ANALYSIS OF CONCRETE'S COMPRESSIVE STRENGTH

Statistics 512: Semester Project Report

Due Date: 26th April, 2019

Team Members: Mayur Deo, Rishi Ganesan

Index

1. Introduction

Concrete is one most used substances used by mankind after water. It forms the basis of modern architecture and all other aspects of building construction. Some concrete wonders include the Burj Khalifa (Dubai) or the Shanghai Tower (China). Being such a crucial material for the advancement of human civilization, there is a constant drive to better understand and improve this material's overall life cycle. The main goal of this project is to identify the relationship between the compressive strength of concrete and its material composition.

The dataset used in this project was obtained from the University of California Irvine Machine Learning Repository[1]. The concrete data was determined in a laboratory setting by measuring the concrete's compressive strength given its mixture's material composition and its mixture's age.

The objectives of the methods and results expounded in the following sections are to: Inspect the effect of the different mixture components on the concrete's compressive strength, figure out the most significant of these components and find out the combination of components in the concrete's mixture that produces the highest compressive strength.

[1] Data obtained from: http://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength

2. Methods

2.1. Dataset Description

The Concrete Strength dataset contains 1030 instances, one response variable (concrete compressive strength), eight input variables and zero missing values. The specific description of each of the variables is shown in the table below:

Table 1: Description of Variables in the Concrete Data

Name	Unit of Measurement	Unit Type	Type
Concrete Compressive Strength	MPa	Quantitative	Response
Age	Days	Quantitative	Input
Fine Aggregate	$\frac{kg}{m^3}$	Quantitative	Input
Coarse Aggregate	$\frac{kg}{m^3}$	Quantitative	Input
Superplasticizer	$\frac{kg}{m^3}$	Quantitative	Input
Water	$\frac{kg}{m^3}$	Quantitative	Input
Fly Ash	$\frac{kg}{m^3}$	Quantitative	Input
Blast Furnace Slag	$\frac{kg}{m^3}$	Quantitative	Input
Cement	$\frac{kg}{m^3}$	Quantitative	Input

2.2. Preliminary Analysis

Before beginning to build any linear regression model, some basic pre-modelling analysis was done. This analysis included:

- Determining column classes in the dataset: All the dataset variables are numeric.

- Histograms of the response variable and the predictors (Figure 1-Figure 9): It can be seen that the response is approximately normally distributed (Figure 1) and the covariates are pretty skewed to the left.

- Summary statistics of the response variable and predictors (Figure 10): It was found out that the age of the concrete sample was very sparse because its variance (and standard deviation) values were very high. Upon further investigation, it was found out that this variable takes few discrete values, thus it could also be modelled as a factor.

- Scatterplot matrix of the entire dataset (Figure 11): From looking at this matrix it was determined that the content of cement in the concrete sample is positively related to the sample's compressive strength. The relationships between the rest of the covariates and the response isn't very clear just from looking at the scatterplot matrix. A better option would be to look at the correlation plot.

- Correlation plot and correlation matrix of the entire dataset (Figure 12): From these it was confirmed that the cement content is the predictor with the largest positive correlation with the response and that water content is the predictor with the largest negative correlation with the response.

The goal of this task was to get more familiar with the dataset before jumping into modelling any relationships. This step helped in the process of making sure the data fed into the models is complete and accurate, checking for any possible mistakes (missing values or outliers) and getting a rough view of the relationships between the variables in the dataset, both, dependent and independent . Thus, enabling to build better models for analyzing the given data. The results of this analysis is outlined in Appendix B.

2.3. Model Building and Selection

After the data exploration phase, the next step was to develop the actual regression model. The goal was to develop the most parsimonious model that would best explain the relationship between the given set of predictors and the response. The modelling process was an iterative process, meaning, the same sequence of procedures were applied in a loop to transform one model into another.

The first model considered was the basic model including all the predictors but considering the variable *Age* as a factor, as determined in the preliminary data analysis. A forward variable selection was performed on this model, this resulted in a reduced model by eliminating the *Superplasticizer, Coarse Aggregate* and *Fine Aggregate* covariates. On conducting a *F-test* and generating a summary of the model, it was obtained that the current model had all significant regressors and a high R-square value, showing its viability for further analysis. The next task was to check for the assumptions used in the process of regression. This was determined from analyzing the model diagnostics, more precisely, by checking if the model residuals satisfied the constant variance and normality assumptions from the Pearson residual plots (Figure 13) and the Q-Q plot (Figure 14) respectively. This resulted in the power transformation of the predictors *Cement, Blast Furnace* and *Fly Ash* to remedy the non-constant variance involved. The current model was a reduced and transformed version of the original model. The homoscedasticity and normality assumptions were once more tested, now on this new model, thus the iterative nature of the model building process. This model satisfied both assumptions, explained *82.39%* of the variability in the data while being a simpler version of the original model, thus this was the selected model to analyze the given dataset.

After selecting a final model, a more intense study of the diagnostics was performed and inferences were made on the results it produced. The additional diagnostic methods included a complete outlier analysis to check how extreme values affect the model and hence the response variable. A single outlier was obtained from the test with a

Bonferonni adjusted p-value less than 0.05 resembling its significance. Despite the low *Bonferonni adjusted p-value*, on testing the model developed from the removal of the outlier, it was identified that the outlier was ineffective on the parameter estimates or any other of the model's results, thus it was decided to continue to include this point and to move forward with the model towards further analysis. To further identify the various influential points, a diagnostic plots and an influence plot (Figures 15 & 16) were obtained for the current model. On analysis of the visually separated points in the diagnostic plots, there was a suspicion that some of these points may be influential points but upon removing each point one by one and re-analyzing the model, it was determined that these points didn't affect the model parameters, hence these points were not considered influential and kept. The diagnostics methods conducted helped test the robustness and the stability of the model generated. Thus, the current model was considered as the final model for further inferential analysis to answer the research questions.

The inferential methods were mainly focused on interpreting the information given by the model's parameter estimates such as their significance, estimating 95% confidence intervals and performing hypothesis tests using ANOVA at a 0.05 significance level, the null hypothesis being the insignificance of the regressors involved and alternative hypothesis being the significance of at least one of the regressors.

3. Results

During the project, the following models were considered and then rejected to obtain the final model shown above. The rejection of previous models was due to multiple reasons as explained in the *Methods* section.

Table 2: Description of Analyzed Models

Model	Predictors Eliminated	Reason
model1	None	-
model2	Coarse Aggregate, Superplasticizer and Fine Aggregate.	Forward variable selection with AIC criterion.
model 3	None	Transformations were conducted to remedy non constant variance.

After completion of the inferential analysis of the model3, we selected it as the final model due to the highly significant regressors, high R^2value and a good linear relationship between the response and the predictors, with no assumptions being violated. The final model with the transformation was as follows:

$$\begin{aligned} &Concrete\ Compressive\ Strength\ (MPa) \\ &\quad = Cement^{\wedge \mathbf{0.33}} + factor(Age\ (days)) + log(Blastfurnaceslag) + Water \\ &\quad + log(flyash) \end{aligned}$$

The final model summary (Figure 17) as well as the 95% Confidence Interval on its parameter estimates (Figure 18) are shown in Appendix B . From the parameter estimates and the significance of regressors in the final model, it can be observed that the regressors cement, Age, Blast Furnace Slag, and Fly ash have a positive relationship with the response variable i.e. Compressive strength of concrete. While the negative parameter estimate of the regressor water indicates it’s inverse relationship with the response variable.

The ANOVA of the final model is shown in the table below:

Table 3: ANOVA of Final Model

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-value	Pr (>F)
Cement	1	70374	70374	1444.638	2.2e-16 ***
Age (days)	13	115975	115975	183.134	2.2e-16 ***
Blast Furnace Slag	1	33683	33683	691.451	2.2e-16 ***
Water (kg/m3)	1	15623	15623	320.715	2.2e-16 ***
Fly Ash	1	2221	2221	45.594	.448e-11 ***
Residuals	1012	49298	49		

4. Discussion

In this section the answers derived from analyzing the concrete strength data was used to answer the objectives outlined at the end of the Introduction. These were answered by interpreting the significance and effect of the final model's parameters (regressor's coefficients and R^2value) and assessing the final model's predictive power through some basic predictive modelling techniques.

