
USA College Admission
Lets get you in a grad school!

Final Project Report

Analysis of US Admissions Datasets

Rashmi Pai

November 13, 2020

Introduction

Getting into a good US University is very critical for aspiring students. Although there are thousands of universities one can apply, there are several parameters that also decide on whether you get an admit. Moreover these factors seem to have strong correlation among themselves. As part of my final project in *Introduction to Data Analytics,* I decided to choose (link to dataset): US Admissions dataset from Kaggle. I've been through such a journey where I found myself at the crossroads of going through the difficult journey of applying to universities of my liking and waiting anxiously to hear back from them. If I'd access to such a dataset I'm sure it would have helped me a lot back then.

About the dataset

The dataset has following columns and their types.

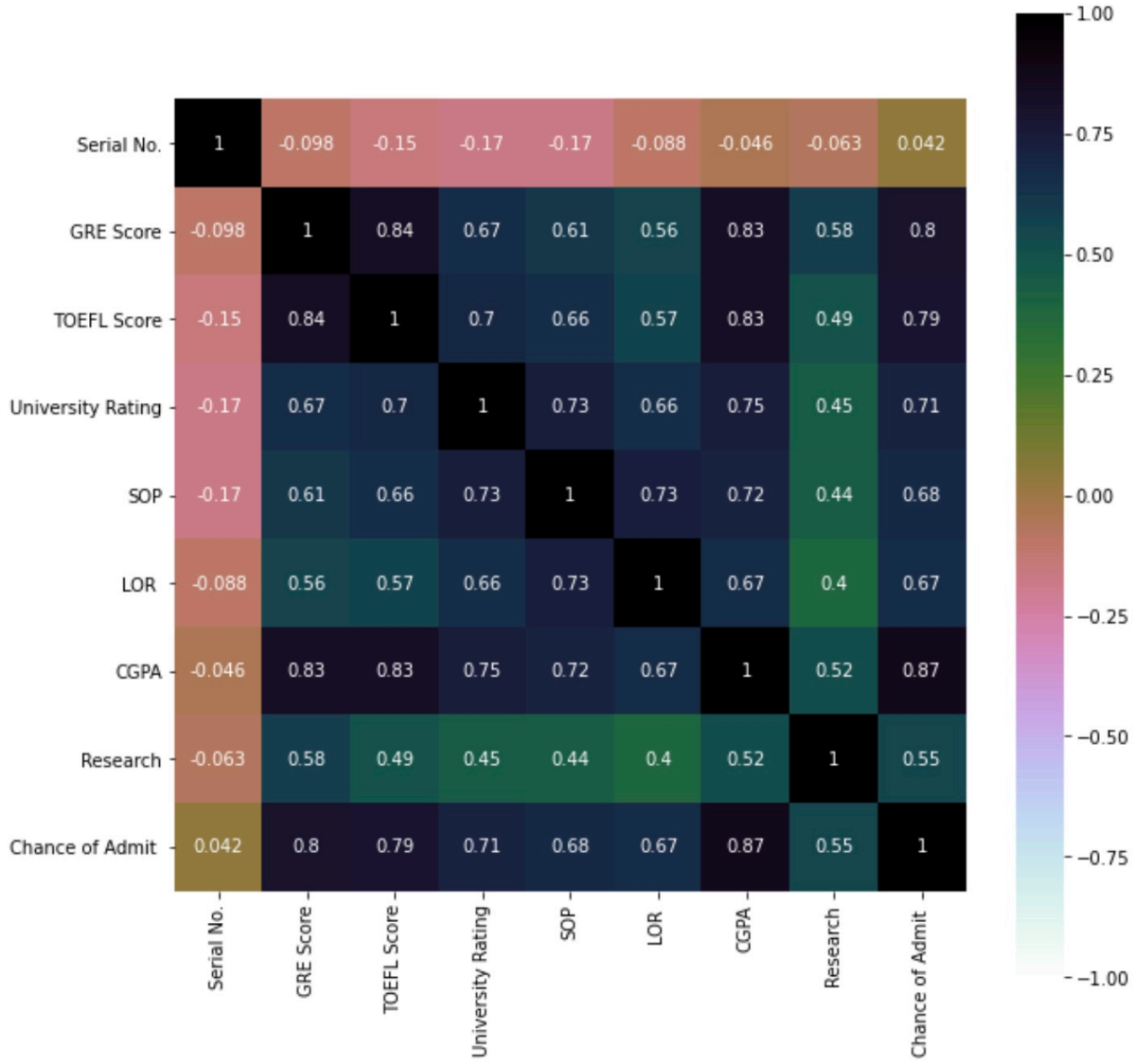
Column	Type
Serial No.	int64
GRE Score	int64
TOEFL Score	int64
University Rating	int64
SOP	float64
LOR	float64
CGPA	float64
Research	int64
Chance of Admit	float64

