# Concrete Compressive Strength Analysis Report

## 1. Background and Scenario Overview

The construction sector is one of the oldest and largest in the world and still expected to increase by 85% to $15.5 trillion by 2030. Concrete strength is a key source of uncertainty during construction. Concrete strength maturity models based on concrete temperature evolution are well-known, but they are not always accurate because the temperature is also affected by the concrete mixture and the weather (European Commission, 2019).

Structural concrete is typically defined depending on its ageing time and material composition (Eskandari-Naddaf, et al 2017). A robust, predictive model that can estimate compressive strength as a function of proportions of mixtures would be useful in enabling high-throughput mixture design and reducing the empirical, labour-intensive nature of "trial batching" approaches that are currently used in industry (young et al, 2019).

## 1.2 Objectives of the analysis

Our construction company is testing a range of concrete mixes and want to better understand how compressive strength relates to the composition of the concrete.

**Primary Objective**

* To analyze the relationships between various concrete mix components and compressive strength, thereby providing insights that can guide the optimization of concrete mixture designs to achieve desired strength levels.

**Specific Objectives**

* **Conduct Exploratory Data Analysis (EDA)**

Summarize data distributions, central tendencies, and variability of concrete mix components.

Identify patterns and ranges within the dataset to understand key characteristics.

* **Perform Correlation Analysis**

Determine which mix components (e.g., cement, aggregate, water) have the strongest relationships with compressive strength. Use correlation coefficients to pinpoint components most likely to influence concrete strength.

* **Develop and Apply Regression Models**

Create regression models to predict concrete compressive strength based on the mix components and curing time.

Validate and refine the models to improve prediction accuracy, addressing assumptions like linearity and normality as needed.

* **Conduct Hypothesis Testing**

Formulate and test hypotheses regarding the impact of specific mix components on compressive strength.

Use statistical testing to verify assumptions and assess the significance of each component’s influence on strength.

## 2. Data Description and Preparation

The dataset consists of several concrete mix components measured in kg per m³ mixture, along with the age of concrete and the compressive strength in megapascals (MPa). This section describes each variable in detail and the preparation steps applied to ensure accuracy in analysis.

Table 2.1

|  |  |  |
| --- | --- | --- |
| Data Column | Description | Type |
| Cement (component 1) | The amount of cement in kg per m³ mixture. | Continuous |
| Blast Furnace Slag (component 2) | The amount of blast furnace slag in kg per m³ mixture. | Continuous |
| Fly Ash (component 3) | The amount of fly ash in kg per m³ mixture. | Continuous |
| Water (component 4) | The amount of water in kg per m³ mixture. | Continuous |
| Superplasticizer (component 5) | The amount of superplasticizer in kg per m³ mixture. | Continuous |
| Coarse Aggregate (component 6) | The amount of coarse aggregate in kg per m³ mixture. | Continuous |
| Fine Aggregate (component 7) | The amount of fine aggregate in kg per m³ mixture. | Continuous |
| Age (day) | The age of the concrete in days, representing the curing time. | Discrete |
| Concrete Category | If the concrete is Coarse or Fine | Character |
| Contains Fly Ash | If it contains Fly Ash | Logical |
| Concrete compressive strength | The target variable, measured in MPa. | Continuous (Target) |

## 2.2.1 Data Cleaning and Preparation

To ensure the integrity and reliability of the dataset for analysis, several data cleaning and preparation steps were conducted:

* **Inspection and Standardization of Column Names**

The dataset was initially inspected for inconsistencies in column names, such as spaces and special characters, which can complicate referencing variables in code. All column names were standardized to a consistent format by replacing spaces with underscores, improving readability and ease of access within the R environment.

* **Missing Value Analysis**

A thorough check for missing values was performed across all variables to assess data completeness. There was no missing value in the dataset.

* **Data Type Checks**

Variables were checked to have their appropriate data types and to ensure they were processed correctly in subsequent analysis stages. Specifically, categorical variables were encoded as factors to reflect their non-numeric, categorical nature. It ensures they are treated as group identifiers rather than numeric values, allowing for meaningful interpretations in regression and visualization.

* **Outlier Detection and Handling**

Outliers detection was performed on continuous variables using boxplots and statistical thresholds. Identified outliers were evaluated. This approach minimized their impact on the analysis while retaining genuine variations in the data.

* **Renaming the columns**

In the dataset, column names were carefully reviewed and adjusted to enhance clarity and ensure consistency when referencing variables in the analysis code. For instance, the column originally named "Cement (component 1)(kg in a m³ mixture)" was renamed to "Cement ". This change improved readability and ease of reference, preventing syntax issues that can arise with complex column names. This standardized naming allowed for more efficient data handling, facilitated automation in code, and reduced the likelihood of errors.

|  |  |
| --- | --- |
| Original Column Name | Renamed Column Name |
| Cement (component 1)(kg in a m³ mixture) | Cement |
| Blast Furnace Slag (component 2)(kg in a m³ mixture) | Blast Furnace Slag |
| Fly Ash (component 3)(kg in a m³ mixture) | Fly Ash |
| Water (component 4)(kg in a m³ mixture) | Water |
| Superplasticizer (component 5)(kg in a m³ mixture) | Superplasticizer |
| Coarse Aggregate (component 6)(kg in a m³ mixture) | Coarse Aggregate |
| Fine Aggregate (component 7)(kg in a m³ mixture) | Fine Aggregate |
| Concrete compressive strength(MPa, megapascals) | Concrete compressive strength |

Through these data cleaning and preparation steps, the dataset was refined to a quality suitable for reliable analysis, minimizing biases and ensuring that results reflect genuine patterns and relationships in the data.

## 3. Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution, central tendency, and variability of the variables in the dataset. Key descriptive statistics such as mean, median, standard deviation, and range for each variable are presented in this section.

