GR521 Team– 4
Sagar B
Dhiab A
Dhruv T
December 9th, 2019

Project Report - Graduate Admission Prediction

Abstract

This project aims to evaluate several parameters which are considered important during the application for Masters Programs. When we went through the process of preparing, applying, and collecting required documents for the admission, we were less informed about how and in what percentage the individual documents will contribute to our selection. Here in this study, we will try to find interesting pattern on how different features relate to *chances of admit*. The results will surely reveal something insightful. In addition, we have emphasized on making this report comprehensible for non-statistical/technical audiences so the results can then be used by the potential students in strengthening their profile accordingly.

Introduction and Data Characteristics

We are looking at the university data and different admission scores, how it's correlated to university ranks and if there is a way to predict the admission based on those scores. The hypothesis question we would like to answer is whether the score will be significantly higher for higher ranking university. The data has 400 different results that contain GRE score, TOEFL score, University Rating, SOP, LOR, CGPA, Research (if there's a research done) and chance of admission based on those scores.

For the graduate admission prediction dataset, following is the brief description of the dataset and its features:

Quantitative features:
GRE Score - ranges from 260 to 340, integer data type.
TOEFL Score - ranges from 0 to 120, integer data type.
SOP strength - ranges from 1 to 5, decimal data type.
LOR strength - ranges from 1 to 5, decimal data type.
CGPA - maximum 10, decimal data type.

Qualitative feature:
Research Experience – 0 or 1.
University Rating - ranges from 1 to 5.

Total number of independent features: 7.

Total number of observations: 400.

Target Variable: Chance of Admit - ranges from 0 to 1, decimal data type.

(Source: **https://www.kaggle.com/mohansacharya/graduate-admissions#Admission_Predict_Ver1.1.csv**)

Here, we can see that the dataset is very rich with useful information. It contains all the major contributing factors which result in an admit or rejection. We have different types of data such as qualitative (ordinal, nominal) and quantitative with different ranges. As we observed, the GRE score has a minimum possible value of 260 and maximum possible value of 340. Similarly, TOEFL score can go as low as 0 and as high as 120. Both of them are integer type i.e., the values are non-decimal. In contrast, the SOP and LOR strength variables are decimal variables. They can go up to one decimal value. Both of them ranges from 1 to 5. At last, CGPA can have up to two decimal values, ranging from 0 to 10. (Note: Although the data may not contain the lowest values for all the quantitative variables, the variables can have those values if we add more data in future).

For qualitative variable, we have university rating and Research experience. Even though the university rating looks like a quantitative variable, on close inspection we can observe that they have an intrinsic order in them, Also, the numbers 1 to 5 do not make much sense in terms of numeric/arithmetic properties i.e., we could have used A to E instead. The properties that matter is that they are distinct and have an order property.

In the next sections, we will start with the descriptive evaluation and visualization of data to get a better understanding of it. After that, we will construct a methodology suitable to the data we have. This will help us in focusing on important features critical to increase the chance of admit and make recommendation accordingly. We will then provide a conclusion from the results and corresponding interpretations we get.

Descriptive summary

First, we calculated averages and mediums of each score. The results are similar for both which means that our data is symmetric. This table below also shows that average scores will give 72% chance of admission. Half of the average admissions include research paper. Median university rank is 3 out of 5. Standard deviation for TOEFL and GRE score shows that there is a dispersion (spread) in the scores, meaning that some admission scores submitted were much better and/or some worse than the average. CGPA scores have very low deviation meaning that they were consistent.

Table 1: Summary statistics of Graduate Admission

Measure	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Average	316.8	107.4	3.1	3.4	3.5	8.6	0.5	72%
Median	317.0	107.0	3.0	3.5	3.5	8.6	1.0	73%
Standard Deviation	11.5	6.1	1.1	1.0	0.9	0.6	0.5	14%

Since the average university rank is 3 and the university rank 3 shows the highest frequency compared to other university ranks shown in the figure below. We decided to look for the average scores' person needs to have a higher chance (90%) of being admitted and what are the lower scores with less chances of being admitted (36% chance). We also looked at the higher rated university scores, what is needed to be admitted to such university and what scores are required. The table below shows that with 90% chance of admission the scores are the same for university that is ranked 3 and 5. We noticed that the chance of admission is much less when there is no research submitted for admission. It is also showing that the highest chance of admission (97%) for the highest ranked university (5 out of 5) requires the highest scores for everything. In other words, whoever receives the highest scores plus submits a research work has almost 97% chance of being admitted.