Further more there were no empty rows or columns.

```
$ data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Serial No.         400 non-null    int64
 1   GRE Score          400 non-null    int64
 2   TOEFL Score        400 non-null    int64
 3   University Rating  400 non-null    int64
 4   SOP                400 non-null    float64
 5   LOR                400 non-null    float64
 6   CGPA               400 non-null    float64
 7   Research           400 non-null    int64
 8   Chance of Admit    400 non-null    float64
dtypes: float64(4), int64(5)
memory usage: 28.2 KB
```

Data Correlation and Statistics

Next I tried to find what are the columns on interest and what are some of the columns I can remove. Plotting the dataframe correlation I could clearly see that the Serial No. had no correlation to other columns whatsoever so I could drop it. Similarly I could see strong correlation among other columns like GRE and Chance of Admit and CGPA and Chance of Admit for e.g.



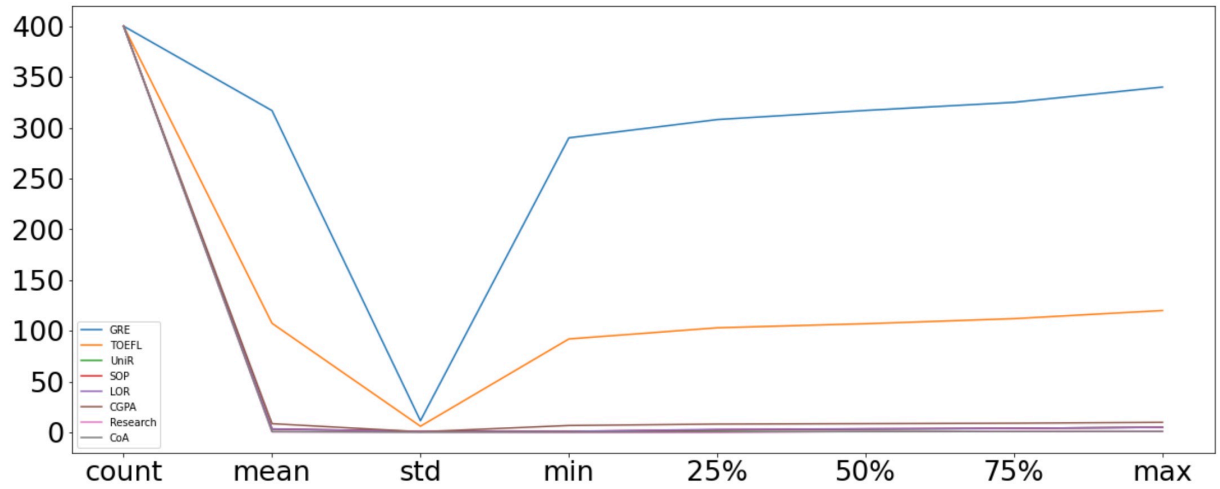
Also since the column names have spaces or are too long, I decided to rename them because I'll be repeatedly using them.

Column	New Column Name
~~Serial No.~~	N/A
~~GRE Score~~	GRE
~~TOEFL Score~~	TOEFL
~~University Rating~~	UniR
SOP	SOP
LOR	LOR
CGPA	CGPA
Research	Research
~~Chance of Admit~~	CoA

Next I plotted the graph on `data.describe()` to get a feel of the data.

```
data.describe().plot(fontsize=27, figsize = (20,8))
```

