Introduction:

One of the most important and crucial building material in the construction industry is concrete. It is widely used as a foundation in producing magnificent structures because of its availability, cheap rate and flexibility of handling and giving shape to any desired form. Concrete is basically a mixture of aggregates and paste which is comprised of cement and water that hardens overtime. This mixture of fine and course aggregate is used in a variety of applications including basic foundations, waste water treatment facilities, parking structures, floor construction and exterior surfaces.

As a construction material, one of the greatest challenges in the civil engineering industry is predicting the compressive strength of concrete which is the resistance to failure of the concrete under compressive forces. While other design properties of concrete such as durability and stability might be important, strength is considered the most important one. It is necessary to study about predicting the compressive strength of concrete since it provides time for concrete form removal, project scheduling and quality control. For designers and engineers, this might help in the application of post-tensioning.

For many years, researchers have proposed various methods for predicting the strength of concrete but these methods take into account only a limited number of data and parameters. For example, prediction models developed using the maturity concept which uses a fixed equation is based on a limited number of parameters. In contrast, Artificial Neural Networks (ANN) are increasingly being used for predicting the strength of concrete. This is due to the fact that ANNs work well with non-linear data and do not require a fixed equation. They can also continuously retrain on the new data to adapt to the new data conveniently by tweaking weights inside the neurons. There are a variety of parameters that may or may not influence the strength of concrete, such as cement, slag, water and coarse aggregate. Since we are not sure which independent variable is most important, a neural network would do the work for us by extracting patterns from within our data. Therefore a clear understanding of how concrete strength is impacted by these variables is vital in order to use it in various engineered structures.

In this paper, an attempt has been made to develop a relationship between concrete strength and its constituents and use a neural network model to predict its compressive strength.

Objectives:

The objective of this paper is to develop an Artificial Neural Network which includes a number of parameters like cement, slag, fly ash, water, super plasticizer, coarse aggregate, fine aggregate and age to predict the compressive strength of concrete.

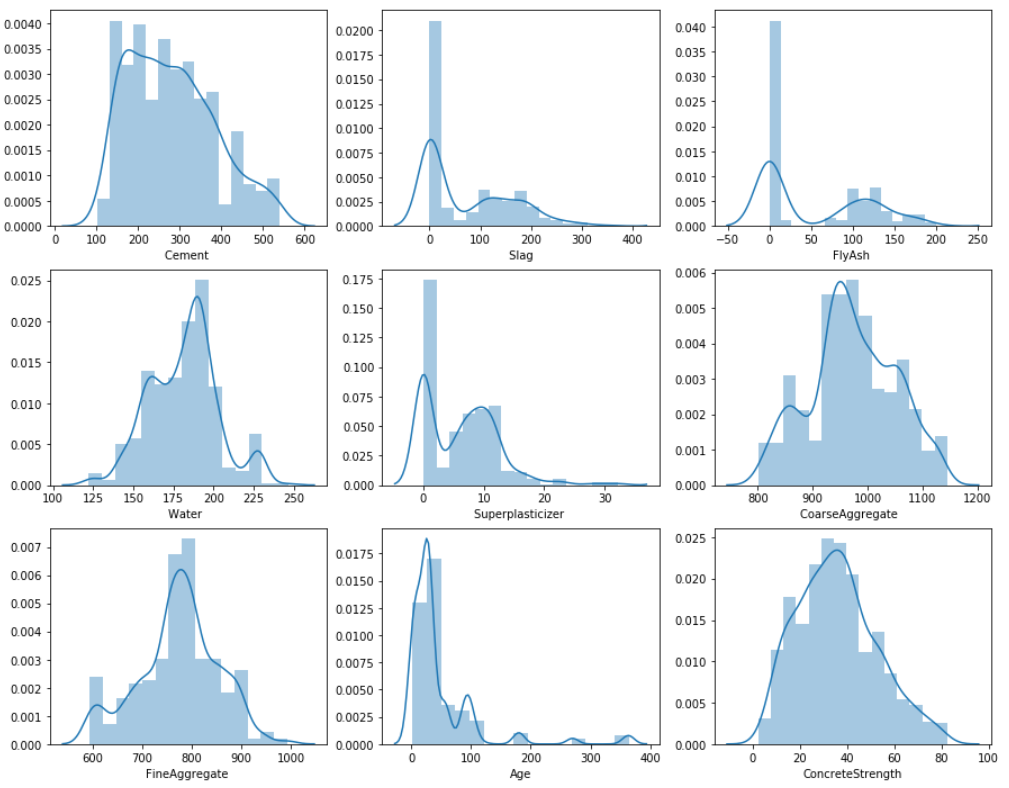
Problem statements:

The research project aims at determining an ample and cost-effective material to make up concrete. This will be done by using an examining various factors that impact the compressive strength of concrete.