Graph 1: Histogram for University Rating

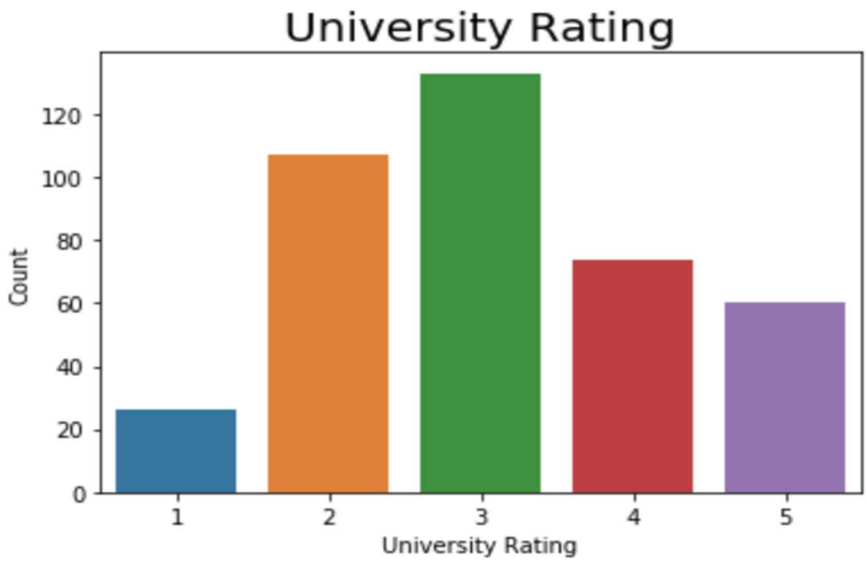


Table 2: Average profiles with highest and lowest chance of admits for 3 and 5 rated universities.

University rating	Chance of Admit	Average of TOEFL Score	Average of CGPA	Average of LOR	Average of SOP	Average of GRE Score	Average of Research
3	90%	114	9	5	5	330	1
3	36%	99	8	3	2	303	0
5	97%	120	10	4	4	337	1
5	90%	113	9	4	4	331	1
5	61%	108	8	3	3	305	0

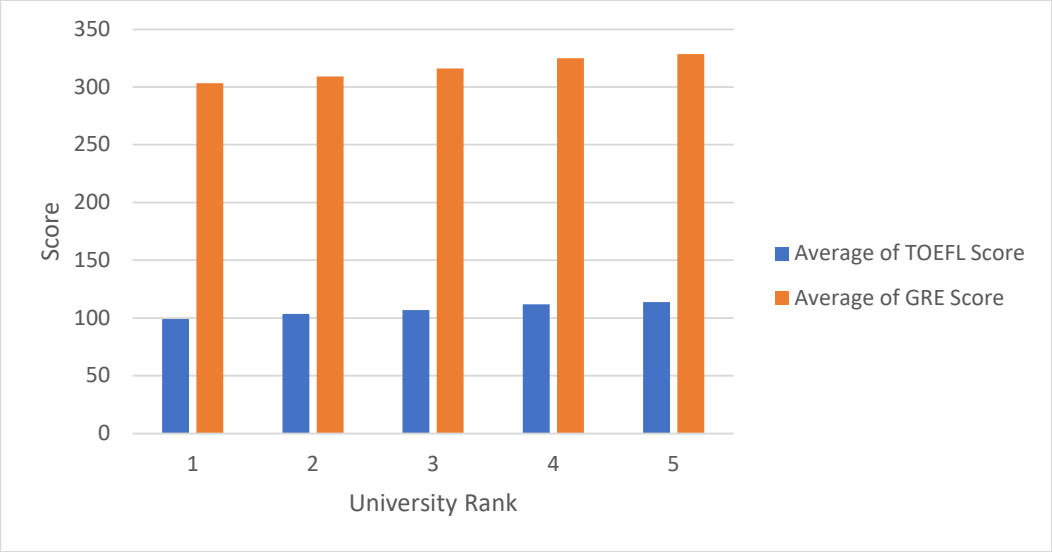
Following table 3 showing values at the 90th percentile for each feature. For universities with rating "3", it suggests that the 325.6 is greater than the 90 percent of student's GRE scores. Similarly, 112 for TOEFL score, 4 for SOP Strength, 4.5 for LOR strength, 9.04 for CGPA and 1 for Research. For the five rated universities, the values are significantly high. It is 338.9 for GRE score, 119 for TOEFL score, 5 for SOP Strength, 5 for LOR strength, 9.778 for CGPA and 1 for Research. It is clear that scores have to be higher for higher ranked universities.