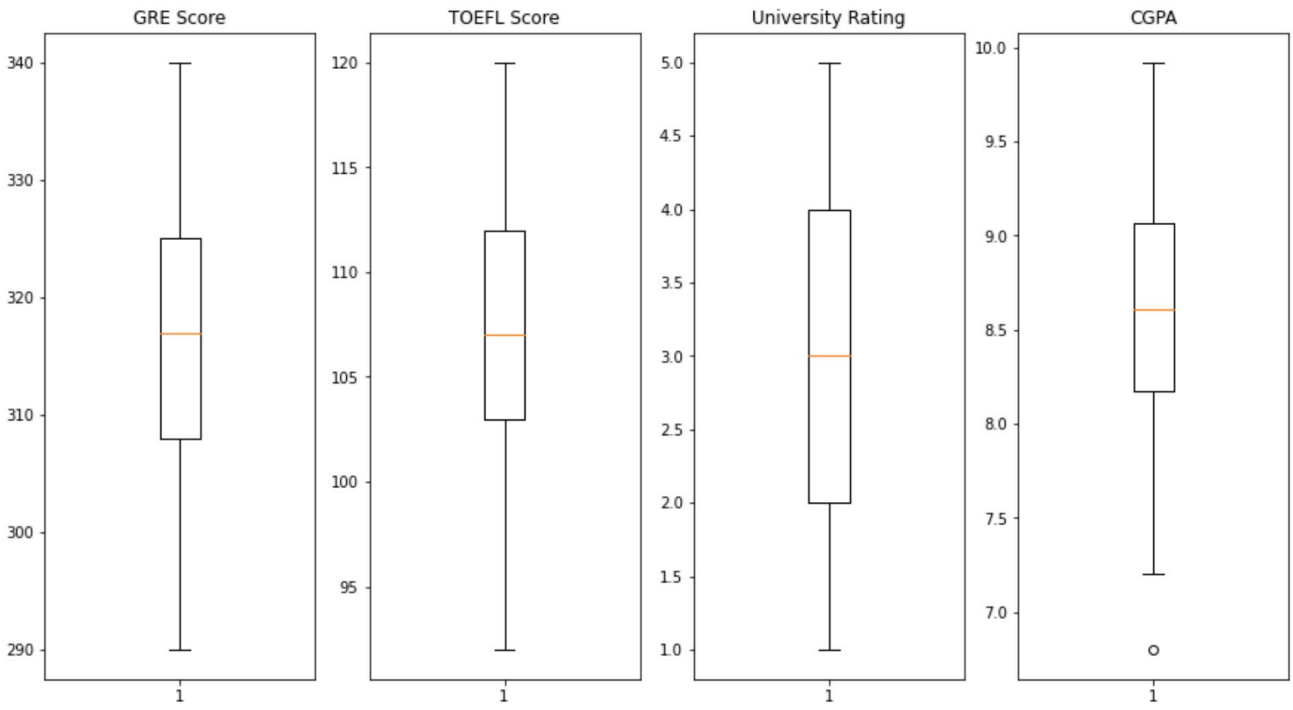
```
<AxesSubplot:>
```



This told the range of the data in various columns as well as the statistics on my dataset.