From Figure 17 in Appendix B it is shown that the components that have a significant effect on the compressive strength are *Cement, Blast Furnace, Water, Fly Ash* and the different factor levels representing the *Age* of the concrete sample. From the coefficient estimates of these predictors, it is observed that the addition of an extra unit of these component quantities in the mixture don't change the response that much but the aging of the sample can significantly affect the strength of the sample, take for example a sample 101 days old increases its compressive strength by 57 MPa approximately. An additional observation was that the regressors specified previously explained around 82 % of the variability in the concrete's compressive strength.

Next, the given data was split into a training and testing set. The training set was used to build the model with the previously selected and transformed regressors. Then this model's predictive accuracy was tested on the unseen data in testing set. The scoring function used to calculate the predictive accuracy was the Root Mean Square Error (RMSE). The trained model gave an out of sample RMSE of 7.01 which is a pretty low value. Therefore, this model will potentially be able to predict the strength of a future concrete sample given its component mixture and age, meaning these variables can be adjusted such that the compressive strength is maximized for each future sample.

5. References

- Kutner, Michael H. Applied linear statistical models.-5th ed. Michael H Kutner... [et al].p. cm. - (McGraw-HillfIrwin series Operations and decision sciences) Rev.ed. of: Applied linear regression models. 4th ed. c2004. ISBN 0-07-238688-6.

- Weisberg, Sanford, Applied linear regression / Sanford Weisberg, School of Statistics, University of Minnesota, Minneapolis, MN.—Fourth edition. ISBN 978-1-118-38608-8.

6. Appendix A: R Code

# BASIC MODEL (ALL PREDICTORS INCLUDED)

```
model1<- lm(df$`Concrete Compressive Strength (MPa)`~., data=df)

par(mfrow=c(2,2)) # Enables to see all four regresion diagnostic plots at the same time.
plot(model1)
```

#FORWARD VARIABLE SELECTION

```
scope <- ~ df$`Cement (kg/m3)`+df$`Blast Furnace (kg/m3)`+df$`Fly Ash
(kg/m3)`+df$`Superplasticizer(kg/m3)`+df$`Water (kg/m3)`+df$`Coarse Aggregate
(kg/m3)`+df$`Fine Aggregate (kg/m3)`+factor(df$`Age (days)`)

model1.forward<-step(model3,scope=scope,direction="forward")
```

#MODEL-2 GENERATION AND TESTING

```
model2<-lm(df$`Concrete Compressive Strength (MPa)` ~ df$`Cement (kg/m3)` +
factor(df$`Age (days)`) + df$`Blast Furnace (kg/m3)` + df$`Water (kg/m3)` + df$`Fly Ash
(kg/m3)`)

anova(model2)
```

#RESIDUAL PLOT (ASSUMPTION CHECK - CONSTANT VARIANCE)

```
residualPlots(model2)
```

#NORMALITY CHECK

```
qqPlot(stepmodel, id.no=2)
```

#POWER TRANSFORM AND TRANSFORMATION TEST

```
summary(b<-powerTransform(cbind(`Cement (kg/m3)`,`Blast Furnace (kg/m3)`,`Fly Ash
(kg/m3)`,`Water (kg/m3)`) ~ 1, df))

testTransform(b,c(0.33,0,0,1))#Testing the transformation being used
```

#TRANSFORMATIONS

```
cementtrans<-(df$`Cement (kg/m3)`^0.33-1)/0.33

blastftrans<-log(df$`Blast Furnace (kg/m3)`)

flyashtransform<-log(df$`Fly Ash (kg/m3)`)
```

#TRANSFORMED MODEL

```
model3<-lm(df$`Concrete Compressive Strength (MPa)` ~ cementtrans + factor(df$`Age
(days)`) + blastftrans + df$`Water (kg/m3)` + flyashtransform)
```

#CHECK FOR ASSUMPTIONS

```
residualPlots(model3)

qqPlot(model 3, id.no=2)
```

#OUTLIER TEST

```
outlierTest(transstepmodel)
```

#DIAGNOSTIC TEST

```
influenceIndexPlot(model 3)

influencePlot(model 3)
```

# TESTING EFFECTS OF REMOVING OUTLIERS

```
df2 <-df[-384,]

cementtrans2<-(df2$`Cement (kg/m3)`^0.33-1)/0.33

blastftrans2<-log(df2$`Blast Furnace (kg/m3)`)

flyashtransform2<-log(df2$`Fly Ash (kg/m3)`)

substitutemodel3<-lm(df2$`Concrete Compressive Strength (MPa)` ~ cementtrans2 +
factor(df2$`Age (days)`) + blastftrans2 + df2$`Water (kg/m3)` + flyashtransform2 )
```

```
anova(substitutemodel3)
summary(substitutemodel3)
confint(substitutemodel3,level=0.95)

# PREDICTIVE MODELLING
df3 <- df

colnames(df3)<- c("cement","blastfurnace","flyash","water","superp","ca",
"fa","age","strength")



# Variable Transformation
df3$cement<-(df3$cement^0.33-1)/0.33

df3$blastfurnace<-log(df3$blastfurnace)

df3$flyash<-log(df3$flyash)

# Data Splitting
ind <-  sample.int(n = nrow(df3), size = floor(.75*nrow(df3)), replace = F)

train<-df3[ind,]

test<-df3[-ind,]

# Model Building
transstepmodel3 <- lm(strength ~ cement+ factor(age) + blastfurnace + water + flyash,
data = train)

# Make Predictions
predictions <- predict(transstepmodel3,newdata = test)

#Calculate Score
RMSE <- sqrt(mean((test$strength-predictions)^2))
```

7. Appendix B: R Output

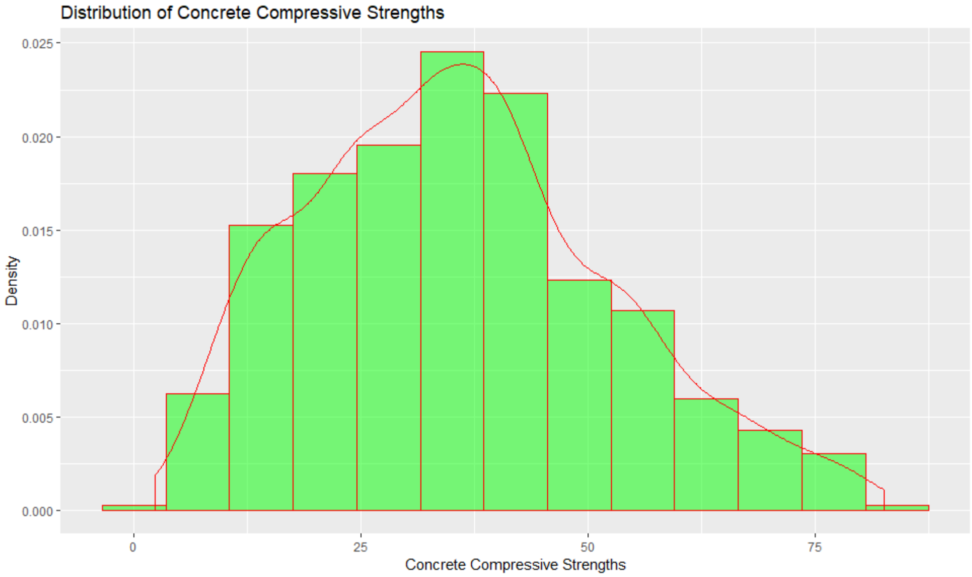


Figure 1: Distribution of Response Variable.

