STATISTICAL ANALYSIS AND PREDICTION OF CONCRETE COMPRESSIVE STRENGTH USING REGRESSION

BY
KRISHNENDU DUTTA
Roll no- B.Sc(Sem-VI)Stat-04
Department of Statistics,
Visva Bharati, Santiniketan

PLAN OF THE PROJECT

- Introduction
- Objective
- Data Description
- Data Visualization
- Statistical Tools and Techniques
- Statistical Analysis
- Conclusion & References

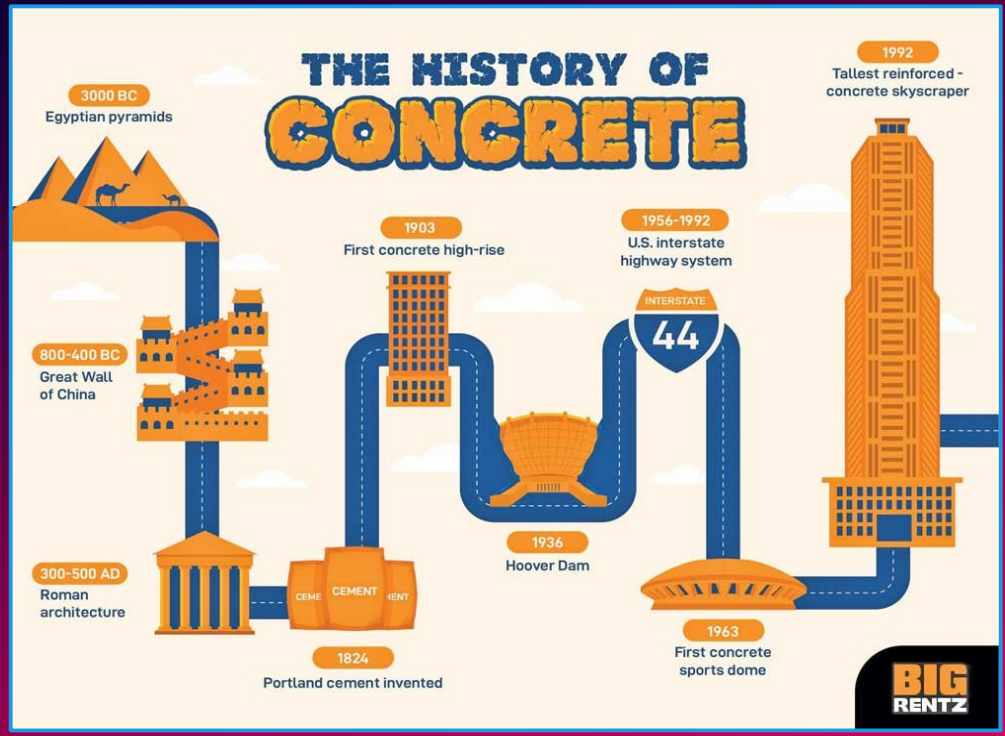
INTRODUCTION :

- Concrete is a composite material composed of fine and coarse aggregate bonded together with a fluid cement (cement paste) that hardens (cures) over time.
- Globally, the ready-mix concrete industry, the largest segment of the concrete market, is projected to exceed 600 billion in revenue by 2025.

- More than 70 % of the world's population lives in a structure that uses concrete. At this point, the amount of concrete in the world almost outweighs all living matter. It's a vital resource for construction and civil engineering.
- Our roads, overpasses and bridges are all laid with concrete. Our parking lots and parking structures are built with it. It is everywhere, and has been for a very long time.

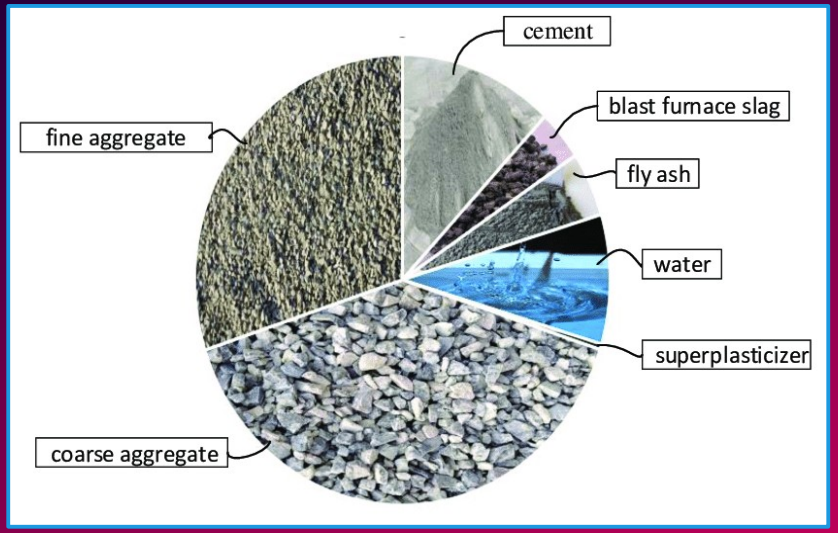
HISTORY OF CONCRETE :

- In the Ancient Egyptian and later Roman eras, builders discovered that adding volcanic ash to the mix allowed it to set underwater.
- The Romans useD concrete extensively from 300 BC to 476 AD.
- From the 11th century, the increased use of stone in church and castle construction led to an increased demand for mortar.
- Reinforced concrete was invented in 1849 by Joseph Monier .The first concrete reinforced bridge was designed and built by Joseph Monier in 1875.



COMPOSITION OF CONCRETE :

- Concrete is an artificial composite material, comprising a matrix of cementitious binder (typically Portland cement paste or asphalt) and a dispersed phase or "filler" of aggregate (typically a rocky material, loose stones, and sand).
- The binder "glues" the filler together to form a synthetic conglomerate. Many types of concrete are available, determined by the formulations of binders and the types of aggregate used to suit the application of the engineered material.
- These variables determine strength and density, as well as chemical and thermal resistance of the finished product.



TYPES OF CONCRETE :

- Normal Strength Concrete
- Plain or Ordinary Concrete
- Reinforced Concrete
- Prestressed Concrete
- Light – Weight Concrete
- Air Entrained Concrete
- Polymer Concrete
- Polymer impregnated concrete
- High-Performance Concrete
- Shotcrete Concrete
- Vacuum Concrete
- Stamped Concrete
- Asphalt Concrete
- Rapid Strength Concrete
- Precast Concrete
- High-Density Concrete
- Ready Mix Concrete
- Polymer cement concrete
- High-Strength Concrete
- Self - Consolidated Concrete
- Pervious Concrete
- Pumped Concrete
- Limecrete
- Roller Compacted Concrete
- Glass Concrete

WHY USE CONCRETE ?

