Group 03

Activity 02

STAT - 31631

Methodology

1.Design

This study employs a quantitative research design to analyze the implications of various concrete mix components on construction quality and sustainability. The primary focus is on understanding how the proportions of different materials and the age of the concrete influence its compressive strength, which is a critical measure of construction quality.

2. Data Collection

- **Sample Selection**: The study will use a dataset containing records of different concrete mixtures with varying proportions of components and their respective compressive strengths.
- **Variables**:
  - **Independent Variables**:
    - Cement (kg in a m^3 mixture)
    - Blast Furnace Slag (kg in a m^3 mixture)
    - Fly Ash (kg in a m^3 mixture)
    - Water (kg in a m^3 mixture)
    - Superplasticizer (kg in a m^3 mixture)
    - Coarse Aggregate (kg in a m^3 mixture)
    - Fine Aggregate (kg in a m^3 mixture)
    - Age (days)
  - **Dependent Variable**:
    - Concrete Compressive Strength (MPa)

3. Experimental Procedure

- **Mix Design Preparation**: Prepare different concrete mixtures by varying the amounts of Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, and Fine Aggregate.
- **Curing Process**: Allow the concrete mixtures to cure over specified time periods to measure the effect of age on compressive strength.

4. **Data Analysis**

Statistical analysis will be performed using R

Descriptive Analysis

Initially, a descriptive analysis will be performed to understand the basic characteristics of the data. This includes:

- Summary statistics (mean, median, mode, standard deviation, etc.)
- Distribution plots for each variable (histograms, box plots, etc.)
- Correlation matrix to identify relationships between variables

Univariate Analysis

Each variable will be analyzed individually to understand its distribution and basic properties. This includes:

- Plotting histograms and density plots
- Calculating measures of central tendency and dispersion
- Identifying any outliers or anomalies

Multiple Linear Regression Analysis

Full Model

A multiple linear regression model will be developed with Concrete Compressive Strength as the response variable and all other variables as predictors. The steps include:

- Fitting the full model
- Checking the overall model fit (R^2, Adjusted R^2, etc.)
- Analyzing the significance of individual predictors using p-values

Residual Analysis

- Examining residual plots to check for homoscedasticity
- Analyzing the normality of residuals using Q-Q plots
- Identifying any patterns or trends in the residuals

Variable Selection

To improve the model, variable selection techniques will be employed. This includes:

- Stepwise regression (both forward and backward selection)
- Comparing the models using criteria such as AIC, BIC, and Adjusted R^2

Model Comparison and Selection

Models with different subsets of variables will be compared to select the one with the highest predictive accuracy. This involves:

- Calculating prediction errors (RMSE, MAE, etc.)
- Performing cross-validation to evaluate model stability
- Selecting the best model based on predictive performance and interpretability

Residual Analysis of the Final Model

A detailed residual analysis will be conducted for the selected model to ensure it meets all the assumptions of linear regression:

- Checking for homoscedasticity, normality, and independence of residuals
- Identifying points with high leverage or influence using Cook's distance

Model Validation

The final model will be validated using an independent dataset (if available) or by splitting the original dataset into training and testing sets. This includes:

- Evaluating the model's performance on the testing set
- Comparing the predictions with actual values to assess accuracy

5.Limitations

- The study is limited to the available data on concrete mix proportions and compressive strength.
- The generalization of findings may be restricted due to variations in concrete production processes and environmental conditions.

6.Implications for Construction Quality and Sustainability

Finally, the results will be interpreted in the context of construction quality and sustainability. This includes:

- Discussing the impact of each significant predictor on concrete compressive strength
- Providing recommendations for optimizing concrete mixtures for better quality and sustainability
- Highlighting any limitations of the study and suggesting areas for future research

Descriptive Analysis

GROUP_03

2024-07-28

```r
#Import the dataset(Copy of Concrete_Data)

Data<-read.csv("C:\\Users\\User\\OneDrive\\Desktop\\Copy of Concrete_Data.csv
")

colnames(Data)<-c("Cement","Blast_Furnace_Slag","Fly_Ash","Water","Superplast
icizer","Coarse_Aggregate","Fine_Aggregate","Age_day","Concrete_compressive_s
trength")

names(Data)
```

```
## [1] "Cement"                         "Blast_Furnace_Slag"
## [3] "Fly_Ash"                        "Water"
## [5] "Superplasticizer"               "Coarse_Aggregate"
## [7] "Fine_Aggregate"                 "Age_day"
## [9] "Concrete_compressive_strength"
```

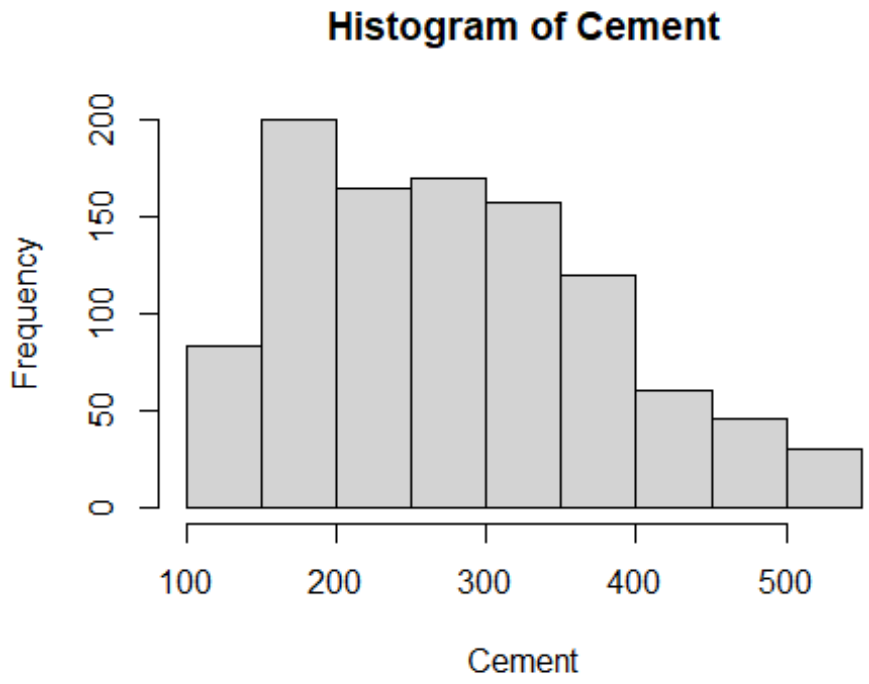
```r
head(Data)
```

```
##   Cement Blast_Furnace_Slag Fly_Ash Water Superplasticizer Coarse_Aggregat
e
## 1  540.0               0.0       0   162              2.5           1040.
0
## 2  540.0               0.0       0   162              2.5           1055.
0
## 3  332.5             142.5       0   228              0.0            932.
0
## 4  332.5             142.5       0   228              0.0            932.
0
## 5  198.6             132.4       0   192              0.0            978.
4
## 6  266.0             114.0       0   228              0.0            932.
0
##   Fine_Aggregate Age_day Concrete_compressive_strength
## 1          676.0      28                         79.99
## 2          676.0      28                         61.89
## 3          594.0     270                         40.27
## 4          594.0     365                         41.05
## 5          825.5     360                         44.30
## 6          670.0      90                         47.03
```

```r
###Missing values

print("Total of Missing values")

## [1] "Total of Missing values"

sum(is.na(Data))

## [1] 0

colSums(is.na(Data))

##                         Cement            Blast_Furnace_Slag
##                              0                             0
##                        Fly_Ash                         Water
##                              0                             0
##              Superplasticizer              Coarse_Aggregate
##                              0                             0
##                Fine_Aggregate                       Age_day
##                              0                             0
## Concrete_compressive_strength
##                              0
```

