**CS6444**

**Big Data and Analytics Spring 2022**

**Class Project #2**

There are several packages in R that allow us, to work with excel files. After some research and following some tutorials, we decided to use package named xlsx to load data from .xlsx file.



As names of the column were too long in original data, we decided to rename them.

Text

Description automatically generated

R has built in functions that can help us get sense of data quickly. For example, **str,summary**

Text

Description automatically generated

We can see that; all columns are numeric which is fine and there is no need to convert them.

It is totally possible that, we have NA values in data, however, they are useless and do not hold any value to our analysis.



To dig deeper into data and make sense out of it can cannot just rely on basic **str** and **summary** functions. Fortunately, it is really easy to visualize data in R.



Diagram

Description automatically generated

To understand, which of the columns does indeed have an impact on result we have to see correlation matrix. But before that we need to normalize columns to have standard view across all of the attributes. We decided to use min-max method for that purpose.



Graphical user interface, text

Description automatically generated with medium confidence

Chart, bubble chart

Description automatically generated

According to the plot some attributes are not significantly connected with the major variable c\_strength. We can assume that **blast\_furnance** ,**fly ash, coarse\_agg and fine\_agg** **are not** relevant for our analysis. We will remove them from our data.

Text

Description automatically generated

**K-Means**

The primary principle underlying cluster partitioning methods like k-means clustering is to define clusters with the least amount of intra-cluster variance (sometimes called total within-cluster variation or total within-cluster sum of square).

The elbow approach is a heuristic for figuring out how many clusters there are in a data collection. Plotting the explained variation as a function of the number of clusters and selecting the elbow of the curve as the number of clusters to utilize is the method.

Text

Description automatically generated

Chart, line chart

Description automatically generated

According, to the plot optimal value of clusters is 5.