

School of Computing, Creative Technology and Engineering

|  |  |
| --- | --- |
| Student ID | 77356725 |
| Student Name | Sadikshya Duwadi |
| Module Name & CRN | Applied Machine Learning-18910 |
| Level | 5 |
| Assessment Name & Part No. | Case Study-3: Concrete Strength, PART-I |
| Project Title | Data Analysis on Concrete Strength |
| Date of Submission | 31st March, 2024 |
| Course | BSc. (Hons) Computing |
| Academic Year | 2024 |

Table of Contents

[PART I 3](#_Toc158150264)-20

Introduction ………………………………………………………………………………………. 3

Literature Review ………………………………………………………………………………… 3

Exploratory Data Analysis (EDA) ……………………………………………………………….. 3-9

Data Pre-processing …………………………………………………………………………….. 9-18

BIBLIOGRAPHY ……………………………………………………………………………... 19-20

List of Figures:

1. Figure 1: Structure of dataset Train and Test
2. Figure 2: Structure of combined dataset
3. Figure 3: Summary of two combined datasets
4. Figure 4: Dimension of combined datasets
5. Figure 1.1.1 Code snippet of Distribution of Cement Content
6. Figure 1.1.2 Histogram of distribution of cement content and their frequency
7. Figure 1.2.1 Code snippet to view density and strength
8. Figure 1.2.2 Histogram to view density and strength
9. Figure 1.3.1 Code snippet of Cement vs. Strength
10. Figure 1.3.2 Scatterplot of Cement vs. Strength
11. Figure1.4 Performing PCA with output
12. Figure 1.5.1 Extracting the first two PCA and creating scatter plot
13. Figure 1.5.2 Scatter plot of first two PCA components
14. Figure 1.6.1 Calculation of missing values in each column and overall
15. Figure 1.6.2 Code snippet of Data Imputation
16. Figure 1.6.3 Output of Data Imputation
17. Figure 1.6.4 md pattern of Final Imputed Data
18. Figure 1.7.1 Output of Outlier detection
19. Figure 1.7.2 Boxplot of Outlier detection
20. Figure 1.7.3 Removing outliers
21. Figure 1.7.4 Visualizations to compare before and after outlier handling
22. Figure 1.7.5 Box plot compare before and after outlier handling
23. Figure 1.8.1 Calculating Multicollinearity
24. Figure 1.8.2 Visualizing correlation matrix
25. Figure 1.8.3 Code snippet of visualizing the relationship between Strength and Cement
26. Figure 1.8.4 Scatterplot between Strength & Cement
27. Figure 1.8.5 Removing variable having low variance
28. Figure 1.9.1 Code Snippet of Impact of Scaling on the target variable ‘Strength’
29. Figure 1.9.2 gg Plot of Original and Scaled value of Strength

# PART I

Data Analysis on Concrete Strength

Sadikshya Duwadi, Bsc (hons) Computing, 2024

**Introduction**:

Concrete, one of the world's most produced material, is structural material made of a hard, chemically inert particle component known as aggregate (often sand and gravel) that is formed by the interaction of cement and water which cures overtime (Concrete | Definition, Composition, Uses, Types, & Facts | Britannica, 2024). When making a building, houses or any structure, it is very important to know the compressive strength of the concrete to make it more durable and last long. The strength of the concrete mainly depends on the amount and quality of the material (Cement, fly ash, Water, Coarse aggregate, etc.) mixed to make it. The use of data analysis or machine learning for knowing the compressive strength of the concrete have all contributed to the progress of concrete mix designs.

Literature review:

The study examined several areas of measuring and optimising concrete strength. It investigated how various local sources of fine aggregates impact concrete characteristics, highlighting the critical importance of aggregates in this respect (Rahman, 2020). Furthermore, the research emphasised the need of attaining the appropriate bond strength between layers of high-strength and lightweight concrete, and they proposed several strategies for improving interlayer bonding (Eisa, Aboul-Nour and Mohamad, 2024). A statistical study of concrete compressive strength measurement methods was also conducted, with the goal of determining the optimal sample sizes and proposing improved evaluation methodologies (Sujeet Kumar Mahato and Kumar, 2024). Predictive methods, such as the IABC-MLP algorithm, were praised for their accuracy in forecasting concrete strength by combining heuristic algorithms and neural networks (Li et al., 2024). Furthermore, the use of machine learning approaches to optimise concrete mix designs was noteworthy, with artificial neural networks and data mining being used to effectively estimate compressive strength (Ziolkowski and Maciej Niedostatkiewicz, 2019). The historical evolution of cement and concrete, as well as their reactivity to environmental conditions, demonstrated the continual need for study and improvement in concrete engineering (Gagg, 2014). Furthermore, research into aggregate grading and natural sand composition revealed their major influence on concrete strength, emphasising the need of knowing them for appropriate mix design (S. Hasdemir, A. Tuğrul and M. Yılmaz, 2016). Finally, the study of the influence of specimen size on compressive strength evaluation demonstrated the need of include specimen features in concrete testing methods (Banarjee, Alam and Ahmad, n.d.).

* **Ee**

EDA employs statistical and visualisation approaches to eewhich can help us to understand the data easily.

The two datasets, train and test are combined to find the compressive strength of the concrete. The combined dataset has1030 number of rows and 10 number of columns. The ultimate variable, 'Strength', indicates the concrete's compressive strength, which we are trying to predict in this case study.

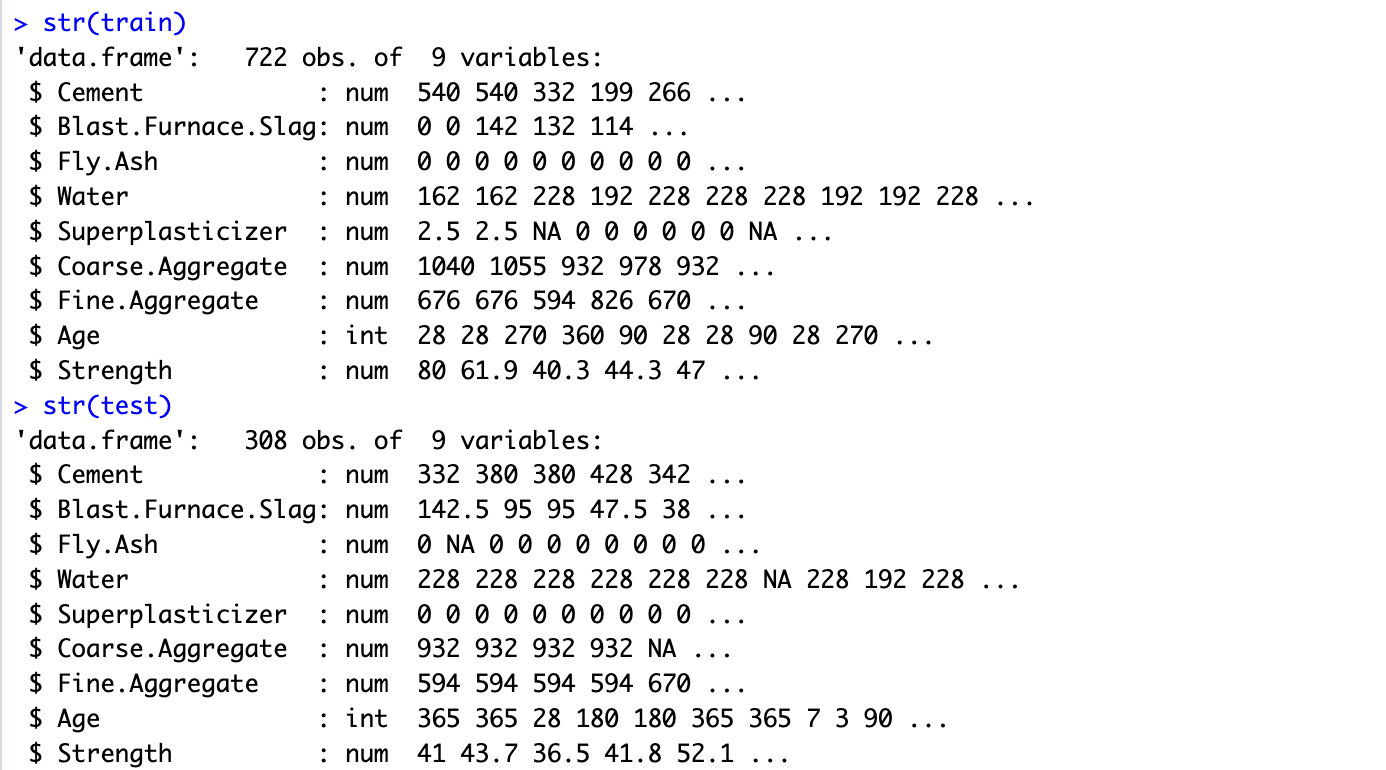


Figure 1: Structure of dataset Train and Test

The above figure 1 provides the structure of two data sets i.e. >str(train) for dataset Train and >str(test) for dataset Test. They both consist same structure with same row and column.

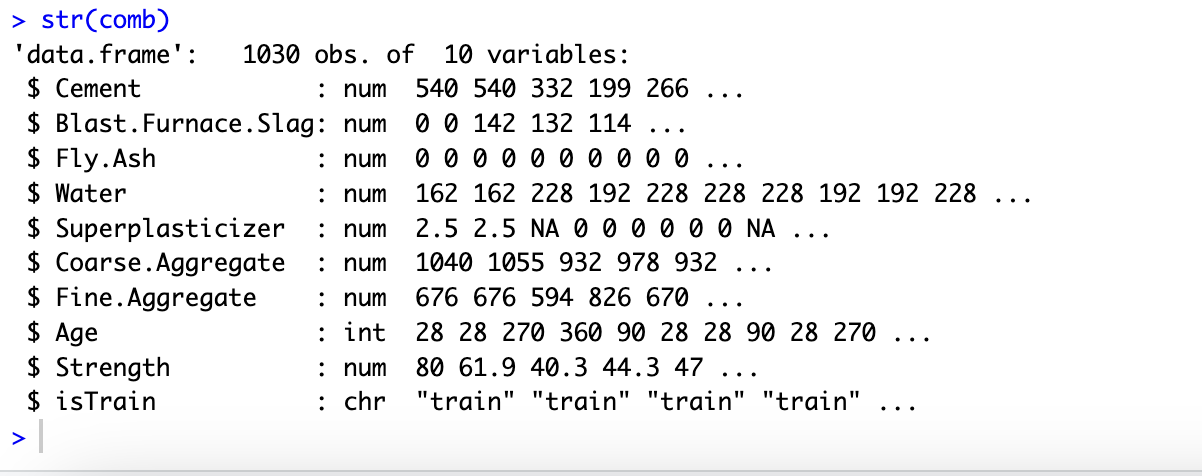
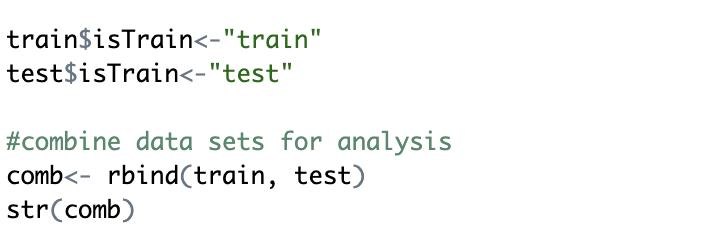


Figure 2: Structure of combined dataset



In fig 2, I have combined the above two datasets into one named as comb. Here, >str(comb) shows the structure of the combined datasets. I used rbind( ) function to combine them. I added an extra column ‘isTrain’ to distinguish them.

