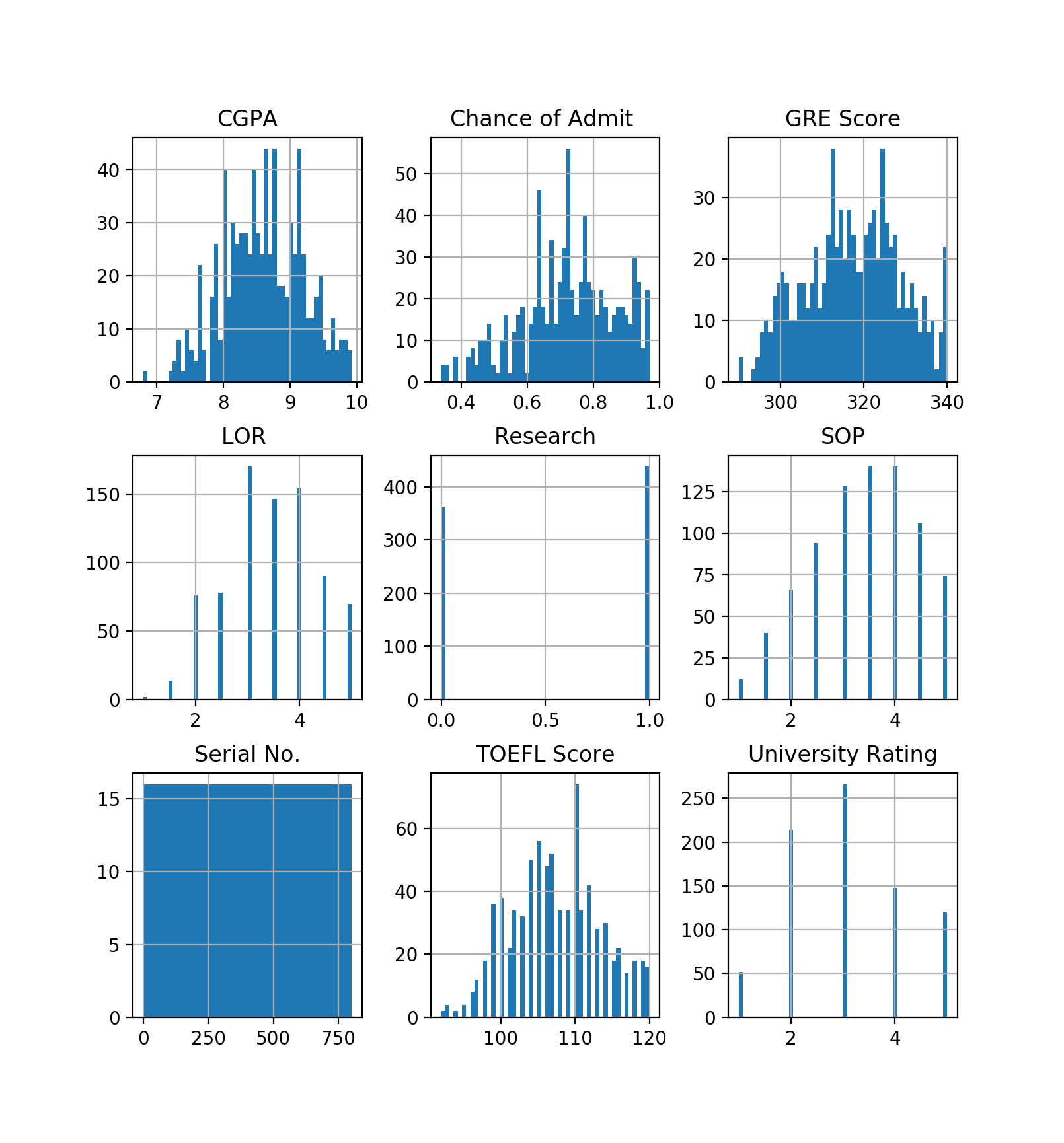
**Visualization**

For the data visualization part, we mainly used pandas dataframe, since pandas is a very useful package in data cleaning, modeling and exploration. It provides high performance and easy to use data structures in data analysis and data visualization and that’s why we choose pandas.

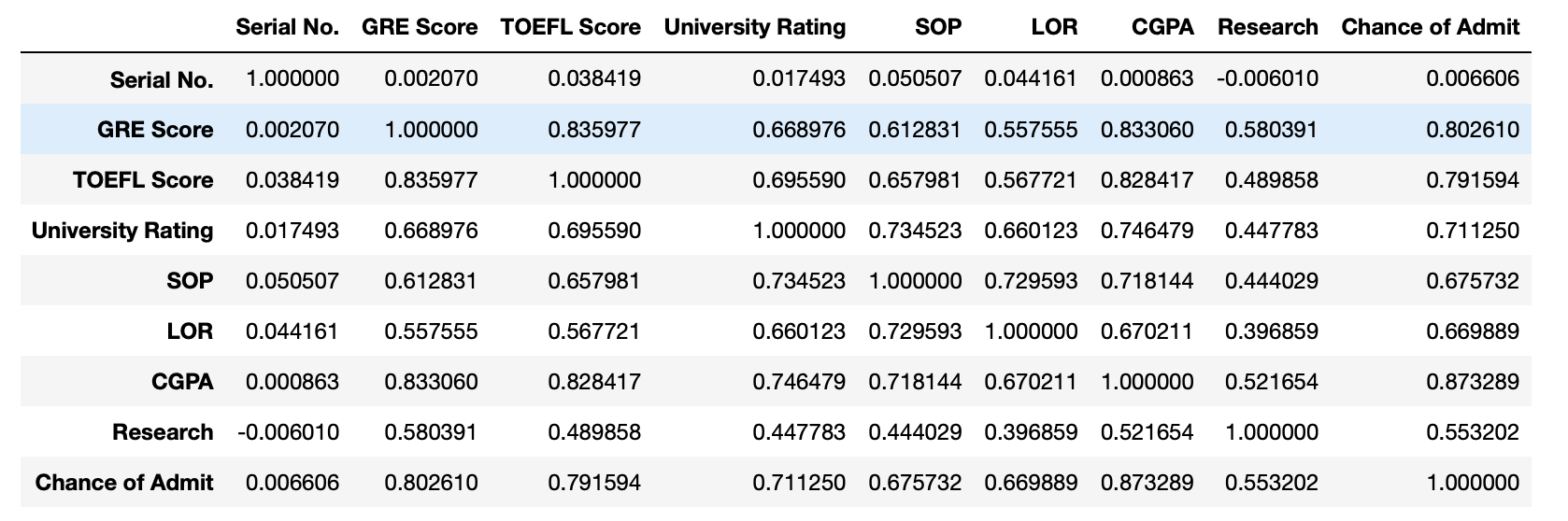
After reading the data, we made a histogram to have a basic idea about how the data looks like and the result plot is shown below:



From the graph, we can see that the dataset contains 3 continuous variables, which are GPA, GRE score and TOEFL score, and 4 categorical variables, which are LOR, Research, SOP and university ranking. The outcome, which is the chance of admission, is also a continuous variable.

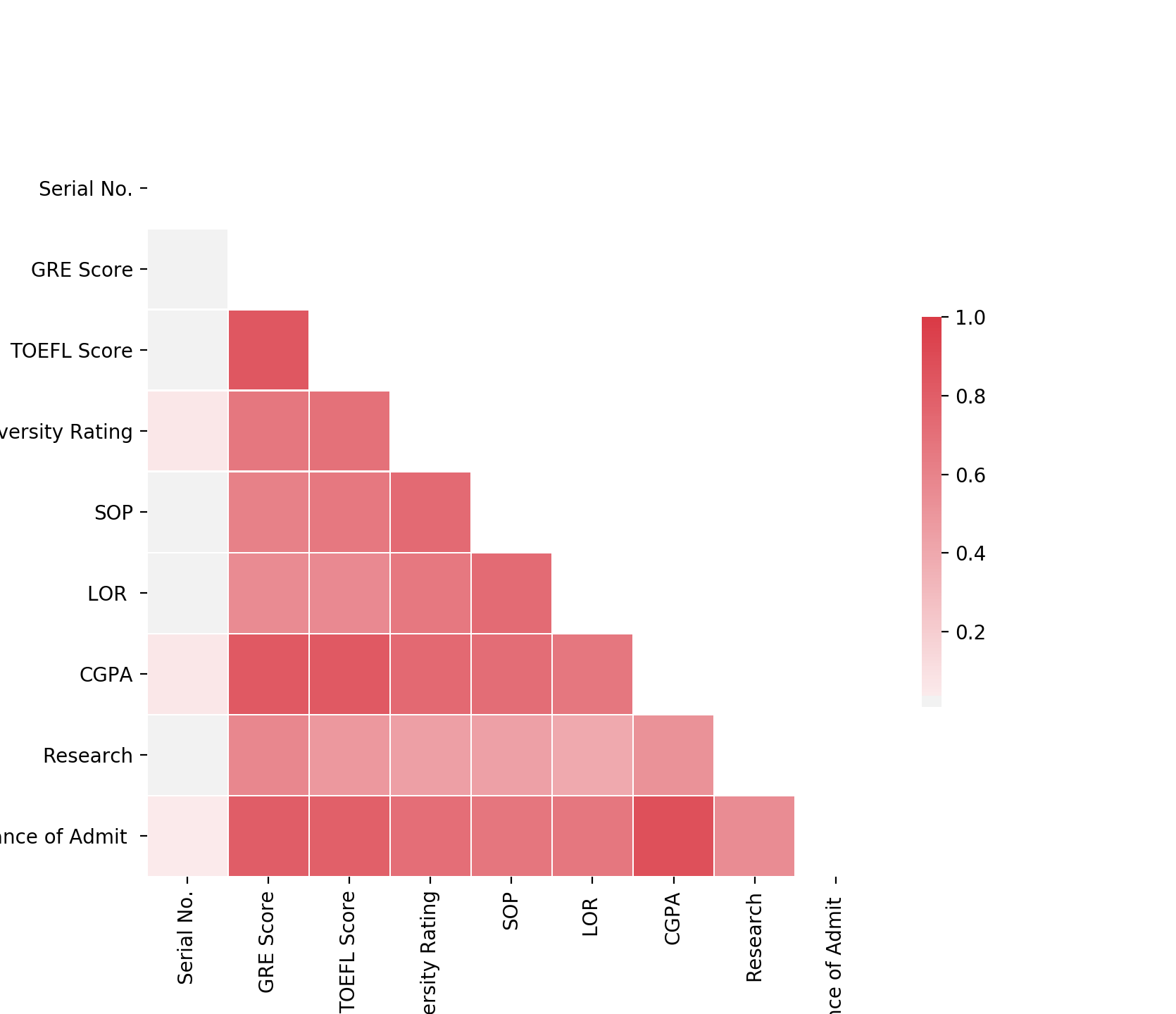
We could notice that the variable Serial No. only represents the case number of each student which does not contain useful information for the prediction model. Based on that, we should omit this variable in the further analysis. This will also be proved in the correlation analysis in the next steps.

After having a basic idea about how the data looks like, we calculate correlation matrix and the result is shown below.