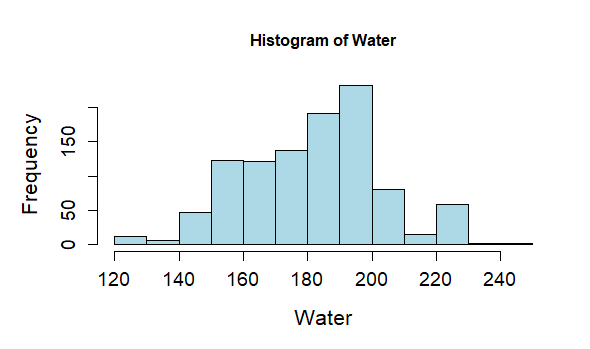
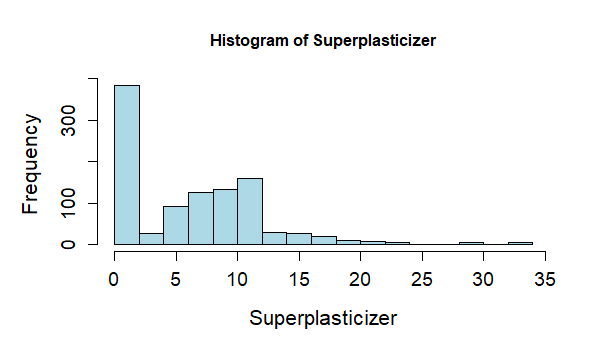
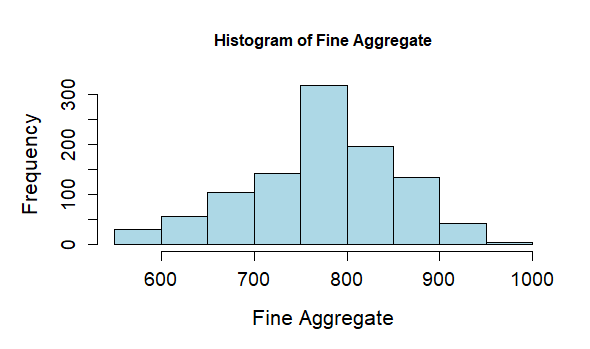
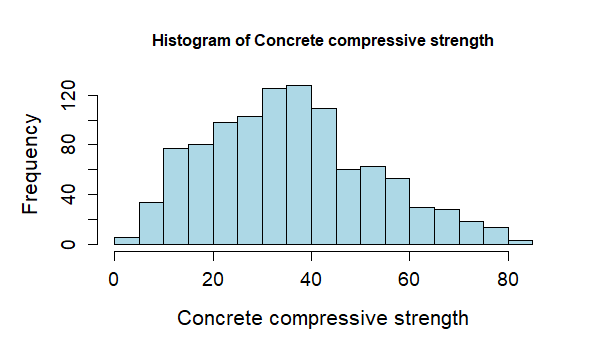
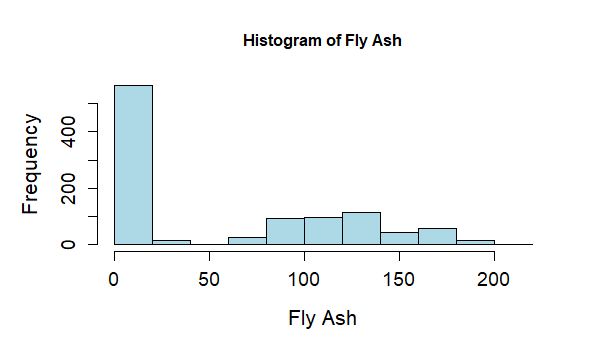
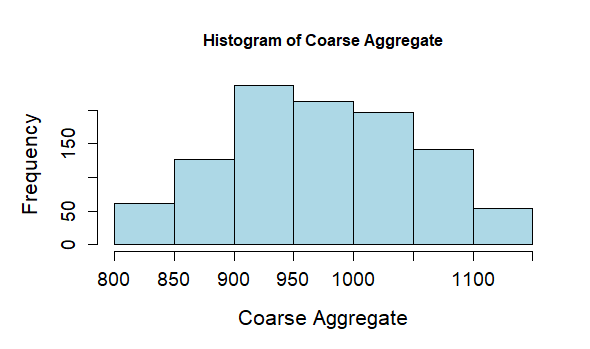
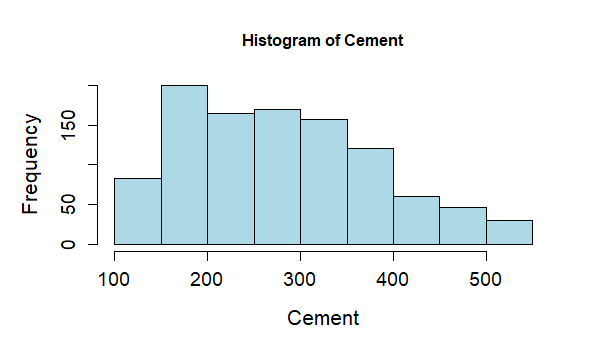
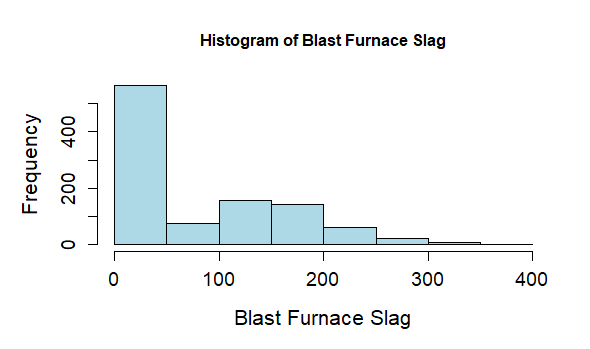
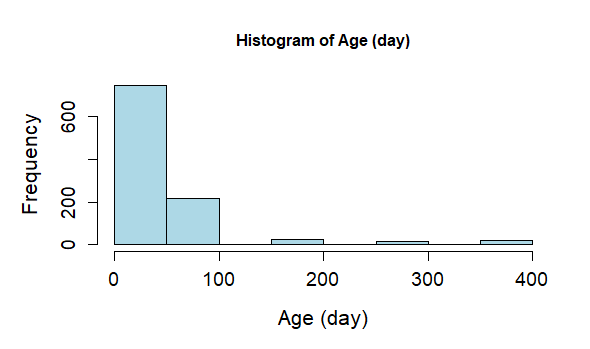
* **Descriptive Statistics**

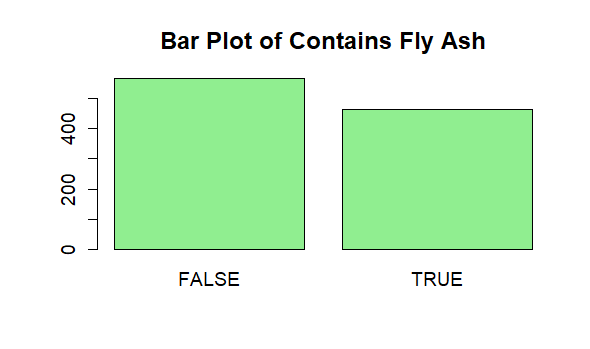
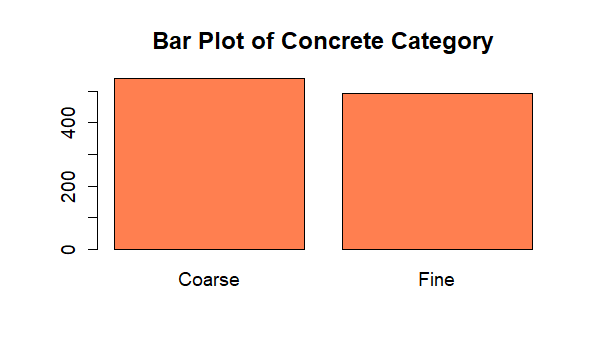
Summary statistics (such as mean, median, standard deviation, minimum, and maximum) were calculated for each variable to provide a snapshot of the data’s central tendencies and variability.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Vars | N | Mean | SD | Median | Trimmed | MAD | Min | Max | Range | Skew | Kurtosis | SE |
| Cement | 1 | 1030 | 281.17 | 104.51 | 272.90 | 273.47 | 117.72 | 102.00 | 540.0 | 438.00 | 0.51 | -0.53 | 3.26 |
| Blast Furnace Slag | 2 | 1030 | 73.90 | 86.28 | 22.00 | 62.43 | 32.62 | 0.00 | 359.4 | 359.40 | 0.80 | -0.52 | 2.69 |
| Fly Ash | 3 | 1030 | 54.19 | 64.00 | 0.00 | 46.85 | 0.00 | 0.00 | 200.1 | 200.10 | 0.54 | -1.33 | 1.99 |
| Water) | 4 | 1030 | 181.57 | 21.36 | 185.00 | 181.19 | 19.27 | 121.75 | 247.0 | 125.25 | 0.07 | 0.11 | 0.67 |
| Superplasticizer | 5 | 1030 | 6.20 | 5.97 | 6.35 | 5.56 | 7.87 | 0.00 | 32.2 | 32.20 | 0.91 | 1.39 | 0.19 |
| Coarse Aggregate | 6 | 1030 | 972.92 | 77.75 | 968.00 | 973.49 | 68.64 | 801.00 | 1145.0 | 344.00 | -0.04 | -0.61 | 2.42 |
| Fine Aggregate | 7 | 1030 | 773.58 | 80.18 | 779.51 | 776.41 | 67.44 | 594.00 | 992.6 | 398.60 | -0.25 | -0.11 | 2.50 |
| Age (day) | 8 | 1030 | 45.66 | 63.17 | 28.00 | 32.53 | 31.13 | 1.00 | 365.0 | 364.00 | 3.26 | 12.07 | 1.97 |
| Concrete Category | 9 | 1030 | 1.48 | 0.50 | 1.00 | 1.47 | 0.00 | 1.00 | 2.0 | 1.00 | 0.09 | -1.99 | 0.02 |
| Contains Fly Ash | 10 | 1030 | NaN | NA | NA | NaN | NA | Inf | -Inf | -Inf | NA | NA | NA |
| Concrete compressive strength | 11 | 1030 | 35.82 | 16.71 | 34.44 | 34.96 | 16.20 | 2.33 | 82.6 | 80.27 | 0.42 | -0.32 | 0.52 |

