# Concrete Mix Statistical Analysis

Jose Leonardo Wong

## Contents

# 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*

```
Source on Save

#####INITIAL EXPLORATORY DATA ANALYSIS#####

# Reading the data
data <- read_excel('concrete compressive strength.xlsx')

# Dimension of the dataset
dim(data)

# Structure of the dataset and data types
str(data)

25:10  INITIAL EXPLORATORY DATA ANALYSIS
```

```
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 0 ...
 $ Water  (component 4)(kg in a m^3 mixture)            : num [1:1030] 162 162 228 228 192 228 228 228 228 228 ...
 $ Superplasticizer (component 5)(kg in a m^3 mixture)  : num [1:1030] 2.5 2.5 0 0 0 0 0 0 0 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 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/m³, water content, ranging from 121.8 to 247 kg/m³ 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*

Source on Save

```r
# Summary statistics
summary_stats <- as.data.frame(summary(data))
print(summary_stats)
# Checking for missing values
missing_values <- colSums(is.na(data))
missing_values_df <- as.data.frame(missing_values)
rownames(missing_values_df) <- colnames(data)
print(missing_values_df)
# 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)
# Count number of duplicate rows
num_duplicates <- nrow(duplicated_rows_df)
cat("Number of duplicate rows:", num_duplicates, "\n")
```

47:24 INITIAL EXPLORATORY DATA ANALYSIS

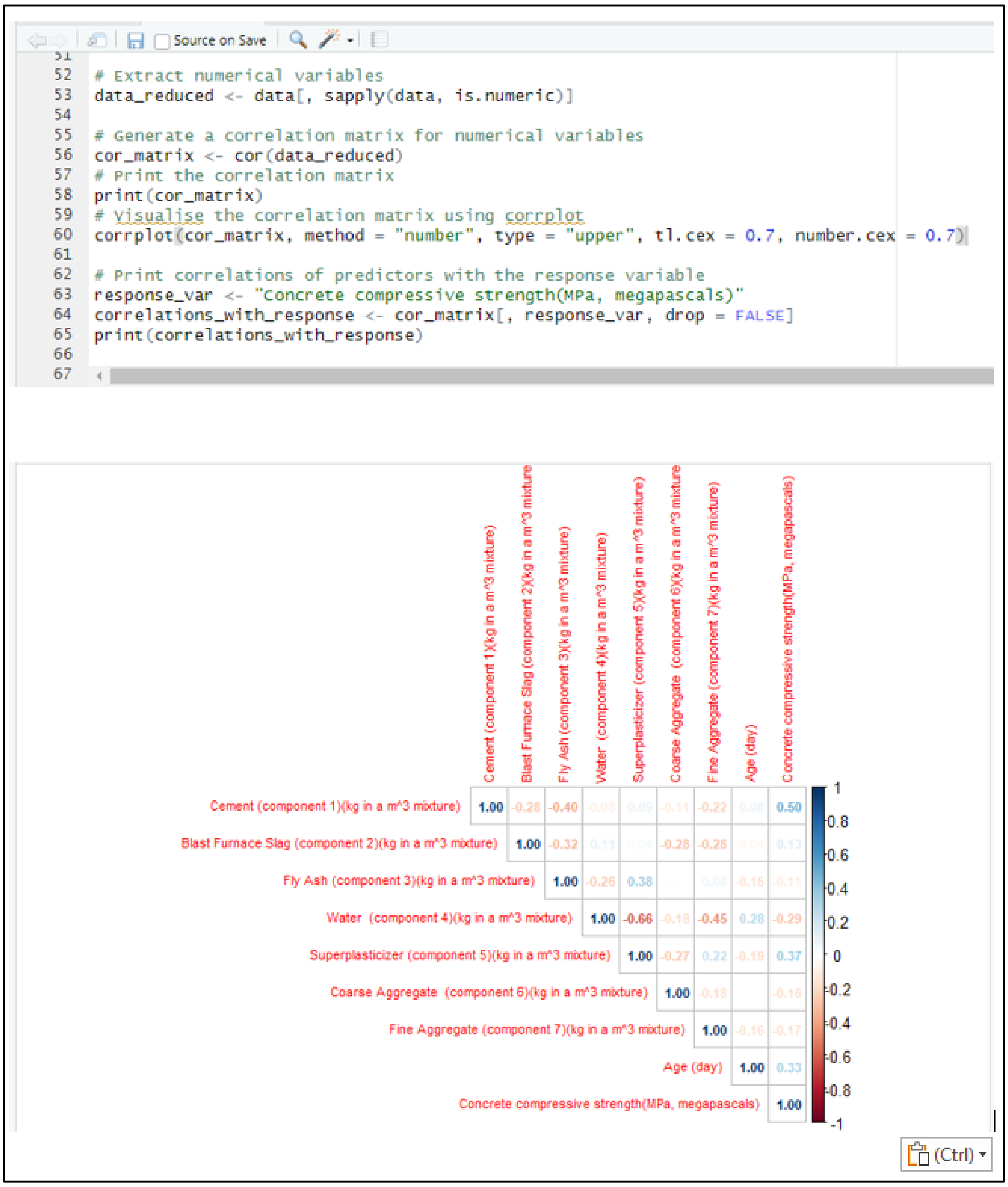
R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSE

```
> rownames(missing_values_df) <- colnames(data)
> print(missing_values_df)
                                                      missing_values
Cement  (component 1)(kg in a m^3 mixture)                         0
Blast Furnace Slag (component 2)(kg in a m^3 mixture)              0
Fly Ash (component 3)(kg in a m^3 mixture)                         0
Water  (component 4)(kg in a m^3 mixture)                          0
Superplasticizer (component 5)(kg in a m^3 mixture)                0
Coarse Aggregate  (component 6)(kg in a m^3 mixture)               0
Fine Aggregate (component 7)(kg in a m^3 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^3 mixture)                      280
Blast Furnace Slag (component 2)(kg in a m^3 mixture)           187
Fly Ash (component 3)(kg in a m^3 mixture)                      163
Water  (component 4)(kg in a m^3 mixture)                       205
Superplasticizer (component 5)(kg in a m^3 mixture)             155
Coarse Aggregate  (component 6)(kg in a m^3 mixture)            284
Fine Aggregate (component 7)(kg in a m^3 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
```

Subsequently, we generated a correlation matrix, finding that cement has the strongest positive correlation with the response variable compressive strength (r=0.498), followed by superplasticizer (r=0.366r) and age (r=0.329). Conversely, water content showed a moderate negative correlation (r=−0.290). Moreover, potential multicollinearity was noted between water and superplasticizer (r=−0.657) (Figure 3)

**Figure 3**

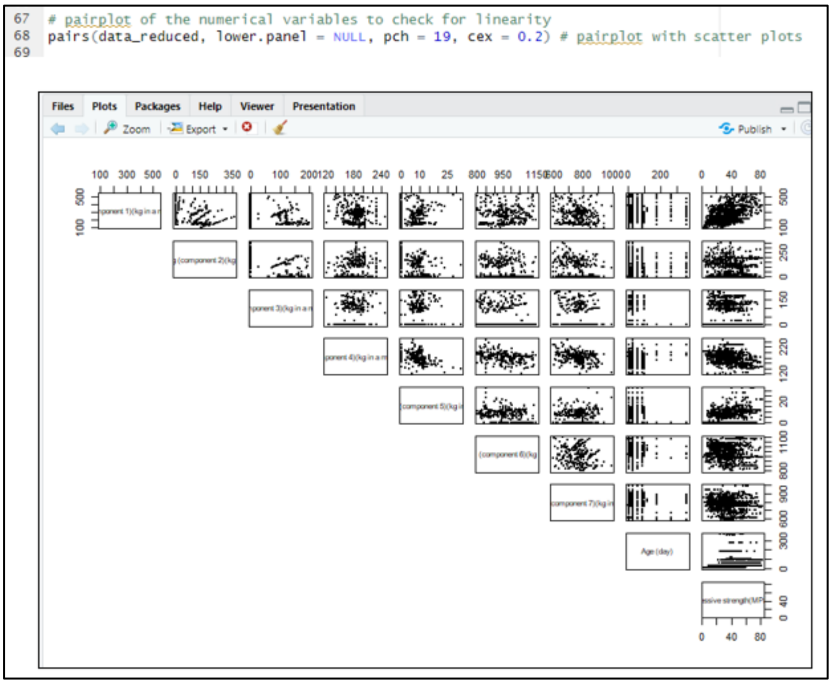
*Correlation Matrix for Concrete Mix Variables*



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**

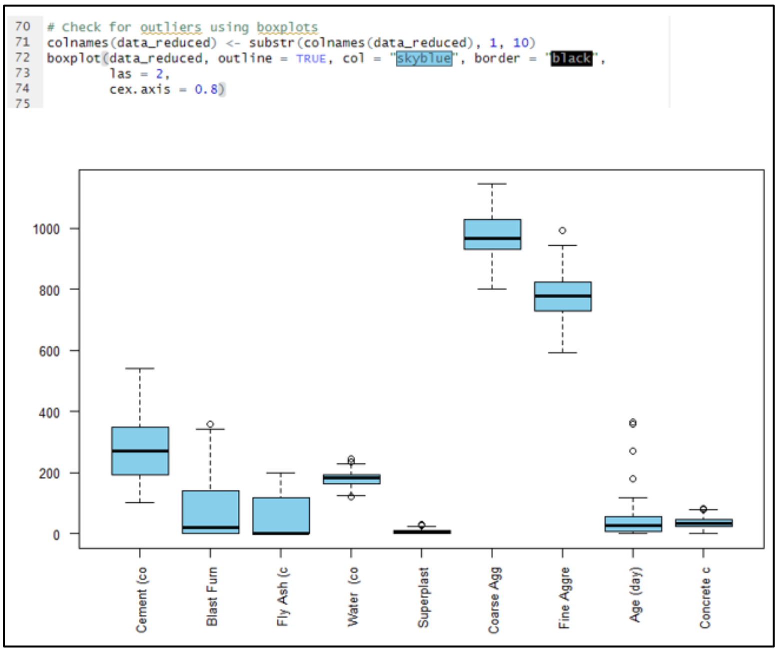
*Pair Plot Matrix of Concrete Mix Variables*




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*

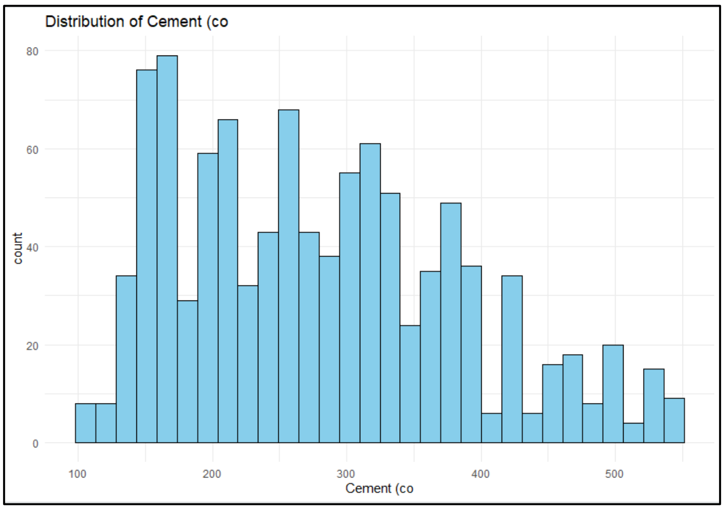



**Figure 7**

*Histogram of Blast Furnace Slag Distribution*

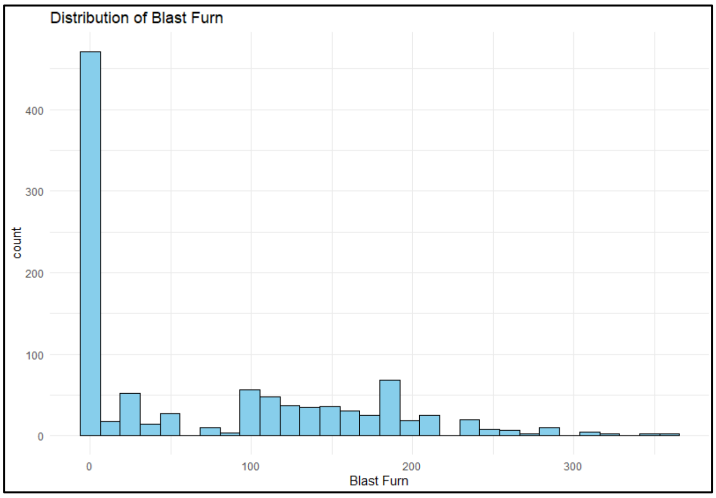



**Figure 8**

*Histogram of Fly Ash Distribution*

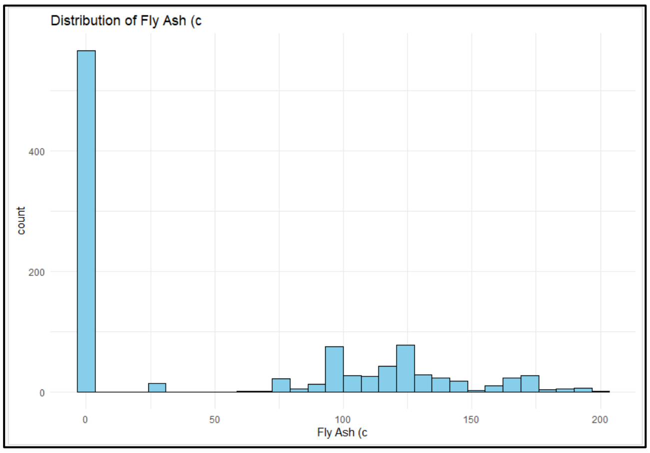


**Figure 9**

*Histogram of Water Content Distribution*

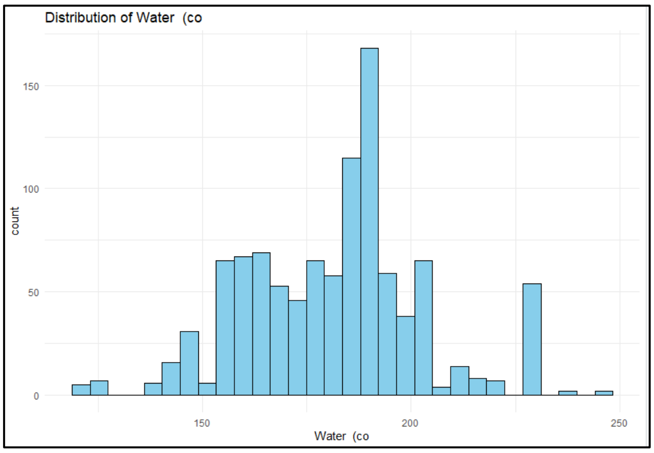



**Figure 10**

*Histogram of Superplasticizer Distribution*

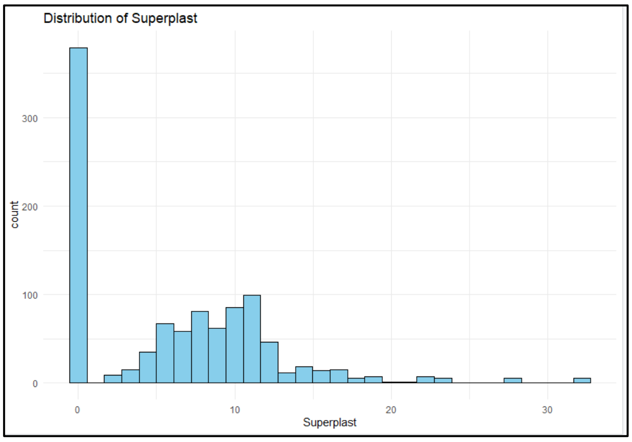



**Figure 11**

*Histogram of Coarse Aggregate Distribution*

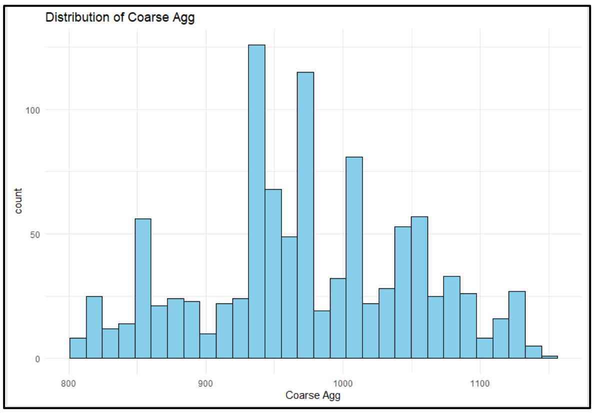


**Figure 12**

*Histogram of Fine Aggregate Distribution*

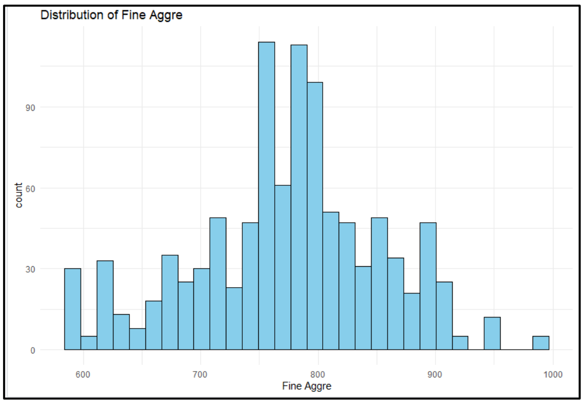



**Figure 13**

*Histogram of Age (Days) Distribution*

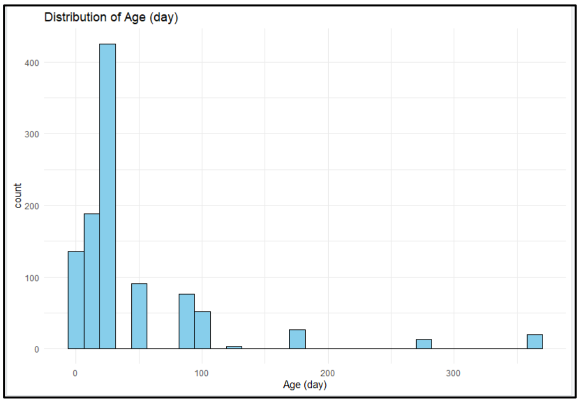
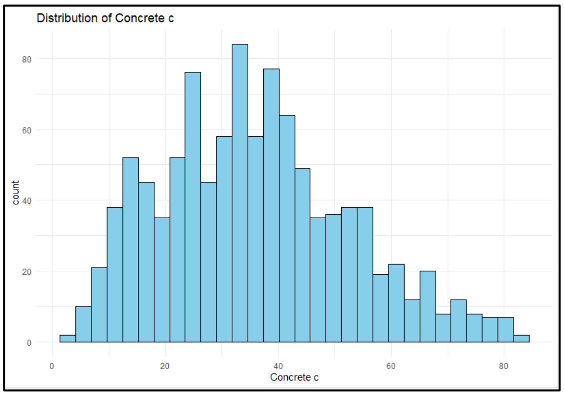



**Figure 14**

*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 varaible should nor 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*