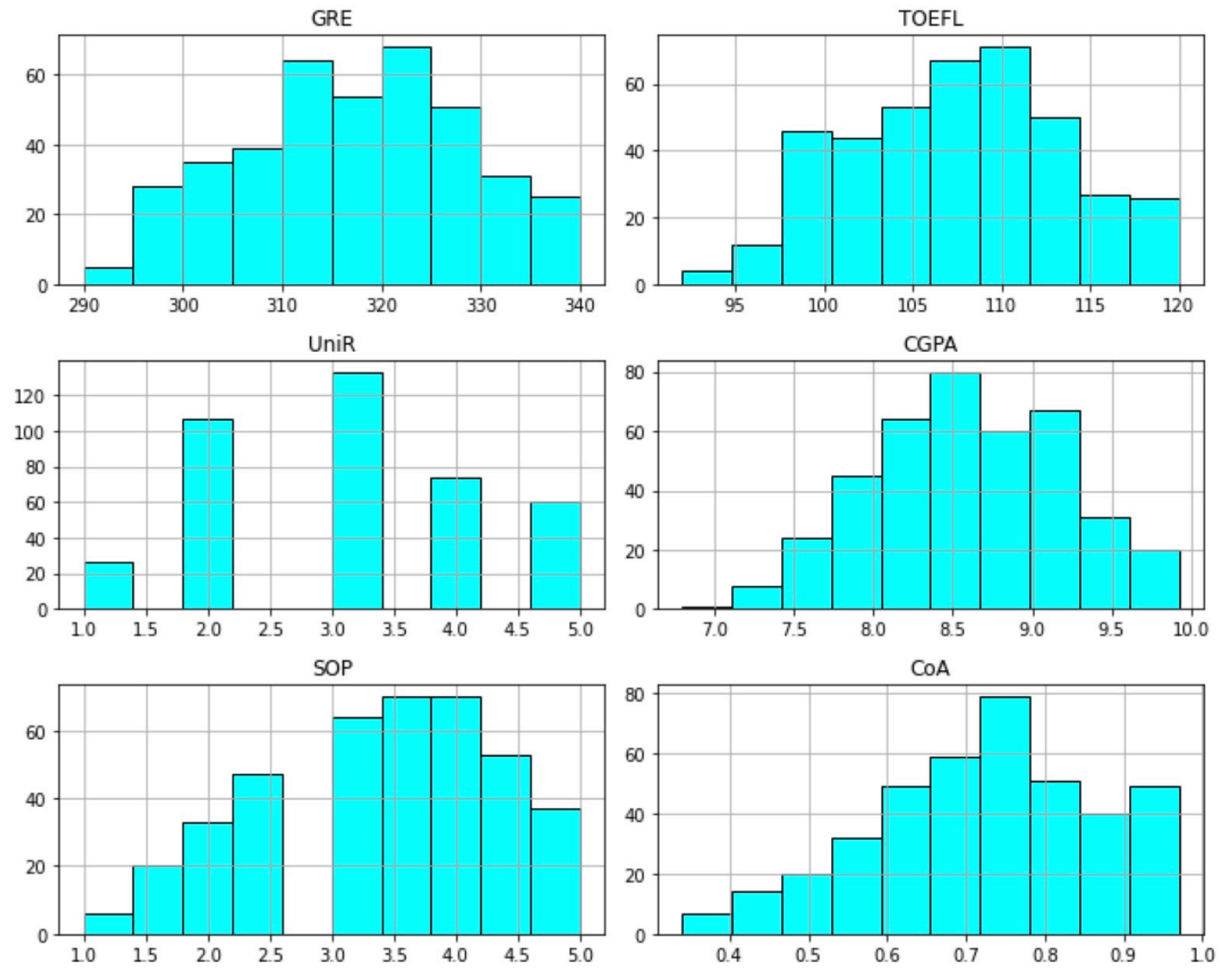
Analysis on the data

Box Plot



The box plot on GRE, TOEFL, University Ranking and CGPA allowed me to visualize the range of data in each feature/column.

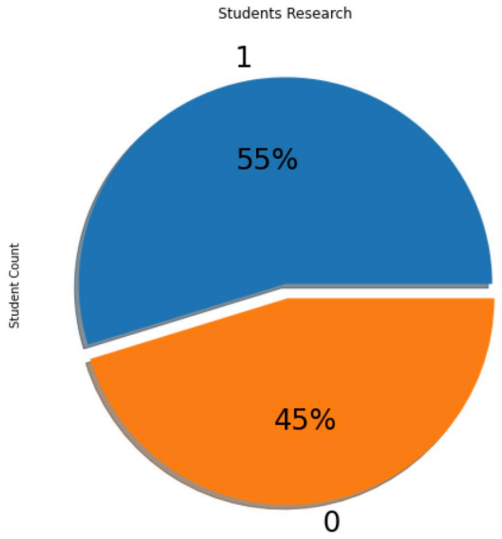
Histogram

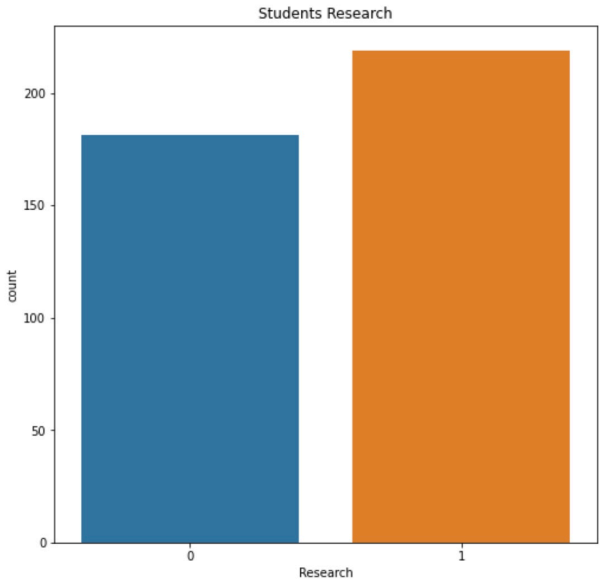


The Histogram allowed me to see what the data distributions across each of the columns. GRE, TOEFL and CGPA showed a well defined bell curve distribution of data.

Importance of Research

Next I looked at whether choosing to do research when applying to universities really mattered. In order to find that, I first tried to see how many students from total rows in the dataset chose research vs how many did not.