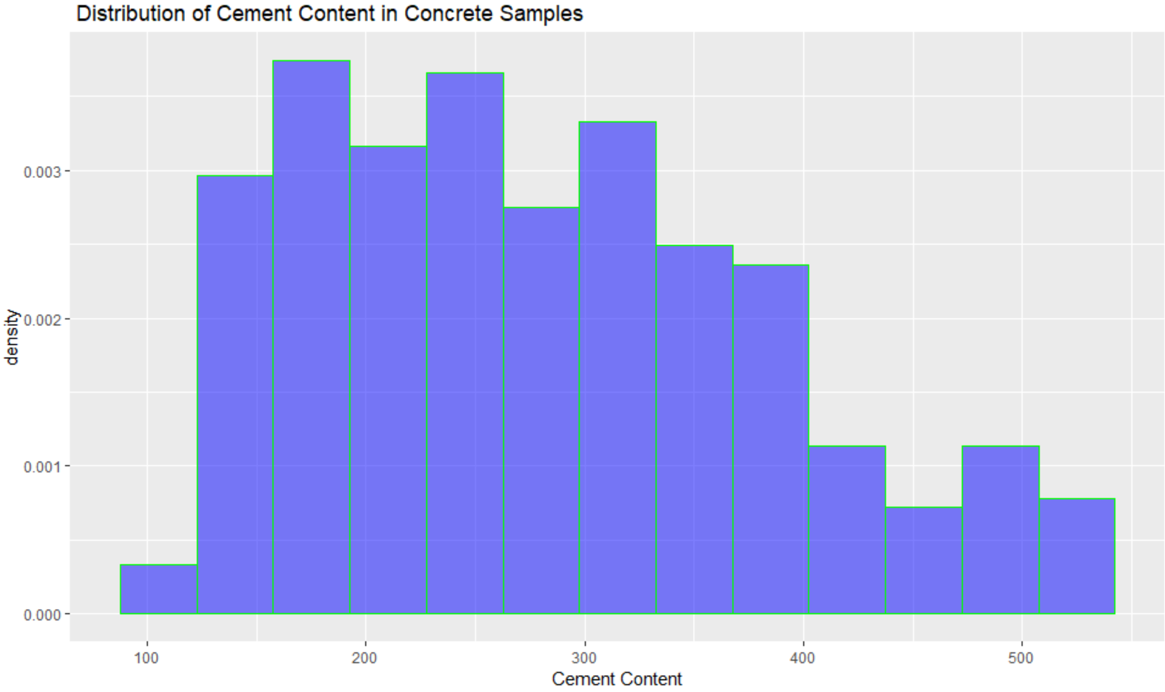


Figure 2: Distribution of Cement Content in Concrete Samples.

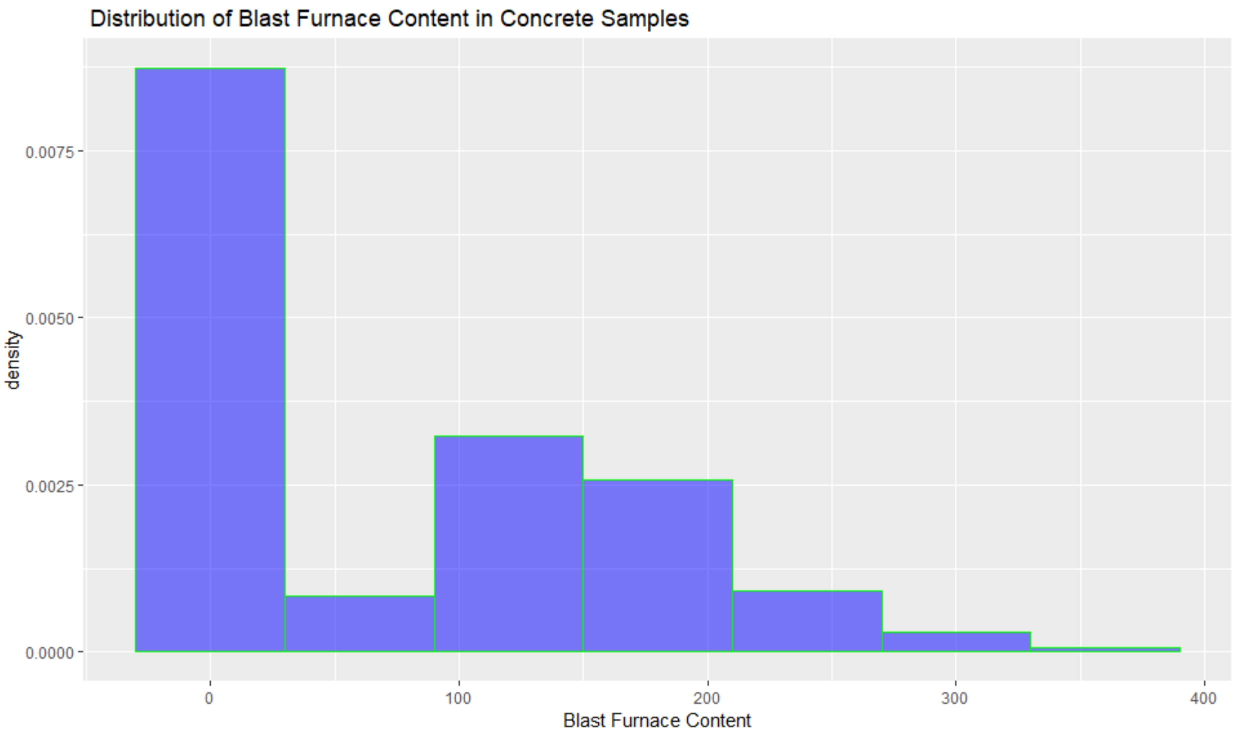


Figure 3: Distribution Content of Blast Furnace in Concrete Samples

.

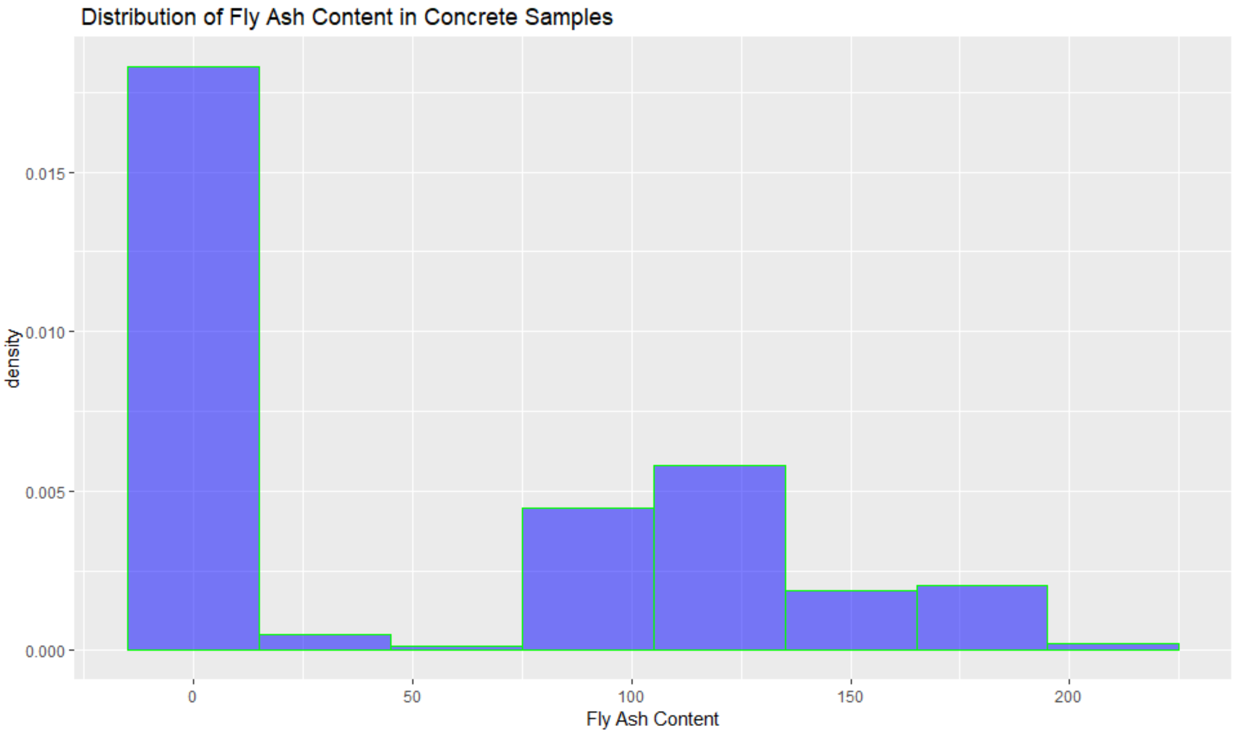


Figure 4: Distribution Content of Fly Ash in Concrete Samples.