```
# Log-transform variables with dominant zeros
data$`Blast Furnace Slag (component 2)(kg in a m^3 mixture)` <- ifelse(
  data$`Blast Furnace Slag (component 2)(kg in a m^3 mixture)` > 0,
  log1p(data$`Blast Furnace Slag (component 2)(kg in a m^3 mixture)`),
  0
)
data$`Fly Ash (component 3)(kg in a m^3 mixture)` <- ifelse(
  data$`Fly Ash (component 3)(kg in a m^3 mixture)` > 0,
  log1p(data$`Fly Ash (component 3)(kg in a m^3 mixture)`),
  0
)
data$`Superplasticizer (component 5)(kg in a m^3 mixture)` <- ifelse(
  data$`Superplasticizer (component 5)(kg in a m^3 mixture)` > 0,
  log1p(data$`Superplasticizer (component 5)(kg in a m^3 mixture)`),
  0
)

# Log-transform the remaining skewed variable (Age)
data$`Age (day)` <- log1p(data$`Age (day)`)

# Check the distributions of predictors after transformation
transformed_vars <- c(
  "Blast Furnace Slag (component 2)(kg in a m^3 mixture)",
  "Fly Ash (component 3)(kg in a m^3 mixture)",
  "Superplasticizer (component 5)(kg in a m^3 mixture)",
  "Age (day)"
)
lapply(transformed_vars, function(var) {
  ggplot(data, aes_string(x = paste0("`", var, "`"))) +
    geom_histogram(bins = 30, fill = "skyblue", color = "black") +
    ggtitle(paste("Distribution of", var, "After Transformation")) +
    theme_minimal()
})
```

## 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*

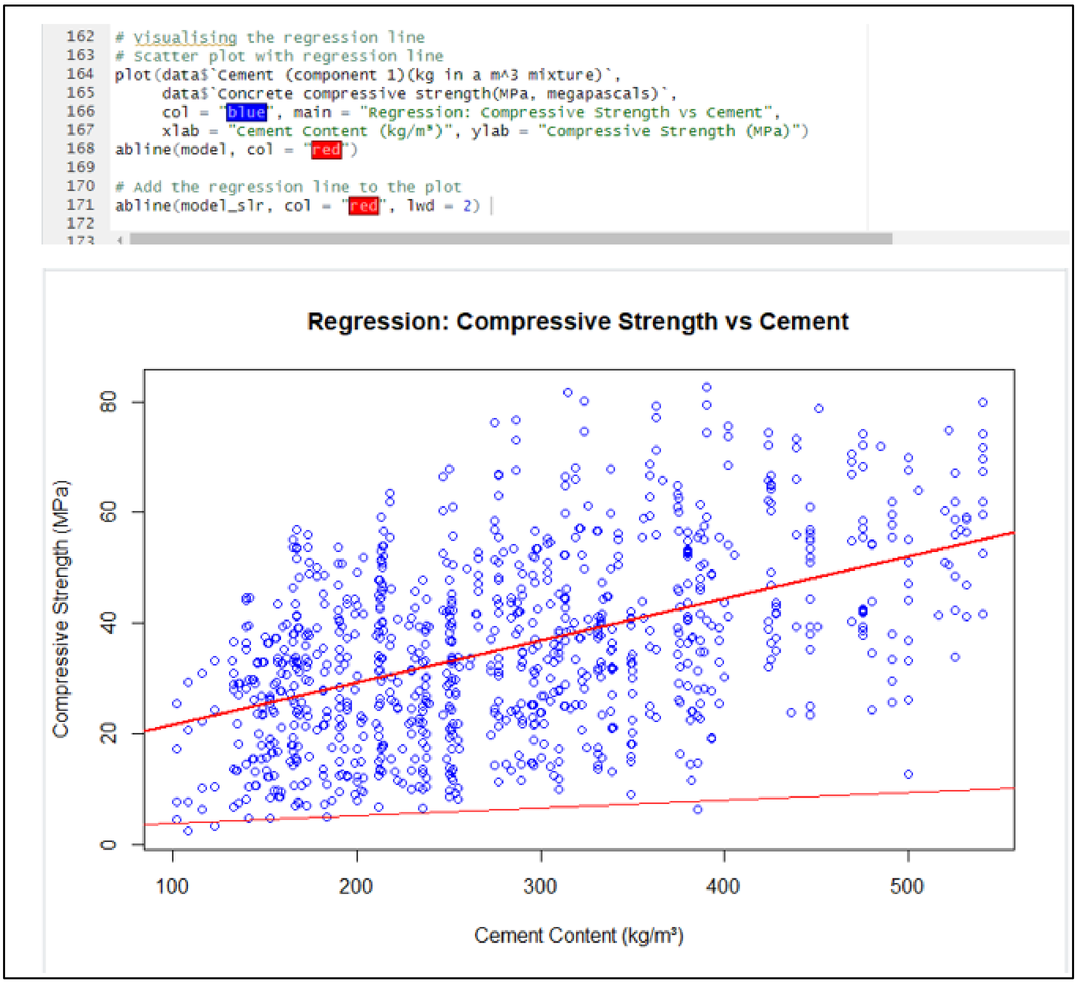
```
#####Simple Linear Regression Model#####

# To predict the concrete compressive strength using the amount of cement as a predictor variab

# Fit a simple linear regression model with cement as predictor variable
model_slr <- lm(`Concrete compressive strength(MPa, megapascals)` ~ `Cement (component 1)(kg in
                data = data)

