

# **Concrete Mix Statistical Analysis**

Jose Leonardo Wong

## Contents

1.	Concrete Mix Statistical Analysis.....	1
1.1	Introduction .....	1
1.2	Exploratory Data Analysis.....	4
1.3	Data Preprocessing.....	12
1.4	Regression Problems.....	13
1.4.1	Linear Regression Model.....	13
1.4.2	Multiple Linear Regression Model .....	16
1.5	Hypothesis Testing .....	31
1.5.1	T-Test .....	31
1.5.2	Chi-Squared Hypothesis Test.....	35
1.5.3	ANOVA Hypothesis Test.....	36
1.6	Conclusion.....	43
	References .....	44

## 1. Concrete Mix Statistical Analysis

### 1.1 Introduction

This task examines the statistical analysis of a dataset containing measurements of concrete compressive strength and corresponding mix composition variables, including cement, water, fine and coarse aggregates, and supplementary materials such as fly ash and blast furnace slag, addressing a scenario where the purpose is to evaluate how the components influence the compressive strength of concrete.

The analysis involves introducing key statistical concepts and then implementing them to address the scenario requirement, using R Studio, including hypothesis testing, regression analysis, and ANOVA, to assess the relationships between the concrete mix components and compressive strength, allowing for evaluation of whether observed differences are statistically significant and whether predictive models can accurately describe the relationships in the data.

A sample is a group of specimens randomly taken from a population to obtain information about a parameter or characteristic of this population under the assumption that a statistic from the sample data, a value obtained by using an estimator method, will be a good point estimate of the population parameter. (Chao, 1980)

There is a level of uncertainty associated with the point estimate's ability to match the value of the corresponding population parameter, derived from the fact that the sample specimens are collected randomly. This uncertainty means that there is a probability that a point estimate does not correspond with the population parameter. This probability is quantified by a significance level ( $\alpha$ ). For instance, a 0.05 significance level implies a 5% probability of making an error. Consequently, a margin of error, depending on a given required confidence level ( $1-\alpha$ ), provides additional information about the reliability of the point estimate to represent the population parameter. (Chao, 1980)

Therefore, a confidence interval is the range of values between the point estimate minus the margin of error, the lower confidence limit, and the point estimate plus the margin of error, the upper confidence limit. Accordingly, we can say that the value of the population parameter is within the confidence interval with a level of certainty equal to the required confidence level and with a chance of error equal to the significance level  $\alpha$ . (Moore, 1989)

The z-confidence and t-confidence intervals are methods to build confidence intervals when the point estimator is the sample's mean, defining the confidence interval as the mean plus or minus the standard error of the mean, which is the standard deviation of the mean's sampling distribution. These methods allow us to estimate the population mean from the sample mean and assess how good the sample mean is as an estimate of the population mean. The z-confidence interval is used when the population standard deviation is known, while the t-confidence interval is for when the population standard deviation is not known. Either way, the population must be approximately normally distributed if the sample is smaller than 30. (Clarke, 1994)

For samples larger than 30, the Central Limit Theorem allows us to assume that the sample mean is one of the means from a hypothetical normally distributed sampling distribution of means, enabling us to construct a confidence interval, even if the population itself is not normally distributed, by using the sample mean, standard deviation, standard error, and a critical value representing the number of standard errors away from the mean required to achieve the required confidence level. The critical value defines the boundary of the rejection region based on the chosen significance level. (Clarke, 1994)

For z-confidence intervals, the critical values correspond to specific points from the normal distribution based on the required confidence level, while for t-confidence intervals, the critical values correspond to points from the t-distribution for the same confidence level. Because of its shape with longer tails, the t-distribution is better able to handle the additional uncertainty that comes from estimating the population standard deviation when using t--confidence intervals where the population standard deviation is not known beforehand. (Fischetti, 2015)

Confidence levels are not only used to estimate unknown population parameters with a range of likely values but also to test specific claims about the parameter, comparing a sample statistic with the claimed value of the parameter. The process involves formulating a null hypothesis that assumes an observed difference between these two values is only due to the randomness of the sampling process, and an alternative hypothesis stating that an observed difference is unlikely caused by randomness but due to the population parameter differing meaningfully from the claimed value instead. (Fischetti, 2015)

Once the hypotheses have been formulated, the allowed maximum probability of wrongly rejecting the null hypothesis when it is true (Type I error) is determined and represented by a significance level. Then, a test statistic is calculated. This is a value that represents the difference between the sample statistic and the claimed value of the parameter.

Hypothesis tests are classified based on the direction of the alternative hypothesis, with a one-tailed test checking if the parameter is either greater or less than the claimed value and a two-tailed test checking for differences in either direction. (Fischetti, 2015)

When the claimed value of the parameter and the sample statistic refer to the mean of the population and sample, respectively, two of the most common test statistics are the z-test and t-test. Similar to the z-confidence and t-confidence intervals, they are based on the standard normal distribution and t-distribution, respectively and are used when the population's standard deviation is either known or unknown, respectively. Moreover, they require the population to be approximately normally distributed if the sample size is 30 or below. This test statistic is compared against a critical value from the corresponding statistical distribution, depending on the test being used, and if the test statistic falls within the rejection region, determined by the critical value and the significance level, the null hypothesis is rejected. The p-value quantifies the probability of obtaining a test statistic as extreme as the obtained one given that the null hypothesis is true. (Fischetti, 2015)

A single mean test compares a sample mean to that of a known population mean; the sample statistic is compared with the claimed value of the population parameter to determine

if a significant difference exists between them. A two-means test compares the means of two independent samples, which may come from the same broader population but are exposed to different conditions. The observed difference between their sample means is compared to the claimed difference in the corresponding population means (the null hypothesis assumes this to be zero) to determine if a significant difference exists between them. (Fischetti, 2015)

When the claimed value of the parameter and the sample statistic refer to the variance of the population and sample, the F-test is used instead of the z-test or t-test. The F-test compares two variances from two independent samples by calculating the ratio between them, determining whether they differ significantly. When more than two variances are compared, an extension of the F-test, known as the Analysis of Variance (ANOVA), is used. Although the test compares means, it does it by analysing the variance between and within groups. ANOVA applies the F-Test to calculate the differences between group means and the random variance between each of the groups, and then compares these two sources of variance. If the variance between groups significantly exceeds the variance within groups, it means that at least one group's mean differs from the others. (Fischetti, 2015)

When the relationships tested are categorical rather than numerical, the Chi-Square Test is used to evaluate if the observed distribution of categories is meaningfully different from what is expected under the null hypothesis, comparing the observed sample frequencies to the expected frequencies based on a theoretical distribution. In this case, the test statistic is calculated as the sum of the squared differences between observed and expected frequencies, divided by the expected frequencies. If the calculated value exceeds a critical value, corresponding to the chosen significance level, from the Chi-Square distribution, the null hypothesis of no association gets rejected. (Fischetti, 2015)

Regression analysis is an extension of hypothesis testing where the purpose is to determine how well a mathematical equation can describe the relationship between one or more independent variables (predictors) and a dependent variable (response). Linear regression implies a linear relationship between a single independent variable and a dependent variable, while multilinear regression considers multiple independent variables.

The regression equation represents the relationship, where the dependent variable to be predicted is on one side of the equation, and the independent variables (predictors) are on the other side. Each predictor is multiplied by its corresponding regression coefficient (slope), which represents the direction and magnitude of the dependent variable changes as that predictor increases by one unit. An additional constant term (intercept) represents the value of the dependent variable when all predictors are zero. An error term represents the random variability which is not explained by the model. (Fischetti, 2015)

In regression models, the role of hypothesis testing is to test how well the equation describes the relationship or predicts new values of the dependent variable by using the ANOVA analysis to partition the total variance of the dependent variable into two components: the explained variance due to the regression model and unexplained variance or error. A large F-statistic resulting from this process indicates that the regression model explains well the variability in the dependent variable, suggesting that the relationship

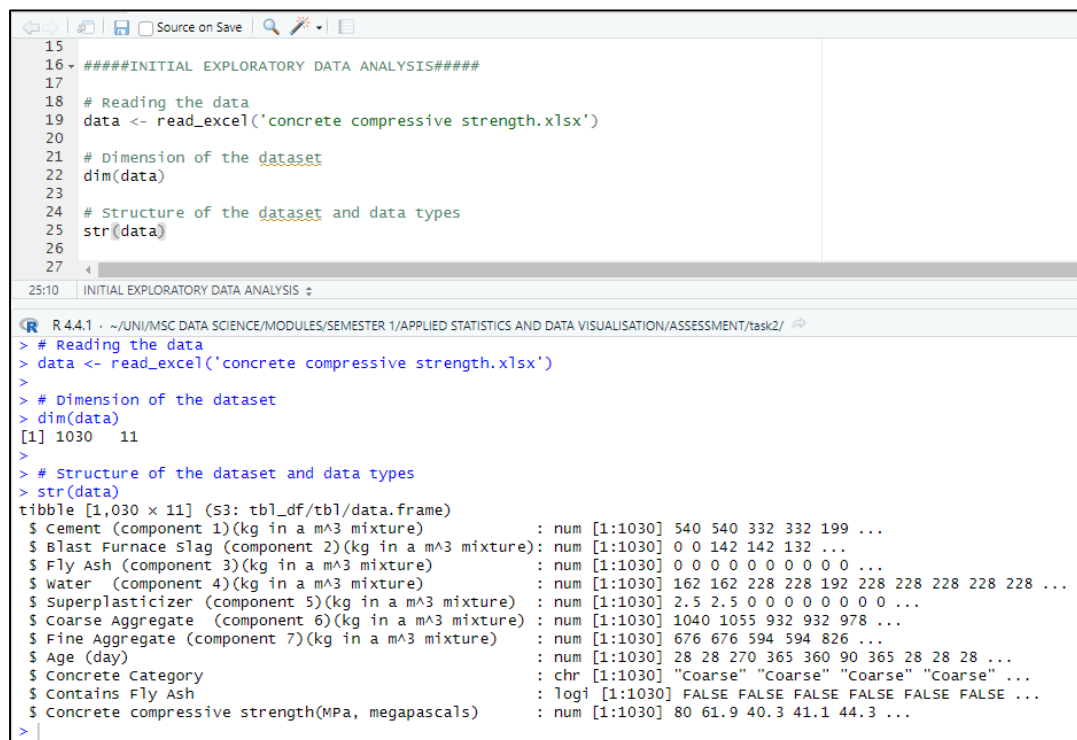
modelled by the equation is unlikely to be caused by random chance. If the p-value associated with the F-test is less than the significance level, the null hypothesis that the model has no explanatory power gets rejected. (Fischetti, 2015)

While the F-value indicates if the explained variance is statistically significant, another indicator, the coefficient of determination ( $R^2$ ) represents the proportion of the variability in the dependent variable that is explained by the model, reflecting the difference between the observed and predicted values of the dependent variable. (Fischetti, 2015)

## 1.2 Exploratory Data Analysis

We began by loading the dataset into R Studio to find that it contains 1030 records, and 11 variables, including key concrete mix components such as cement, water, aggregates, supplementary materials, and the corresponding concrete compressive strength. The inspection of the data structure showed that all the variables have appropriate data types, including numeric variables for mixture components and compressive strength, and categorical variables for Concrete category, coarse and fine and fly ash presence, true and false (Figure 1).

**Figure 1**  
*Data Structure Inspection*



```

15
16 #####INITIAL EXPLORATORY DATA ANALYSIS#####
17
18 # Reading the data
19 data <- read_excel('concrete compressive strength.xlsx')
20
21 # Dimension of the dataset
22 dim(data)
23
24 # Structure of the dataset and data types
25 str(data)
26
27

```

```

R 4.4.1 ~\UNI\MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/
> # Reading the data
> data <- read_excel('concrete compressive strength.xlsx')
>
> # Dimension of the dataset
> dim(data)
[1] 1030 11
>
> # Structure of the dataset and data types
> str(data)
tibble [1,030 × 11] (s3: tbl_df/tbl/data.frame)
 $ cement (component 1)(kg in a m^3 mixture) : num [1:1030] 540 540 332 332 199 ...
 $ Blast Furnace slag (component 2)(kg in a m^3 mixture): num [1:1030] 0 0 142 142 132 ...
 $ Fly Ash (component 3)(kg in a m^3 mixture) : num [1:1030] 0 0 0 0 0 0 0 0 0 ...
 $ water (component 4)(kg in a m^3 mixture) : num [1:1030] 162 162 228 228 192 228 228 228 228 ...
 $ Superplasticizer (component 5)(kg in a m^3 mixture) : num [1:1030] 2.5 2.5 0 0 0 0 0 0 0 ...
 $ Coarse Aggregate (component 6)(kg in a m^3 mixture) : num [1:1030] 1040 1055 932 932 978 ...
 $ Fine Aggregate (component 7)(kg in a m^3 mixture) : num [1:1030] 676 676 594 594 826 ...
 $ Age (day) : num [1:1030] 28 28 270 365 360 90 365 28 28 28 ...
 $ Concrete Category : chr [1:1030] "coarse" "coarse" "coarse" "coarse" ...
 $ Contains Fly Ash : logi [1:1030] FALSE FALSE FALSE FALSE FALSE ...
 $ Concrete compressive strength(MPa, megapascals) : num [1:1030] 80 61.9 40.3 41.1 44.3 ...
>

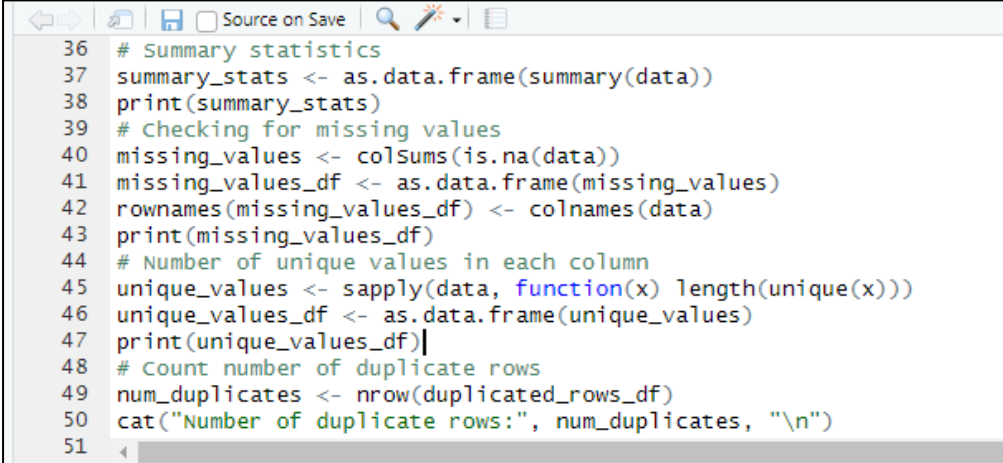
```

We then continued with inspecting the summary statistics, which showed substantial variability in variables such as cement content, ranging from 102 to 540 kg/m<sup>3</sup>, water content, ranging from 121.8 to 247 kg/m<sup>3</sup> and the target variable, concrete compressive strength,

ranging from 2.3 to 82.6 MPa. The check for unique values confirmed this finding. Moreover, no missing values and 25 duplicated rows were found (Figure 2).