Note that, 90th percentile for 3 and 5 rated universities is different than having the 90% chance of admission for 3 and 5 rated universities. The later is actually the value of our target variable as "Chances of Admit" tells us the percentage. Whereas 90th percentile is the calculation done on all of the values and getting the one having 90% of the values below it.

Table 3: Student profiles at 90th percentile of chance of admit for 3 and 5 rated universities

	University Rating	GRE Score	TOEFL Score	SOP	LOR	CGPA	Research
90th percentile	3	325.6	112	4	4.5	9.04	1
90th percentile	5	338.9	119	5	5	9.778	1

Graph 2: Average TOEFL scores and average GRE scores by university rank



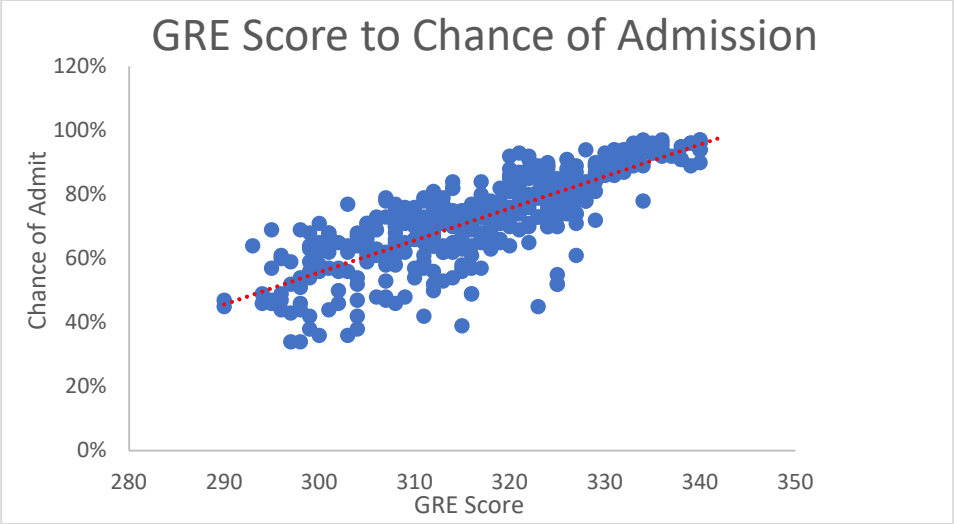
The Graph 1 above displaying two admission measures TOEFL and GRE exams from the 400 samples that shows that there is a definite increase of needed scores to be admitted to a higher ranked university. However, our team decided to exclude the independent variable TOEFL Scores as we believe that not all students are required to take the TOEFL test to join Graduate School. We based this off of our experience when applying to different Grad Schools. TOEFL is usually only required for Undergraduate admissions which is not the case of our dataset. It makes sense to exclude TOEFL score as we will see later on in our methodology section that it proved not to be a useful predictor of our model. Moreover, the variations in the TOEFL score distorts some of our model assumptions such as the assumption of the constant variance.

Methodology

First, we start off by exploring the correlation of our independent variables with our dependent variable "Chance of Admit" by using some scatterplots and correlation matrices.

