STATISTICAL ANALYSIS ON CONCRETE STRENGTH IN RELATION TO ITS CONSTITUENTS

Detailed Project Report:

Prepared by: Sushanth S

MAIB 2021 Batch

Contents

Introduction:

Concrete is a mixture of sand or fly ash, water and some other aggregates in smaller quantity. Today concrete is used in very huge quantity, such that concrete usage stands just after water. This study focusses on the impact of each constituent element added to form a concrete mixture. How each constituent effect the final concrete strength.

The objective of the study is to predict the concrete strength based on constituents used and age of the concrete. This will allow us to find the remaining utility of any structure and decide upon the types of maintenance to be performed. We have a data set consisting of historical concrete strength data based on test results.

The data consists of following attributes:

1. Cement
2. Blast Furnace slag
3. Fly Ash
4. Water
5. Superplasticiser
6. Coarse Aggregate
7. Fine Aggregate
8. Age
9. Strength

Strength is the dependant variable to be studied based on the input variables.

Data Understanding:

Sample Data:

The sample data is as shown below:

Note: All the figure henceforth shown in the report will be an output from R-Code

```
> head(mydata)
  Cement Blast.Furnace.Slag Fly.Ash Water Superplasticizer Coarse.Aggregate Fine.Aggregate Age Strength
1  540.0                0.0       0   162              2.5           1040.0          676.0  28    79.99
2  540.0                0.0       0   162              2.5           1055.0          676.0  28    61.89
3  332.5              142.5       0   228              0.0            932.0          594.0 270    40.27
4  332.5              142.5       0   228              0.0            932.0          594.0 365    41.05
5  198.6              132.4       0   192              0.0            978.4          825.5 360    44.30
6  266.0              114.0       0   228              0.0            932.0          670.0  90    47.03
```

As per the **NOIR classification** (Nominal, Ordinal, Interval and Ratio classification) the data in dataset can be classified into **Interval data** of **continuous type**.

NULL value test was performed on the dataset. No NULL values where present in the dataset.

Key Statistics for Data:

Before we proceed let us find the key parameters of the data attribute:

```
> describe(mydata)
                   vars    n   mean     sd median trimmed    mad    min    max  range  skew kurtosis
Cement                1 1030 281.17 104.51 272.90  273.47 117.72 102.00  540.0 438.00  0.51    -0.53
Blast.Furnace.Slag    2 1030  73.90  86.28  22.00   62.43  32.62   0.00  359.4 359.40  0.80    -0.52
Fly.Ash               3 1030  54.19  64.00   0.00   46.86   0.00   0.00  200.1 200.10  0.54    -1.33
Water                 4 1030 181.57  21.35 185.00  181.19  19.27 121.80  247.0 125.20  0.07     0.11
Superplasticizer      5 1030   6.20   5.97   6.40    5.56   7.86   0.00   32.2  32.20  0.90     1.39
Coarse.Aggregate      6 1030 972.92  77.75 968.00  973.49  68.64 801.00 1145.0 344.00 -0.04    -0.61
Fine.Aggregate        7 1030 773.58  80.18 779.50  776.41  67.46 594.00  992.6 398.60 -0.25    -0.11
Age                   8 1030  45.66  63.17  28.00   32.53  31.13   1.00  365.0 364.00  3.26    12.07
Strength              9 1030  35.82  16.71  34.45   34.96  16.20   2.33   82.6  80.27  0.42    -0.32
```

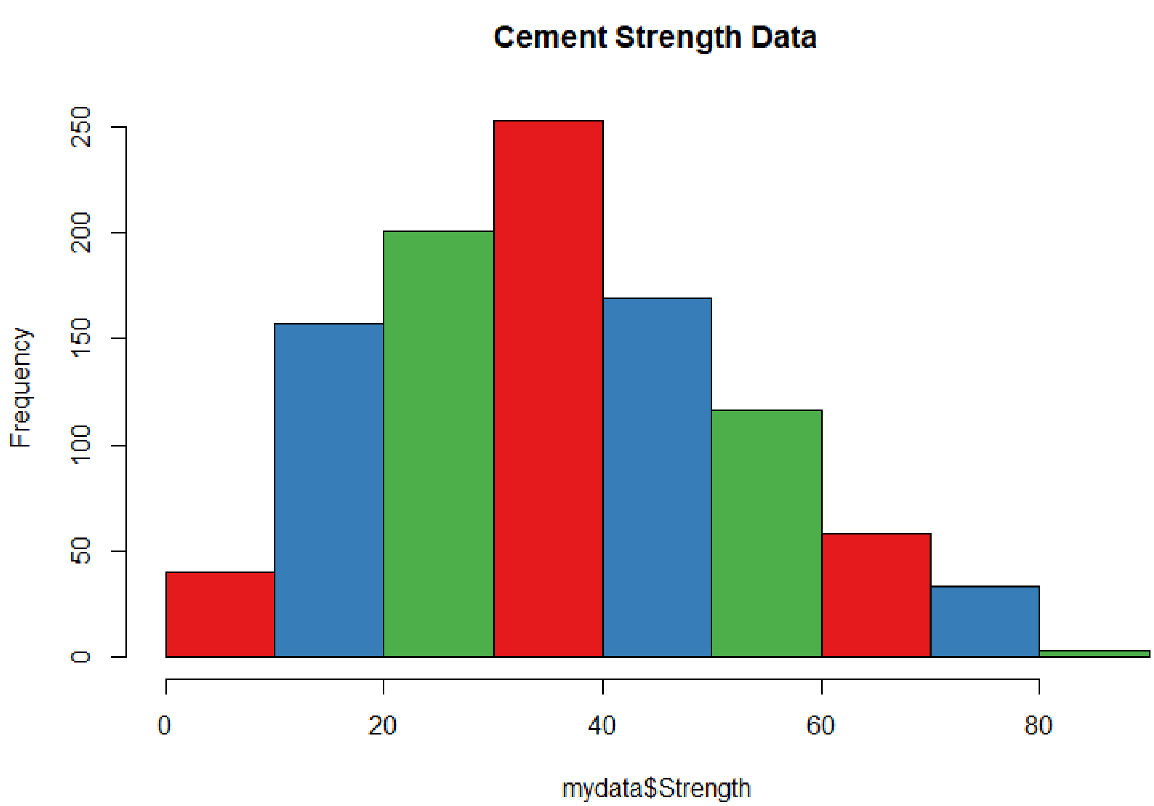
From the above table output:

