

**Report on**

**Penguin Datasets EDA & its Predictive Model**

**Course Title: Bigdata Analysis LAB**

**Course Code: CSE 4460**

**Submitted by:**

**Md Abrar Saief Safat, 203014020**

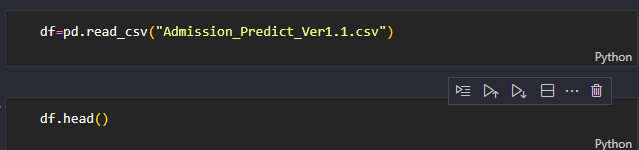
**Submitted to:**

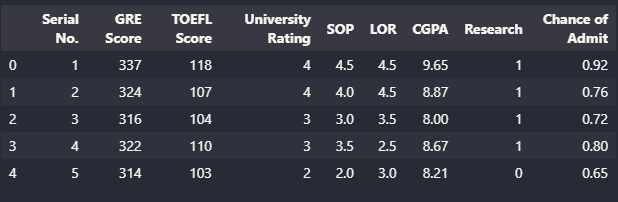
**Fahim Morshed**

**Visiting Lecturer**

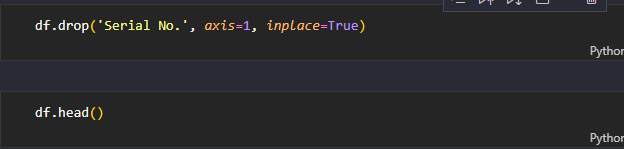
**School of Science & Engineering**

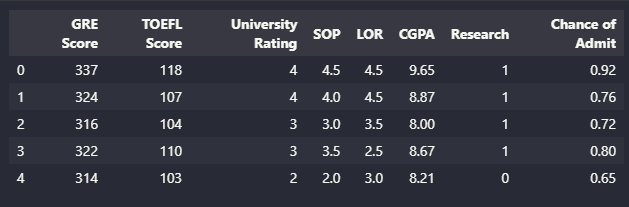
**Loading and displaying default format of dataset:**



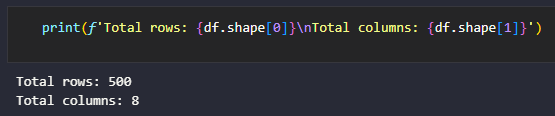


**Dropping ‘Serial No.’ feature as it is deemed as unnecessary:**

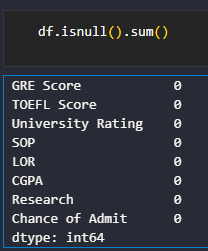
****

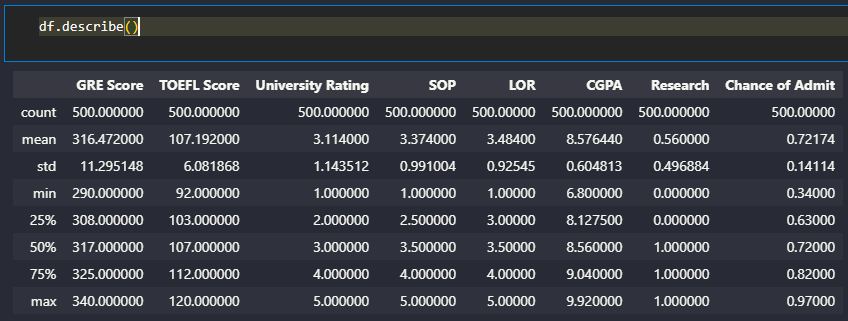
****

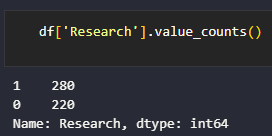
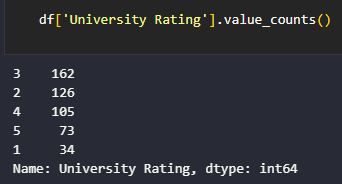
**Printing total dimension of dataset:**

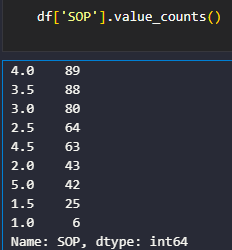
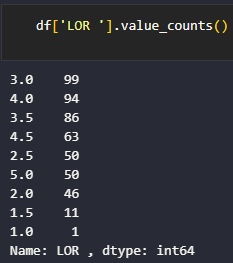
****