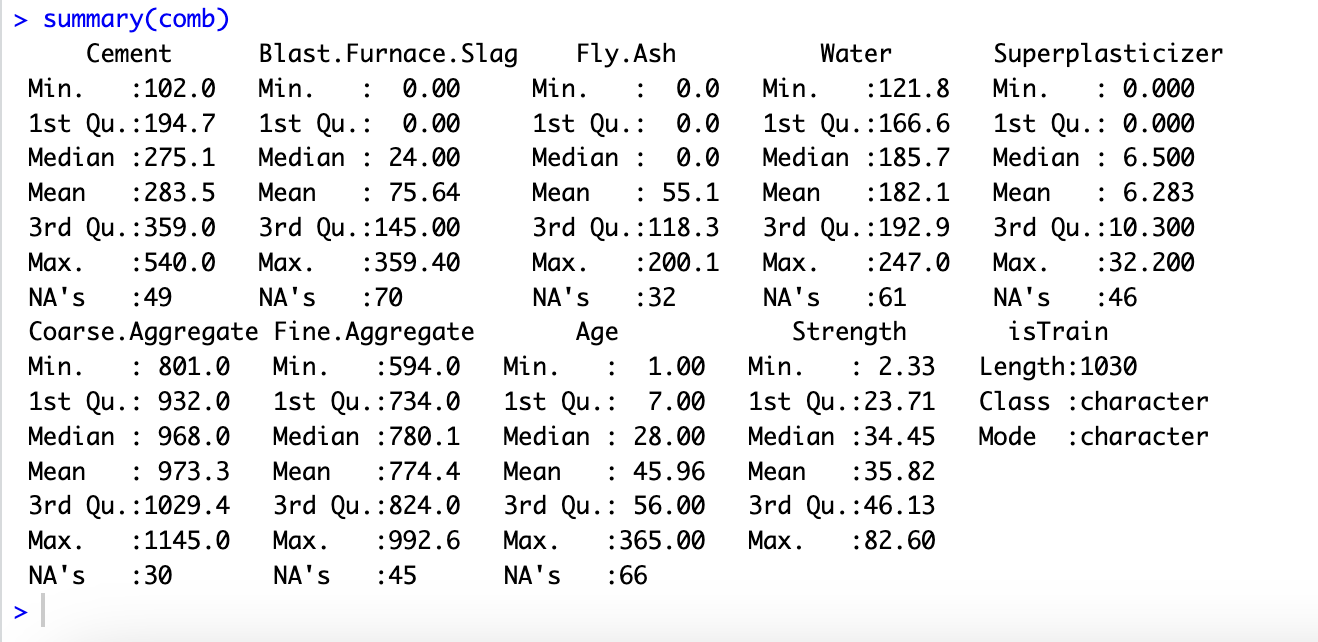


Figure 3: Summary of two combined datasets

Fig 3, provides the summary of two combined datasets. The summary consists of Minimum value, 1st Quartile, Median, Mean, 3rd Quartile, Maximum value and Missing value for each column.

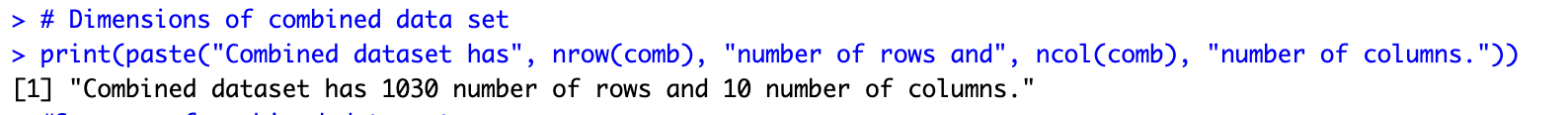


Figure 4: Dimension of combined datasets

Above fig 4, prints the dimension of combined datasets. Combined dataset has 1030 number of rows and 10 number of columns.

* **Visualization**

1. **Distribution of Cement Content**

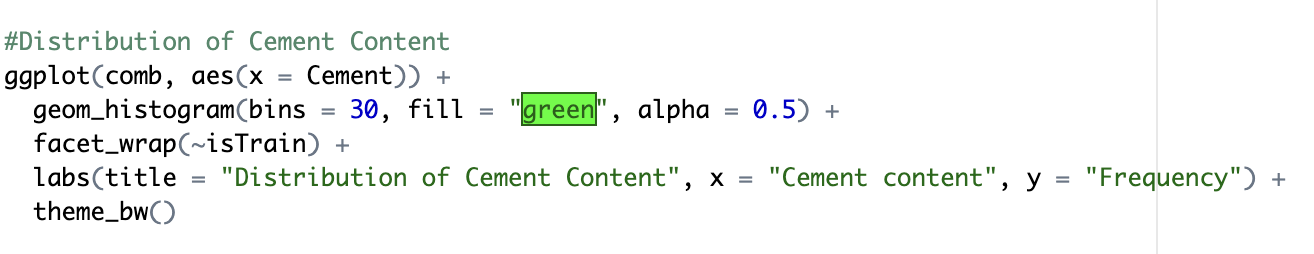


Figure 1.1.1 Code snippet of Distribution of Cement Content

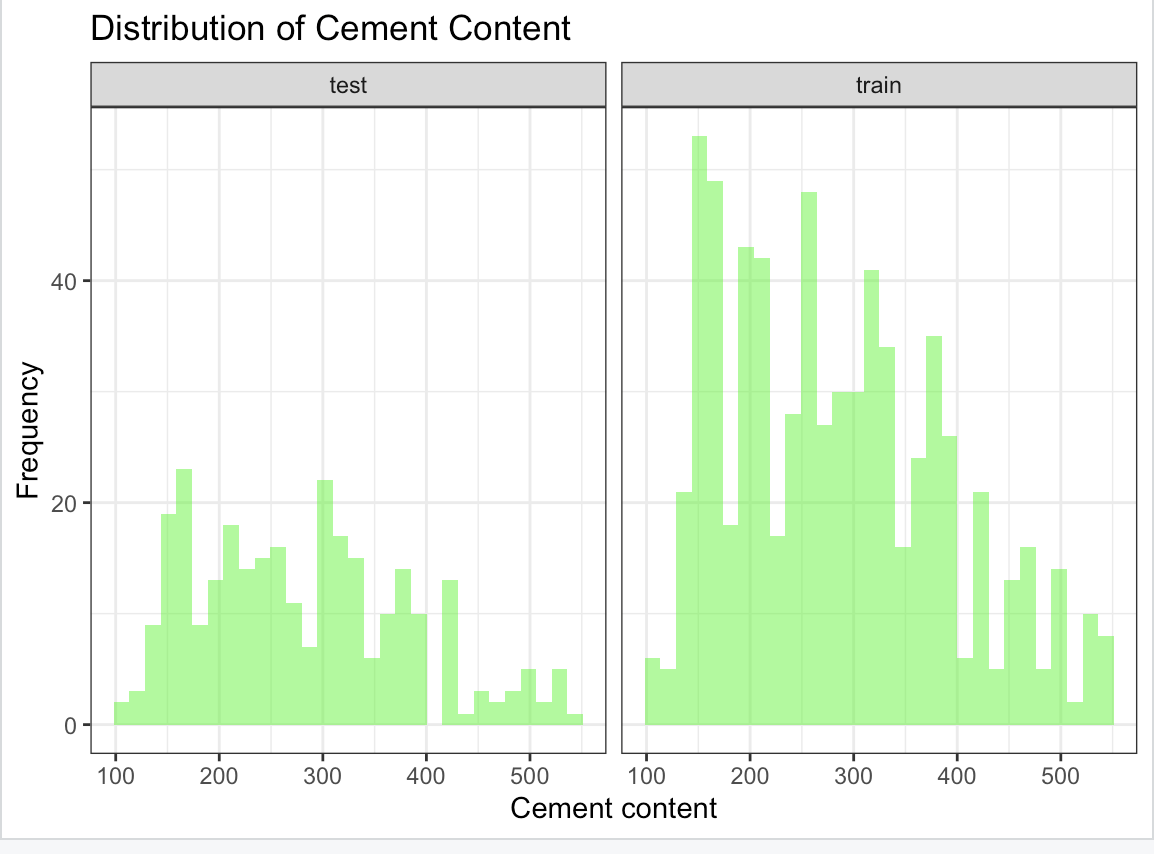


Figure 1.1.2 Histogram of distribution of cement content and their frequency

Figure 1.1.1 represents the ggplot code to give histogram of two combined datasets, where x-axis shows the cement content and x-axis shows the frequency in fig 1.1.2.