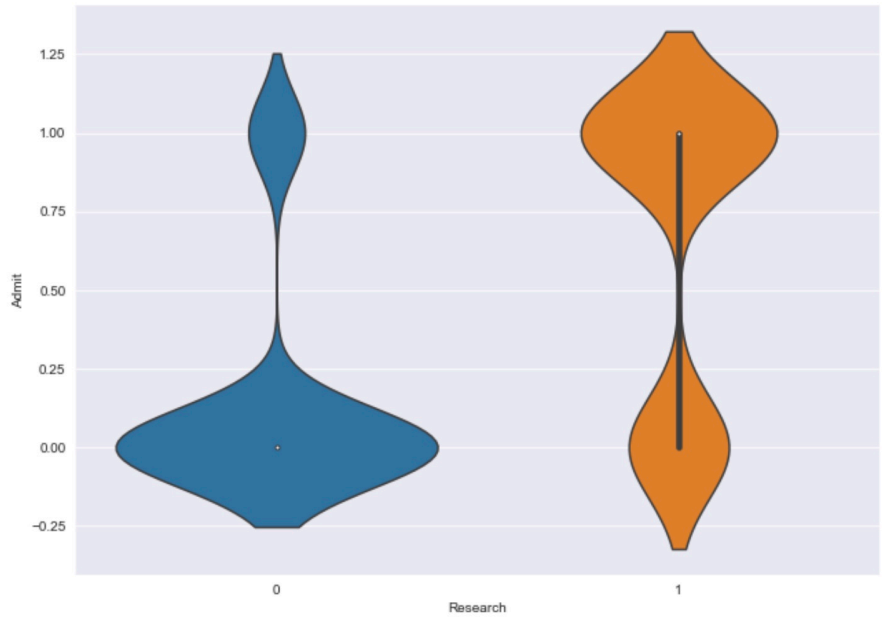




The dataset has more student data who chose research vs those who did not. Although the distribution was 55%:45% I was able to use this data to find some interesting observations in the dataset.

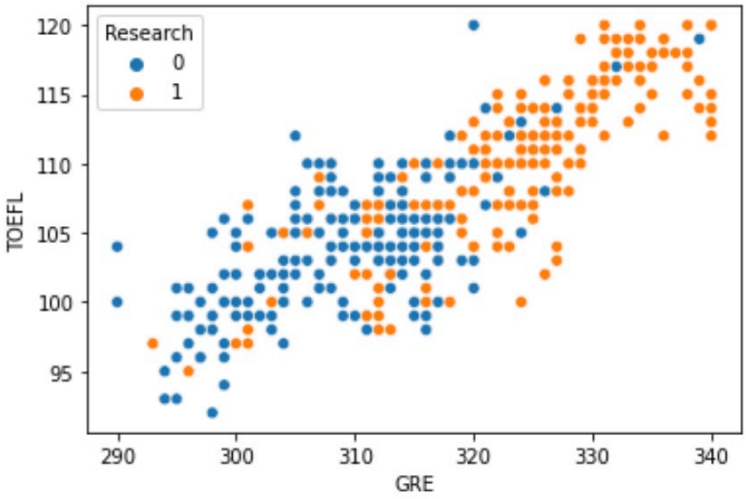
In order to correlate importance of research to chance of admission, I decided to split the chance of admission into a binary data distribution where any chance of 0.75 or greater I mapped as 1 and rest as 0 and inserted a new column called Admit. I then plotted some graphs that showed me correlation of other columns in regards to research when it came to students getting admit or not.

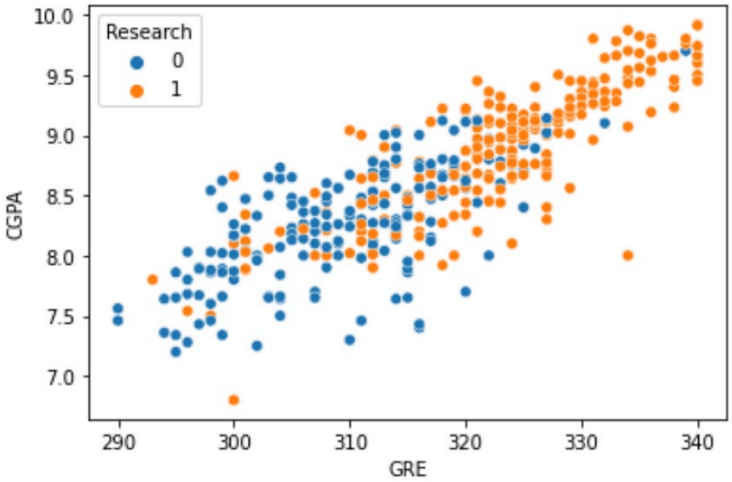
If I just looked at whether choosing research landed the students admit, this is what the data visualization appeared.



