Graduate Admission Data Analysis

Introduction

Every year thousands of students apply to universities in the US for their master's degree. The process of selecting a university is tedious and time-consuming as there are a lot of factors that are considered while applying to a university. Using this dataset we plan to understand the student selection model used by one of the most reputed colleges in the USA- UCLA. This model will help students understand the weight of each type p-score in the final decision and also how they can together influence the chances of any student getting an admit from the university of their choice.

Our analysis will consist of three stages. The first stage of data will be preprocessed. However, there are various analyses and graphs. In the next step, the weight of each variable will be showcased by using many supervised algorithms. Then we will reverse engineer the model that UCLA uses.

1. About the DataSet

This dataset is inspired by the UCLA Graduate Dataset. The test scores and GPA are in the older format. The dataset is owned by Mohan S Acharya. We have obtained the same from Mr. Mohan contribution on Kaggle

This dataset link is below:

https://www.kaggle.com/mohansacharya/graduate-admissions

2. Features in the dataset

The dataset contains several parameters which are considered important during the application for Masters Programs in a US university.

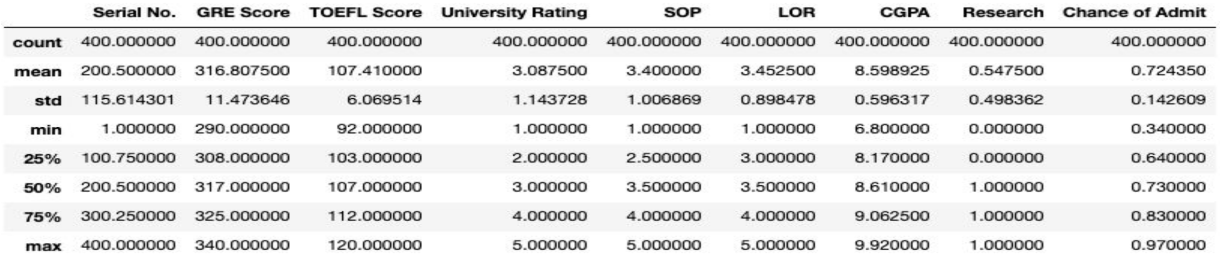
The features included are:

1. GRE Scores (290 to 340)
2. TOEFL Scores (92 to 120)
3. University Rating (1 to 5)
4. Statement of Purpose (1 to 5)
5. Letter of Recommendation Strength (1 to 5)
6. Undergraduate CGPA (6.8 to 9.92)
7. Research Experience (0 or 1)
8. Chance of Admit (0.34 to 0.97)

3. Pre Processing and data exploration

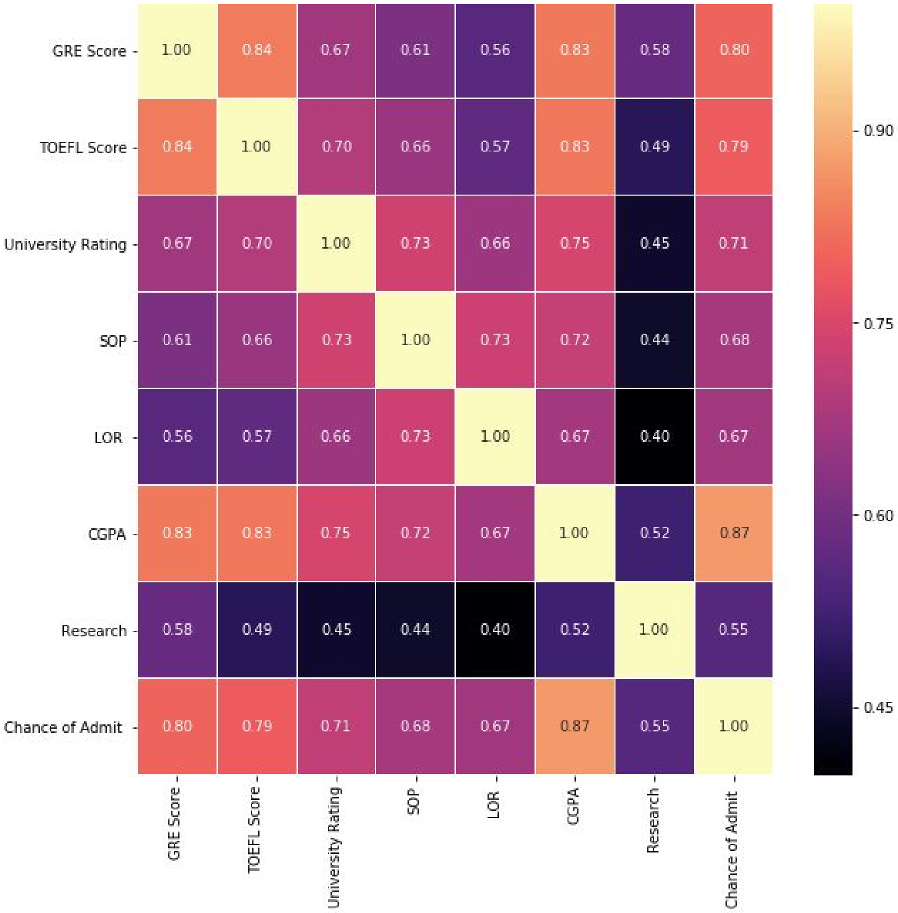
Taking a look at our dataset for understanding:

- **Using the describe function to understand the values present in each variable:**

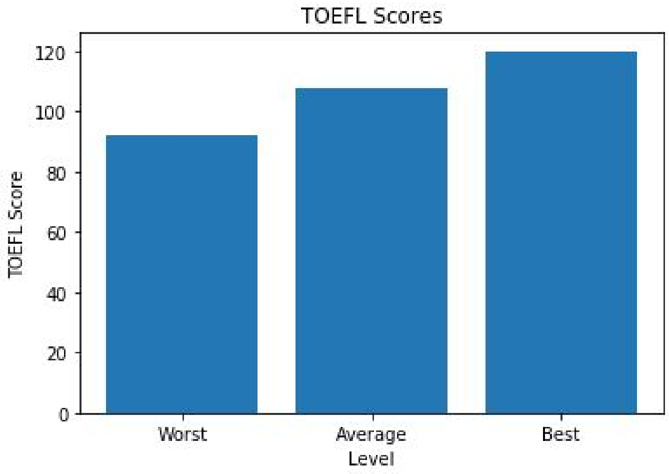


	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000
mean	200.500000	316.807500	107.410000	3.087500	3.400000	3.452500	8.598925	0.547500	0.724350
std	115.614301	11.473646	6.069514	1.143728	1.006869	0.898478	0.596317	0.498362	0.142609
min	1.000000	290.000000	92.000000	1.000000	1.000000	1.000000	6.800000	0.000000	0.340000
25%	100.750000	308.000000	103.000000	2.000000	2.500000	3.000000	8.170000	0.000000	0.640000
50%	200.500000	317.000000	107.000000	3.000000	3.500000	3.500000	8.610000	1.000000	0.730000
75%	300.250000	325.000000	112.000000	4.000000	4.000000	4.000000	9.062500	1.000000	0.830000
max	400.000000	340.000000	120.000000	5.000000	5.000000	5.000000	9.920000	1.000000	0.970000

- **Correlation of the Variables:**
  - A high correlation is visible in most of the variables. Values are ranging from .40-.84. Hence any interpretations of the coefficient values needs to be done with caution.

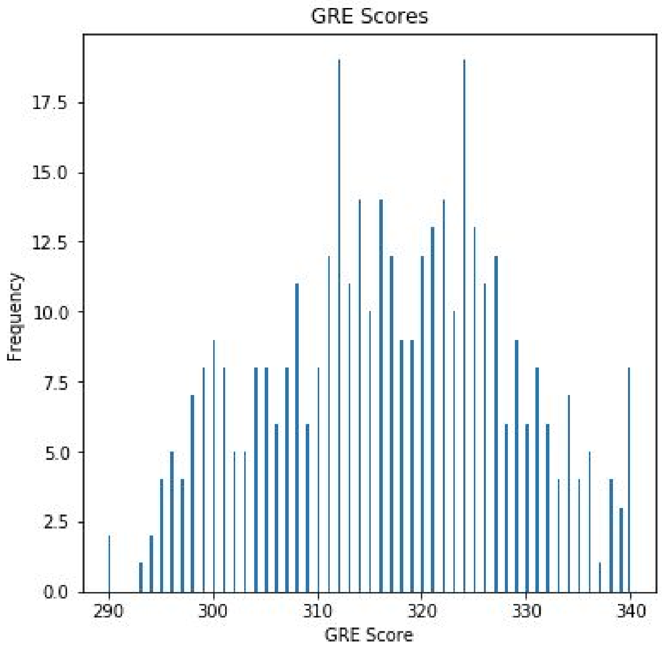


- **TOEFL Score:**
  - The lowest TOEFL score is 92 and the highest Toefl score is 120. The average is 107.41.