This data set has no missing values.

```r
##########Descriptive Analysis##########

#################Cement############

##Summary

summary(Data$Cement)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   102.0   192.4   272.9   281.2   350.0   540.0

##Histogram
hist(Data$Cement,xlab = "Cement",main="Histogram of Cement")
```



Interpretation :-

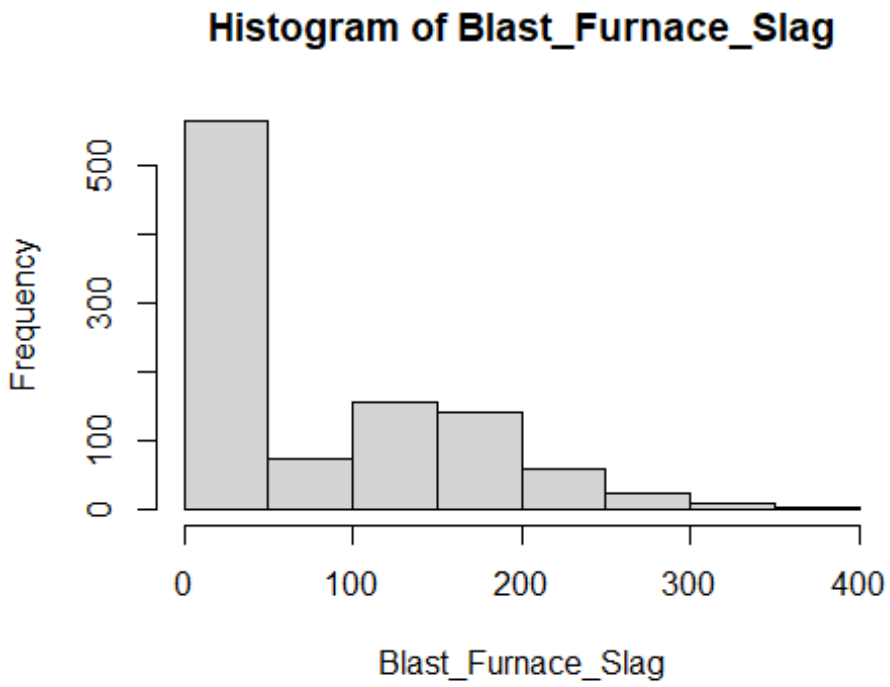
*. There is a peak in the cement content distribution between 150 and 200 units, which is right-skewed.

*. The values of cement content are roughly between 100 and 550 units.

*. As the value rises, the frequency of cement content falls, suggesting a greater percentage of samples with lower cement concentration.

*. A few possible outliers on the higher end of the cement content range may exist, and they need additional research.

```
##################Blast_Furnace_Slag#############

##Summary

summary(Data$Blast_Furnace_Slag)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0    22.0    73.9   142.9   359.4

##Histogram
hist(Data$Blast_Furnace_Slag,xlab = "Blast_Furnace_Slag",main="Histogram of B
last_Furnace_Slag")
```



Interpretation :-

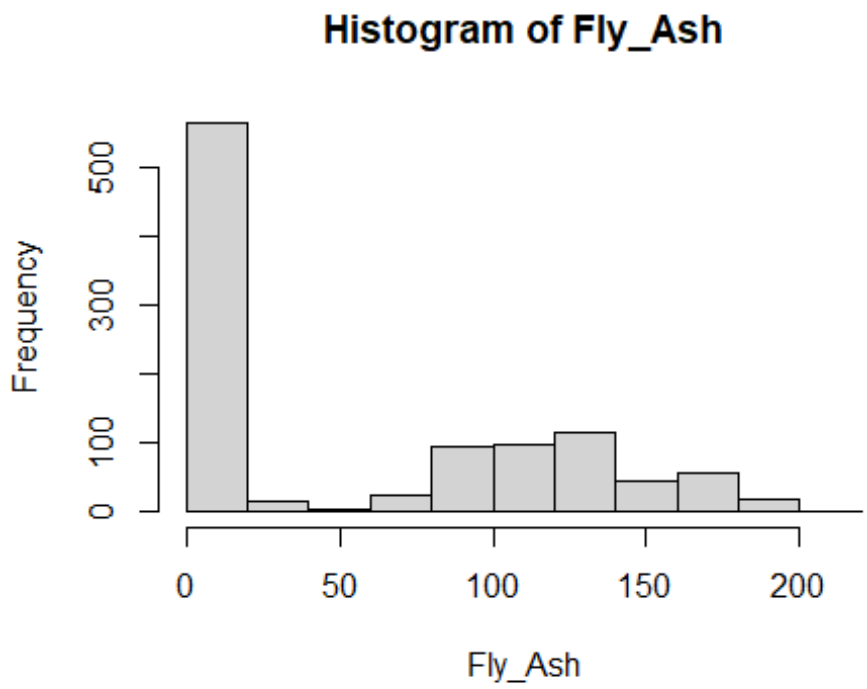
*. The data has a lengthy tail that points towards greater amounts of blast furnace slag and is biased to the right (positively skewed).

*. Most samples fell between 0 and 100, suggesting that this is the most common range for blast furnace slag concentration.

*. With a few outliers going above 300, the Blast Furnace Slag content ranges from roughly 0 to about 400.

*. There appears to be a single main concentration of data, as indicated by the unimodal histogram.

```
##################Fly_Ash#############

##Summary

summary(Data$Fly_Ash)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   54.19  118.30  200.10

##Histogram
hist(Data$Fly_Ash,xlab = "Fly_Ash",main="Histogram of Fly_Ash")
```