**Figure 2**

*Summary Statistics and Missing and Unique Values Checks*



```

36 # Summary statistics
37 summary_stats <- as.data.frame(summary(data))
38 print(summary_stats)
39 # Checking for missing values
40 missing_values <- colSums(is.na(data))
41 missing_values_df <- as.data.frame(missing_values)
42 rownames(missing_values_df) <- colnames(data)
43 print(missing_values_df)
44 # Number of unique values in each column
45 unique_values <- sapply(data, function(x) length(unique(x)))
46 unique_values_df <- as.data.frame(unique_values)
47 print(unique_values_df)
48 # Count number of duplicate rows
49 num_duplicates <- nrow(duplicated_rows_df)
50 cat("Number of duplicate rows:", num_duplicates, "\n")
51
47:24 INITIAL EXPLORATORY DATA ANALYSIS

```

R 4.4.1 ~ /UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSE

```

> rownames(missing_values_df) <- colnames(data)
> print(missing_values_df)

```

	missing_values
Cement (component 1)(kg in a m <sup>3</sup> mixture)	0
Blast Furnace Slag (component 2)(kg in a m <sup>3</sup> mixture)	0
Fly Ash (component 3)(kg in a m <sup>3</sup> mixture)	0
Water (component 4)(kg in a m <sup>3</sup> mixture)	0
Superplasticizer (component 5)(kg in a m <sup>3</sup> mixture)	0
Coarse Aggregate (component 6)(kg in a m <sup>3</sup> mixture)	0
Fine Aggregate (component 7)(kg in a m <sup>3</sup> mixture)	0
Age (day)	0
Concrete Category	0
Contains Fly Ash	0
Concrete compressive strength(MPa, megapascals)	0

```

>
> # Number of unique values in each column
> unique_values <- sapply(data, function(x) length(unique(x)))
> unique_values_df <- as.data.frame(unique_values)
> print(unique_values_df)

```

	unique_values
Cement (component 1)(kg in a m <sup>3</sup> mixture)	280
Blast Furnace Slag (component 2)(kg in a m <sup>3</sup> mixture)	187
Fly Ash (component 3)(kg in a m <sup>3</sup> mixture)	163
Water (component 4)(kg in a m <sup>3</sup> mixture)	205
Superplasticizer (component 5)(kg in a m <sup>3</sup> mixture)	155
Coarse Aggregate (component 6)(kg in a m <sup>3</sup> mixture)	284
Fine Aggregate (component 7)(kg in a m <sup>3</sup> mixture)	304
Age (day)	14
Concrete Category	2
Contains Fly Ash	2
Concrete compressive strength(MPa, megapascals)	938

```

>
> # Count number of duplicate rows
> num_duplicates <- nrow(duplicated_rows_df)
> cat("Number of duplicate rows:", num_duplicates, "\n")
Number of duplicate rows: 25

```

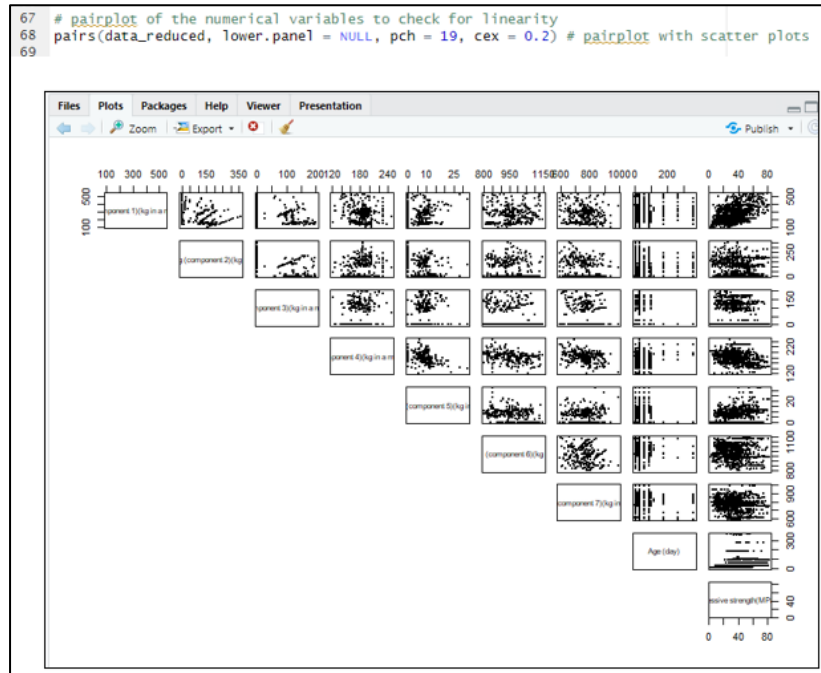




Then we generated a pair plot matrix to visually explore linear relationships between the numerical predictors and the response variable compressive strength, finding that Cement showed a clear positive linear relationship, while other variables showed less recognisable patterns (Figure 4).

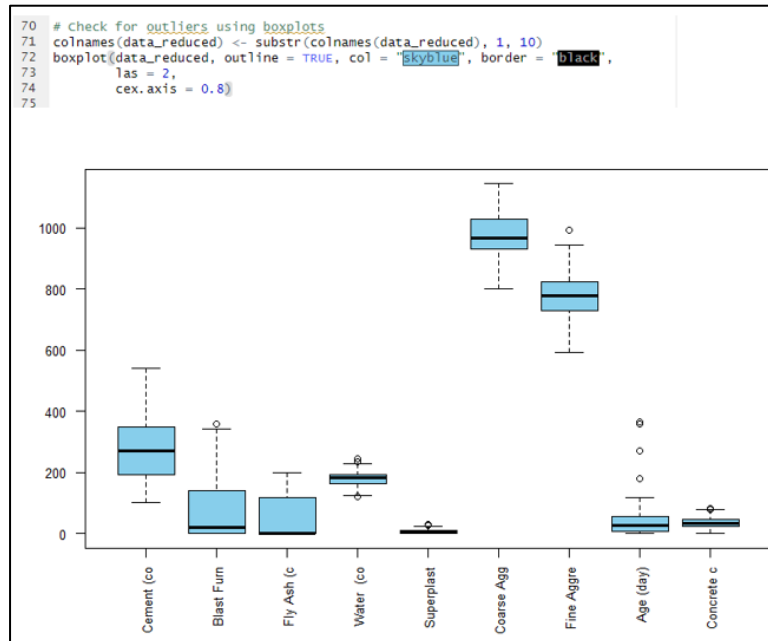
**Figure 4**

*Pair Plot Matrix of Concrete Mix Variables*



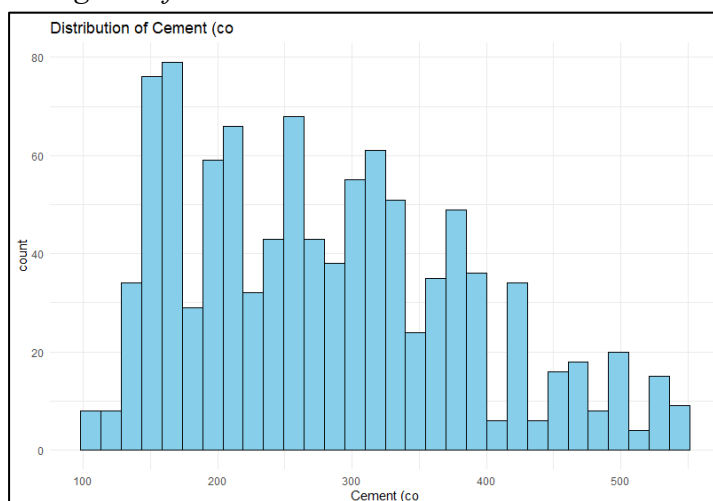
Then we created Boxplots finding the presence of outliers in variables such as Blast Furnace Slag, Water, Superplasticizer, Fine Aggregate and Age (Figure 5)

**Figure 5**  
*Outlier Detection Using Boxplots*

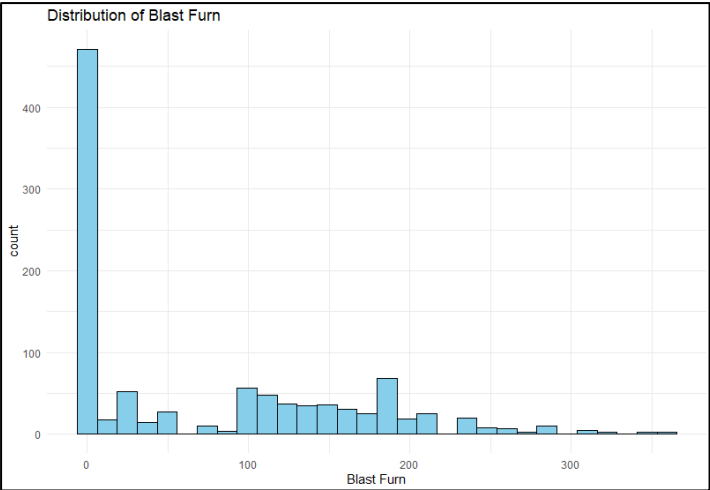


Then we generated histograms for all the numeric variables, shown in figures from 6 to 14. It was found that Blast Furnace, Fly Ash, and Superplasticizer were highly skewed with most values clustered at zero perhaps because these are optional concrete mix components. Similarly, the Age variable showed most of its values close to zero probably indicating that the obtention of samples was made shortly after the production date. The response variable, compressive strength, followed a near-normal distribution

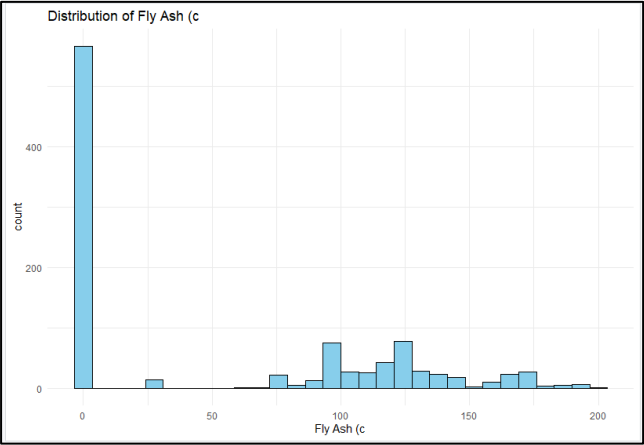
**Figure 6**  
*Histogram of Cement Content Distribution*

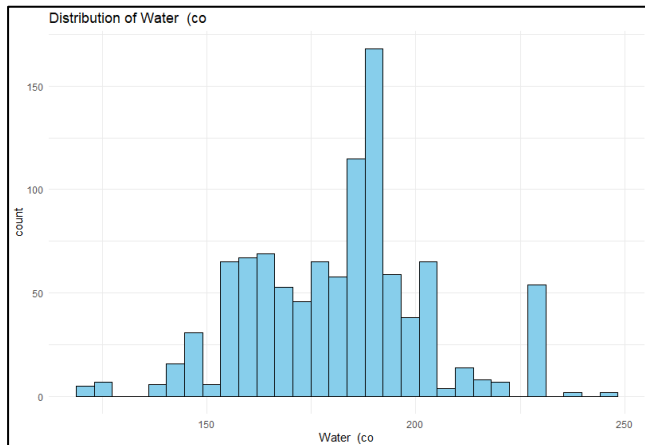
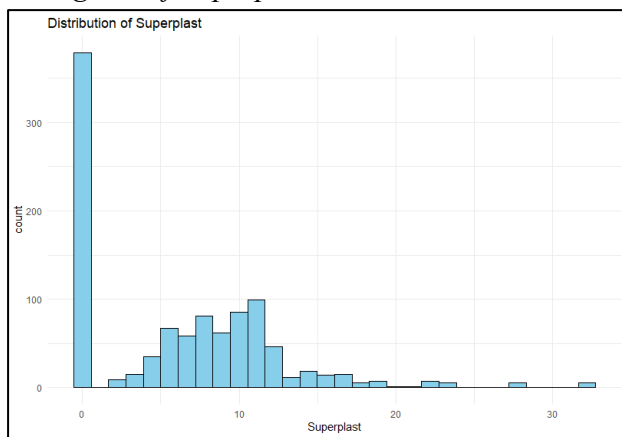
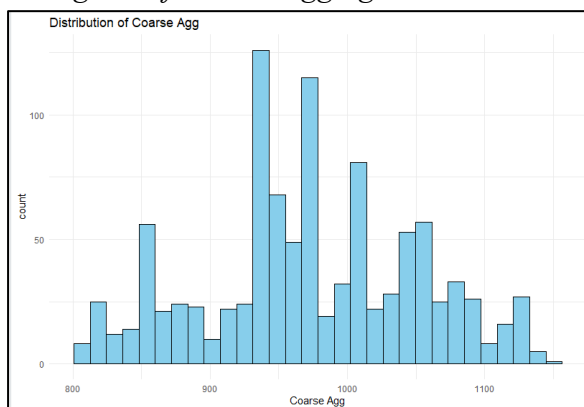


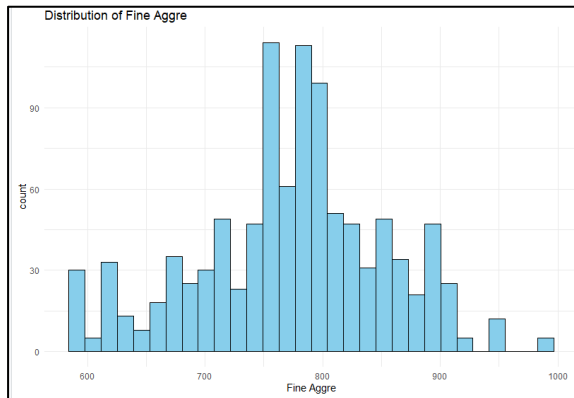
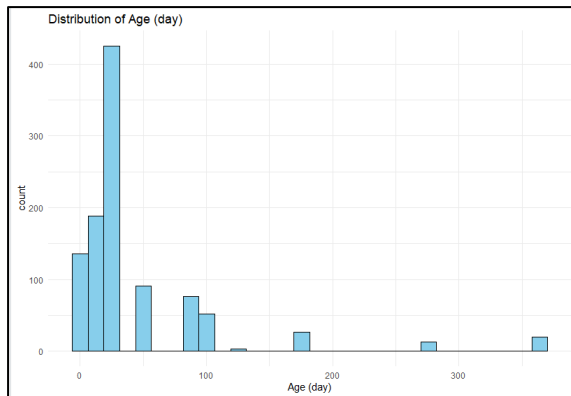
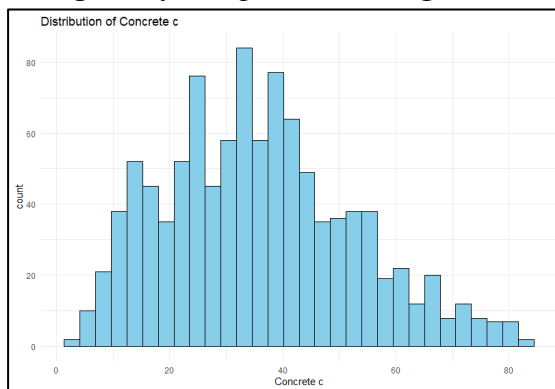
**Figure 7**  
*Histogram of Blast Furnace Slag Distribution*



**Figure 8**  
*Histogram of Fly Ash Distribution*



**Figure 9***Histogram of Water Content Distribution***Figure 10***Histogram of Superplasticizer Distribution***Figure 11***Histogram of Coarse Aggregate Distribution*

**Figure 12***Histogram of Fine Aggregate Distribution***Figure 13***Histogram of Age (Days) Distribution***Figure 14***Histogram of Compressive Strength Distribution*

### 1.3 Data Preprocessing

Moving into the data preprocessing stage, we removed duplicates from the dataset since they would not provide additional information for the models (Figure 15). Following this, we created dummy binary variables for predictors with dominant zeros corresponding to optional components in a concrete mix, such as Blast Furnace Slag, Fly Ash, and Superplasticizer, to denote their presence or absence with the intention of enhancing the models' predictive capabilities in the modelling stage (Figure 15).

**Figure 15**

*Duplicate Removal and Creation of Dummy Variables*

```
# Remove duplicates
data <- unique(data)

# Create dummy variables for predictors with dominant zeros (These are optional components in a concrete mixture)
# The original variable and its dummy variable should not be used in the same model as they have high multicollinearity.
data$Blast_Furnace_Slag_Present <- ifelse(data$`Blast Furnace Slag (component 2)(kg in a m^3 mixture)` > 0, 1, 0)
data$Fly_Ash_Present <- ifelse(data$`Fly Ash (component 3)(kg in a m^3 mixture)` > 0, 1, 0)
data$superplasticizer_Present <- ifelse(data$`superplasticizer (component 5)(kg in a m^3 mixture)` > 0, 1, 0)
```

The remaining values different from zero, in addition to the variable Age (day), were transformed using a logarithmic transformation to help with outliers and make their distributions closer to a normal distribution (Figure 16). Histograms were then plotted to verify the effects of the transformations. Additionally, a new correlation matrix was generated, given the new dummy variables added to the dataset.