Following key observations can be made:

1. Mean of Strength 35.8 is higher than median 34.4, indicating a positive skewness 0.42.
    a. We will check for outliers and make the mean closer to median
2. Fine Aggregate and Superplasticizer are having high kurtosis (4^{th} derivative of moment generating function) 12.07 and 1.39 respectively

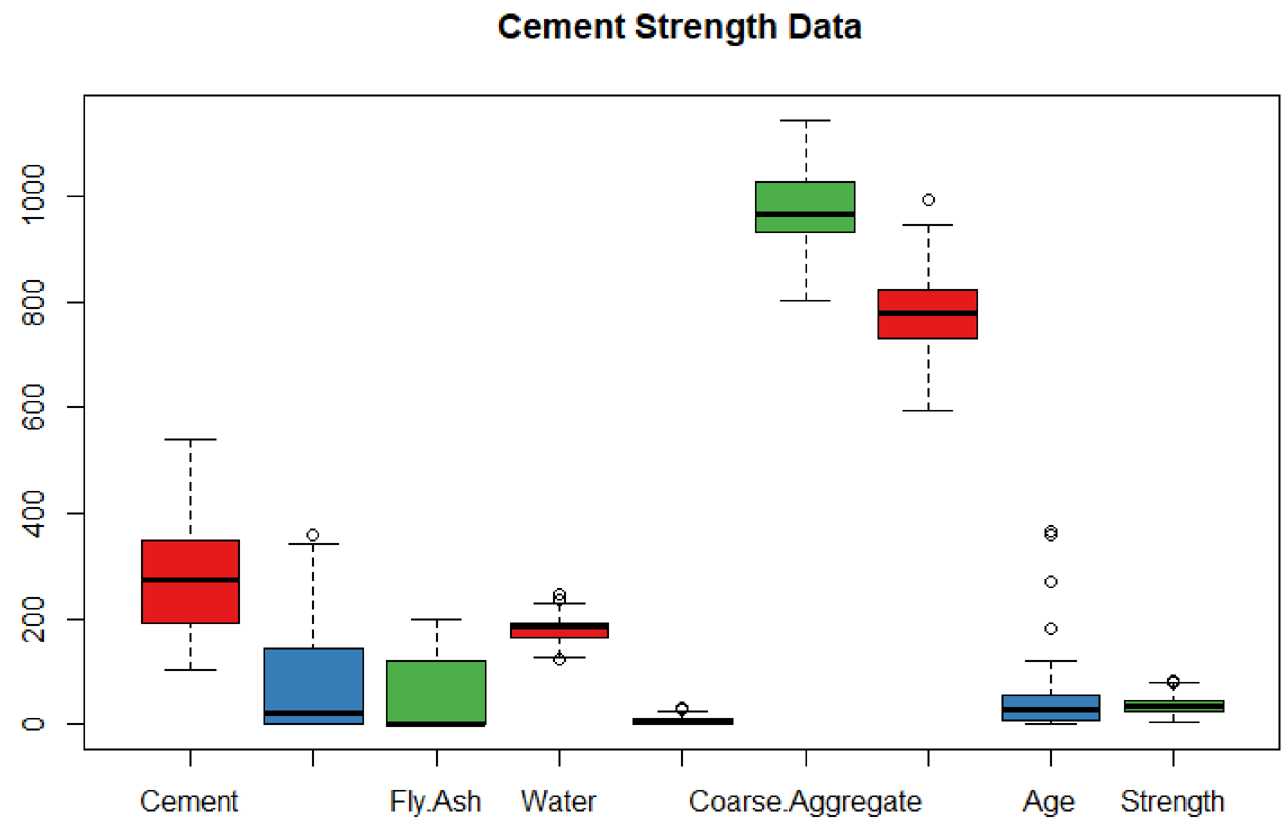
Histogram & Box Plots:

Let us draw some important plots to understand the distribution of our dependent variable:



As the above figure is a normal approximation but slightly positively skewed.

The box plot below shows how each attribute is classified and how much outliers are present.



There are very few outliers in the data set. After examining the dataset, it was found that only age is one parameter for which the outliers can be removed. As there was no correlation observed between the output and age for the outlier values.

The new data set was imported into the R environment for further study.

After removal of outliers, the mean value became more representable.

