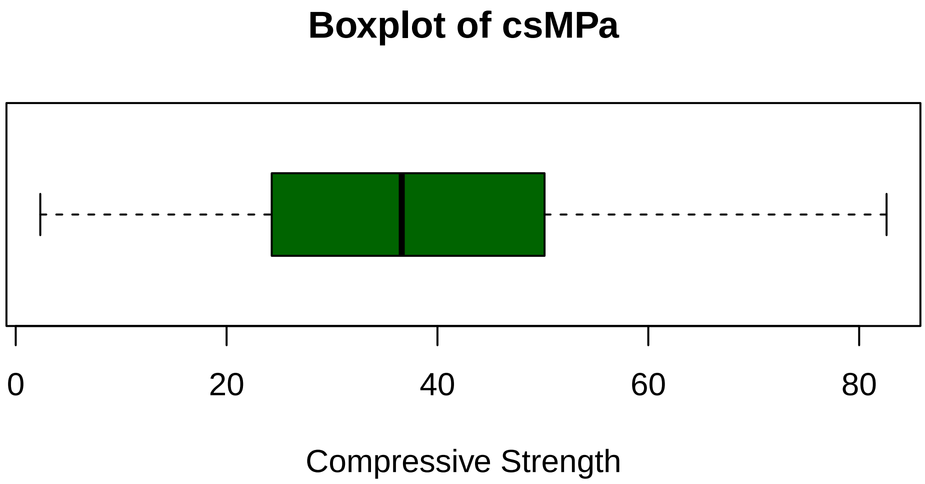
From the above graphs, it is seen that the values in age columns are highly skewed to the right(positively skewed) with the presence of outliers. The coefficient of variation is 139.56%. Kurtosis is much higher than 3 which means the distribution is leptokurtic.The distribution has a sharp peak to its left.

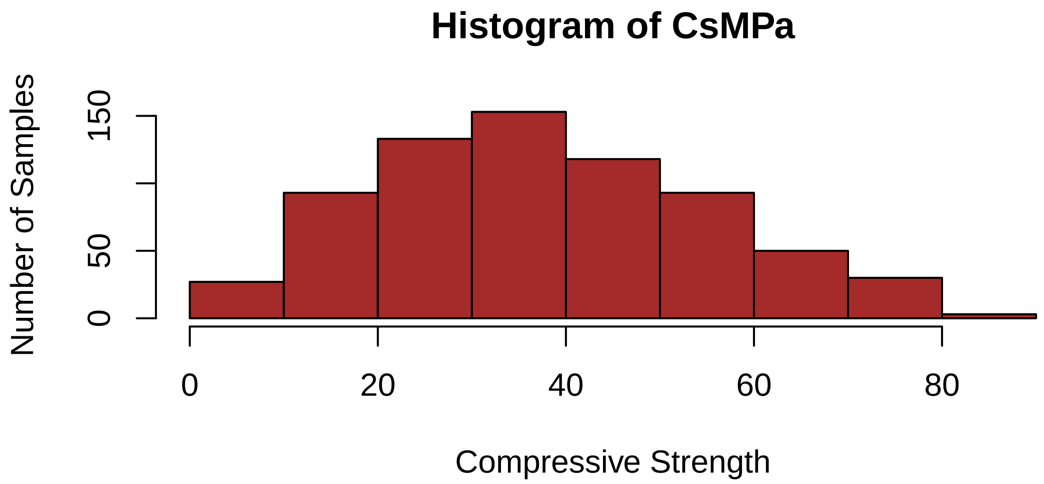
5.1.9 Feature - CsMPa

```
par(mfrow = c(2,1))

boxplot(train$csMPa,horizontal = T, xlab = 'Compressive Strength',
        main = 'Boxplot of csMPa',
        col = 'dark green', border = 'black')

hist(train$csMPa,xlab = 'Compressive Strength', ylab = 'Number of Samples',
     main = 'Histogram of CsMPa', col = 'brown', border = 'black')
```





```
summary(train$csMPa)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.33   24.28   36.62   37.64   50.12   82.60
```

```
sd(train$csMPa) #standard deviation
```

```
## [1] 17.60793
```

```
sd(train$csMPa)/mean(train$csMPa) * 100 #coefficient of variation
```

```
## [1] 46.7767
```

```
skewness(train$csMPa) #coefficient of skewness
```

```
## [1] 0.3173037
```

```
kurtosis(train$csMPa) #kurtosis
```

```
## [1] 2.44881
```

From the above graphs, it is seen that the values in csMPa columns are skewed to the right(positively skewed) without the presence of outliers. The coefficient of variation is 46.77%. Kurtosis is less than 3 which means the distribution is platykurtic.It is spread out and a roughly flat distribution.