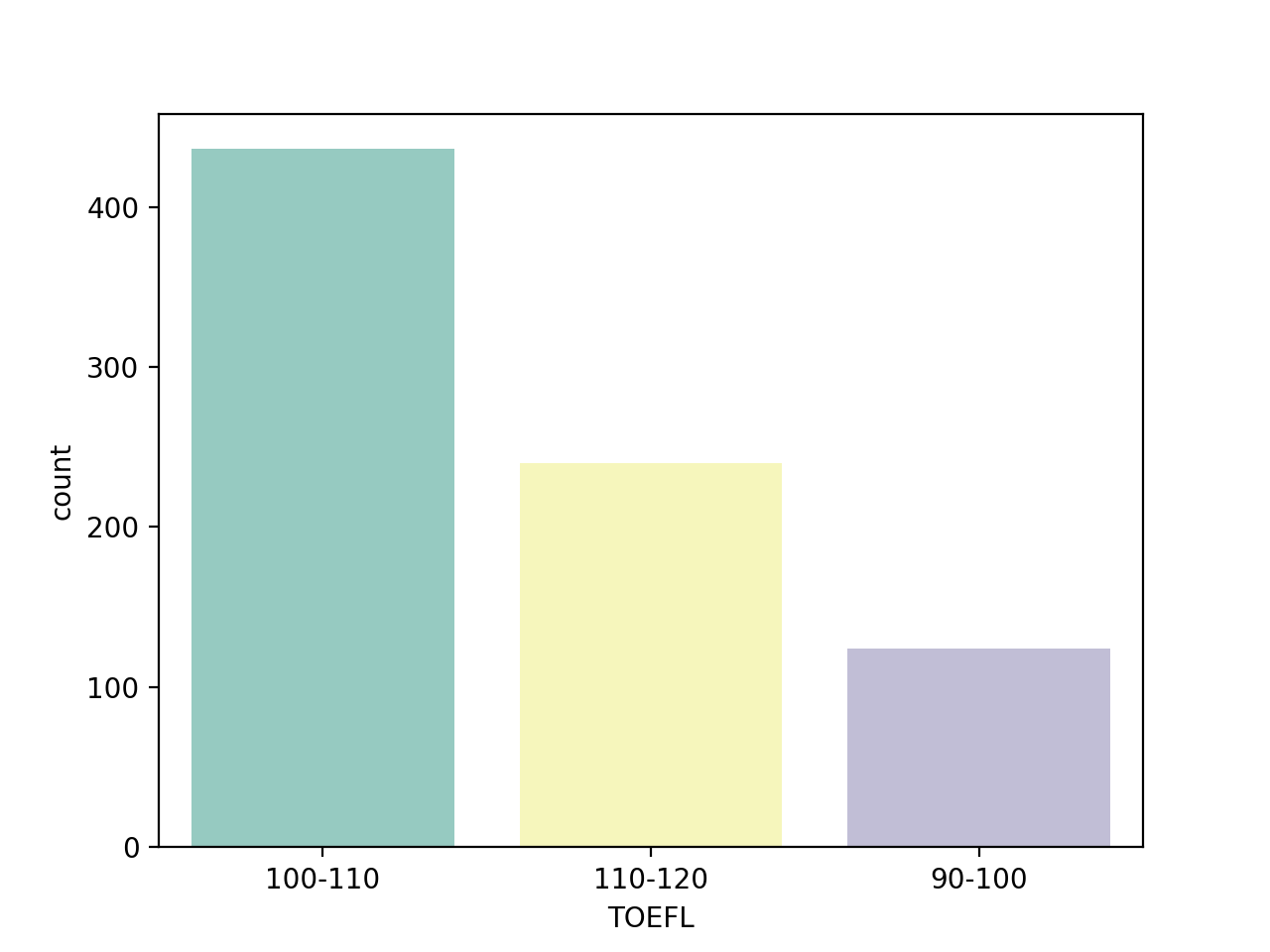
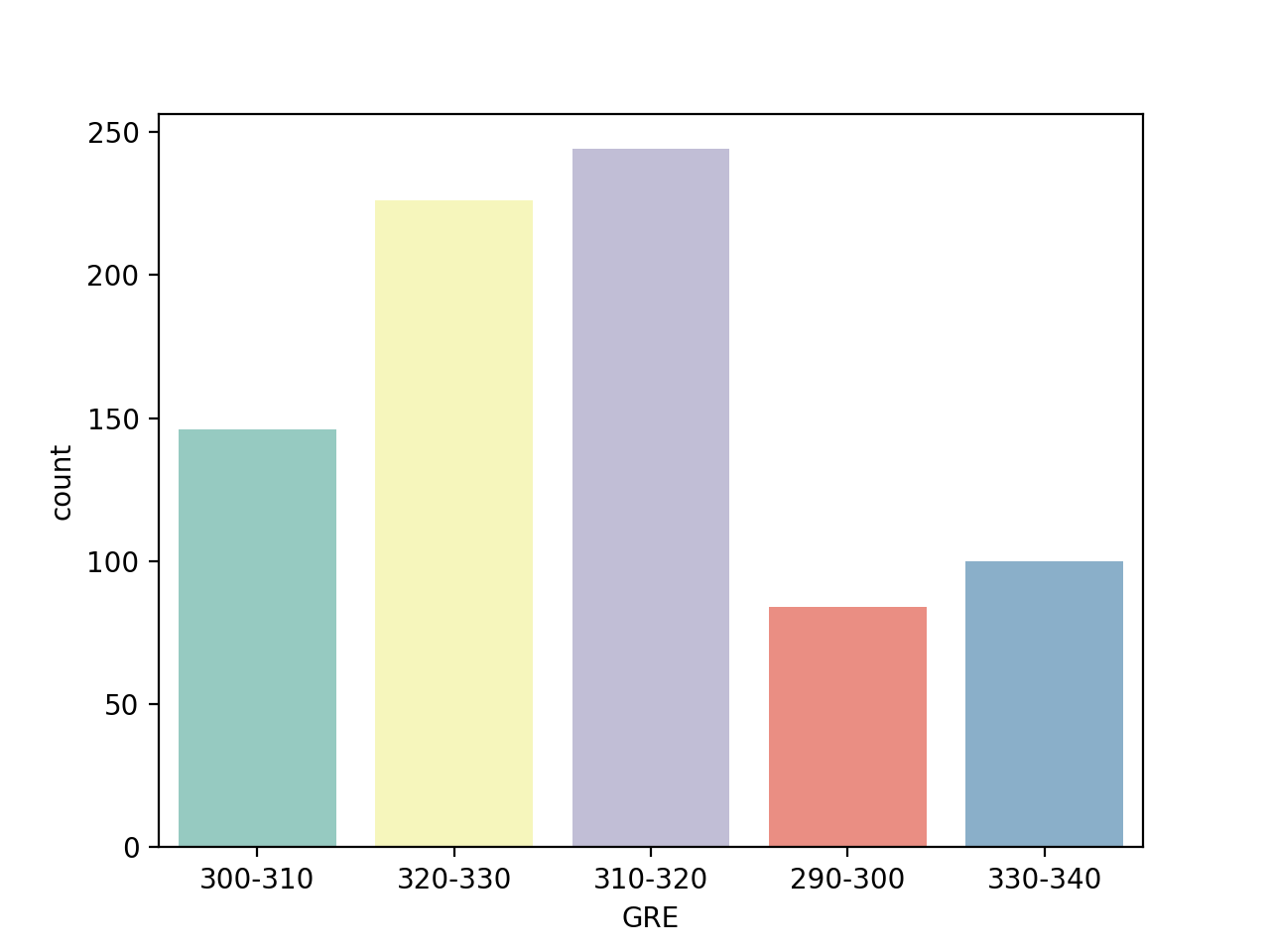