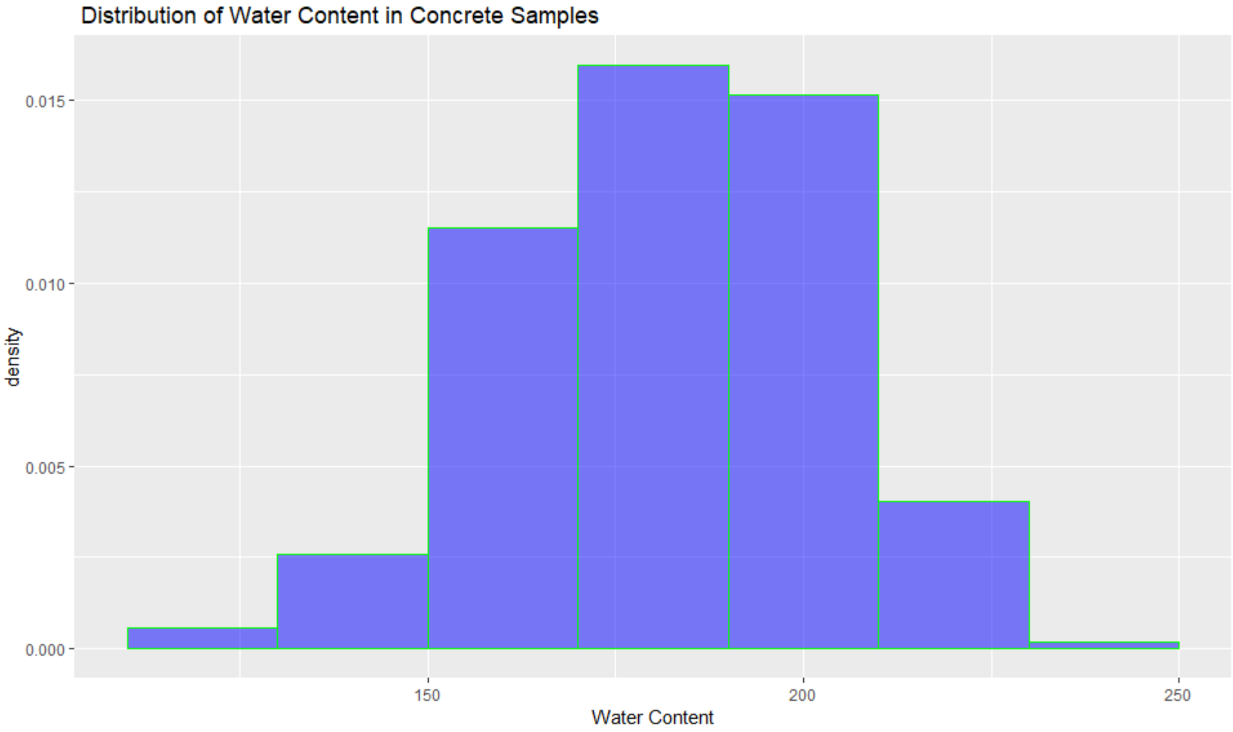


Figure 5: Distribution Content of Water in Concrete Samples.

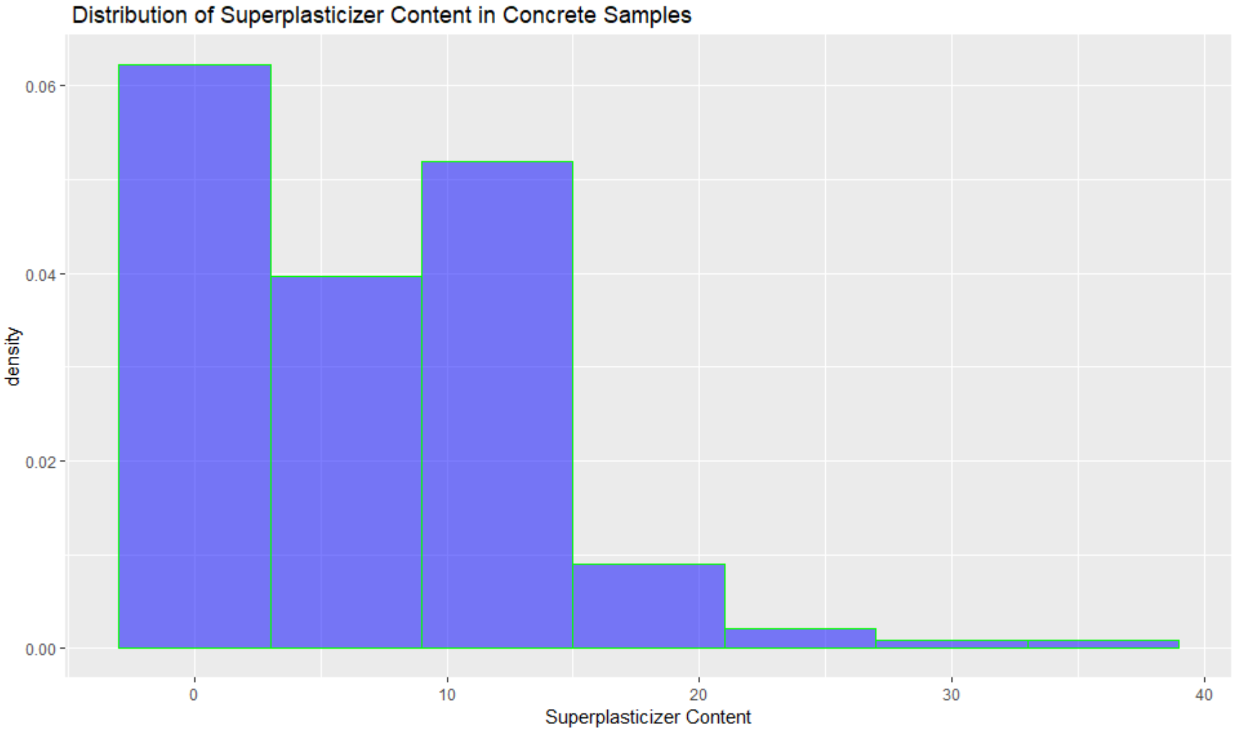


Figure 6: Distribution Content of Superplasticizer in Concrete Samples.