Mean of Strength (before) = 35.8 (when median = 34.9)

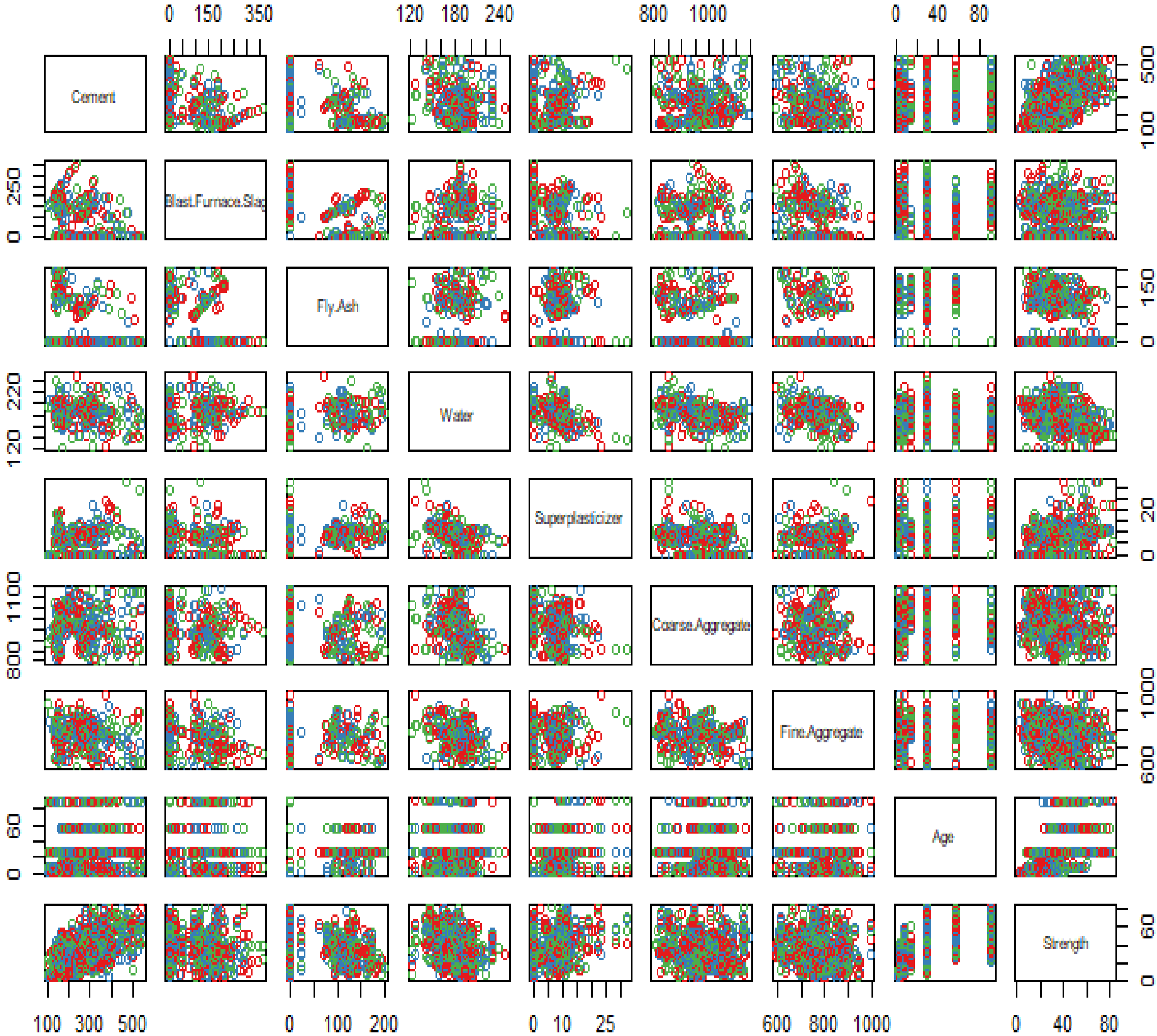
Mean of Strength (After) = 34.5 (closer to the median of sample)

Correlation Scatter Plot and Correlation Matix:

- Correlation coefficient between two random variables X and Y, usually denoted by $r(X,Y)\ or\ r_{XY}$ is a numerical measure of linear relationship between them and is defined as:

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- r_{XY} provided a measure of linear relationship between X and Y.
- It is a measure of degree of relationship.