* **Distribution Analysis**

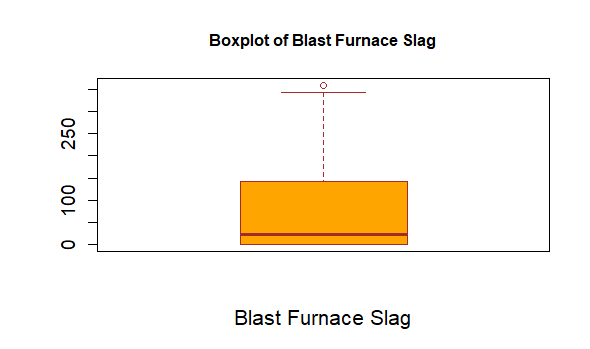
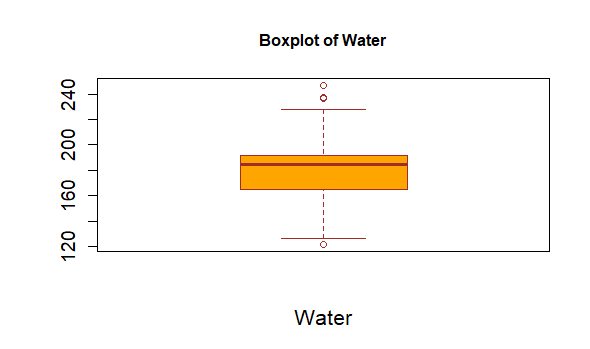
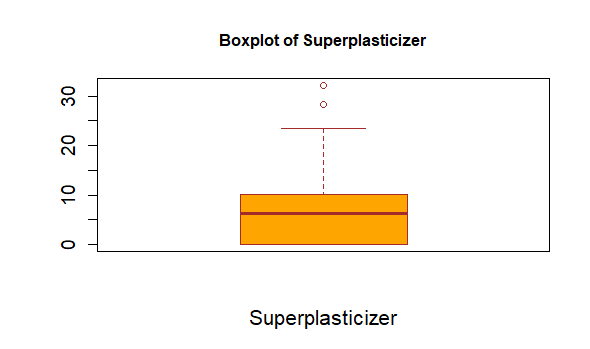
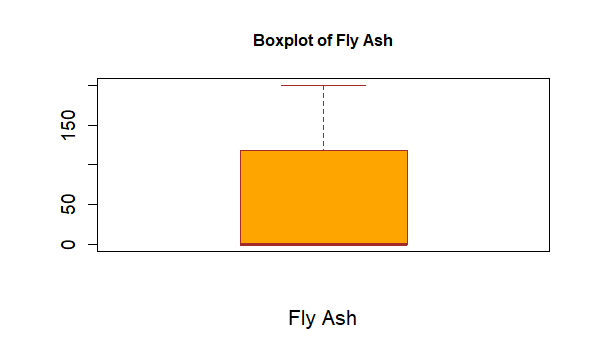
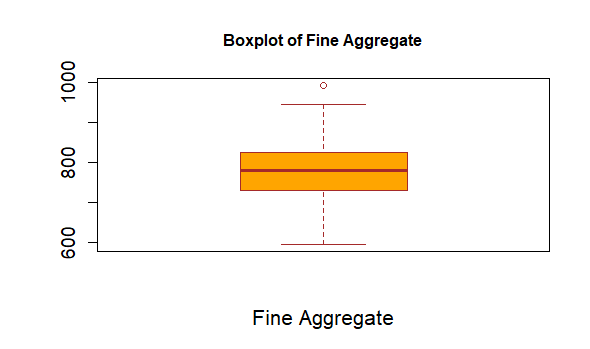
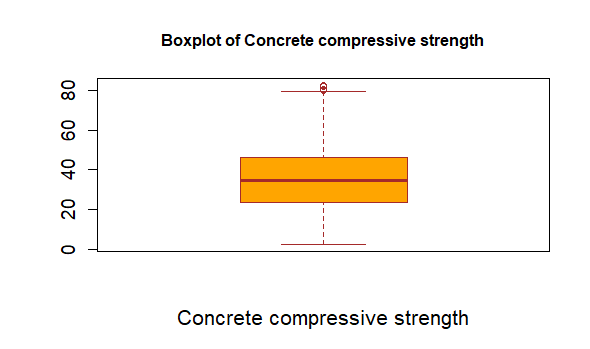
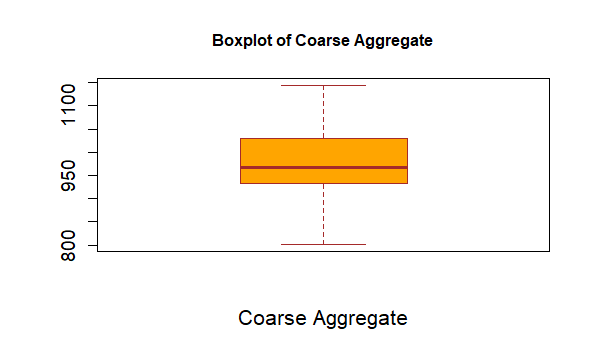
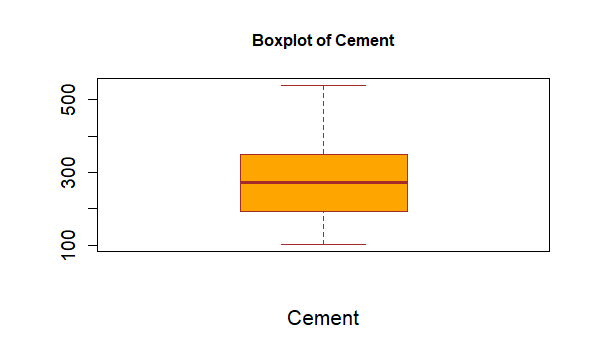
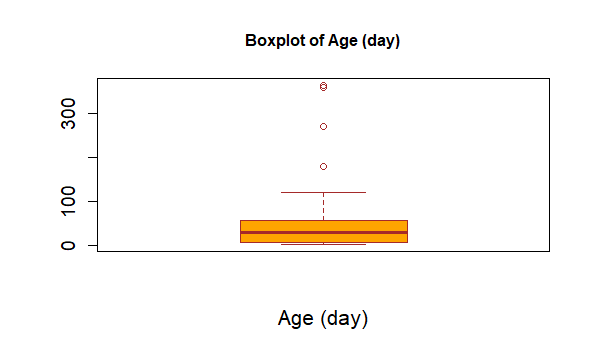
The distribution of each variable was analyzed using histograms and bar plot to assess whether they were normally distributed, skewed, or had any unusual patterns. This understanding is crucial for selecting appropriate modeling techniques, as many statistical methods assume normally distributed data. For instance, cement, water, and other component variables were plotted to observe any skewness or peaks, which may suggest specific patterns in the concrete mix proportions.



****

* **Outlier Detection**

Outliers were identified through boxplots, which allowed for the visualization of the spread of data and identification of any extreme values.

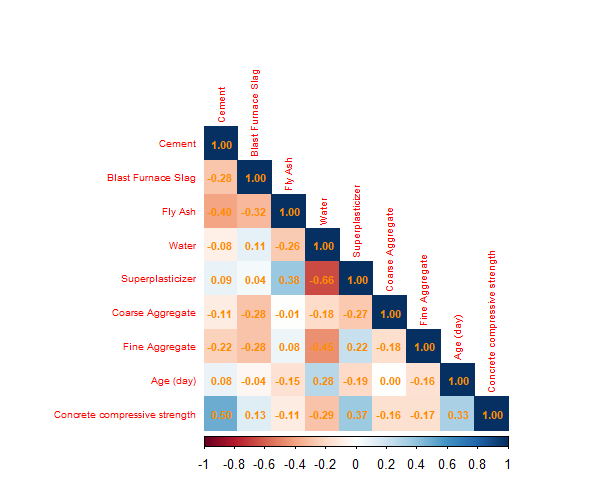


* **Correlation Analysis**

A correlation matrix was created to examine the relationships between the numerical variables and concrete compressive strength. This analysis helped in identifying which variables are likely to influence compressive strength significantly. High correlations between variables suggest possible dependencies, which was useful in feature selection for regression modeling.

* **Visual Exploration of Relationships**

In visualizing the relationships, Heat map was generated to visualize relationships between the component quantities and the compressive strength. This visual exploration aided in spotting trends, such as a positive relationship between cement content and compressive strength, or the influence of water content on the strength of the concrete.