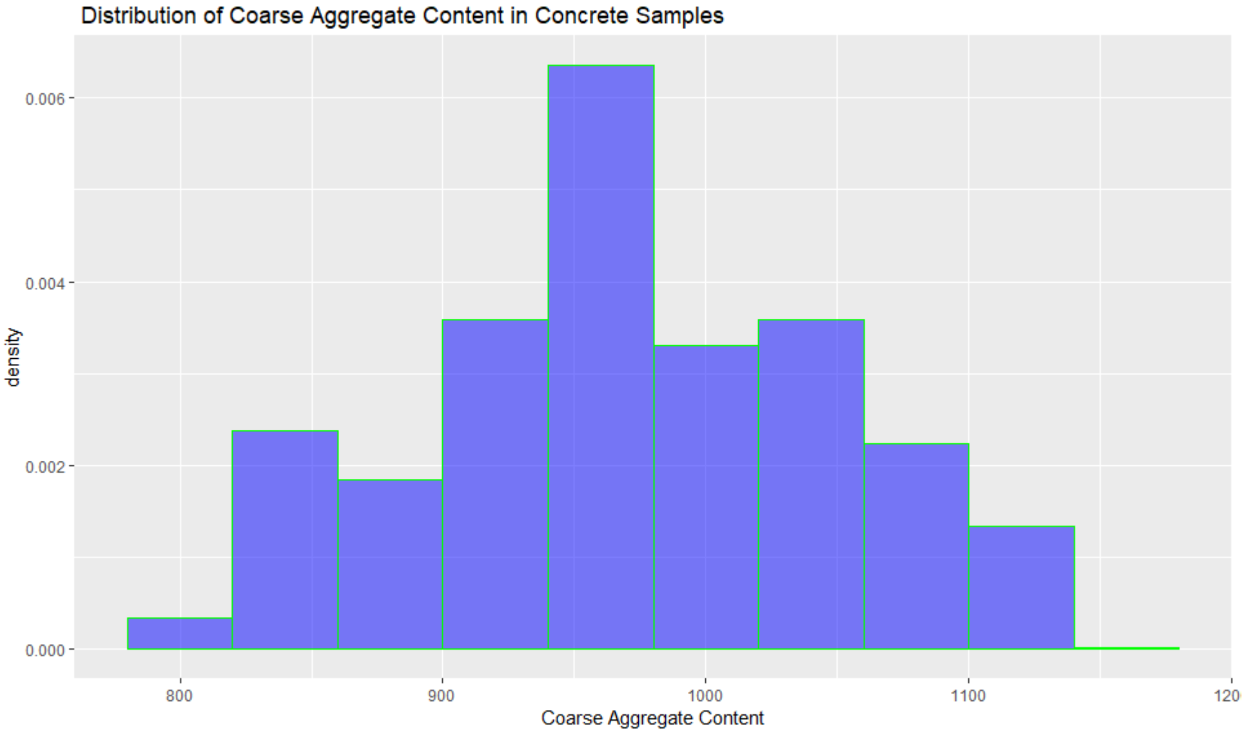


Figure 7: Distribution Content Coarse Aggregate in Concrete Samples.

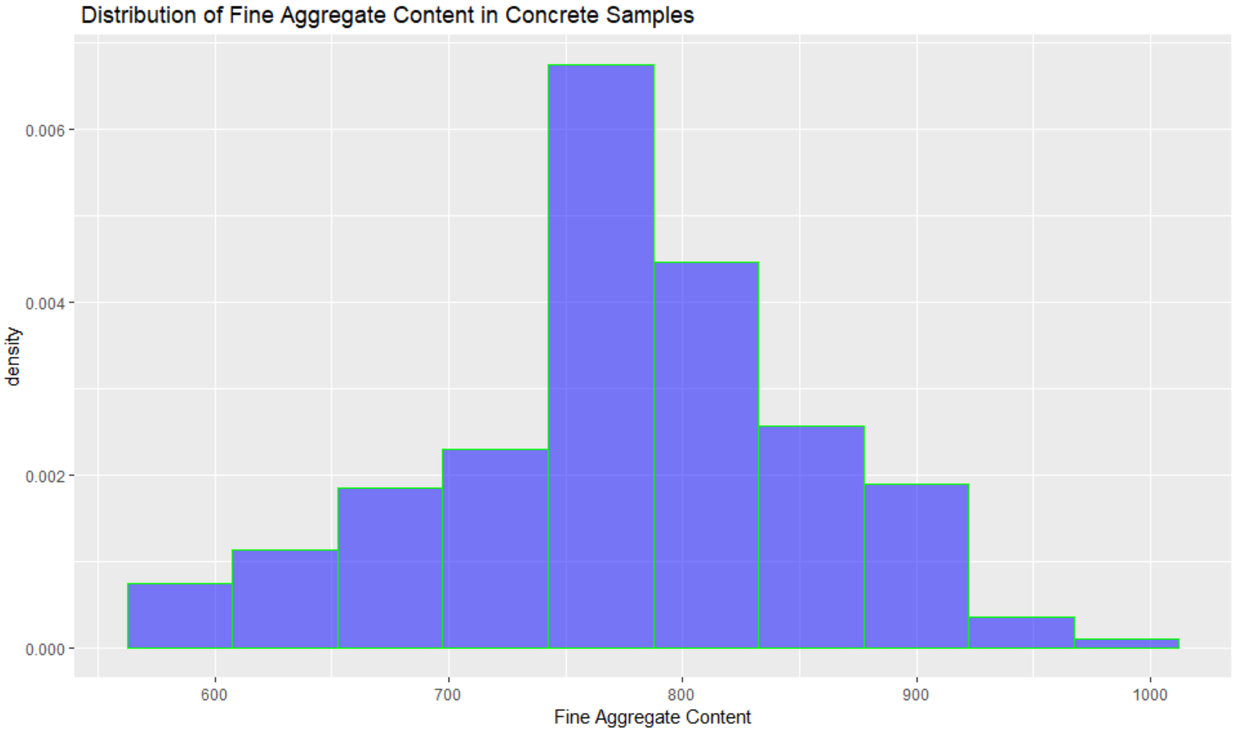


Figure 8: Distribution Content Fine Aggregate in Concrete Samples.

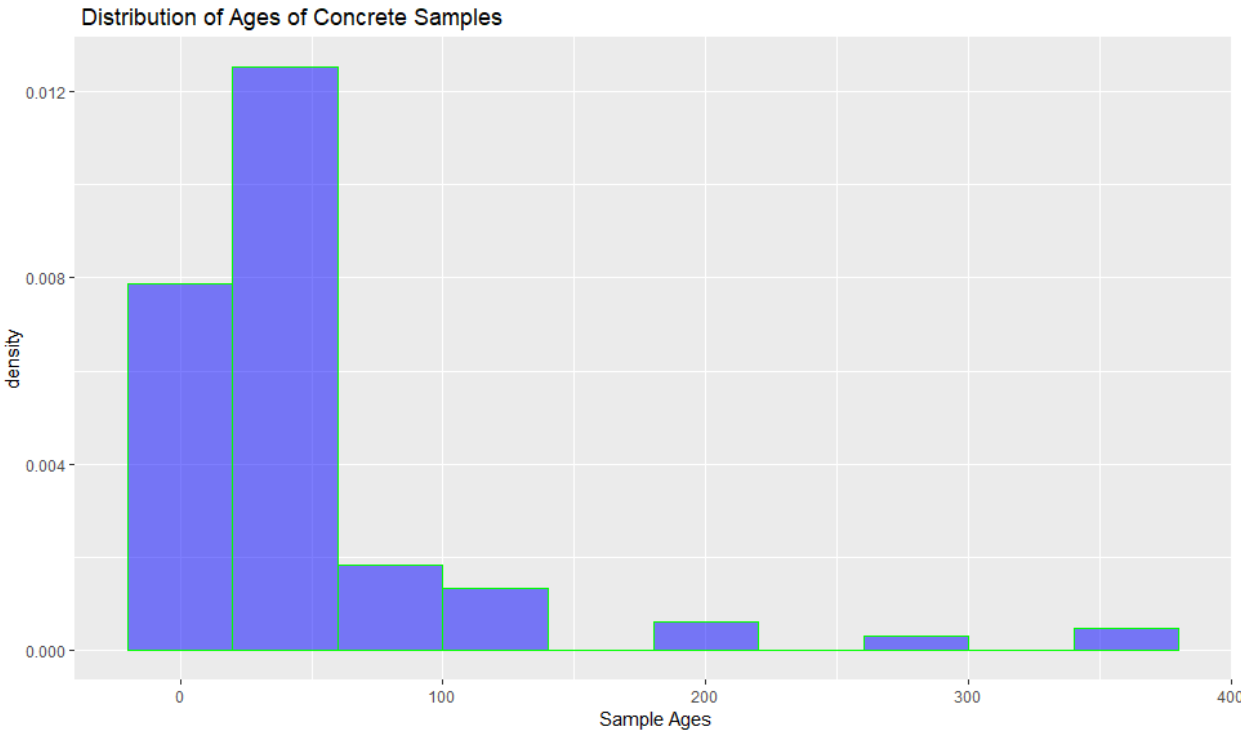


Figure 9: Distribution Ages of Concrete Samples.