1. **Histogram of Density and Strength**

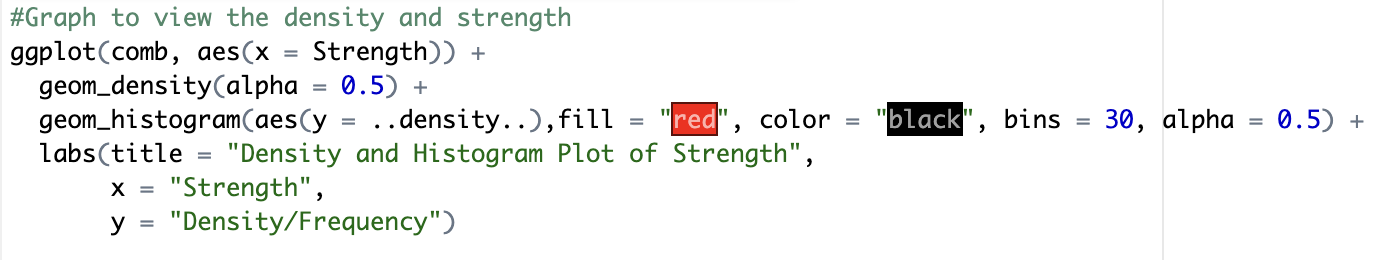
****

Figure 1.2.1 Code snippet to view density and strength

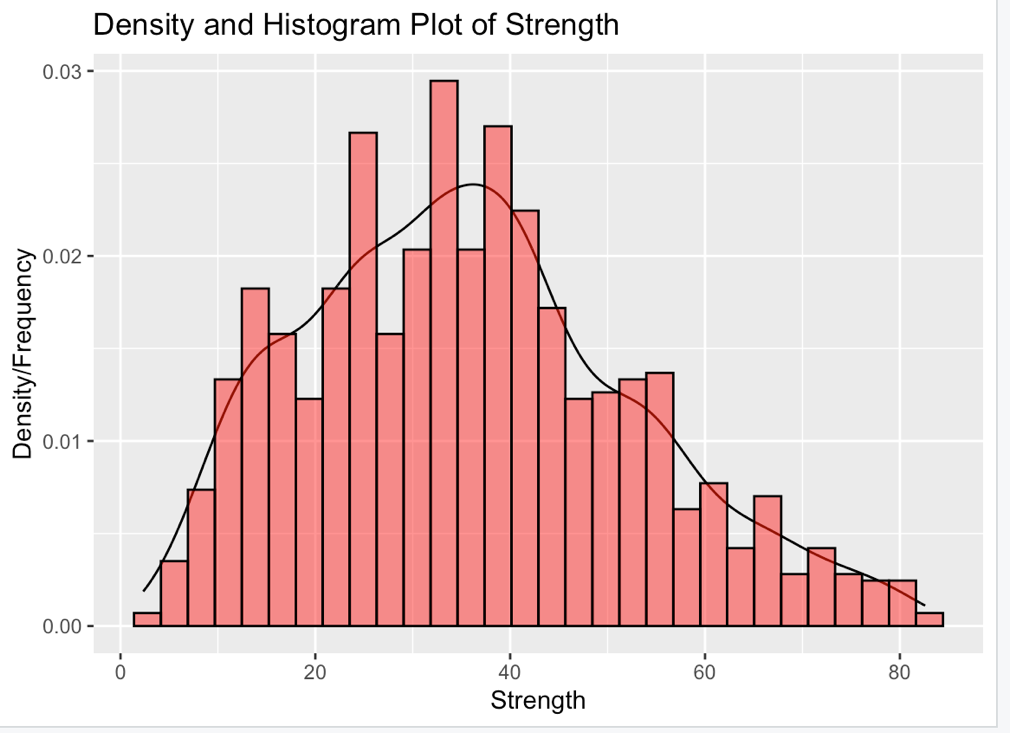
****

Figure 1.2.2 Histogram to view density and strength

Figure 1.2.2 shows the distribution of 'Strength' in the concrete dataset using both a density plot and a histogram. The smooth curve in the density plot represents a continuous picture of the data's probability distribution and histogram's bars provide a segmented representation of the data's frequency distribution.

1. **Scatter plot for Cement vs. Strength**

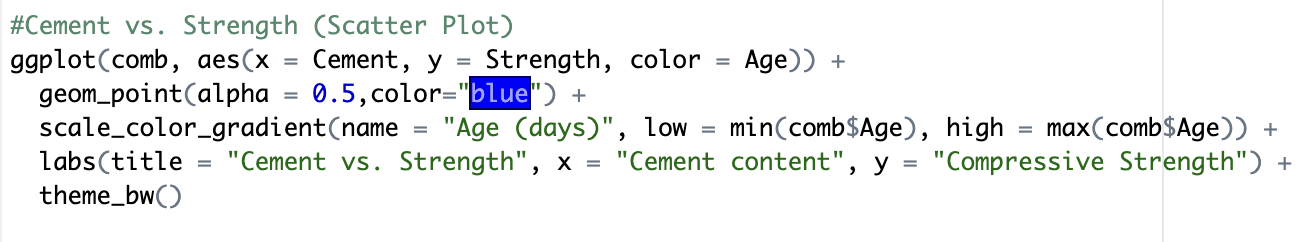
****

Figure 1.3.1 Code snippet of Cement vs. Strength

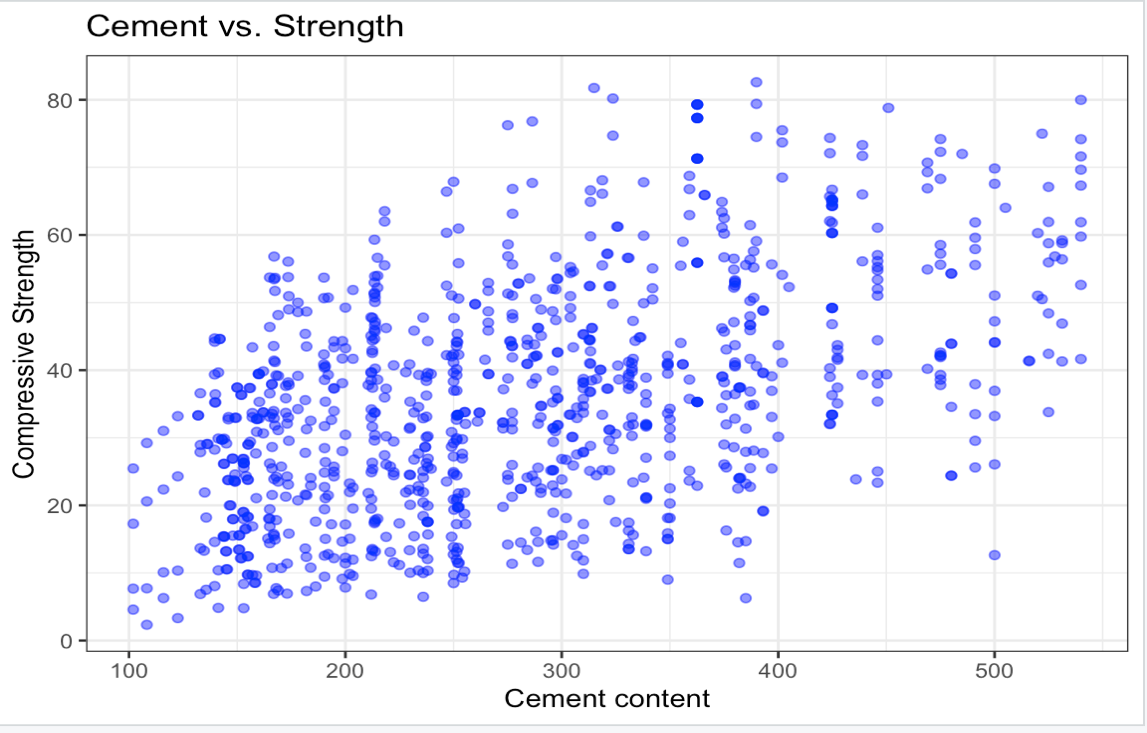
****

Figure 1.3.2 Scatterplot of Cement vs. Strength

Figure 1.3.1 is a scatterplot that examines the connection between 'Cement' and 'Strength' in the concrete dataset. The y-axis shows compressive strength and the x-axis shows cement concentration.

* + - **PCA**

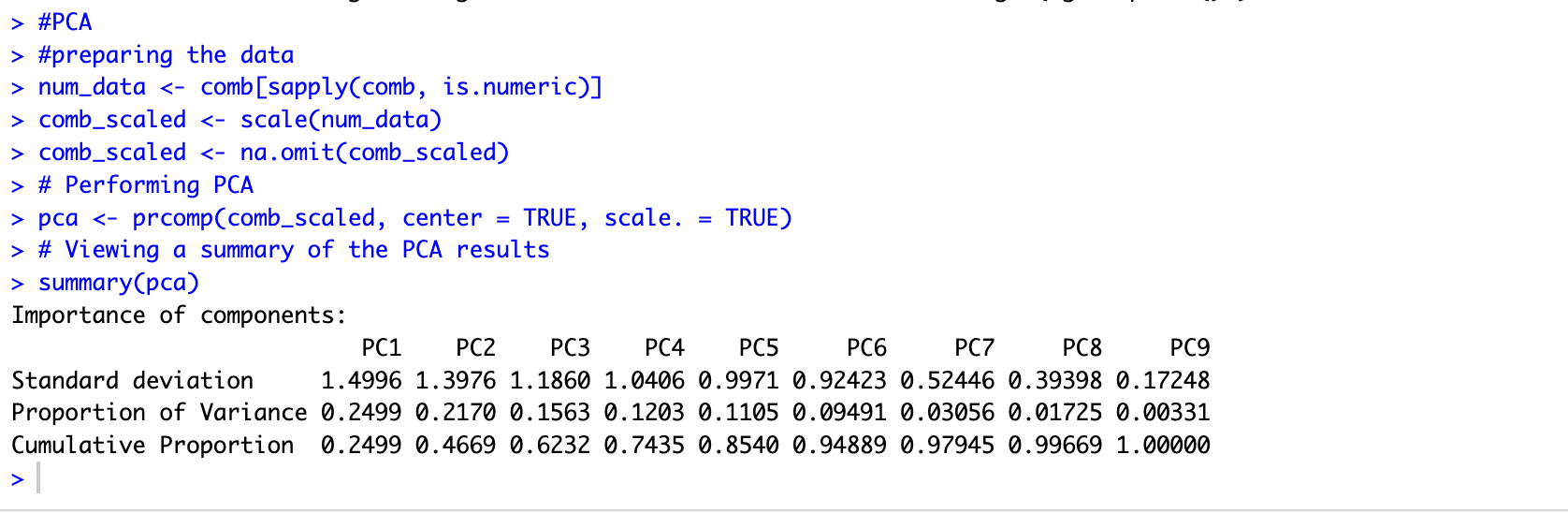
****

Figure1.4 Performing PCA with output

In the above fig 1.4, PCA is performed on the scaled numerical basis of combined dataset to comprehend the data's underlying structure. The result provides: Standard Deviation (PC1 and PC2 have the highest standard deviations, implying that they describe the majority of the variance in the data.), Proportion of Variance (PC1 makes up 24.99% of the overall variance, while PC2 contributes an additional 21.70%.) and Cumulative Proportion (It shows as we include more components, the variance percentage increases.)