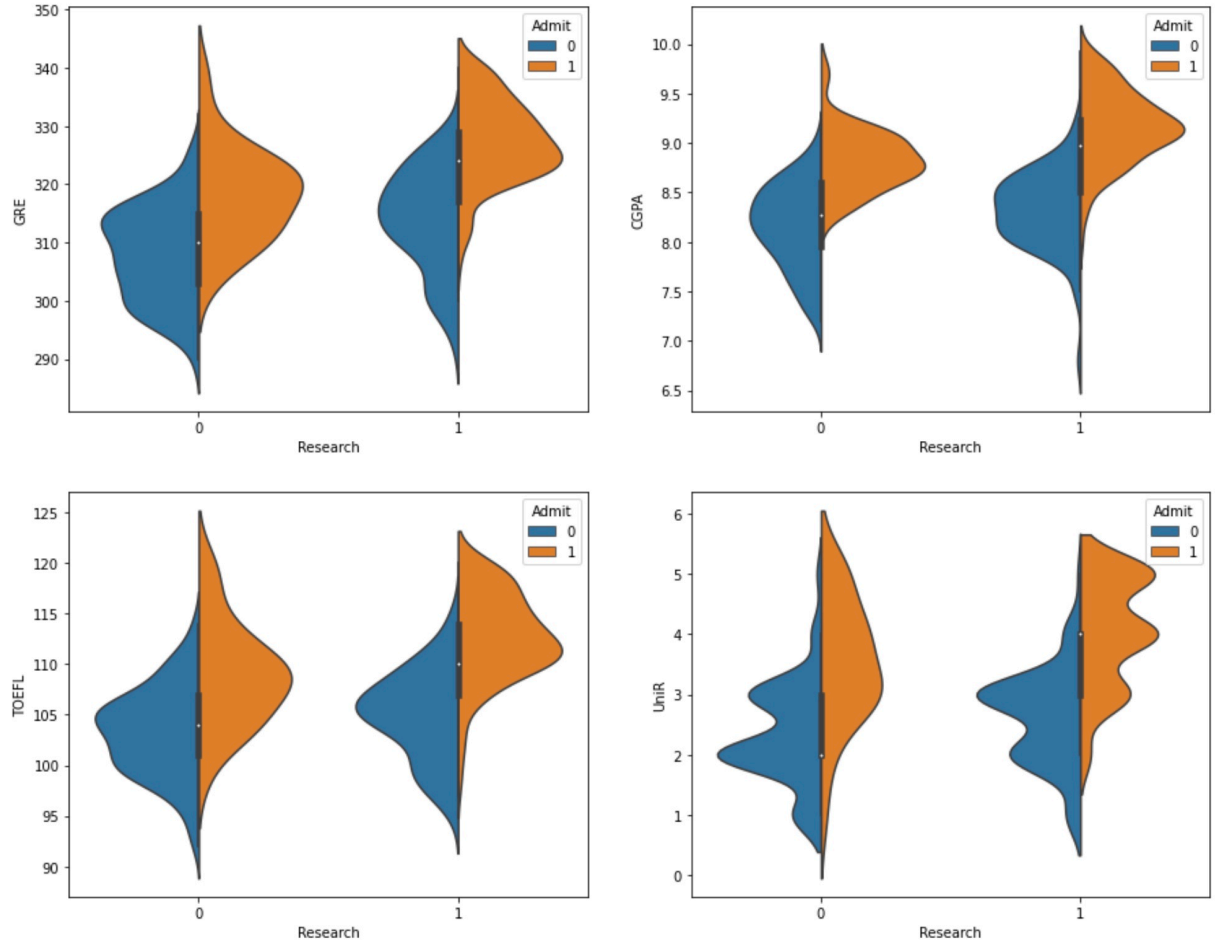
This showed me that for this data set, students who chose research had a serious advantage to their counterparts who opted not to do research.

Furthermore I found that students who opted research tend to have higher GRE/TOEFL and CGPA as can be seen below.





Lastly, I tried to see how choosing research reflected on scores on all the columns vs students with chance of admit mapped to a binary yes/no.



All features/columns favored doing research and procuring higher/better scores when it came to a chance of admit greater than 0.75.