**Finding total null values in the dataset:**

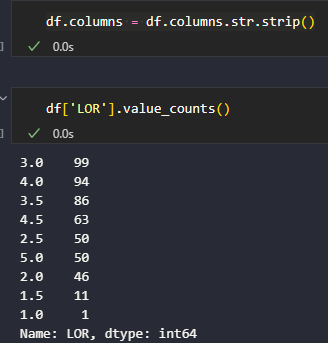
****

**Standard deviation details in the dataset:  
**

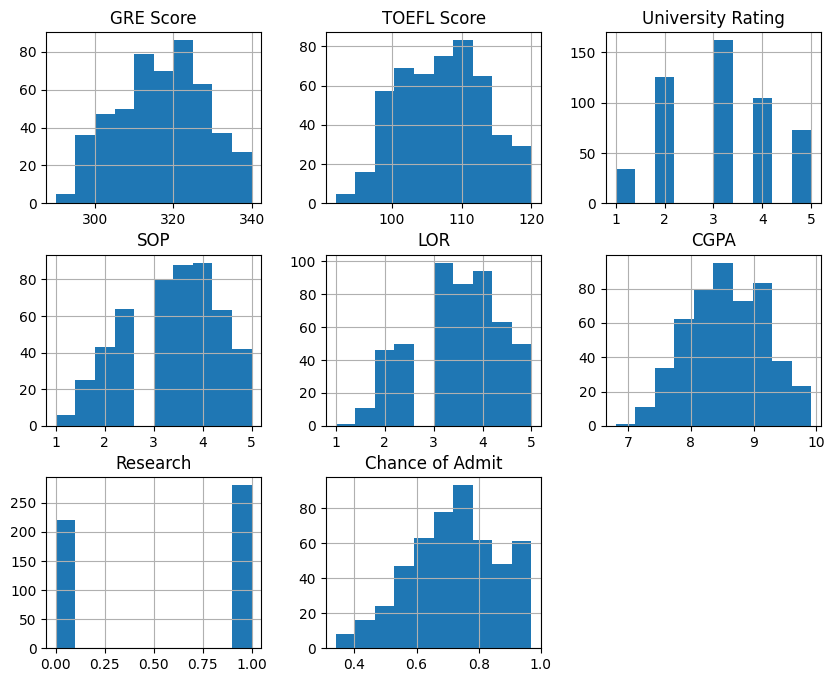
**Unique value calculation:  
**

****

There was a whitespace on the LOR feature, it later revealed as shown below:

****

**Histogram of features:**

****

The GRE score’s distribution is right skewed that indicates that most applicants have the GRE score in that range. There are a few applicants that are in the extremes.

The TOFL score distribution is also right skewed. Most clusters appear within 100 and 120 score which suggests that most applicants have a high TOEFL score.

The university rating’s distribution is showing that higher rated universities have fewer applicants which means that most applicants prefer applying to mid or lower tier universities.

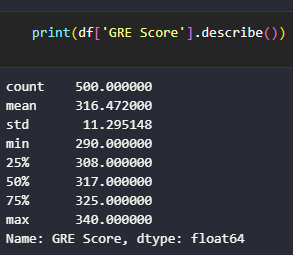
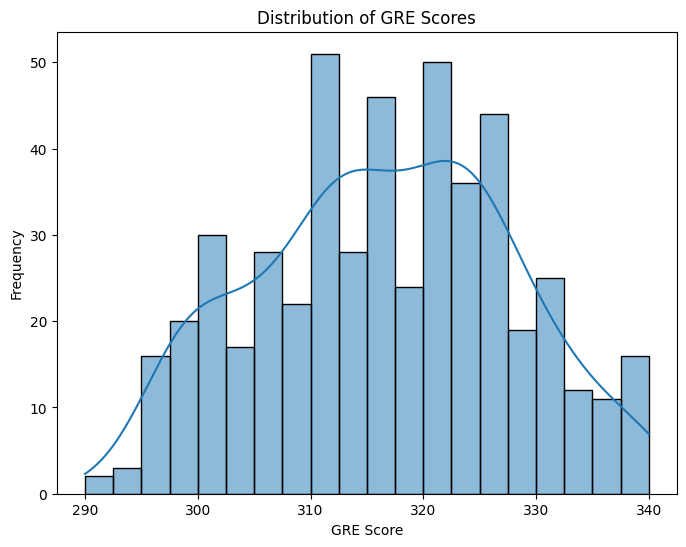
The SOP’s distribution is left skewed, this shows that most SOP’s submitted have moderately higher ratings.

LOR’s distribution is similar to SOP’s distribution.

CGPA’s distribution is right skewed, this indicates that most applicants have a high CGPA, however, there are a few instances/outliers.