**Figure 16**

*Log Transformation of Variables*

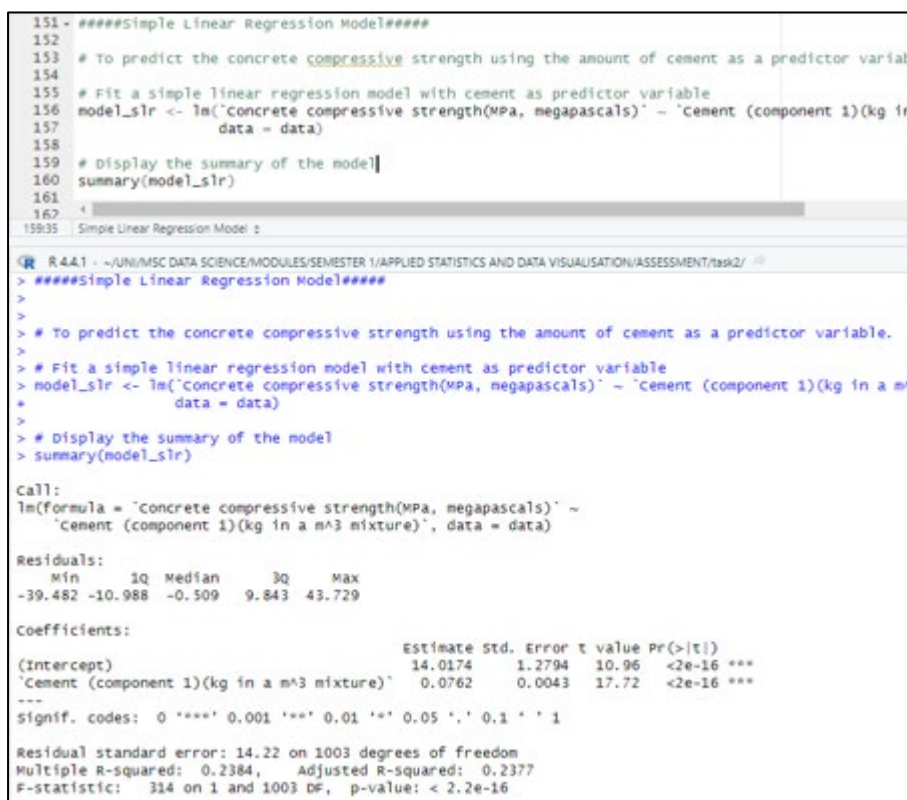
```
97 # Log-transform variables with dominant zeros
98 data$`Blast Furnace Slag (component 2)(kg in a m^3 mixture)` <- ifelse(
99   data$`Blast Furnace Slag (component 2)(kg in a m^3 mixture)` > 0,
100   log1p(data$`Blast Furnace Slag (component 2)(kg in a m^3 mixture)`),
101   0
102 )
103 data$`Fly Ash (component 3)(kg in a m^3 mixture)` <- ifelse(
104   data$`Fly Ash (component 3)(kg in a m^3 mixture)` > 0,
105   log1p(data$`Fly Ash (component 3)(kg in a m^3 mixture)`),
106   0
107 )
108 data$`Superplasticizer (component 5)(kg in a m^3 mixture)` <- ifelse(
109   data$`superplasticizer (component 5)(kg in a m^3 mixture)` > 0,
110   log1p(data$`superplasticizer (component 5)(kg in a m^3 mixture)`),
111   0
112 )
113 |
114 # Log-transform the remaining skewed variable (Age)
115 data$`Age (day)` <- log1p(data$`Age (day)` )
116
117 # Check the distributions of predictors after transformation
118 transformed_vars <- c(
119   "Blast Furnace Slag (component 2)(kg in a m^3 mixture)",
120   "Fly Ash (component 3)(kg in a m^3 mixture)",
121   "Superplasticizer (component 5)(kg in a m^3 mixture)",
122   "Age (day)"
123 )
124 ~ lapply(transformed_vars, function(var) {
125   ggplot(data, aes_string(x = paste0("`", var, "`"))) +
126     geom_histogram(bins = 30, fill = "skyblue", color = "black") +
127     ggtitle(paste("Distribution of", var, "After Transformation")) +
128     theme_minimal()
129 ~ })
130
```

## 1.4 Regression Problems

### 1.4.1 Linear Regression Model

This exercise involved a simple linear regression model to predict compressive strength using the amount of cement as the predictor, which was chosen as it has the highest correlation with the response variable. The result showed that the coefficient in the equation for cement was significant ( $p < 0.001$ ) and estimated at 0.0762, which means that increasing a kilogram of cement per cubic meter increases the concrete compressive strength by 0.0762 MPa. The  $R^2$  value of 0.2384 showed that cement explains 23.8% of the variability in compressive strength, which highlights the importance of cement in predicting strength but suggests the need for additional predictors for more explanatory power (Figure 17).

**Figure 17**  
*Linear Regression Model Summary*

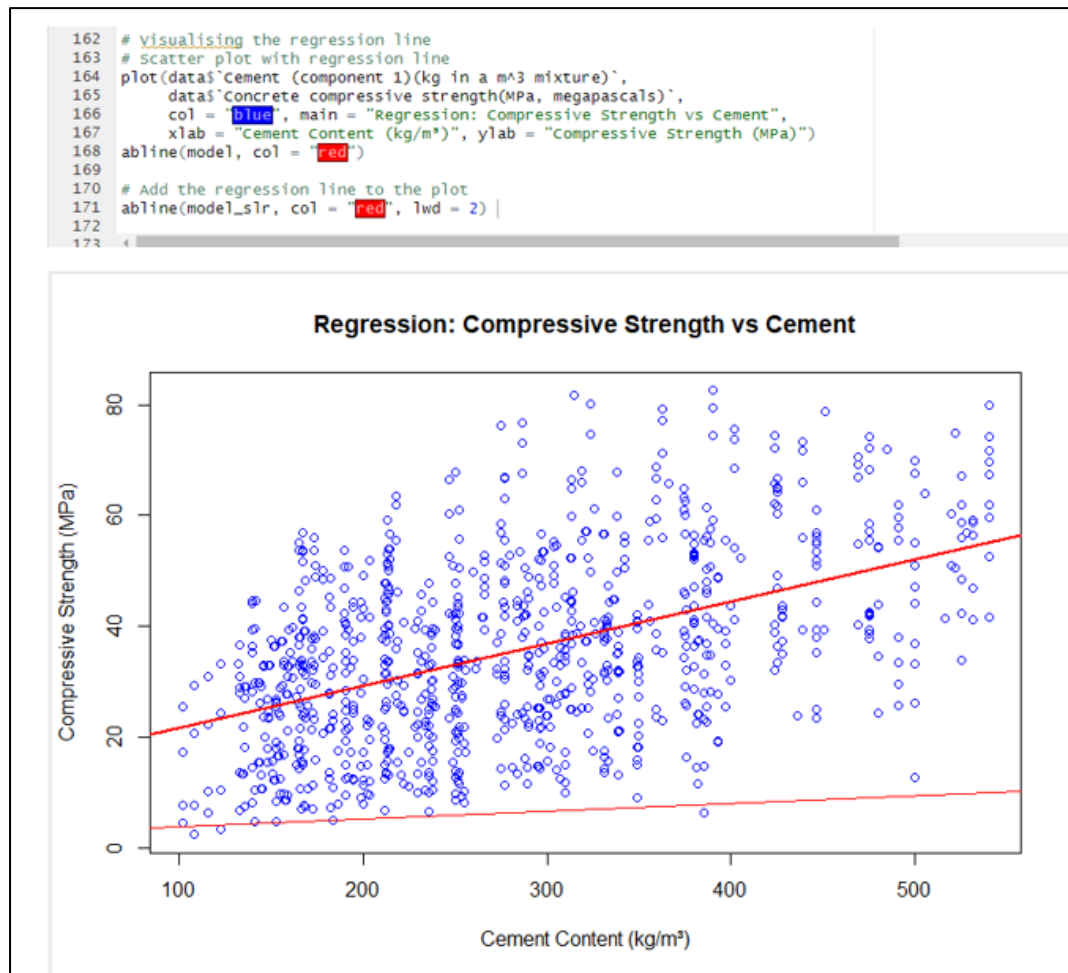


To visualise the relationship, we generated a scatter plot of cement content versus compressive strength, including the regression line, which demonstrated a positive linear relation where an increase in cement content corresponds to higher compressive strength.

However, the dispersion of points around the line suggests variability in the data not captured by the model, confirming the need for additional predictors (Figure 18).

**Figure 18**

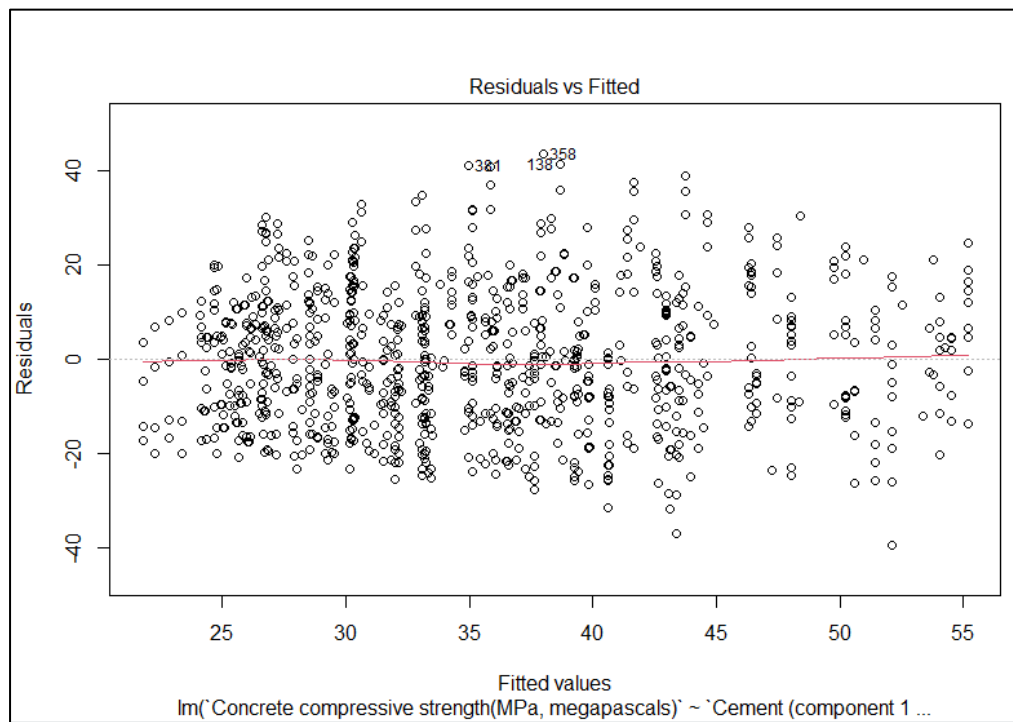
*Scatter Plot of Cement Content vs Compressive Strength with Regression Line*



Then we proceeded to check that the model complies with the assumptions of linear regression. The first assumption is the linear relationship between the predictor and the response variable for which we checked using the residuals vs. fitted values plot which indicated a linear relationship by showing that the residuals (the differences between the observed and the predicted values of the dependent variable) were scattered randomly around the zero-line which represents where residuals would perfectly align if the model were a perfect fit. (Figure 19)

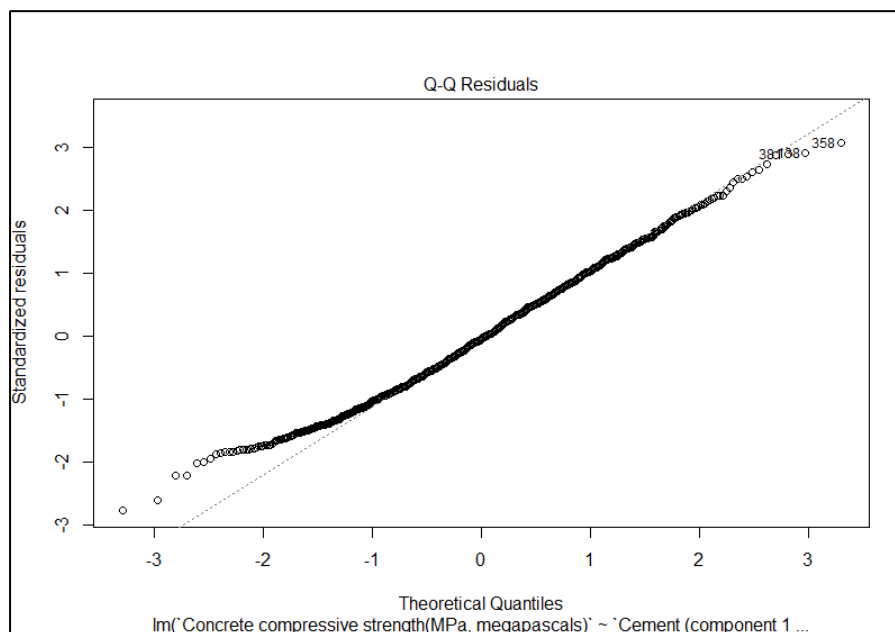


**Figure 19**  
*Residuals vs. Fitted Values Plot*



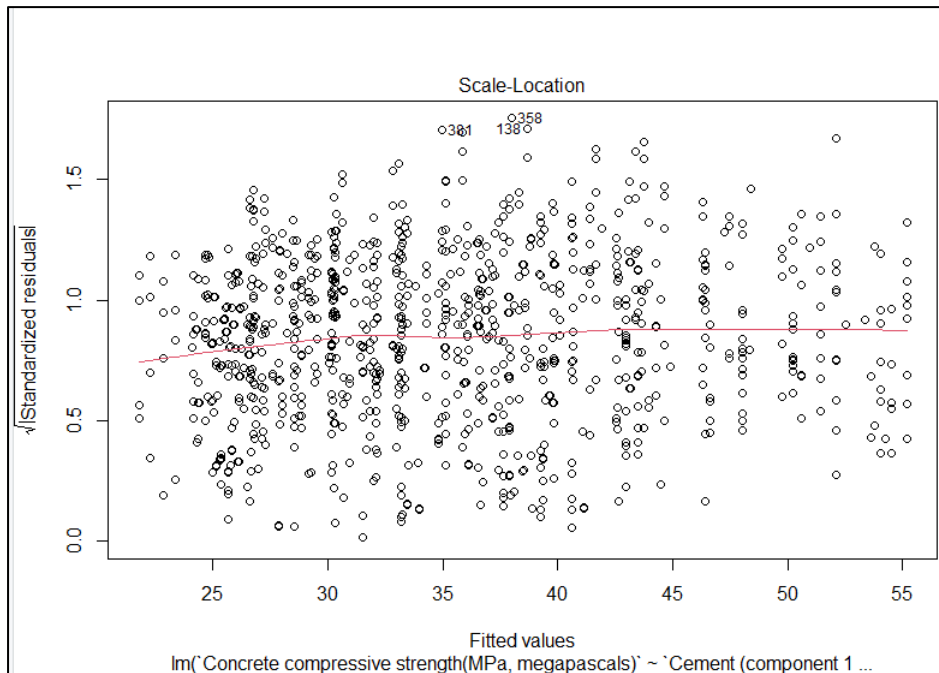
Similarly, we used a Q-Q plot, which compares the distribution of residuals to a normal distribution, to evaluate the second assumption, normality of residuals, with most residuals closely following the diagonal line, indicating that the residuals are approximately normally distributed (Figure 20).

**Figure 20**  
*Q-Q Plot for Model Residuals*



Lastly, we assessed for homoscedasticity, the assumption that the variance of residuals is constant across all the predicted values of the dependent variable. For this, we used the scale-location plot showing that the spread of residuals across fitted values was consistent, validating the assumption of homoscedasticity (Figure 21).