# Display the summary of the model
summary(model_slr)

159:35  Simple Linear Regression Model

R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/
> #####Simple Linear Regression Model#####
>
>
> # To predict the concrete compressive strength using the amount of cement as a predictor variable.
>
> # Fit a simple linear regression model with cement as predictor variable
> model_slr <- lm(`Concrete compressive strength(MPa, megapascals)` ~ `Cement (component 1)(kg in a m^
+                 data = data)
>
> # Display the summary of the model
> summary(model_slr)

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

Residuals:
    Min      1Q  Median      3Q     Max
-39.482 -10.988  -0.509   9.843  43.729

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  14.0174     1.2794   10.96   <2e-16 ***
`Cement (component 1)(kg in a m^3 mixture)`   0.0762     0.0043   17.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.22 on 1003 degrees of freedom
Multiple R-squared:  0.2384,	Adjusted R-squared:  0.2377
F-statistic:   314 on 1 and 1003 DF,  p-value: < 2.2e-16
```

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**

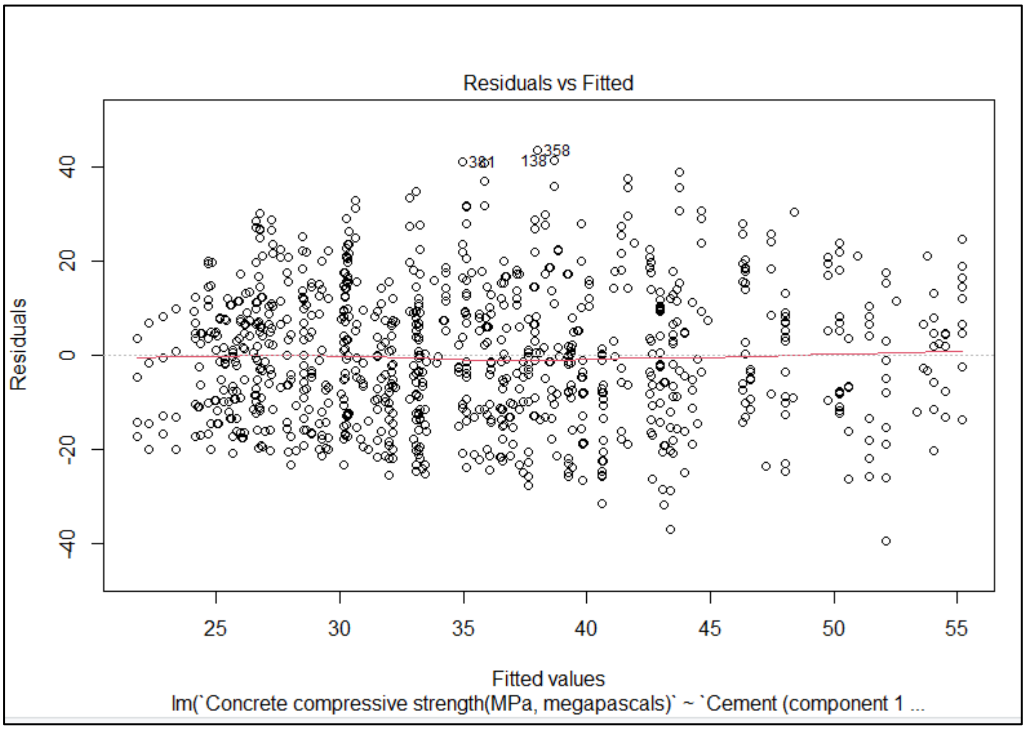
*Scatter Plot of Cement Content vs Compressive Strength with Regression Line*




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**

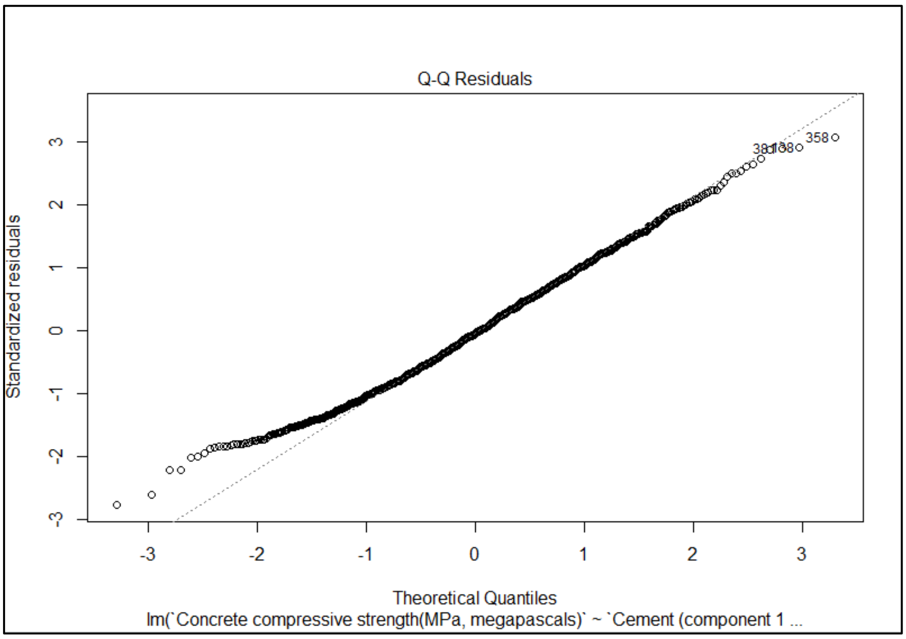
*Residuals vs. Fitted Values Plot*




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).

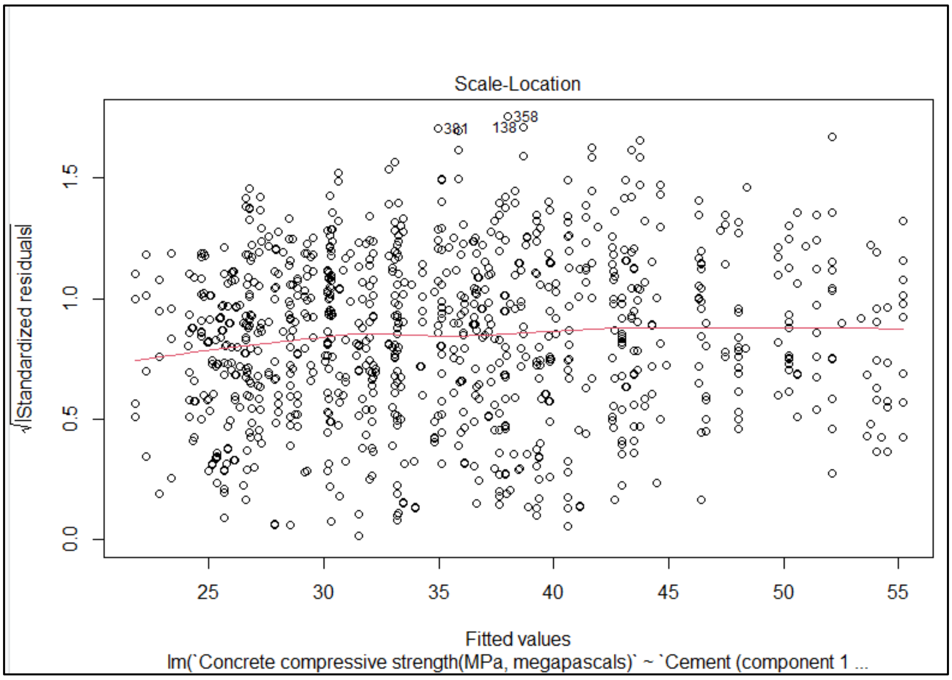
**Figure 20**

*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*

```r
#####Multiple Linear Regression Model#####

# 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)


213:45   Multiple Linear Regression Model
```

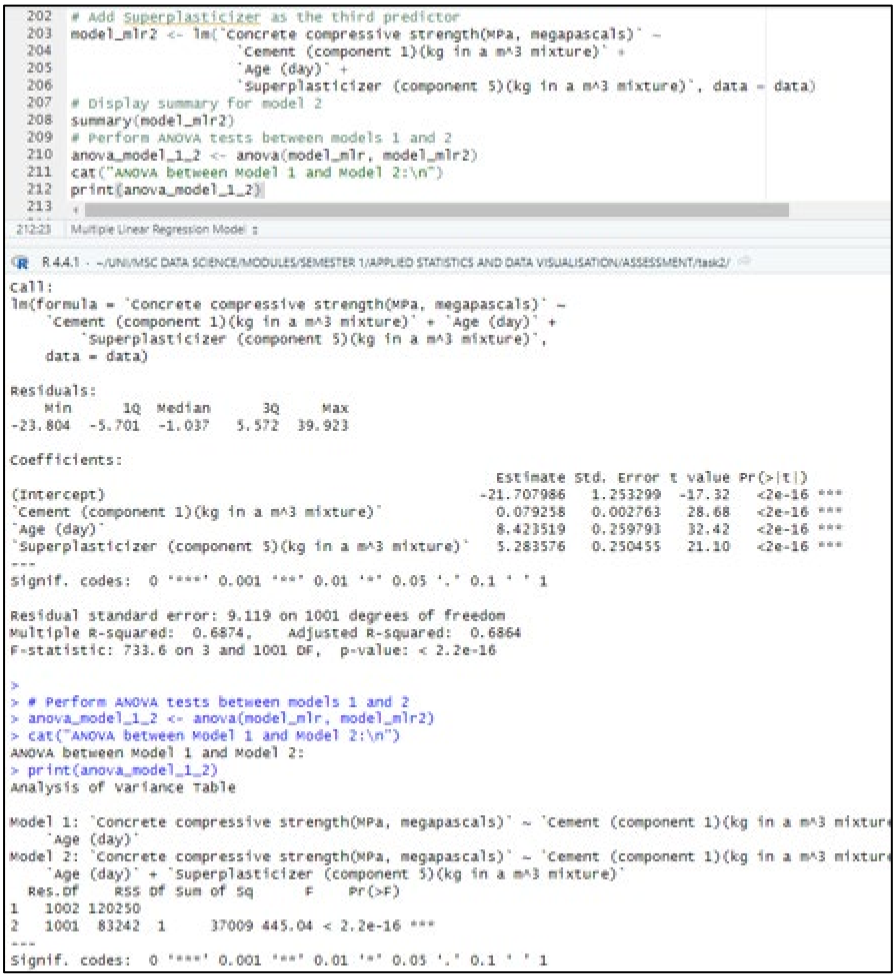
```
R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/
> # To predict the concrete compressive strength using multiple predictor variables.
>
> # Fit initial model with Cement and Age as predictor variables
> model_mlr <- lm(`Concrete compressive 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*



```
# Add Superplasticizer as the third predictor
model_mlr2 <- 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)`, data = data)
# Display summary for model 2
summary(model_mlr2)
# 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")
print(anova_model_1_2)
```