From the correlation matrix, we can see that the serial number is barely correlated with the chance of admission or other features. This means that we can delete this variable in the future analysis because it does not provide useful information.

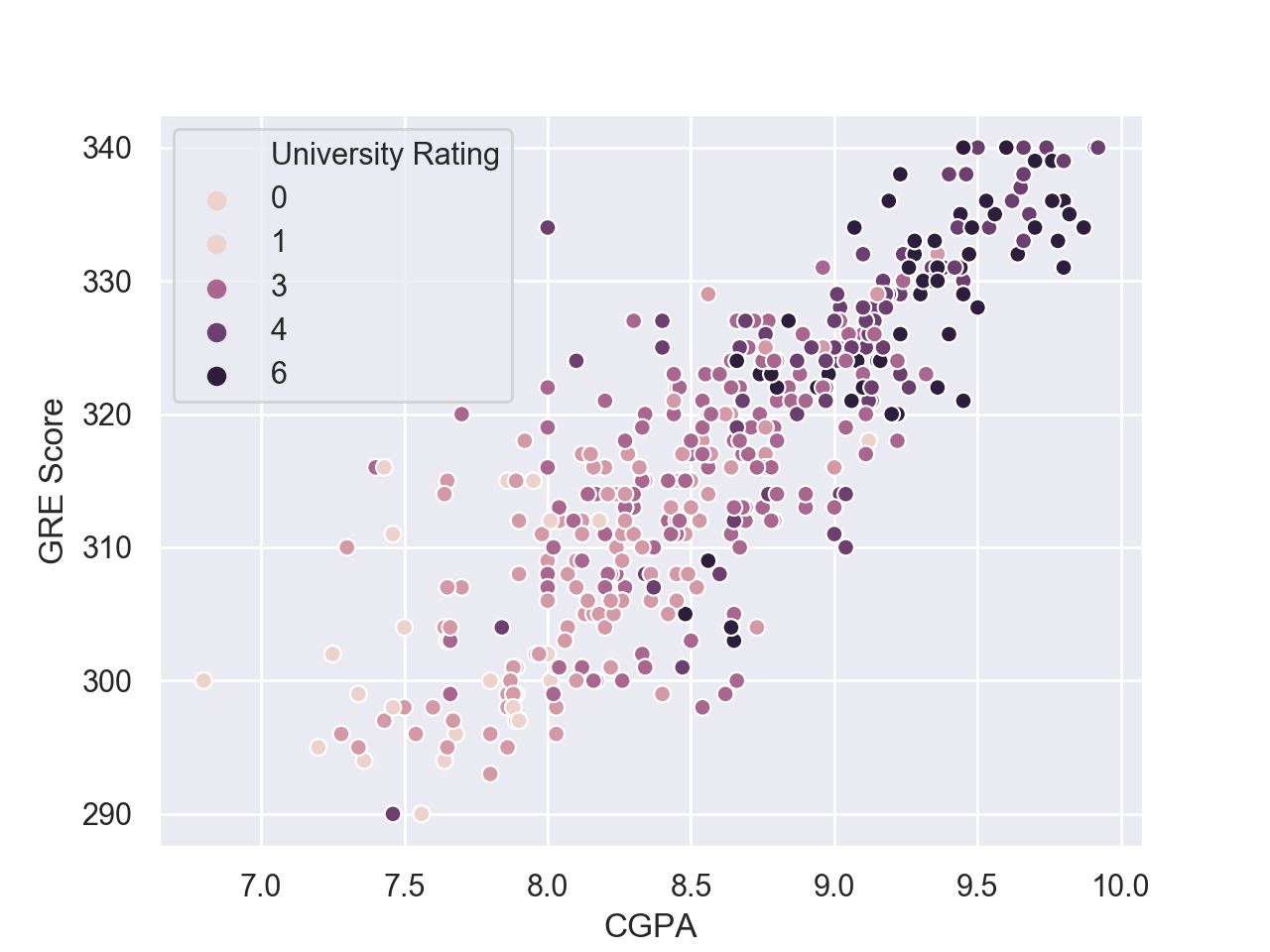
After that, we generate a custom diverging colormap. From the graph, we can see that the features that are most positively correlated with the chance of admission are GPA, GRE score and TOEFL score. The features that are least positively correlated with the chance of admission are Research, LOR and SOP.



