DATA WRANGLING

Capstone project – 1

*Title:*

Graduate Admission 2 - Predicting admission from important parameters

Cleaning steps performed:

Inconsistent column names:

Two column names are inconsistent. It has a space ‘LOR ‘ and Chance of Admit ‘

I have removed the space by renaming the columns as below

df.rename(columns = {'Chance of Admit ':'Chance of Admit'}, inplace = True),

df.rename(columns = {'LOR ':'LOR'}, inplace = True)

Missing values:

There are no missing values

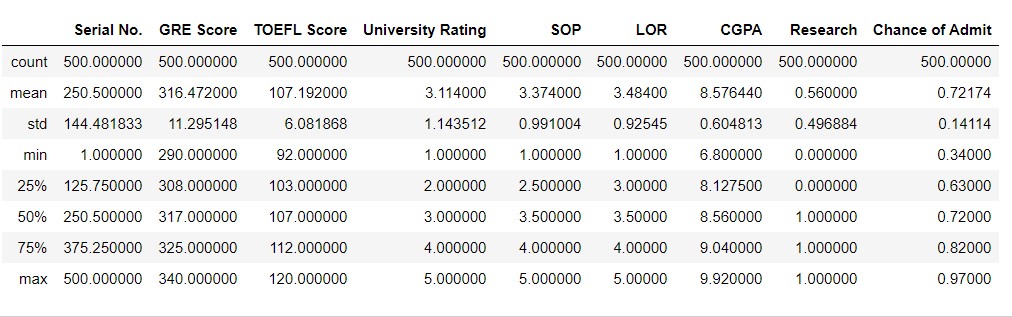
This was checked with df.info() method

Outliers:

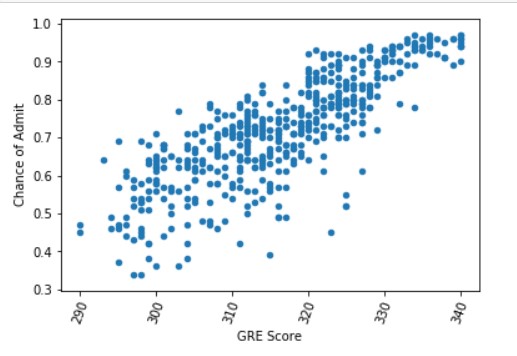
There are no outliers

This was checked by comparing aggregate mean and max values

Obtained by df.describe()



This was also checked by scatter plot



There is a linear relation between 'GRE Score' and 'Chance of Admit'

I have taken the data and then did some cleaning on it. Like I observed for missing values,

inconsistent column names, and for any outliers from the data set named 'df' .

Then I named the cleaned dataset as df\_cleaned. Later I have done with the Exploratory Data Analysis

along with some visualization to check the relation between all the fields with the "Chance of Admit" column to predict some facts if anything.

""Visualizations are GOOD ENOUGH to confirm that the relation with all the fields with "Chance of Admit" is linear or directly proportional.

Finally, I have come up with those records that who has highest chance of admission and have lowest chance of admission