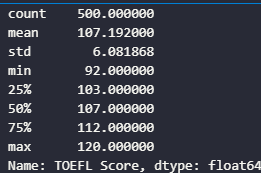
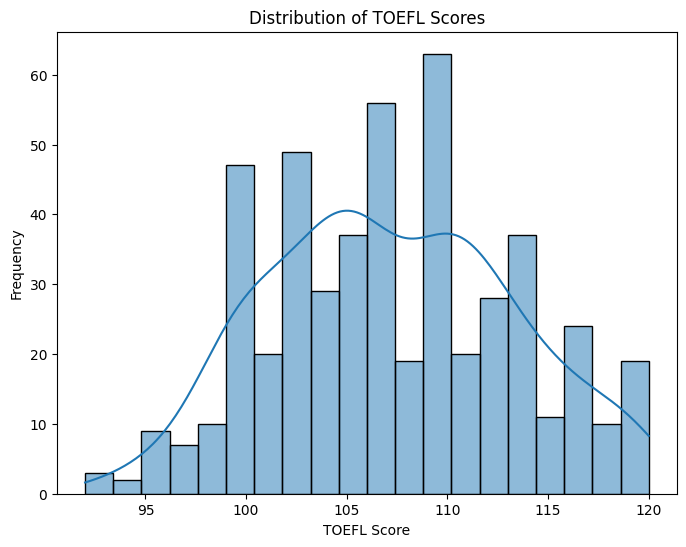
Research’s distribution is shows that the majority of applicants have research experience.

**Detailed distribution of GRE scores:**

****

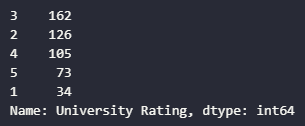
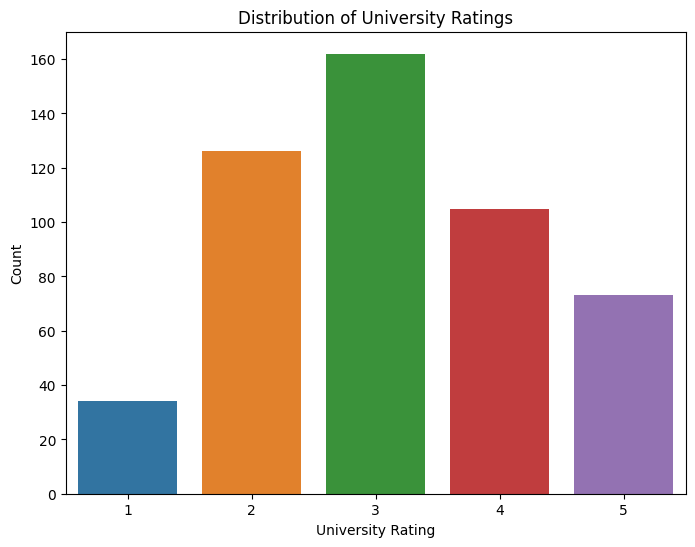
This shows that the majority peaks about 320-330. There were fewer applicants scoring below 300 or above 340 that suggests that the distribution is slightly skewed to the right.

**Detailed distribution of TOEFL scores:**

****

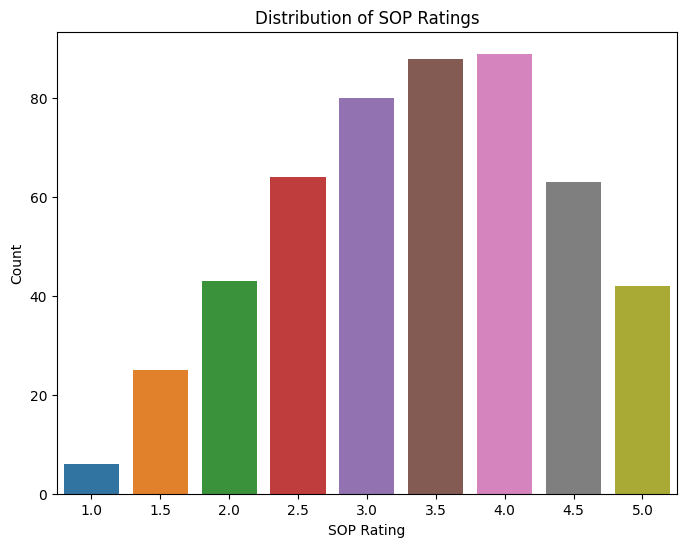
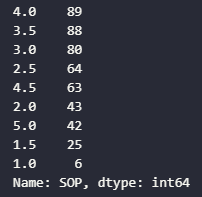
The peak of the distribution is around 110-115 that shows that most applicants have a TOEFL score in this range and it is similar to GRE scores. The distribution is somewhat normal with a small amount of skew due to the fewer amount of scores below 100 and above 120.