- ***CONCRETE IS WATER-RESISTANT :***

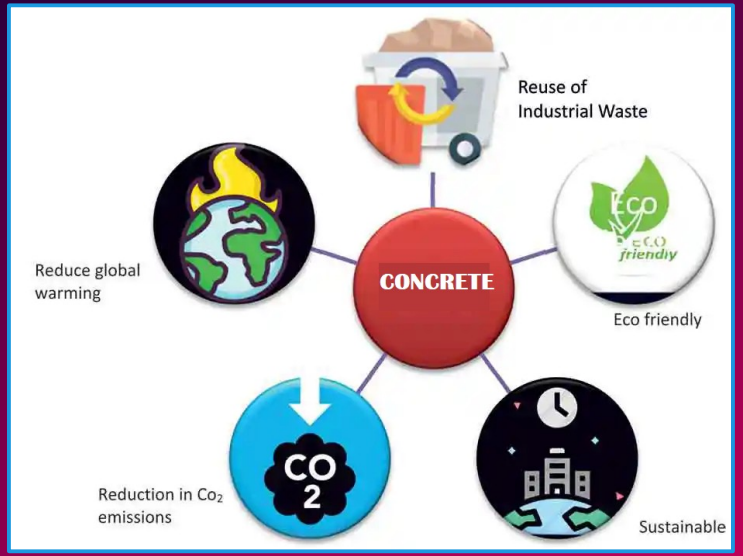
  As a complex and porous substance, concrete is resistant to erosion by water

- ***CONCRETE IS EASY TO WORK WITH :***

  In general, concrete requires less energy to produce and less effort to upkeep in comparison to other building materials.

- ***CONCRETE IS RESISTANT TO HEAT AND PRESSURE :***

  Unlike steel and wood, concrete does not transfer heat very well. It can withstand extremely high temperatures for hours on end. Similarly, the compound structure of normal concrete makes it incredibly resistant to the application of pressure.

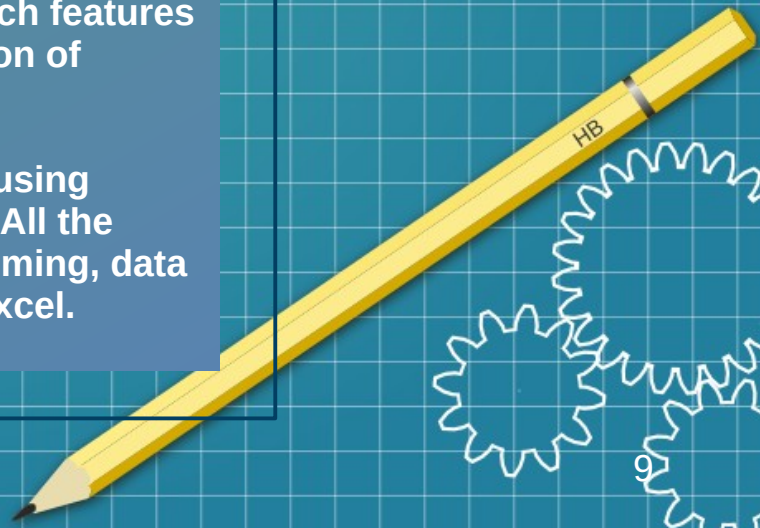


CONCRETE COMPRESSIVE STRENGTH

- Compressive strength of concrete is the Strength of hardened concrete measured by the compression test. The compression strength of concrete is a measure of the concrete's ability to resist loads which tend to compress it.
- It is measured by crushing cylindrical concrete specimens in compression testing machine. The compressive strength of concrete can be calculated by the failure load divided with the cross sectional area resisting the load and reported in mega pascals (MPa) in SI units.
- Concrete's compressive strength requirements can vary from 2500 psi (17 MPa) for residential concrete to 4000psi (28 MPa) and higher in commercial structures.

PROJECT OBJECTIVE

- In this project work our objective is to model the Compressive Strength of Concrete based on several associated factors (Cement, Blast Furnace Slag ,Fly Ash, Water, Coarse Aggregate, Superplasticizer ,Fine Aggregate ,Age).
- We have to find which model has the greatest accuracy and identify correlations in our data. Finally, we also have to determine which features are the most influential in our prediction of Compressive Strength of Concrete.
- We will perform Multivariate Analysis using various statistical methods and tools. All the analysis are being done by R Programming, data trimmings are done using Microsoft Excel.



DATA DESCRIPTION

Concrete is the most important material in civil engineering. The actual concrete compressive strength (MPa-megapascals) for a given mixture under a specific age (days) was determined from laboratory.

Given is the variable name, variable type, the measurement unit and a brief description. The concrete compressive strength is the regression problem. The order of this listing corresponds to the order of numerals along the rows of the database.

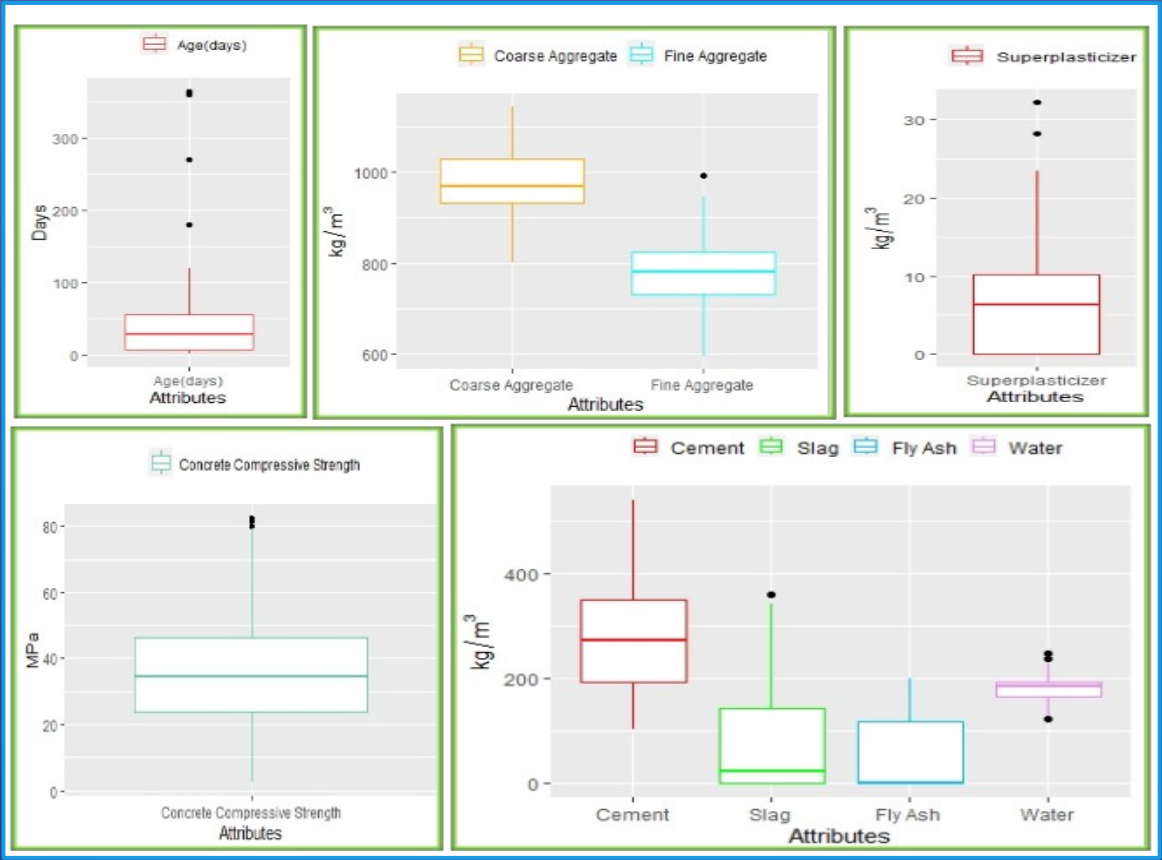
Data Source: https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength

Name	Data Type	Measurement	Description
Cement	quantitative	kg in a m3 mixture	Input Variable
Blast Furnace Slag	quantitative	kg in a m3 mixture	Input Variable
Fly Ash	quantitative	kg in a m3 mixture	Input Variable
Water	quantitative	kg in a m3 mixture	Input Variable
Superplasticizer	quantitative	kg in a m3 mixture	Input Variable
Coarse Aggregate	quantitative	kg in a m3 mixture	Input Variable
Fine Aggregate	quantitative	kg in a m3 mixture	Input Variable
Age	quantitative	Day (1~365)	Input Variable
Concrete compressive strength	**quantitative**	**MPa**	**Output Variable**

A GLIMPSE OF THE DATASET

	A	B	C	D	E	F	G	H	I	J
1	s.no	Cement (component 1)(kg in a m^3 mixture)	Blast Furnace Slag (component 2)(kg in a m^3 mixture)	Fly Ash (component 3)(kg in a m^3 mixture)	Water (component 4)(kg in a m^3 mixture)	Superplasticizer (component 5)(kg in a m^3 mixture)	Coarse Aggregate (component 6)(kg in a m^3 mixture)	Fine Aggregate (component 7)(kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
2	1	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
3	2	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
4	3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
5	4	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
6	5	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30
7	6	266.0	114.0	0.0	228.0	0.0	932.0	670.0	90	47.03
8	7	380.0	95.0	0.0	228.0	0.0	932.0	594.0	365	43.70
9	8	380.0	95.0	0.0	228.0	0.0	932.0	594.0	28	36.45
10	9	266.0	114.0	0.0	228.0	0.0	932.0	670.0	28	45.85
11	10	475.0	0.0	0.0	228.0	0.0	932.0	594.0	28	39.29
12	11	198.6	132.4	0.0	192.0	0.0	978.4	825.5	90	38.07
13	12	198.6	132.4	0.0	192.0	0.0	978.4	825.5	28	28.02
14	13	427.5	47.5	0.0	228.0	0.0	932.0	594.0	270	43.01
15	14	190.0	190.0	0.0	228.0	0.0	932.0	670.0	90	42.33
16	15	304.0	76.0	0.0	228.0	0.0	932.0	670.0	28	47.81
17	16	380.0	0.0	0.0	228.0	0.0	932.0	670.0	90	52.91
18	17	139.6	209.4	0.0	192.0	0.0	1047.0	806.9	90	39.36
19	18	342.0	38.0	0.0	228.0	0.0	932.0	670.0	365	56.14

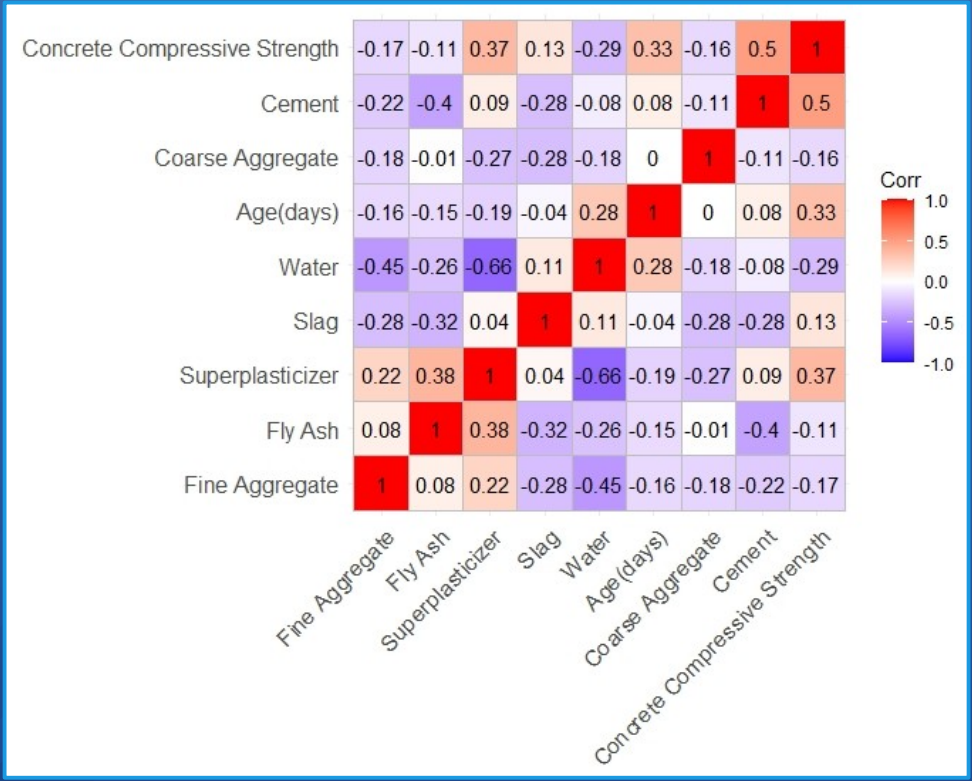
GRAPHICAL REPRESENTATION



BOXPLOT

From the boxplot of the attributes we can detect the presence of outliers in several of the attributes namely Blast Furnace Slag, Water, Age, Superplasticizer, Fine Aggregate and Concrete Compressive Strength.

GRAPHICAL REPRESENTATION



CORRELATION PLOT

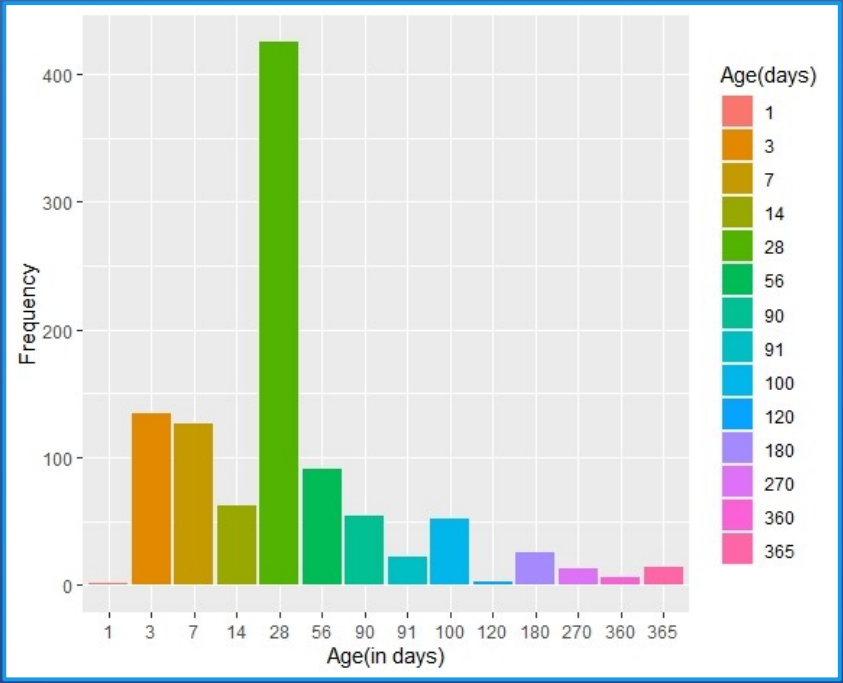
From the above correlation plot of the attributes we can determine the presence of highest positive moderate correlation between Cement and Concrete Compressive Strength.