Interpretation:-

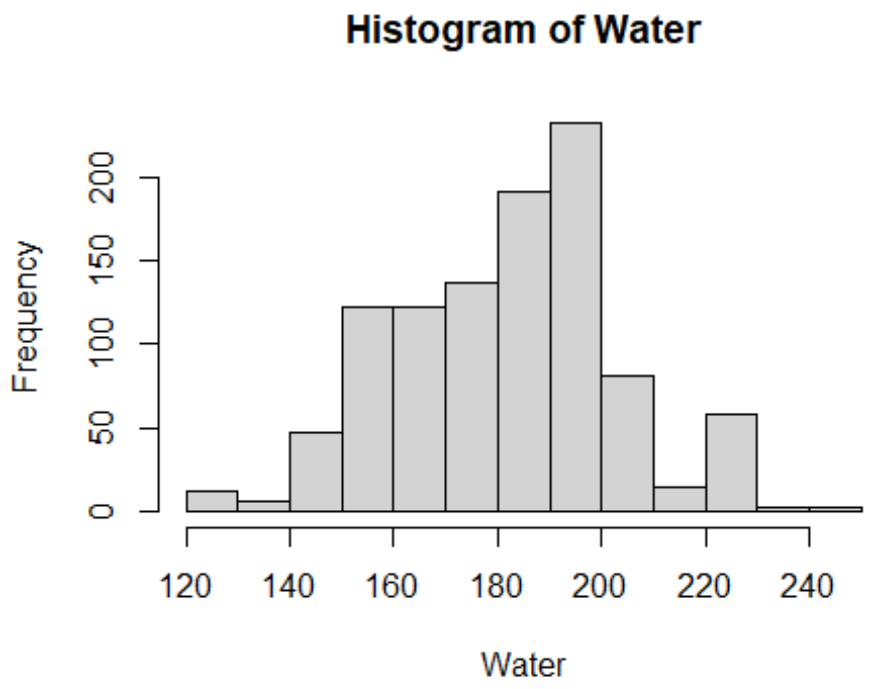
∗. A long tail points in the direction of higher values, and the distribution is biased to the right.

∗. Fly_Ash concentrations range from 0 to 50 units in most concrete samples.

∗. Increasing value results in a decrease in the frequency of fly_ash content.

∗. Several samples may be deemed outliers due to their extremely high fly_ash composition.

```r
##################Water#############

##Summary

summary(Data$Water)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   121.8   164.9   185.0   181.6   192.0   247.0

##Histogram
hist(Data$Water,xlab = "Water",main="Histogram of Water")
```



Interpretation :-

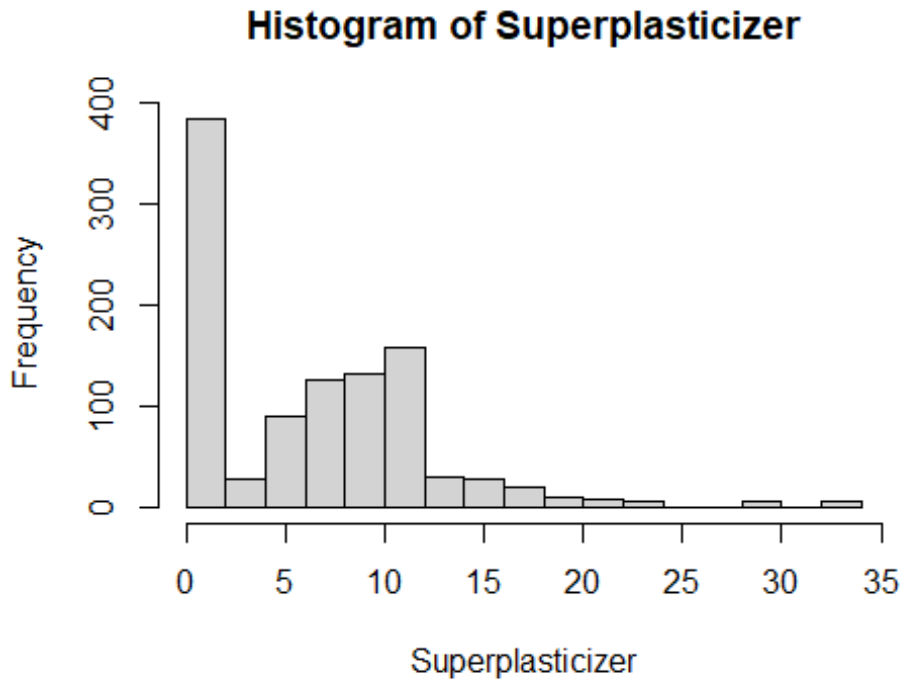
*. The histogram shows the distribution of water amounts used in different concrete mixes.

*. The most common water amount is between 180 and 200 units.

*. The distribution is slightly skewed to the right, indicating that a few concrete mixes used significantly higher amounts of water.

*. There are a few outliers with very low water usage.

```
###############Superplasticizer############

##Summary

summary(Data$Superplasticizer)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   6.400   6.205  10.200  32.200

##Histogram
hist(Data$Superplasticizer,xlab = "Superplasticizer",main="Histogram of Super
plasticizer")
```



Interpretation :-

*. There is a lengthy tail pointing towards higher values and a skew to the right in the data.

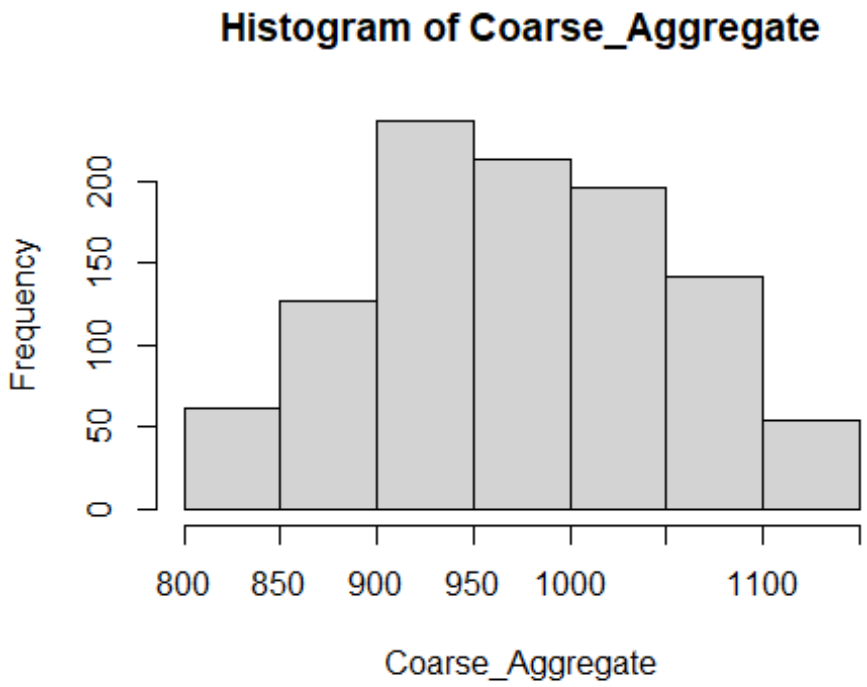
*. For superplasticizer content, most data values lie between 0 and 5.

*. There are a few outliers that fall outside of the range of 0 to 35 for the superplasticizer content.

*. There is some fluctuation in the distribution, and it is not entirely symmetrical.