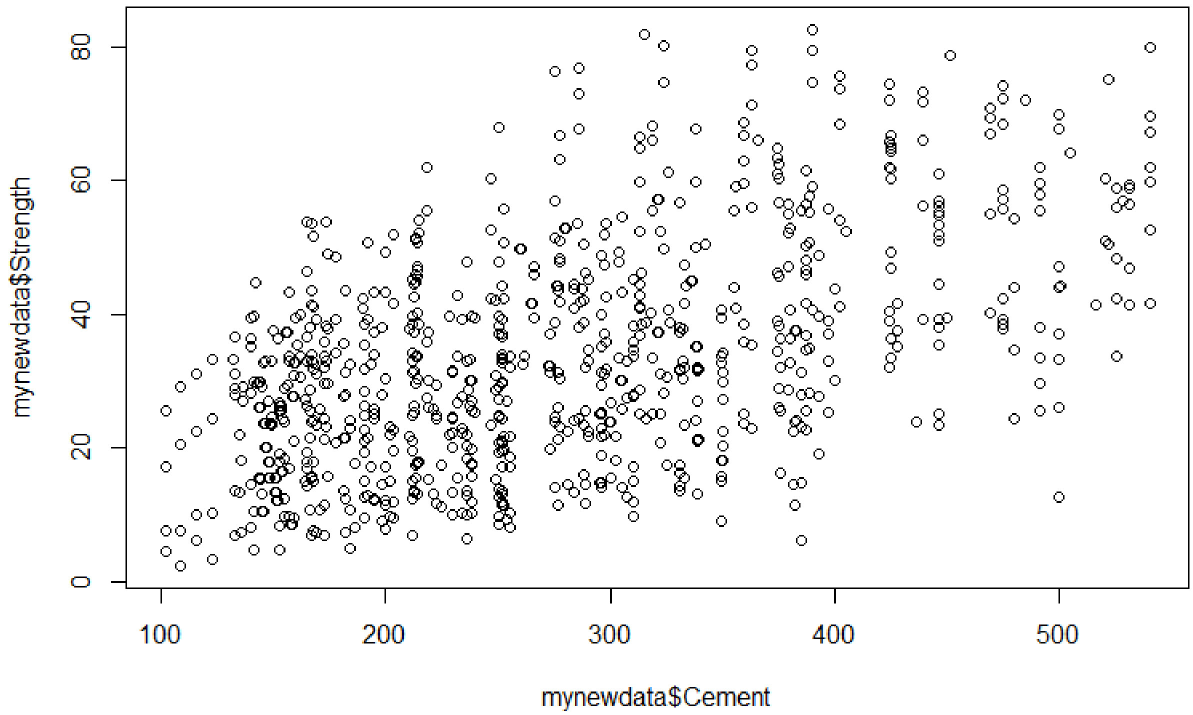


	Cement	Blast.Furnace.Slag	Fly.Ash	Water	Superplasticizer	Coarse.Aggregate	Fine.Aggregate	Age	Strength
Cement	1.000	-0.278	-0.374	-0.163	0.143	-0.114	-0.175	0.051	**0.532**
Blast.Furnace.Slag	-0.278	1.000	-0.334	0.103	0.041	-0.270	-0.284	0.069	0.172
Fly.Ash	-0.374	-0.334	1.000	-0.180	0.341	-0.049	0.025	-0.081	-0.127
Water	-0.163	0.103	-0.180	1.000	**-0.643**	-0.151	-0.398	0.036	-0.358
Superplasticizer	0.143	0.041	0.341	-0.643	1.000	-0.299	0.188	0.044	0.416
Coarse.Aggregate	-0.114	-0.270	-0.049	-0.151	-0.299	1.000	-0.217	-0.099	-0.221
Fine.Aggregate	-0.175	-0.284	0.025	-0.398	0.188	-0.217	1.000	-0.016	-0.147
Age	0.051	0.069	-0.081	0.036	0.044	-0.099	-0.016	1.000	0.524
Strength	0.532	0.172	-0.127	-0.358	0.416	-0.221	-0.147	0.524	1.000

From the correlation data we can conclude only two parameters are correlation with low significance level.

We can also see a some step pattern with age and other parameters.

Below is the scatter plot of cement quantity and concrete strength. Both are dependent and positively correlated.



Trian and Test Split:

The final data after outlier removal is taken for analysis. Now data is split into training and testing data with following commands.

```
library(caTools)

set.seed(123)
split = sample.split(mynewdata$Strength, SplitRatio = 0.70)

train_data <- subset(mynewdata, split==T)  # Created training data for analysis

test_data <- subset(mynewdata, split==F)   # Created testing data for final verification
```

Data is split in 70:30 ratio, 70 % data is considered for training and remaining 30% data is considered for testing.

Regression Analysis:

Now that the data is split into training and testing part. We will take the training data for our regression model.

- A multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon \quad \ldots \ldots \ldots \ldots (1)$$

- ☐ For Hypothesis testing and the setting of confidence limits, we also assume that ε is normally distributed.
- ☐ The linearity of the model (1) is defined with respect to the regression coefficients

X variables $\beta_1, \beta_2\ etc.$ ... in the test are as follows:

1. Cement
2. Blast Furnace slag
3. Fly Ash
4. Water
5. Super Plasticizer
6. Coarse Aggregate
7. Fine Aggregate
8. Age

Y variable for the model is:

1. Concrete Strength percentage

Regression model Trial -1:

Model-1

Output of the model:

Statistic	Value	Criteria
Residual standard error	8.578	
Multiple R-squared	0.753	> 0.6
Adjusted R-squared	0.749	> 0.6

Model	df	F	p value
Regression	8	240.8	2.2e-16
Residual	632		
Total	641		

Criteria:

P value < 0.05 of the above F test indicates that the Model-1 holds good for predicting the output.

Regression Output Coefficients and p-value:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         5.499199  28.155521   0.195    0.845    
Cement              0.114625   0.008864  12.932  < 2e-16 ***
Blast.Furnace.Slag  0.090671   0.010730   8.451  < 2e-16 ***
Fly.Ash             0.075851   0.013237   5.730 1.55e-08 ***
Water              -0.191243   0.041823  -4.573 5.80e-06 ***
Superplasticizer    0.089428   0.095848   0.933    0.351    
Coarse.Aggregate    0.004854   0.010000   0.485    0.628    
Fine.Aggregate      0.006324   0.011411   0.554    0.580    
Age                 0.350541   0.014253  24.594  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above result show that for attributes Superplasticizer, coarse Aggregate and Fine Aggregate the P-Value is not significant. This gives clear indication that these parameters have less or no impact on the output. Also, the intercept for this model is not significant.

Let us proceed further and improve the model in the following steps.

Model-2: (After removing the insignificant attributes)

X variables $\beta_1, \beta_2\ etc.$ ... in the test are as follows:

1. Cement
2. Blast Furnace slag
3. Fly Ash
4. Water
5. ~~Super Plasticizer~~
6. ~~Coarse Aggregate~~
7. ~~Fine Aggregate~~
8. Age

Y variable for the model is:

2. Concrete Strength percentage

Output of the model:

Statistic	Value	Criteria
Residual standard error	8.566	

Multiple R-squared	0.752	> 0.6
Adjusted R-squared	0.75	> 0.6

Model	df	F	p value
Regression	5	386	2.2e-16
Residual	635		
Total	641		

Only a small improvement in Adjusted R^2 could be achieved in the second iteration.