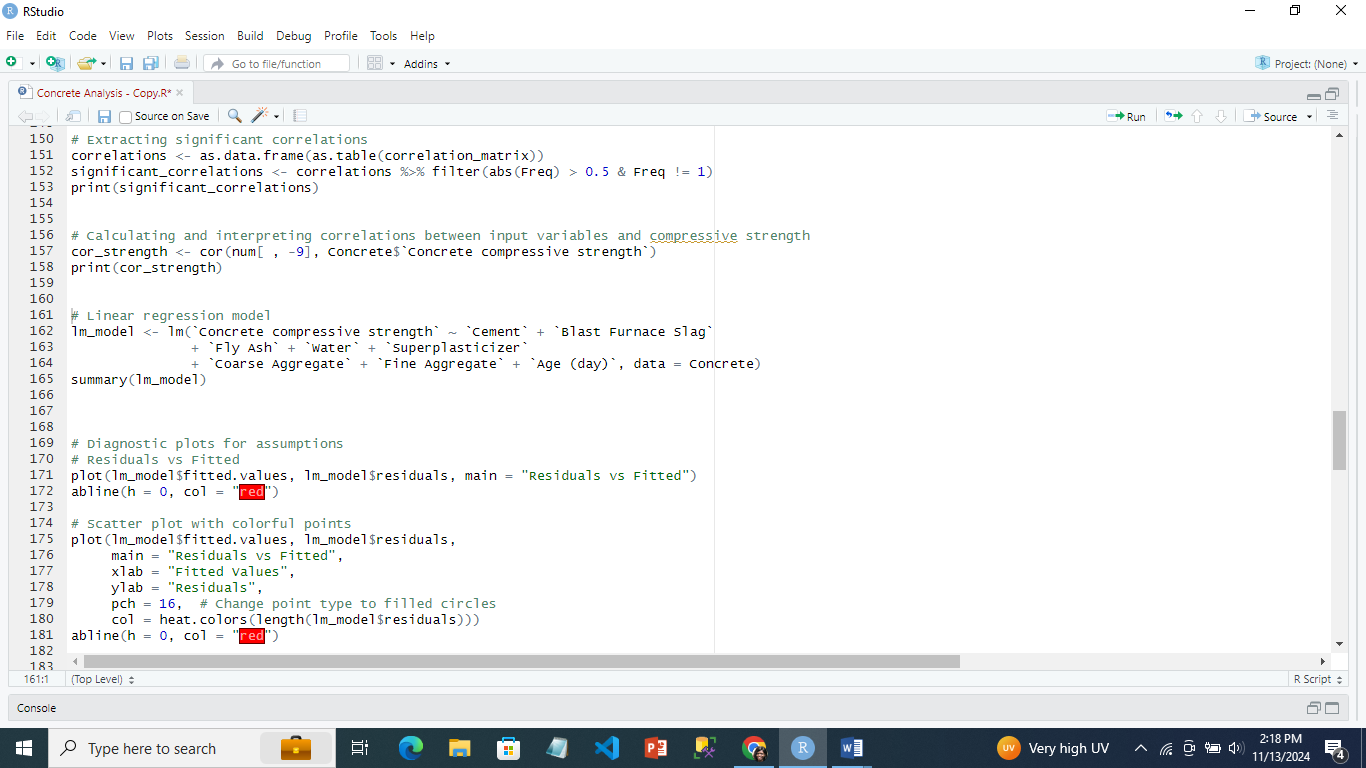
Through these EDA steps, a comprehensive understanding of the dataset was developed, guiding the subsequent modeling and hypothesis testing. EDA revealed important trends and patterns within the data, informing decisions on data transformations, feature selection, and potential variable interactions to explore further in the analysis.

## 4. Correlation Analysis

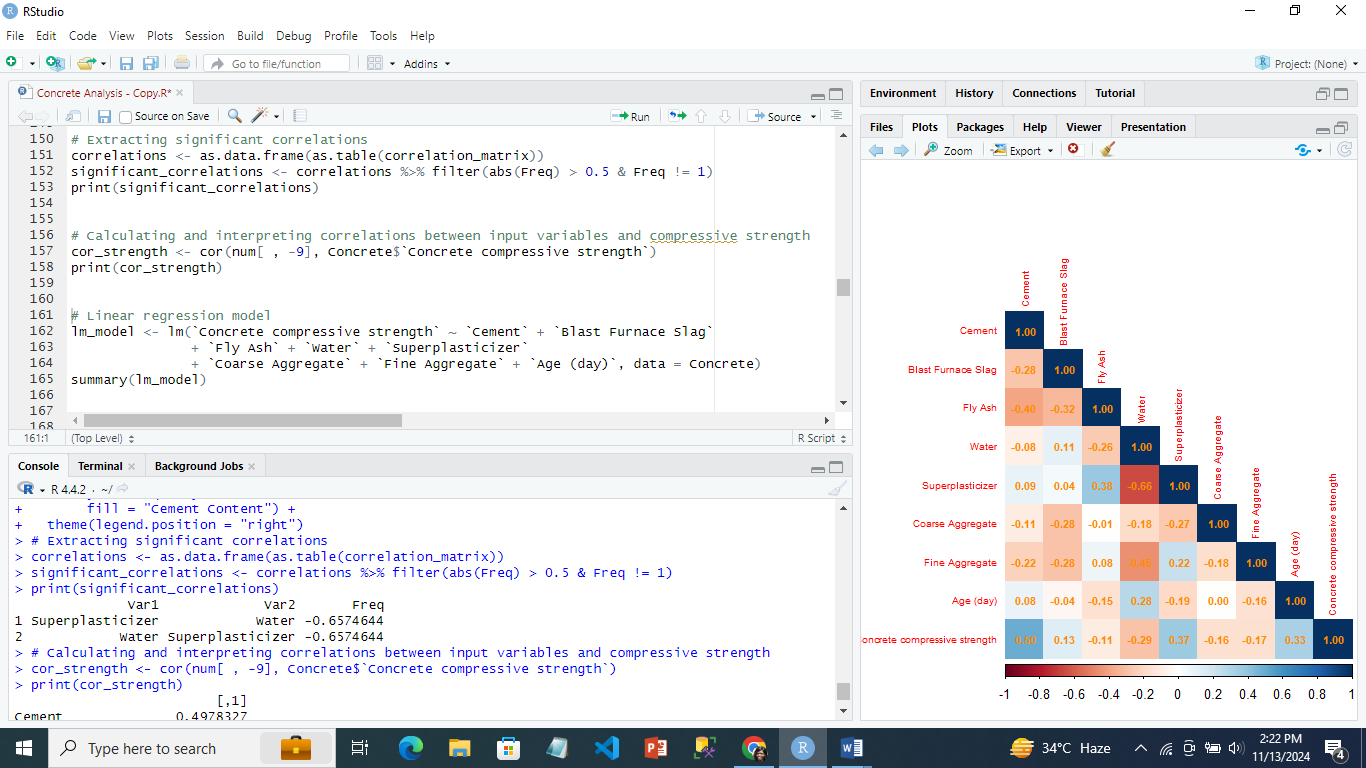
### 4.1 Extracting Significant correlations

The first thing that was done was the extraction and identification of significant correlations from the correlation matrix