Increasing your chance of admit

In order to achieve over 95% of chance of admit, you need to have GRE=337.45, TOEFL=117.27, SOP=4.54, LOR=4.5, CGPA=9.78.

```
data_sort=data.sort_values(by=data.columns[-1],ascending=False)
data_sort.head()
```

	GRE	TOEFL	UniR	SOP	LOR	CGPA	Research	CoA	Admit
0	337	118	4	4.5	4.5	9.65	1	0.92	1
222	324	113	4	4.5	4.0	8.79	0	0.76	1
235	326	111	5	4.5	4.0	9.23	1	0.88	1
234	330	113	5	5.0	4.0	9.31	1	0.91	1
231	319	106	3	3.5	2.5	8.33	1	0.74	1

```
data_sort[(data_sort['CoA']>0.95)].mean().reset_index()
```

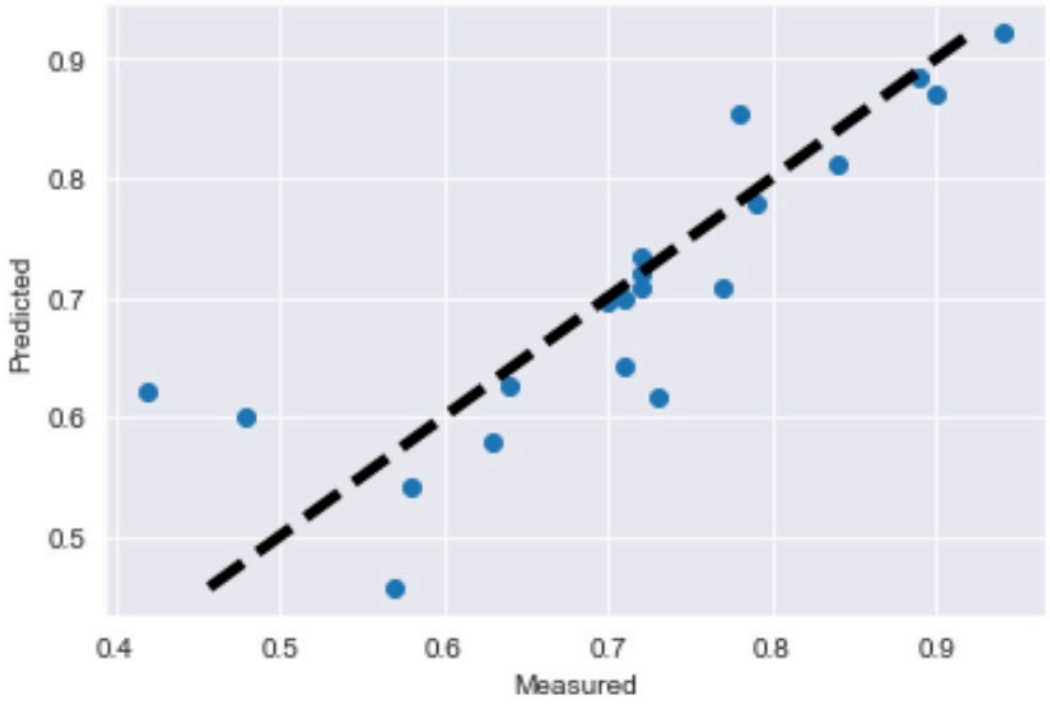
	index	0
0	GRE	337.454545
1	TOEFL	117.272727
2	UniR	4.636364
3	SOP	4.545455
4	LOR	4.500000
5	CGPA	9.787273
6	Research	1.000000
7	CoA	0.963636
8	Admit	1.000000

Predict admission using Linear Regression

I tried to fit a Linear Regression model on my dataset by splitting the data set into training and test data using the sklearn package. The MSE (Mean Square Error) on the linear regression model came out to 0.005. When I put my model to test I saw following observations.