212:23 Multiple Linear Regression Model

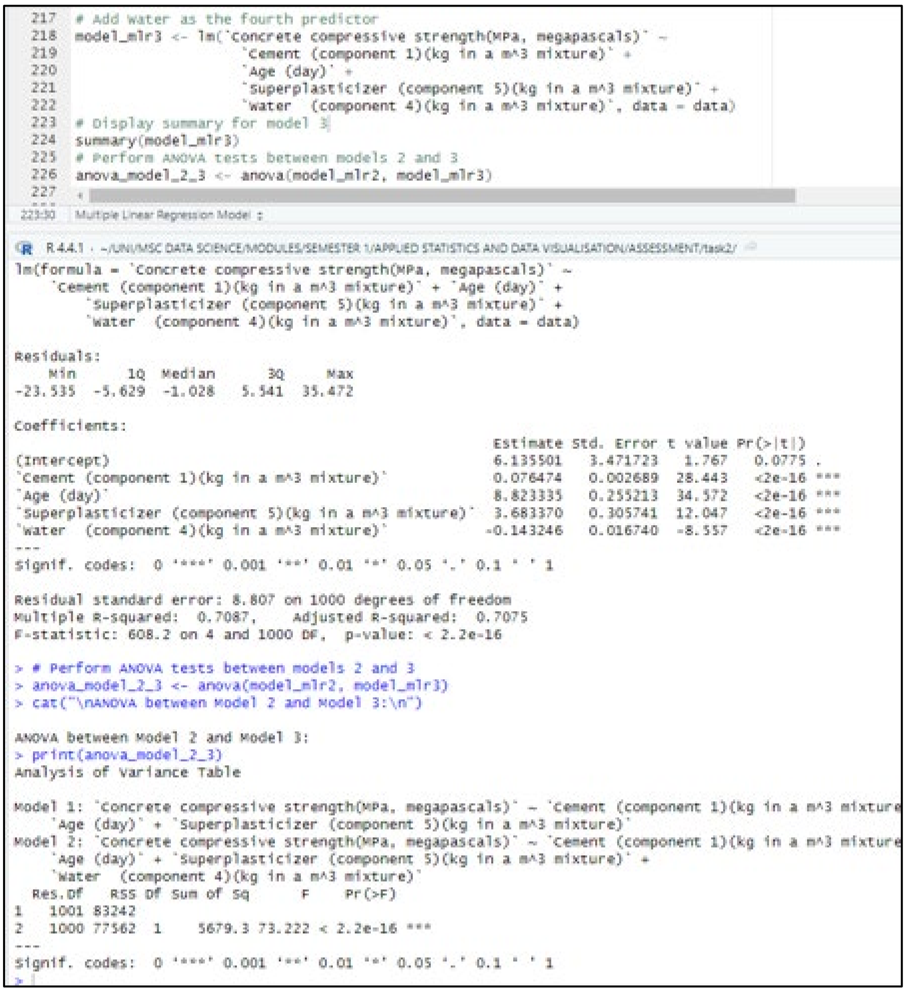
R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/

```
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 mixtur
    `Age (day)`
Model 2: `Concrete compressive strength(MPa, megapascals)` ~ `Cement (component 1)(kg in a m^3 mixtur
    `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 kg/m³) was found to be statistically significant ($p$-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 ($p$-value $< 0.001$) (Figure 24).

**Figure 24**

*Water Content Effect in Multiple Regression Mode*

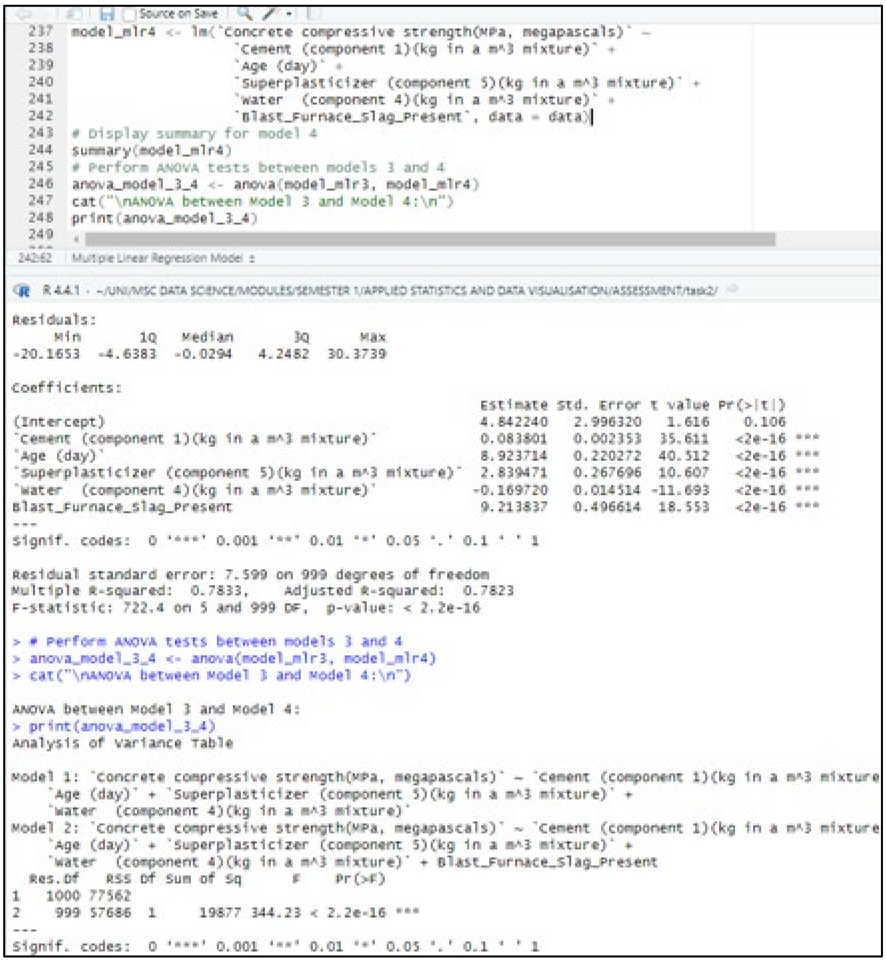


```
# Add water as the fourth predictor
model_mlr3 <- lm(`Concrete compressive strength(MPa, megapascals)` ~
                   `Cement (component 1)(kg in a m^3 mixture)` +
                   `Age (day)` +
                   `Superplasticizer (component 5)(kg in a m^3 mixture)` +
                   `Water  (component 4)(kg in a m^3 mixture)`, data = data)
# Display summary for model 3
summary(model_mlr3)
# Perform ANOVA tests between models 2 and 3
anova_model_2_3 <- anova(model_mlr2, model_mlr3)

223:30  Multiple Linear Regression Model :

R 4.4.1 · ~/UNI/MSC 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*


```
model_mlr4 <- 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`, data = data)
# Display summary for model 4
summary(model_mlr4)
# Perform ANOVA tests between models 3 and 4
anova_model_3_4 <- anova(model_mlr3, model_mlr4)
cat("\nANOVA between Model 3 and Model 4:\n")
print(anova_model_3_4)
```

Multiple Linear Regression Model

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

```
Residuals:
     Min       1Q   Median       3Q      Max 
-20.1653  -4.6383  -0.0294   4.2482  30.3739 

Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                           4.842240   2.996320   1.616    0.106    
`Cement (component 1)(kg in a m^3 mixture)`           0.083801   0.002353  35.611   <2e-16 ***
`Age (day)`                                           8.923714   0.220272  40.512   <2e-16 ***
`Superplasticizer (component 5)(kg in a m^3 mixture)` 2.839471   0.267696  10.607   <2e-16 ***
`Water  (component 4)(kg in a m^3 mixture)`          -0.169720   0.014514 -11.693   <2e-16 ***
Blast_Furnace_Slag_Present                            9.213837   0.496614  18.553   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.599 on 999 degrees of freedom
Multiple R-squared:  0.7833,	Adjusted R-squared:  0.7823 
F-statistic: 722.4 on 5 and 999 DF,  p-value: < 2.2e-16

> # Perform ANOVA tests between models 3 and 4
> anova_model_3_4 <- anova(model_mlr3, model_mlr4)
> cat("\nANOVA between Model 3 and Model 4:\n")

ANOVA between Model 3 and Model 4:
> print(anova_model_3_4)
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)` + 
    `Water  (component 4)(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)` + Blast_Furnace_Slag_Present
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)    
1   1000 77562                                  
2    999 57686  1     19877 344.23 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Subsequently, adding Fine Aggregate as a model predictor resulted in a marginal improvement in model performance, with the adjusted $R^2$ increasing from 0.7823 to 0.7872, indicating a minimal increase in predictive power. Based on this finding, no further predictors were added, and the decision was taken to keep the previous model, including Cement, Age, Superplasticizer, Water, and Blast Furnace Slag Presence, representing a more optimal model avoiding the unnecessary increase in model complexity generated by adding Fine Aggregate (Figure 26)

**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*

```r
best_subsets <- regsubsets(`Concrete compressive strength(MPa, megapascals)` ~
                             `Age (day)` +
                             `Cement (component 1)(kg in a m^3 mixture)` +
                             `Superplasticizer (component 5)(kg in a m^3 mixture)` +
                             `Blast_Furnace_Slag_Present` +
                             `Water  (component 4)(kg in a m^3 mixture)`,
                           data = data, nbest = 1, method = "exhaustive")

# Extract the best model by Adjusted R²
best_model_summary <- summary(best_subsets)
best_model_index <- which.max(best_model_summary$adjr2)

# Extract selected predictors
selected_predictors <- names(coef(best_subsets, best_model_index))[-1]

# Check if any predictors were selected
if (length(selected_predictors) == 0) {
  best_model_formula <- as.formula("`Concrete compressive strength(MPa, megapascals)` ~ 1")
} else {
  best_model_formula <- as.formula(
    paste("`Concrete compressive strength(MPa, megapascals)` ~",
          paste(selected_predictors, collapse = " + "))
  )
}

# Refit and summarize the best model
best_model_refit <- lm(best_model_formula, data = data)
summary(best_model_refit)
```

291:66 Multiple Linear Regression Model

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

```
Residuals:
     Min       1Q   Median       3Q      Max 
-20.1653  -4.6383  -0.0294   4.2482  30.3739 