*. In the 0–5 range, there seems to be just one mode, or peak.

```r
###############Coarse_Aggregate############

##Summary

summary(Data$Coarse_Aggregate)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   801.0   932.0   968.0   972.9  1029.4  1145.0

##Histogram
hist(Data$Coarse_Aggregate,xlab = "Coarse_Aggregate",main="Histogram of Coars
e_Aggregate")
```



Interpretation :-

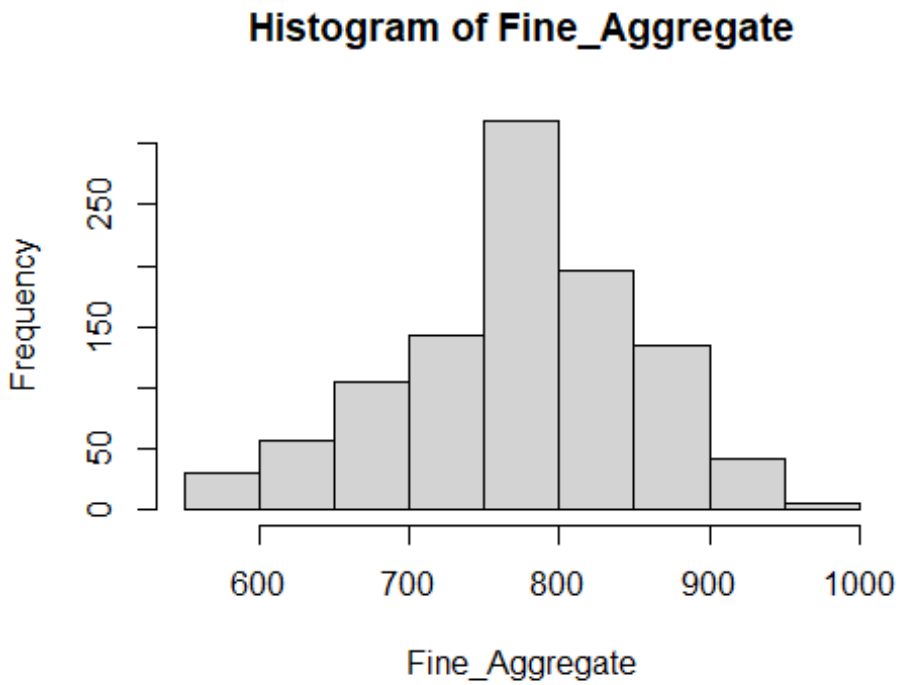
*. A normal distribution of coarse aggregate sizes in the concrete mix is suggested by the histogram, which has a bell-shaped approximate form.

*. The distribution seems to have a central tendency in the 900–950 range, suggesting that most of the coarse aggregates are in this size range.

*. There appears to be some variance in the aggregate sizes, with the data spreading fairly widely.

*. The data shows no obvious outliers, indicating that the coarse aggregate sizes utilized in the concrete mix are within a regular range.

```r
###############Fine_Aggregate############

##Summary

summary(Data$Fine_Aggregate)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   594.0   731.0   779.5   773.6   824.0   992.6

##Histogram
hist(Data$Fine_Aggregate,xlab = "Fine_Aggregate",main="Histogram of Fine_Aggr
egate")
```



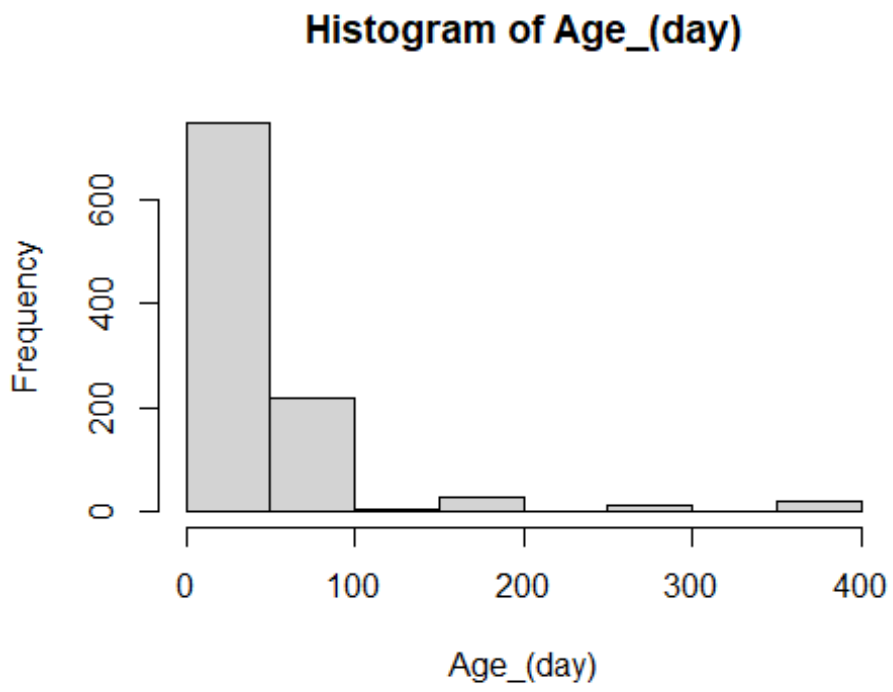
Interpretation :-

*. There are more samples with lower fine aggregate values than higher ones, indicating a right-skewed (or positively skewed) distribution of fine aggregate values.

*. The data seems to be concentrated in the 750–800 range, indicating that this could be the dataset's typical or average fine aggregate value.

*. The values of fine aggregate fall roughly between 600 and 1000.

```r
###############`Age_day`############

##Summary

summary(Data$Age_day)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   28.00   45.66   56.00  365.00

##Histogram
hist(Data$Age_day,xlab = "Age_(day)",main="Histogram of Age_(day)")
```



Interpretation :-

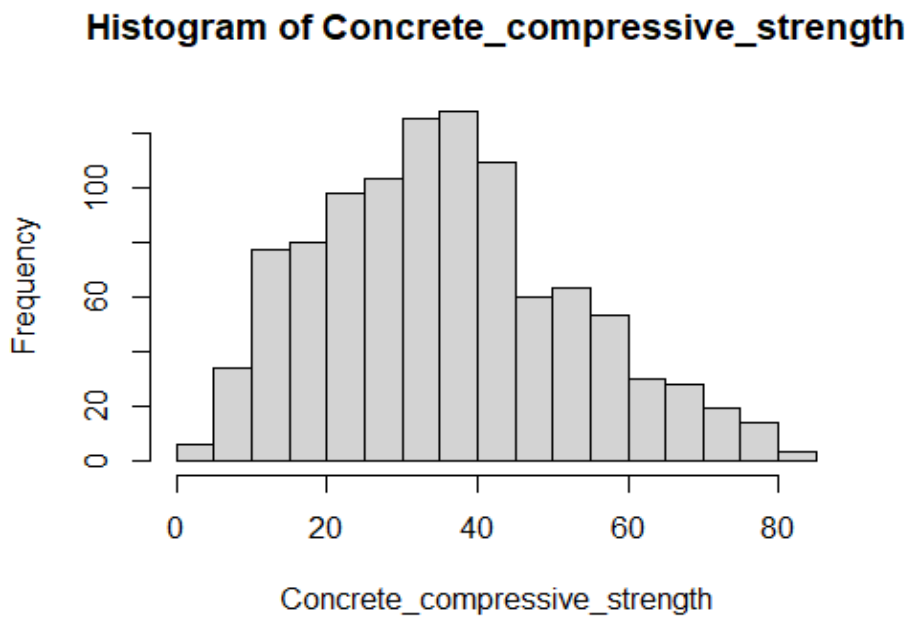
*. There is a significant leftward bias in the data, with many samples falling within the beginning age range of 0–100 days.