**Distribution of University Ratings:**

****

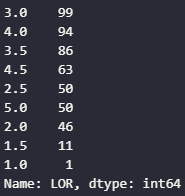
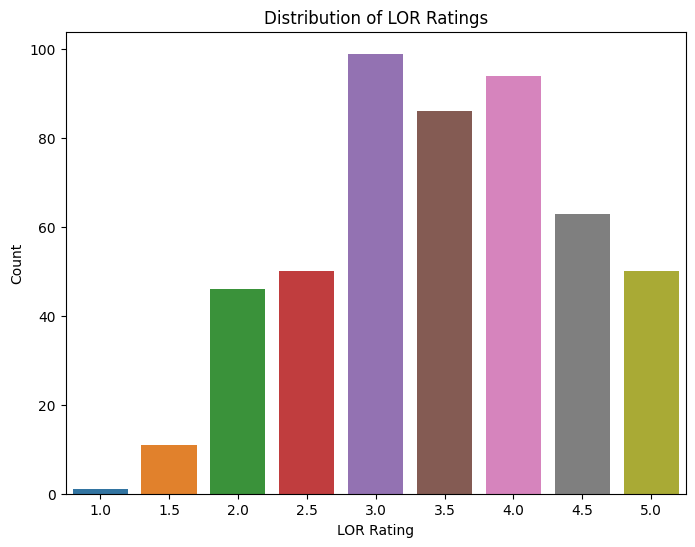
The distribution shows that universities with the rating of 3 has the highest number of applicants indicating a high preference for mid tier universities.

**Distribution of SOP Ratings:**

****

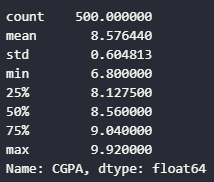
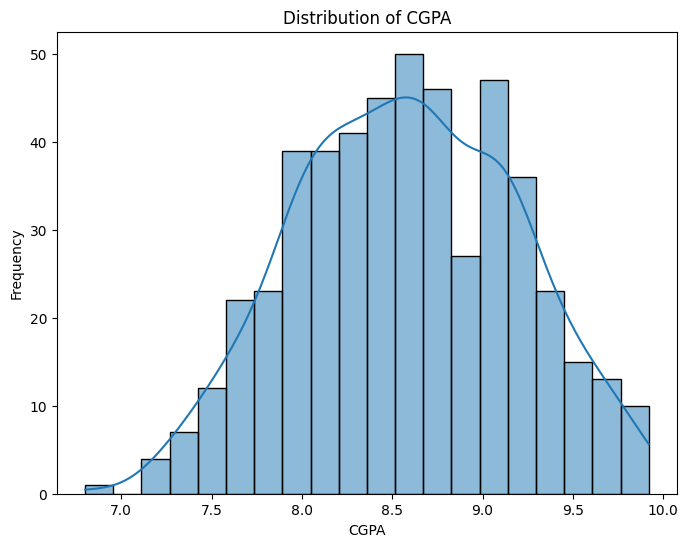
The distribution suggests that most applicants have a strong SOP rating as the distribution is left skewed.

**Distribution of LOR Ratings:**



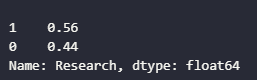
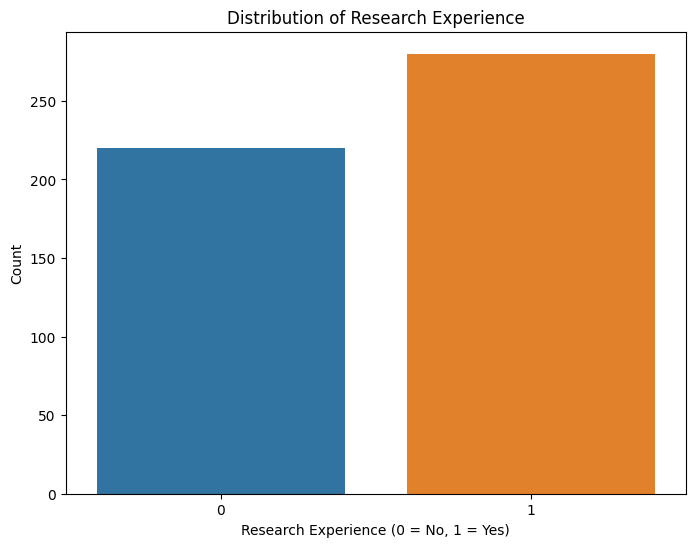
The LOR ratings distribution is also left skewed although there is a slight decrease at 4.5 and a gradual decrease at 5.0 and ratings below 2.0 are rare/outliers.

**Distribution of CGPA:**



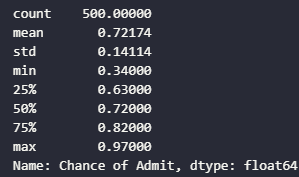
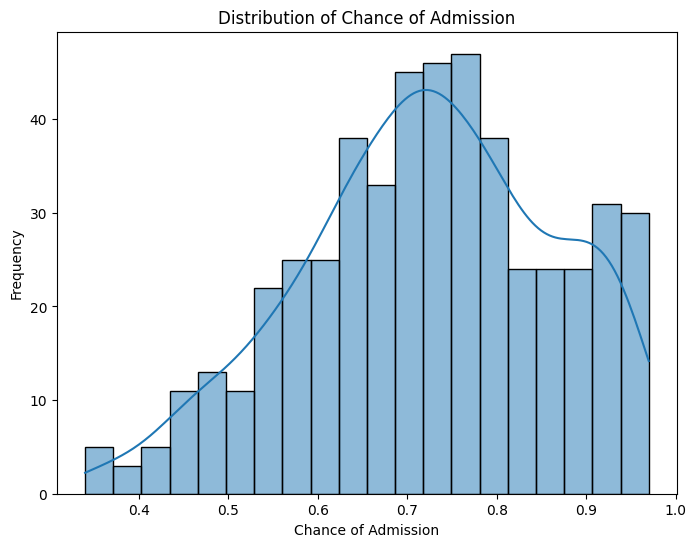
This distribution suggests that most applicants have a CGPA of 8.0 to 9.0. There are few applicants below 7.5 or above 9.5.