5.2 Other Univariate analysis

Some of the descriptive statistics are explored below

```
sort(lengths(lapply(train,unique))) #number of unique values in all columns in ascending order
```

```
##              age           flyash superplasticizer             slag 
##               12               53               74               89 
##            water  coarseaggregate    fineaggregate           cement 
##              108              138              145              155 
##            csMPa 
##              631
```

```
sort(tapply(train$age,train$age, FUN = length)) #count of samples (grouped by age)
```

```
## 360 270 365 180  91  90 100  14  56   7   3  28 
##   3  10  14  17  22  33  52  55  90  97 116 191
```

```
round(sapply(train, function(x) mean(x)),2) #mean of all columns
```

```
##           cement             slag           flyash            water 
##           286.92            71.99            54.64           178.06 
## superplasticizer  coarseaggregate    fineaggregate              age 
##             6.51           979.87           781.06            49.01 
##            csMPa 
##            37.64
```

```
round(sapply(train, function(x) sd(x) / mean(x) * 100),2) #coefficient of variation of all columns
```

```
##           cement             slag           flyash            water 
##            35.33           119.71           115.09            12.93 
## superplasticizer  coarseaggregate    fineaggregate              age 
##            96.66             7.29            10.59           139.57 
##            csMPa 
##            46.78
```

```
quantile(train$csMPa,seq(0.10,1.00,0.10)) #percentile distribution of csMPa
```

```
##    10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
## 14.585 21.852 25.946 32.094 36.615 40.936 46.680 53.596 61.900 82.600
```

```
quantile(train$cement,seq(0.10,1.00,0.10)) #percentile distribution of cement
```

```
##    10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
## 168.90 194.46 213.70 237.50 254.00 297.20 349.30 385.40 427.50 540.00
```

```
quantile(train$slag,seq(0.10,1.00,0.10)) #percentile distribution of slag
```

```
##    10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
##   0.00   0.00   0.00   0.00  24.00  82.96 116.00 157.00 192.15 359.40
```

```
quantile(train$flyash,seq(0.10,1.00,0.10)) #percentile distribution of fly ash
```

```
##   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
##   0.0   0.0   0.0   0.0   0.0  94.1 100.5 123.0 141.0 174.7
```

```
quantile(train$water,seq(0.10,1.00,0.10)) #percentile distribution of water
```

```
##   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
## 153.5 158.1 162.0 169.6 178.5 185.7 191.8 192.9 203.5 228.0
```

```
quantile(train$superplasticizer,seq(0.10,1.00,0.10)) #percentile distribution of superplasticizer
```

```
##  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##  0.0  0.0  0.0  4.5  6.5  8.2  9.9 11.6 12.8 32.2
```

```
quantile(train$coarseaggregate,seq(0.10,1.00,0.10)) #percentile distribution of coarse aggregate
```

```
##     10%     20%     30%     40%     50%     60%     70%     80%     90%    100%
##  882.00  932.00  942.00  956.90  968.00 1004.60 1023.46 1052.30 1075.70 1145.00
```

```
quantile(train$fineaggregate,seq(0.10,1.00,0.10)) #percentile distribution of fine aggregate
```

```
##    10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
## 670.00 715.74 754.30 762.40 780.70 799.14 812.00 856.40 889.00 992.60
```