*The code:*



*The output:*

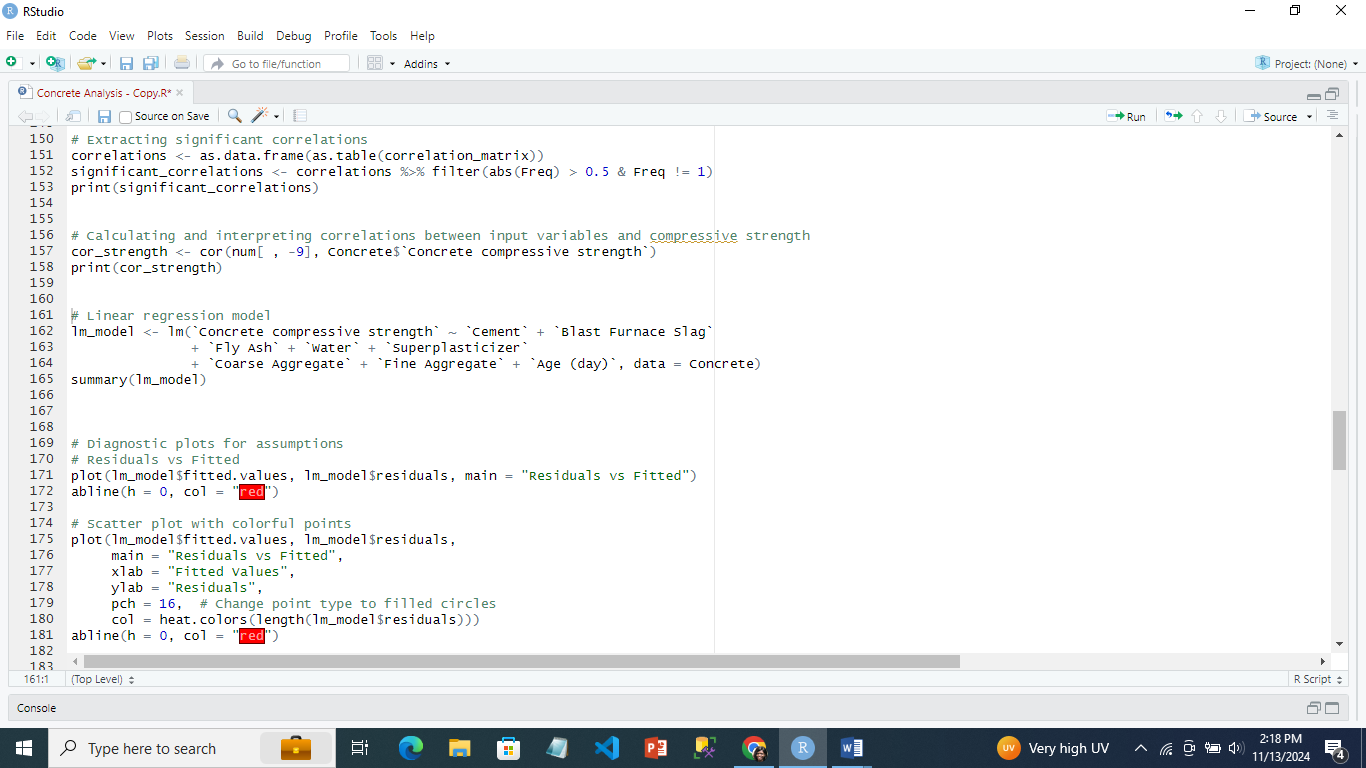


The analysis extracted significant correlations from a correlation matrix by filtering for correlations with an absolute value greater than 0.5 and excluding perfect self-correlations. The key finding is a strong negative correlation of -0.657 between **Superplasticizer** and **Water**, indicating that as the amount of superplasticizer increases, the amount of water tends to decrease, and vice versa.

### 4.2 Pearson Correlation

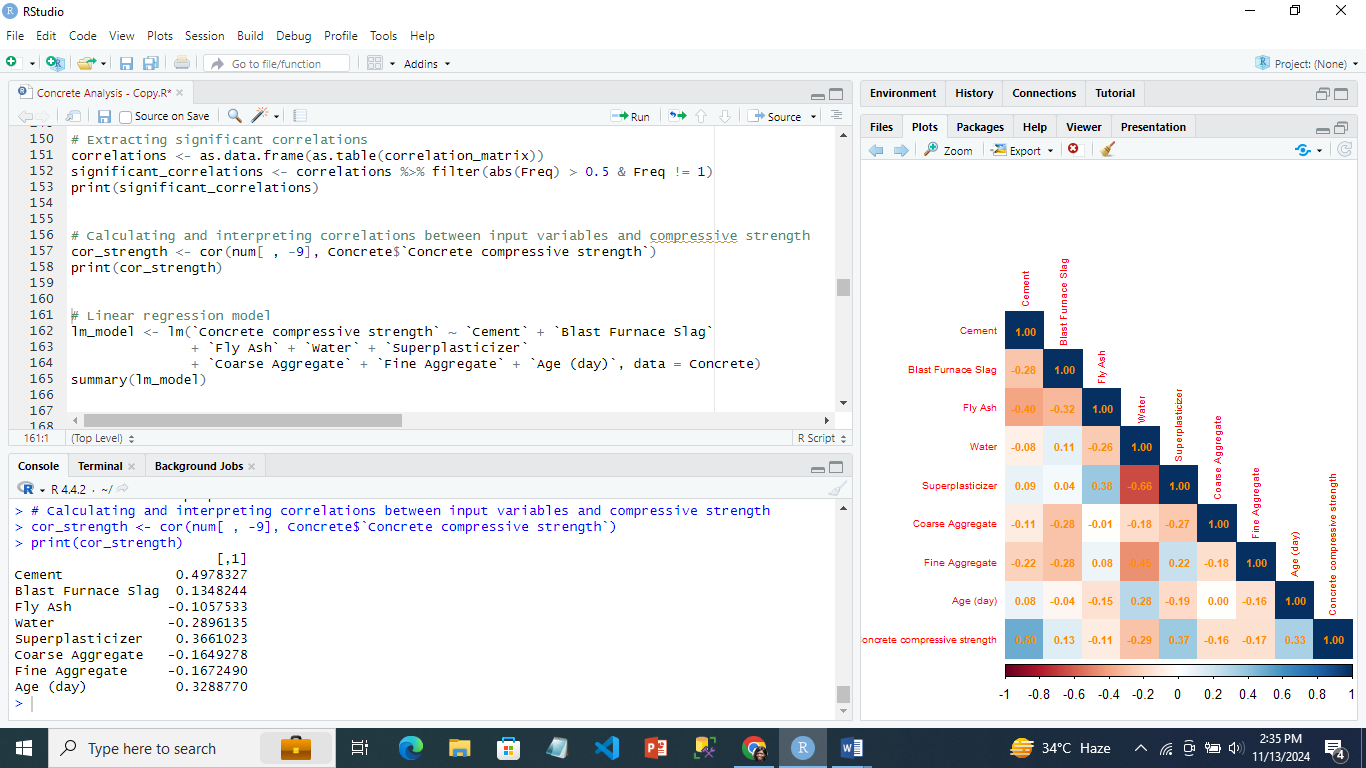
A Pearson correlation analysis was performed to identify the relationships between the mix components and concrete compressive strength.

*The code:*



the correlation coefficients between the input variables and **Concrete Compressive Strength** were calculated.

*The output:*



This shows that:

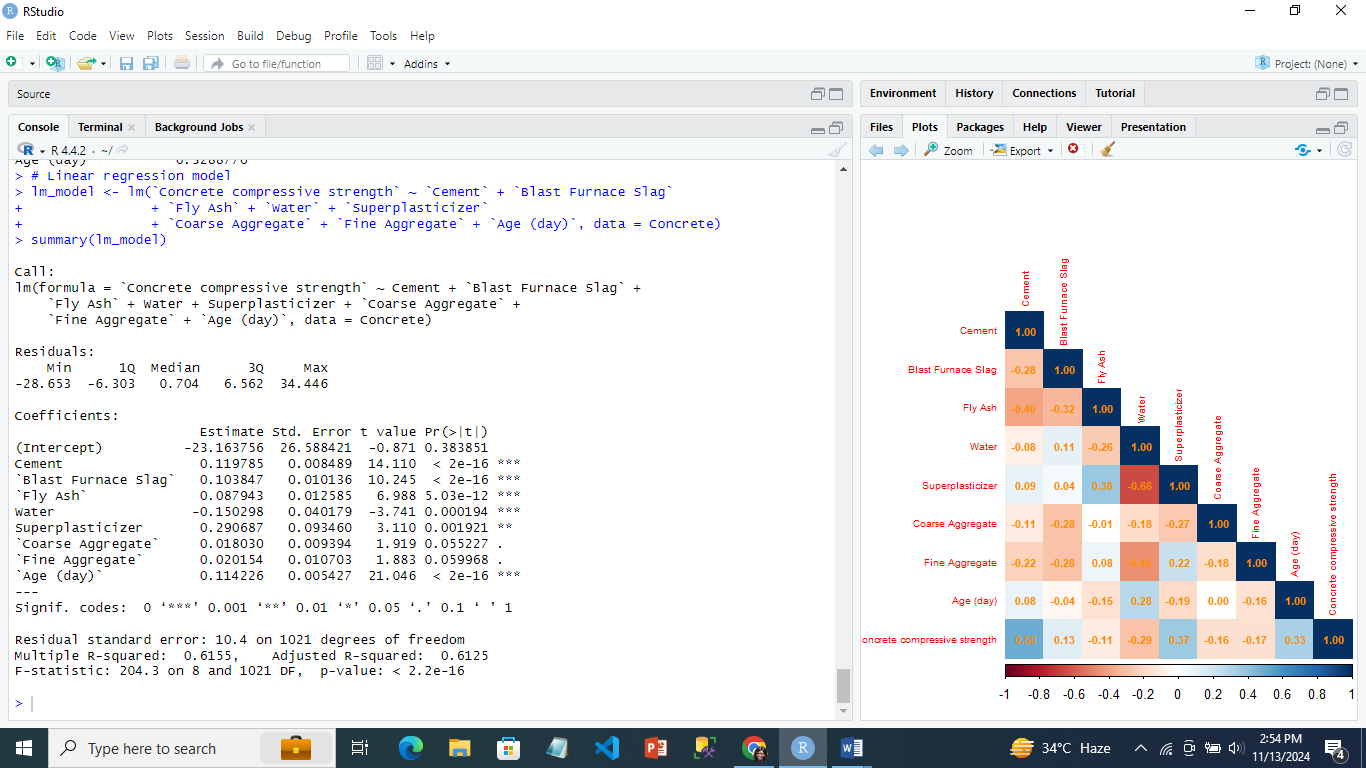
* Compressive strength and cement had the greatest positive correlation (0.498). This implies that the concrete's compressive strength tends to rise in tandem with an increase in cement content.
* Slag from blast furnaces (0.135): A marginally favourable association. Although there is a little positive correlation between compressive strength and slag concentration, it is not significant enough to be a contributing factor.
* Fly Ash: A weakly negative correlation (-0.106). This suggests that the fly ash content and compressive strength have a minor negative relationship, suggesting that adding more fly ash may marginally lower the compressive strength.
* Water: A moderately negative correlation (-0.290). Compressive strength decreases with increasing water content, which is consistent with the general rule that too much water dilutes the combination and weakens it.
* Superplasticizer: A relatively favourable correlation (0.366). This implies that adding more superplasticizer improves compressive strength, most likely because it makes workability better while keeping water content low.
* There are weak negative associations between coarse aggregate (-0.165) and fine aggregate (-0.167). Although there is a modest inverse link between these factors and compressive strength, it is not as significant as that of water or cement.
* There is a fairly favourable connection between age (day) (0.329). This illustrates the widely accepted idea that as concrete cures, its compressive strength rises.