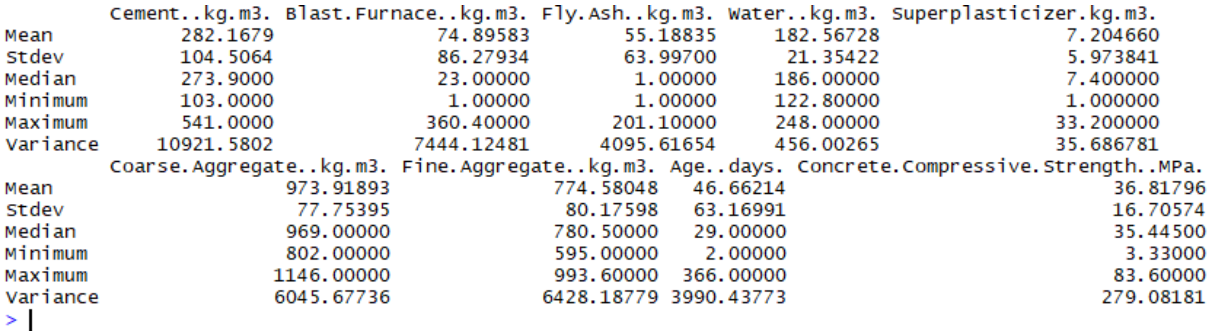

```
         Cement..kg.m3. Blast.Furnace..kg.m3. Fly.Ash..kg.m3. Water..kg.m3. Superplasticizer.kg.m3.
Mean           282.1679              74.89583        55.18835     182.56728                7.204660
Stdev          104.5064              86.27934        63.99700      21.35422                5.973841
Median         273.9000              23.00000         1.00000     186.00000                7.400000
Minimum        103.0000               1.00000         1.00000     122.80000                1.000000
Maximum        541.0000             360.40000       201.10000     248.00000               33.200000
Variance     10921.5802            7444.12481      4095.61654     456.00265               35.686781
         Coarse.Aggregate..kg.m3. Fine.Aggregate..kg.m3. Age..days. Concrete.Compressive.Strength..MPa.
Mean                    973.91893              774.58048   46.66214                            36.81796
Stdev                    77.75395               80.17598   63.16991                            16.70574
Median                  969.00000              780.50000   29.00000                            35.44500
Minimum                 802.00000              595.00000    2.00000                             3.33000
Maximum                1146.00000              993.60000  366.00000                            83.60000
Variance               6045.67736             6428.18779 3990.43773                           279.08181
> |
```

Figure 10: Summary Statistics of Dataset Variables.

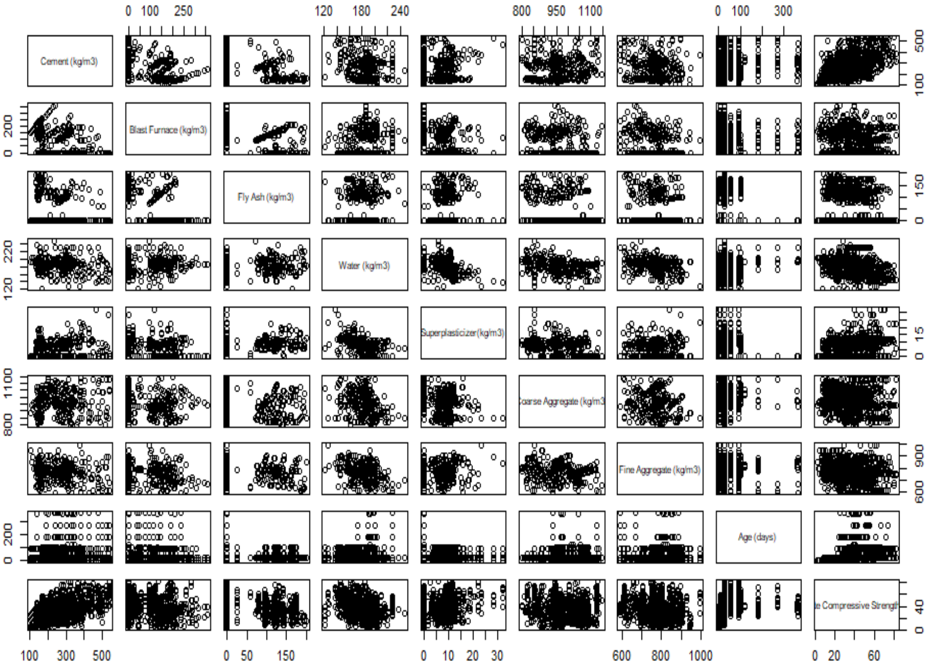


Figure 11: Scatterplot Matrix.

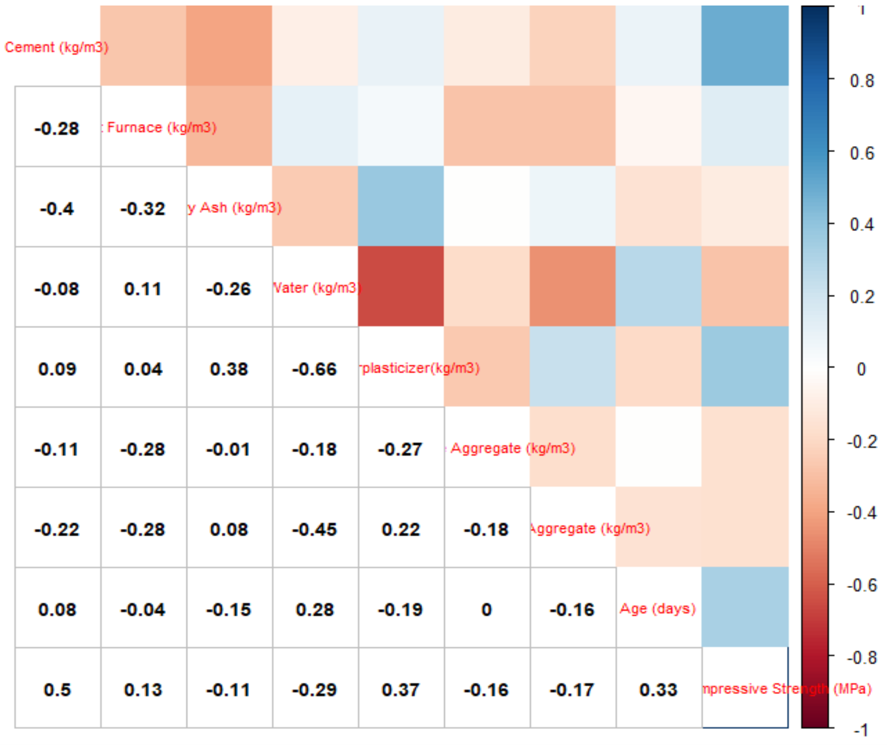


Figure 12: Correlation Plot.