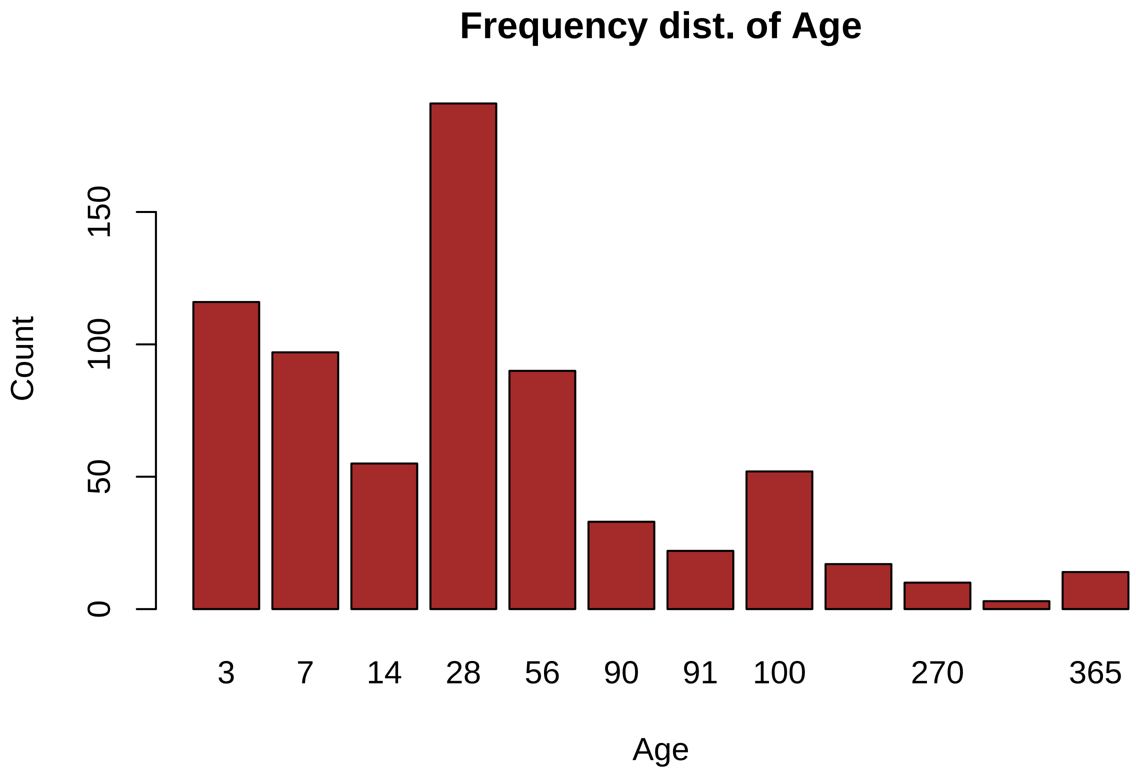
```
quantile(train$age,seq(0.10,1.00,0.10)) #percentile of age
```

```
##  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    3    7    7   28   28   28   56   90  100  365
```

```
round(prop.table(table(train$age))*100,2) #proportion of different age samples
```

```
##
##     3     7    14    28    56    90    91   100   180   270   360   365
## 16.57 13.86  7.86 27.29 12.86  4.71  3.14  7.43  2.43  1.43  0.43  2.00
```

```
barplot(table(train$age),col='brown',border='black',main='Frequency dist. of Age',
        xlab = 'Age', ylab = 'Count') #frequency plot of different age samples
```



As you can see from the above graph, number of observations available after an age of 100 days is very low.

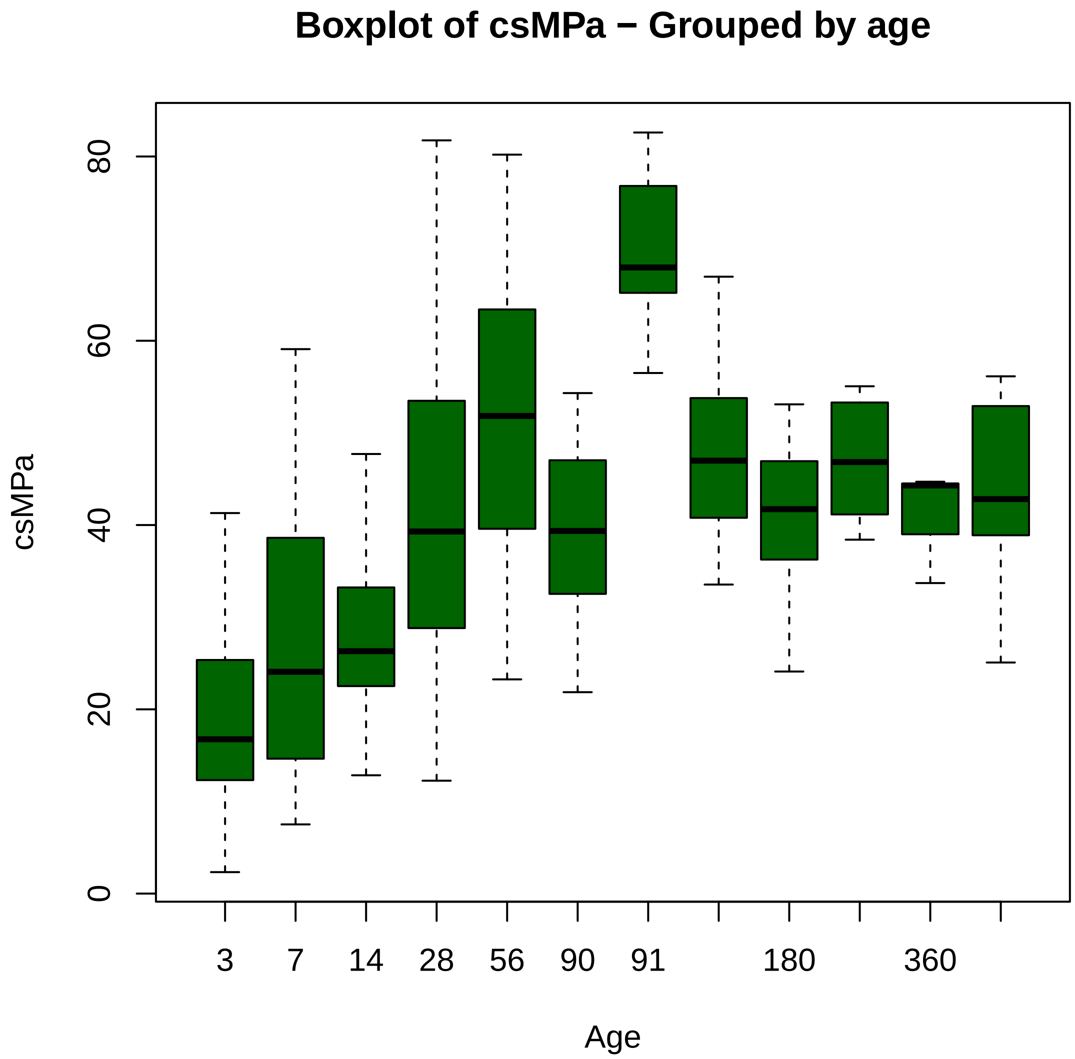
6. Bivariate Analysis:

The features are compared with other features to derive insights on the data. Response of target variable with respect to change in value of predictors is mainly discussed in this section.

6.1. Plot box-plot for csMPa grouped by Age

Trying to find out if the average csMPa is same across all ages of concrete

```
boxplot(train$csMPa~train$age,horizontal = F, xlab = 'Age', ylab = 'csMPa',
        main = 'Boxplot of csMPa - Grouped by age',
        col = 'dark green', border = 'black')
```



We can see from the above graph that the average csMPa is different across different sample age.Also we can infer that:

1. 28 days csMPa has the highest range.
2. First quartile of 56 days csMPa is higher than the second quartile(median) of 28 days csMPa.
3. All the samples are not tested for csMPa in all available days of the 'age' categorical column.
4. csMPa of the highest grade concretes are measured only till 91 days.
5. Minimum value of 91 days csMPa is higher than the maximum value of 90 days csMPa. This is due to the fact that higher grade concretes are tested only till 91 MPa and the lower grade concretes are tested directly on 100th day after testing it on 90th day. Daily testing of concrete samples does not occur in the industry and hence the observations/samples in 90th day & 91st day are mutually exclusive.
6. The average csMPa is roughly the same across different ages post 100 days of curing.
7. Maximum value of csMPa is in the range of 80 MPa.