Cement and superplasticizer have the largest positive relationships with compressive strength, but water has a considerable negative association. Coarse aggregate, fly ash, blast furnace slag, and fine aggregate are among the materials with lesser relationships. In order to achieve increased compressive strength, curing time is crucial, as seen by the moderately beneficial influence of age.

## 5. Regression Modeling

The linear regression model was developed for predicting concrete compressive strength using the input variables - cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age (day). Assumptions of linearity, normality, and multicollinearity were tested and addressed as needed.

*Regression code and output:*



*This shows that:*

* **Residuals:**

Min: -28.653

1st Quartile: -6.303

Median: 0.704

3rd Quartile: 6.562

Max: 34.446

Residuals indicate the difference between the observed and predicted values.

**Coefficients:** The coefficients denote the change in compressive strength for a unit change in each predictor variable:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Estimate | Std. Error | t Value | p-Value | Significance |
| (Intercept) | -23.164 | 26.588 | -0.871 | 0.384 |  |
| Cement | 0.120 | 0.00849 | 14.110 | < 2e-16 | \* |
| Blast Furnace Slag | 0.104 | 0.0101 | 10.245 | < 2e-16 | \* |
| Fly Ash | 0.088 | 0.0126 | 6.988 | 5.03e-12 | \* |
| Water | -0.150 | 0.0402 | -3.741 | 0.000194 |  |
| Superplasticizer | 0.291 | 0.0935 | 3.110 | 0.001921 |  |
| Coarse Aggregate | 0.018 | 0.00939 | 1.919 | 0.055227 | . |
| Fine Aggregate | 0.020 | 0.0107 | 1.883 | 0.059968 | . |
| Age (day) | 0.114 | 0.00543 | 21.046 | < 2e-16 | \* |

This shows that:

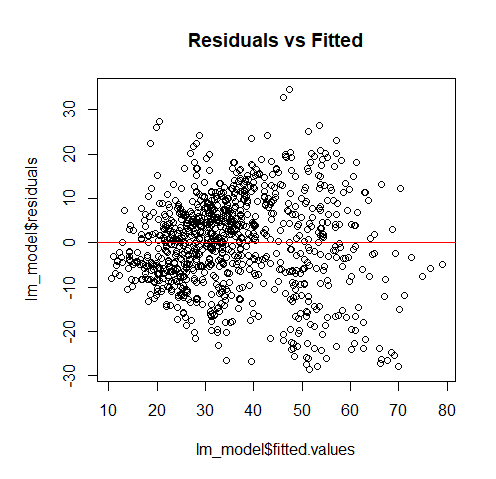
* The effects of cement, blast furnace slag, fly ash, water, superplasticizer, and age (day) on compressive strength are statistically significant (p-value 0.05), with cement and blast furnace slag having the most beneficial influence.
* Coarse and fine aggregates had lower significance (p-values near to 0.05), indicating a minor influence on compressive strength.
* Water has a negative correlation, which means that additional water reduces compressive strength.

## 5.1 Model performance:

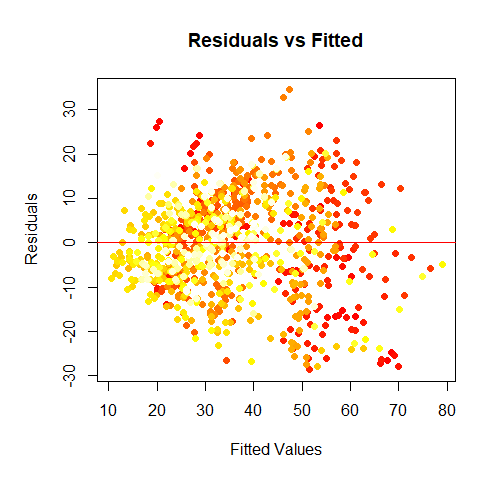
* The residual standard error is 10.4 on 1021 degrees of freedom. This shows the average difference between the observed and projected values.
* Multiple R-squared = 0.6155, indicating that the model explains about 61.55% of the variance in compressive strength.
* Adjusted R-squared: 0.6125, which modifies the R-squared value according to the number of predictors in the model.
* F-statistic: 204.3 on 8 and 1021 DF, p-value < 2.2e-16. This shows that the model as a whole is statistically significant.

The regression analysis shows that water has a detrimental influence, but cement and superplasticizer have significant beneficial effects on compressive strength. Coarse and fine aggregates, have smaller correlations.

## 5.2 Residual vs Fitted Visualizations

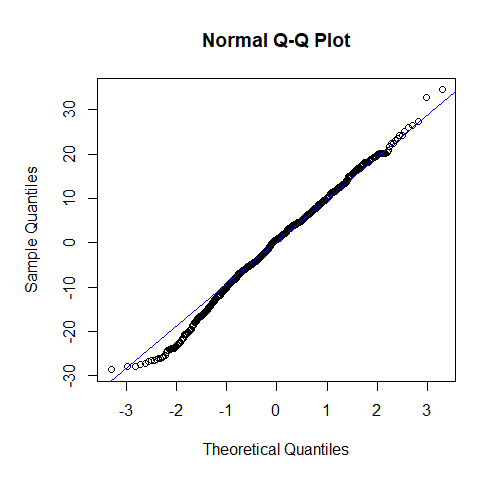
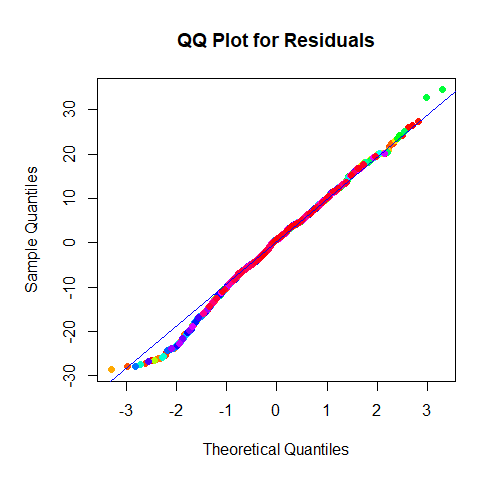
The residuals were plotted against the fitted values, with a red horizontal line at zero. This plot was used to check for homoscedasticity or the constant variance of residuals and linearity.

The scatter plot with colored points was created to provide a clear visualization of the residuals. The colors represent the intensity of the residuals, with the red line indicating the zero-residual line. This allows for a more detailed inspection of any patterns or outliers.



## 5.3 ****Quantile-Quantile for**** Normality of Residuals

The Quantile-Quantile (QQ) plot was used to evaluate whether the residuals of the linear regression model follow a normal distribution. The points on the QQ plot should lie along a straight line, indicating that the residuals follow a normal distribution. The colorful plot shows a better visualization for the residual distribution.