**Distribution of Research Experience:**

****

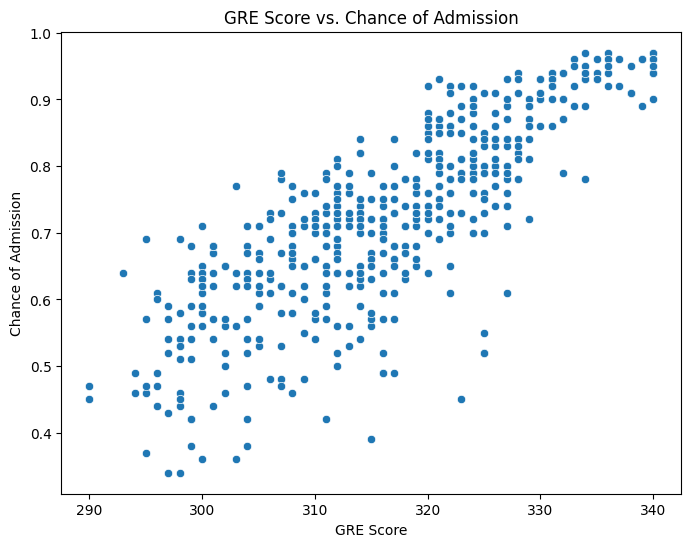
This binary data shows that most applicants come with research experience.

**Distribution of Chance of Admission:**

****

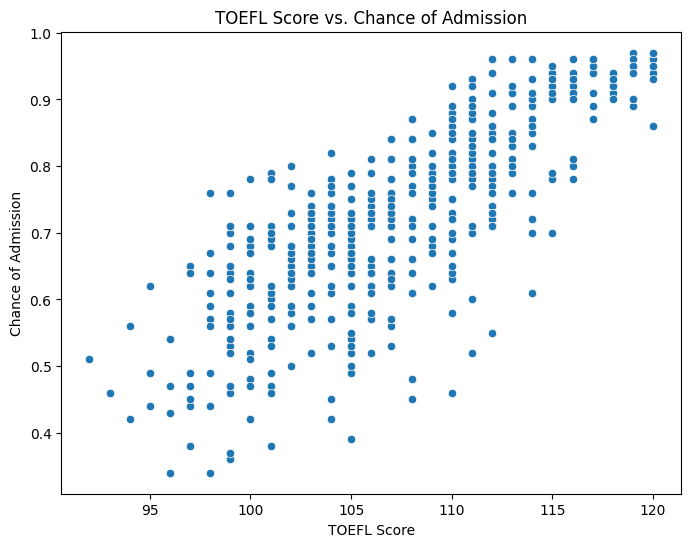
This distribution has a peak in around 0.7-0.8 and it shows that most applicants have a very high chance of admission as the data is skewed to the left.

**Scatter plot of GRE Score VS Chance of Admission:**

****

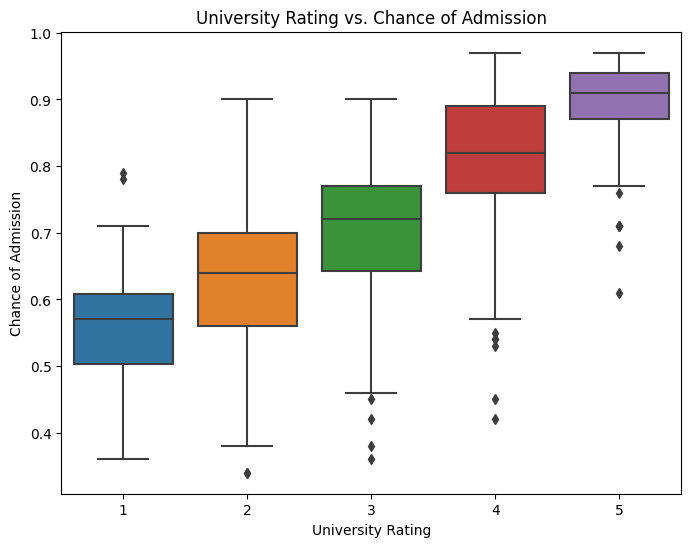
This graph shows that the correlation between the GRE Scores and the Chances of Admission are positive and they are highly connected. The scores above 320 has a higher chance of admission chances (although there are some variabilities in that conjecture).