6.1.1 Statistical test using One-way anova

```
summary(aov(csMPa ~ age, data = train))
```

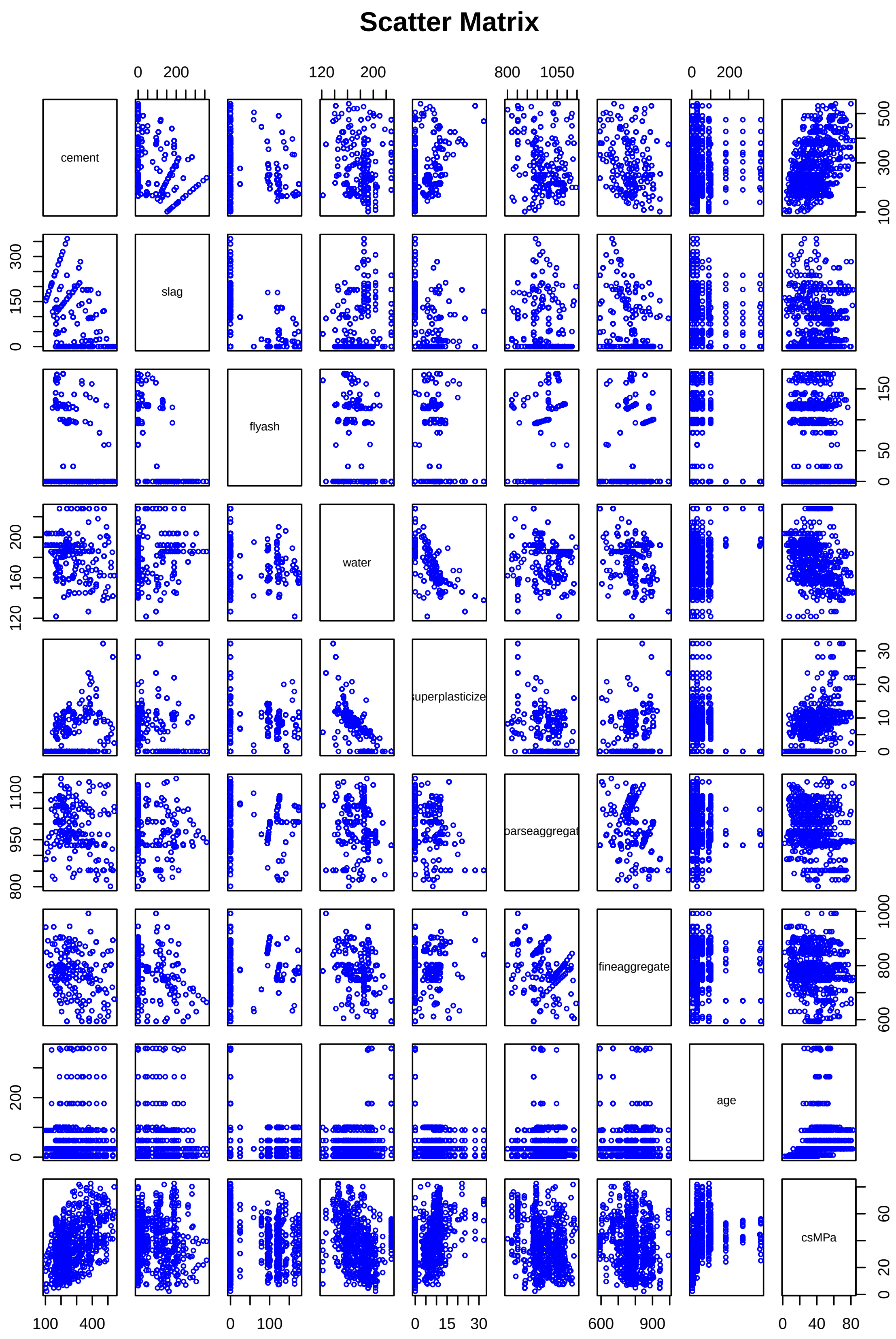
```
##              Df Sum Sq Mean Sq F value Pr(>F)    
## age           1  21170   21170   75.57 <2e-16 ***
## Residuals   698 195547     280                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above summary shows us that the null hypothesis(average csMPa is same across all ages) is rejected in favour of alternate hypothesis at a significance level of 5%.

6.2 Scatter plot & Correlation for all numerical variables

We will get an idea about the strength of linearity through the below scatter plots

```
plot(train,main = 'Scatter Matrix',cex=0.6,col='blue')
```


Scatter Matrix
cement
slag
flyash
water
superplasticize
coarseaggregat
fineaggregate
age
csMPa
0
200
120
200
800
1050
0
200
100
300
500
0
150
300
0
50
150
120
160
200
0
10
20
30
800
950
1100
600
800
1000
0
200
20
60
100
400
0
100
0
15
30
600
900
0
40
80