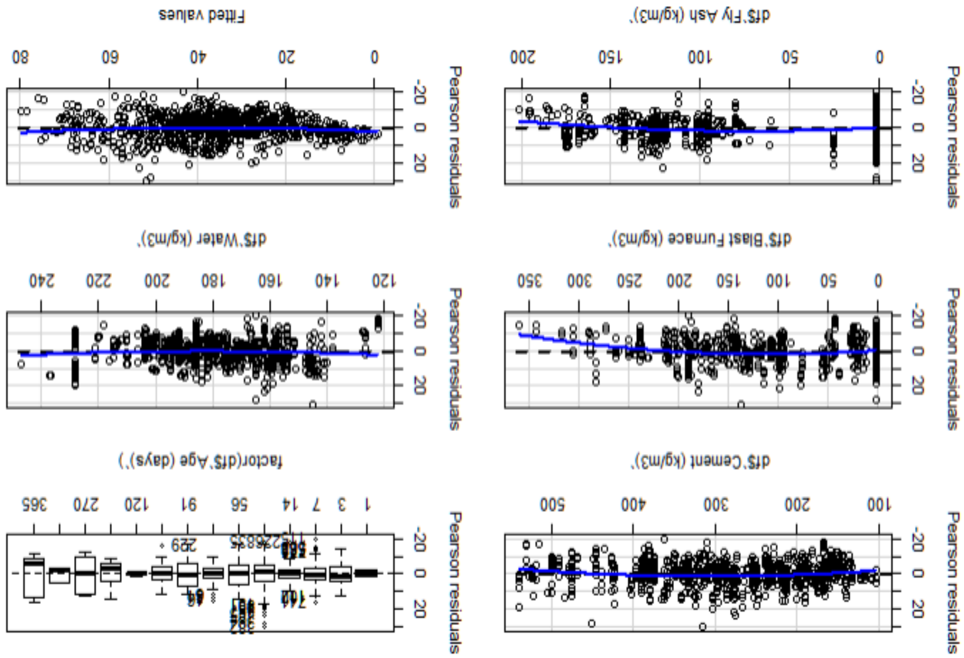


Figure 13: Pearson's Residual Plots for Basic Model

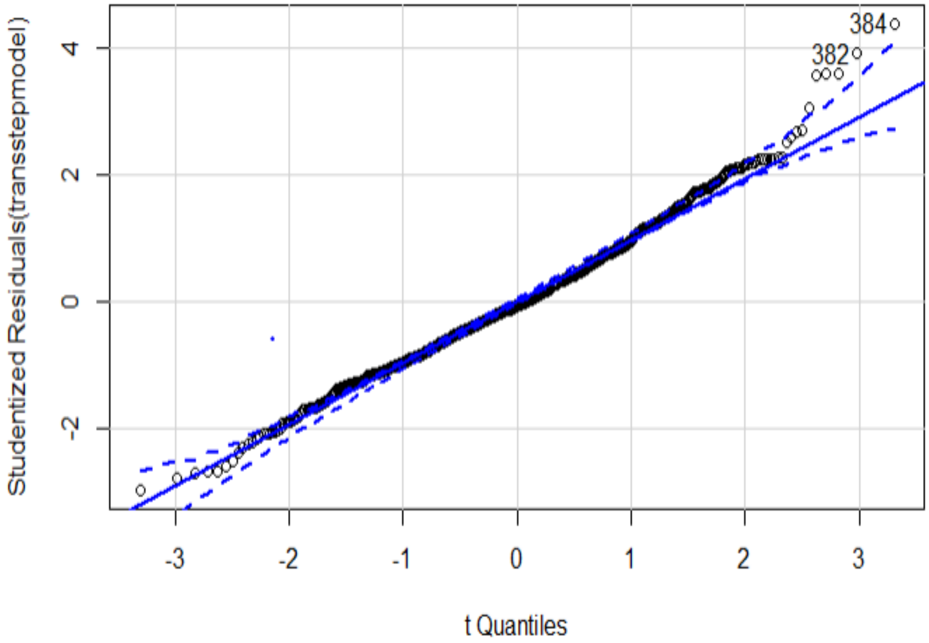


Figure 14: QQ Plot

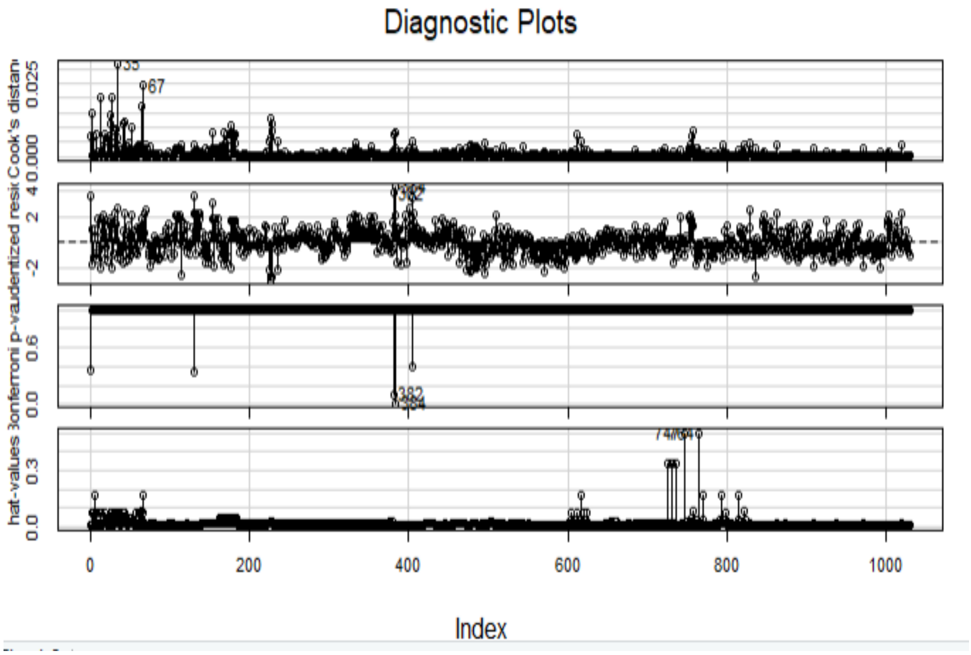


Figure 15: Diagnostic Plot

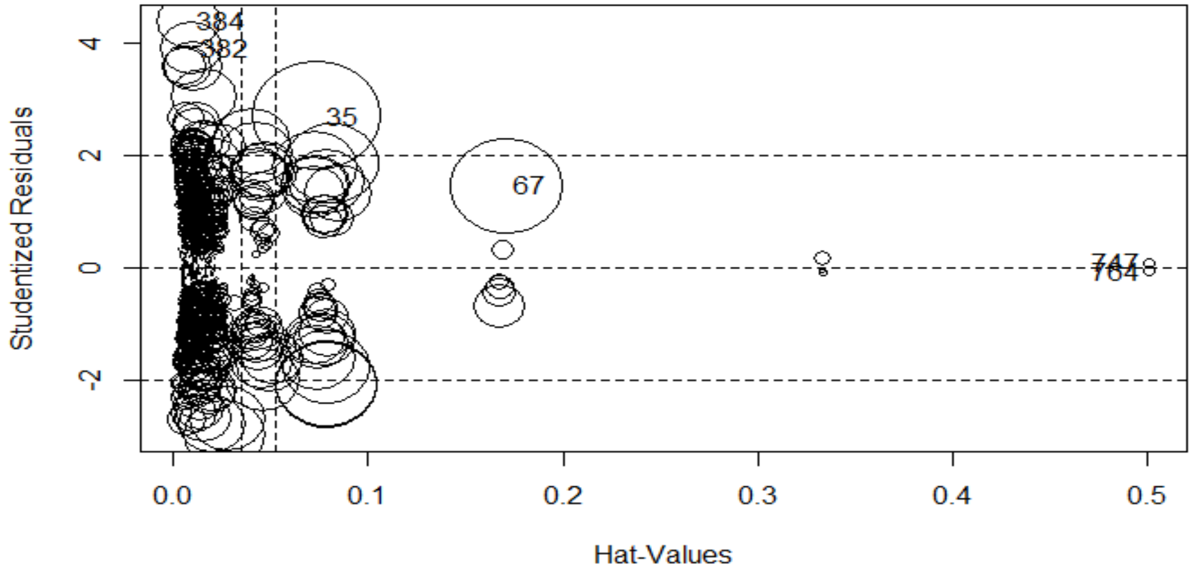
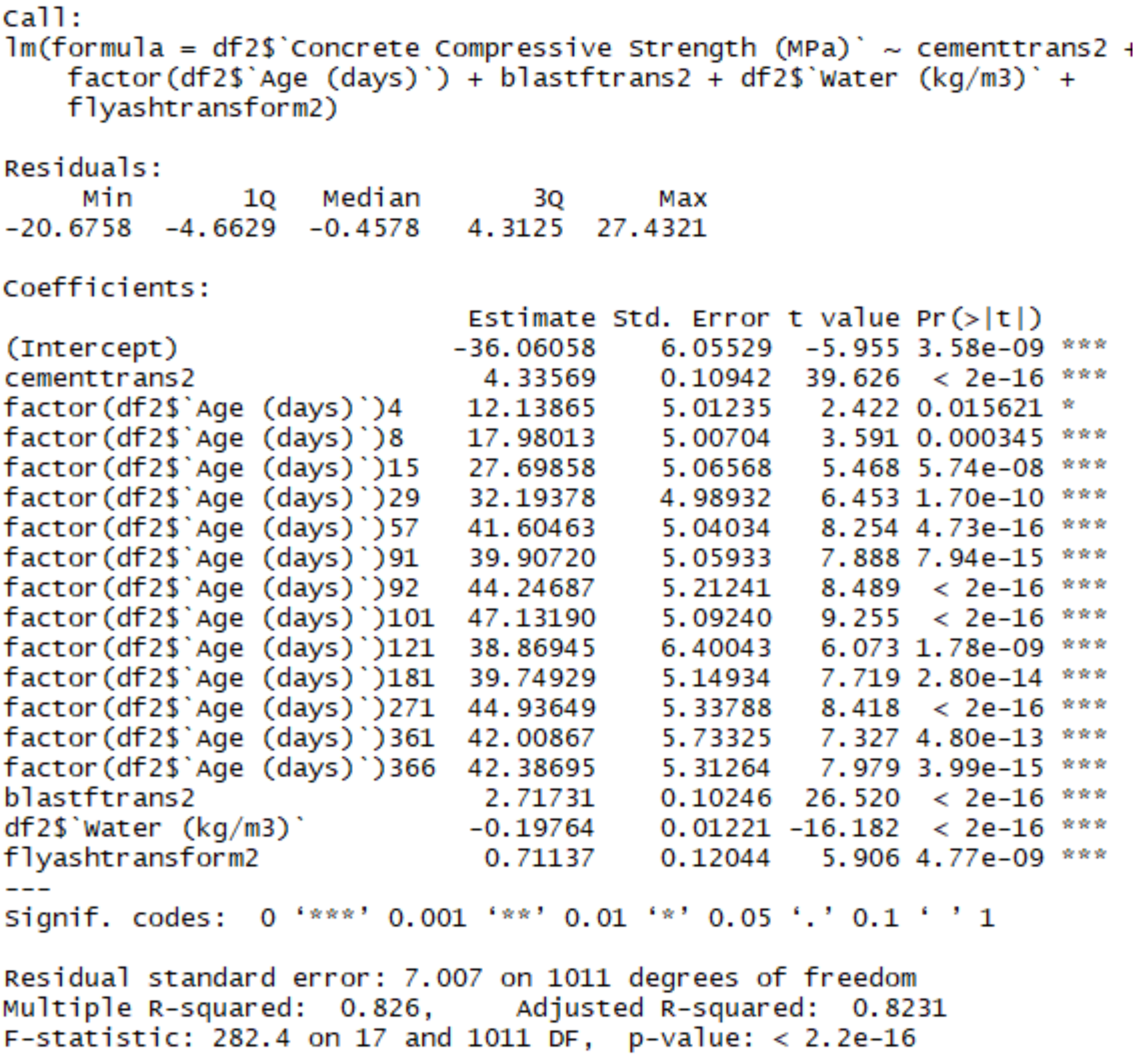