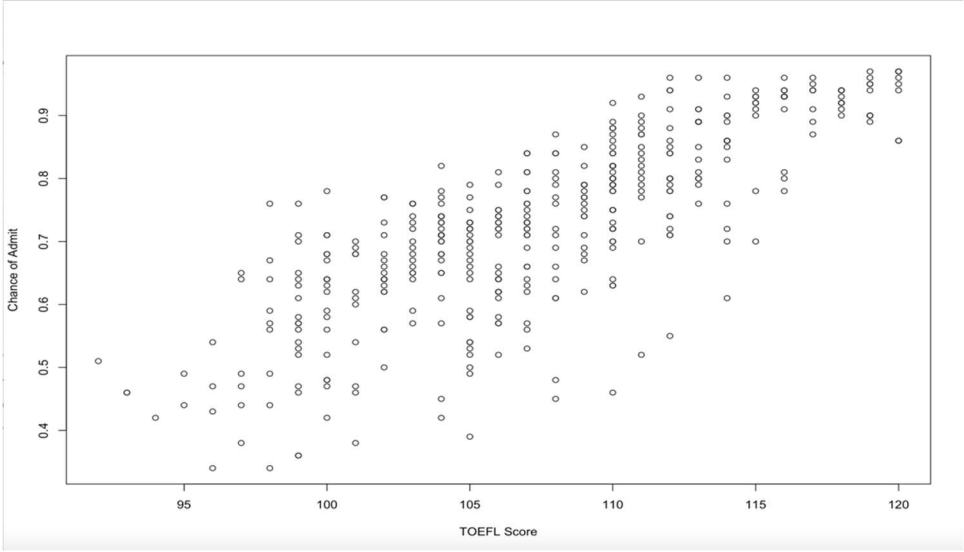
The diagram below shows GRE scores to Chance of Admission that shows us that the higher the score the higher the chance to be admitted.

Graph 3: Scatter plot between chance of admit and GRE score



The diagram below shows TOEFL scores to Chance of Admission that shows the higher the score the higher the chance of admission. However, we can see the extreme variations in the TOEFL scores with regards to the Chance of Admit which solidifies our decision statistically for not using it as a variable to predict the chance of admit.

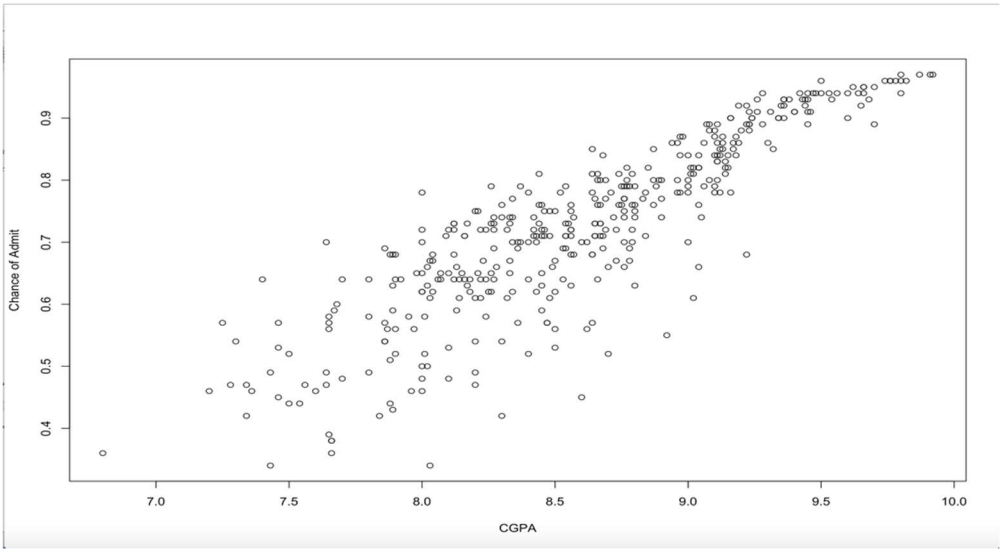
Graph 4: Scatter plot between chance of admit and TOEFL score