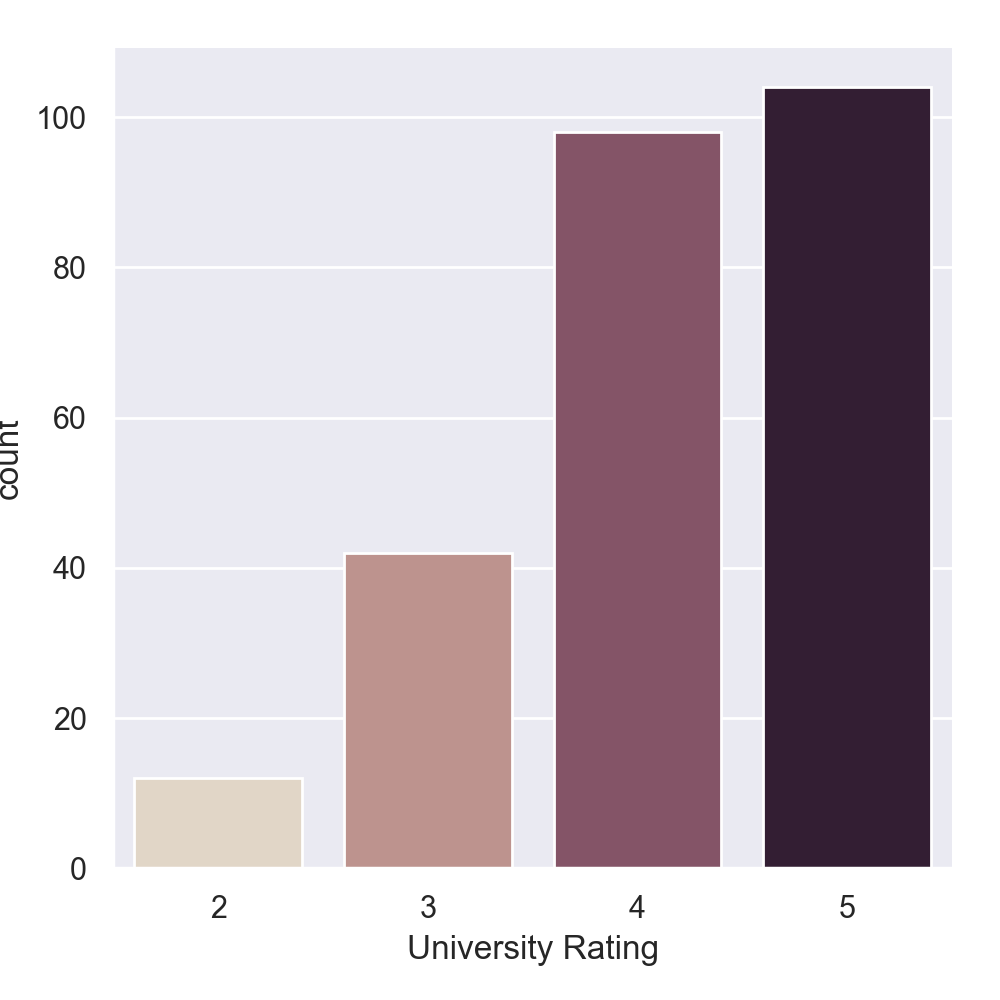
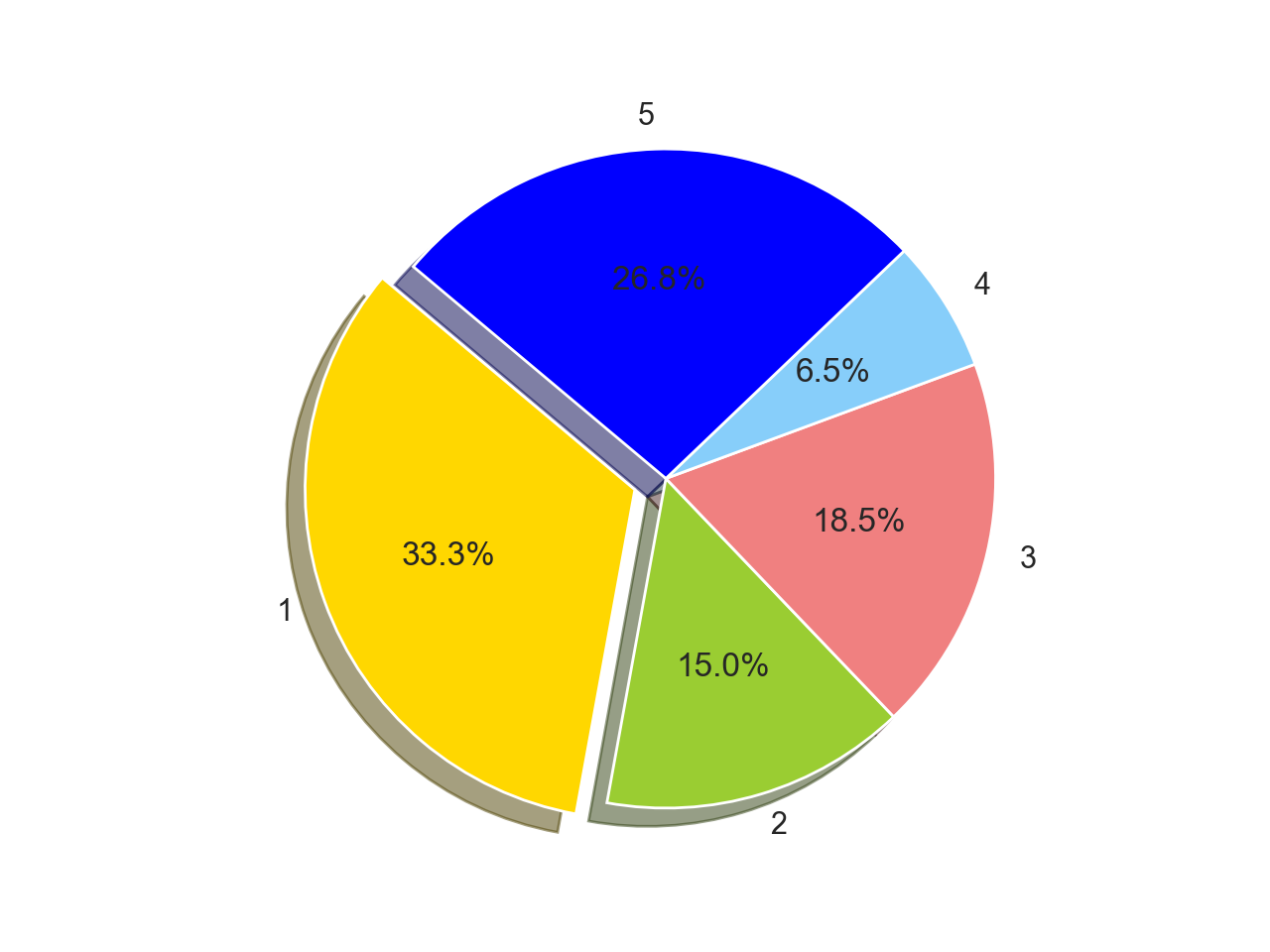
We also draw a bar plot about GRE and TOEFL score to see the frequency of test scores among the candidates. From the graph, we can see that the most frequent GRE score is in the range from 310-320. And the most frequent TOEFL score is in the range from 100-110.