**Scatter plot of TOEFL Scores VS Chance of Admission:**



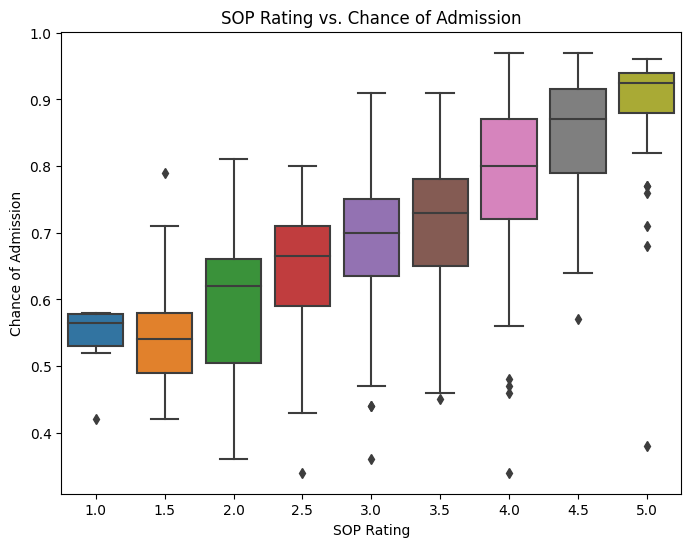
This positive correlation of TOEFL scores and Chance of Admission shows that applicants with a TOEFL score more than 110 usually have a higher chance of admission. There is a cluster between 105 and 115 that shows that they have a higher chance of admission with a little variability.

**Box Plot for University Rating VS Chance of Admission:**



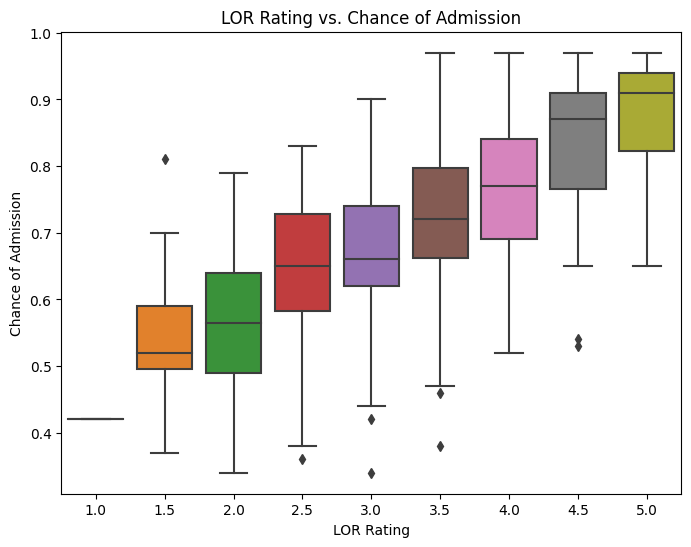
This box plot shows that the median of the chance of admission gradually increases the university’s rating. Low rated universities have higher standard deviation with wider interquartile ranges and a lower median chance of admission compared to high rated universities that show higher chances with some outliers.

**Box Plot for SOP Ratings VS Chance of Admission**

****

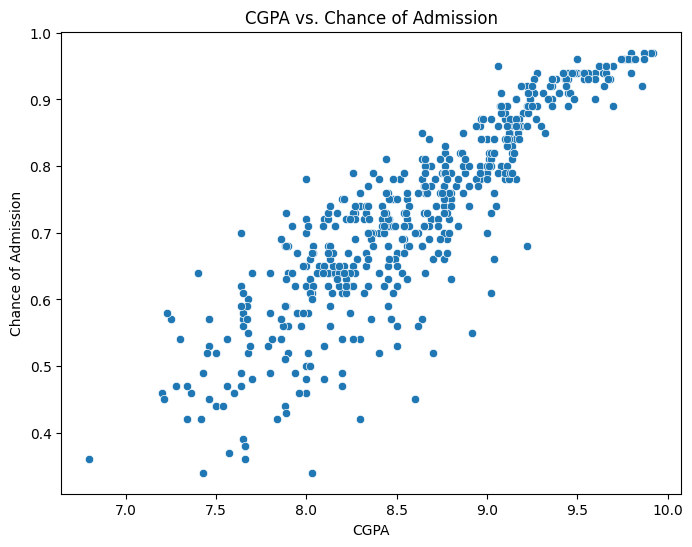
This box plot shows that the general increase in the median chance of admission as the SOP rating goes up, and the lower the SOP ratings the lower the median chances and a wider standard deviation of the feature.