```
round(cor(train),2) #correlation matrix
```

```
##                   cement  slag flyash water superplasticizer coarseaggregate
## cement              1.00 -0.12  -0.35 -0.07             0.34           -0.46
## slag               -0.12  1.00  -0.50  0.14            -0.06           -0.21
## flyash             -0.35 -0.50   1.00 -0.37             0.25            0.28
## water              -0.07  0.14  -0.37  1.00            -0.77           -0.14
## superplasticizer    0.34 -0.06   0.25 -0.77             1.00           -0.18
## coarseaggregate    -0.46 -0.21   0.28 -0.14            -0.18            1.00
## fineaggregate      -0.22 -0.35   0.17 -0.50             0.30           -0.15
## age                 0.02  0.00  -0.15  0.36            -0.19           -0.10
## csMPa               0.49  0.14  -0.07 -0.28             0.47           -0.26
##                   fineaggregate   age csMPa
## cement                    -0.22  0.02  0.49
## slag                      -0.35  0.00  0.14
## flyash                     0.17 -0.15 -0.07
## water                     -0.50  0.36 -0.28
## superplasticizer           0.30 -0.19  0.47
## coarseaggregate           -0.15 -0.10 -0.26
## fineaggregate              1.00 -0.21 -0.15
## age                       -0.21  1.00  0.31
## csMPa                     -0.15  0.31  1.00
```

From the above plot and correlation matrix we can infer that:

1. There aren't any high correlations between csMPa and other features except for cement, which should be the case for more strength.
2. Age and superplasticizer are the other two features which are strongly correlated with csMPa.
3. Superplasticizer roughly has a negative correlation with water as both are used interchangeably. Also, superplasticizer has positive correlations with fly ash and fine aggregate.

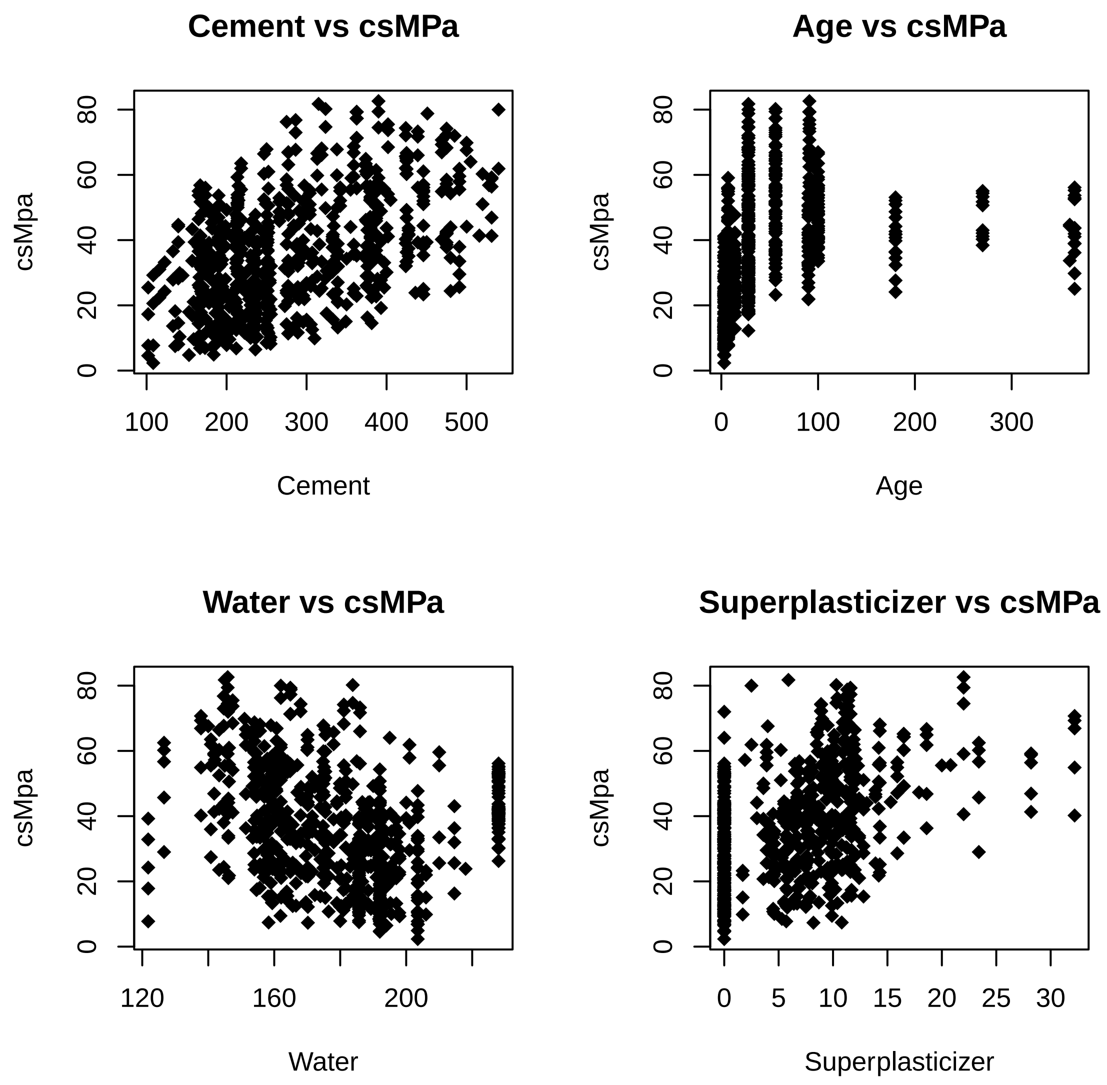
```
par(mfrow = c(2,2))

plot(train$cement,train$csMPa, xlab = 'Cement',ylab = 'csMpa',
        main = 'Cement vs csMPa',col = 'black',cex=1.2,pch=18)

plot(train$age,train$csMPa, xlab = 'Age',ylab = 'csMpa',
        main = 'Age vs csMPa',col = 'black',cex=1.2,pch=18)

plot(train$water,train$csMPa, xlab = 'Water',ylab = 'csMpa',
        main = 'Water vs csMPa',col = 'black',cex=1.2,pch=18)

plot(train$superplasticizer,train$csMPa, xlab = 'Superplasticizer',ylab = 'csMpa',
        main = 'Superplasticizer vs csMPa',col = 'black',cex=1.2,pch=18)
```



The above scatter plots show us that the Concrete strength roughly increases when less water is used in preparing it. Also it roughly increases with increase in age,cement and superplasticizer content.

6.3 Other Bivariate analysis

Cross tab of descriptive statistics are analyzed in this section.