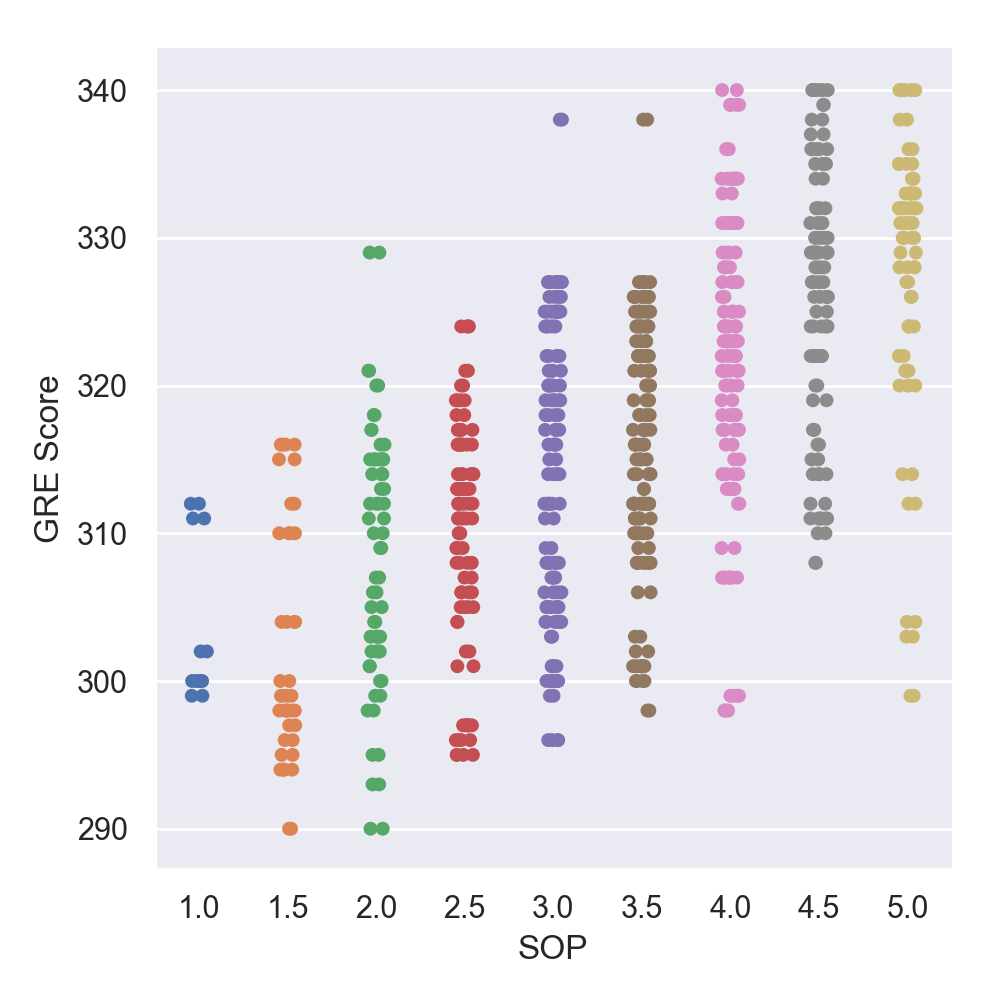
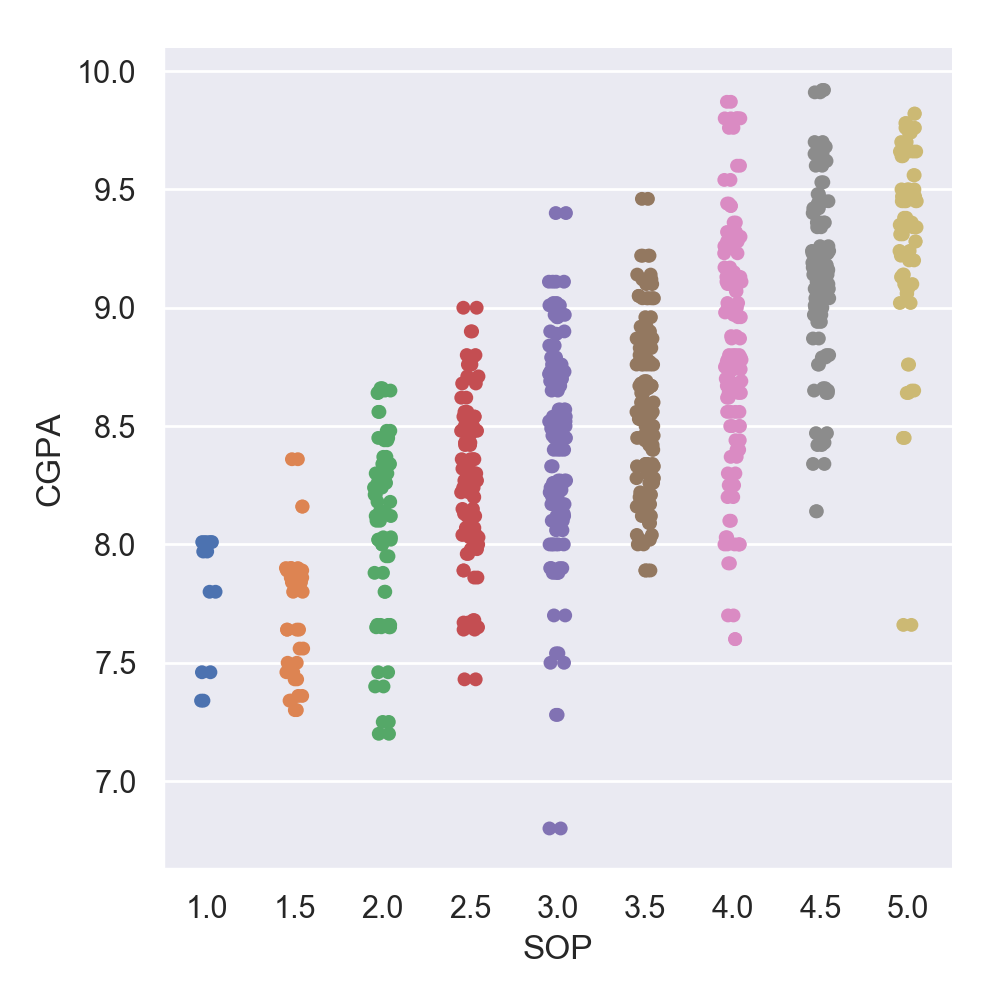
We want to see a comparison about GPA vs GRE, so we draw a scatterplot by using seaborn package to get the comparison. The graph shows a positive linearity relationship between GPA and GRE, better GPA is correlated with better GRE score. Also candidates with higher university ranking tend to have higher GPA and higher GRE score.