Data Analysis and Methodology:

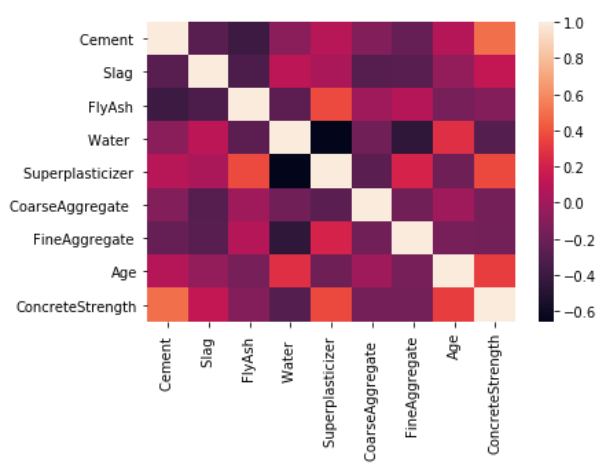
To get a clear picture about the dataset at hand, we started with some exploratory data analysis. Table 1.1 describes all the variables in the dataset with concrete strength also included. Since the count is the same for all of the variables, it can be realized that there were no missing values in the dataset. Secondly, Cement was seen as having the largest standard deviation amongst the rest which is further signified by the huge difference between its minimum and maximum values. Coarse aggregate was observed as having the highest mean in the dataset. Apart from this, we can infer that most of the dataset we were working with consisted of concrete with age around 36 years old.