The diagram below shows a scatterplot of x = CGPA and y = Chance of Admit. We can see a high correlation between CGPA and the chance of admit. We see an interesting trend here as the CGPA gets higher we see a lower variation in the chance of admit. Contextually speaking this makes sense because universities usually require a range of CGPA. For example, universities won't require a specific CGPA like exactly 9 or 10 to admit students they ask to have a CGPA of 9 or above. Therefore, a CGPA of 9 could have the same chance of admit as a student with a 9.5 CGPA. Moreover, the reason why sees less variations in the chance of admit could also be because the number of students with a high CGPA are less than students with an average or lower CGPA. Overall, we think

CGPA will be a good predictor for our model because of our experience and we will see below the correlation matrix shows the highest correlation with the chance of admit (0.87328910).

Graph 5: Scatter plot between chance of admit and CGPA



Graph 6: Correlation Matrix of Variables

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR
Serial No.	1.00000000	-0.09752579	-0.1479317	-0.1699479	-0.1669324	-0.08822139
GRE Score	-0.09752579	1.00000000	0.8359768	0.6689759	0.6128307	0.55755452
TOEFL Score	-0.14793170	0.83597680	1.0000000	0.6955898	0.6579805	0.56772092
University Rating	-0.16994786	0.66897585	0.6955898	1.0000000	0.7345228	0.66012345
SOP	-0.16693236	0.61283074	0.6579805	0.7345228	1.0000000	0.72959254
LOR	-0.08822139	0.55755452	0.5677209	0.6601235	0.7295925	1.00000000
CGPA	-0.04560845	0.83306045	0.8284174	0.7464787	0.7181440	0.67021130
Research	-0.06313754	0.58039064	0.4898579	0.4477825	0.4440288	0.39685926
Chance of Admit	0.04233586	0.80261046	0.7915940	0.7112503	0.6757319	0.66988879
y	0.07023445	0.38589273	0.3920317	0.2695692	0.2683720	0.33854404

	CGPA	Research	Chance of Admit	y
Serial No.	-0.04560845	-0.06313754	0.04233586	0.07023445
GRE Score	0.83306045	0.58039064	0.80261046	0.38589273
TOEFL Score	0.82841742	0.48985785	0.79159399	0.39203167
University Rating	0.74647869	0.44778251	0.71125025	0.26956920
SOP	0.71814396	0.44402881	0.67573186	0.26837199
LOR	0.67021130	0.39685926	0.66988879	0.33854404
CGPA	1.00000000	0.52165423	0.87328910	0.45031051
Research	0.52165423	1.00000000	0.55320214	0.22030581
Chance of Admit	0.87328910	0.55320214	1.00000000	0.59928769
y	0.45031051	0.22030581	0.59928769	1.00000000

Table 4: Step Wise Regression Analysis

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change
1	.873[a]	.762	.762	.06964	.762
2	.884[b]	.781	.780	.06690	.019
3	.891[c]	.794	.792	.06499	.013
4	.894[d]	.799	.797	.06431	.005

a. Predictors: (Constant), CGPA

b. Predictors: (Constant), CGPA, GREScore

c. Predictors: (Constant), CGPA, GREScore, LOR

d. Predictors: (Constant), CGPA, GREScore, LOR, Research

Step Wise Analysis is a process of fitting regression model in which the choice of predictive model is carried out by an automatic process.

From the above table 4, we can see a *Stepwise Analysis* to figure out which are the most important variables which impacts the chance of admit and to what extent. To run a step wise analysis, we consider "Chance of Admit" as the dependent variable, while remainder of the factors are considered as independent variables. From the Table 4, we can see that we have CGPA, GRE Score, LOR, and Research as the most important variables on which the Chances of Admit depends. Again, if we look at the Table 4, and if we see the second column *R square*, we can interpret that CGPA explains 76.2% of variation in Chance of Admit making CGPA the most important factor while calculating chances of admission.
Four of the variables combined help is explaining the variation by ~80% which is significantly high and thus, tells us that we have a good chance of predicting the model.

Remaining 20% of the variation can include other factors such as Ambition of a student or maybe, other subjective factors which we don't have the data for. Further Market Research study will help us to enable more insightful actions/recommendations to build an even robust model.