Figure 16: Influence Plot



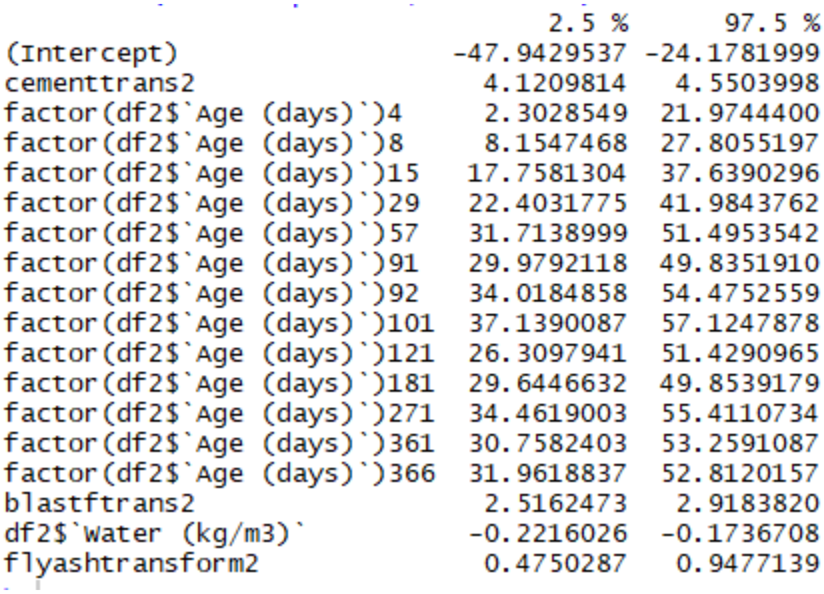
```
Call:
lm(formula = df2$`Concrete Compressive Strength (MPa)` ~ cementtrans2 +
    factor(df2$`Age (days)`) + blastftrans2 + df2$`Water (kg/m3)` +
    flyashtransform2)

Residuals:
     Min       1Q   Median       3Q      Max
-20.6758  -4.6629  -0.4578   4.3125  27.4321

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -36.06058    6.05529  -5.955 3.58e-09 ***
cementtrans2                  4.33569    0.10942  39.626  < 2e-16 ***
factor(df2$`Age (days)`)4    12.13865    5.01235   2.422 0.015621 *
factor(df2$`Age (days)`)8    17.98013    5.00704   3.591 0.000345 ***
factor(df2$`Age (days)`)15   27.69858    5.06568   5.468 5.74e-08 ***
factor(df2$`Age (days)`)29   32.19378    4.98932   6.453 1.70e-10 ***
factor(df2$`Age (days)`)57   41.60463    5.04034   8.254 4.73e-16 ***
factor(df2$`Age (days)`)91   39.90720    5.05933   7.888 7.94e-15 ***
factor(df2$`Age (days)`)92   44.24687    5.21241   8.489  < 2e-16 ***
factor(df2$`Age (days)`)101  47.13190    5.09240   9.255  < 2e-16 ***
factor(df2$`Age (days)`)121  38.86945    6.40043   6.073 1.78e-09 ***
factor(df2$`Age (days)`)181  39.74929    5.14934   7.719 2.80e-14 ***
factor(df2$`Age (days)`)271  44.93649    5.33788   8.418  < 2e-16 ***
factor(df2$`Age (days)`)361  42.00867    5.73325   7.327 4.80e-13 ***
factor(df2$`Age (days)`)366  42.38695    5.31264   7.979 3.99e-15 ***
blastftrans2                  2.71731    0.10246  26.520  < 2e-16 ***
df2$`Water (kg/m3)`          -0.19764    0.01221 -16.182  < 2e-16 ***
flyashtransform2              0.71137    0.12044   5.906 4.77e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.007 on 1011 degrees of freedom
Multiple R-squared:  0.826,	Adjusted R-squared:  0.8231
F-statistic: 282.4 on 17 and 1011 DF,  p-value: < 2.2e-16
```

Figure 17: Final Model Summary



```
                                  2.5 %      97.5 %
(Intercept)                 -47.9429537 -24.1781999
cementtrans2                  4.1209814   4.5503998
factor(df2$`Age (days)`)4     2.3028549  21.9744400
factor(df2$`Age (days)`)8     8.1547468  27.8055197
factor(df2$`Age (days)`)15   17.7581304  37.6390296
factor(df2$`Age (days)`)29   22.4031775  41.9843762
factor(df2$`Age (days)`)57   31.7138999  51.4953542
factor(df2$`Age (days)`)91   29.9792118  49.8351910
factor(df2$`Age (days)`)92   34.0184858  54.4752559
factor(df2$`Age (days)`)101  37.1390087  57.1247878
factor(df2$`Age (days)`)121  26.3097941  51.4290965
factor(df2$`Age (days)`)181  29.6446632  49.8539179
factor(df2$`Age (days)`)271  34.4619003  55.4110734
factor(df2$`Age (days)`)361  30.7582403  53.2591087
factor(df2$`Age (days)`)366  31.9618837  52.8120157
blastftrans2                  2.5162473   2.9183820
df2$`Water (kg/m3)`          -0.2216026  -0.1736708
flyashtransform2              0.4750287   0.9477139
```

Figure 18: 95 % Confidence Interval on Model Parameter Estimates