**Box Plot for LOR Rating VS Chance of Admission:**



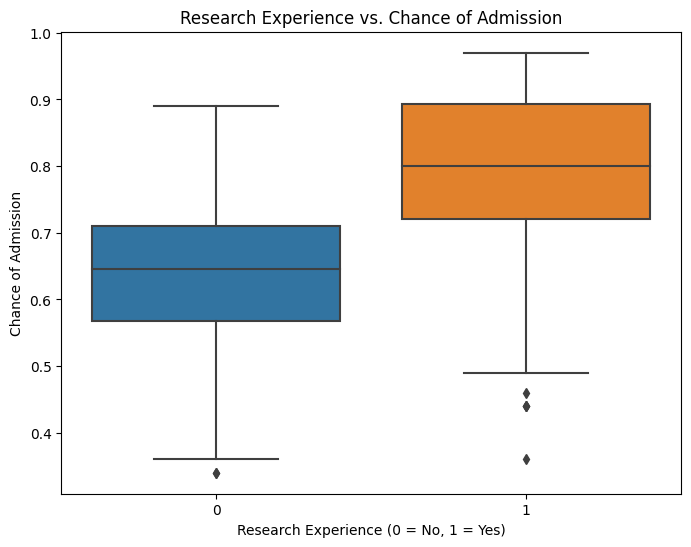
This box plot shows that the positive trend of the increase of the chance of admission with the increase in the LOR ratings. Lower LOR ratings have larger whiskers indicating a higher variety and outliers which higher rating give a better chance of admission.

**Scatter Plot for CGPA VS Chance of Admission**



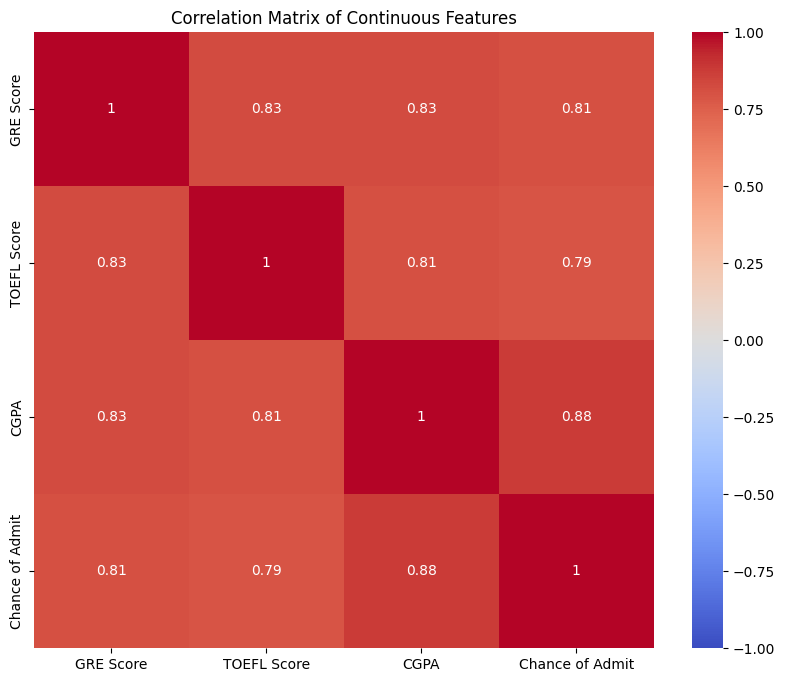
The scatterplot of CGPA and Chance of admission shows that they are closely associated and CGPA with 8.5+ have a higher chance of admission. The cluster of points exist between 8.0 and 9.5 CGPA that indicates the significant amount of chance of admission lies in that range.