Moreover moderate negative correlation exists between water and Superplastizer.

Also there seems to be no correlation between Coarse Aggregate and the Age of the Concrete.

Therefore we can come to the conclusion that the Compressive strength of Concrete is influenced by the factors such as Cement, Age of the Concrete and Superplasticizer.

GRAPHICAL REPRESENTATION



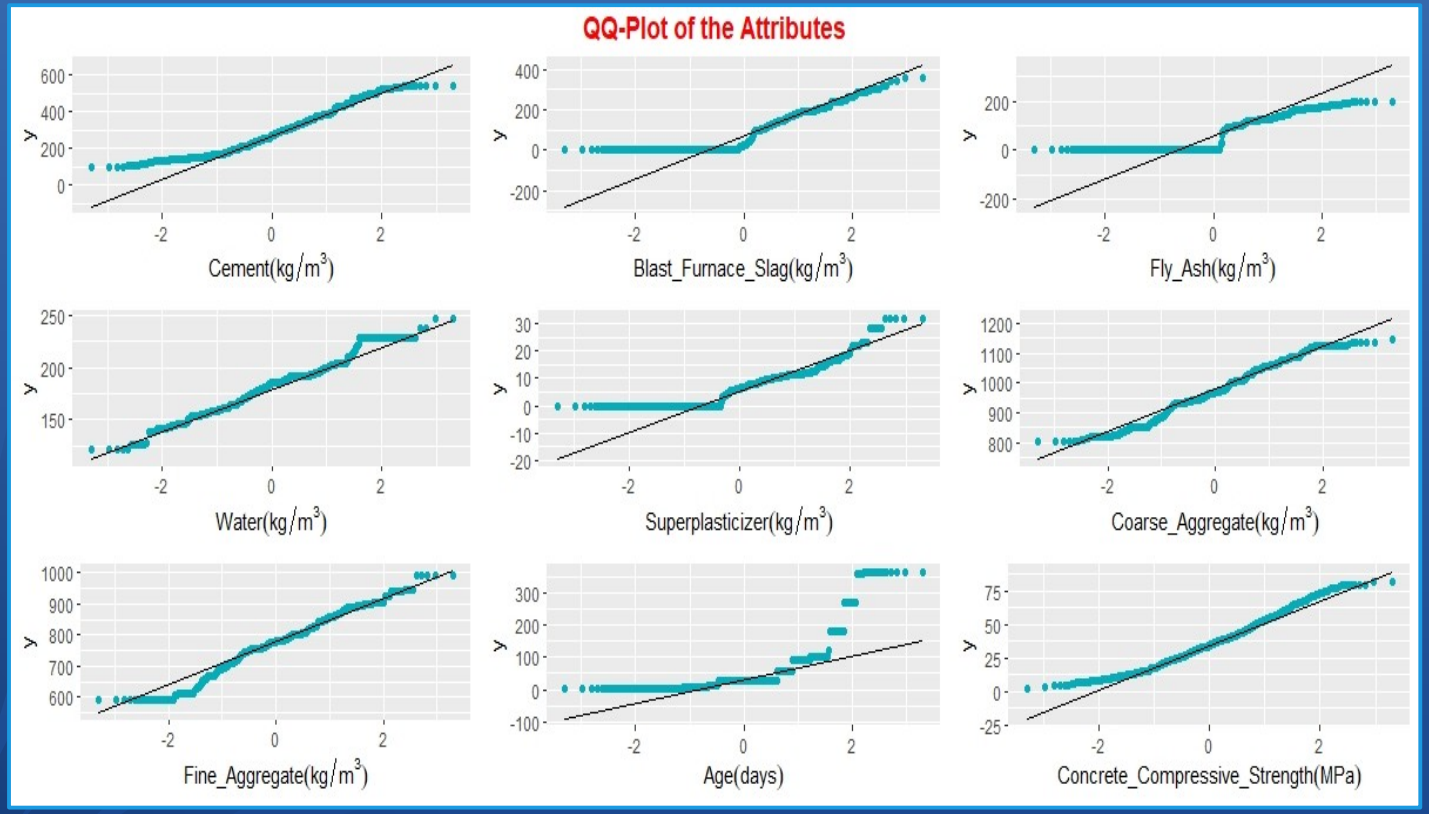
AGE PLOT OF THE CONCRETE SAMPLES

The above diagram gives us an idea about the age distribution of the concrete samples.

The x-axis represents the age of the concrete in days and the y axis represents the no of concrete samples of that particular age.

For example we have 425 samples of concrete which are of age 28 days. Similarly the number of samples for other age groups are given in this graph.