* **Extracting the first two PCA and plotting**

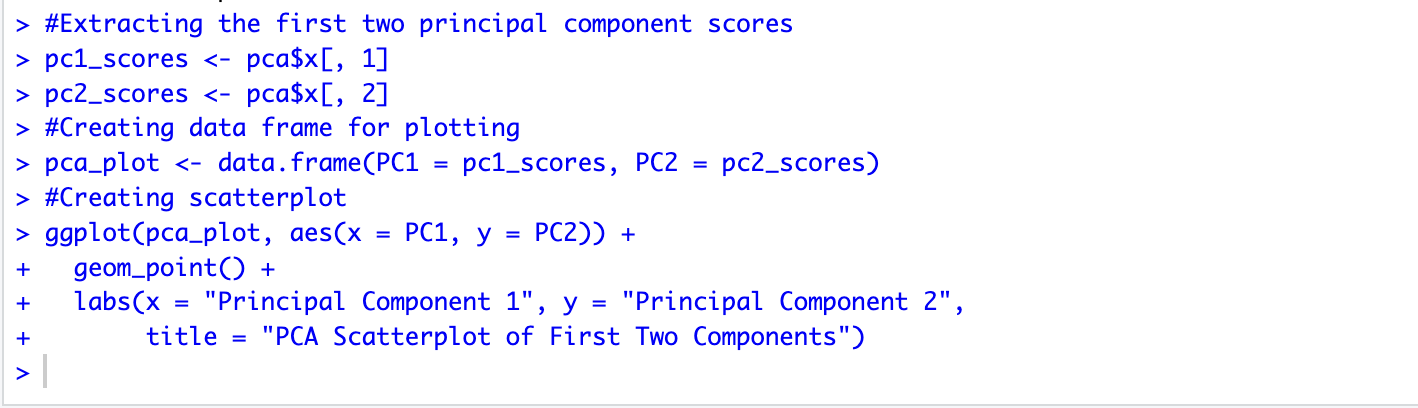
****

Figure 1.5.1 Extracting the first two PCA and creating scatter plot

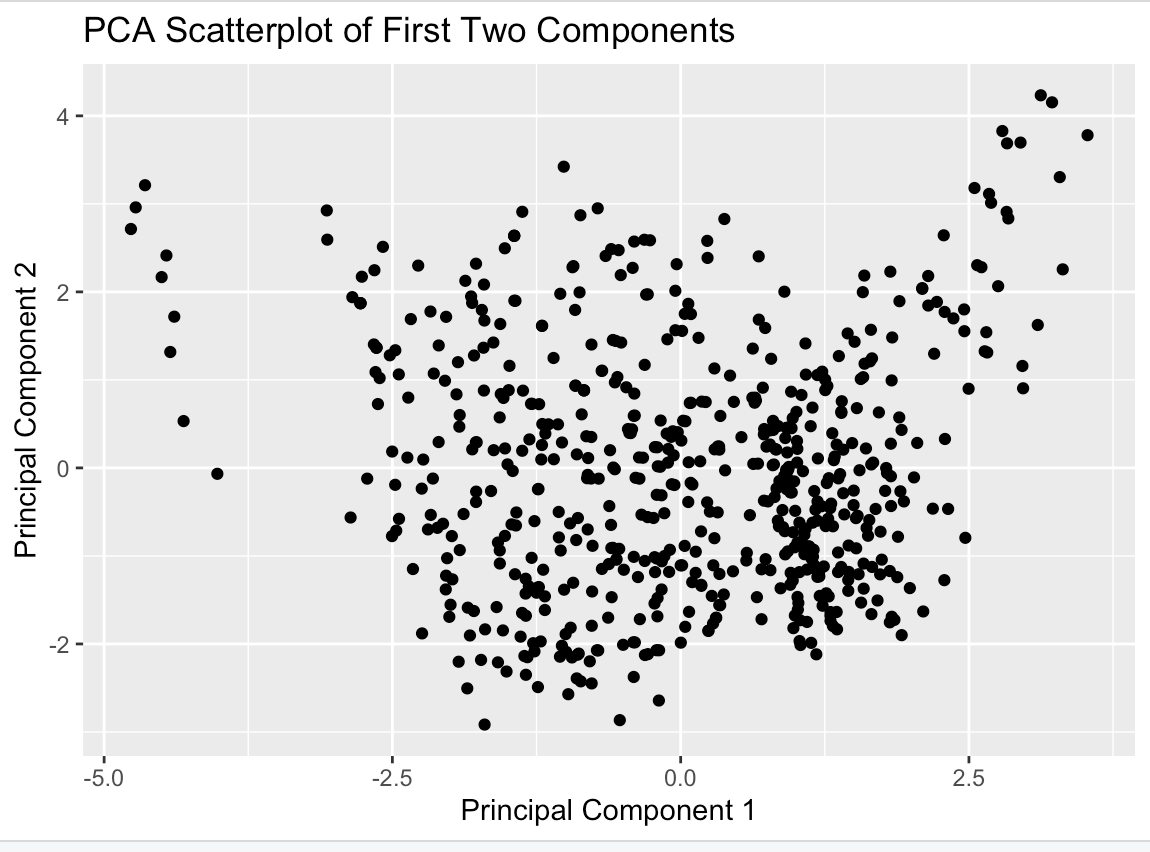
****

Figure 1.5.2 Scatter plot of first two PCA components

Figure 1.5.2 shows the scatter plot of first two PCA components i.e. PC1 and PC2 that we extracted in fig 1.5.1. Every data point is a physical sample represented into the dimension specified by these two main components.

* **Data Pre-processing**

1. **Missing Values**

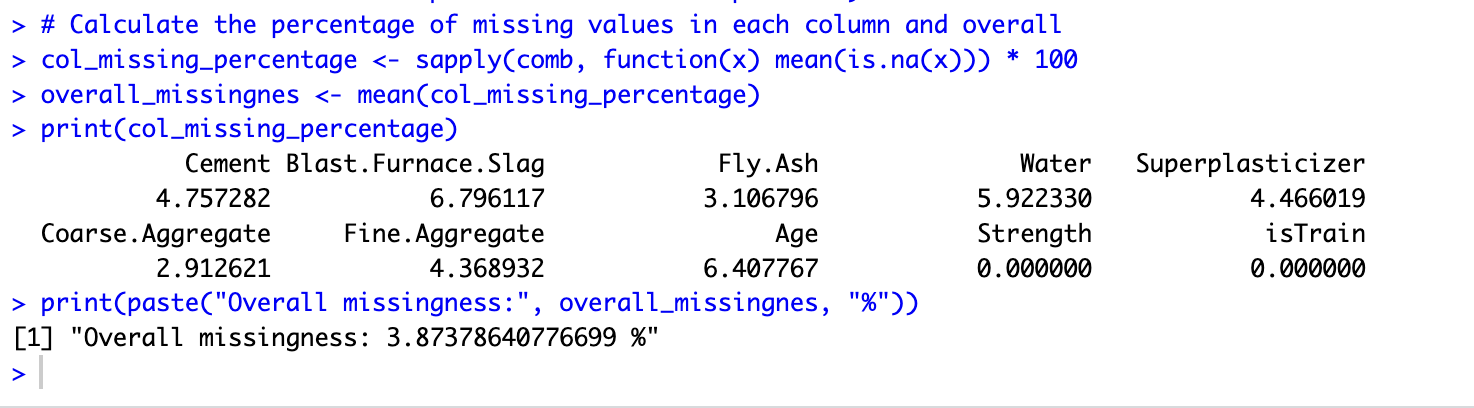
****

Figure 1.6.1 Calculation of missing values in each column and overall

Missing Data Percentage for each component: Cement'(4.76%), Blast Furnace Slag (6.8%), Fly Ash (3.11%), Water (5.92%), Superplasticizer (4.47%), Coarse Aggregate (2.91%), Fine Aggregate (4.37%), and Age'(6.41%).

Overall Missing Data Percentage: 3.873%.

* + **Imputation**

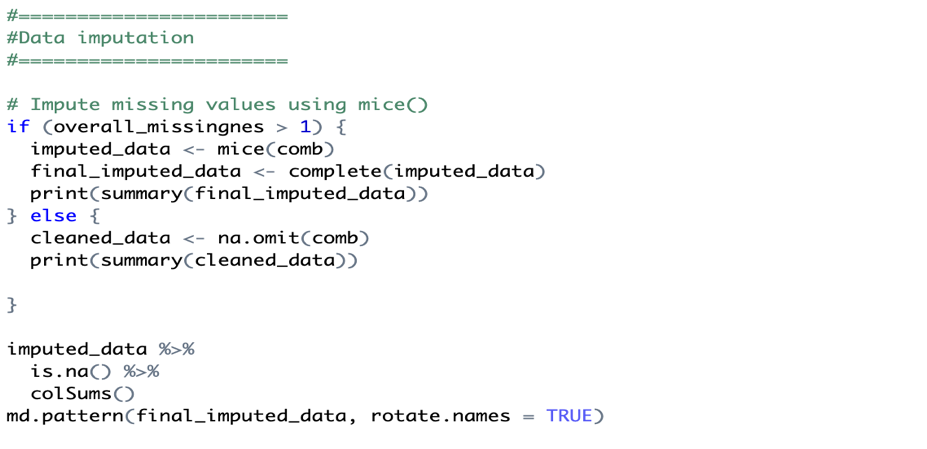
****

Figure 1.6.2 Code snippet of Data Imputation

In above fig 1.6.2, the mice function is used for cleaning data rather than omit function to generate numerous imputed datasets by filling in missing values based on known correlations between variables, if the average of missing values in column is greater than 1%, as the omit function may result in the loss of data. The md pattern for the final imputed data was also generated which can be found in fig 1.6.4.