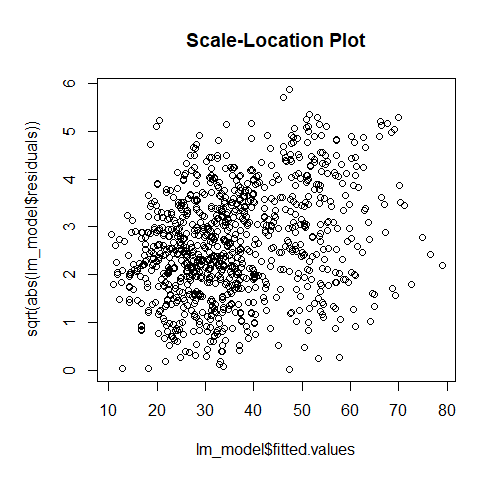


## 5.4 Homoscedasticity

## 5.4 ****Scale-Location Plot for Homoscedasticity****

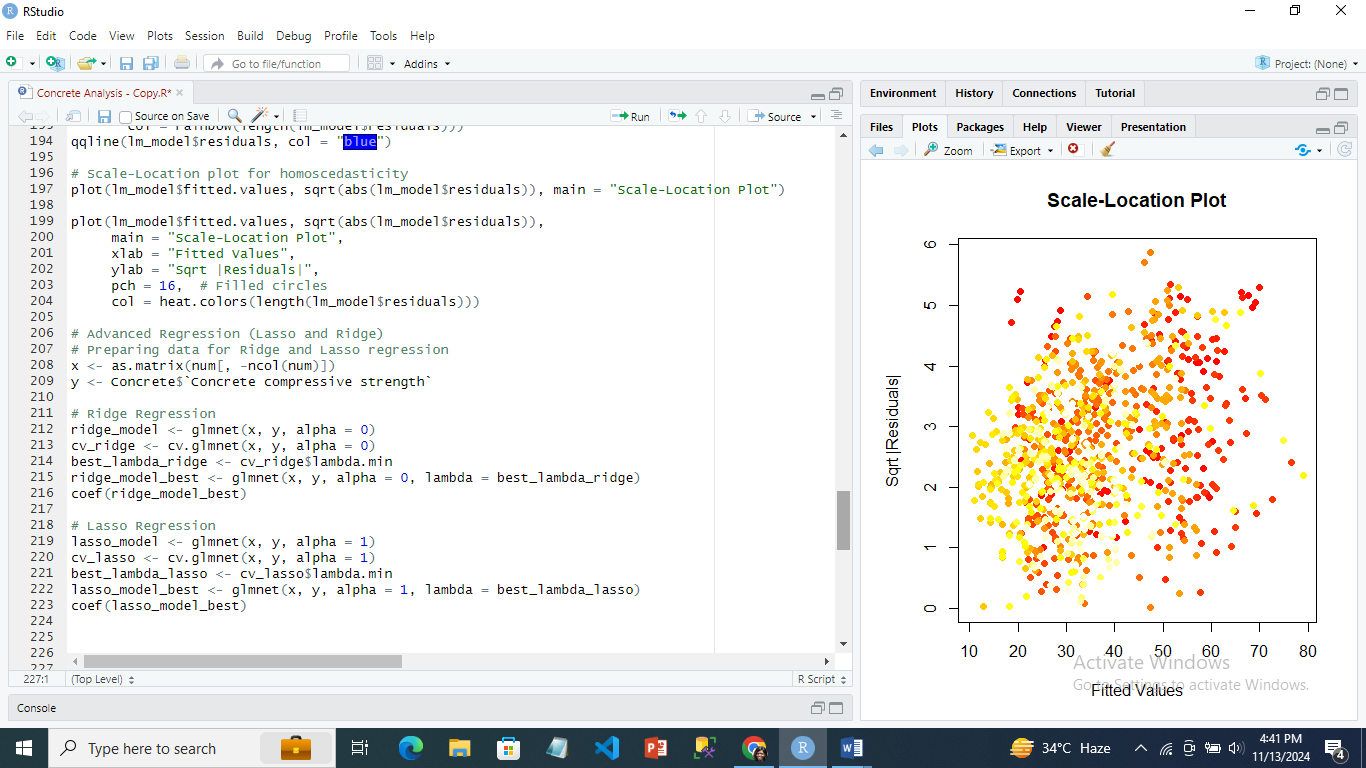
## 5.4 ****Scale-Location Plot for Homoscedasticity****

The **Scale-Location plot** was used to check the assumption of homoscedasticity, which checks if the residuals had constant variance across fitted values. The plot displayed a random scatter of residuals with no clear patterns which shows that the residuals have constant variance across fitted values



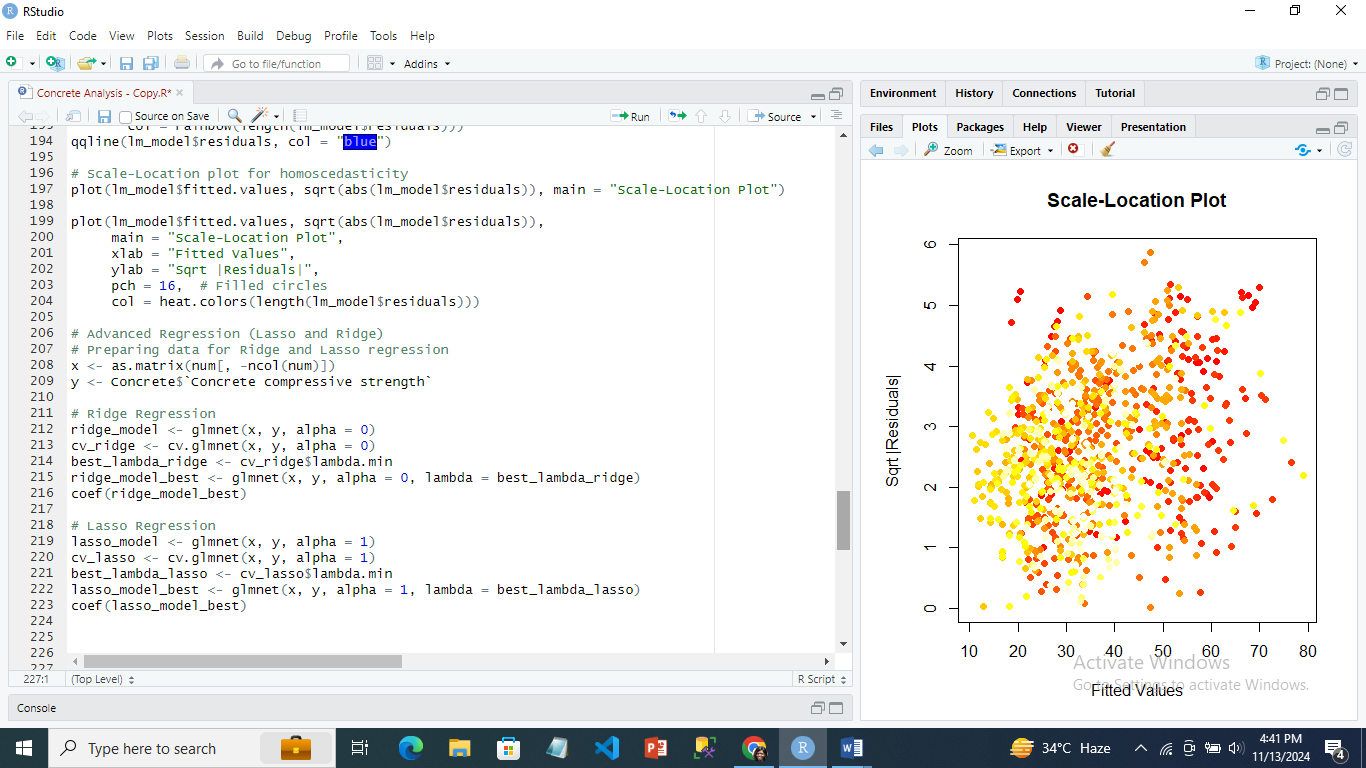
## 5.5 Advanced Regression Models

The Ridge and Lasso Regressions wre used to address potential overfitting and improve model performance, these techniques were used for regularization



#### 5.51. ****Ridge Regression****

Ridge Regression applies an L2 penalty to the model coefficients to prevent overfitting by shrinking coefficients towards zero. The **best lambda value** was determined using cross-validation.