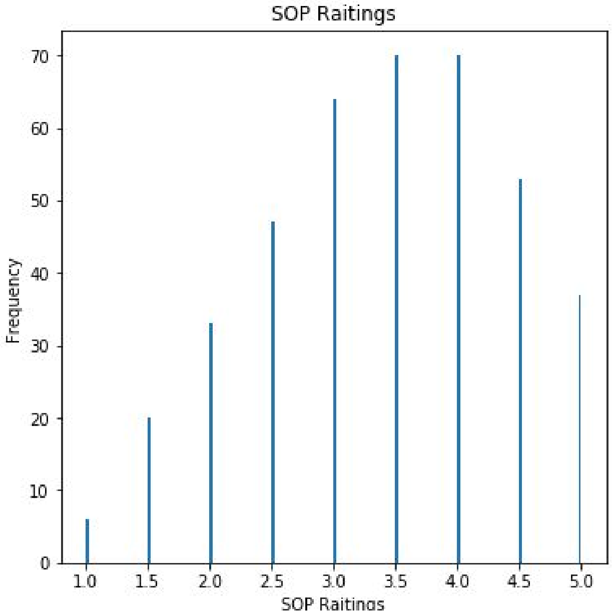


- **GRE Score:**

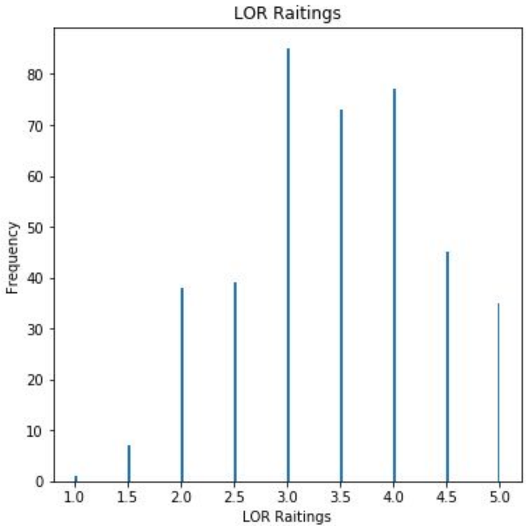
This histogram shows the frequency for GRE scores. There is a density between 310 and 330. Being above this range would be a good feature for a candidate to stand out.



- **SOP:**
  - Most of the SOP rating is in range 3-4. They have been calculated by an expert on a scale from 1-5 as per the quality of the SOP

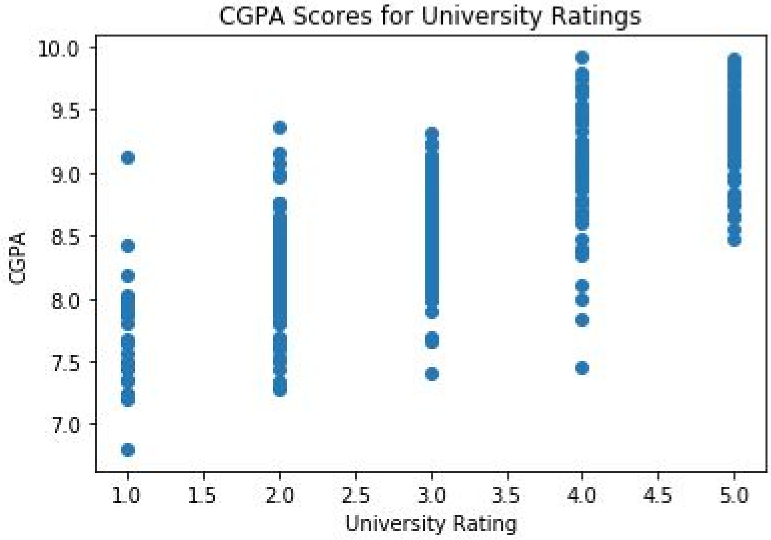


- **LOR:**
  - 75% of the LOR ratings are above 3.0.They have been calculated by an expert on a scale from 1-5 as per the quality of the LOR. For more than 1 LOR the average of all the ratings is taken.



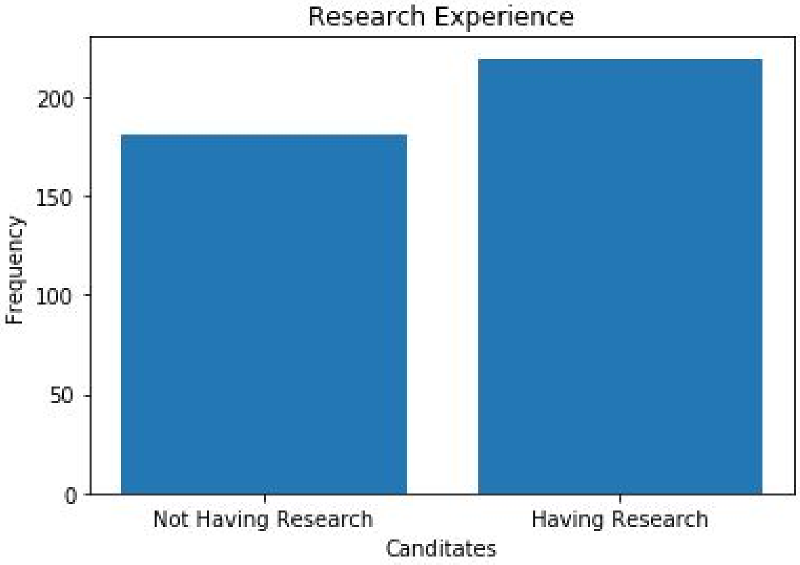
- **CGPA Scores for University Ratings:**

As the quality of the university increases, the CGPA score increases.