*. As age rises, sample frequency rapidly declines, suggesting a reduction in samples in older age groups.

*. The histogram shows a lengthy tail to the right, indicating that some concrete samples are considerably older than the others.

*. A small number of samples that fall into the older age range (beyond 300 days) may indicate the existence of possible outliers that need closer examination.

```r
###############Concrete_compressive_strength############

##Summary

summary(Data$Concrete_compressive_strength)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.33   23.71   34.45   35.82   46.13   82.60

##Histogram
hist(Data$Concrete_compressive_strength,xlab = "Concrete_compressive_strength
",main="Histogram of Concrete_compressive_strength")
```



Interpretation :-

∗. The histogram is roughly bell-shaped, suggesting that the compressive strength values are approximately normally distributed.

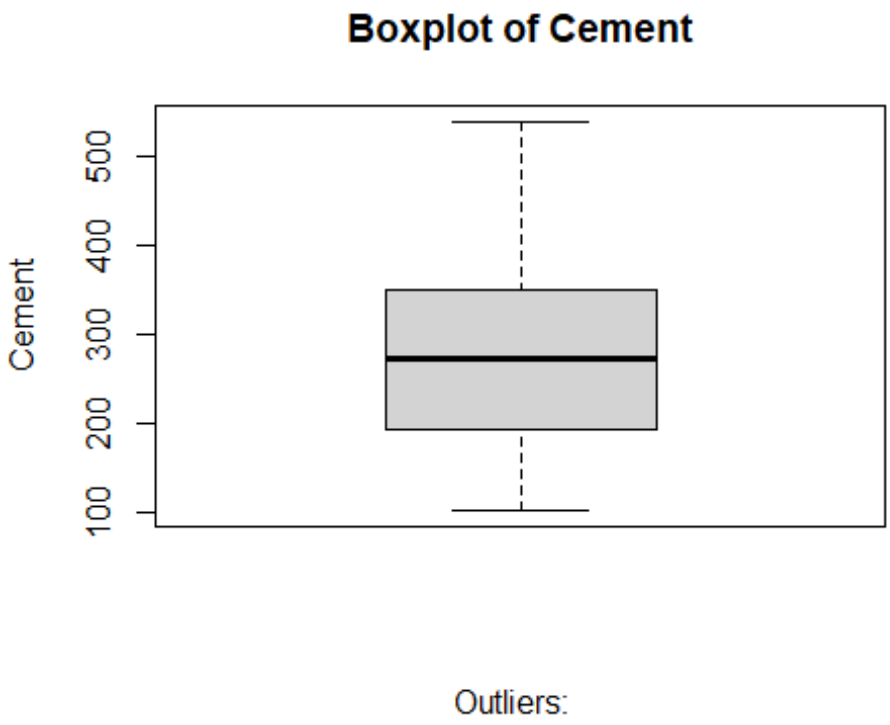
∗. The distribution appears to be centered around the 30-40 MPa range, indicating that a majority of the concrete samples have compressive strengths within this range.

∗. The data is spread across a range of approximately 0 to 80 MPa, with a visible concentration between 20 and 60 MPa.

∗. The frequency of concrete samples decreases as the compressive strength moves away from the central range, both towards lower and higher values.

∗. There are a few data points with very low compressive strength (below 10 MPa), which could be considered potential outliers.

```r
####Cement
##Boxplot
boxplot(Data$Cement,ylab="Cement",main="Boxplot of Cement",sub=paste("Outlier
s: ",boxplot.stats(Data$Cement)$out))
```



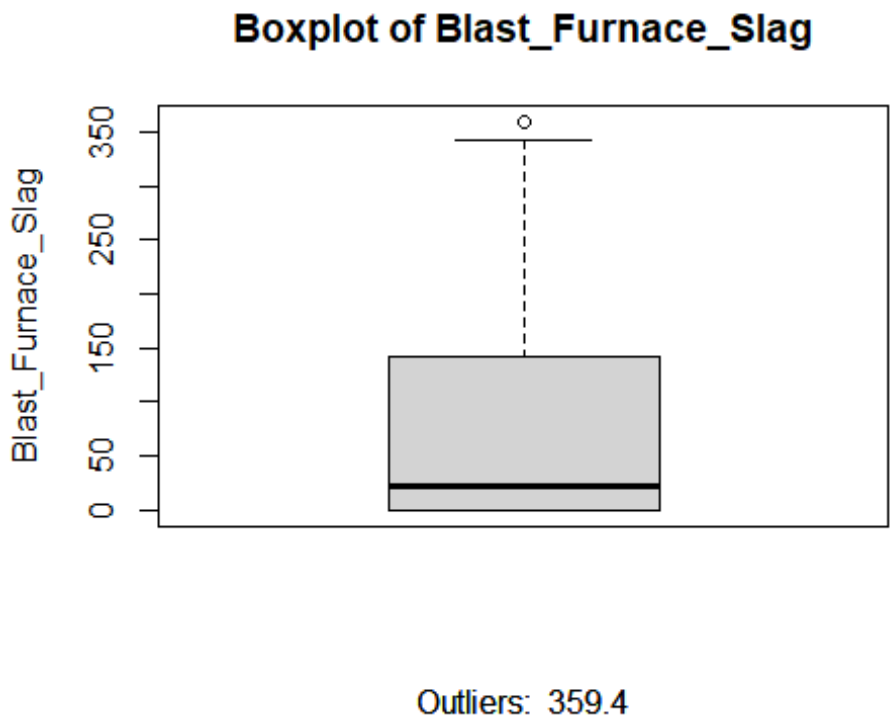
Interpretation :-

*. The median amount of cement used falls between 200 and 300 units.

*. The data is moderately spread out, with the box representing the interquartile range (IQR) extending from approximately 150 to 350 units.

*. The overall range of cement usage is substantial, extending from around 100 to 500 units.

```r
#This boxplot seems that no outliers

#Blast_Furnace_Slag
##Boxplot
boxplot(Data$Blast_Furnace_Slag,ylab="Blast_Furnace_Slag",main="Boxplot of Bl
ast_Furnace_Slag",sub=paste("Outliers: ",boxplot.stats(Data$Blast_Furnace_Sla
g)$out))
```



Interpretation :-

*. The median Blast Furnace Slag value is approximately 125.

*. The data is positively skewed, with a longer tail towards higher values.