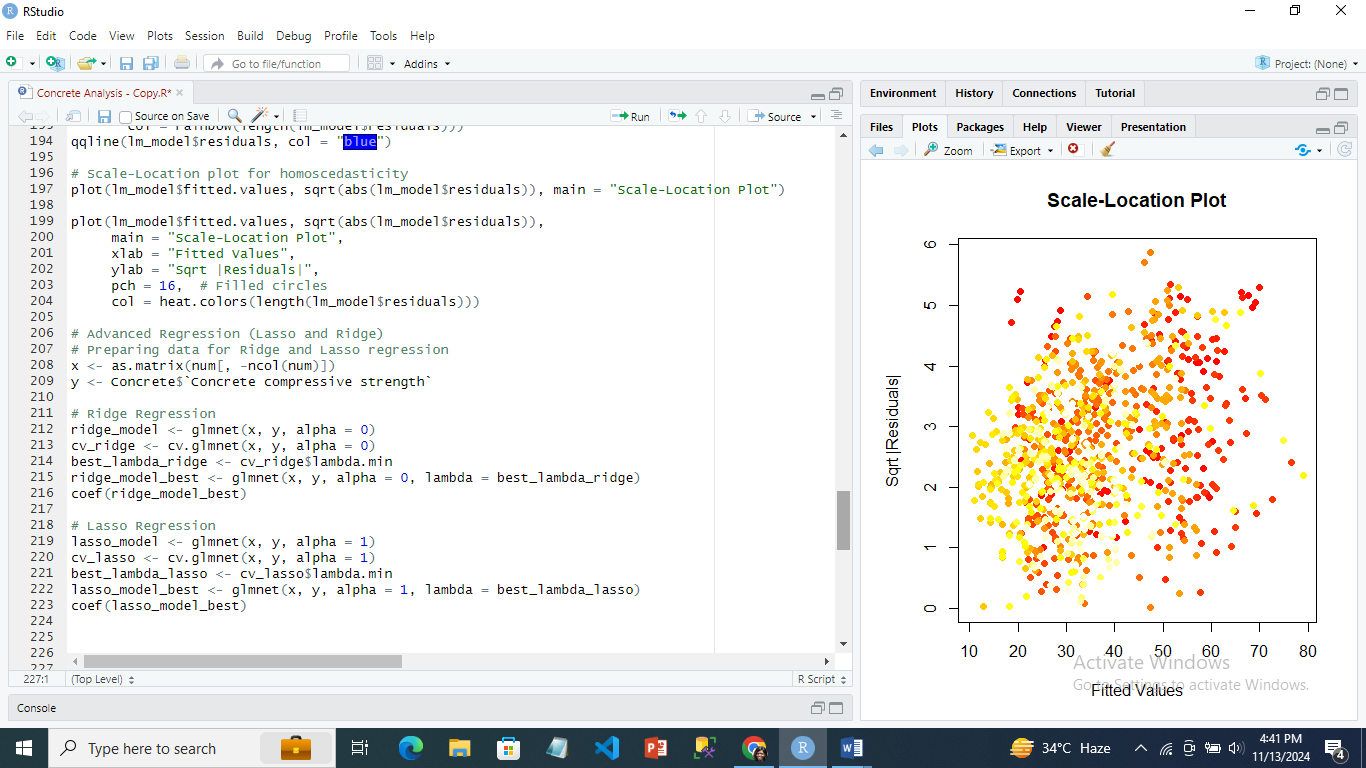


|  |  |
| --- | --- |
| Variable | Coefficient |
| Intercept | 65.4464 |
| Cement | 0.0830 |
| Blast Furnace Slag | 0.0606 |
| Fly Ash | 0.0355 |
| Water | -0.2362 |
| Superplasticizer | 0.3629 |
| Coarse Aggregate | -0.0106 |
| Fine Aggregate | -0.0171 |
| Age (day) | 0.1051 |

Positive coefficients imply a direct link with compressive strength (e.g., superplasticizer, cement), whereas negative coefficients indicate the opposite (e.g., water). Variables with modest coefficients, such as Fine Aggregate and Coarse Aggregate, appear to have a lower influence on the result than other variables.

#### 5.52. ****Lasso Regression****

Lasso Regression applied the L1 penalty that shrink some coefficients exactly to zero, which is useful for feature selection. The **best lambda value** was also selected through cross-validation.



|  |  |
| --- | --- |
| Variable | Coefficient |
| Intercept | 28.8977 |
| Cement | 0.1022 |
| Blast Furnace Slag | 0.0825 |
| Fly Ash | 0.0626 |
| Water | -0.2102 |
| Superplasticizer | 0.2701 |
| Coarse Aggregate | 0.0001 |
| Fine Aggregate | 0.0000 |
| Age (day) | 0.1110 |

Lasso Regression reduced the coefficient for Fine Aggregate to zero, indicating it does not contribute to the model. The Lasso model retained only the most influential variables, such as Cement, Fly Ash, Superplasticizer, and Age (day), emphasizing the most significant predictors of compressive strength.

## 6. Hypothesis Testing

Hypotheses about the impact of key components on compressive strength were formulated and tested using appropriate statistical tests. Non-parametric tests were applied when model assumptions were not met.

## 6.1 Does Cement Content Affect Compressive Strength?

* **Pearson Correlation Test**
  + Correlation Coefficient 0.4978
  + p-value: < 2.2e-16 (statistically significant)
* **T-Test**
  + Groups: High Cement vs Low Cement (based on median split)
  + p-value: < 2.2e-16 (statistically significant)
  + Confidence Interval: [12.36, 16.06]
  + Mean in High Cement Group: 42.92
  + Mean in Low Cement Group: 28.71

This hypothesis tests whether the amount of cement in the concrete affects how strong it is. This hypothesis is to check if cement is a key ingredient in concrete and to find out if more cement makes the concrete stronger. After analyzing the data, the results show a moderate positive relationship. This means that as the amount of cement increases, the compressive strength tends to increase as well.

It was also found that the higher cement content, the stronger the compressive strength.

In conclusion, this hypothesis is satisfied, as the analysis shows that an increase in cement content leads to higher compressive strength.

## 6.2 Does Water Content Affect Compressive Strength?

* T-Test:
  + Groups: High Water vs Low Water (based on median split)
  + p-value: < 2.2e-16 (statistically significant)
  + Confidence Interval: [-12.80, -8.94]
  + Mean in High Water Group: 30.33
  + Mean in Low Water Group: 41.20
* Pearson Correlation Test:
  + Correlation Coefficient: -0.2896 (moderate negative correlation)
  + p-value: < 2.2e-16 (statistically significant)
* Linear Regression:
  + Coefficient for Water: -0.2266 (each unit increase in Water content is associated with a decrease in compressive strength)
  + p-value: < 2e-16 (statistically significant)
  + Multiple R-squared: 0.08388 (explains about 8.39% of the variance in compressive strength)

This hypothesis looked at the impact of water content on the strength of the concrete. Water is also an important ingredient in the mix, The analysis showed that the more water makes the compressive strength weaker. In the t-test, the group with lower water content had significantly higher strength than the group with higher water content. The regression model also confirmed that adding more water tends to reduce the compressive strength.

In conclusion, it satisfies the hypothesis because the results indicate that higher water content negatively affects compressive strength.

## 6.3 Does Superplasticizer Content Affect Compressive Strength?

* Pearson Correlation Test:
  + Correlation Coefficient: 0.3661 (moderate positive correlation)
  + p-value: < 2.2e-16 (statistically significant)
* Linear Regression:
  + Coefficient for Superplasticizer: 1.02385 (each unit increase in Superplasticizer content is associated with an increase in compressive strength)
  + p-value: < 2e-16 (statistically significant)
  + Multiple R-squared: 0.134 (explains about 13.4% of the variance in compressive strength)

This hypothesis tests if the amount of superplasticizer in the concrete affects its strength. Superplasticizers are chemical additives used to improve the workability of concrete without needing more water. The analysis revealed a moderate positive relationship, showing that more superplasticizer content leads to stronger concrete. The regression model confirmed this, showing that as the superplasticizer content increased, so did the concrete’s strength, and this relationship was statistically significant.

In conclusion, it satisfies the hypothesis because the findings revealed that using superplasticizer improves compressive strength, likely due to enhanced workability and mix efficiency.

## 7. Summary and Recommendations

The research demonstrates that both cement and superplasticizer have a positive impact on concrete compressive strength, which means that as their amounts grow, so does the concrete's capacity to endure pressure. Cement is a key component with a moderately favourable association with strength, whereas superplasticizer, an addition to increase workability, makes a considerable contribution. In contrast, water reduces compressive strength; increasing water content weakens the concrete.

Following these observations, a few suggestions can be made. To make stronger concrete, use larger quantities of cement and superplasticizer, as these chemicals directly contribute to compressive strength. At the same time, water content must be carefully regulated since excess water reduces strength. These changes in material proportions can help produce a more durable and high-strength concrete mix, which is especially significant in building projects that require higher structural integrity.

# References

Young, B. A., Hall, A., Pilon, L., Gupta, P., & Sant, G. (2019). Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods. *Cement and concrete research*, *115*, 379-388.

European Commission. (2019). Internet of Things (IoT) platform for concrete strength measurement and Machine Learning (ML) predictions in construction *Official Journal of the European Communities* Retrieved from <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projects-details/31045243/876017/H2020>

Eskandari-Naddaf, H., & Kazemi, R. (2017). ANN prediction of cement mortar compressive strength, influence of cement strength class. *Construction and Building Materials*, *138*, 1-11.