Criteria:

3. P value < 0.05 of the above F test indicates that the Model-2 holds good for predicting the output.

Model Validation:

In order to validate the model we will conduct VIC test (Variance Inflation factor) and step AIC to see whether the model is optimum:

Variance Inflation Factor (VIF)

Measures the correlation (linear association) between each x variable with other x's

$VIF_i = 1/(1- R_i^2)$

Where R_i is the coefficient for regressing x_i on other x's

Criteria: VIF > 5 can be an indication of multi collinearity.

Tackling Multicollinearity: Remove one or more of highly correlated independent variable.

Method: Removing highly correlated variable – Stepwise Regression

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2$$

Results:

```
car :: vif(regressor1)    #vif <20 so no collinearity
          Cement Blast.Furnace.Slag            Fly.Ash              Water   Superplasticizer
        7.743983           7.854438           6.312516           6.361523           2.920702
Coarse.Aggregate     Fine.Aggregate                Age
        5.659706           6.773512           1.024179
car :: vif(regressor2)    #vif <20 so no collinearity
          Cement Blast.Furnace.Slag            Fly.Ash              Water                Age
        1.582490           1.459059           1.640583           1.104120           1.015092
```

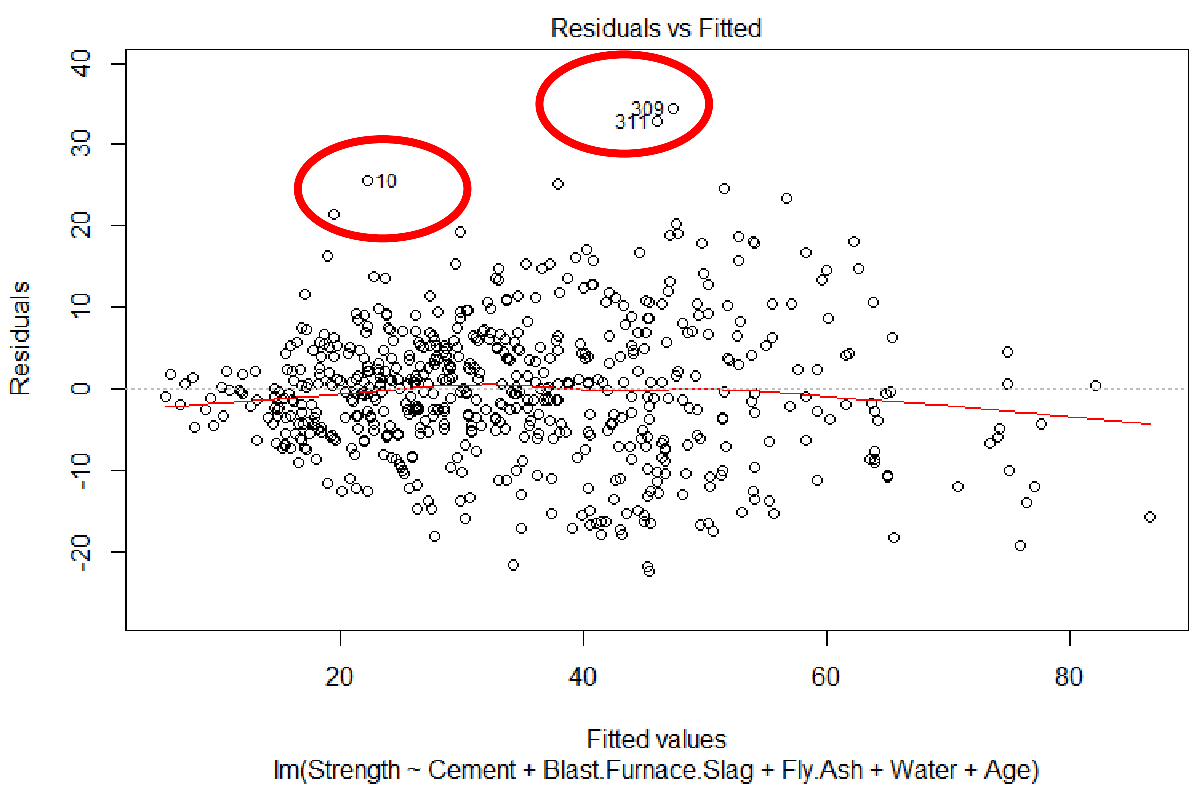
VIF values are higher for many parameters, lets validate it further by conducting step AIC.