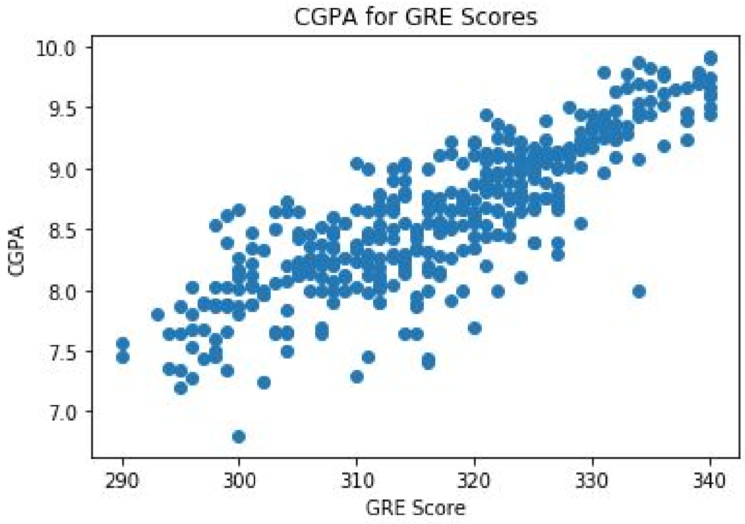


- **Having Research or not:**

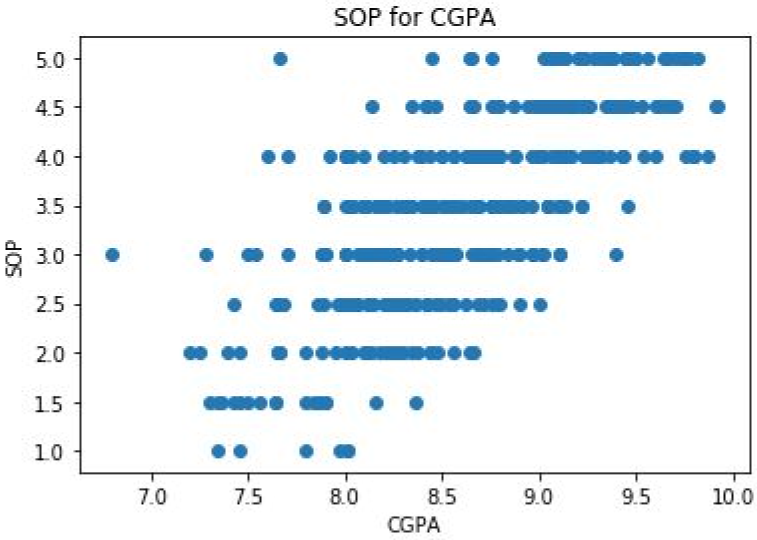
The majority of the candidates in the dataset have research experience. Therefore, the Research will be an unimportant feature for the Chance of Admit. The correlation between Chance of Admit and Research was already lower than other correlation values.
Not Having Research: 181
Having Research: 219



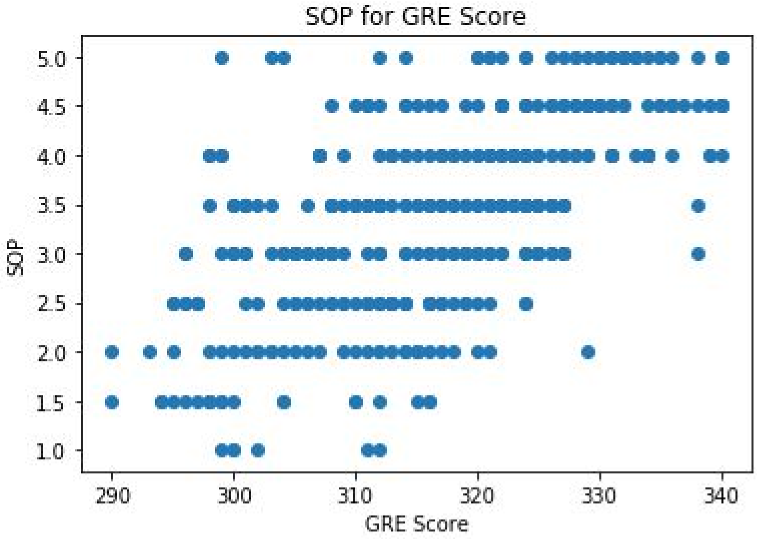
- **Candidates with high GRE scores usually have a high CGPA score:**