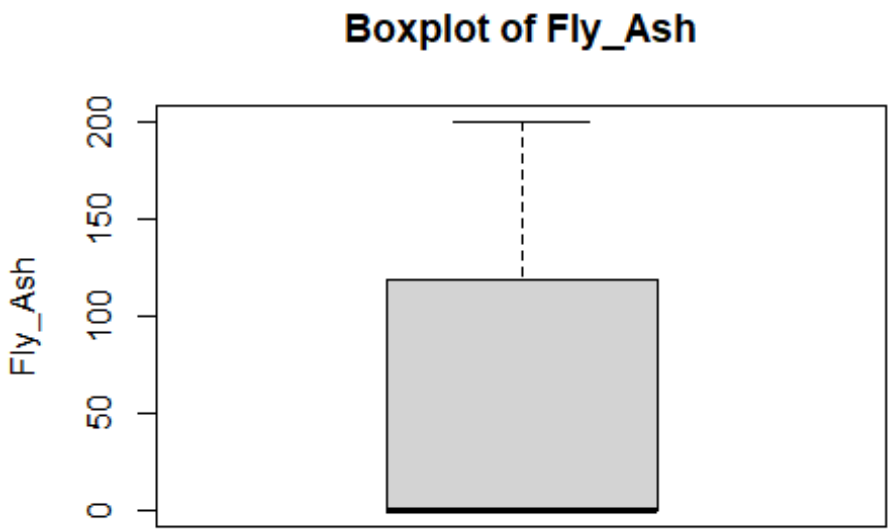
*. The interquartile range (IQR) is approximately 75, indicating moderate variability in the Blast Furnace Slag content.

*. One outlier is present at 359.4, which is significantly higher than the rest of the data.

*. The range of Blast Furnace Slag values extends from approximately 50 to 359.4.

```r
#This box plot seems that one outlier point.
#Remove the outlier row.
outlier_rows_Blast_Furnace_Slag<-boxplot.stats(Data$Blast_Furnace_Slag)$out

#Clean Outliers
Data_cleaned_1<-subset(Data,!(Blast_Furnace_Slag%in%outlier_rows_Blast_Furnac
e_Slag))
```

```r
##Fly_Ash
##Boxplot
boxplot(Data_cleaned_1$Fly_Ash,ylab="Fly_Ash",main="Boxplot of Fly_Ash",sub=p
aste("Outliers: ",boxplot.stats(Data_cleaned_1$Fly_Ash)$out))
```



Interpretation :-

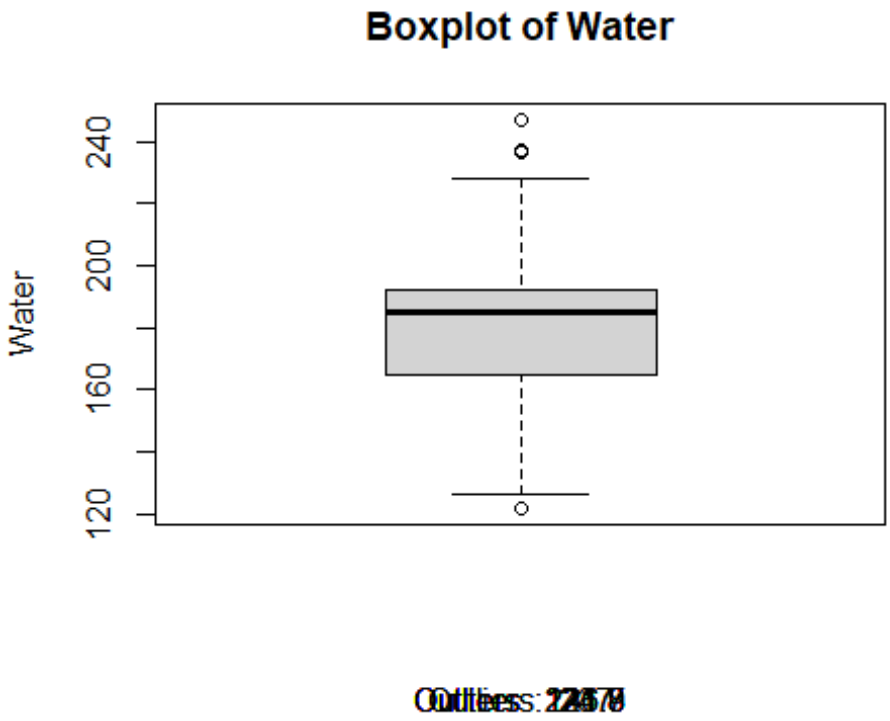
*. The middle 50% of data points fall in 0.

*. The data spans from 0 to around 200.

*. The data is slightly skewed to the right, with a longer tail above the median.

*. There are no visible outliers in the data.

```r
##Water
##Boxplot
boxplot(Data_cleaned_1$Water,ylab="Water",main="Boxplot of Water",sub=paste("
Outliers: ",boxplot.stats(Data_cleaned_1$Water)$out))
```



Interpretation :-

*. Based on the horizontal line inside the box, it looks that the concrete mix's median water content value is 185 units.

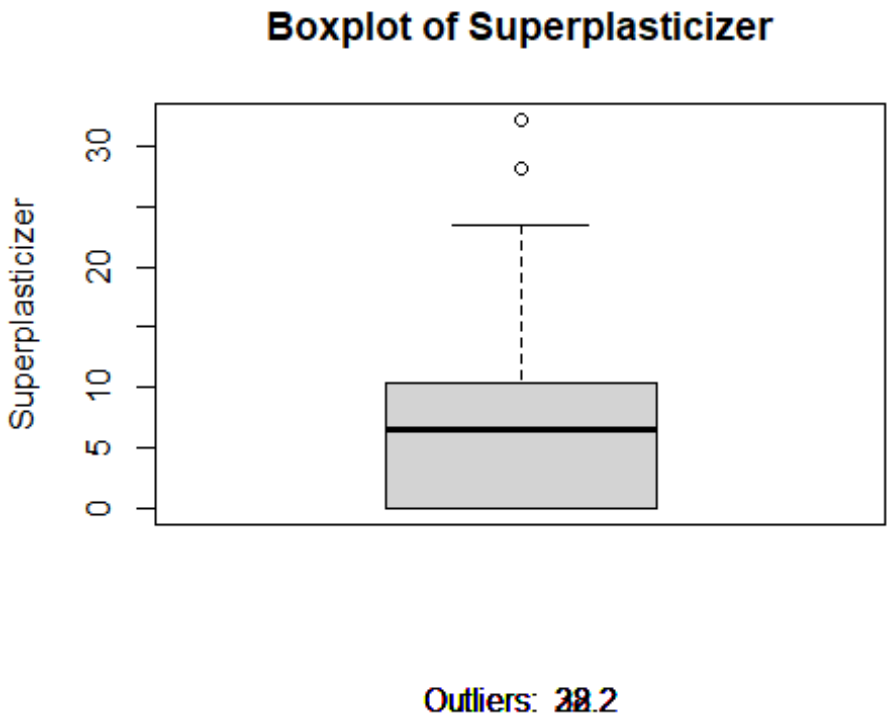
*. The water content of 25% of the samples is below the first quartile (Q1), which is about at 164.9 units. About 192 units is the third quartile (Q3), which indicates that 25% of the samples had a water content higher than this.

*. The IQR is approximately 27.1 units (192 – 164.9), and it represents the range that contains the middle 50% of the data.