Coefficients:
                                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                          4.842240   2.996320   1.616    0.106    
`Age (day)`                                          8.923714   0.220272  40.512   <2e-16 ***
`Cement (component 1)(kg in a m^3 mixture)`          0.083801   0.002353  35.611   <2e-16 ***
`Superplasticizer (component 5)(kg in a m^3 mixture)` 2.839471   0.267696  10.607   <2e-16 ***
Blast_Furnace_Slag_Present                           9.213837   0.496614  18.553   <2e-16 ***
`Water  (component 4)(kg in a m^3 mixture)`         -0.169720   0.014514 -11.693   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.599 on 999 degrees of freedom
Multiple R-squared:  0.7833,	Adjusted R-squared:  0.7823 
F-statistic: 722.4 on 5 and 999 DF,  p-value: < 2.2e-16
```

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*

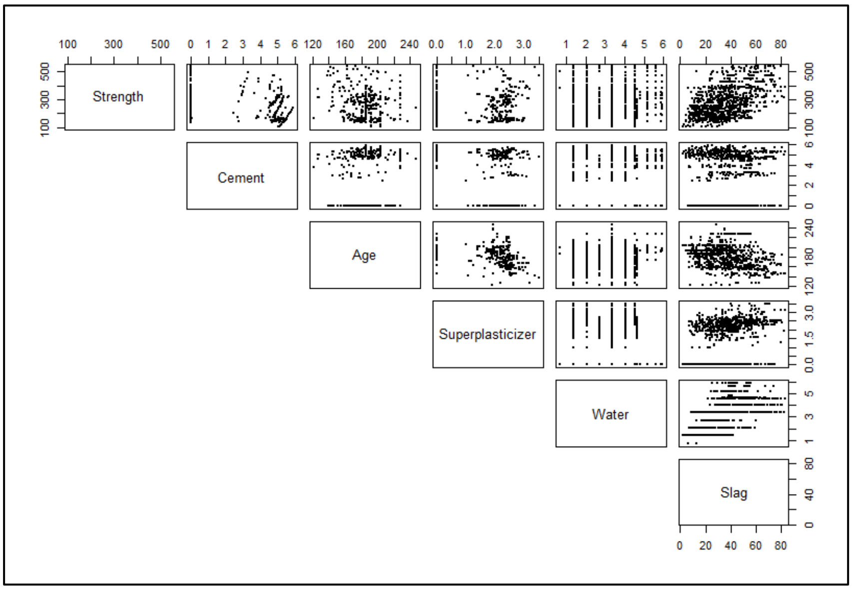



**Figure 29**

*Residuals vs. Fitted Values for Multiple Regression Model*

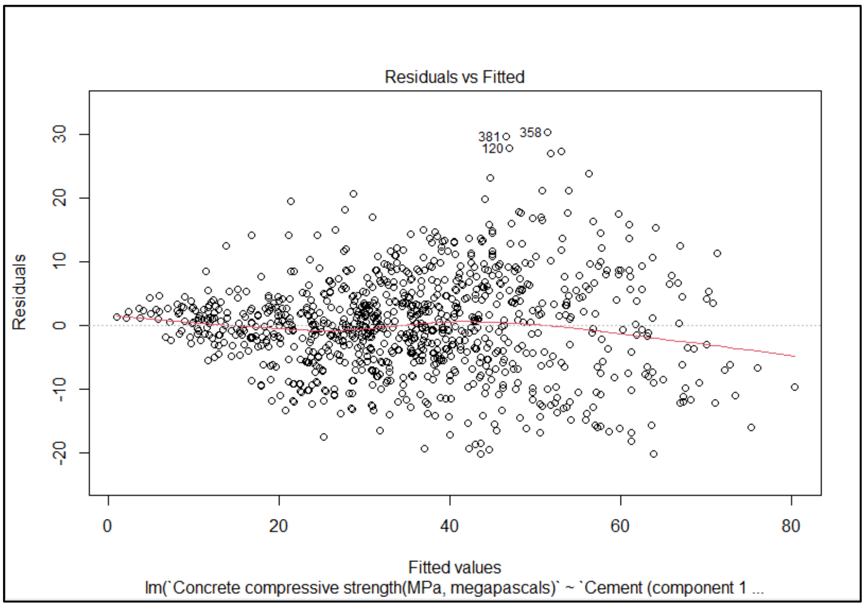


**Figure 30**
*Q-Q Plot for Multiple Regression Model Residuals*

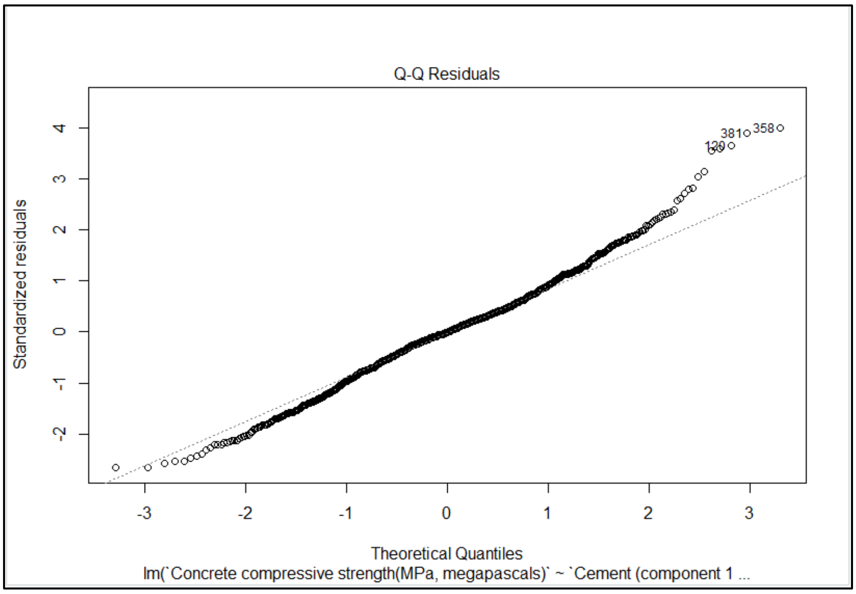
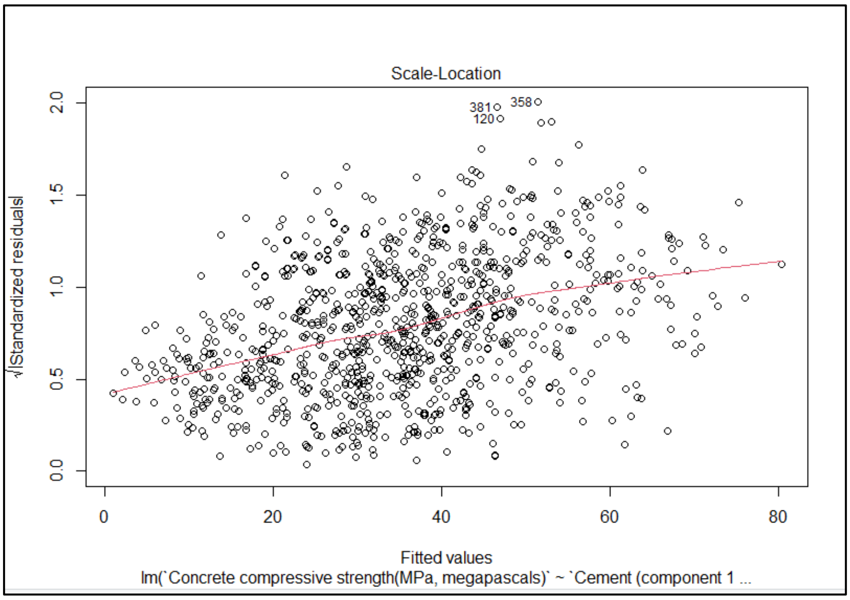



**Figure 31**
*Scale-Location Plot for Multiple Regression Model*




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 stepwise selection
# method to include only significant interactions in the model.

# Standardize numeric predictors before using the interaction terms.
data$Cement_scaled <- scale(data$`Cement (component 1)(kg in a m^3 mixture)`, center = TRUE, scale = TRUE)