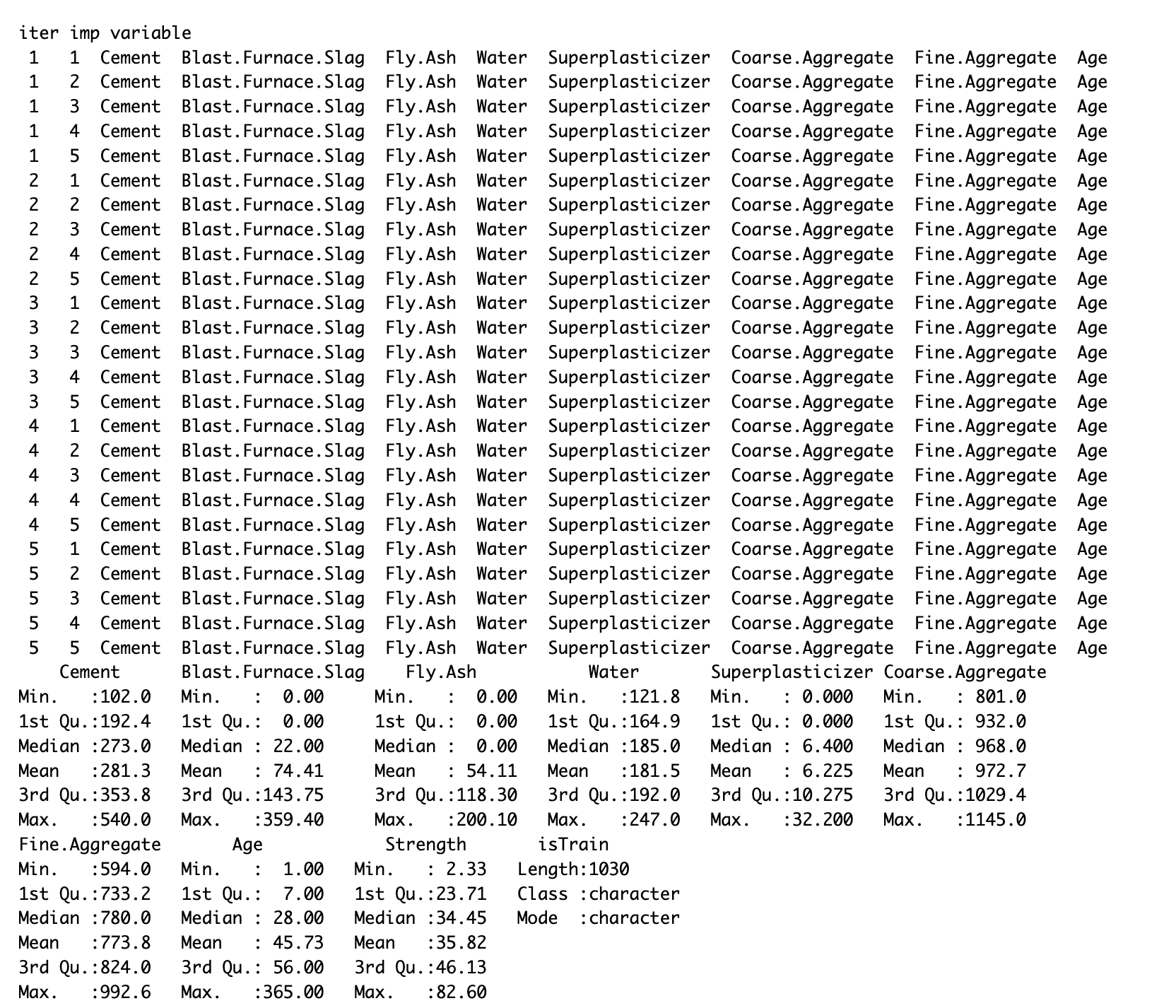
****

Figure 1.6.3 Output of Data Imputation

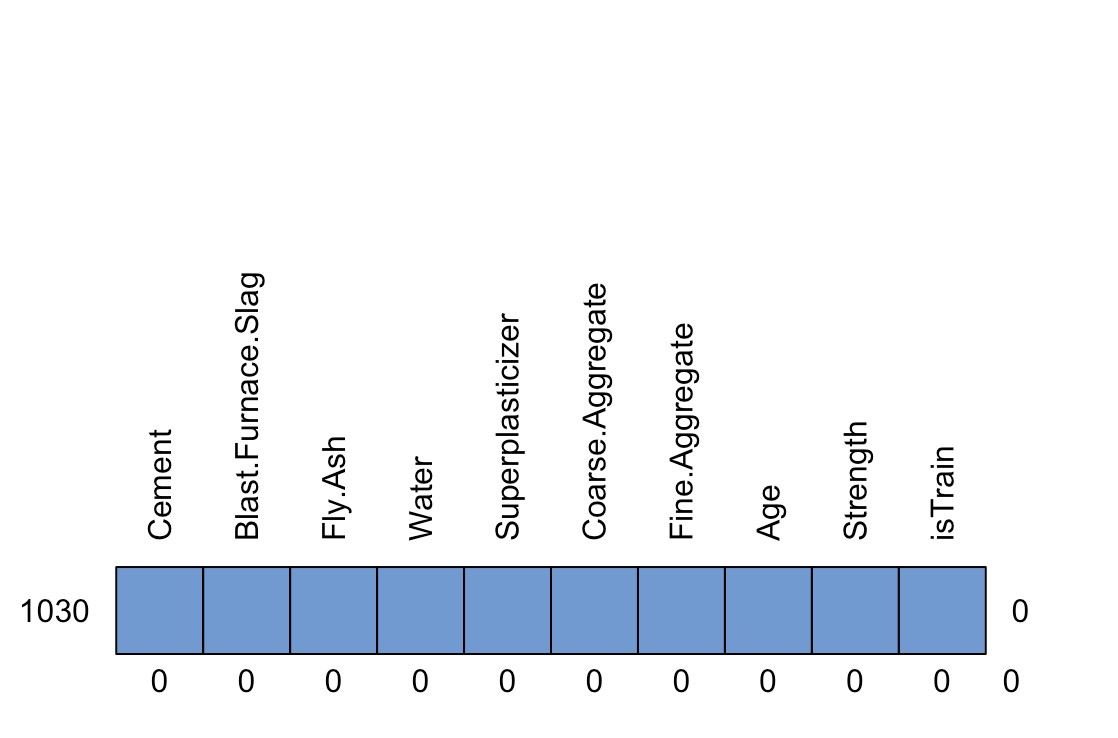
****

Figure 1.6.4 md pattern of Final Imputed Data

1. **Outliers**

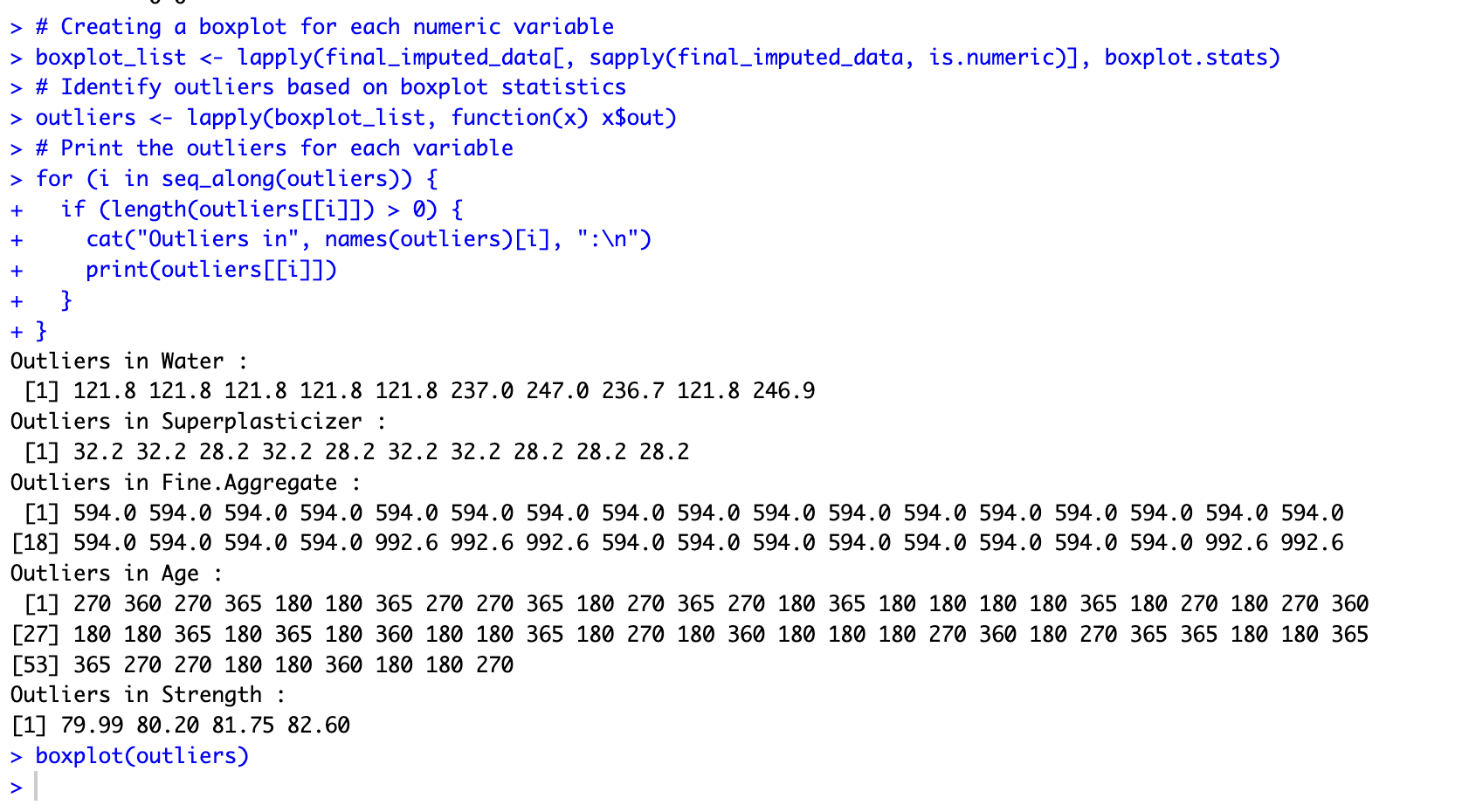
****

Figure 1.7.1 Output of Outlier detection

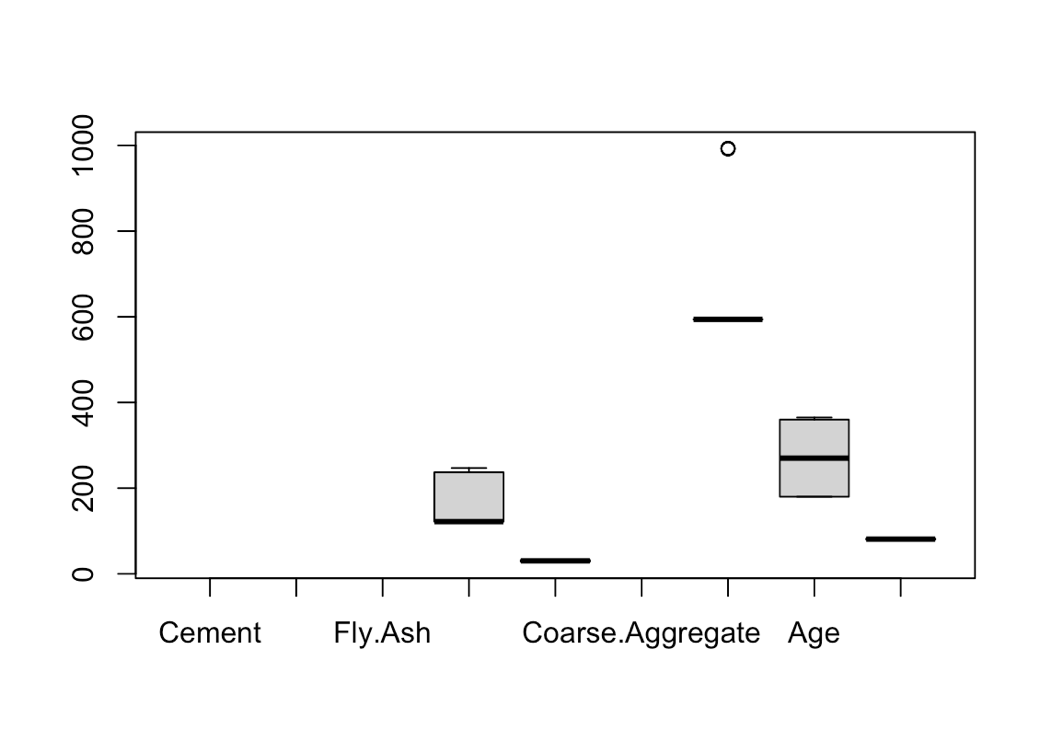
****

Figure 1.7.2 Boxplot of Outlier detection

To detect outliers in fig 1.7.1, Boxplot statistic method is used. Outliers can be detected in several components as shown in fig 1.7.2.