- **Candidates with high CGPA scores usually have a high SOP score:**

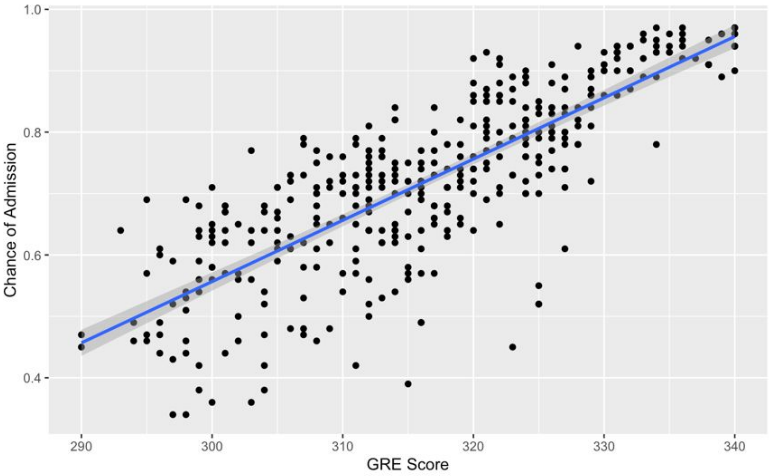


- **Candidates with high GRE scores usually have a high SOP score.**



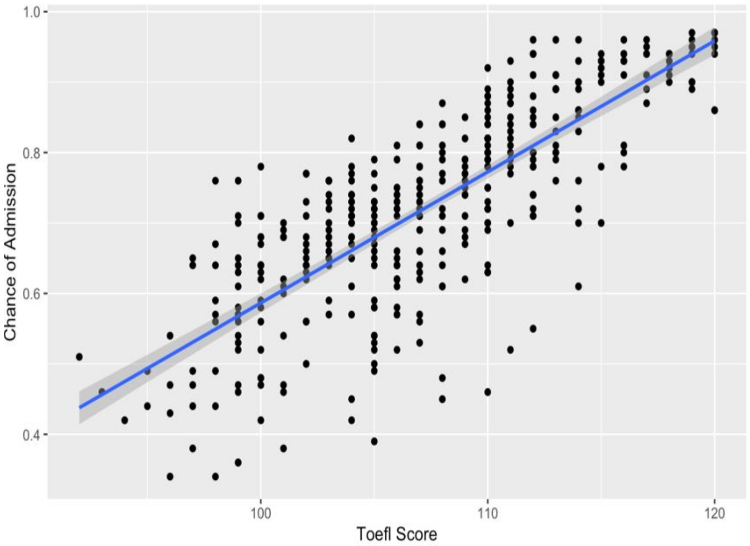
Analyses

4. Regression Analysis



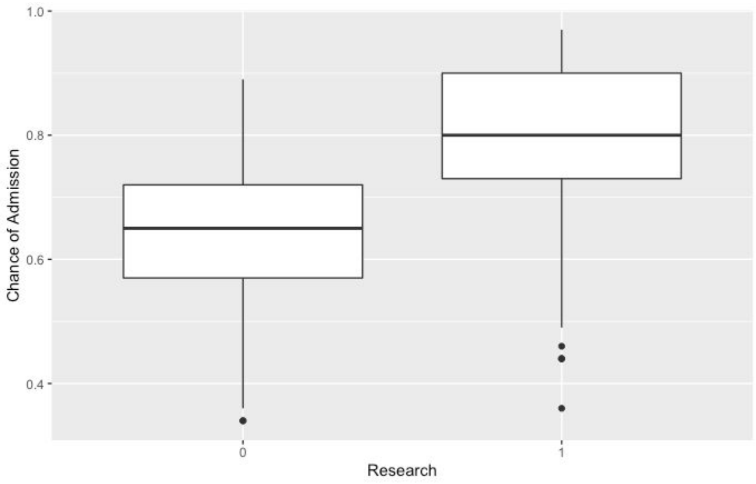
Term	Estimate Coefficient	Standard Error	Statistic	P-value
(Intercept)	-2.436084245	0.1178141186	-20.67735	4.876246e-65
GRE Score	0.009975882	0.0003716362	26.84314	2.458112e-91

From the above analysis, we can get the conclusion that given there is no relationship between GRE Score and Chance of Admission, there is a $2.46*10^{-91}$ probability of chance we get this sample results. Thus, we could reject the null hypothesis that there is no relationship between GRE Score and Chance of Admission, and conclude that there is a positive relationship between GRE score and the chance of admission for students. When GRE score increases 1, the chance of admission will increase 1% on average.



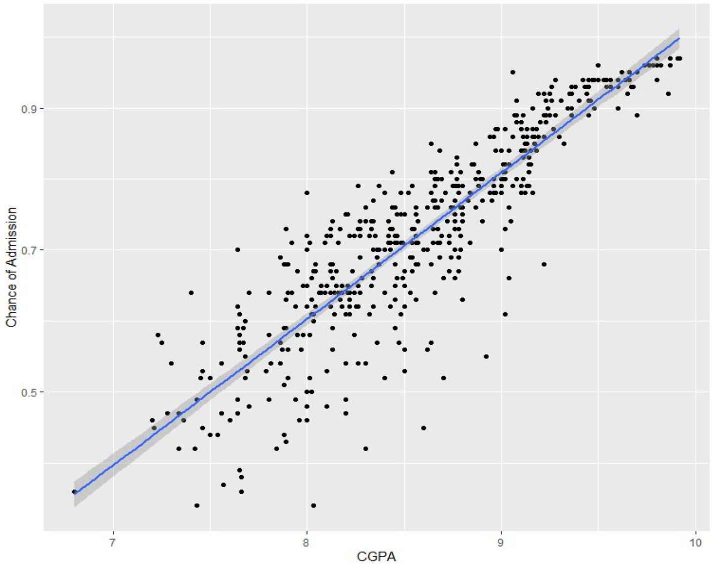
Term	Estimate Coefficient	Standard Error	Statistic	P-value
(Intercept)	-1.2734005	0.0774216975	-16.44759	9.443661e-47
TOEFL Score	0.0185993	0.0007196601	25.84456	3.634102e-87

From the above analysis, we can get the conclusion that given there is no relationship between TOEFL Score and Chance of Admission, there is a 3.63-87 probability of chance we get this sample results. Thus, we could reject the null hypothesis that there is no relationship between TOEFL Score and Chance of Admission, and conclude that there is a positive relationship between TOEFL score and the chance of admission for students. When the TOEFL score increases 1, the chance of admission will increase 1.86% on average.