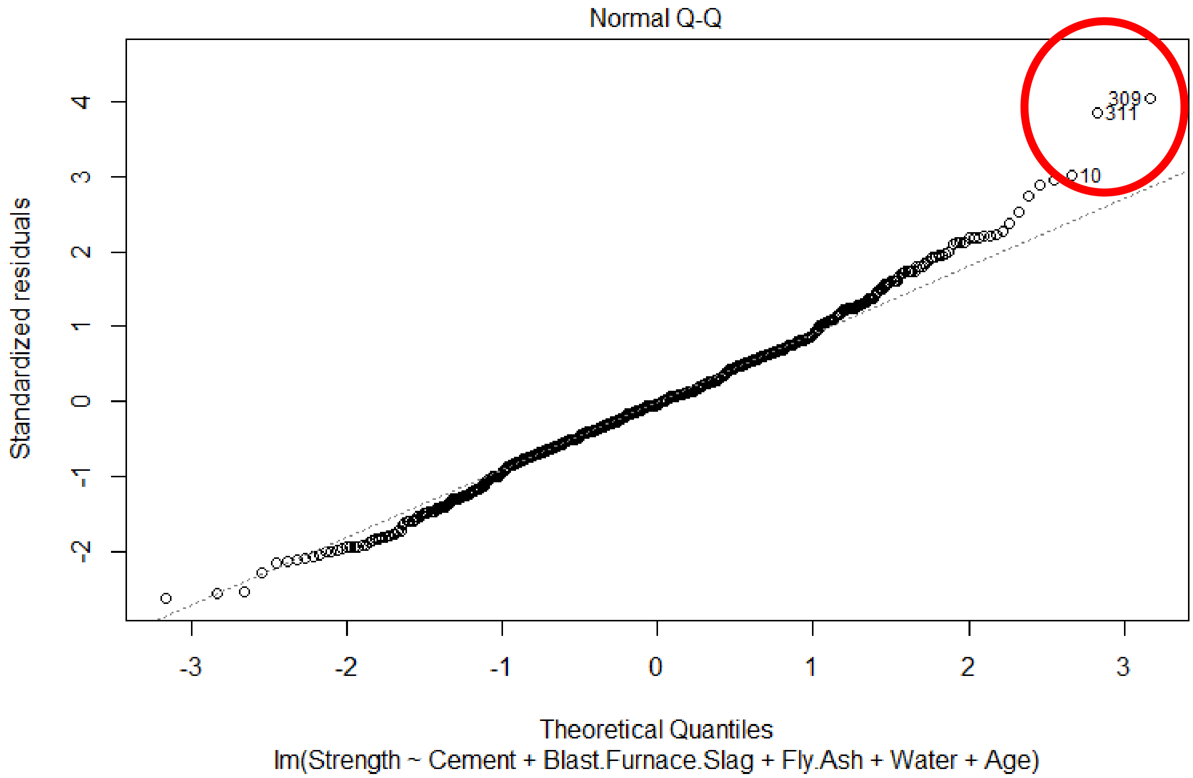
Step AIC is performed on model-1 only considering all the input parameters:

```
Step:  AIC=2759.38
Strength ~ Cement + Blast.Furnace.Slag + Fly.Ash + Water + Age

                     Df Sum of Sq    RSS    AIC
<none>                             46589 2759.4
- Fly.Ash             1      8746  55334 2867.7
- Water               1     12092  58681 2905.3
- Blast.Furnace.Slag  1     26343  72932 3044.7
- Age                 1     45057  91646 3191.1
- Cement              1     57273 103862 3271.3

Call:
lm(formula = Strength ~ Cement + Blast.Furnace.Slag + Fly.Ash +
    Water + Age, data = train_data)

Coefficients:
       (Intercept)              Cement  Blast.Furnace.Slag             Fly.Ash               Water
          22.64075             0.11179             0.08750             0.07357            -0.22336
               Age
           0.35113
```

It is evident that our model-2 output is a corollary of step AIC. We have already deleted 3 parameters from the input of model-2.

Residuals and QQ plot



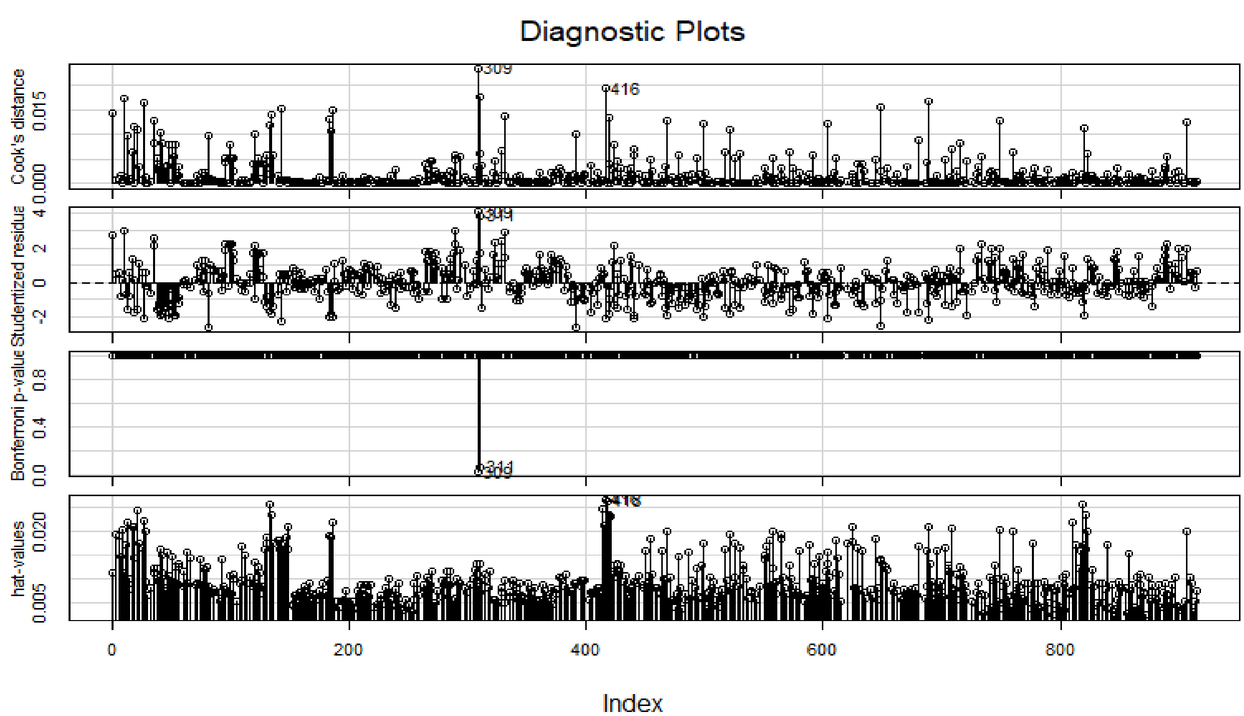


Inference: From both residual vs fitted plot and Normal QQ plot, we can see the assumption that data is linearly distributed is correct. But there are some outliers that are highlighted I both the above plots.