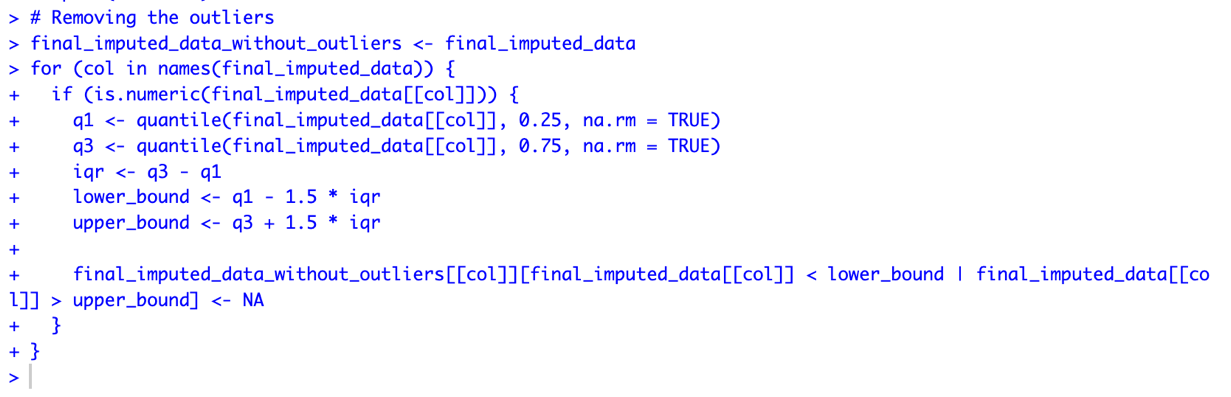
****

Figure 1.7.3 Removing outliers

Removing outliers can help clean up data and focus study on key patterns. But rather than removing the outliers completely, the above code flags the outliers, as they may represent valid events within datasets. It computes the interquartile range and establishes outlier bounds. It substitutes values that fall outside of certain boundaries with missing values (NA) to identify outliers and generates a new data frame with the original data and highlighted outliers.

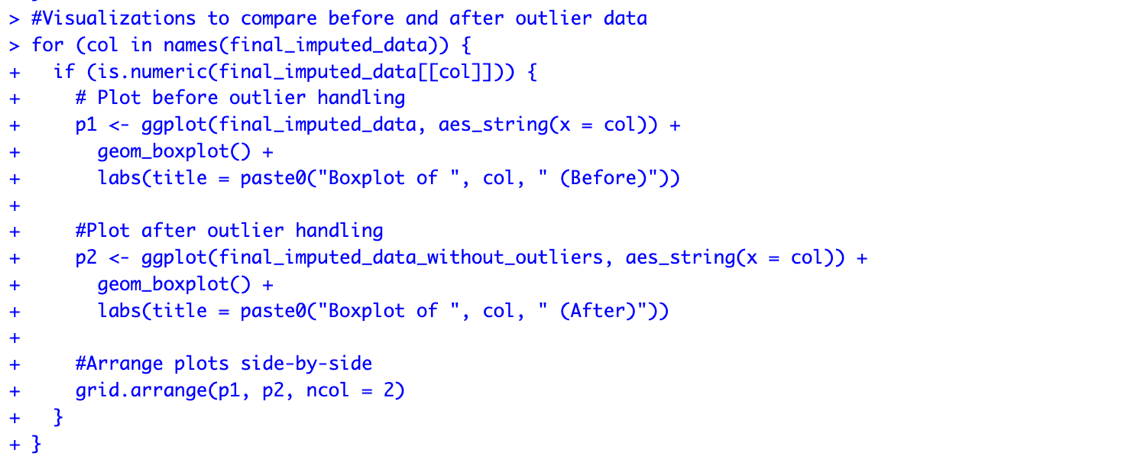
****

Figure 1.7.4 Visualizations to compare before and after outlier handling

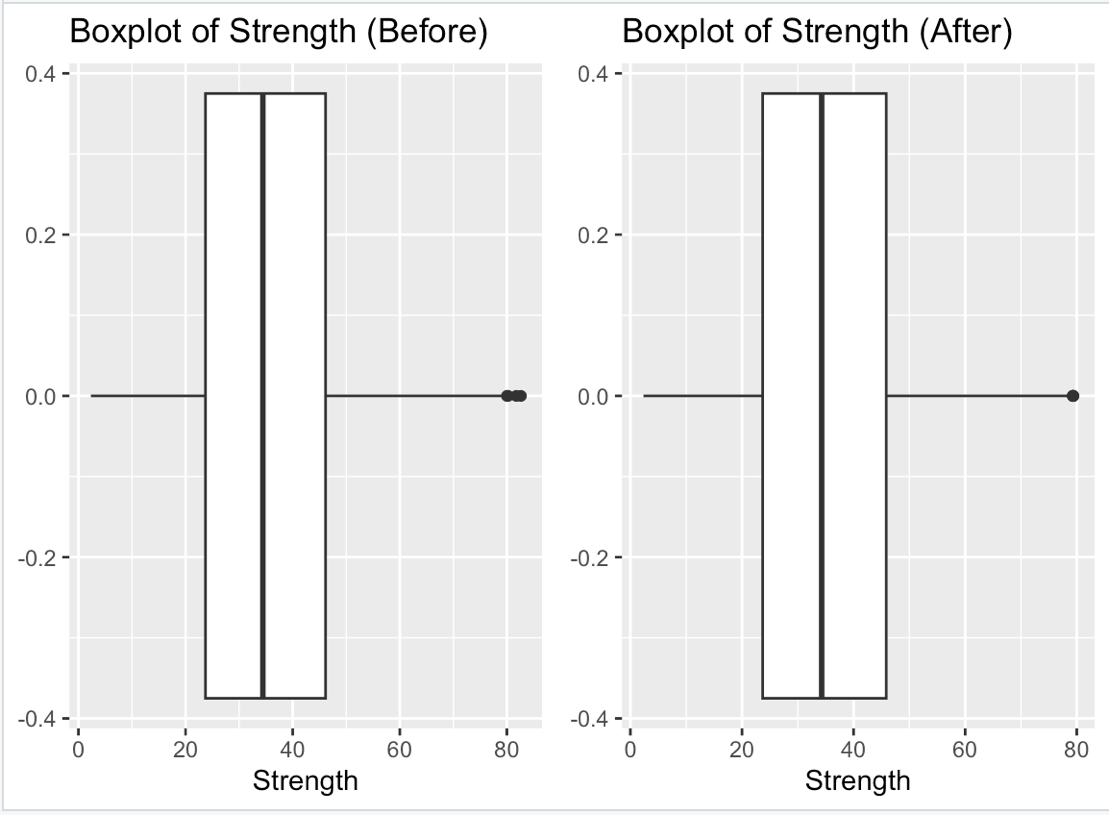
****

Figure 1.7.5 Box plot compare before and after outlier handling

Above fig 1.7.5, represents box plot to compare before and after outlier handling.

1. **Multicollinearity**

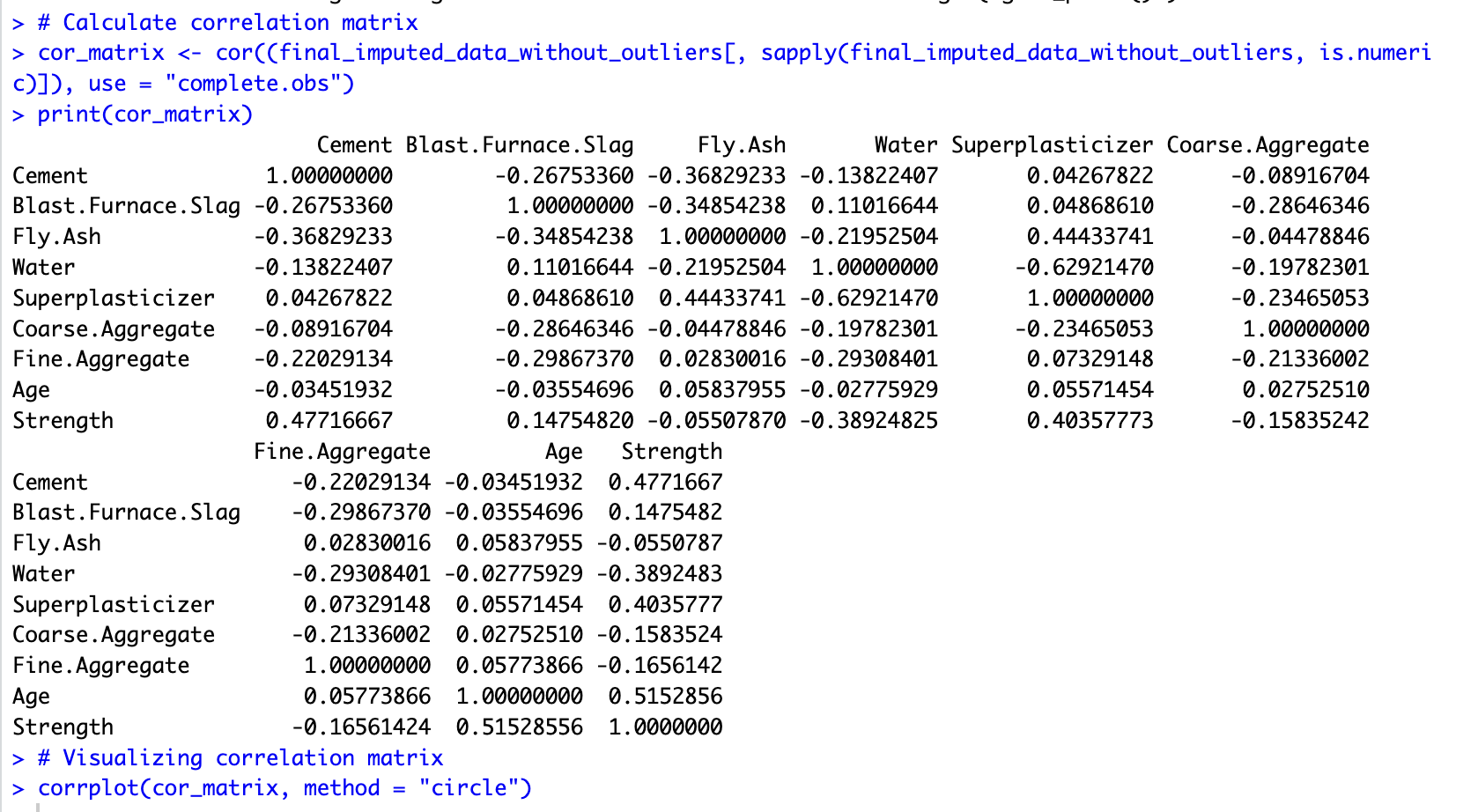
****

Figure 1.8.1 Calculating Multicollinearity

The output in fig 1.8.1 represents a correlation matrix, a table displaying the correlation coefficients between pairs of variables in a dataset.