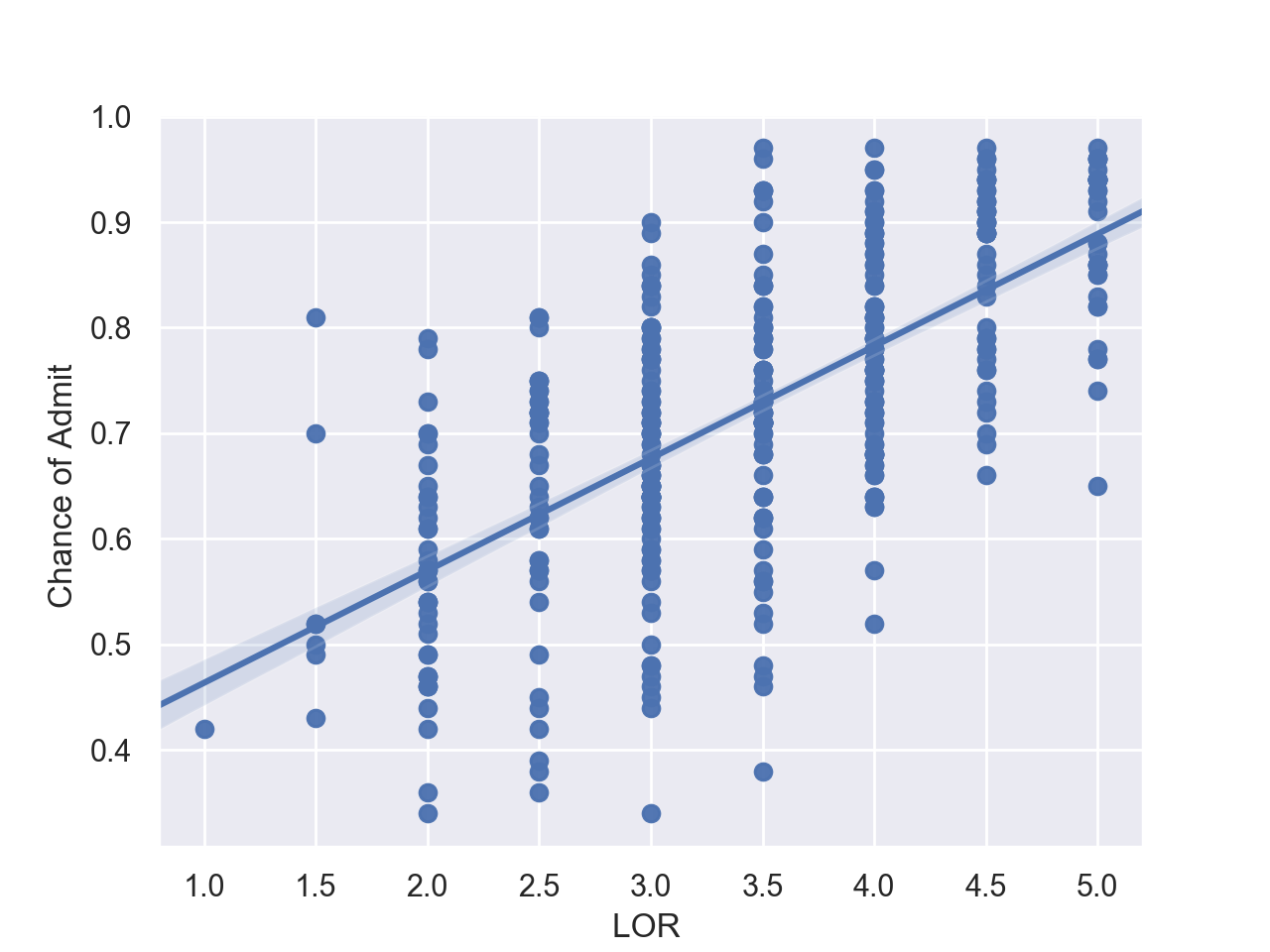
The pie plot below shows the distribution of each students’ university ranking. We could a pretty interesting points that the university ranked as 1 and university ranked as 5 does not have a pretty big difference. From the bar plot we can see among the candidates with high admission change, the most of them are coming from higher-ranked colleges. (Nearly a quarter of the data)