**Figure 21**  
*Scale-Location Plot for Model Residuals*



### 1.4.2 Multiple Linear Regression Model

We implemented a multiple linear regression model to predict concrete compressive strength using multiple predictors using a forward stepwise approach, implying the addition of predictors sequentially to the model in an order determined by the absolute value of their correlation with the response variable. We began with Cement and Age as the initial predictor variables, both resulting in statistically significant ( $p$ -values  $< 0.001$ ). The coefficient for Cement indicated a 0.076 MPa increase in compressive strength per additional kilogram per cubic meter, consistent with our previous model results, while Age showed an 8.18 MPa increase per day, reflecting the effect of concrete mix curing. We found the model explained 54.8% of the variance in compressive strength (Adjusted  $R^2 = 0.5475$ ) (Figure 22)

**Figure 22**  
*Initial Multiple Linear Regression Model Summary*

```

191 #####Multiple Linear Regression Model#####
192
193 # To predict the concrete compressive strength using multiple predictor variables.
194
195 # Fit initial model with Cement and Age as predictor variables
196 model_mlr <- lm(`Concrete compressive strength(MPa, megapascals)` ~
197               `Cement (component 1)(kg in a m^3 mixture)` +
198               `Age (day)`, data = data)
199 # Display summary for Model 1
200 summary(model_mlr)
201
202

```

213:45 Multiple Linear Regression Model

R 4.4.1 · ~/UNI/MSc DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/

```

> # To predict the concrete compressive strength using multiple predictor variables.
>
> # Fit initial model with Cement and Age as predictor variables
> model_mlr <- lm(`Concrete compressive strength(MPa, megapascals)` ~
+               `Cement (component 1)(kg in a m^3 mixture)` +
+               `Age (day)`, data = data)
> # Display summary for Model 1
> summary(model_mlr)

```

Call:

```
lm(formula = `Concrete compressive strength(MPa, megapascals)` ~
    `Cement (component 1)(kg in a m^3 mixture)` + `Age (day)`,
    data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-29.898	-7.969	-0.769	6.615	42.750

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-12.356595	1.408285	-8.774	<2e-16 ***
`Cement (component 1)(kg in a m^3 mixture)`	0.075646	0.003313	22.830	<2e-16 ***
`Age (day)`	8.175637	0.311774	26.223	<2e-16 ***

---  
 signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.95 on 1002 degrees of freedom  
 Multiple R-squared: 0.5484, Adjusted R-squared: 0.5475  
 F-statistic: 608.3 on 2 and 1002 DF, p-value: < 2.2e-16

The next step involved adding Superplasticizer as the third predictor to the multiple linear regression model, which improved the model's predictive power, reflected by an increase in the  $R^2$  value from 0.5484 to 0.6874 (Adjusted  $R^2 = 0.6864$ ). The coefficient for Superplasticiser (5.28 MPa per unit) was statistically significant (p-value < 0.001) (Figure 23).

To validate the improvement, an ANOVA test compared Model 1 with Model 2, resulting in a significant F-statistic of 445.04 (p-value < 0.001), confirming that the addition of Superplasticizer enhanced the model significantly. (Figure 23).

**Figure 23**  
*Superplasticizer Impact on Compressive Strength*

```

202 # Add Superplasticizer as the third predictor
203 model_mlr2 <- lm('Concrete compressive strength(MPa, megapascals)' ~
204   'Cement (component 1)(kg in a m^3 mixture)' +
205   'Age (day)' +
206   'Superplasticizer (component 5)(kg in a m^3 mixture)', data = data)
207 # Display summary for model 2
208 summary(model_mlr2)
209 # Perform ANOVA tests between models 1 and 2
210 anova_model_1_2 <- anova(model_mlr, model_mlr2)
211 cat("ANOVA between Model 1 and Model 2:\n")
212 print(anova_model_1_2)
213 +

```

Multiple Linear Regression Model 2

```

R 4.4.1 - ~/UNIMSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/
Call:
lm(formula = "concrete compressive strength(MPa, megapascals)" ~
    "Cement (component 1)(kg in a m^3 mixture)" + "Age (day)" +
    "Superplasticizer (component 5)(kg in a m^3 mixture)",
    data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-23.804  -5.701  -1.037   5.572  39.923

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -21.707986    1.253299   -17.32  <2e-16 ***
'Cement (component 1)(kg in a m^3 mixture)'    0.079258    0.002763    28.68  <2e-16 ***
'Age (day)'      8.423519    0.259793    32.42  <2e-16 ***
'Superplasticizer (component 5)(kg in a m^3 mixture)'  5.283576    0.250455    21.10  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.119 on 1001 degrees of freedom
Multiple R-squared:  0.6874,    Adjusted R-squared:  0.6864
F-statistic: 733.6 on 3 and 1001 DF,  p-value: < 2.2e-16

>
> # Perform ANOVA tests between models 1 and 2
> anova_model_1_2 <- anova(model_mlr, model_mlr2)
> cat("ANOVA between Model 1 and Model 2:\n")
ANOVA between Model 1 and Model 2:
> print(anova_model_1_2)
Analysis of Variance Table

Model 1: 'Concrete compressive strength(MPa, megapascals)' ~ 'Cement (component 1)(kg in a m^3 mixture)'
          'Age (day)'
Model 2: 'Concrete compressive strength(MPa, megapascals)' ~ 'Cement (component 1)(kg in a m^3 mixture)'
          'Age (day)' + 'Superplasticizer (component 5)(kg in a m^3 mixture)'
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1    1002 120250
2    1001  83242   1    37009 445.04 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The next step involved adding Water as the fourth predictor in the multiple linear regression model, showing further improvement of the  $R^2$ , increasing from 0.6874 to 0.7087 (Adjusted  $R^2 = 0.7075$ ) (Figure 24). Additionally, the coefficient for Water (-0.14 MPa per  $\text{kg/m}^3$ ) was found to be statistically significant ( $p\text{-value} < 0.001$ ) with a negative association with compressive strength, coherently with the intuition that adding excessive water weakens the concrete matrix. The ANOVA test between Models 2 and 3 produced a highly significant F-statistic of 73.22 ( $p\text{-value} < 0.001$ ) (Figure 24).

**Figure 24**  
*Water Content Effect in Multiple Regression Mode*

```

217 # Add water as the fourth predictor
218 model_mlr3 <- lm('concrete compressive strength(MPa, megapascals)' ~
219   'Cement (component 1)(kg in a m^3 mixture)' +
220   'Age (day)' +
221   'Superplasticizer (component 5)(kg in a m^3 mixture)' +
222   'water (component 4)(kg in a m^3 mixture)', data = data)
223 # Display summary for model 3:
224 summary(model_mlr3)
225 # Perform ANOVA tests between models 2 and 3
226 anova_model_2_3 <- anova(model_mlr2, model_mlr3)
227 +

```

223:30 Multiple Linear Regression Model :

```

R 4.4.1 - ~/UN/MS DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/
lm(formula = "concrete compressive strength(MPa, megapascals)" ~
  "Cement (component 1)(kg in a m^3 mixture)" + "Age (day)" +
  "Superplasticizer (component 5)(kg in a m^3 mixture)" +
  "water (component 4)(kg in a m^3 mixture)", data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-23.535  -5.629  -1.028    5.541   35.472

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.135501    3.471723   1.767  0.0775 .
'Cement (component 1)(kg in a m^3 mixture)'  0.076474    0.002689   28.443 <2e-16 ***
'Age (day)'      8.823335    0.255213   34.572 <2e-16 ***
'Superplasticizer (component 5)(kg in a m^3 mixture)'  3.683370    0.305741   12.047 <2e-16 ***
'water (component 4)(kg in a m^3 mixture)'  -0.143246    0.016740  -8.557 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.807 on 1000 degrees of freedom
Multiple R-squared:  0.7087,    Adjusted R-squared:  0.7075
F-statistic: 608.2 on 4 and 1000 DF,  p-value: < 2.2e-16

> # Perform ANOVA tests between models 2 and 3
> anova_model_2_3 <- anova(model_mlr2, model_mlr3)
> cat("\nANOVA between Model 2 and Model 3:\n")

ANOVA between Model 2 and Model 3:
> print(anova_model_2_3)
Analysis of Variance Table

Model 1: 'concrete compressive strength(MPa, megapascals)' ~ 'Cement (component 1)(kg in a m^3 mixture)'
  'Age (day)' + 'Superplasticizer (component 5)(kg in a m^3 mixture)'
Model 2: 'concrete compressive strength(MPa, megapascals)' ~ 'Cement (component 1)(kg in a m^3 mixture)'
  'Age (day)' + 'Superplasticizer (component 5)(kg in a m^3 mixture)' +
  'water (component 4)(kg in a m^3 mixture)'
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1    1001 83242
2    1000 77562  1    5679.3 73.222 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Next, adding the Blast Furnace Slag presence predictor improved the model's  $R^2$  from 0.7087 to 0.7833 (adjusted  $R^2 = 0.7823$ ) (Figure 25). The coefficient for Blast Furnace Slag presence was positive (9.21 MPa,  $p$ -value  $< 0.001$ ), indicating that adding it to the concrete mix enhances compressive strength. Moreover, the ANOVA comparison showed a highly significant F-statistic of 344.23 ( $p$ -value  $< 0.001$ ), formally confirming the finding (Figure 25)

**Figure 25**  
*Blast Furnace Slag Presence Impact Analysis*

```

237 model_mlr4 <- lm('concrete compressive strength(MPa, megapascals)' ~
238   'Cement (component 1)(kg in a m^3 mixture)' +
239   'Age (day)' +
240   'Superplasticizer (component 5)(kg in a m^3 mixture)' +
241   'water (component 4)(kg in a m^3 mixture)' +
242   'Blast_Furnace_slag_Present', data = data)
243 # Display summary for model 4
244 summary(model_mlr4)
245 # Perform ANOVA tests between models 3 and 4
246 anova_model_3_4 <- anova(model_mlr3, model_mlr4)
247 cat("\nANOVA between Model 3 and Model 4:\n")
248 print(anova_model_3_4)
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2800
2801
2802
2803
2804
2805
2806

```



**Figure 26**  
*Final Selected Regression Model Overview*