Table 5: Coefficient Summary for Step Wise Analysis

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-1.331	.118		-11.307	.000
	GREScore	.003	.001	.207	4.816	.000
	LOR	.023	.005	.145	4.751	.000
	CGPA	.134	.011	.559	12.255	.000
	Research	.025	.008	.086	3.080	.002

a. Dependent Variable: ChanceofAdmit

In the table 5, we can observe the **coefficient value** (Unstandardized B) of each of the independent variables. Considering that we have an equation as $\mathbf{Y = m_1x_1 + m_2x_2 + m_3x_3 + m_4x_4 + c}$

$\mathbf{x_1}$ = GRE Score
$\mathbf{x_2}$ = LOR
$\mathbf{x_3}$ = CGPA
$\mathbf{x_4}$ = Research

The sign of Unstandardized B – gives us the direction of the relationship independent variables has with the dependent variables. Since, all the independent variables are positive, this implies that they have a positive impact on the Chance of Admit.

Basis the above equation, we can come up with an equation such as:

Chance of Admit = 0.003*GRE + 0.023*LOR + 0.134*CGPA + 0.025*Research

a) Confidence Interval

Here, we will construct the confidence interval for each of the independent feature which are significant based on the previous Step Wise Analysis we performed as well as the target variable. Those features are GRE Score, LOR, CGPA and Research. We know from the Table 1, the mean for each of them. They are 316.8, 3.5, 8.6 and 0.5 for GRE Score, LOR, CGPA and Research respectively. We then calculated the Margin of error with alpha as 0.05 (95% confidence level), Standard deviation (specific to each of them), and the sample size of 400.

Table 6: Summary statistics of Graduate Admission

Features	Mean	Margin of error	Confidence Interval with 95% confidence level
GRE Score	316.8	1.12	[315.68, 317.92]
LOR	3.5	0.09	[3.41, 3.59]
CGPA	8.6	0.06	[8.54, 8.66]
Research	0.5	0.05	[0.45, 0.55]
Chance of Admit	72	1.37	[70.63, 73.37]

Interpretations:

1) GRE Score:
   At 5% significance level, we can say that the average population value of the GRE Score will lie in between 315.68 and 317.92.

2) LOR:
   At 5% significance level, we can say that the average population value of the LOR will lie in between 3.41 and 3.59.

3) CGPA:

At 5% significance level, we can say that the average population value of the CGPA will lie in between 8.54 and 8.66.

4) Research:
   At 5% significance level, we can say that the average population value of the Research will lie in between 0.45 and 0.55.

5) Chance of Admit:
   At 5% significance level, we can say that the average population value of our target variable (Chance of Admit) will lie in between **70.63% and 73.37%**.

These intervals will help us in inferring the possible mean value range for different samples we evaluate for the same population. In addition, we get an important insight about our target variable *"Chance of Admit"*. We noted that on average the mean of chance of admit varies between **70.63% and 73.37%**. This is very important because the range is **significantly far from 50%** i.e., many more than half of the students have a good chance to get an admit from the university they apply and they actually get it. This information can also be used by the future applicants as a **motivation** to apply and expect to get an admit.

b) Hypothesis Testing / ANOVA

Early on in our introduction to the data set we mentioned about the hypothesis to establish that higher university ranking, the higher the chance of admit. Now we will do a hypothesis test to establish the fact that there exists a relationship between University Ranking and Chance of Admit using one way ANOVA.

*- For simplification, going forward, we will call **University Ranking as UR, Chance of Admit as CoA!**

For Hypothesis Testing, our first step is to create *Null* & *Alternative* hypothesis.

H_0 : $\mu 1 = \mu 2$ (There exists no relation whatsoever b/w UR and CoA)
H_a : $\mu 1$ is not equal to $\mu 2$ (There exists a relationship b/w UR and CoA)
where $\mu 1$ = Average of University Ranking
and $\mu 2$ = Average of Chances of Admit

For this hypothesis, we will use $\alpha = 0.05$ = Significance Value and now, we will check which of the hypothesis is true using ANOVA! We will be doing an ANOVA test between University Rating and Chance of Admit.