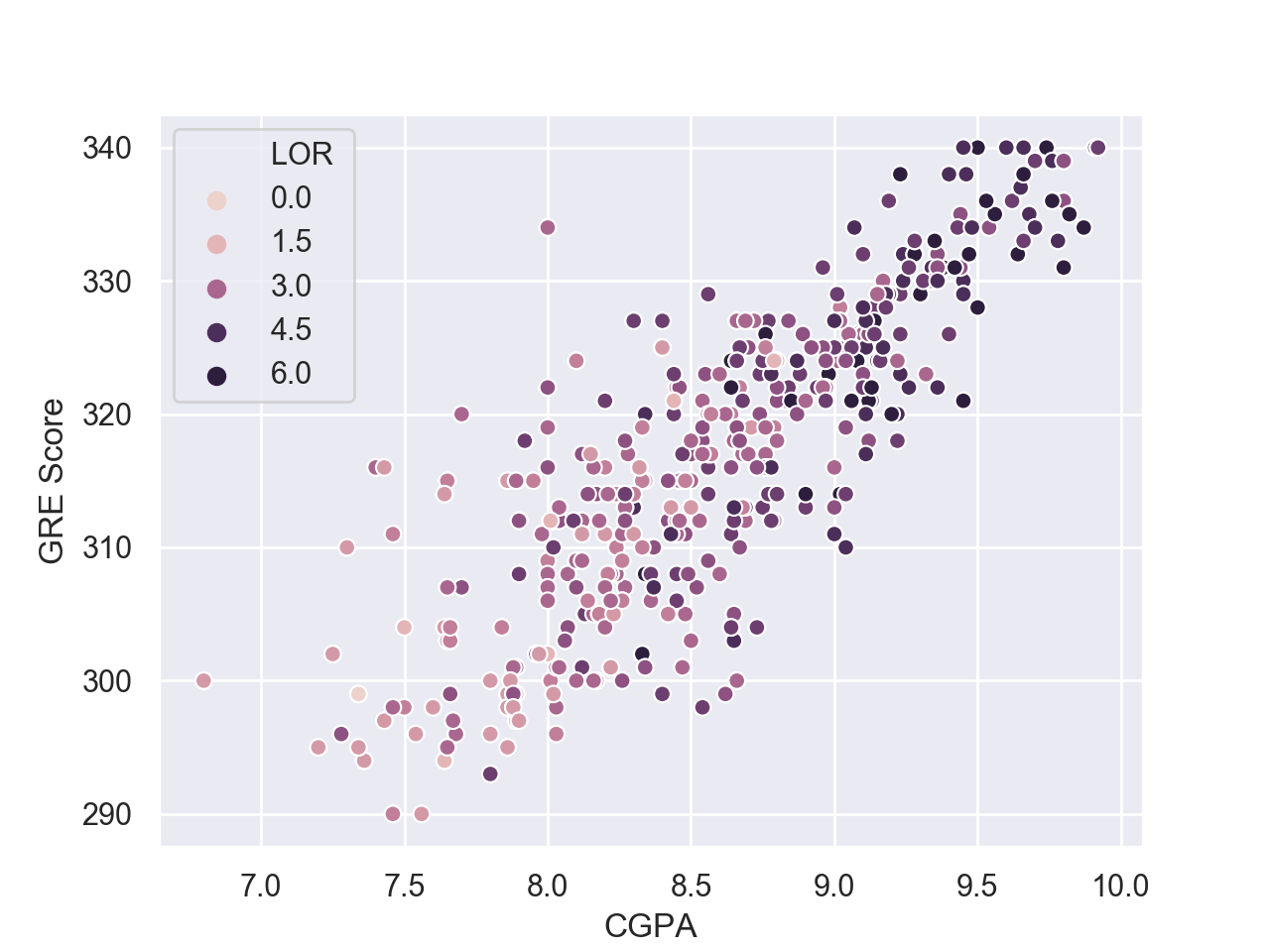
We made a comparison about GPA vs SOP as well as GRE vs SOP and the result plots can be seen below. From the plot we could see better SOP associated with better GPA as well as GRE.



In university admissions, letter of recommendation also counts a great part to influence whether a student can be admitted or not, so we visualize it to see the trend and the plot is shown below. From the plot we found stronger letter of recommendation associated with higher chance of admission.



Lastly, we draw a scatterplot to compare GPA and GRE. The graph shows a positive linearity relationship between GPA and GRE. Better GPA is correlated with better GRE score also candidates with stronger recommendation letter tend to have higher GPA and higher GRE score.



**Model Comparison**

In this project, we used linear regression, decision tree, random forest and neural network. Below is the comparison about these four models.

Linear regression attempts to find relationship between exploratory variable and dependent variable. If the model has multiple exploratory variables, it’s called multiple linear regression. Linear regression is a pretty useful, easy to use and implement regression model. Since we have a pretty small and clear dataset, the first model we have thought of was linear regression.

Decision tree is a model that pretty straightforward in visualizations. There are many advantages of using decision tree. The internal workings are capable of being observed and thus we could use decision tree to reproduce work. We can use decision tree to analyze both numerical and categorical data. If the dataset is large, decision tree algorithm can also perform pretty well. The disadvantage includes choose of each node (it can only choose the best result in each step rather than the whole step, which easily leads to local optimum rather than global optimum) and overfitting (we could use prune to somewhat get rid of it).

Random forest can decrease both error due to bias and error due to variance. It is simply a collection of decision trees and aggregated into one final result. This can mitigate overfitting problem and also won’t substantially increase error due to bias. Thus, it is a better algorithm in avoid overfitting comparing to decision tree.

Based on this, if you want to make the model simple and easy to explain, and you also don’t care a lot about the multi-collinearity and overfitting problem it’s better to use decision tree than random forest.

Deep learning has been more and more popular nowadays. It is a subset of machine learning that achieves great power to represent world as nested hierarchy of concepts. In this project, we built three-layer neural network to train the model (input layer, hidden layer and output layer). Deep learning works extremely well in large dataset. Since our dataset is really small, it might not be a good idea to use very complex deep learning model to train, that’s why we simply use one hidden layer in our model building part. Basically, if the dataset is simple, and we build very complex neural network it can hardly receive well-performed result. If the dataset is simple and we use simple neural network to train, the result can be pretty similar with machine learning models. That’s why in our project, we would prefer using machine learning algorithms rather than deep learning.