```
> # Add Fine Aggregate as the sixth predictor
> model_mlr5 <- lm('Concrete compressive strength(MPa, megapascals)' ~
+ 'Cement (component 1)(kg in a m^3 mixture)' +
+ 'Age (day)' +
+ 'Superplasticizer (component 5)(kg in a m^3 mixture)' +
+ 'water (component 4)(kg in a m^3 mixture)' +
+ 'Blast_Furnace_Slag_Present' +
+ 'Fine Aggregate (component 7)(kg in a m^3 mixture)', data = data)
> # Display summary for model 5
> summary(model_mlr5)
```

Call:  
lm(formula = 'Concrete compressive strength(MPa, megapascals)' ~  
'Cement (component 1)(kg in a m^3 mixture)' + 'Age (day)' +  
'Superplasticizer (component 5)(kg in a m^3 mixture)' +  
'water (component 4)(kg in a m^3 mixture)' + Blast\_Furnace\_Slag\_Present +  
'Fine Aggregate (component 7)(kg in a m^3 mixture)',  
data = data)

Residuals:

Min	1Q	Median	3Q	Max
-21.3644	-4.6688	0.1028	4.3987	29.3168

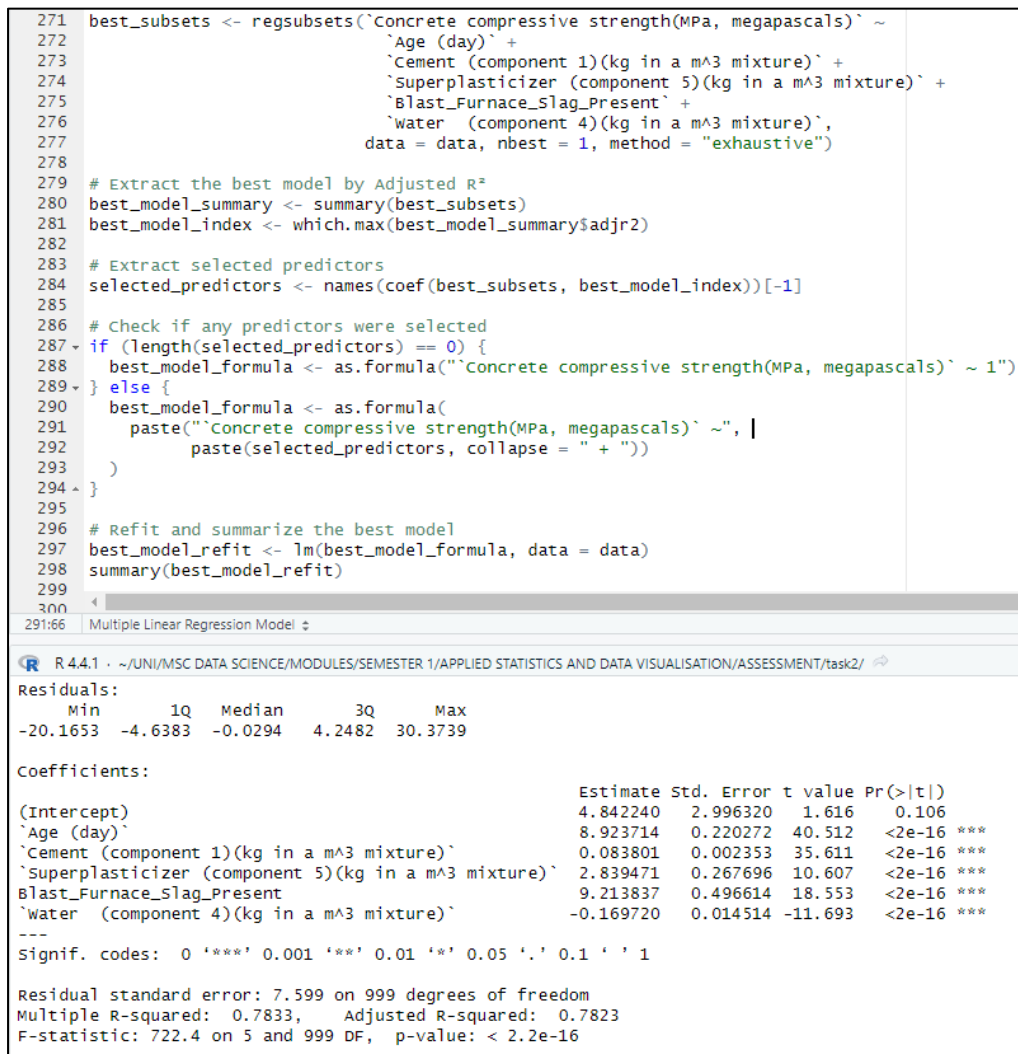
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	27.493017	5.471193	5.025	5.96e-07	***
'Cement (component 1)(kg in a m^3 mixture)'	0.079157	0.002510	31.534	< 2e-16	***
'Age (day)'	8.881825	0.217920	40.757	< 2e-16	***
'Superplasticizer (component 5)(kg in a m^3 mixture)'	2.692820	0.266305	10.112	< 2e-16	***
'water (component 4)(kg in a m^3 mixture)'	-0.205799	0.016111	-12.774	< 2e-16	***
Blast_Furnace_slag_Present	8.429031	0.516159	16.330	< 2e-16	***
'Fine Aggregate (component 7)(kg in a m^3 mixture)'	-0.018146	0.003685	-4.924	9.92e-07	***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.512 on 998 degrees of freedom  
Multiple R-squared: 0.7885, Adjusted R-squared: 0.7872  
F-statistic: 620.1 on 6 and 998 DF, p-value: < 2.2e-16

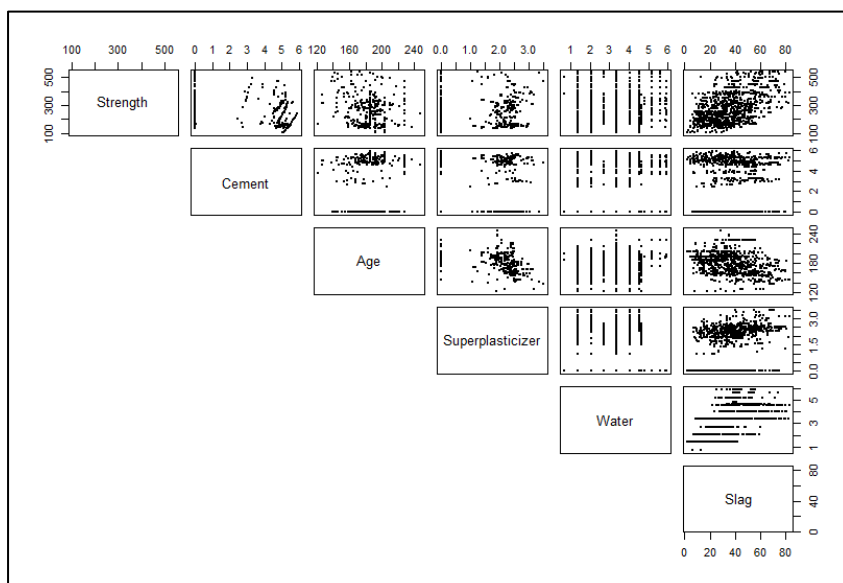
To explore potential improvements in model performance while reducing complexity, we conducted an exhaustive and programmatic search of all possible subsets of predictors in the model using the adjusted  $R^2$  as a criterion and an implementation with the “regsubsets” function from the leaps package. However, the search did not identify any subset that outperformed the previously selected model. (Figure 27).

**Figure 27***Model Selection Process Using Subset Search*

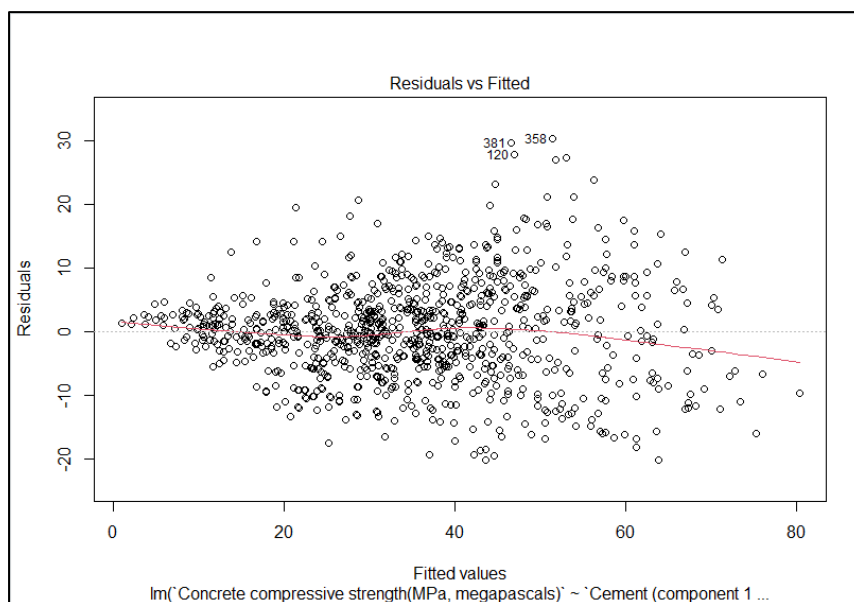
We validated the assumptions of the multiple linear regression model, starting with the linearity assumption, using a scatterplot matrix (Figure 28), which revealed deviations from linearity in some predictor relationships. Additionally, residual independence was evaluated with the residuals vs. fitted values plot (Figure 29), showing residuals randomly scattered around the zero-line, indicating linearity. Moreover, the normality of residuals was checked with the Q-Q plot (Figure 30), where most residuals aligned closely with the diagonal line, meaning that they are approximately normally distributed. The Homoscedasticity assumption of constant variance of residuals was tested using the scale-location plot (Figure 31), which showed a consistent spread of residuals across fitted values, validating the assumption compliance. Lastly, multicollinearity among predictors was examined using the Variance Inflation Factor (VIF), which measures how much the variance of a regression coefficient is increased due to correlations among predictors, with all values below 2, indicating no significant multicollinearity issues.



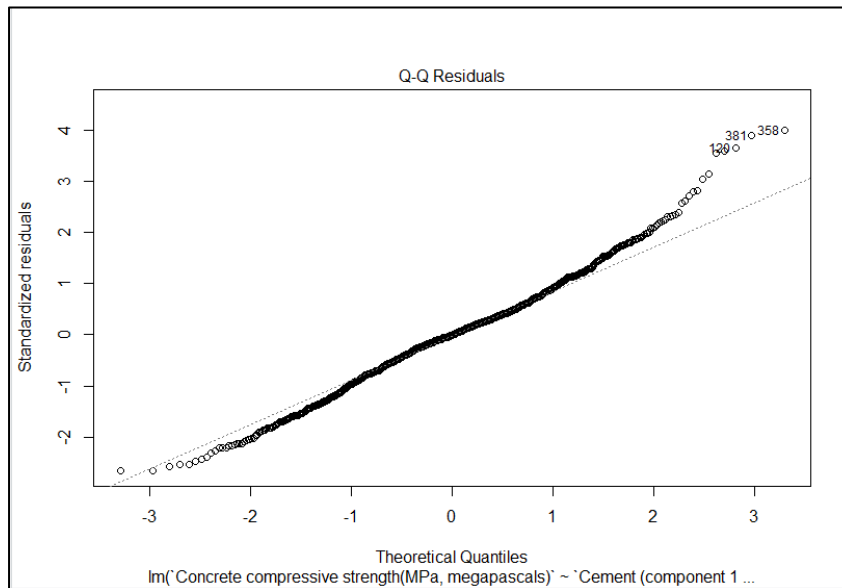
**Figure 28**  
*Scatterplot Matrix for Linear Relationship Validation*



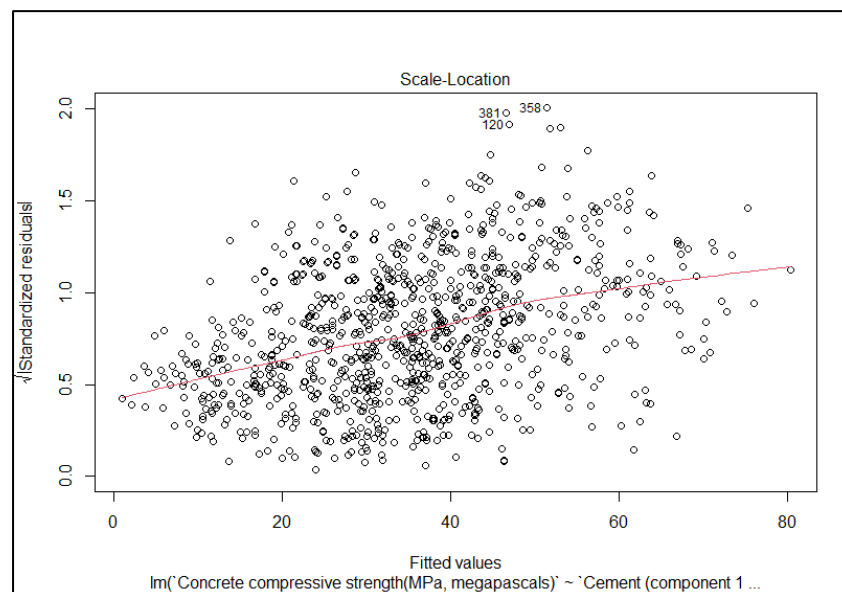
**Figure 29**  
*Residuals vs. Fitted Values for Multiple Regression Model*



**Figure 30**  
*Q-Q Plot for Multiple Regression Model Residuals*



**Figure 31**  
*Scale-Location Plot for Multiple Regression Model*



To address the violation of the linearity assumption, we used interaction terms between predictors, which represent the interplay of two or more predictors on the response variable. These interactions are represented in the regression model equation by multiplying the predictors involved. To perform this, we rescaled the numeric predictors to standardise their ranges and fit a model including all possible interaction terms. We used the `step()` function to perform a selection of the most significant interaction terms (Figure 32). The

interaction terms significantly improved the model's performance, increasing the adjusted  $R^2$  value to 0.833. (Figure 32).

**Figure 32**

*Selection of The Most Significant Interaction Terms*

```
# We found violation of linearity assumption in the model. We can address this
# by adding interaction terms between the predictors.
# Generates interaction terms programmatically using the stepAIC selection
# method to include only significant interactions in the model.

# Standardize numeric predictors before using the interaction terms.
dataCement_scaled <- scale(data$Cement (component 1)(kg in a m^3 mixture)', center = TRUE, scale = TRUE)
dataAge_scaled <- scale(data$Age (day)', center = TRUE, scale = TRUE)
dataSuperplasticizer_scaled <- scale(data$Superplasticizer (component 5)(kg in a m^3 mixture)',
                                     center = TRUE, scale = TRUE)
dataWater_scaled <- scale(data$Water (component 4)(kg in a m^3 mixture)', center = TRUE, scale = TRUE)

# Fit a model with all interactions using scaled predictors
interaction_model <- lm('Concrete compressive strength(MPa, megapascals)' ~
  (Cement_scaled *
    Age_scaled *
    Superplasticizer_scaled *
    Water_scaled *
    Blast_Furnace_Slag_Present),
  data = data)

# Perform stepwise selection to include only significant interactions
interaction_stepwise <- stepAIC(interaction_model, direction = "both", trace = TRUE)

# view the summary of the selected model
summary(interaction_stepwise)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-20.6545	-3.9494	-0.0391	3.6946	30.8817

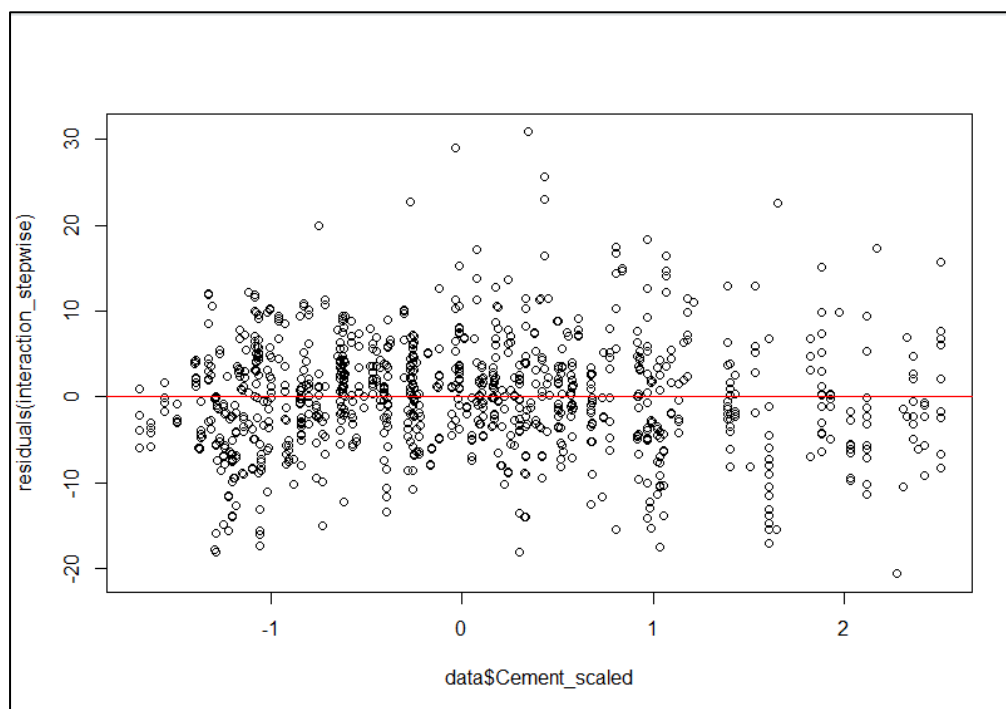
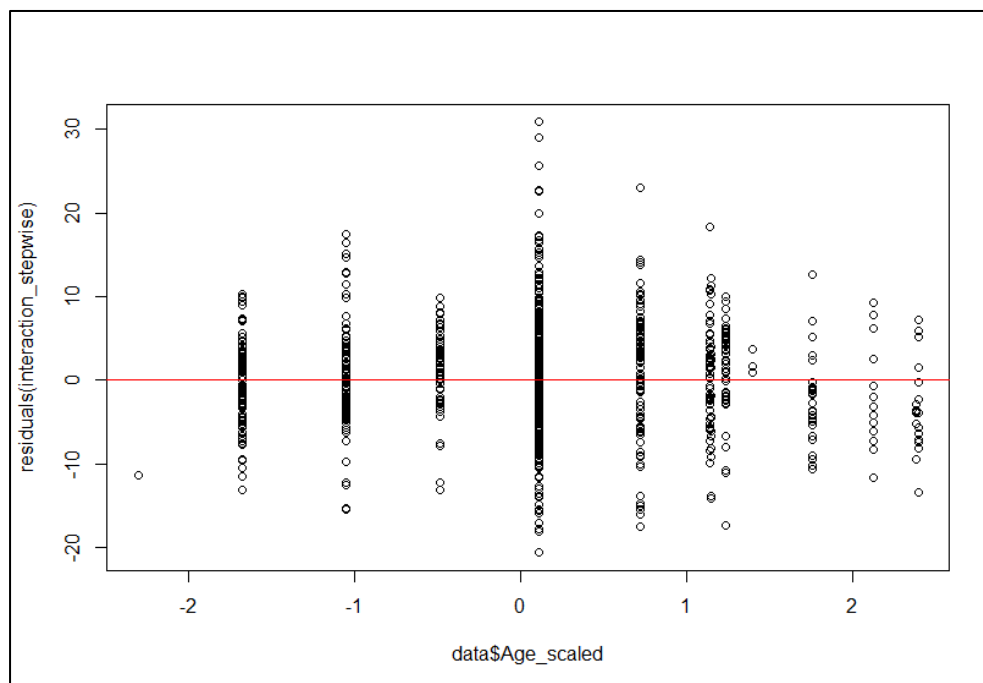
Coefficients:

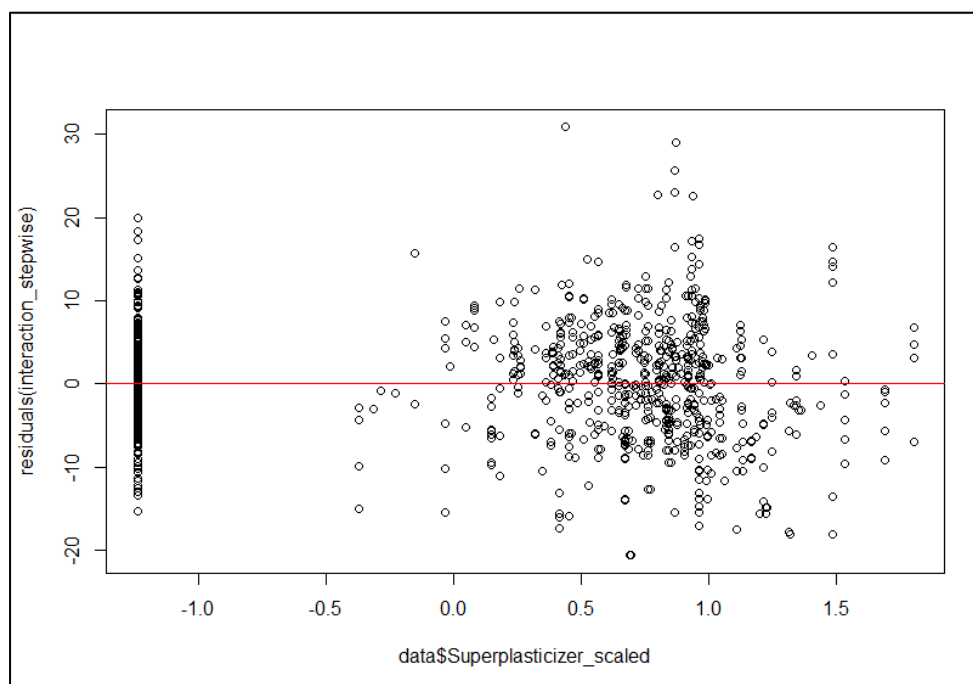
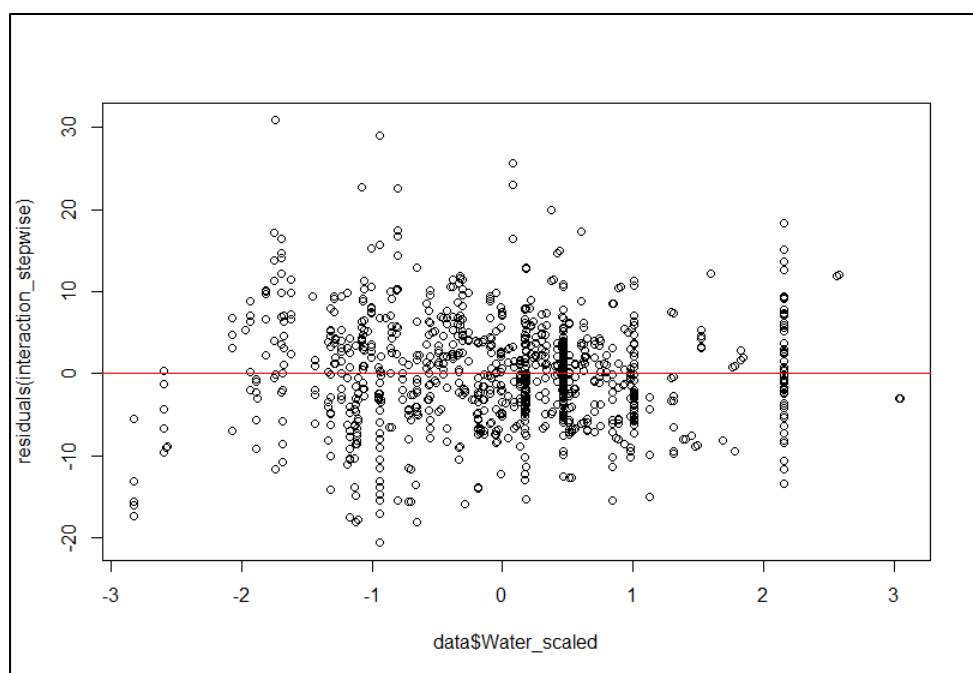
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	29.70046	0.43920	67.624	< 2e-16 ***
Cement_scaled	12.28781	0.43152	28.476	< 2e-16 ***
Age_scaled	8.44991	0.52524	16.088	< 2e-16 ***
Superplasticizer_scaled	5.26856	0.46685	11.285	< 2e-16 ***
Water_scaled	-4.19669	0.65468	-6.410	2.25e-10 ***
Blast_Furnace_Slag_Present	10.42763	0.69533	14.997	< 2e-16 ***
Cement_scaled:Age_scaled	1.34196	0.41662	3.221	0.001319 **
Cement_scaled:Superplasticizer_scaled	-0.25572	0.37343	-0.685	0.493644 .
Age_scaled:Superplasticizer_scaled	2.48625	0.57300	4.339	1.58e-05 ***
Cement_scaled:Water_scaled	0.60656	0.51790	1.171	0.241809 .
Age_scaled:Water_scaled	1.07486	0.62506	1.720	0.085817 .
Superplasticizer_scaled:Water_scaled	-0.61879	0.67726	-0.914	0.361120 .
Cement_scaled:Blast_Furnace_Slag_Present	-2.35526	0.61157	-3.851	0.000125 ***
Age_scaled:Blast_Furnace_Slag_Present	4.45522	0.81532	5.464	5.89e-08 ***
Superplasticizer_scaled:Blast_Furnace_Slag_Present	-1.37999	0.74956	-1.841	0.065913 .
Water_scaled:Blast_Furnace_Slag_Present	2.16568	0.81744	2.649	0.008194 **
Cement_scaled:Age_scaled:Superplasticizer_scaled	-0.52184	0.39755	-1.313	0.189616 .
Cement_scaled:Age_scaled:Water_scaled	-0.52173	0.35880	-1.454	0.146238 .
Cement_scaled:Superplasticizer_scaled:Water_scaled	2.16922	0.28982	7.485	1.59e-13 ***
Age_scaled:Superplasticizer_scaled:Water_scaled	-1.65344	0.53782	-3.074	0.002168 **
Age_scaled:Superplasticizer_scaled:Blast_Furnace_Slag_Present	-2.25541	0.91542	-2.464	0.013918 *.
Cement_scaled:Water_scaled:Blast_Furnace_Slag_Present	-1.39070	0.50178	-2.772	0.005685 **
Age_scaled:Water_scaled:Blast_Furnace_Slag_Present	-2.46590	0.83463	-2.954	0.003207 **
Superplasticizer_scaled:Water_scaled:Blast_Furnace_Slag_Present	-0.05277	0.76377	-0.069	0.944935 .
Cement_scaled:Age_scaled:Superplasticizer_scaled:Water_scaled	1.30391	0.29115	4.478	8.40e-06 ***
Age_scaled:Superplasticizer_scaled:Water_scaled:Blast_Furnace_Slag_Present	2.40716	0.62822	3.832	0.000135 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

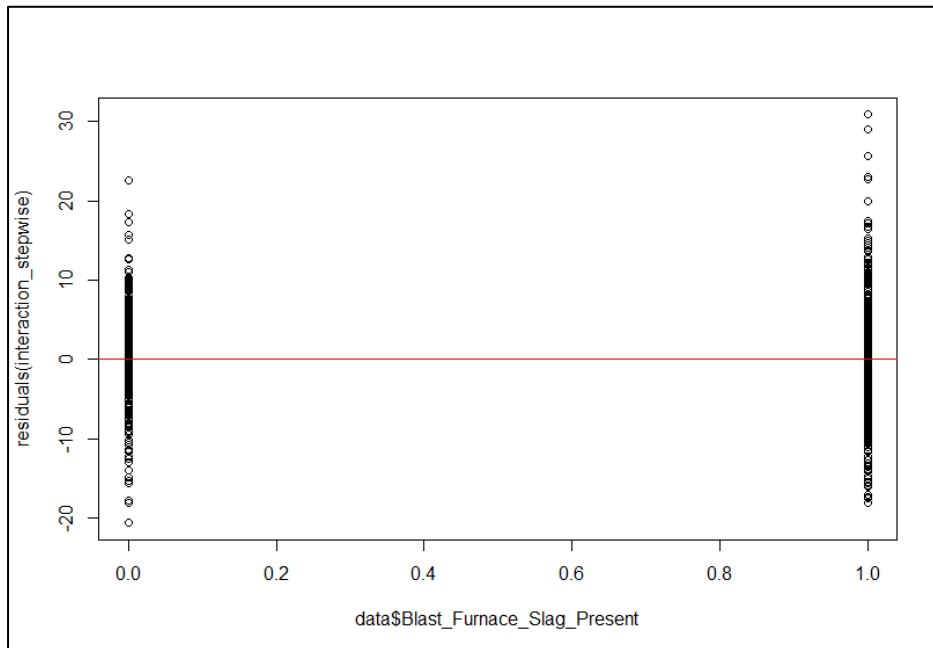
Residual standard error: 6.647 on 979 degrees of freedom  
Multiple R-squared: 0.8375, Adjusted R-squared: 0.8334  
F-statistic: 201.9 on 25 and 979 DF, p-value: < 2.2e-16

To validate the model's compliance with assumptions after adding interaction terms, the following checks were performed. Scatter plots of residuals against each predictor were generated to check for linearity, as shown in Figures 33 to 37. The absence of curved patterns or clustering indicates that the relationship between predictors and the response variable is linear in the model.

**Figure 33***Residuals vs. Cement Content Scatter Plot***Figure 34***Residuals vs. Water Content Scatter Plot*

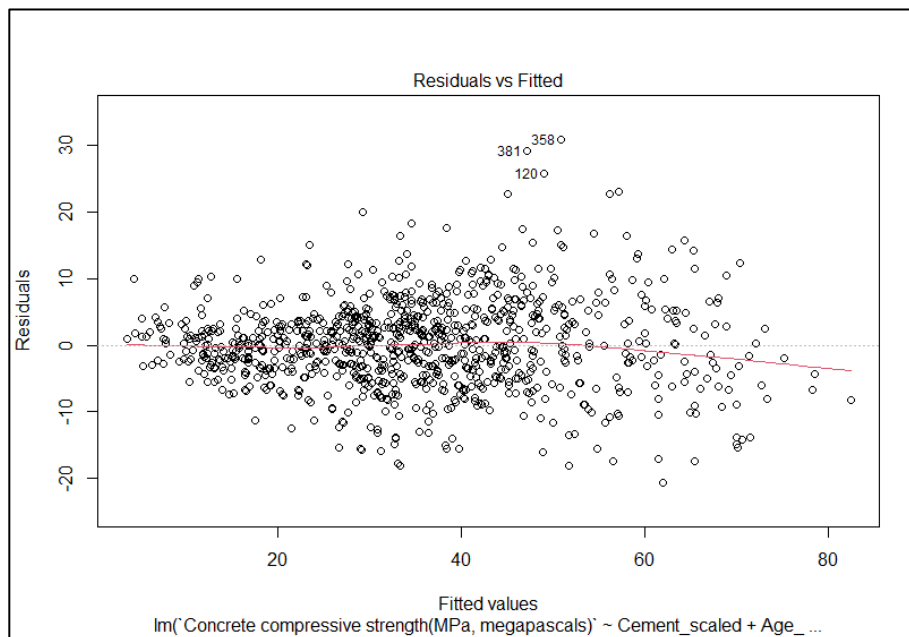
**Figure 35***Residuals vs. Superplasticizer Content Scatter Plot***Figure 36***Residuals vs. Age (Days) Scatter Plot*

**Figure 37**  
*Residuals vs. Blast Furnace Slag Presence Scatter Plot*



Similarly, the residuals vs. fitted values plot was generated to check for residual independence, showing a random scatter of points, which confirms that the residuals are independent and not influenced by the predicted values (Figure 38).

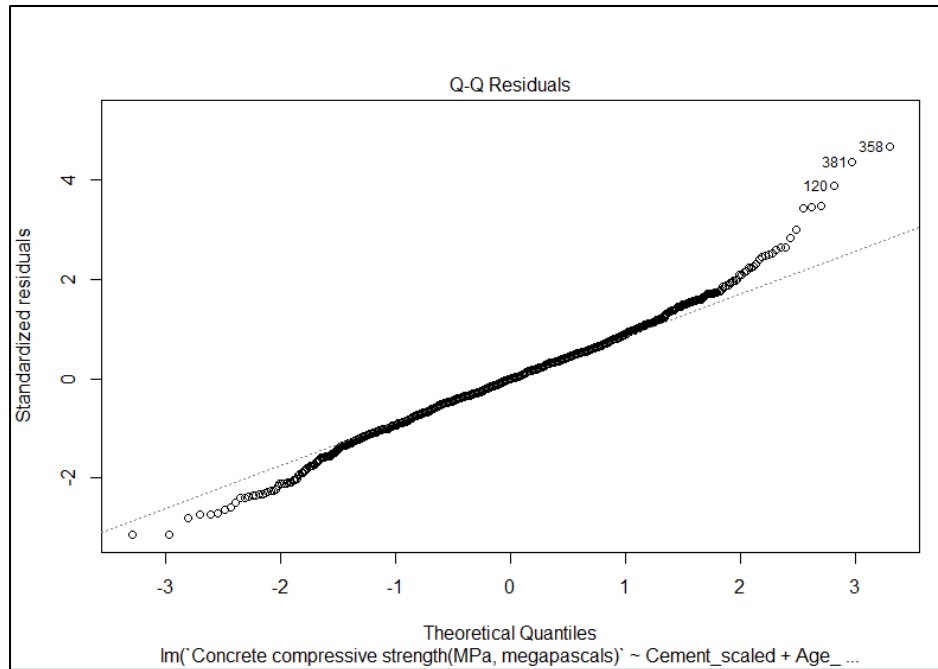
**Figure 38**  
*Residuals vs. Fitted Values for Improved Model*



A Q-Q plot of the residuals was generated to ensure they follow a normal distribution. The showed points aligned closely with the diagonal reference line, except for minor

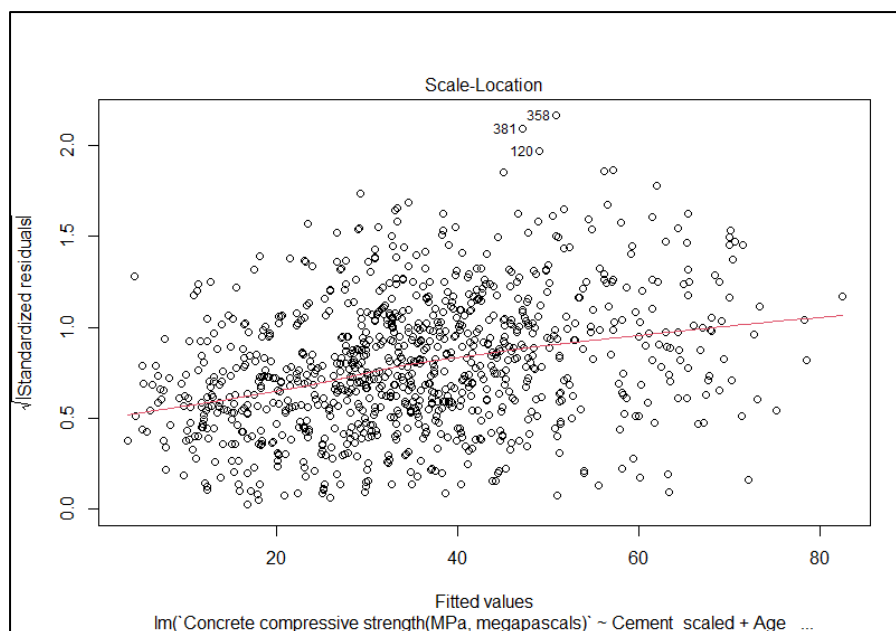
deviations at the extremes, consistent with approximately normally distributed residuals (Figure 39)

**Figure 39**  
*Q-Q Plot for Improved Model Residuals*



The Scale-Location plot was inspected to validate for heteroscedasticity, showing that the spread of residuals was consistent across all fitted values, with no noticeable increase or decrease in variance (Figure 40)

**Figure 40**  
*Scale-Location Plot for Improved Model Residuals*



Lastly, the generalised Variance Inflation Factor (GVIF) values were calculated for each predictor and interaction term. All GVIF values were approximately 1, indicating that the predictors and their interactions are not highly correlated (Figure 41).

**Figure 41**  
*Generalised Variance Inflation Factor (GVIF) Results*

```

399 # No multicollinearity (Predictor variables should not be highly correlated with each other.)
400 # Check for multicollinearity using the VIF (Variance Inflation Factor)
401 vif(interaction_stepwise, type = "predictor")
402
1:1 (Top Level)

```

---

R 4.4.1 · ~/UNI/MSc DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task3/

```

GVIFs computed for predictors

```

	GVIF	DF	GVIF^(1/(2*DF))
Cement_scaled	1.25	1	1
Age_scaled	1.25	1	1
Superplasticizer_scaled	1.25	1	1
Water_scaled	1.25	1	1
Slag_Furnace_slag_present	1.25	1	1

```

Interacts with other predictors
Cement_scaled      Age_scaled, Superplasticizer_scaled, Water_scaled, Slag_Furnace_slag_present --
Age_scaled         Cement_scaled, Superplasticizer_scaled, Water_scaled, Slag_Furnace_slag_present --
Superplasticizer_scaled  Cement_scaled, Age_scaled, Water_scaled, Slag_Furnace_slag_present --
Water_scaled       Cement_scaled, Age_scaled, Superplasticizer_scaled, Slag_Furnace_slag_present --
Slag_Furnace_slag_present  Cement_scaled, Age_scaled, Superplasticizer_scaled, Water_scaled --

```



## 1.5 Hypothesis Testing

### 1.5.1 T-Test

The purpose of this test was to determine whether the mean concrete compressive strength of coarse aggregate is greater than that of fine aggregate. We formulated a null hypothesis stating that the mean compressive strength of coarse aggregate is less than or equal to that of fine aggregate and an alternative hypothesis claiming that the mean compressive strength of coarse aggregate is greater. Given that the independent variable was a two-level categorical variable and the dependent variable was a numerical continuous variable, we selected an independent two-sample t-test.

To carry out the test, we split the dataset into two based on the concrete categories, Coarse and Fine. To ensure compliance with the requirements of using a t-test, several checks were conducted. In this respect, we verified the number of observations in each group to confirm sufficient sample sizes (Figure 42). Similarly, Histograms were used to assess the data distributions (Figure 43) (Figure 44), Q-Q plots to check for normality (Figure 45) (Figure 46), and boxplots to compare the spread of the data between the groups (Figure 47). Moreover, a Levene's test, which checks whether the variances of two or more groups are equal, was performed to test the equality of variances between the two groups, resulting in a p-value of 0.4797, consistent with no significant difference in variances (Figure 48).