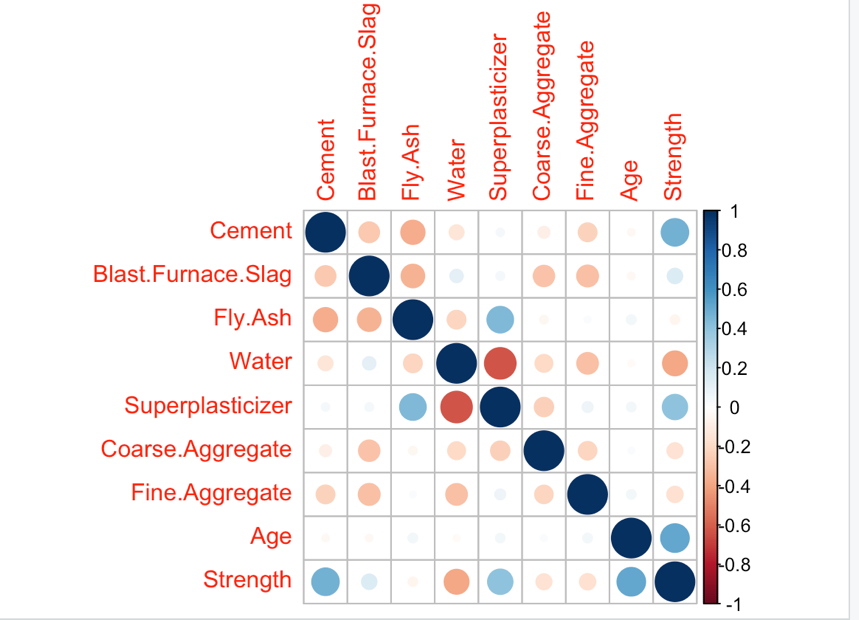
****

Figure 1.8.2 Visualizing correlation matrix

Figure 1.8.2 illustrates a correlation matrix with a circle packing visual representation. The blue circles represent strong positive correlation and the red circles represents strong negative correlation.

* **Scatterplot visualizing the relationship between Strength and Cement**

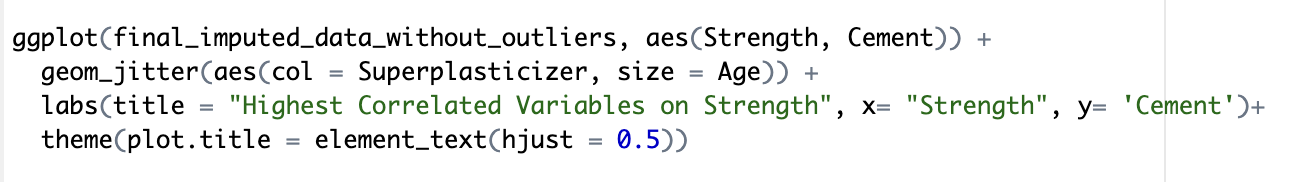
****

Figure 1.8.3 Code snippet of visualizing the relationship between Strength and Cement

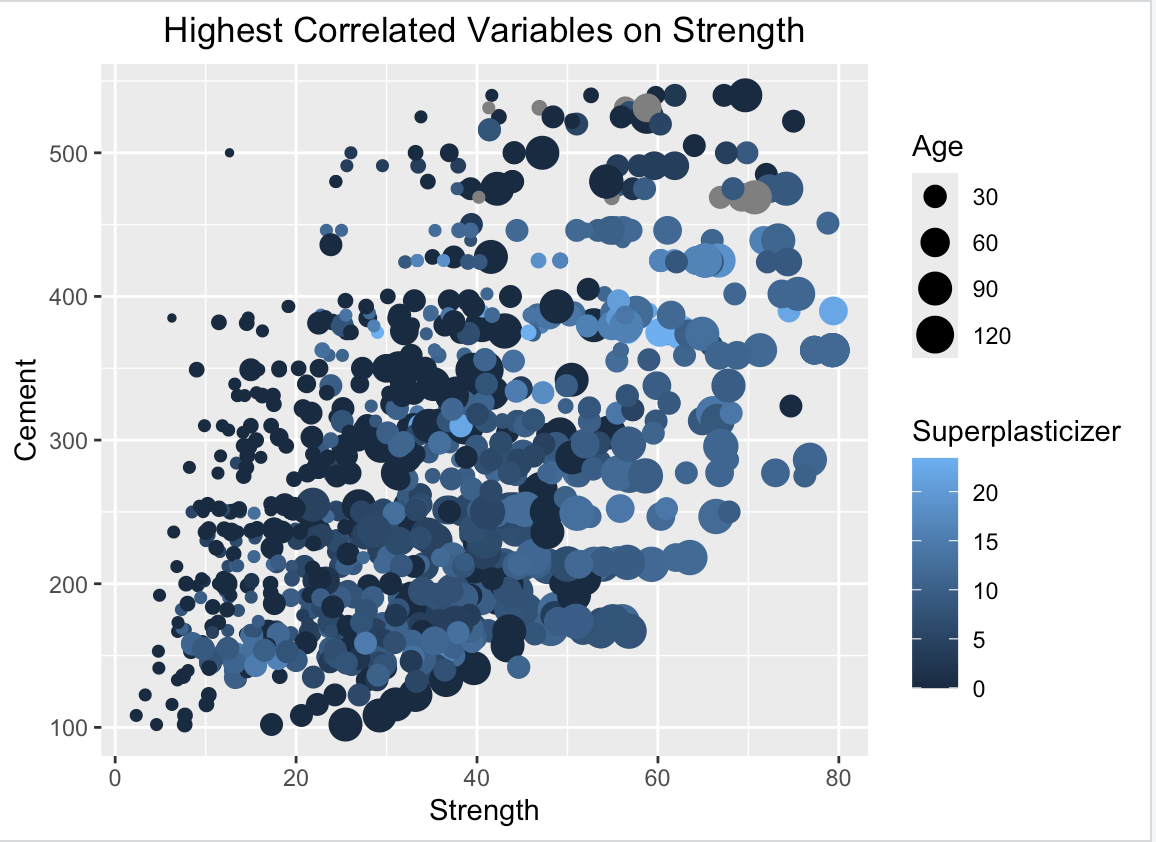
****

Figure 1.8.4 Scatterplot between Strength & Cement

Figure 1.8.4 shows the scatterplot between Strength & Cement, highlighted as highest correlated variables. The colour and size of the data points (black for age and blue for superplasticizer) may be used to visually investigate the effects of age and superplasticizer in this plot.

* **Investigating Low Variance Variable**

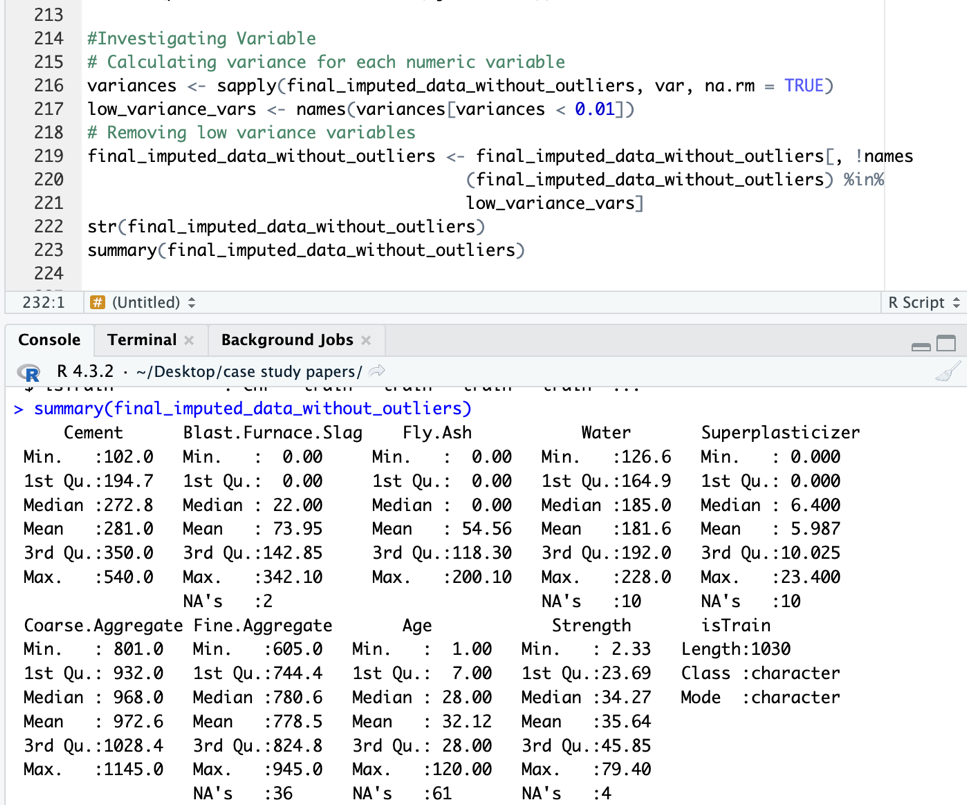


Figure 1.8.5 Removing variable having low variance

The code in fig 1.8.5 seeks to recognise and delete variables with low variance, which may not contribute much to modelling or analysis. Removing low variance variables can minimise noise, simplify models, and perhaps enhance performance.