```
round(tapply(train$csMPa,train$age,FUN = mean),2) #Average csMPa across different sample ages
```

```
##     3     7    14    28    56    90    91   100   180   270   360   365
## 19.47 27.74 28.15 41.73 52.13 38.75 69.81 47.67 40.82 47.01 40.90 43.56
```

```
round(tapply(train$csMPa,train$age,FUN = median),2) #Median of csMPa across different sample ages
```

```
##     3     7    14    28    56    90    91   100   180   270   360   365
## 16.75 24.07 26.31 39.30 51.84 39.36 67.95 46.98 41.72 46.83 44.30 42.81
```

```
round(tapply(train$csMPa,train$age,FUN = var),2)  #Variance of csMPa across different sample ages
```

```
##      3      7     14     28     56     90     91    100    180    270    360
##  95.62 226.01  53.72 255.73 201.58  78.79  59.25  70.59  68.99  43.09  38.92
##    365
##  92.55
```

From the above analysis, it is inferred that the average strength of concrete is higher at 91 days in the considered samples.

7. Anomaly/Outlier analysis:

In this section we will find out the presence of soft and hard outliers,if any, in all columns. Necessary outlier correction techniques, if required, will be used to minimize them. Presence of any anomaly will be rectified with business intuitive solutions.

7.1 Soft/Hard Outliers

Detecting the presence of hard/soft outliers, specifying the upper/lower benchmarks and their counts are discussed in this section.

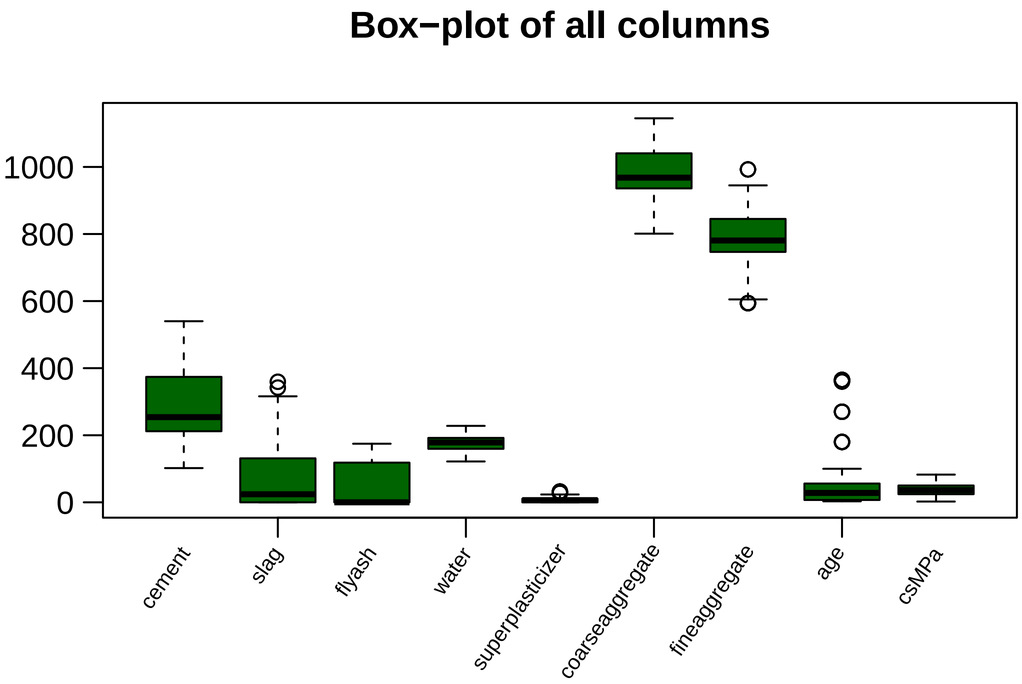
```
#Boxplot of all features

boxplot(train,col='dark green',main='Box-plot of all columns',
        border='black',xaxt='n',yaxt='n')

## Draw x-axis without labels.
axis(side = 1, labels = FALSE)

## Draw y-axis.
axis(side = 2,
     ## Rotate labels perpendicular to y-axis.
     las = 2,
     ## Adjust y-axis label positions.
     mgp = c(3, 0.75, 0))

## Draw the x-axis labels.
text(x = 1:length(train),
     ## Move labels to just below bottom of chart.
     y = par("usr")[3] - 100,
     ## Use names from the data list.
     labels = names(train),
     ## Change the clipping region.
     xpd = NA,
     ## Rotate the labels by 90 degrees.
     srt = 55,
     ## Adjust the labels to almost 100% right-justified.
     adj = 0.965,
     ## Increase label size.
     cex = 0.7)
```