**Figure 42**

*Split the Dataset into Two Groups*

```
# Dependent Variable: Concrete Compressive Strength
# Independent variable: Concrete category (Categorical: Coarse, Fine)

# Separate the data into two groups based on the "Concrete Category" variable.

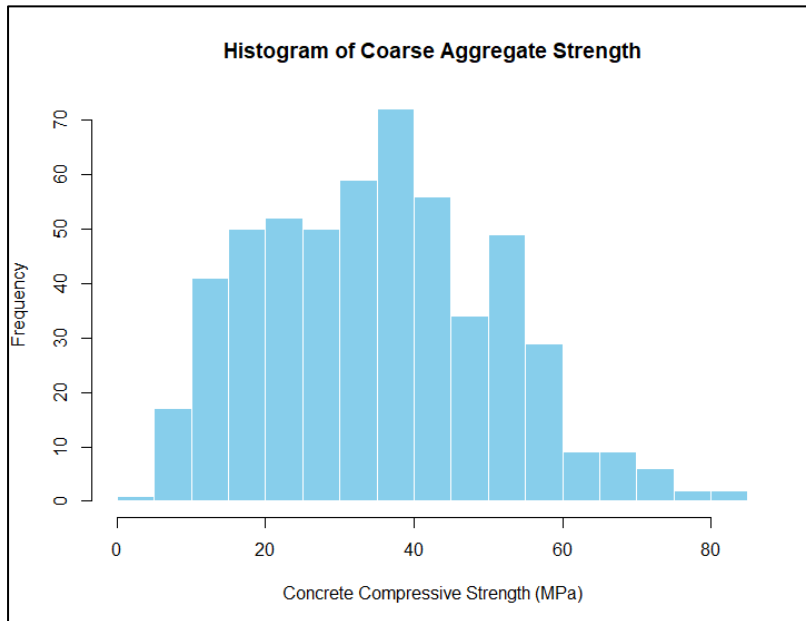
# Coarse Aggregate subset
coarse_data <- subset(data, `Concrete Category` == "Coarse")

# Fine Aggregate subset
fine_data <- subset(data, `Concrete Category` == "Fine")

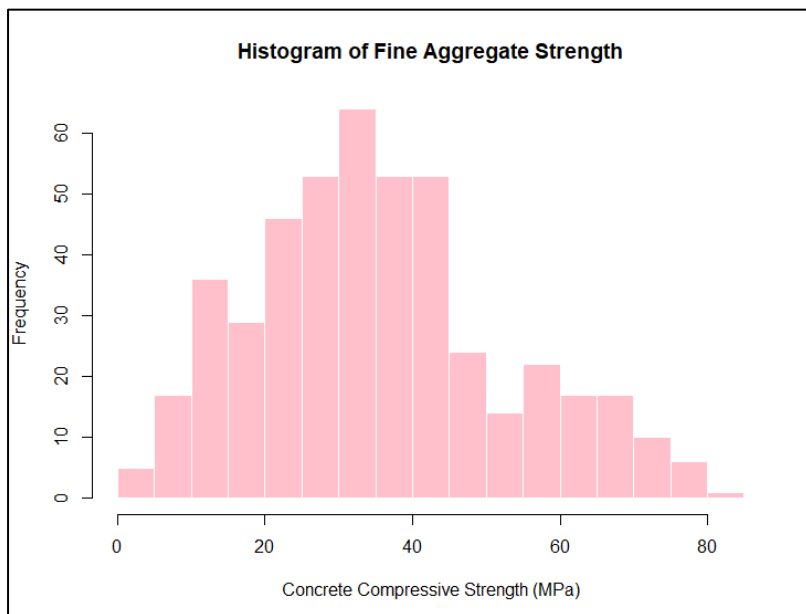
# Extract the compressive strength column from each group
coarse_strength <- coarse_data$`Concrete compressive strength(MPa, megapascals)`
fine_strength <- fine_data$`Concrete compressive strength(MPa, megapascals)`

# Check the number of observations in each group
cat("Number of observations in Coarse group:", length(coarse_strength), "\n")
cat("Number of observations in Fine group:", length(fine_strength), "\n")
```

**Figure 43**  
*Histogram of Compressive Strength (Coarse Aggregate)*

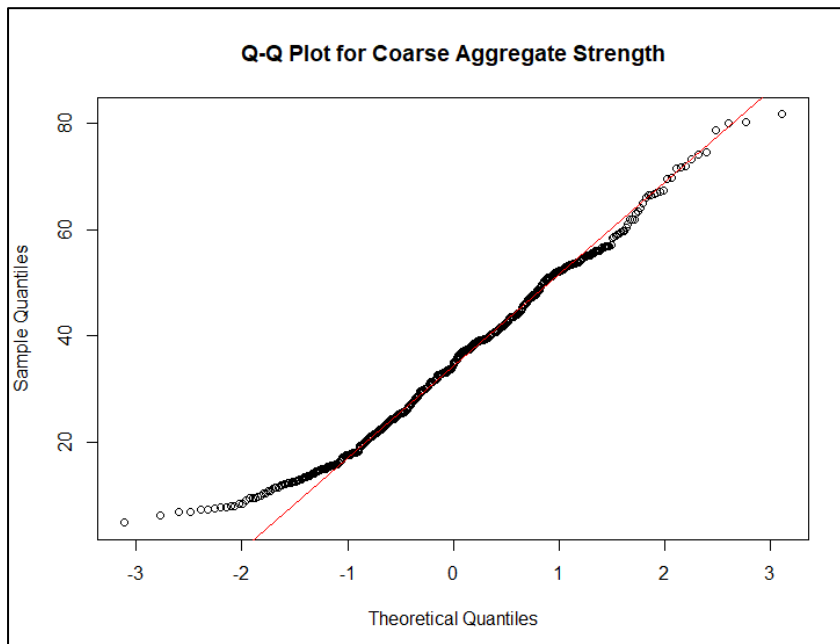


**Figure 44**  
*Histogram of Compressive Strength (Fine Aggregate)*

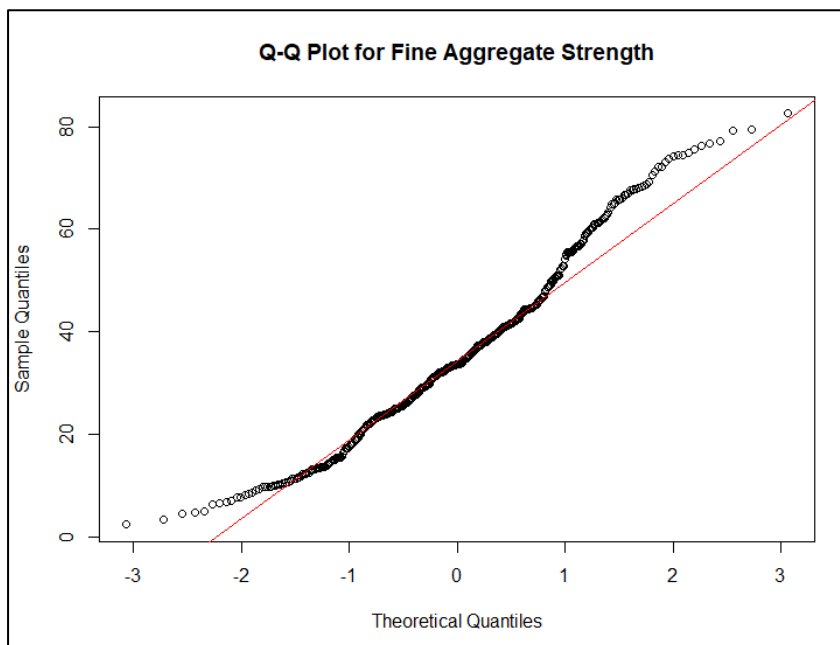


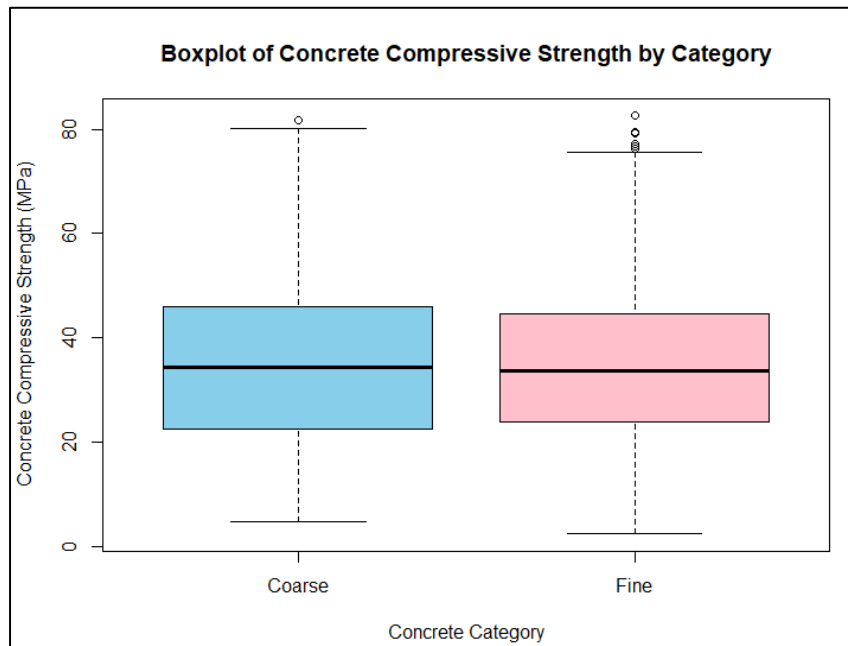
**Figure 45**

*Q-Q Plot for Compressive Strength (Coarse Aggregate)*

**Figure 46**

*Q-Q Plot for Compressive Strength (Fine Aggregate)*



**Figure 47***Boxplot Comparison of Compressive Strength by Aggregate Type***Figure 48***Levene's Test for Equality of Variances*

```

124 # Combine the data into a single data frame for Levene's Test
125 combined_data <- data.frame(
126   strength = c(coarse_strength, fine_strength),
127   category = rep(c("Coarse", "Fine"), c(length(coarse_strength), length(fine_strength)))
128 )
129
130 # Perform Levene's Test for equality of variances
131
132 levene_test <- leveneTest(strength ~ category, data = combined_data)
133 print(levene_test)
134
135
150:1 Hypothesis Testing 1

```

R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/

```

> print(levene_test)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1    0.5 0.4797
1003

```

The t-test results showed a p-value of 0.6594, which was much greater than the significance level of 0.05, concluding that there is no statistically significant evidence to reject the null hypothesis and consequently that the mean compressive strength of coarse aggregate is not greater than that of fine aggregate (Figure 49).

**Figure 49**  
*T-Test Results Summary*

```

135 # Check the result of Levene's Test
136- if (levne_test$`Pr(>F)`[1] > 0.05) {
137   cat("Levene's Test p-value:", levne_test$`Pr(>F)`[1], "\n")
138   cat("Variances are equal. Performing standard t-test...\n")
139
140   # Perform standard t-test (equal variances)
141   t_test <- t.test(coarse_strength, fine_strength, alternative = "greater", var.equal = TRUE)
142
143- } else {
144   cat("Levene's Test p-value:", levne_test$`Pr(>F)`[1], "\n")
145   cat("Variances are not equal. welch's t-test is more appropriate...\n")
146
147   # Perform t-test with unequal variances (welch's t-test)
148   t_test <- t.test(coarse_strength, fine_strength, alternative = "greater", var.equal = FALSE)
149- }
150
151 # Display the results of the t-test
152 print(t_test)
153
154

```

150:1 Hypothesis Testing 1

R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/

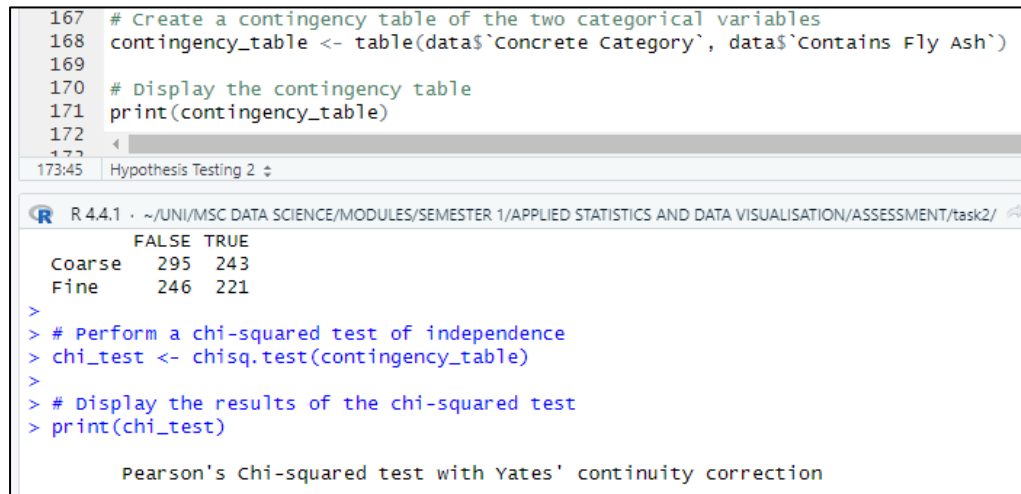
Two Sample t-test

data: coarse\_strength and fine\_strength  
t = -0.41107, df = 1003, p-value = 0.6594  
alternative hypothesis: true difference in means is greater than 0  
95 percent confidence interval:  
-2.119937 Inf  
sample estimates:  
mean of x mean of y  
35.05346 35.47701

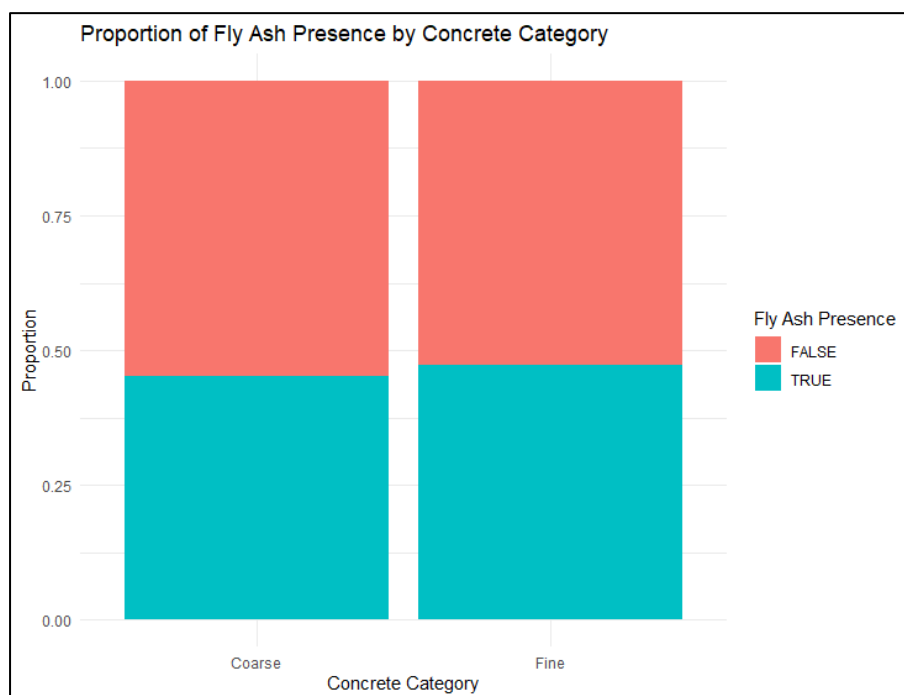
### 1.5.2 Chi-Squared Hypothesis Test

The purpose of this test was to determine whether the presence of Fly Ash is associated with the Concrete Category. The formulated null hypothesis stated that the presence of Fly Ash is independent of the Concrete Category, and the alternative hypothesis argued that the presence of Fly Ash is associated with the Concrete Category. A chi-squared test of independence was chosen because both variables involved, Fly Ash presence (TRUE/FALSE) and Concrete Category (Coarse/Fine), are categorical. A contingency table was created, outlining the frequencies of Fly Ash presence within each Concrete Category (Figure 50). The chi-squared test returned a test statistic of 0.38487 and a p-value of 0.535, which is greater than the significance level of 0.05, indicating no statistically significant association between the presence of Fly Ash and the Concrete Category. To further illustrate the result, we generated a stacked bar chart showing that the proportions of Fly Ash presence (TRUE/FALSE) were similar across the two categories (Figure 51).