data$Age_scaled <- scale(data$`Age (day)`, center = TRUE, scale = TRUE)
data$Superplasticizer_scaled <- scale(data$`Superplasticizer (component 5)(kg in a m^3 mixture)`,
                                       center = TRUE, scale = TRUE)
data$Water_scaled <- scale(data$`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 <- step(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*

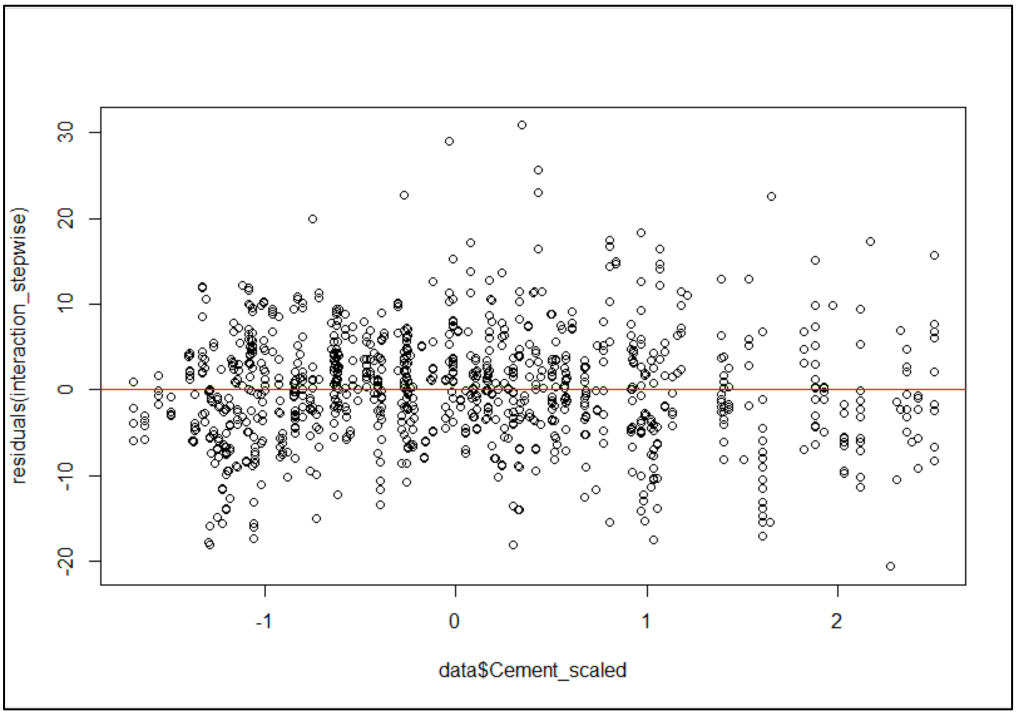



**Figure 34**

*Residuals vs. Water Content Scatter Plot*

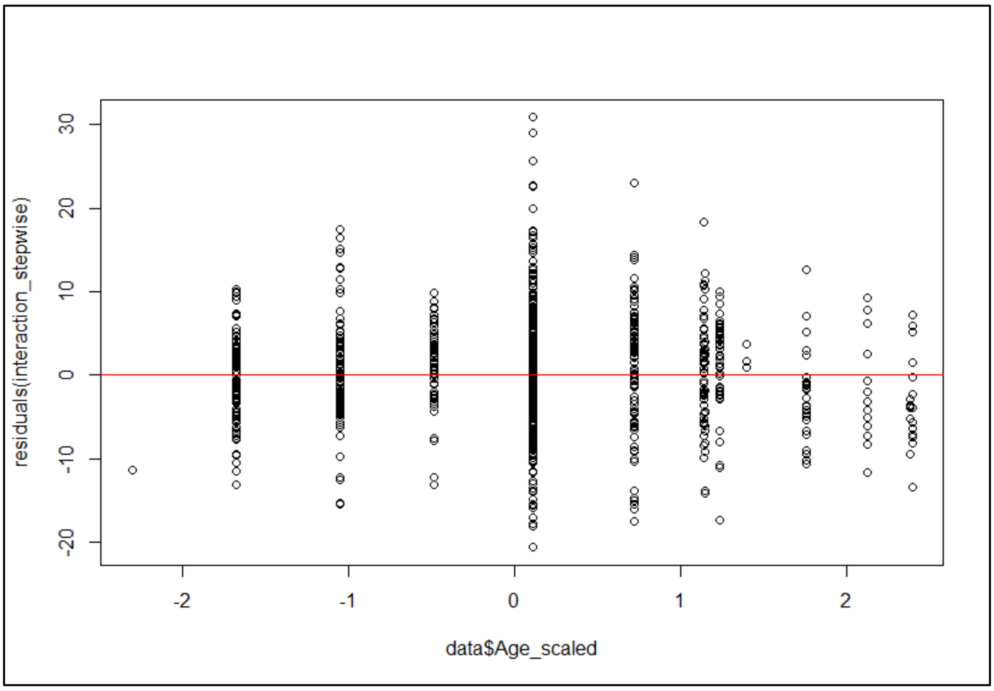


**Figure 35**
*Residuals vs. Superplasticizer Content Scatter Plot*

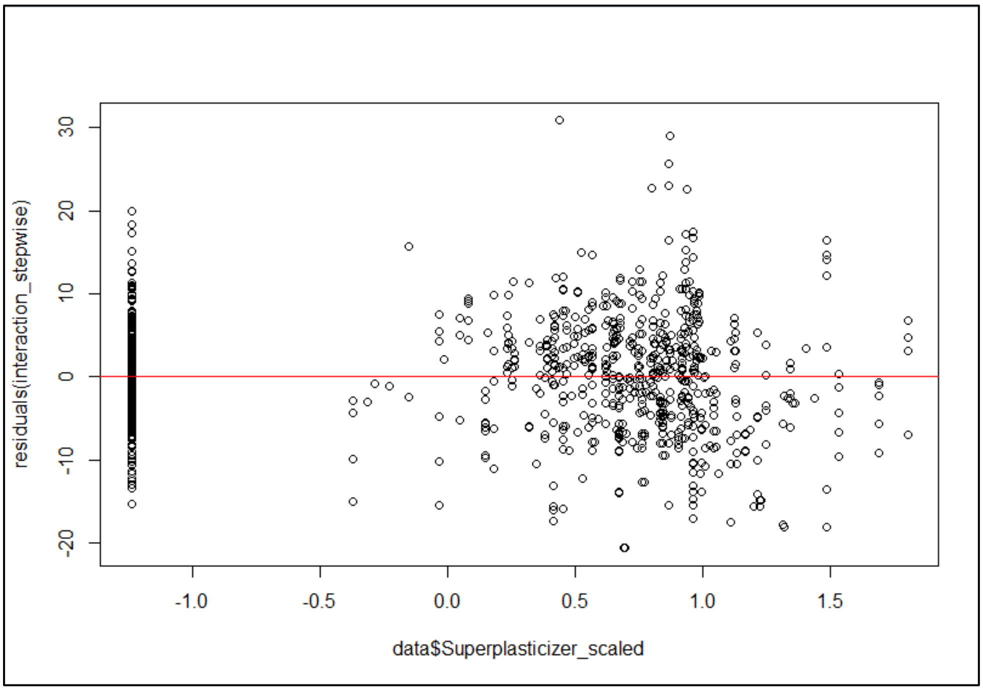



**Figure 36**
*Residuals vs. Age (Days) Scatter Plot*

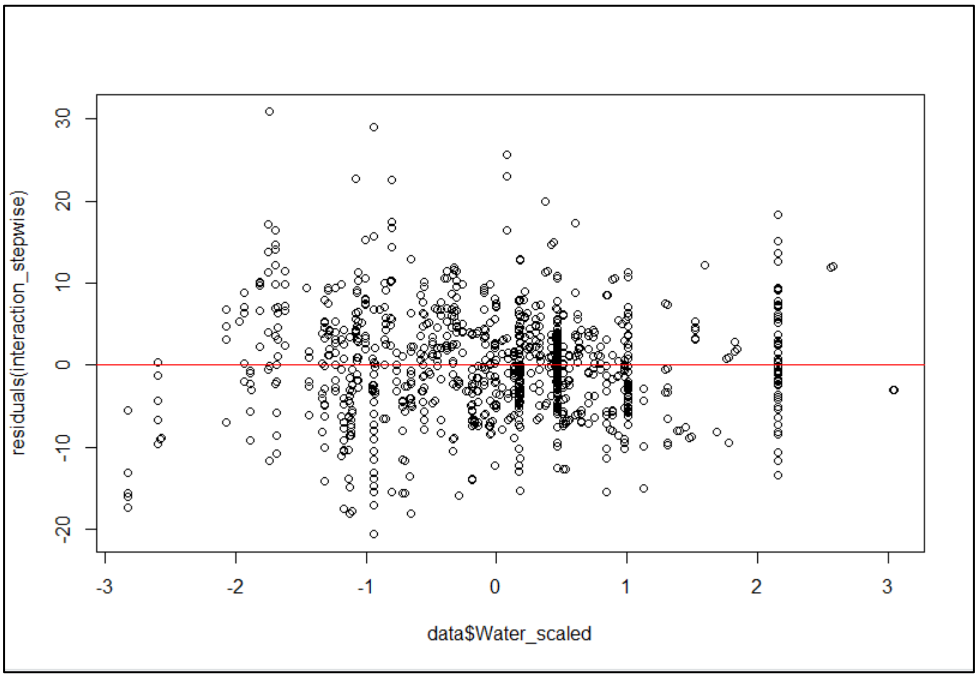
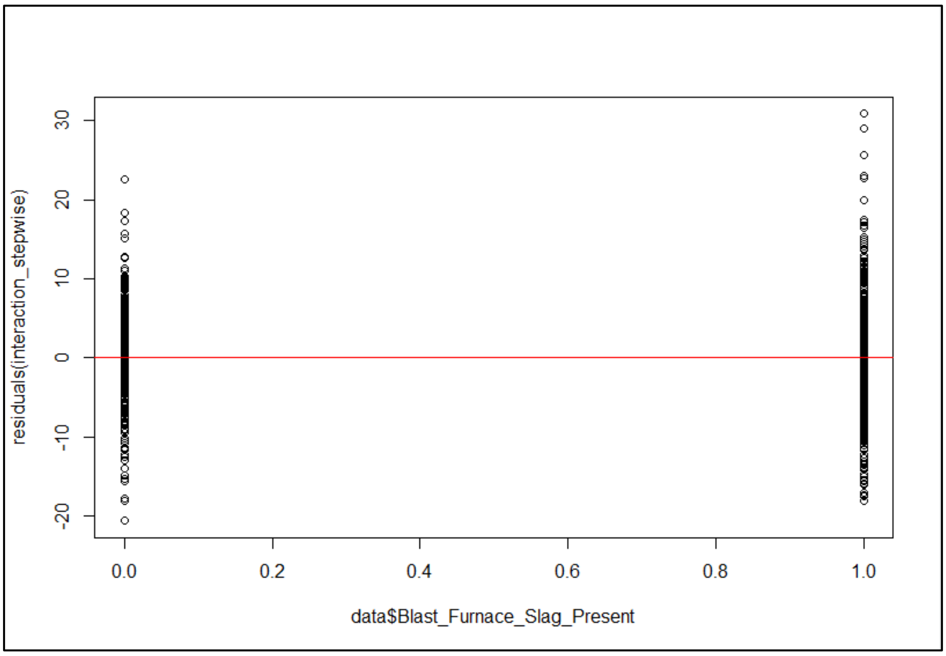


**Figure 37**

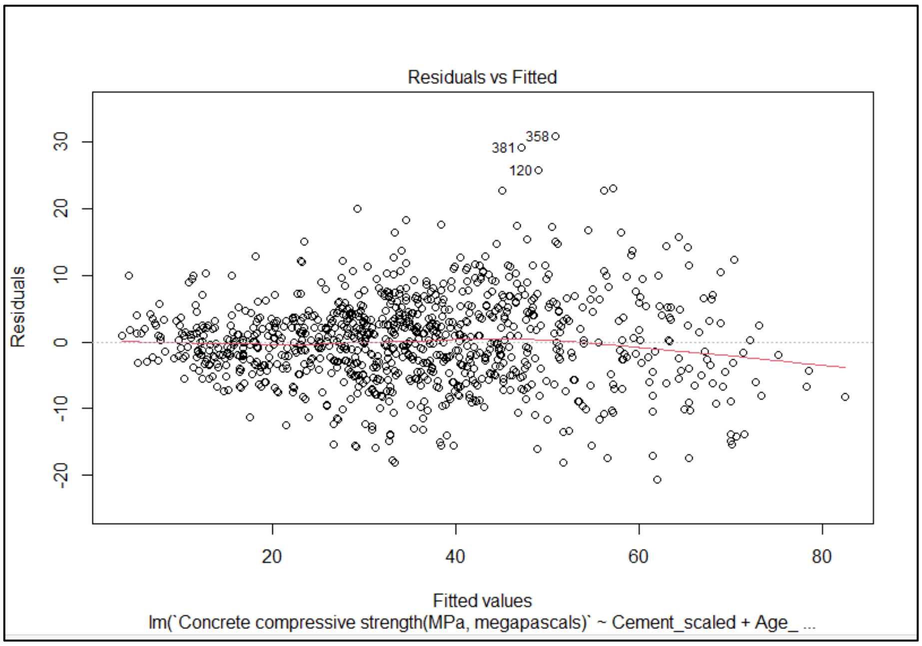
*Residuals vs. Blast Furnace Slag Presence Scatter Plot*




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).

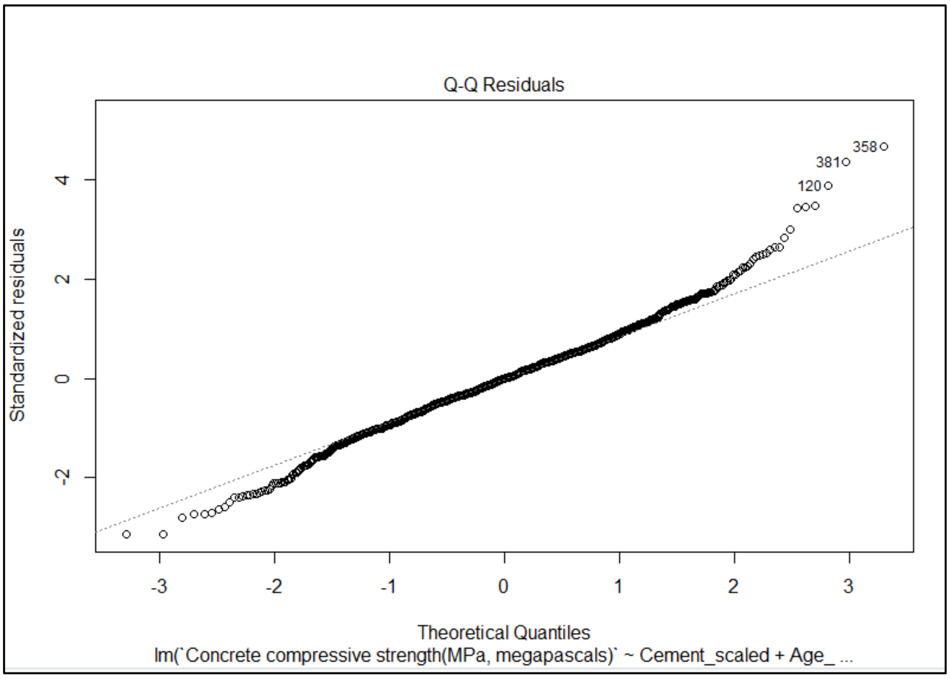
**Figure 38**

*Residuals vs. Fitted Values for Improved Model*



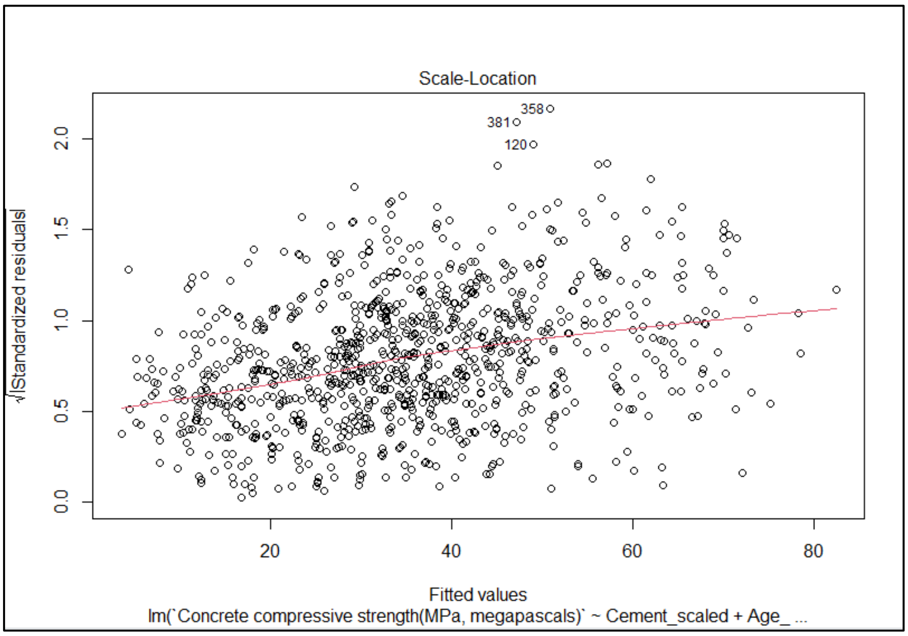

A Q-Q plot of the residuals was generated to ensure they follow a normal distribution. The showed points aligned closely with the diagonal reference line, except for minor deviations at the extremes, consistent with approximately normally distributed residuals (Figure 39)

**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)

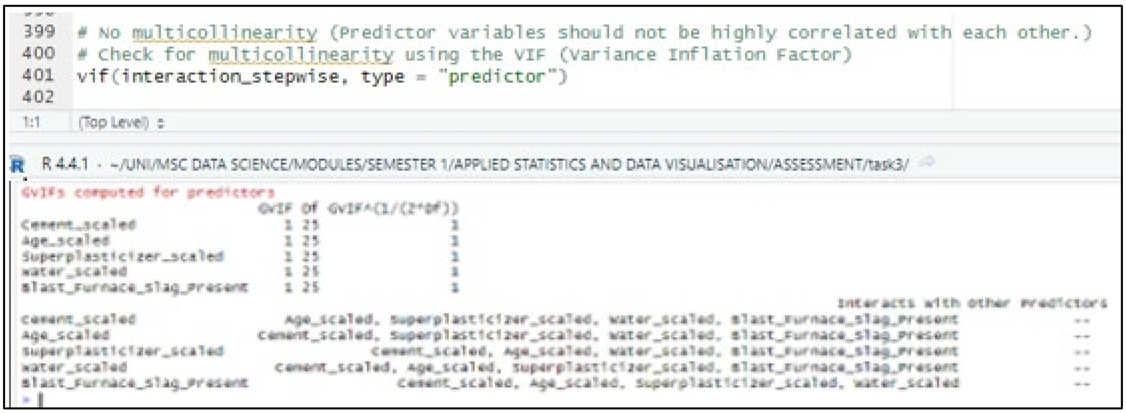
**Figure 40**
*Scale-Location Plot for Improved Model Residuals*



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*