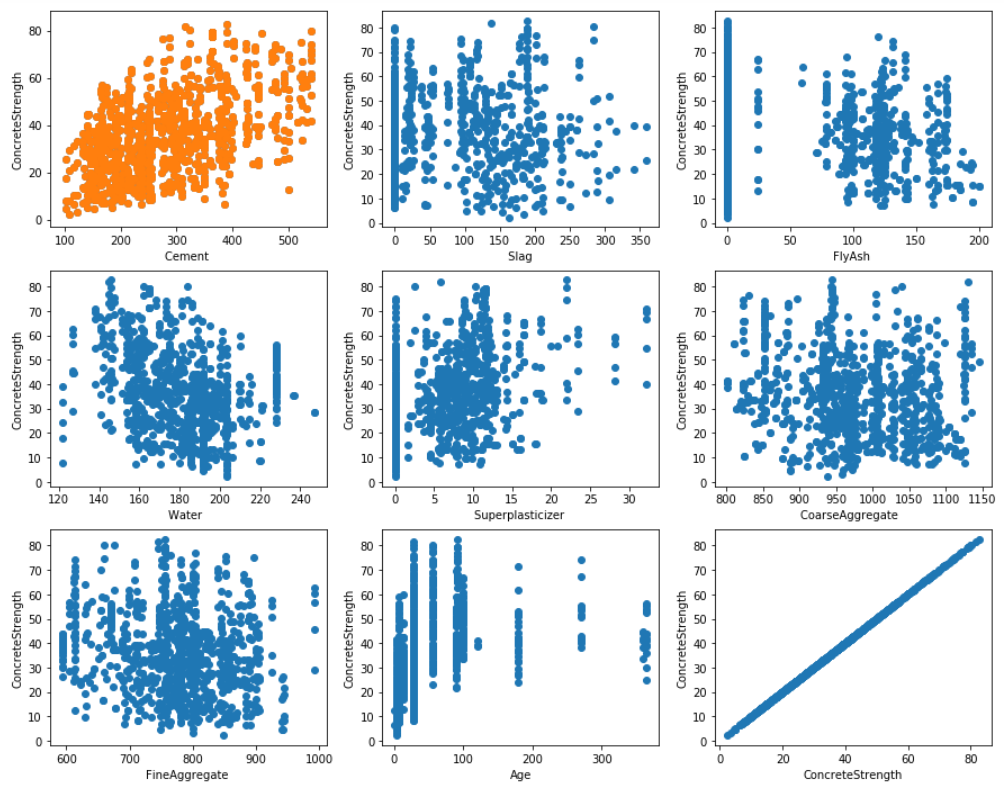


Next, we moved on to figure out the correlation between the independent (predictor) variables and dependent variable. Table 1.2 summarizes this by displaying how the correlation coefficient for each variable relates to the other. It is evident that cement has the highest positive correlation with concrete strength while water has the strongest negative correlation. To give direction to our analysis, we extracted four features with significant correlation coefficients against concrete strength, namely, Cement, Superplasticizer, Age and Water. The rest of the predictor variables have a very poor correlation with our dependent variable.



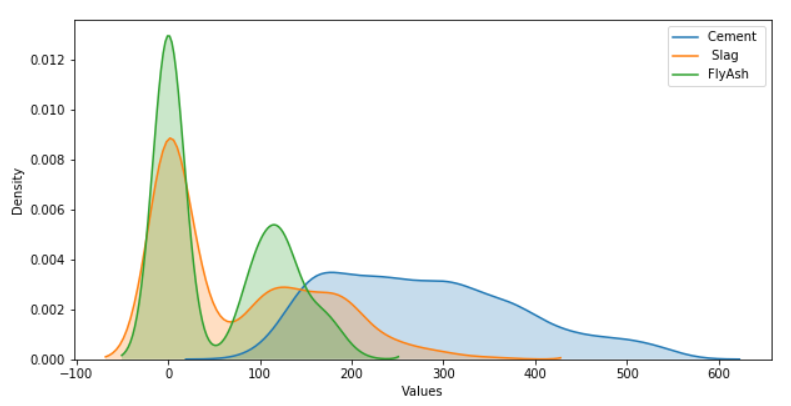


To further confirm and reinforce the information we derived from the correlation table, we used scatter plots to help us visualize exactly how each predictor variable impacts the compressive strength of concrete. We see that in figure 1.2 that Cement, as realized earlier, had the strongest positive correlation. However, no obvious relationship can be seen for the plots of Age and Superplasticizer against concrete strength. Furthermore, the remaining features also show no significant or obvious relationship. Since, concrete strength with itself has a correlation coefficient of 1.0, it has a straight line passing through the origin.



To further explore our data, we performed some more univariate analysis by building box plots for each variable as shown in figure 1.3. For Cement, we can see that the box moves towards the lower whisker. This suggests that most of the Cement values lie somewhere around 250 kg/m^3. As for Slag and Fly Ash, high density of the dataset lies around the zero point as also depicted by the kernel density estimate plot in fig 1.4.





Likewise, we also performed box plot visualization on Course aggregate and Fine aggregate variables to get a sense of how they are distributed in the dataset. It is can be seen in figure 1.5 that Course aggregate has a lower whisker at 800 and upper whisker above 1100 with the median lying at roughly 950. It seems to be normally distributed since the box is not moving to either of the two whiskers. Similarly, Fine aggregate has lower and upper whiskers at 600 and 1000 respectively with the median touching nearly 800. Lastly, the box plots of two crucial variables, Age and Concrete strength can be seen in fig 1.6. The age box is moving towards the lower whisker indicating that most of the concrete age is close to zero. However, we can notice a huge standard deviation since the upper whisker is at a point above 350. Conversely, concrete strength seems to be normally distributed with a box right in between the two whiskers with median around 40.