GRAPHICAL REPRESENTATION



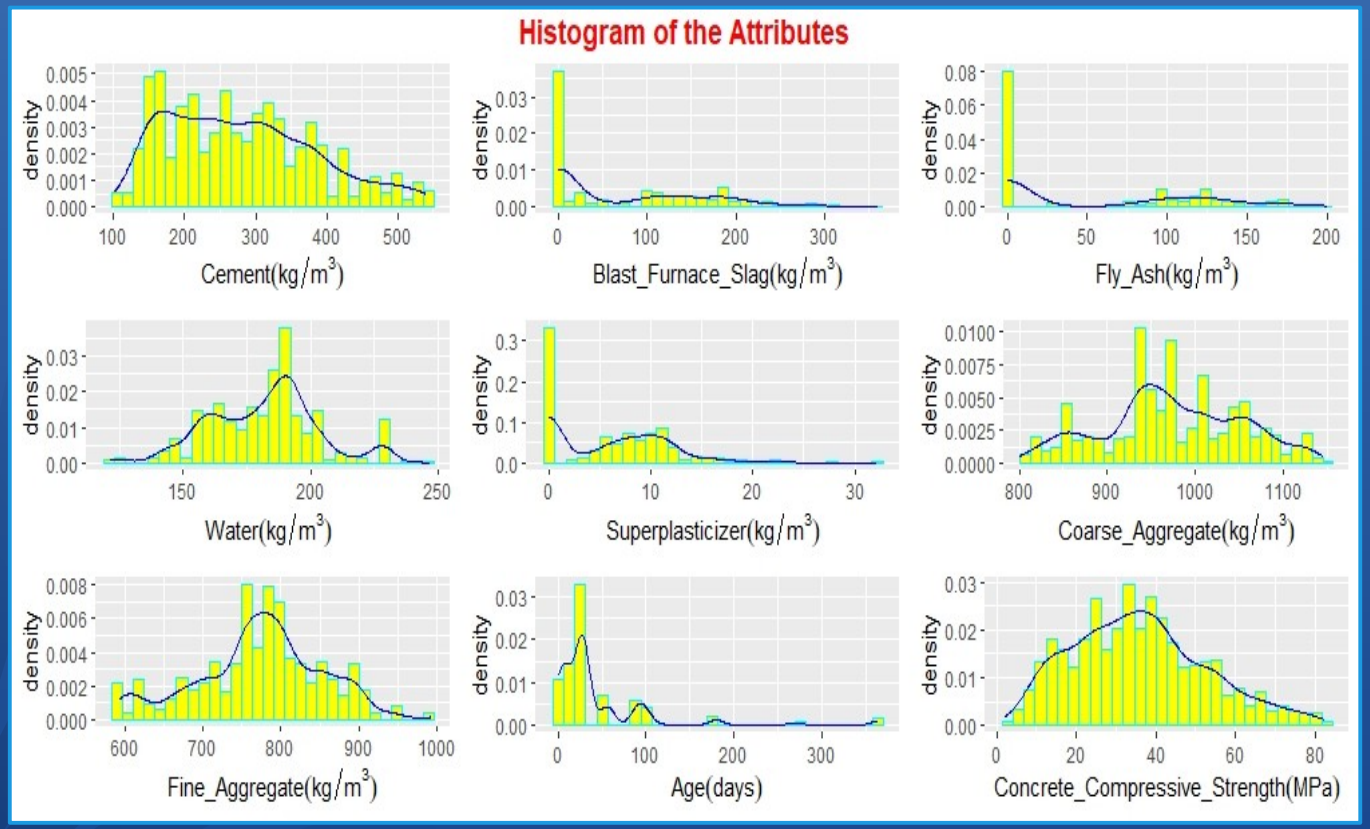
QQ-PLOT

A Quantile Quantile plot (or QQ Plot) helps us to check whether the data follows a normal distribution.

The data is assumed to be normally distributed when the points approximately follows the 45 degree reference line.

In the above diagram we can see that the our response variable Concrete Compressive strength approximately follows a normal distribution since almost all the points follows the reference line.

GRAPHICAL REPRESENTATION



HISTOGRAM

From the above plot of the histogram of the attributes we can get a general idea about the distribution of the data for each of the attributes.

Moreover addition of a density curve over the histogram helps us in better visualization of the plot.

Therefore by visual inspection we can infer that the response variable Concrete Compressive Strength more or less follows a normal distribution while this is not the case for the other remaining attributes.

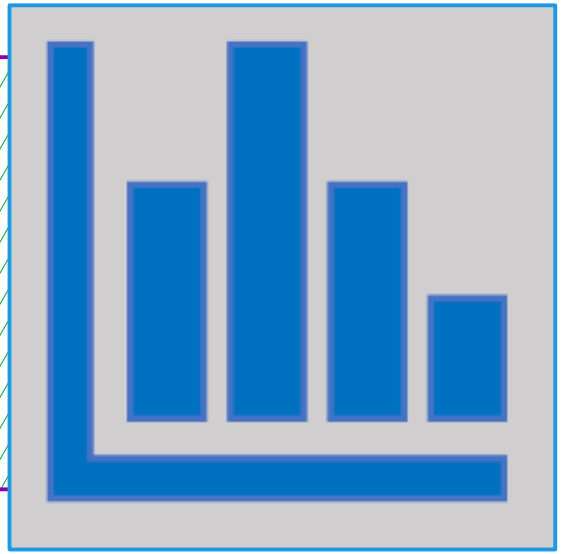
STATISTICAL TOOLS AND TECHNIQUES

SHAPIRO-WILK TEST

ANOVA

LINEAR REGRESSION

RIDGE REGRESSION



SHAPIRO-WILK TEST

The Shapiro–Wilk test is a test of normality. The test gives us a W value; small values indicate our sample is *not* normally distributed (we can reject the null hypothesis that our population is normally distributed if our values are under a certain threshold).

- **x_(i) are the ordered random sample values**
- **a_i are constants generated from the covariances, variances and means of the sample (size n) from a normally distributed sample**

The test has limitations, most importantly that the test has a bias by sample size. The larger the sample, the more likely we'll get a statistically significant result.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

MULTIPLE LINEAR REGRESSION

The multiple linear regression model is given by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Least Squares Estimation of the Parameters

- The least squares function is given by

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

- The least squares estimates must satisfy

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

and

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \ldots, k$$

RIDGE REGRESSION

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity.When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

Hoerl and Kennard (1970) proposed that potential instability in the LS estimator

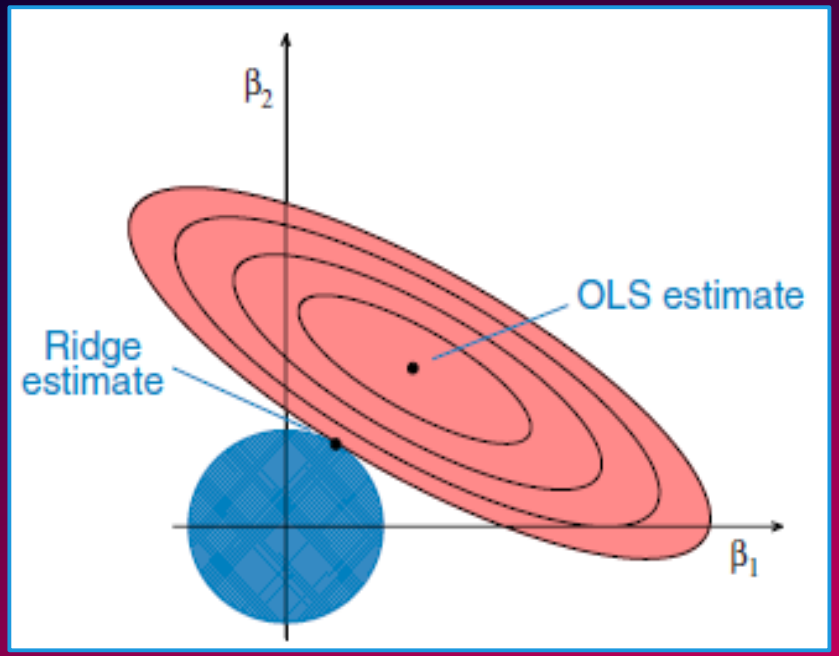
$$\hat{\beta}=(X'X)^{-1}X'Y$$

could be improved by adding a small constant value λ to the diagonal entries of the matrix X`X before taking its inverse.

The result is the ridge regression estimator

$$\hat{\beta}_{ridge}=(X'X+\lambda I_p)^{-1}X'Y$$

Therefore, ridge regression puts further constraints on the parameters, βj's, in the linear model. In this case, what we are doing is that instead of just minimizing the residual sum of squares we also have a penalty term on the β's. This penalty term is λ (a pre-chosen constant) times the squared norm of the β vector.



ANOVA(multiple linear regression)

Tests for multiple linear regression model

Let us suppose that we have a set of k independent variables, x_1, x_2, ..., x_k, and the dependent variable y. As the model, we take

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + e_i$$
$$(i = 1, 2, ..., n),$$

where e_i are independently normal, each with mean 0 and variance σ_e^2. We are interested in the null hypothesis H_0 : all $\beta_j = 0$, which means that there is no dependence of y on the k fixed variates x_1, x_2, ..., x_k.