1. **Scaling**

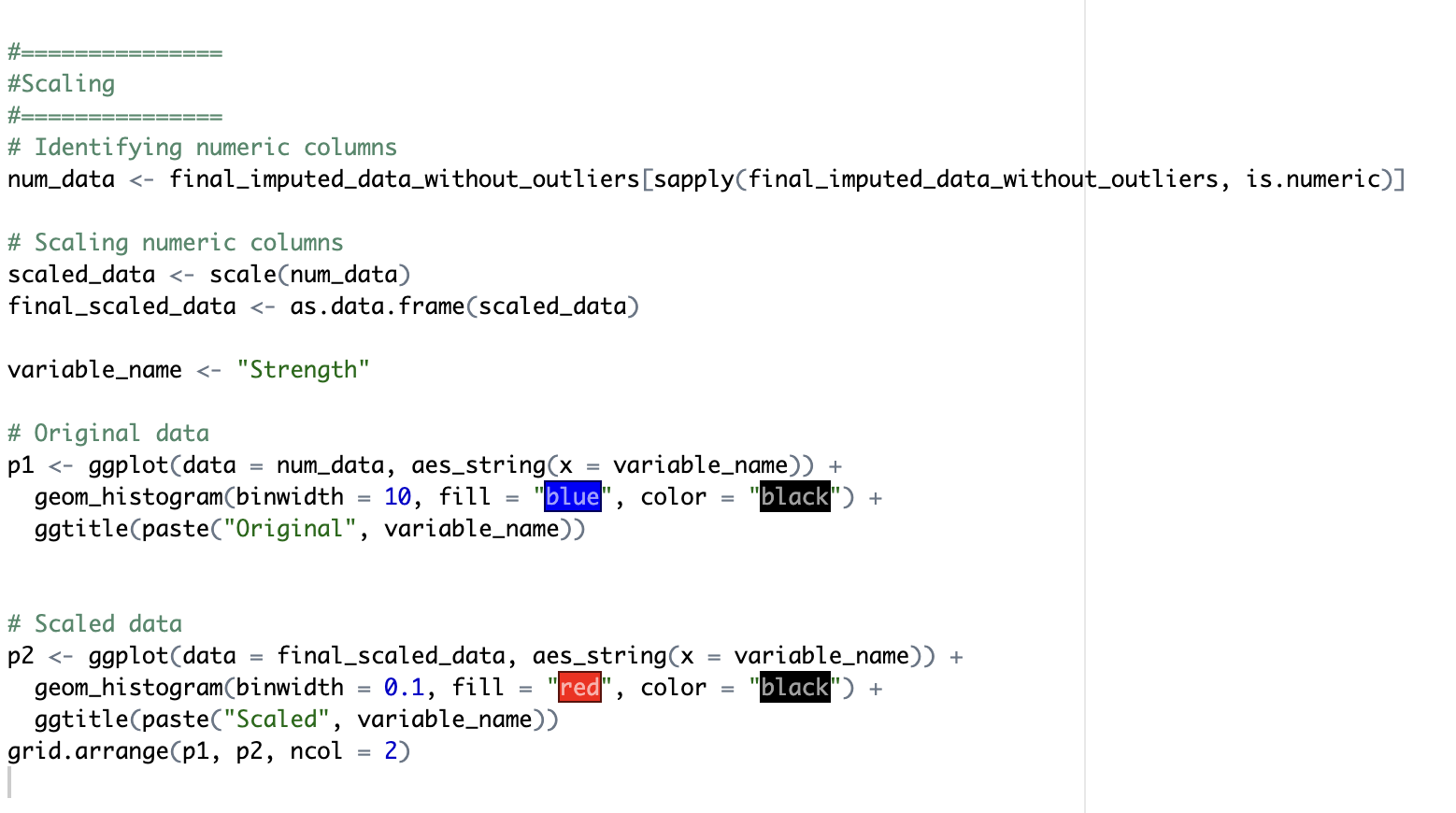
****

Figure 1.9.1 Code Snippet of Impact of Scaling on the target variable ‘Strength’

The above code in fig 1.9. compares the distribution of the "Strength" variable before and after scaling. Scaling may dramatically alter the shape and spread of a variable's distribution. Visualising this helps us grasp how scale impacts the data.

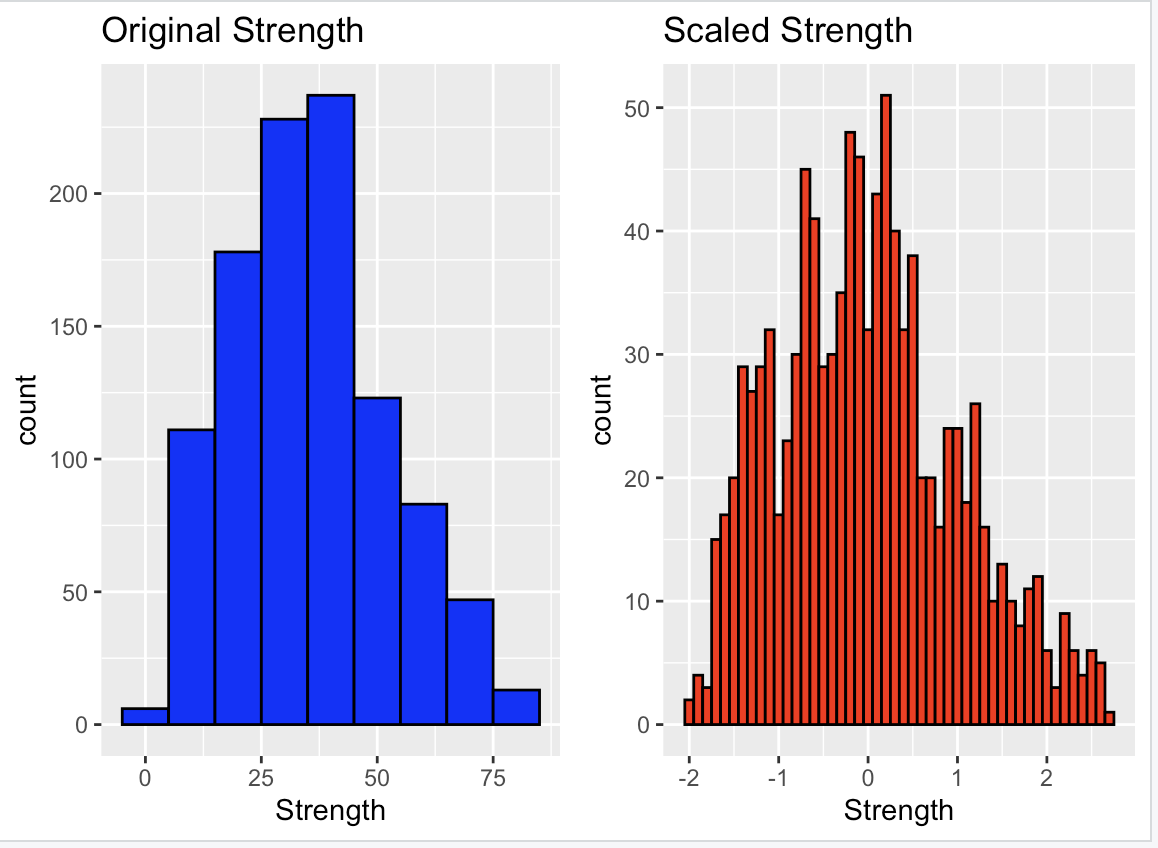
****

Figure 1.9.2 gg Plot of Original and Scaled value of Strength

Figure 1.9.2 shows the histogram of Original and Scaled value of Strength variable before and after scaling.

* **Bibliography**

1. Rahman, S. (2020). *Analysis on Compressive Strength of Concrete Using Different Sources of Fine Aggregates*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/338548502\_Analysis\_on\_Compressive\_Strength\_of\_Concrete\_Using\_Different\_Sources\_of\_Fine\_Aggregates [Accessed 26 Mar. 2024].
2. Eisa, A., Aboul-Nour, L.A. and Mohamad, A. (2024). *Experimental and theoretical investigation on the bond strength between high-strength and lightweight concrete*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/377993550\_Experimental\_and\_theoretical\_investigation\_on\_the\_bond\_strength\_between\_high-strength\_and\_lightweight\_concrete [Accessed 26 Mar. 2024].
3. Sujeet Kumar Mahato and Kumar, A. (2024). *Statistical Analysis of Compressive Strength of Concrete*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/378149940\_Statistical\_Analysis\_of\_Compressive\_Strength\_of\_Concrete [Accessed 26 Mar. 2024].
4. Li, P., Zhang, Y., Gu, J. and Duan, S. (2024). *Prediction of compressive strength of concrete based on IABC-MLP algorithm*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/377308680\_Prediction\_of\_compressive\_strength\_of\_concrete\_based\_on\_IABC-MLP\_algorithm [Accessed 26 Mar. 2024].
5. Ziolkowski, P. and Maciej Niedostatkiewicz (2019). Machine Learning Techniques in Concrete Mix Design. *Materials*, [online] 12(8), pp.1256–1256. doi:https://doi.org/10.3390/ma12081256.
6. Gagg, C.R. (2014). *Cement and Concrete as an engineering material: an historic appraisal and case study analysis*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/260439461\_Cement\_and\_Concrete\_as\_an\_engineering\_material\_an\_historic\_appraisal\_and\_case\_study\_analysis [Accessed 26 Mar. 2024].
7. Kumar, D.P. and Biable, A. (2020). *Experimental Investigation on the Compressive Strength of Concrete With Different Sizes of Coarse Aggregate*. [online] Social Science Research Network. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3713640.
8. Wang, S. and Baxter, L. (2007). Comprehensive study of biomass fly ash in concrete: Strength, microscopy, kinetics and durability. *Fuel Processing Technology*, [online] 88(11-12), pp.1165–1170. doi:https://doi.org/10.1016/j.fuproc.2007.06.016.
9. S. Hasdemir, A. Tuğrul and M. Yılmaz (2016). The effect of natural sand composition on concrete strength. *Construction and Building Materials*, [online] 112, pp.940–948. doi:https://doi.org/10.1016/j.conbuildmat.2016.02.188.
10. Banarjee, R., Alam, M. and Ahmad, Z. (n.d.). *Study of Compressive Strength of Various Grades of Concrete using Different Sizes of Cubes*. [online] Available at: <https://www.ijert.org/research/study-of-compressive-strength-of-various-grades-of-concrete-using-different-sizes-of-cubes-IJERTV4IS070455.pdf>.
11. Concrete | Definition, Composition, Uses, Types, & Facts | Britannica. (2024). In: *Encyclopædia Britannica*. [online] Available at: https://www.britannica.com/technology/concrete-building-material [Accessed 31 Mar. 2024].