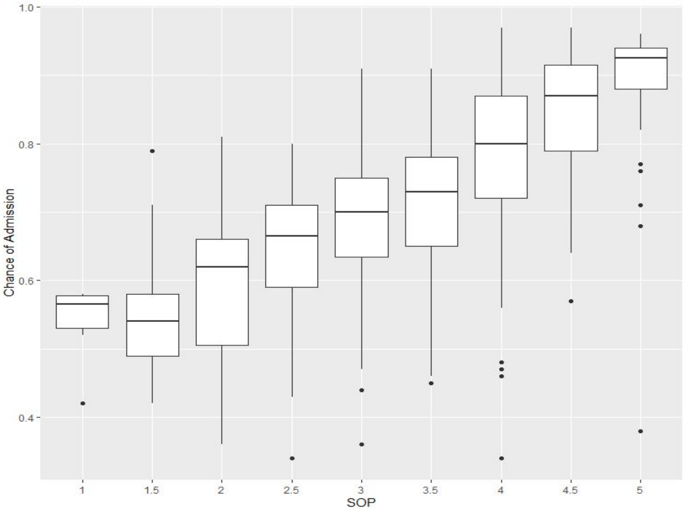
Term	Estimate Coefficient	Standard Error	Statistic	P-value
(Intercept)	0.6376796	0.008841442	72.12393	1.279681e-230
as.factor(adm$Research)1	0.1583022	0.011948986	13.24817	1.918173e-33

From the above analysis, we can conclude that given there is no relationship between Research and Chance of Admission, there is a 1.92*10-33 probability of chance that we can get this sample results. Thus, we could reject the null hypothesis that there is no relationship between Research and Chance of a student. For students who have done research, the chance of admission will be higher. On average, the chance of admission will increase by 15.83%.



Term	Estimate Coefficient	Std Error	Statistic	p-value
Intercept	-1.04	0.0423	-24.7	<2e-16
CGPA	0.206	0.00492	41.9	<2e-16

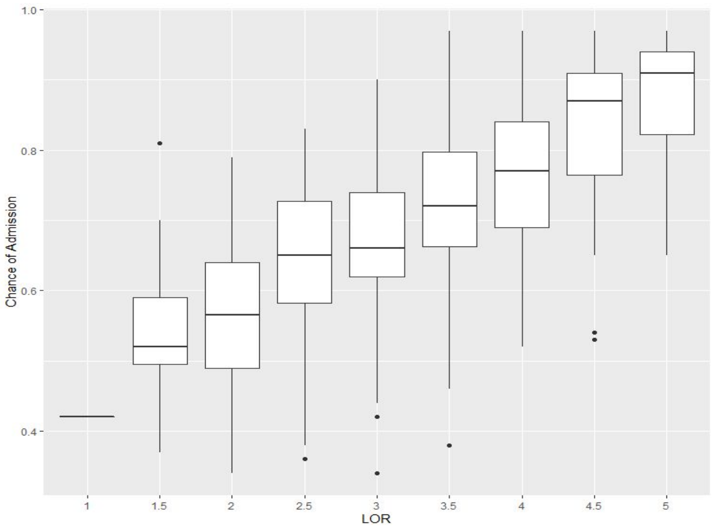
From the above analysis, we can reject the null hypothesis that is no relationship between CGPA and Chance of Admission, there is a 2* 10^-16 probability of chance that we can get this sample results. For students who have higher CGPA, the chance of admission will be higher. On average, the chance of admission will increase by 20.6%(0.206) for an increase of 1 point in CGPA



Term	Estimate Coefficient	Std Error	Statistic	p-value
(Intercept)	0.538	0.042	12.8	1.24E-31
as.factor(SOP)1.5	0.00807	0.0468	0.172	8.63e- 1
as.factor(SOP)2	0.0512	0.0448	1.14	2.54e- 1
as.factor(SOP)2.5	0.107	0.0439	2.44	1.52e- 2
as.factor(SOP)3	0.14	0.0436	3.22	1.38e- 3
as.factor(SOP)3.5	0.174	0.0434	4	7.28e- 5
as.factor(SOP)4	0.244	0.0434	5.63	2.99e- 8
as.factor(SOP)4.5	0.312	0.044	7.09	4.72E-12

as.factor(SOP)5	0.347	0.0449	7.72	6.60E-14

From the above analysis, we can reject the null hypothesis between SOP and Chance of Admission, there is a probability of chance that we can get this sample results. Thus, we could reject the null hypothesis that there is no relationship between SOP and Chance of admit. For students who have higher CGPA, the chance of admission will be higher.