ANALYSIS OF VARIANCE FOR TESTING FOR THE MULTIPLE LINEAR REGRESSION

Source of variation	*df*	*SS*	*MS*	F_0
Due to multiple linear regression	k	$\sum b_j P_j$	*MSR*	$\frac{MSR}{MSE}$
Error	$n-k-1$	$\sum_i (y_i - \bar{y})^2 - \sum_j b_j P_j$	*MSE*	
Total	$n-1$	$\sum_i (y_i - \bar{y})^2$	—	

DATA ANALYSIS

SUMMARY STATISTICS

```
Cement(comp_1)  Blast_Furnace_Slag(comp_2) Fly_Ash(comp_3)  Water(comp_4)
Min.   :102.0   Min.   :  0.0             Min.   :  0.00   Min.   :121.8
1st Qu.:192.4   1st Qu.:  0.0             1st Qu.:  0.00   1st Qu.:164.9
Median :272.9   Median : 22.0             Median :  0.00   Median :185.0
Mean   :281.2   Mean   : 73.9             Mean   : 54.19   Mean   :181.6
3rd Qu.:350.0   3rd Qu.:142.9             3rd Qu.:118.30   3rd Qu.:192.0
Max.   :540.0   Max.   :359.4             Max.   :200.10   Max.   :247.0
Superplasticizer(comp_5) Coarse_Aggregate(comp_6) Fine_Aggregate(comp_7)    Age(day)
Min.   : 0.000           Min.   : 801.0           Min.   :594.0           Min.   :  1.00
1st Qu.: 0.000           1st Qu.: 932.0           1st Qu.:731.0           1st Qu.:  7.00
Median : 6.400           Median : 968.0           Median :779.5           Median : 28.00
Mean   : 6.205           Mean   : 972.9           Mean   :773.6           Mean   : 45.66
3rd Qu.:10.200           3rd Qu.:1029.4           3rd Qu.:824.0           3rd Qu.: 56.00
Max.   :32.200           Max.   :1145.0           Max.   :992.6           Max.   :365.00
Concrete_compressive_strength
Min.   : 2.33
1st Qu.:23.71
Median :34.45
Mean   :35.82
3rd Qu.:46.13
Max.   :82.60
```

SHAPIRO-WILK TEST

Variable	W-value	p-value
Cement	0.95896	< 2.2e-16
Blast Furnace Slag	0.81241	< 2.2e-16
Fly Ash	0.762	< 2.2e-16
Water	0.98039	1.463e-10
Superplasticizer	0.86603	< 2.2e-16
Coarse Aggregate	0.98245	8.347e-10
Fine Aggregate	0.98067	1.842e-10
Age(day)	0.59071	< 2.2e-16
Concrete Compressive Strength	0.97979	9.01e-11

Since the p-values of the variables are less than 0.05, we can say that the variables deviate from a normal distribution.

DATA ANALYSIS

ANOVA

Null hypothesis : There is no effect of the amount of each factor used on the overall strength of the concrete mixture.

Against

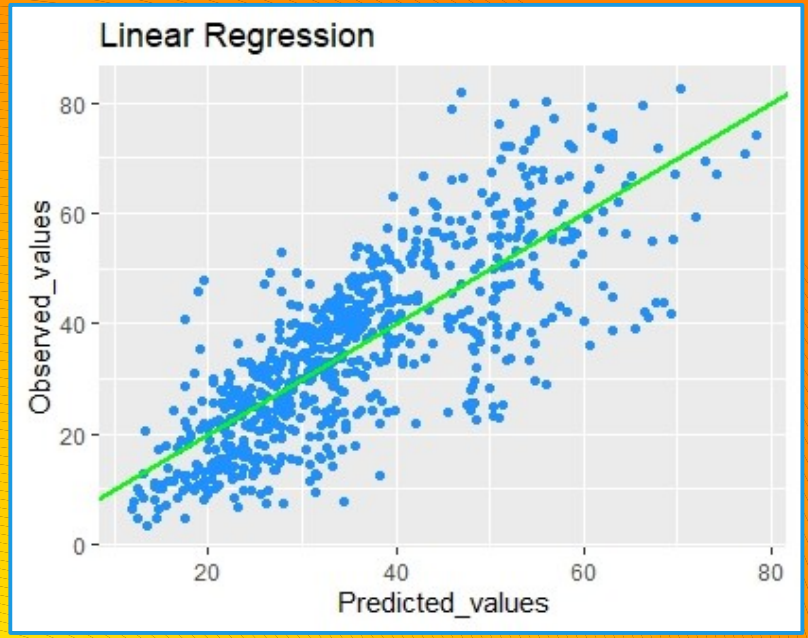
Alternative hypothesis : There is a significant effect of the amount of each factor used on the overall strength of the concrete mixture.

```
Analysis of Variance Table

Response: Concrete_compressive_strength
                              Df Sum Sq Mean Sq  F value    Pr(>F)
Cement.comp_1.                 1  71173   71173 658.1385 < 2.2e-16 ***
Blast_Furnace_Slag.comp_2.     1  22961   22961 212.3185 < 2.2e-16 ***
Fly_Ash.comp_3.                1  21639   21639 200.0939 < 2.2e-16 ***
Water.comp_4.                  1  11458   11458 105.9500 < 2.2e-16 ***
Superplasticizer.comp_5.       1   1374    1374  12.7097 0.0003806 ***
Coarse_Aggregate.comp_6.       1    256     256   2.3646 0.1244260
Fine_Aggregate.comp_7.         1      1       1   0.0066 0.9353645
Age.day.                       1  47902   47902 442.9521 < 2.2e-16 ***
Residuals                   1021 110413     108
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above anova table results,we can infer that there is a significant effect of the factors Water ,Cement, Blast Furnace Slag, Fly Ash, Superplasticizer, Age on the response variable Concrete Compressive Strength .Based on the above observations We reject the null hypothesis that there is no effect of the amount of each factor used on the overall strength of the concrete mixture.

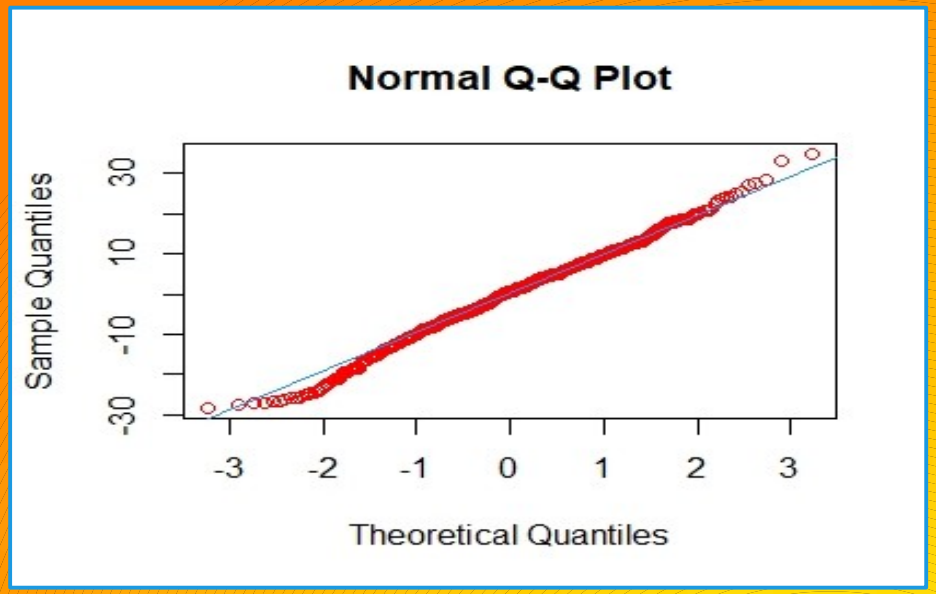
MULTIPLE LINEAR REGRESSION