Table 7: Summary statistics of Graduate Admission

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
University Rating	400	1235	3.0875	1.308114035
Chance of Admit	400	289.74	0.72435	0.020337421

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1116.895585	1	1116.895585	1681.500034	1.2099E-198	3.853138126
Within Groups	530.052131	798	0.664225728			
Total	1646.947716	799				

From the above ANOVA analysis, we can observe that
p-value = 1.2099E-198

Since our p-value is less than α, we will reject the null hypothesis and can call out that our result is statically significant.

Now, let's go back a few steps and recall, what did we say in our Alternate Hypothesis (H_a), our alternate hypothesis calls out that $\mu 1$ is not equal to $\mu 2$ and hence, there exists a relation between University Ranking and Chance of Admit.

From the above analysis of variance test (ANOVA), we can conclude that the higher ***the university ranking, the higher the chances of admit.***

c) Regression Analysis

We had to do some outside research for what regression method would be best and we decided to go with regression through the origin with no Beta-0 intercept. This is because it makes sense contextually that there should be no intercept for the chance of admit because if a student has 0 scores for the variables analyzed above his/her chance of admit would be zero. Moreover, after exploring the data and the relation between variables we found that students that are applying to higher ranked universities need higher scores for GRE. Therefore, we decided to try out an interaction term using University Rating with different variables. The interaction term between GRE Scores*University Rating appeared to be the most significant. Below is the model initiated using R-studio.

model = lm (`Chance of Admit` ~ 0 +`GRE Score`*University Rating + CGPA + LOR+ Research)

Y	Chance of Admit
X1 = 0.0232654	LOR
X2 = 0.1313719	CGPA
X3 = - 0.1313719	GRE score
X4 = -0.3802179	University Rating
X5 = 0.0012303	University Rating*GRE Score

Y = 0.0232654x1 + 0.1313719x2 – 0.1313719x3 – 0.3802179x4 + 0.0012303

```
> model = lm(`Chance of Admit`~  0 + LOR + CGPA +`GRE Score`*`University Rating`)
> summary(model)

Call:
lm(formula = `Chance of Admit` ~ 0 + LOR + CGPA + `GRE Score` *
    `University Rating`)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26509 -0.03257  0.01089  0.04030  0.16961

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
LOR                             0.0232654  0.0052346   4.445 1.15e-05 ***
CGPA                            0.1313719  0.0118022  11.131  < 2e-16 ***
`GRE Score`                    -0.0016596  0.0002923  -5.678 2.65e-08 ***
`University Rating`            -0.3802179  0.0360077 -10.559  < 2e-16 ***
`GRE Score`:`University Rating` 0.0012303  0.0001074  11.460  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.066 on 395 degrees of freedom
Multiple R-squared:  0.9921,    Adjusted R-squared:  0.992
F-statistic:  9930 on 5 and 395 DF,  p-value: < 2.2e-16
```