Influence Index Plot:

To identify the outliers in data and rejecting them to improve the model performance.

We will pass the model-2 regressor into the plot and check its outcome:



Model-3 (final optimization after outlier treatment):

Output of the model:

Statistic	Value	Criteria
Residual standard error	7.99	
Multiple R-squared	0.782	> 0.6
Adjusted R-squared	0.781	> 0.6

Model	df	F	p value
Regression	5	451.5	2.2e-16
Residual	626		
Total	632		

Model-3 shows significant improvement by approx. 3%.

Criteria:

P value < 0.05 of the above F test indicates that the Model-3 holds good for predicting the output.

Regression Output Coefficients and p-value:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       20.349137   3.784777   5.377 1.07e-07 ***
Cement             0.113417   0.003842  29.517  < 2e-16 ***
Blast.Furnace.Slag 0.086603   0.004317  20.060  < 2e-16 ***
Fly.Ash            0.075866   0.006487  11.694  < 2e-16 ***
Water             -0.215465   0.016913 -12.740  < 2e-16 ***
Age                0.359114   0.013750  26.118  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

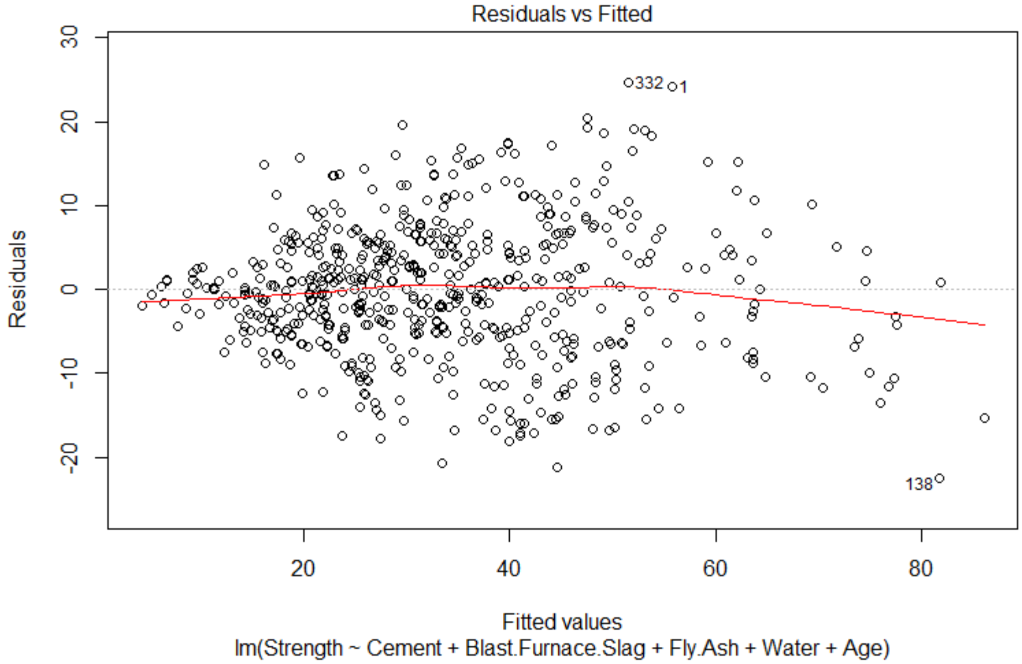
All the coefficients have significant p-value. The intercept also has a significant value. The model is a good fit for prediction.

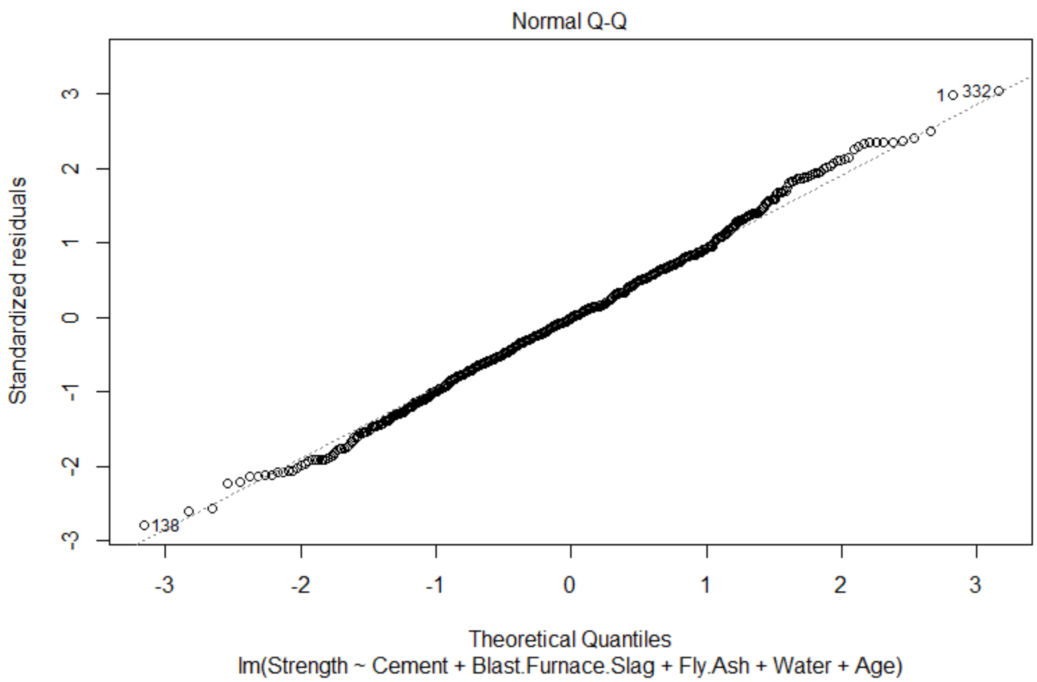
Let us perform model validation and perform hypothesis testing on the model validity.