```
Call:
lm(formula = Concrete_compressive_strength ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-28.186  -6.266   0.723   6.739  34.713

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                33.6914     0.7431  45.340  < 2e-16 ***
Cement.comp_1.             12.3668     1.0045  12.311  < 2e-16 ***
Blast_Furnace_Slag.comp_2.  8.7816     0.9902   8.868  < 2e-16 ***
Fly_Ash.comp_3.             5.6156     0.9221   6.090 1.76e-09 ***
Water.comp_4.              -3.3240     0.9476  -3.508 0.000477 ***
Superplasticizer.comp_5.    0.3181     0.1039   3.061 0.002281 **
Coarse_Aggregate.comp_6.    1.4424     0.8193   1.761 0.078700 .
Fine_Aggregate.comp_7.      1.8890     0.9613   1.965 0.049770 *
Age.day.                    7.1244     0.3876  18.383  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.31 on 793 degrees of freedom
Multiple R-squared:  0.6205,    Adjusted R-squared:  0.6167
F-statistic: 162.1 on 8 and 793 DF,  p-value: < 2.2e-16
```



From the above summary we can see that that R^2 value is 0.6205. This means that 62% of the variability observed in the target variable (Concrete Compressive Strength) can be explained by this linear regression model. Therefore we can say that this model fitting is moderately good .

MULTIPLE LINEAR REGRESSION



Shapiro-Wilk Test For Residuals :

Variable	W-value	p-value
Residuals	0.99528	0.0144

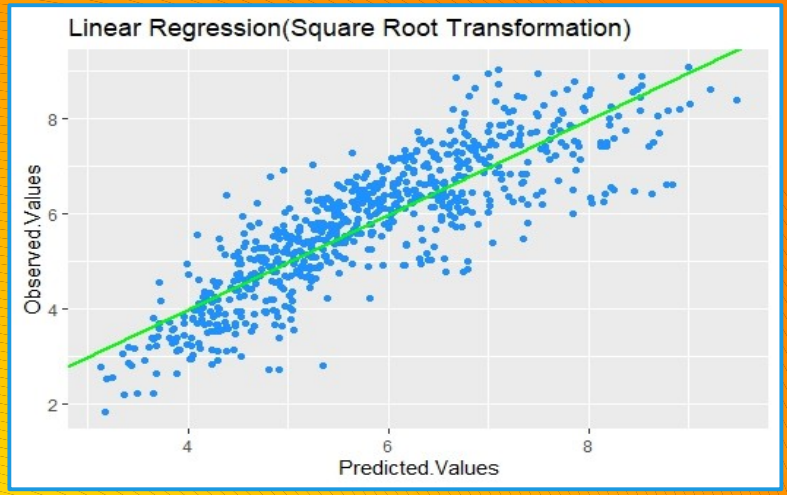
From the above table we can conclude that as that p-value of the residuals is less that 0.05,the residuals deviate from a normal distribution. However the deviation is not significant and the residuals are very much close to a normal distribution.

MULTIPLE LINEAR REGRESSION

SQUARE ROOT TRANSFORMATION:

$$\sqrt{y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip} + \epsilon_i \quad \text{for } i = 1,2,\ldots,n$$

```
Call:
lm(formula = Concrete_compressive_strength ~ ., data = strain)

Residuals:
     Min       1Q   Median       3Q      Max
-2.56517 -0.44820  0.05759  0.50578  2.20151

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.749527   3.184312  -0.235   0.8140
Cement.comp_1.             0.299337   0.017885  16.737  < 2e-16 ***
Blast_Furnace_Slag.comp_2. 0.103443   0.009325  11.093  < 2e-16 ***
Fly_Ash.comp_3.            0.059146   0.010071   5.873 6.31e-09 ***
Water.comp_4.             -0.345373   0.070809  -4.878 1.30e-06 ***
Superplasticizer.comp_5.   0.153254   0.031192   4.913 1.09e-06 ***
Coarse_Aggregate.comp_6.   0.078472   0.038358   2.046   0.0411 *
Fine_Aggregate.comp_7.     0.046024   0.036620   1.257   0.2092
Age.day.                   0.229534   0.007843  29.265  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7383 on 793 degrees of freedom
Multiple R-squared:  0.7394,    Adjusted R-squared:  0.7367
F-statistic: 281.2 on 8 and 793 DF,  p-value: < 2.2e-16
```



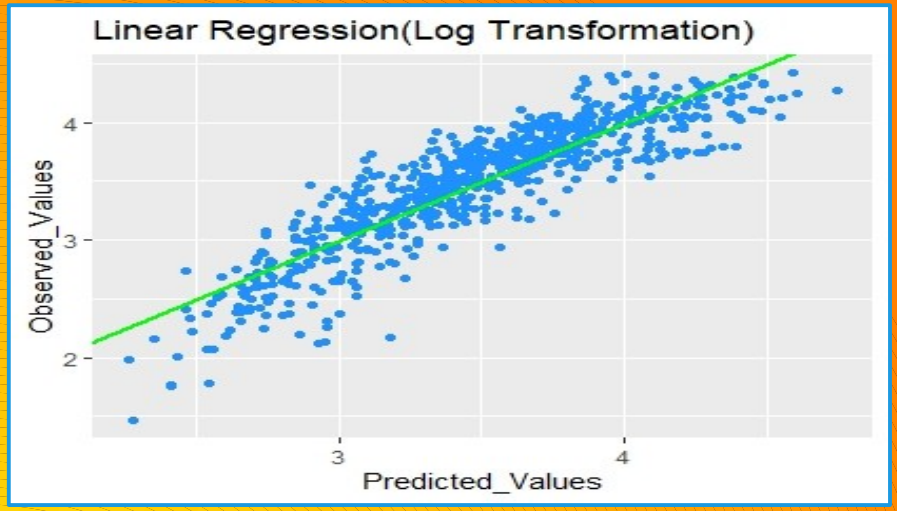
From the above summary we can see that that R2 value is 0.7383. This means that around 74 % of the variability observed in the target variable (Concrete Compressive Strength) can be explained by this linear regression model. Therefore we can say that this model fitting is a good fiitting. Moreover it has a better performance than simple Multiple regression and ridge Regression models.

MULTIPLE LINEAR REGRESSION

LOG TRANSFORMATION:

$$\log y_i = \beta_0 + \beta_1 \log x_{i1} + \beta_2 \log x_{i2} + \beta_3 \log x_{i3} + \ldots + \beta_p \log x_{ip} + \epsilon_i \quad \text{for } i=1,2,\ldots,n$$

```
Call:
lm(formula = Concrete_compressive_strength ~ ., data = mmtrain)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00822 -0.13245  0.02108  0.15205  0.62134