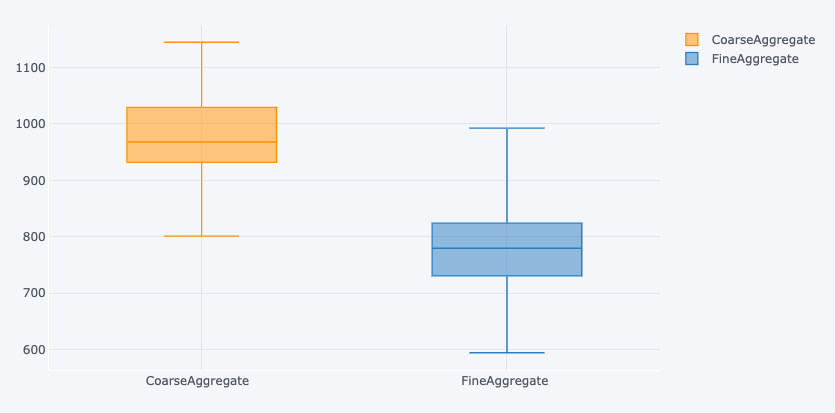
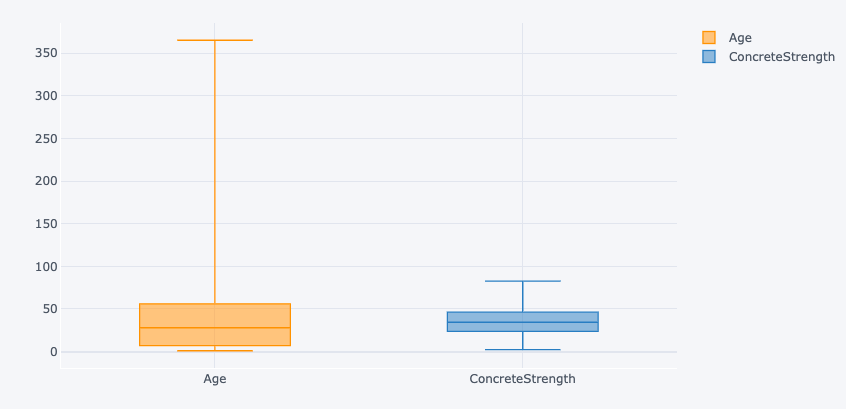
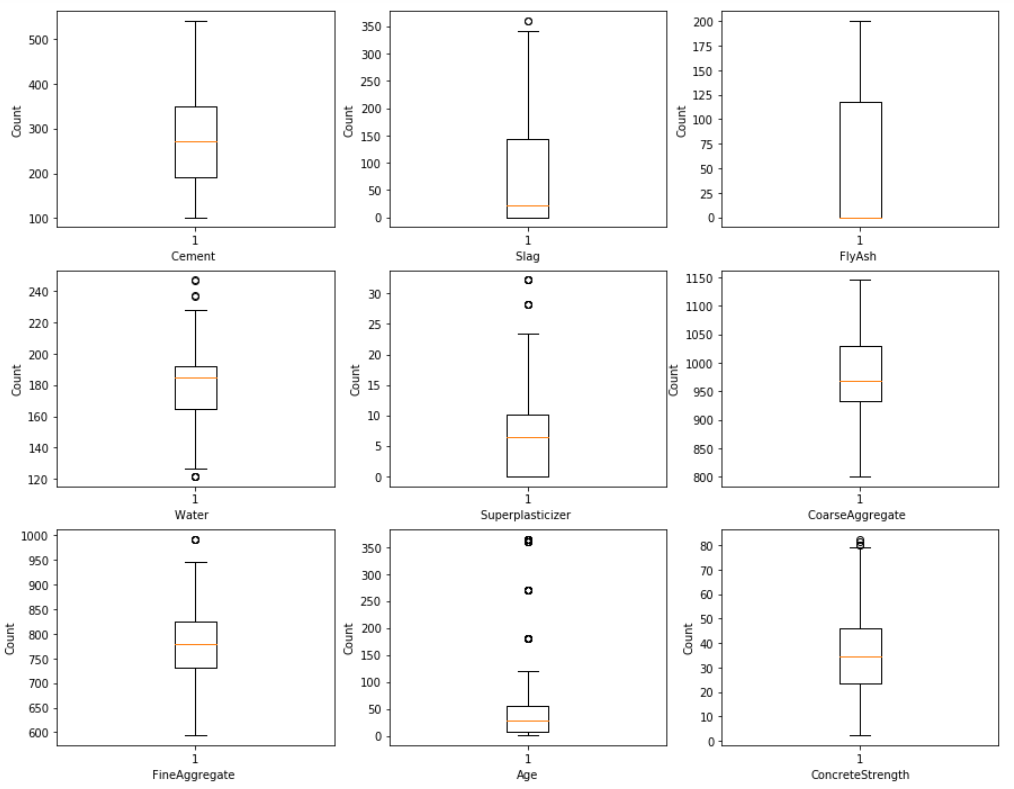
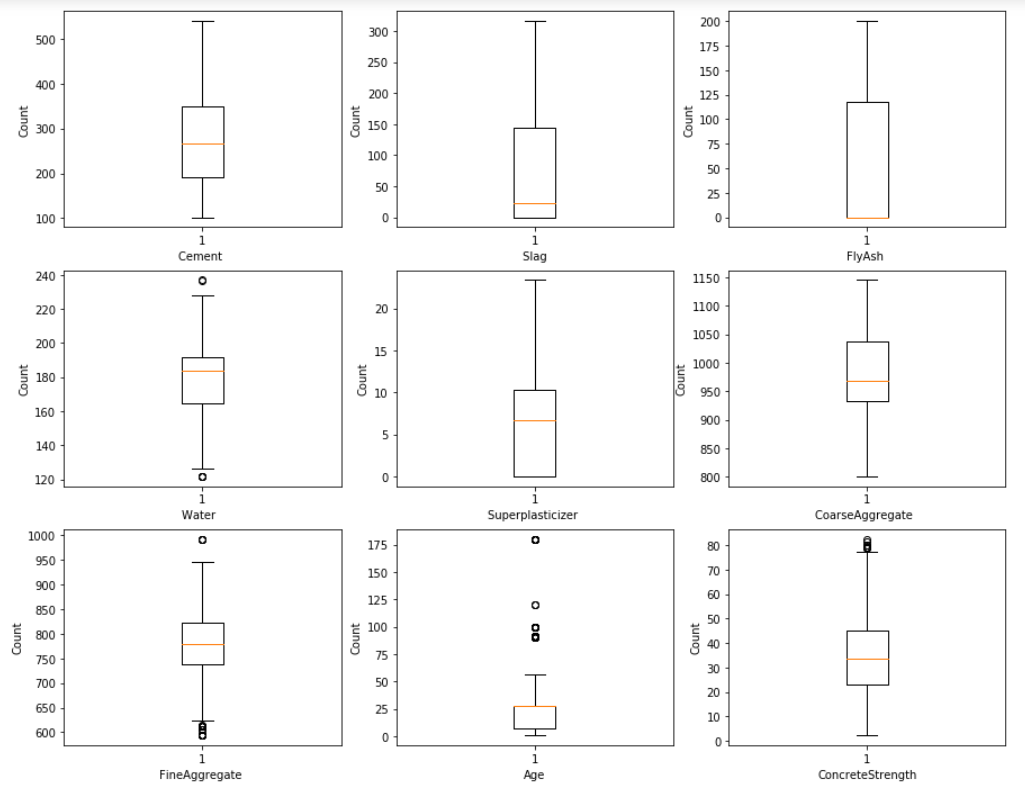
 

Figure 1.7: Box plot for Age and Concrete strength

Figure 1.6: Box plot for Coarse and fine aggregate

Our next step was to determine if any outliers were present which would skew our data. Again, the best approach to figuring this out is by the boxplot method. Below figure 1.7 shows the box plots of all the nine variables present in our dataset. We can see that Age variable has the most number of outliers present. This was also realized earlier in fig 1.6 as age had a huge difference between the upper and lower whiskers. Other than this, evident outlier data points can be observed in Slag (1 points), Water (2 points), Superplasticizer (2 points), Fine aggregate (one point), and Concrete strength (2 points). To prevent these outliers from influencing our data and model, we filtered out some of these data points using the z-score method. A threshold of 3 was used and anything data points having a z score of above 3 were filtered out from our dataset. Figure 1.8 shows the boxplots of the same nine variables after applying the z-score method. It is visible that outliers have been significantly reduced in some of the variables. For example, slag and superplasticizer have no outliers at all now.





Our next step in the data preprocessing was to normalize the data using the MinMaxScaler. We normalized our data between 0 and 1 so that all variables would be equally accounted for by our model without having any sort of bias.