**Figure 50**  
*Contingency Table for Fly Ash and Concrete Category*



**Figure 51**  
*Stacked Bar Chart for Fly Ash Proportion by Category*



### 1.5.3 ANOVA Hypothesis Test

This test involved testing three hypotheses: whether the mean compressive strength is different between Concrete Categories or between samples with and without Fly Ash, and whether an interaction effect exists between these two factors influencing the mean compressive strength. Because the predictors were two categorical variables, each with two

levels, and the response variable was a continuous numerical variable, we selected a two-way ANOVA test.

To ensure the categorical predictors were treated as such, they were converted into factors before proceeding to create a two-way ANOVA model (Figure 52).

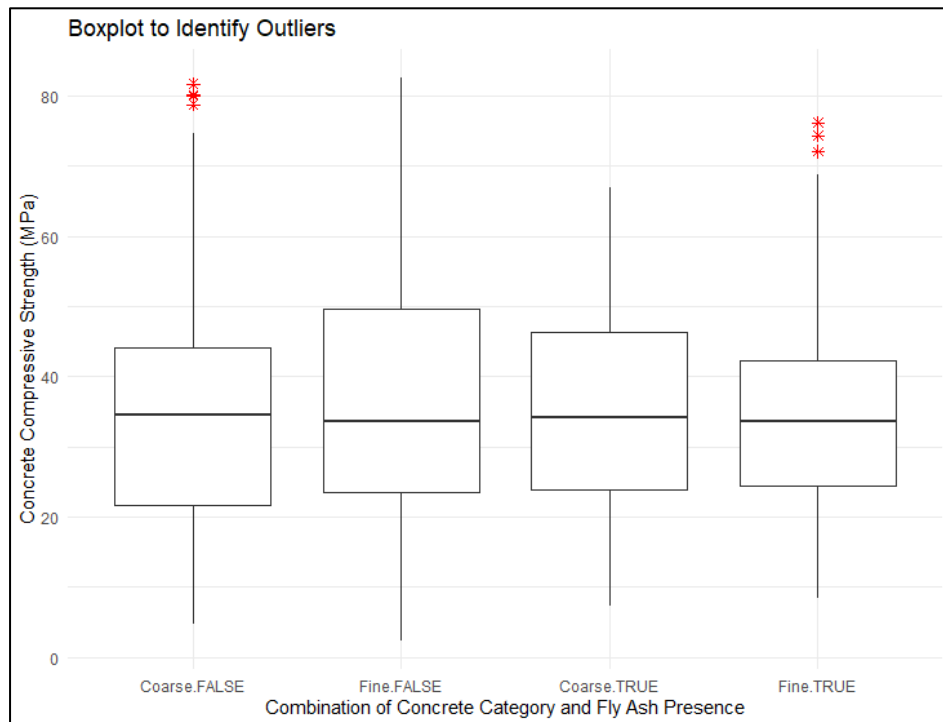
**Figure 52**

*ANOVA Model Definition*

```
222 data$`Concrete Category` <- as.factor(data$`Concrete Category`)
223 data$`Contains Fly Ash` <- as.factor(data$`Contains Fly Ash`)
224
225
226 # create a two-way ANOVA model
227 model <- aov(`Concrete compressive strength(MPa, megapascals)` ~
228             `Concrete Category` * `Contains Fly Ash`, data = data)
229
```

To ensure the model's validity, we checked its compliance with the ANOVA assumptions, starting with the assumption of Independence of observations, which requires that there is no relationship between the observations within or between groups. We considered this to be satisfied as the data is assumed to have been collected randomly.

The next assumption we checked is the absence of significant outliers, as they can affect the results of the ANOVA by affecting the means and variances within groups. For this, a boxplot was created to visually assess the presence of outliers in the dependent variable, concrete compressive strength, across all combinations of the independent variables (Concrete Category and Fly Ash presence), resulting in the finding of outliers in the Coarse.FALSE and Fine.TRUE groups. (Figure 53)

**Figure 53***Boxplot of Compressive Strength by ANOVA Groups*

The next assumption checked is that the dependent variable, concrete compressive strength, should be approximately normally distributed for each combination of the levels of the independent variables. This was assessed using the Shapiro-Wilk test, which evaluates whether a dataset is normally distributed by comparing the data's distribution to a normal distribution. The test results closer to 1 indicate normality and provide a p-value to determine statistical significance. In our case, the test was applied to the residuals of the two-way ANOVA model, showing a result of 0.98445 and a p-value smaller than the significance level of 0.05, meaning a violation of the required ANOVA assumption (Figure 54).



**Figure 54***Shapiro-Wilk Test for Normality of Residuals*

```

262 residuals <- residuals(model)
263 shapiro_test <- shapiro.test(residuals)
264 print(shapiro_test)
265
266
273:1 Hypothesis Testing 3
R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTI
shapiro-wilk normality test
data: residuals
W = 0.98445, p-value = 7.3e-09

```

The next assumption checked was the homogeneity of variances, which requires that the variance of the dependent variable, concrete compressive strength, is equal across all combinations of the independent variables. To check this, we used Levene's test, which evaluates whether the variances in the groups are significantly different by comparing the deviation of data points from their group median. The test returned a value of 10.707 and a p-value below the significance level of 0.05, indicating that the variances are not equal across the groups, violating the assumption required by ANOVA (Figure 55)

**Figure 55***Levene's Test for Homogeneity of Variances*

```

270 levene_test <- leveneTest('Concrete compressive strength(MPa, megapascals)' ~
271                             'Concrete Category' * 'Contains Fly Ash', data = data)
272 print(levene_test)
273
268:3 Hypothesis Testing 3
R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/
Levene's Test for Homogeneity of Variance (center = median)
  Df F value    Pr(>F)
group 3 10.707 6.253e-07 ***
1001
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

To address the violations of the ANOVA assumptions, the dependent variable, concrete compressive strength, was log-transformed to stabilise its variance, reduce the influence of outliers, and improve the normality of the data. We then build a new model using this log-transformed variable (Figure 56).

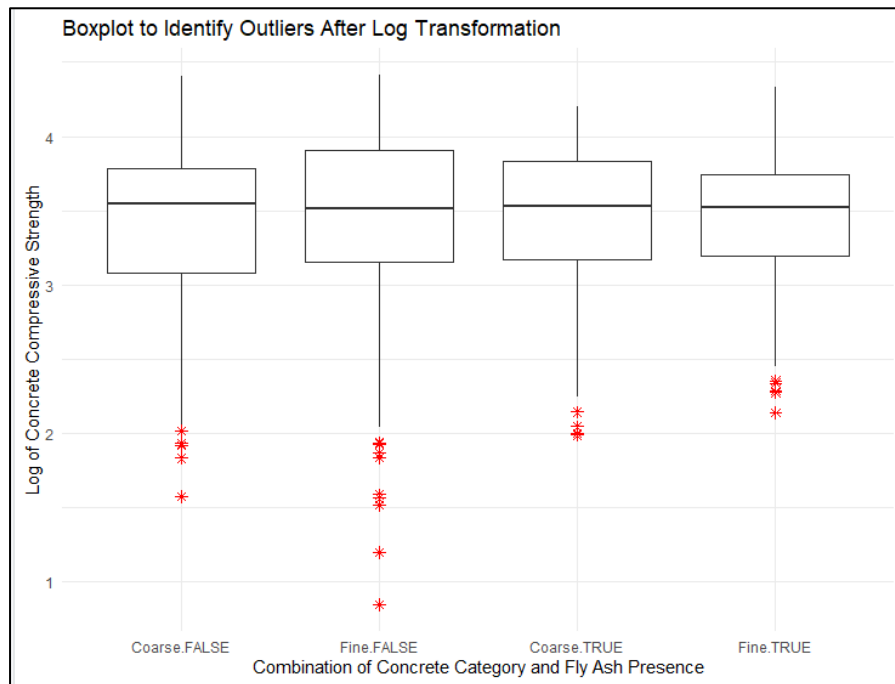
**Figure 56***Log-Transformed ANOVA Model Definition*

```

283 data$log_strength <- log(data$`Concrete compressive strength(MPa, megapascals)`)
284
285 # Build a new model with log-transformed dependent variable
286 model_log <- aov(log_strength ~ `Concrete Category` * `Contains Fly Ash`, data = data)
287

```

After applying the log transformation to the dependent variable, the assumptions of ANOVA were re-checked to assess the effectiveness of the transformation, starting by using boxplots to check for outliers in the log-transformed compressive strength across all group combinations. The check showed that outliers remained in all groups (Figure 57).

**Figure 57***Boxplot of Log-Transformed Compressive Strength*

Similarly, the normality of residuals was reassessed using the Shapiro-Wilk test, which yielded a value below the significance level of 0.05, indicating that the residuals remain significantly non-normal. (Figure 58)

**Figure 58**  
*Shapiro-Wilk Test for Log-Transformed Residuals*

```

302 # Check for normality of residuals after log transformation
303 # Extract residuals
304 residuals_log <- residuals(model_log)
305 # Perform Shapiro-Wilk test
306 shapiro_test_log <- shapiro.test(residuals_log)
307 cat("Shapiro-Wilk Test for Normality (After Log Transformation):\n")
308 print(shapiro_test_log)
309
310 # Check for homogeneity of variances after log transformation
311 # Levene's test for homogeneity of variances with log-transformed variable
312 levene_test_log <- leveneTest(log_strength ~ `Concrete Category` * `Contains Fly Ash`, data = data)
313 cat("\nLevene's Test for Homogeneity of Variances (After Log Transformation):\n")
314 print(levene_test_log)
315
316
317

```

Hypothesis Testing 3

R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/

Shapiro-Wilk Test for Normality (After Log Transformation):  
> print(shapiro\_test\_log)

Shapiro-Wilk normality test

data: residuals\_log  
W = 0.95307, p-value < 2.2e-16

>

> # Check for homogeneity of variances after log transformation  
> # Levene's test for homogeneity of variances with log-transformed variable  
> levene\_test\_log <- leveneTest(log\_strength ~ `Concrete Category` \* `Contains Fly Ash`, data = data)  
> cat("\nLevene's Test for Homogeneity of Variances (After Log Transformation):\n")

Levene's Test for Homogeneity of Variances (After Log Transformation):  
> print(levene\_test\_log)

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	3	8.0025	2.836e-05 ***
	1001		

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since the log transformation did not resolve the violations of normality and homogeneity of variances, we decided to use a robust two-way ANOVA implemented via the `t2way` function from the `WRS2` package. This approach provides reliable results in the presence of outliers and unequal variances. The robust analysis showed no statistically significant main effects (all values greater than 0.05 significance level) for Concrete Category or Fly Ash presence, as well as no significant interaction effect, indicating that neither factor, independently nor in combination, has a meaningful influence on the compressive strength, consequently supporting the null hypotheses for all three hypotheses' tests. (Figure 59)

**Figure 59**  
*Robust ANOVA Model Results Summary*

```

325 robust_anova <- t2way(log_strength ~ `Concrete Category` * `Contains Fly Ash`, data = data)
326
327

```

Hypothesis Testing 3

R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/

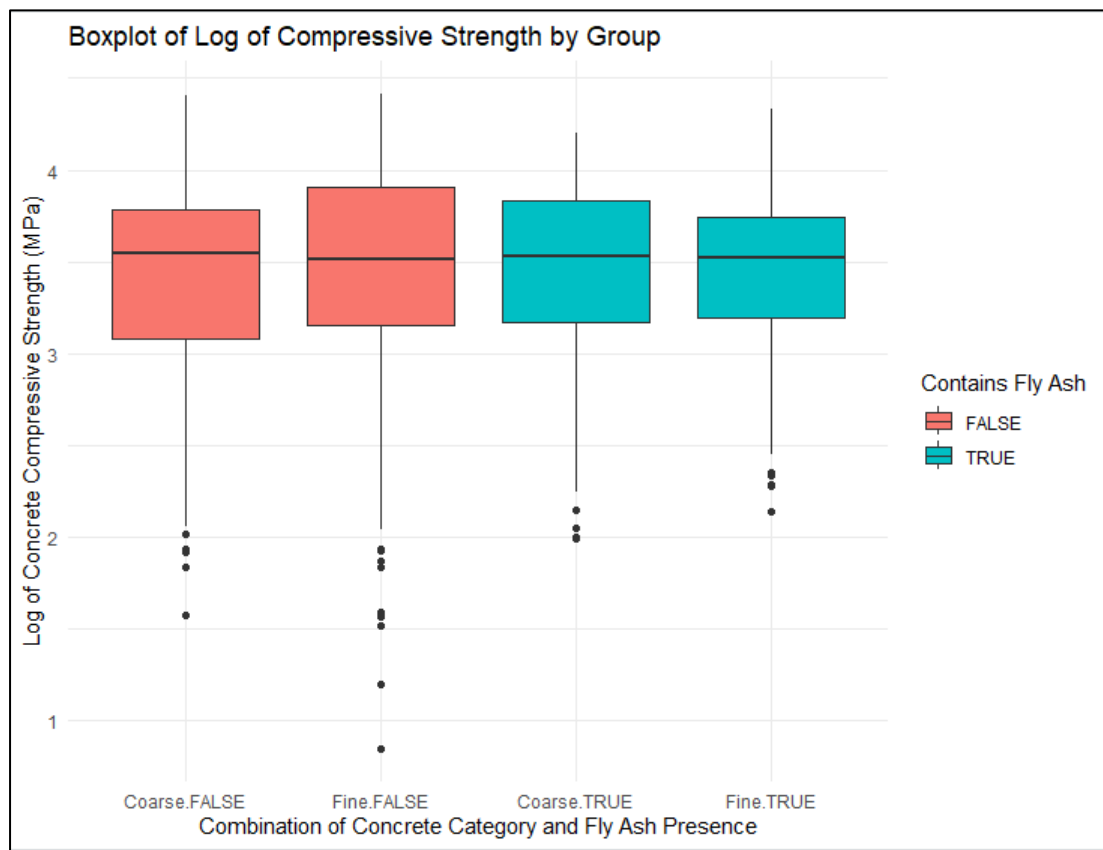
call:  
t2way(formula = log\_strength ~ `Concrete Category` \* `Contains Fly Ash`,  
data = data)

	value	p.value
Concrete Category	0.0024	0.961
Contains Fly Ash	0.0015	0.969
Concrete Category:Contains Fly Ash	0.5330	0.466

Lastly, we generated a boxplot to visualise the log-transformed compressive strength across all group combinations confirming the presence of minimal differences between groups and reinforcing the conclusion that neither the individual factors nor their interaction significantly affects the compressive strength (Figure 60)

**Figure 60**

*Boxplot of Log-Transformed Strength by ANOVA Groups*



## 1.6 Conclusion

The statistical analysis carried out provides key insights for the concrete mix company. For instance, the first regression model identified cement as a significant factor in increasing compressive strength by 0.0762 MPa per kilogram. However, it accounted for 23.8% of the variability requiring additional predictors in the model.

Additionally, the second regression model provided a more comprehensive approach, identifying key predictors such as Cement, Water, Superplasticizer, Age, and Blast Furnace Slag influencing the concrete mix's compressive strength. The complete regression equation allows the company to predict the compressive strength, assess component trade-offs, as well as create simulations of different concrete mixes.

Moreover, the t-test showed that aggregate type, whether it is coarse or fine, does not significantly impact strength, and the Chi-squared test proved the independence of Fly Ash and aggregate type, allowing the company to focus on these two components separately for further analysis, such as exploring their impact on compressive strength. Finally, the ANOVA results indicate that neither aggregate type nor Fly Ash presence significantly affects compressive strength, allowing the company to focus on other factors for optimising the concrete mix.

## References

- Chao, L. L. (1980). *Introduction to Statistics*. California State University.
- Clarke, G. M. (1994). *Statistics & Experimental Design* (3rd ed.). Arnold.
- Fischetti, T. (2015). *Data Analysis with R*. Packt Publishing.
- Moore, D. S. (1989). *Introduction to the Practice of Statistics*. Purdue University.