*. The circles above the upper whisker point to two possible outliers. The water content values in these samples are substantially greater than those in the remaining data.

```r
#This box plot seems that three outlier points.
#Remove the outlier row.
outlier_rows_Water<-boxplot.stats(Data_cleaned_1$Water)$out

Data_cleaned_2<-subset(Data_cleaned_1,!(Water%in%outlier_rows_Water))
```

```r
#Superplasticizer
##Boxplot
boxplot(Data_cleaned_2$Superplasticizer,ylab="Superplasticizer",main="Boxplot
of Superplasticizer",sub=paste("Outliers: ",boxplot.stats(Data_cleaned_2$Supe
rplasticizer)$out))
```



Interpretation :-

*. The superplasticizer value that is median is 6.4.

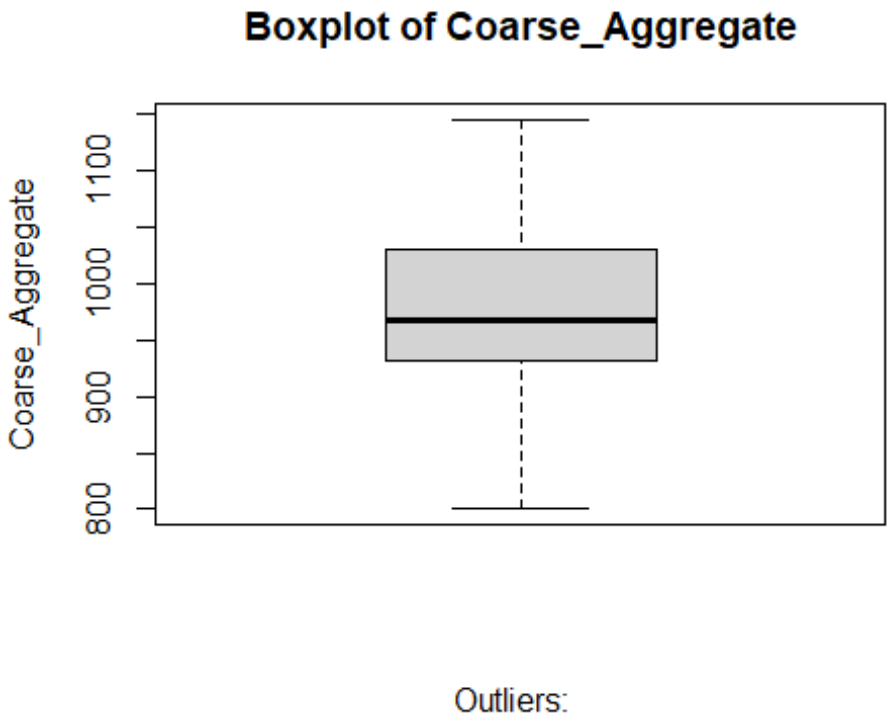
*. The range of the data is around 0 to 30, with a notable difference between the upper quartile and the outlier.

*. 75% of the data points are in the range of 6.4 to 10.2.

*. There are two outliers, suggesting that the superplasticizer value is atypically high.

```r
#This box plot seems that two outlier points.
#Remove the outlier row.
outlier_rows_Superplasticizer<-boxplot.stats(Data_cleaned_2$Superplasticizer)
$out


Data_cleaned_3<-subset(Data_cleaned_2,!(Superplasticizer%in%outlier_rows_Supe
rplasticizer))
```

```r
##Coarse_Aggregate
##Boxplot
boxplot(Data_cleaned_3$Coarse_Aggregate,ylab="Coarse_Aggregate",main="Boxplot
of Coarse_Aggregate",sub=paste("Outliers: ",boxplot.stats(Data_cleaned_3$Coar
se_Aggregate)$out))
```



Interpretation :-

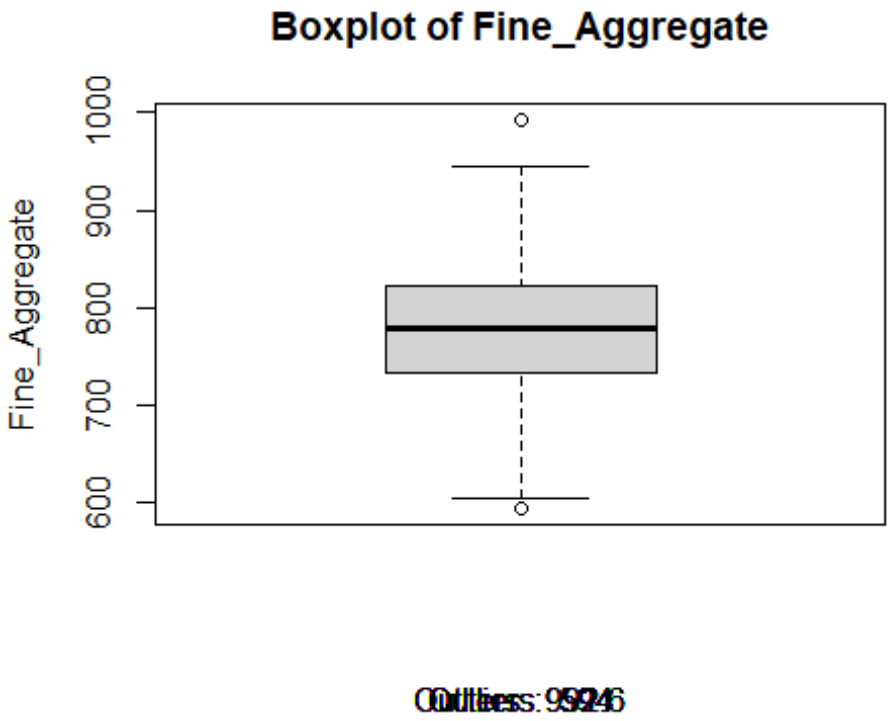
*. 968 is the Coarse_Aggregate median value.

*. There is a small tilt towards higher numbers, but overall the data is quite symmetrical around the median.

*. There are roughly 97.4 units separating the box, or the 25th and 75th percentiles.

*. This boxplot seems that no outlier points

```r
##Fine_Aggregate
##Boxplot
boxplot(Data_cleaned_3$Fine_Aggregate,ylab="Fine_Aggregate",main="Boxplot of
Fine_Aggregate",sub=paste("Outliers: ",boxplot.stats(Data_cleaned_3$Fine_Aggr
egate)$out))
```



Interpretation :-

*. 779.5 is the fine aggregate value that is median.