Actual Value	Model Prediction
0.92	0.95
0.90	0.86
0.88	0.86

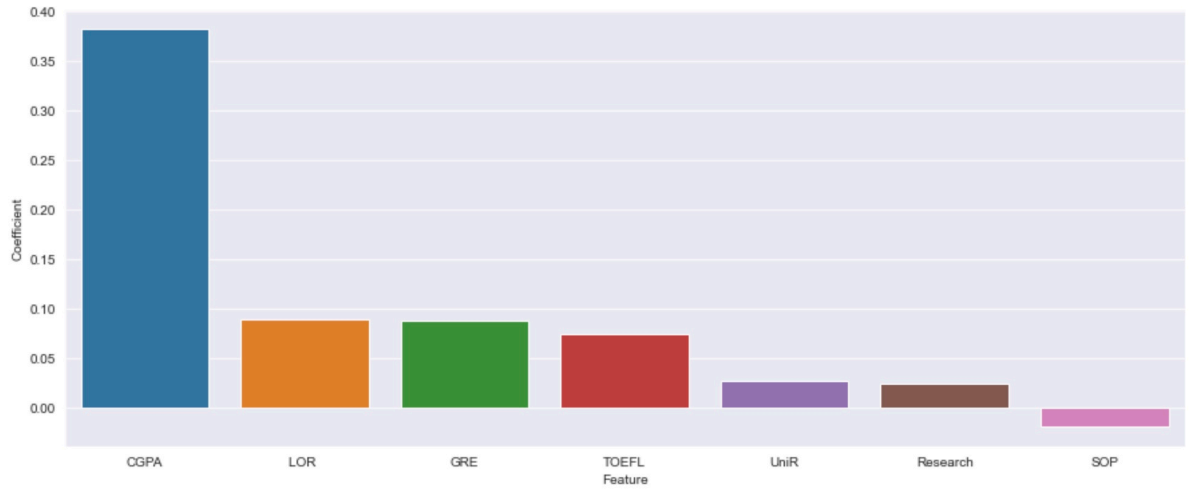
Following was the graph of Predicted vs Test Values correlating data.



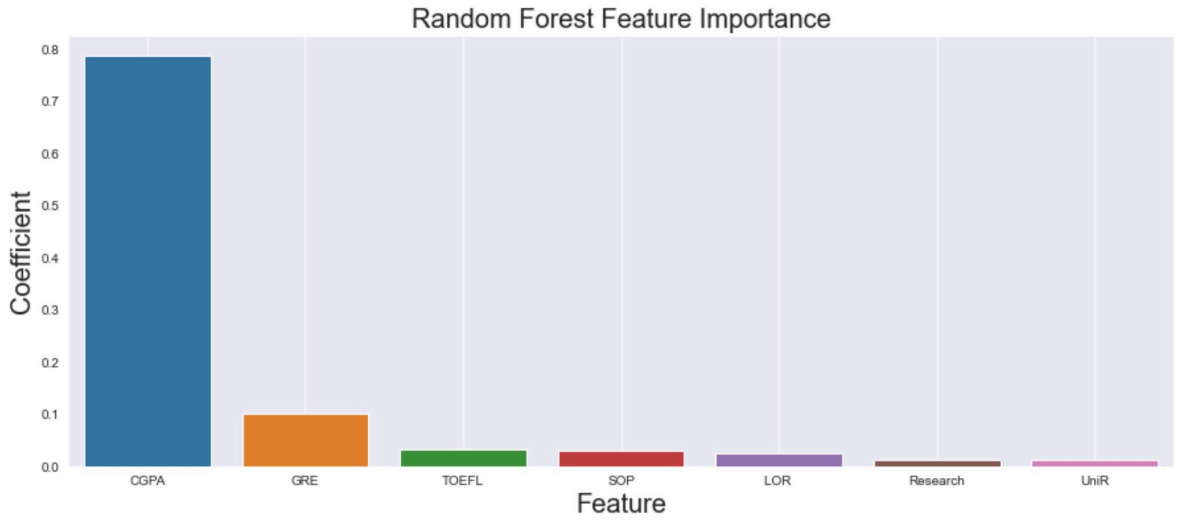
Feature Importance

Both Linear Regression and Random Forest Regression models showed similar feature importance with CGPA playing the most important role while others not so much.

Linear Regression



Random Forest Regression



Conclusion

I drew following conclusion from above analysis.

- *Chance of admit hugely depends on CGPA*: Both Linear Regression and Random Forest Regression models showed CGPA to have high importance when it came to chances of admit.
- *Students with good CGPA and research also tend to have better GRE/TOEFL scores*: As I correlated research and CGPA, I saw those with high CGPA and who choose research tend to have good GRE/TOEFL scores (more than mean)
- *Chance of admit favors students who chose to do research*: Finally the chance of admit was highly skewed in the favor of students who choose to do research.

Appendix/Citations

- Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019
- https://matplotlib.org/3.3.1/gallery/pyplots/boxplot_demo_pyplot.html
- https://stackoverflow.com/questions/30263627/python-pandas-summary-table-plot
- https://seaborn.pydata.org/generated/seaborn.heatmap.html
- https://seaborn.pydata.org/generated/seaborn.catplot.html
- https://machinelearningmastery.com/calculate-feature-importance-with-python/
- https://scikit-learn.org/0.18/auto_examples/plot_cv_predict.html
- https://www.kaggle.com/mohansacharya/graduate-admissions?select=Admission_Predict.csv