```
# No multicollinearity (Predictor variables should not be highly correlated with each other.)
# Check for multicollinearity using the VIF (Variance Inflation Factor)
vif(interaction_stepwise, type = "predictor")
```

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
Age_scaled                     1 25               1
Superplasticizer_scaled        1 25               1
Water_scaled                   1 25               1
Blast_Furnace_Slag_Present     1 25               1
                                                                                         Interacts With Other Predictors
Cement_scaled                     Age_scaled, Superplasticizer_scaled, Water_scaled, Blast_Furnace_Slag_Present          --
Age_scaled                     Cement_scaled, Superplasticizer_scaled, Water_scaled, Blast_Furnace_Slag_Present          --
Superplasticizer_scaled                 Cement_scaled, Age_scaled, Water_scaled, Blast_Furnace_Slag_Present          --
Water_scaled                   Cement_scaled, Age_scaled, Superplasticizer_scaled, Blast_Furnace_Slag_Present          --
Blast_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)*

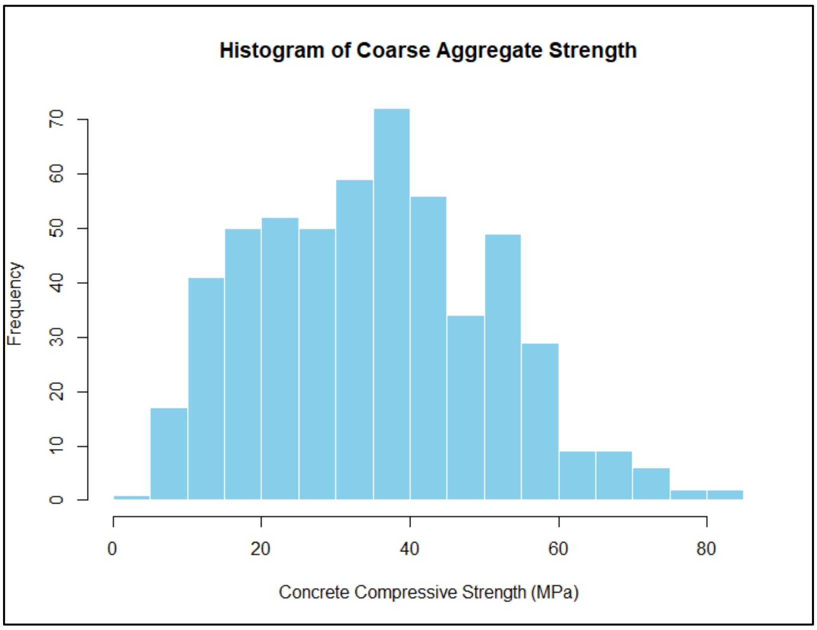



**Figure 44**

*Histogram of Compressive Strength (Fine Aggregate)*

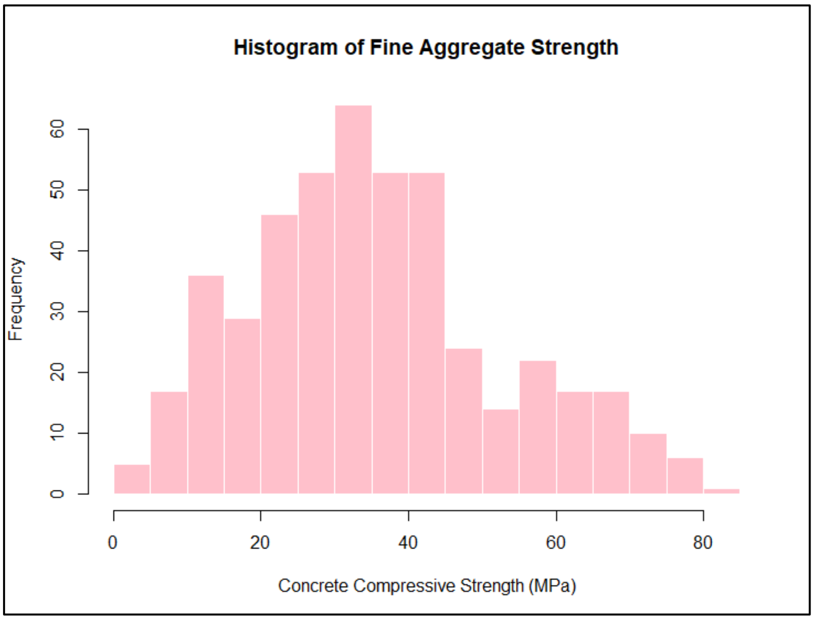


**Figure 45**

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

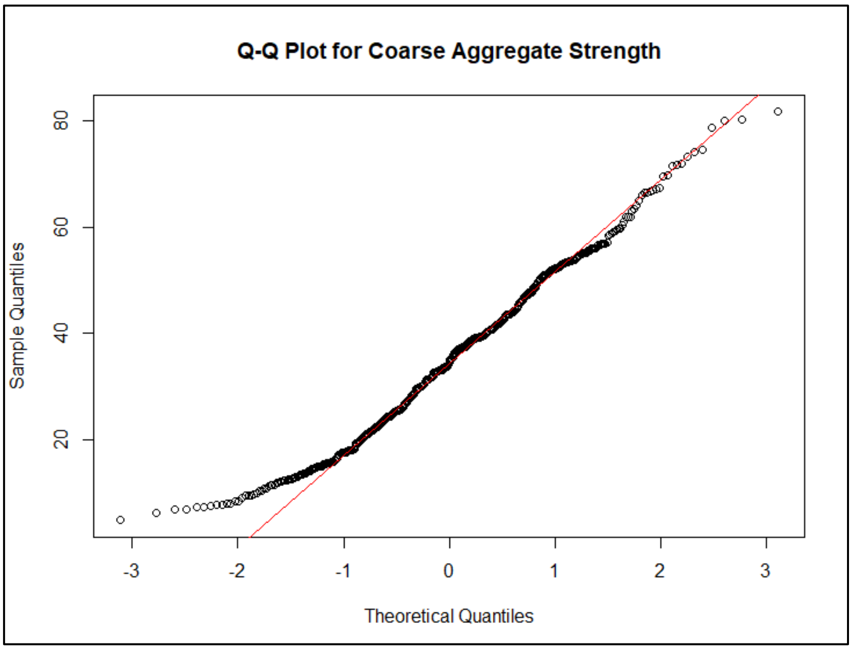



**Figure 46**

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

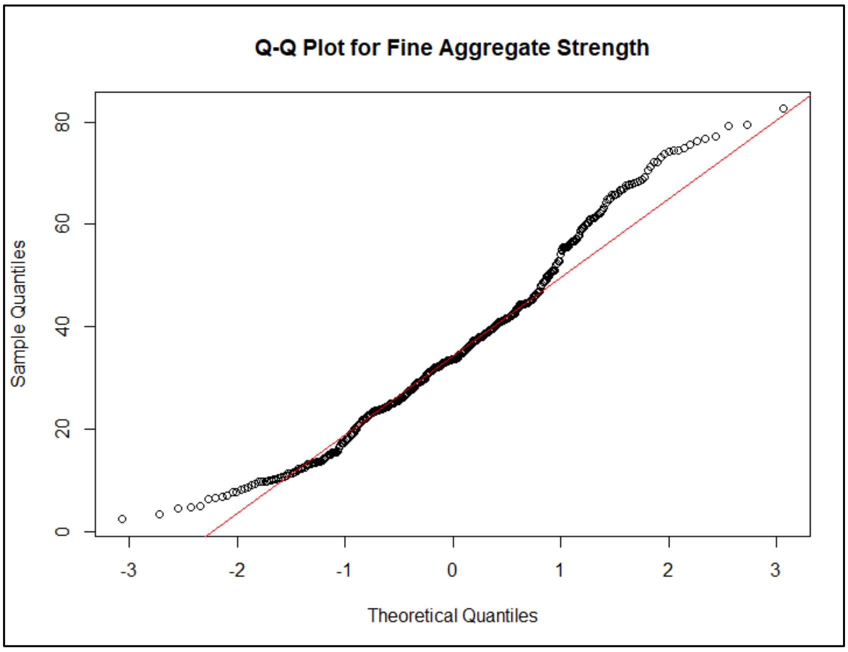


**Figure 47**

*Boxplot Comparison of Compressive Strength by Aggregate Type*

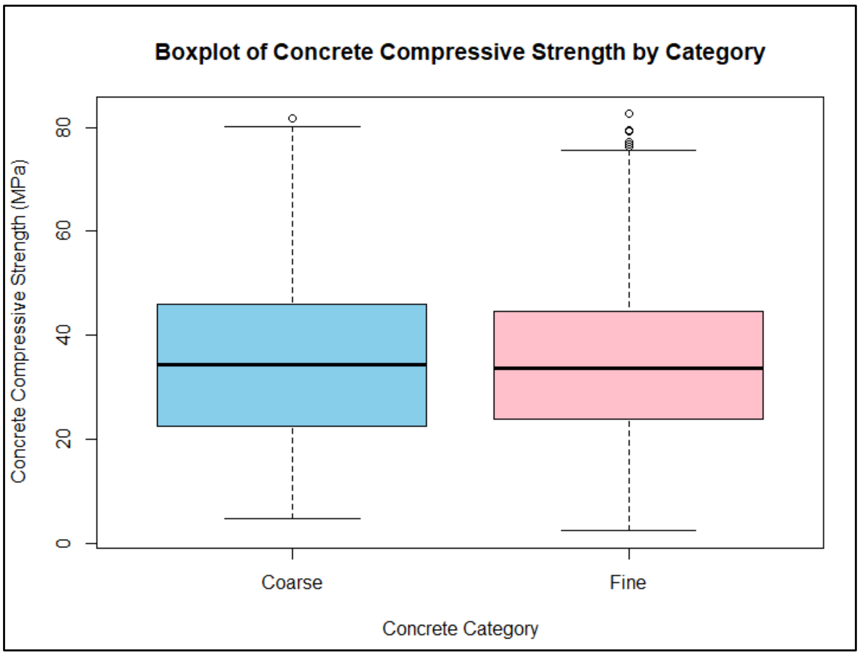
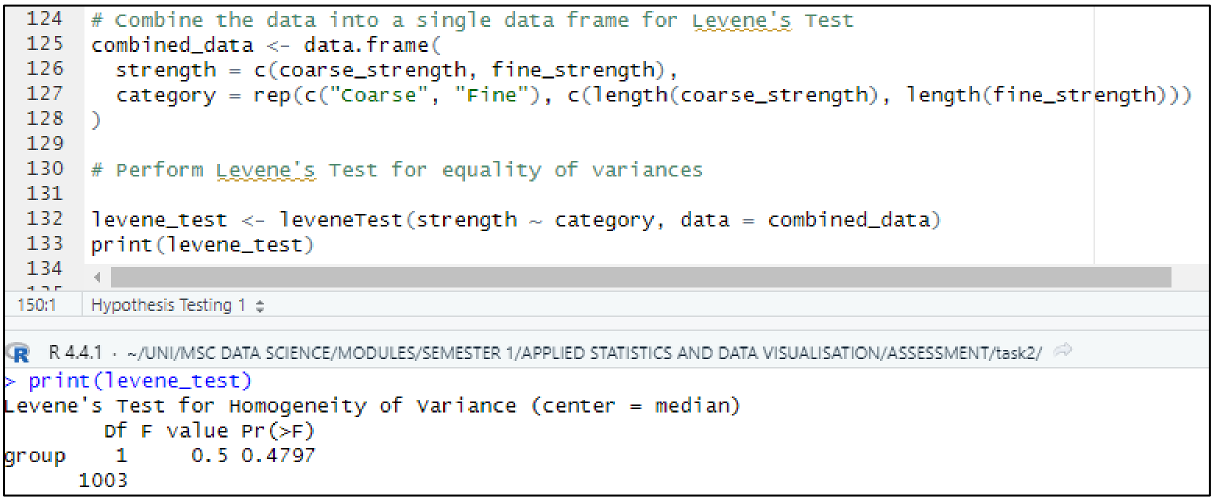



**Figure 48**

*Levene's Test for Equality of Variances*