**Box Plot for Research Experience VS Chance of Admission:**



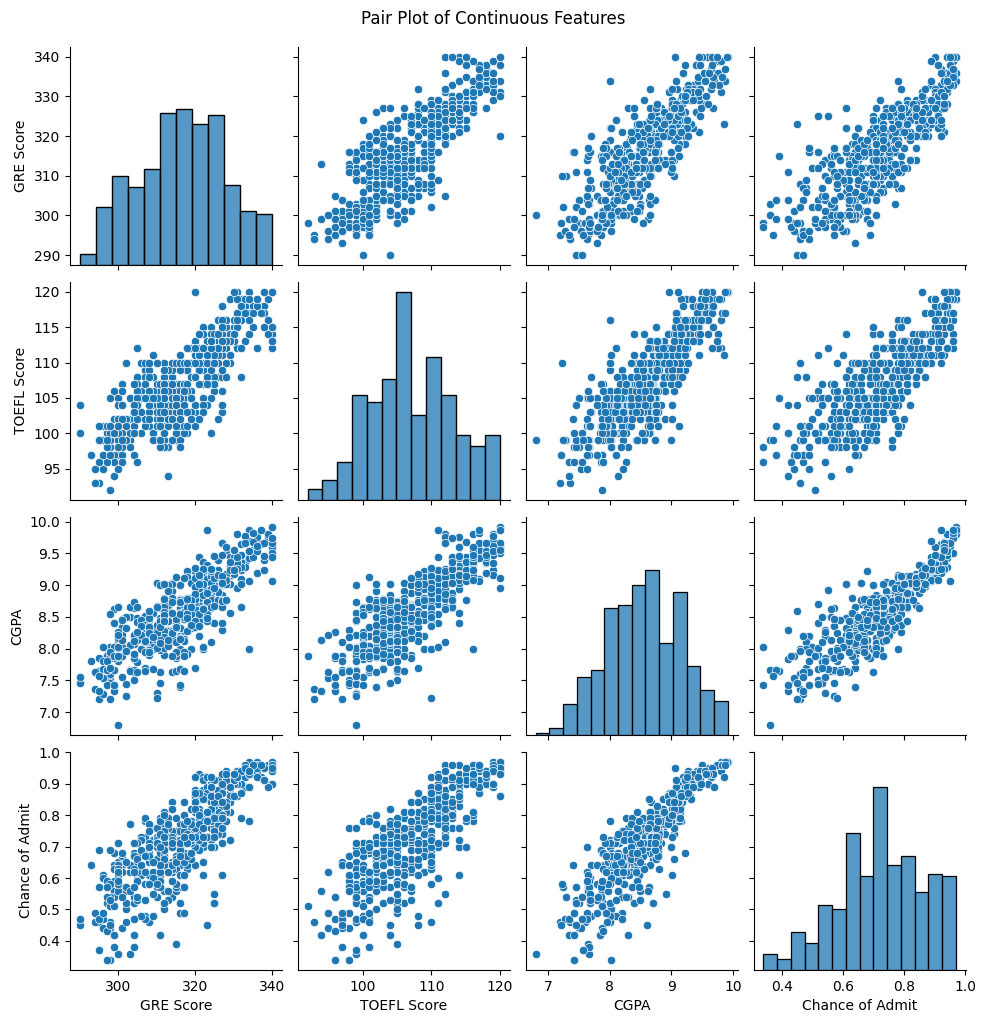
The box plot depicts the applicants with research experience have a higher median chance of admission compared to the applicants without the research experience. However, both groups have a similar interquartile range, but the applicants with research experience have an edge.

**Heatmap Correlation of Continuous Features:**

****

All the features existing in the dataset shows positive correlation between each other, with the strongest relations between CGPA and Chance of Admission, followed by GRE Score and TOEFL Score, this indicates the depending factors that are highly influential for Chances of Admission.

**Pair plot of Continuous Features:**

****

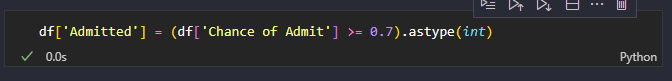
As explained in the correlation heatmap, the Pair plot confirms the correlations between all the features existing in the dataset. The density of samples in the scatterplot indicates that the higher the values of the dependent features, the higher the chances of admission (target feature).

**Conclusion:**

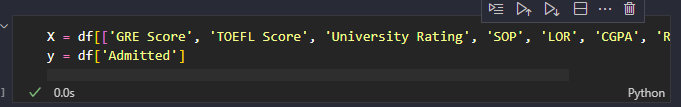
On the basis of the EDA conducted on the Graduate Admission dataset and its numerous trends as well as patterns, I have found that most applicants have met the base requirements as they bring academic excellence in multiple sectors: GRE, TOEFL, and CGPA. On the other hand, factors such as SOP, LOR and research experience also has positive influence in the chance of getting admission for the applicant.

**Implementation of Logistic regression as supervised predictive model:**

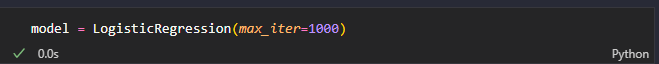
Converting the target variable into binary and putting a 0.7 threshold:



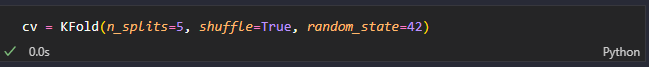
Declaring the features as X and target as y:



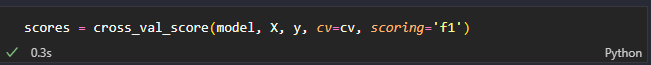
Logistic model initialization with increased max\_iter for convergence if needed:

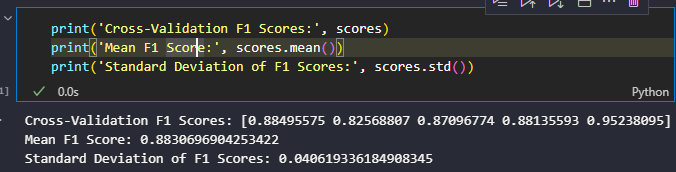


Setting up 5-fold cross validation for evaluation:



Performing cross validation along with f1 score as evaluation metric:



Result outputs:

The individual f1 scores shows that the model’s five fold performance is fairly good. However, the mean f1 score is approximately 88.3% which is good balance between precision and recall. The standard deviation of the results is 4.6% which shows stability in the results.