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.810461   2.559428   1.880   0.0605 .
Cement.comp_1.               0.730826   0.036156  20.213  < 2e-16 ***
Blast_Furnace_Slag.comp_2.   0.065080   0.005966  10.909  < 2e-16 ***
Fly_Ash.comp_3.              0.024992   0.006193   4.035 5.98e-05 ***
Water.comp_4.               -0.981764   0.137034  -7.164 1.79e-12 ***
Superplasticizer.comp_5.     0.068137   0.013809   4.934 9.81e-07 ***
Coarse_Aggregate.comp_6.     0.020009   0.165774   0.121   0.9040
Fine_Aggregate.comp_7.      -0.256455   0.133045  -1.928   0.0543 .
Age.day.                     0.293773   0.007881  37.277  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2395 on 793 degrees of freedom
Multiple R-squared:  0.7898,    Adjusted R-squared:  0.7876
F-statistic: 372.3 on 8 and 793 DF,  p-value: < 2.2e-16
```



From the above summary we can see that that R2 value is 0.7898. This means that around 79 % of the variability observed in the target variable (Concrete Compressive Strength) can be explained by this linear regression model. Therefore we can say that this model fitting is a good fiitting. Moreover it has a better performance than simple Multiple regression and ridge Regression models.

RIDGE REGRESSION

Multicollinearity : It is defined as a problem in predictive modelling if there is a strong linear relationship among the independent variables

Consequences of multicollinearity :

- Parameters of the model are highly unstable as standard error of their estimates are highly inflated
- The model may look ideal for a given sample but it fails to predict accurately for out of sample data

We use variance inflation factor(VIF) to check for multicollinearity.

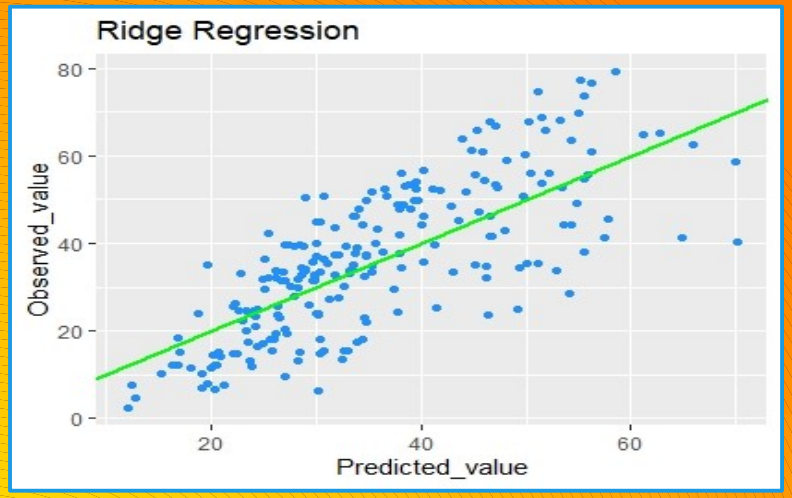
The variance inflation factor (VIF) is the ratio (quotient) of the variance of estimating some parameter in a model that includes multiple other terms (parameters) by the variance of a model constructed using only one term.

```
vif(lin.fit)
        Cement.comp_1. Blast_Furnace_Slag.comp_2.          Fly_Ash.comp_3.
              7.607310                  7.392309                 6.410713
         Water.comp_4.  Superplasticizer.comp_5.  Coarse_Aggregate.comp_6.
              6.769262                  2.900870                 5.060637
Fine_Aggregate.comp_7.                 Age.day.
              6.967396                  1.132361
```

Since the values for VIF of Cement, Blast Furnace Slag , Fly Ash, Water ,Fine Aggregate are greater than 5 ,hence as a remedial measure we will use ridge regression.

RIDGE REGRESSION

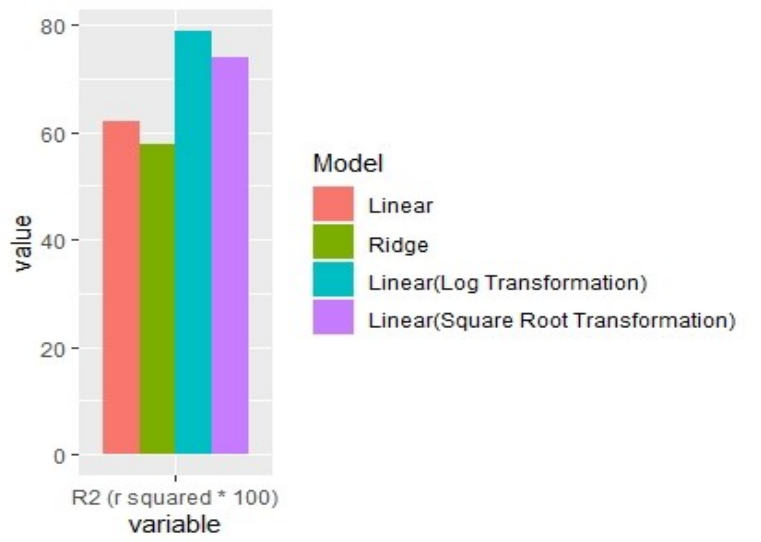
```
9 x 1 sparse Matrix of class "dgCMatrix"
                                       s0
(Intercept)                    32.8725231
Cement.comp_1.                  8.4761595
Blast_Furnace_Slag.comp_2.      5.0135270
Fly_Ash.comp_3.                 2.1353983
Water.comp_4.                  -4.9245779
Superplasticizer.comp_5.        0.4494073
Coarse_Aggregate.comp_6.       -0.6862986
Fine_Aggregate.comp_7.         -1.0642751
Age.day.                        6.5832110
```




```
> R
[1] 0.5783586
```

The estimate of intercept in case of Ridge regression is almost close to Linear regression (33.6914).The R^2 value calculated in this case is 0.5783586. This means that almost 58% of the variability observed in the target variable (Concrete Compressive Strength) can be explained by this regression model. Therefore we can say that this model fitting is moderately good .

CONCLUSION



MODEL	R-SQUARED
LINEAR	0.62
RIDGE	0.58
LINEAR(LOG TRANSFORMATION)	0.79
LINEAR(SQUARE ROOT TRANSFORMATION)	0.74

- The main objective in model building situations is to choose from a large set of factors those that should be included in the 'best' model. Various factors have different effect in model building.
- From the above r squared values it is clearly visible that among all the various models, the Linear Model(Log Transformation) model is the best one as about 79 % of the variability observed in the target variable (Concrete Compressive Strength) can be explained by this linear regression model.
- Therefore we can say that this model fitting is a good fitting. Moreover it has a better performance than simple Multiple regression and ridge Regression models.

REFERENCE!

The information used in project have been collected from the sources given below:

- https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength
- https://en.wikipedia.org/wiki/Concrete#:~:text=Concrete%20is%20an%20artificial%20composite,loose%20stones%2C%20and%20sand).
- https://www.statology.org/ridge-regression-in-r/
- http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm
- http://www.stat.yale.edu/Courses/1997-98/101/anovareg.htm
- Introduction to Linear Regression Analysis
  Book by Douglas Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining
- https://statsandr.com/blog/multiple-linear-regression-made-simple/
- https://rc2e.com/linearregressionandanova

THANK YOU