```
# Combine the data into a single data frame for Levene's Test
combined_data <- data.frame(
  strength = c(coarse_strength, fine_strength),
  category = rep(c("Coarse", "Fine"), c(length(coarse_strength), length(fine_strength)))
)

# Perform Levene's Test for equality of variances

levene_test <- leveneTest(strength ~ category, data = combined_data)
print(levene_test)

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*

```r
# Check the result of Levene's Test
if (levene_test$`Pr(>F)`[1] > 0.05) {
  cat("Levene's Test p-value:", levene_test$`Pr(>F)`[1], "\n")
  cat("Variances are equal. Performing standard t-test...\n")

  # Perform standard t-test (equal variances)
  t_test <- t.test(coarse_strength, fine_strength, alternative = "greater", var.equal = TRUE)

} else {
  cat("Levene's Test p-value:", levene_test$`Pr(>F)`[1], "\n")
  cat("Variances are not equal. Welch's t-test is more appropriate...\n")

  # Perform t-test with unequal variances (Welch's t-test)
  t_test <- t.test(coarse_strength, fine_strength, alternative = "greater", var.equal = FALSE)
}

# Display the results of the t-test
print(t_test)
```

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).

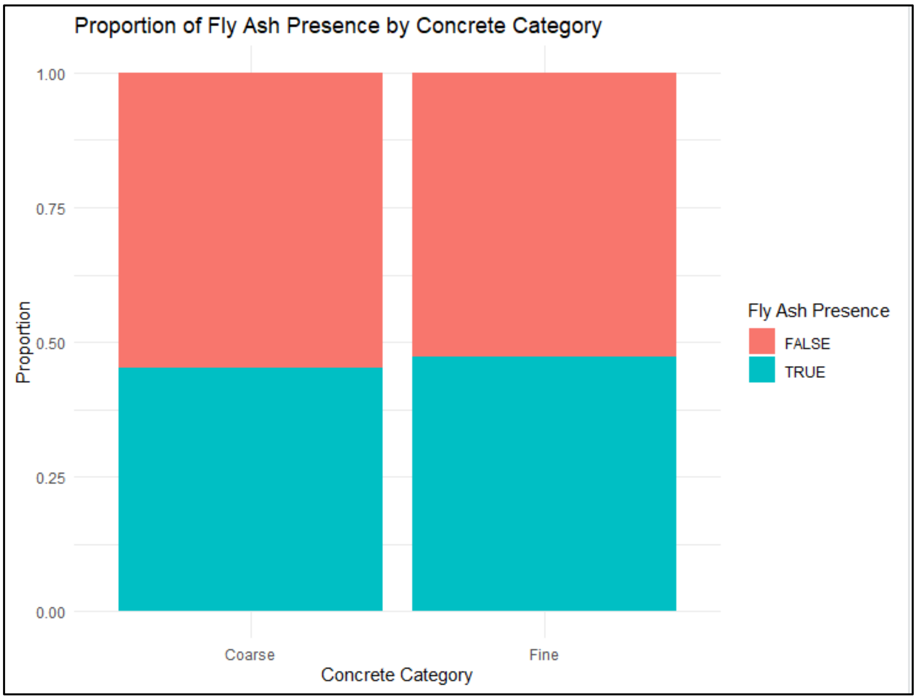
**Figure 50**

*Contingency Table for Fly Ash and Concrete Category*

```
# Create a contingency table of the two categorical variables
contingency_table <- table(data$`Concrete Category`, data$`Contains Fly Ash`)

# Display the contingency table
print(contingency_table)

173:45   Hypothesis Testing 2

R 4.4.1 · ~/UNI/MSC DATA SCIENCE/MODULES/SEMESTER 1/APPLIED STATISTICS AND DATA VISUALISATION/ASSESSMENT/task2/
         FALSE  TRUE
  Coarse   295   243
  Fine     246   221
>
> # Perform a chi-squared test of independence
> chi_test <- chisq.test(contingency_table)
>
> # Display the results of the chi-squared test
> print(chi_test)

        Pearson's Chi-squared test with Yates' continuity correction
```

**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*

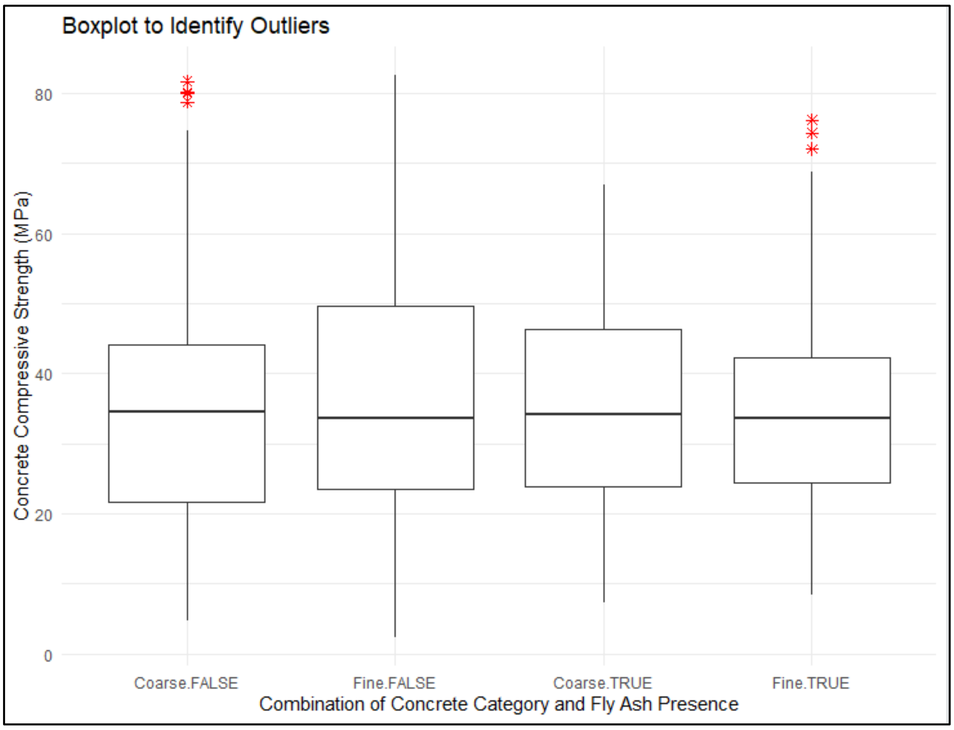
```
data$`Concrete Category` <- as.factor(data$`Concrete Category`)
data$`Contains Fly Ash` <- as.factor(data$`Contains Fly Ash`)


# create a two-way ANOVA model
model <- aov(`Concrete compressive strength(MPa, megapascals)` ~
               `Concrete Category` * `Contains Fly Ash`, data = data)
```

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*

```
residuals <- residuals(model)
shapiro_test <- shapiro.test(residuals)
print(shapiro_test)
```

```
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*

```
levene_test <- leveneTest(`Concrete compressive strength(MPa, megapascals)` ~
                          `Concrete Category` * `Contains Fly Ash`, data = data)
print(levene_test)
```

```
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**

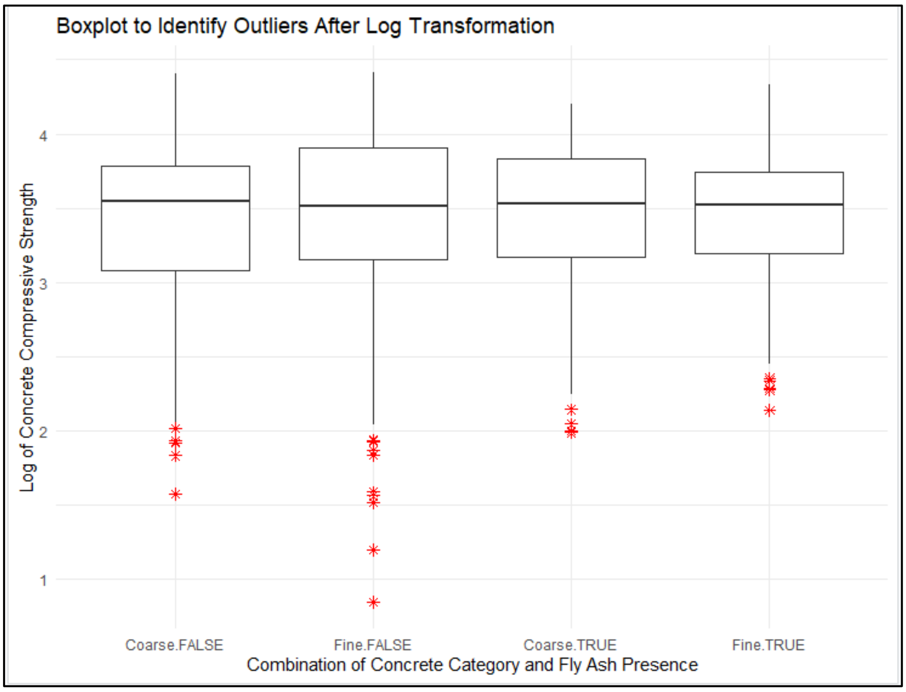
*Log-Transformed ANOVA Model Definition*

```r
data$log_strength <- log(data$`Concrete compressive strength(MPa, megapascals)`)

# Build a new model with log-transformed dependent variable
model_log <- aov(log_strength ~ `Concrete Category` * `Contains Fly Ash`, data = data)

```

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*

```
# Check for normality of residuals after log transformation
# Extract residuals
residuals_log <- residuals(model_log)
# Perform Shapiro-Wilk test
shapiro_test_log <- shapiro.test(residuals_log)
cat("Shapiro-Wilk Test for Normality (After Log Transformation):\n")
print(shapiro_test_log)

# 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")
print(levene_test_log)
```

```
316:1    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*

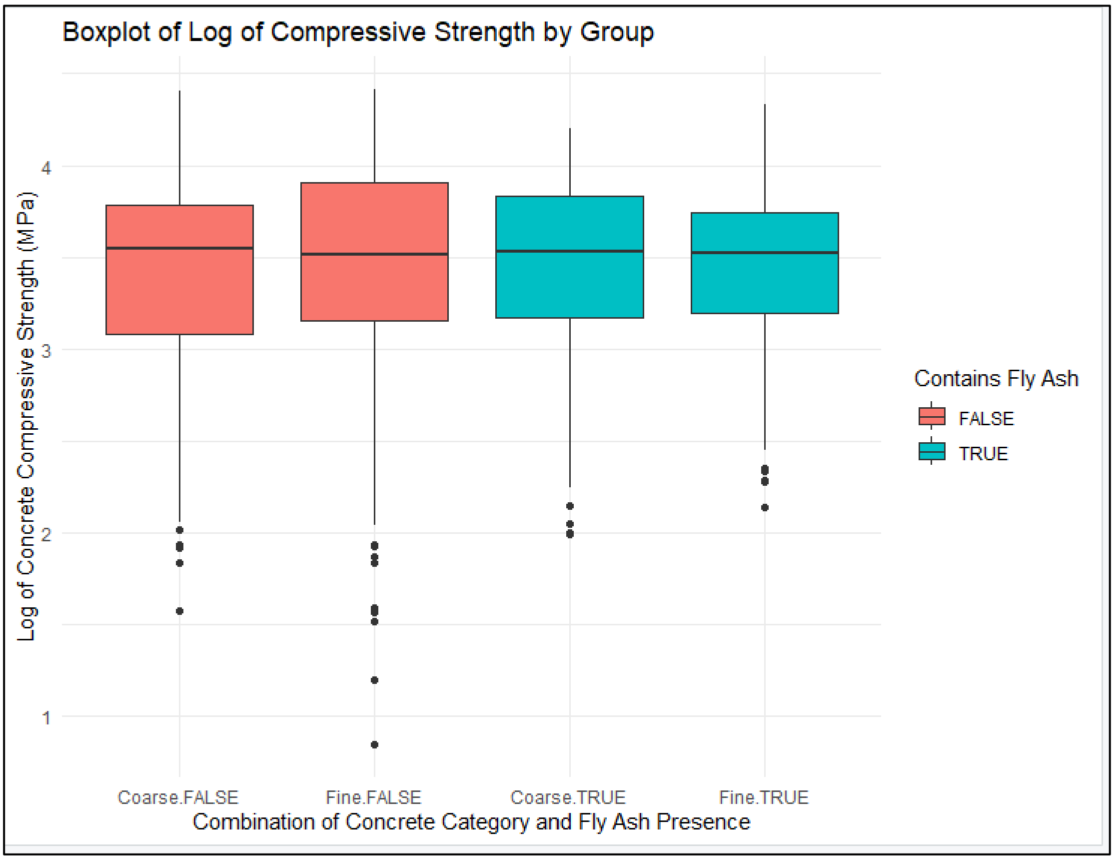
```
robust_anova <- t2way(log_strength ~ `Concrete Category` * `Contains Fly Ash`, data = data)
```

```
329:1    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.