```r
#Benchmark for 'age' hard outliers
ub_age <- train$age[train$age > quantile(train$age,0.75) + 2.5 * IQR(train$age)]
lb_age <- train$age[train$age < quantile(train$age,0.25) - 2.5 * IQR(train$age)]

#Benchmark for 'fine aggregate' soft outliers
ub_fineaggregate <- train$fineaggregate[train$fineaggregate >
            quantile(train$fineaggregate,0.75) + 1.5 * IQR(train$fineaggregate)]
lb_fineaggregate <- train$fineaggregate[train$fineaggregate <
            quantile(train$fineaggregate,0.25) - 1.5 * IQR(train$fineaggregate)]

#Benchmark for 'slag' soft outliers
ub_slag <- train$slag[train$slag > quantile(train$slag,0.75) + 1.5 * IQR(train$slag)]
lb_slag <- train$slag[train$slag < quantile(train$slag,0.25) - 1.5 * IQR(train$slag)]

#Benchmark for 'superplasticizer' soft outliers
ub_superplasticizer <- train$superplasticizer[train$superplasticizer >
      quantile(train$superplasticizer,0.75) + 1.5 * IQR(train$superplasticizer)]

lb_superplasticizer <- train$superplasticizer[train$superplasticizer <
      quantile(train$superplasticizer,0.25) - 1.5 * IQR(train$superplasticizer)]



#Count of Outliers
outlier_age = length(ub_age) + length(lb_age)
outlier_age
```

```
## [1] 44
```

```
outlier_fineaggregate = length(ub_fineaggregate) + length(lb_fineaggregate)
outlier_fineaggregate
```

```
## [1] 35
```

```
outlier_slag = length(ub_slag) + length(lb_slag)
outlier_slag
```

```
## [1] 4
```

```
outlier_superplasticizer = length(ub_superplasticizer) + length(lb_superplasticizer)
outlier_superplasticizer
```

```
## [1] 10
```

```
#Proportion of outliers in each columns
outliers <- round(c(outlier_age,outlier_fineaggregate,outlier_slag,
outlier_superplasticizer)/nrow(train)*100,2)

names(outliers) <- c('age','fineaggregate','slag','superplasticizer')
outliers
```

```
##              age    fineaggregate             slag superplasticizer
##             6.29             5.00             0.57             1.43
```

7.2 Treatment of Outliers

Columns having outlier count of more than 1% of length of dataset are considered for outlier correction. Above results show us that the columns (‘age’,‘fineaggregate’,‘superplasticizer’) fall under this category. We are not going ahead with the outlier treatment of column ‘age’ as that particular feature, business intuitively, can also be considered as a categorical variable. Hence only the columns (‘fineaggregate’,‘superplasticizer’) are treated for outliers using winsorization technique.

```
#Using winzorization technique we replace the outliers with UB or LB whichever is nearer.

#fineaggregate
train$fineaggregate[train$fineaggregate > quantile(train$fineaggregate,0.75) +
1.5 * IQR(train$fineaggregate)] <- quantile(
train$fineaggregate,0.75) + 1.5 * IQR(train$fineaggregate)

train$fineaggregate[train$fineaggregate < quantile(train$fineaggregate,0.25) -
1.5 * IQR(train$fineaggregate)] <- quantile(
train$fineaggregate,0.25) - 1.5 * IQR(train$fineaggregate)


#superplasticizer
train$superplasticizer[train$superplasticizer > quantile(train$superplasticizer,
0.75) + 1.5 * IQR(train$superplasticizer)] <- quantile(
train$superplasticizer,0.75) + 1.5 * IQR(train$superplasticizer)
```

```
train$superplasticizer[train$superplasticizer < quantile(train$superplasticizer,
0.25) - 1.5 * IQR(train$superplasticizer)] <- quantile(
train$superplasticizer,0.25) - 1.5 * IQR(train$superplasticizer)
```

All the necessary outliers are thus treated using winsorization technique. Also, from our analysis, it is observed that the training dataset is free of anamolies.

8. Conclusion:

Exploratory data analysis and data cleaning is completed for the training dataset. The same will be used in python in the next part to fit various models and the csMPa value of test observations will be predicted with minimum errors possible.

```
#Exporting cleaned training dataset as csv file
write.csv(train,'D:\\DS\\Python\\Praxis\\AML\\Project\\train.csv',row.names=F)

#Exporting test dataset as csv file
write.csv(test,'D:\\DS\\Python\\Praxis\\AML\\Project\\test.csv',row.names=F)
```