Model Equation:

Y (Concrete Strength) = 20.3 + 0.11 * Cement + 0.08 * Blast Furnace Slag + 0.075 * Fly Ash – 0.215 * Water + 0.359 * Age

Residuals and QQ plot for Model-3:





The above residual and QQ plot show comparatively lesser distortion than the previous results.

Hypothesis Testing on Model-3 outcome:

Null Hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$$

$$ag. \quad H_1 \ : \ \beta_j \neq 0, \ for \ atleast \ one \ j.$$

ANOVA Output:

```
Analysis of Variance Table

Response: Strength
                   Df Sum Sq Mean Sq F value    Pr(>F)    
Cement              1  54868   54868  858.55 < 2.2e-16 ***
Blast.Furnace.Slag  1  18679   18679  292.28 < 2.2e-16 ***
Fly.Ash             1  16657   16657  260.64 < 2.2e-16 ***
Water               1  10468   10468  163.79 < 2.2e-16 ***
Age                 1  43595   43595  682.15 < 2.2e-16 ***
Residuals         626  40006      64                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As the above results show all the p values are significant. **We can reject the NULL hypothesis. Model can be used for prediction.**

Conclusion:

Three regression models were trained with different input variable

Regression model trained with all input variables:

Model-1 Accuracy: 75.3 %

Regression model trained after removing insignificant parameters:

Superplasticizer	0.089428	0.095848	0.933	0.351
Coarse.Aggregate	0.004854	0.010000	0.485	0.628
Fine.Aggregate	0.006324	0.011411	0.554	0.580

Model-2 Accuracy: 75.2 %

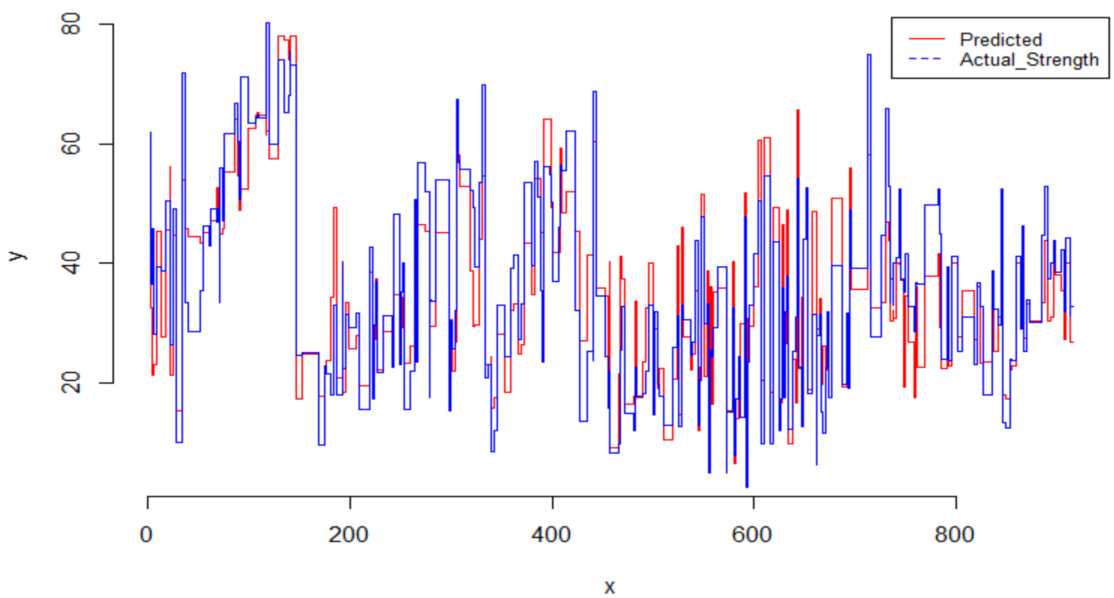
Prediction for test data:

The output Y label (concrete strength) was predicted for test data.

Model-3 Accuracy (after removing the outliers): 78.3 %

The result is as follows: **Accuracy of prediction = 86.6 %**

However, the accuracy of prediction on test data indicated the model is Underfitted with respect to training data. Further training needs to be done based on cross validation techniques to improve the prediction and resolve the underfitting in model.



References:

https://www.sciencemuseum.org.uk/objects-and-stories/everyday-wonders/building-modern-world-concrete-and-our-environment

https://www.sciencedirect.com/science/article/abs/pii/S1350630714000387#:~:text=Today%2C%20second%20only%20to%20water,all%20other%20building%20materials%20combined

https://online.stat.psu.edu/stat462/node/117/

https://www.kaggle.com/c/dat300-2018-concrete/data

Appendix:

R-Script (submitted along with this report)