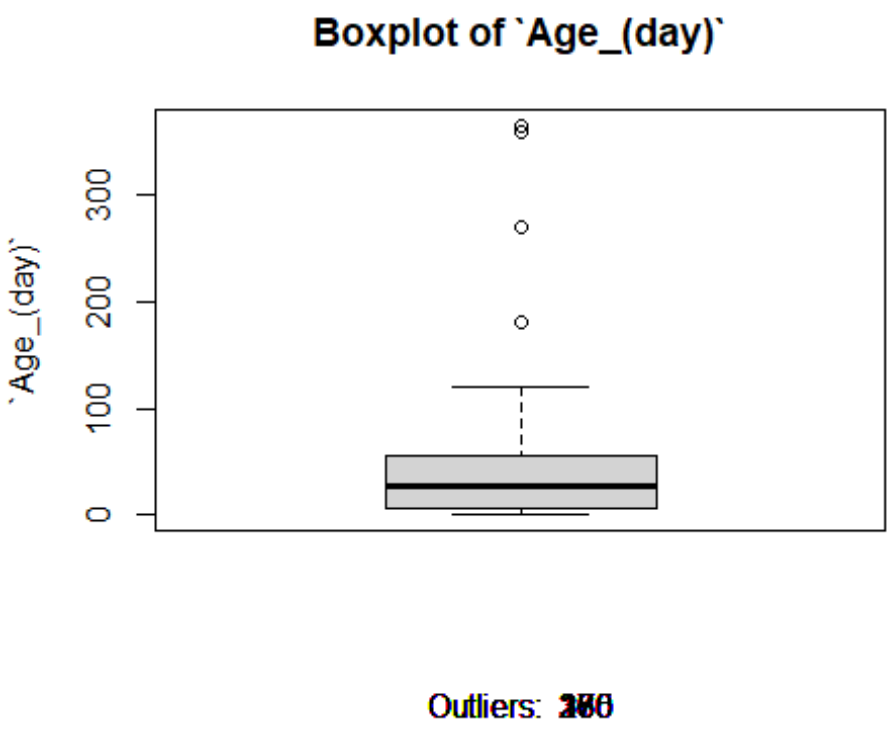
*. There is a longer tail pointing towards higher values, and the data is slightly biased to the right.

*. There are roughly 400 units in the fine aggregate value range.

*. At around 600 and 1000, respectively, there are two possible outliers.

```r
#This box plot seems that one outlier point.
#Remove the outlier row.
outlier_rows_Fine_Aggregate<-boxplot.stats(Data_cleaned_3$Fine_Aggregate)$out


Data_cleaned_4<-subset(Data_cleaned_3,!(Fine_Aggregate%in%outlier_rows_Fine_A
ggregate))
```

```r
#Age_(day)
##Boxplot
boxplot(Data_cleaned_4$Age_day,ylab="`Age_(day)`",main="Boxplot of `Age_(day)
`",sub=paste("Outliers: ",boxplot.stats(Data_cleaned_4$Age_day)$out))
```



Interpretation :-

*. The median age of the concrete samples at testing is 28 days. This is indicated by the horizontal line within the box.

*. The distribution of ages is slightly skewed to the right, as indicated by the longer whisker extending towards the higher age values.

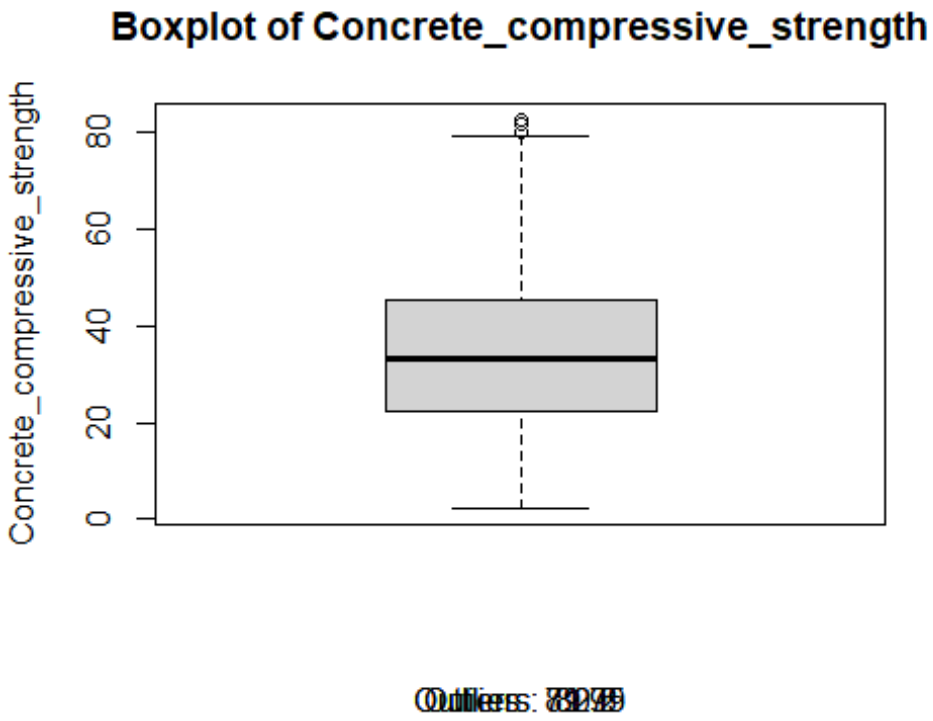
*. The ages of the concrete samples range from approximately 0 days to around 300 days.

*. There are four outliers.

*. The IQR, representing the middle 50% of the data, is approximately 50 days. This suggests that a significant portion of the samples were tested

```r
#This box plot seems that four outlier points.
#Remove the outlier row.
outlier_rows_Age<-boxplot.stats(Data_cleaned_4$Age_day)$out

Data_cleaned_5<-subset(Data_cleaned_4,!(Age_day%in%outlier_rows_Age))
```

```r
#Concrete_compressive_strength
##Boxplot
boxplot(Data_cleaned_5$Concrete_compressive_strength,ylab="Concrete_compressive_strength",main="Boxplot of Concrete_compressive_strength",sub=paste("Outliers: ",boxplot.stats(Data_cleaned_5$Concrete_compressive_strength)$out))
```



Interpretation :-

*. The median compressive strength is around 35 MPa (megapascals). This indicates that half of the concrete samples have a compressive strength below 34.45 MPa, and the other half have a strength above it.

*. The IQR, represented by the box's height, is approximately 22 MPa. This suggests a moderate spread in the compressive strengths of the concrete samples.

*. The distribution is slightly skewed to the right, as indicated by the longer whisker above the box compared to the lower one. This means that there are more concrete samples with higher compressive strengths than lower ones.

*. There are two potential outliers

```r
#This box plot seems that four outlier points.
#Remove the outlier row.
outlier_rows_Concrete_compressive_strength<-boxplot.stats(Data_cleaned_5$Concrete_compressive_strength)$out
```

```r
Data_cleaned_6<-subset(Data_cleaned_5,!(Concrete_compressive_strength%in%outl
ier_rows_Concrete_compressive_strength))

###This is our cleaned dataset

cleaned_data<-Data_cleaned_6
head(cleaned_data)
```

```
##    Cement Blast_Furnace_Slag Fly_Ash Water Superplasticizer Coarse_Aggrega
te
## 2   540.0                0.0       0   162             2.5           1055
.0
## 6   266.0              114.0       0   228             0.0            932
.0
## 9   266.0              114.0       0   228             0.0            932
.0
## 11  198.6              132.4       0   192             0.0            978
.4
## 12  198.6              132.4       0   192             0.0            978
.4
## 14  190.0              190.0       0   228             0.0            932
.0
##    Fine_Aggregate Age_day Concrete_compressive_strength
## 2           676.0      28                         61.89
## 6           670.0      90                         47.03
## 9           670.0      28                         45.85
## 11          825.5      90                         38.07
## 12          825.5      28                         28.02
## 14          670.0      90                         42.33
```

```r
nrow(cleaned_data)
```

```
## [1] 926
```