Term	Estimate Coefficient	Std Error	Statistic	p-value
(Intercept)	0.42	0.108	3.88	0.00012
as.factor(LOR)1.5	0.13	0.113	1.15	0.251
as.factor(LOR)2	0.148	0.109	1.35	0.176
as.factor(LOR)2.5	0.221	0.109	2.02	0.0443
as.factor(LOR)3	0.248	0.109	2.28	0.0229
as.factor(LOR)3.5	0.303	0.109	2.78	0.00562
as.factor(LOR)4	0.344	0.109	3.16	0.00167
as.factor(LOR)4.5	0.412	0.109	3.77	0.000181
as.factor(LOR)5	0.453	0.109	4.14	4.13E-05

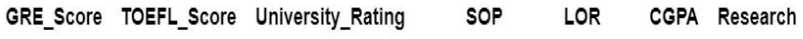
From the above analysis, we can reject the null hypothesis that given there is no relationship between LOR and Chance of Admission, there is an increase in the probability of chance that we can get this sample results. Thus, we could reject the null hypothesis that there is no relationship between LOR and Chance of admit. For students who have good-rated LOR, the chance of admission will be higher.

5. Model Selection

For selecting the model, we designed three models and then decided to come up with the best one out of these models.

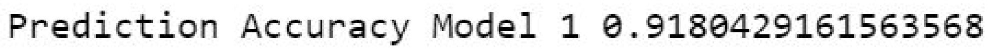
Model 1:

Model 1 included all the independent variables , namely:



```
GRE_Score   TOEFL_Score   University_Rating   SOP   LOR   CGPA   Research
```

The precision we got for model-1 was:

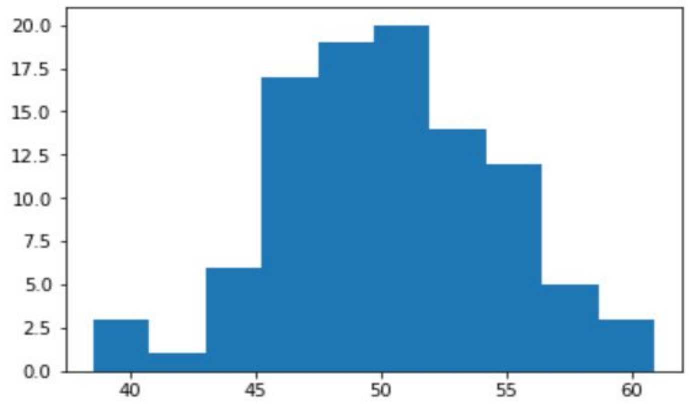


```
Prediction Accuracy Model 1 0.9180429161563568
```

We also did a normality test on the data to rule out any biases in the data.

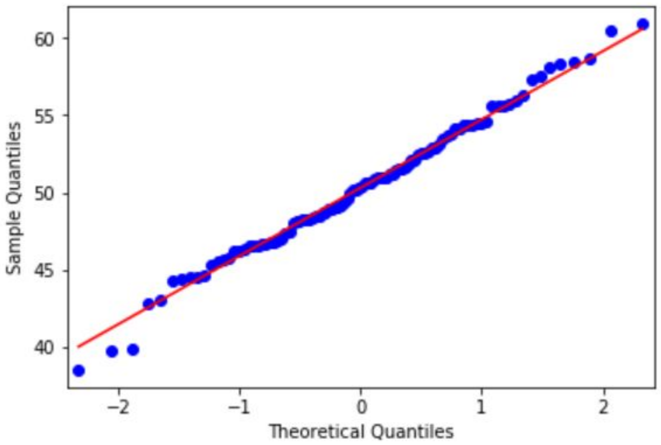
The methods for visually inspecting a dataset to check if it was drawn from a Gaussian distribution were:

Histogram:



Here we can see a Gaussian-like shape to the data, that although is not strongly the familiar bell-shape, is a rough approximation.

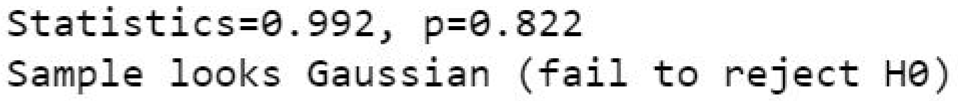
Quantile-Quantile Plot:



Here we get a QQ plot showing the scatter plot of points in a diagonal line, closely fitting the expected diagonal pattern for a sample from a Gaussian distribution.

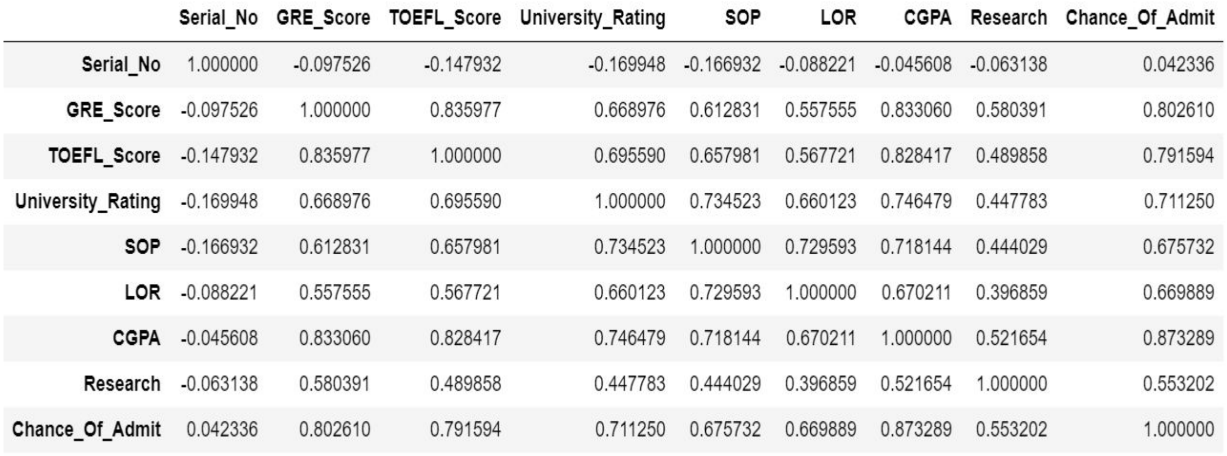
Shapiro-Wilk test:

The Shapiro-Wilk test evaluates a data sample and quantifies how likely it is that the data was drawn from a Gaussian distribution, named for Samuel Shapiro and Martin Wilk.


```
Statistics=0.992, p=0.822
Sample looks Gaussian (fail to reject H0)
```

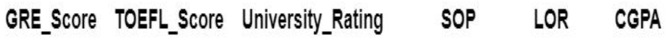
Model 2:

For Model-2 we selected the predictors determined by the correlation matrix. We dropped "Research" which had the least correlation with the dependent variable "Chance of Admit".



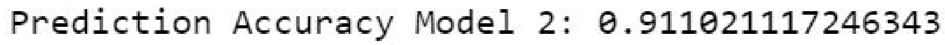
	Serial_No	GRE_Score	TOEFL_Score	University_Rating	SOP	LOR	CGPA	Research	Chance_Of_Admit
Serial_No	1.000000	-0.097526	-0.147932	-0.169948	-0.166932	-0.088221	-0.045608	-0.063138	0.042336
GRE_Score	-0.097526	1.000000	0.835977	0.668976	0.612831	0.557555	0.833060	0.580391	0.802610
TOEFL_Score	-0.147932	0.835977	1.000000	0.695590	0.657981	0.567721	0.828417	0.489858	0.791594
University_Rating	-0.169948	0.668976	0.695590	1.000000	0.734523	0.660123	0.746479	0.447783	0.711250
SOP	-0.166932	0.612831	0.657981	0.734523	1.000000	0.729593	0.718144	0.444029	0.675732
LOR	-0.088221	0.557555	0.567721	0.660123	0.729593	1.000000	0.670211	0.396859	0.669889
CGPA	-0.045608	0.833060	0.828417	0.746479	0.718144	0.670211	1.000000	0.521654	0.873289
Research	-0.063138	0.580391	0.489858	0.447783	0.444029	0.396859	0.521654	1.000000	0.553202
Chance_Of_Admit	0.042336	0.802610	0.791594	0.711250	0.675732	0.669889	0.873289	0.553202	1.000000

Thus our predictor variables for Model-2 were:



```
GRE_Score   TOEFL_Score   University_Rating      SOP      LOR      CGPA
```

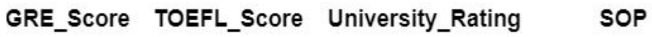
The precision we got for Model-2l was:



```
Prediction Accuracy Model 2: 0.911021117246343
```

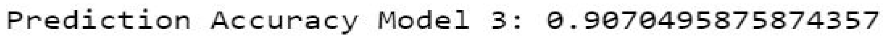
Model 3:

The third model was designed by dropping the least two correlated predictors, which were "Research" and "LOR".
The predictors used were:



```
GRE_Score   TOEFL_Score   University_Rating        SOP
```

And the precision accuracy we got for Model-3 was:



```
Prediction Accuracy Model 3: 0.9070495875874357
```

Conclusion

From the above 3 designed models, we see that the precision is almost the same, with a difference in the precision of maximum +/ - (0.1). Now the deciding factor for us to zero down on a model was the simplicity of the model. We concluded selecting Model-3 for our use as Model-3 was the simplest model with only 4 predictors but a very good precision.of 0.90.

The benefits of selecting a simple model are:
1. Fewer chances of overfitting: The data models with a large number of predictors (also referred to as complex models) often suffer from the problem of overfitting, in which case the data model performs great on training data, but performs poorly on test data.

2. Productivity: With a large number of variables, we may be many times enticed to use all the predictors to have a data model with very high success rate, but practical considerations (such as the amount of data available, storage and compute resources, time is taken for completion, etc.) make it nearly impossible.

Thus, even when you have a large number of relevant predictor variables, it is a good idea to work with fewer predictors (shortlisted through feature selection or developed through feature extraction). This is essentially similar to the Pareto principle, which states that for many events, roughly 80% of the effects come from 20% of the causes.

Focusing on those 20% most significant predictor variables will be of great help in building data models with a considerable success rate in a reasonable time, without needing a non-practical amount of data or other resources.

3.Understandability: This factor is particularly important if at the end of your project you need to present your results to someone interested in not just high success rate, but also in understanding what is happening “under the hood”.