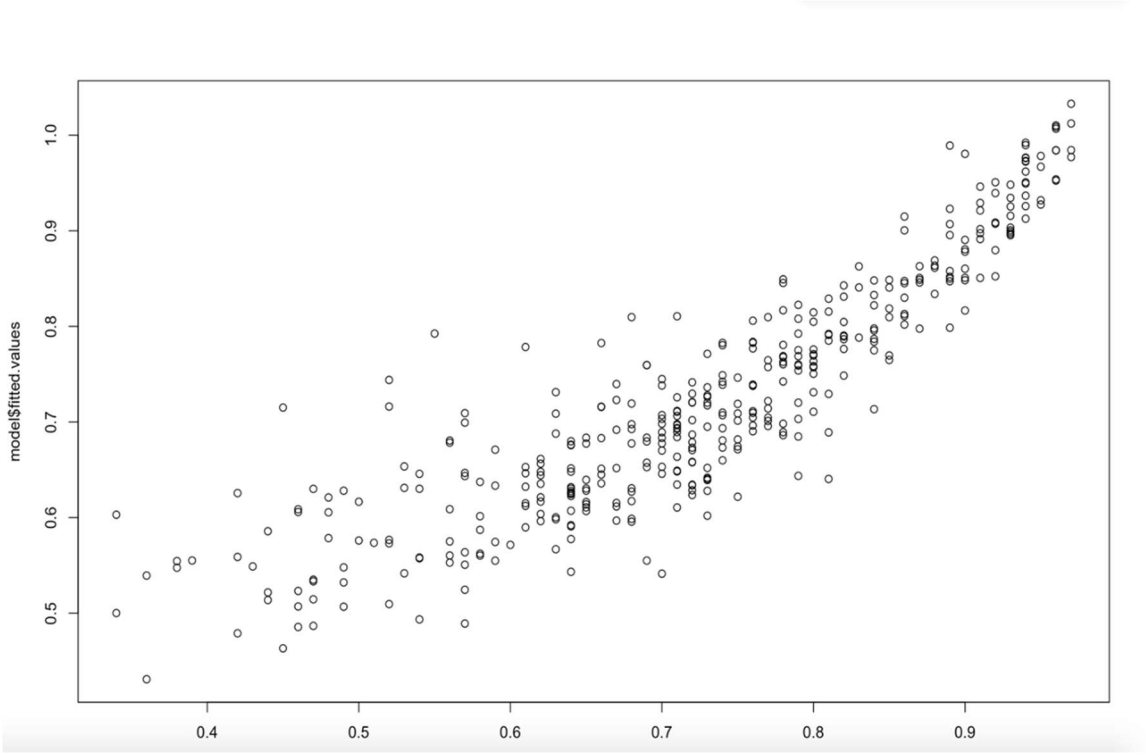
Above we can see the out- put of R shows a highly significant F-test (p-value < 2.2*10^-16) which tells us that this is a highly useful model for predicting the chance of admission. Moreover, the individual t-tests for our independent variables show that our predictors are highly significant at the alpha = .01 level. The adjusted R-squared of our Model shows that our model can predict the chances of admission with high precision by explaining approximately 99% of the observed variations in the chances of admissions in our dataset. Next, we start checking for any Multicollinearities using the Vif function in R which represents the variance inflation factors of the model. Any Vif > 5 is usually representing that multicollinearity exists in our model. Below we can see that there is Multicollinearity. However, after doing some research we learnt that Multicollinearity is not always a problem. It

is a problem when the global F-test shows a significant p-value and the individual t-tests are distorted. This is not our case because all the individual t-tests are very highly significant. Therefore, in this case Multicollinearity is not causing any problems to our model.

```
> vif(model)
                     LOR                     CGPA                `GRE Score`
                32.01926                950.35110                 788.58080
`University Rating` `GRE Score`:`University Rating`
                1290.34895                1189.99926
```

Checking Observed values vs Predicted Values (Y vs Y-hat) and Model Assumption of Constant Variance

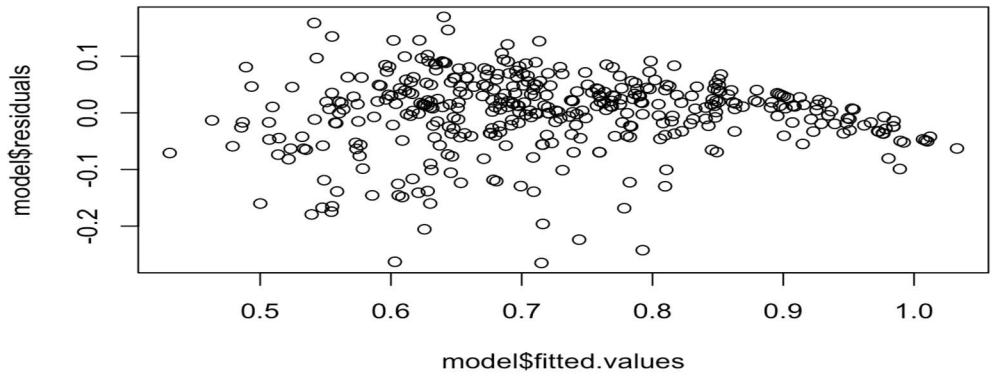
Graph 6: Plot(Chance of Admit, model$fittedvalues)



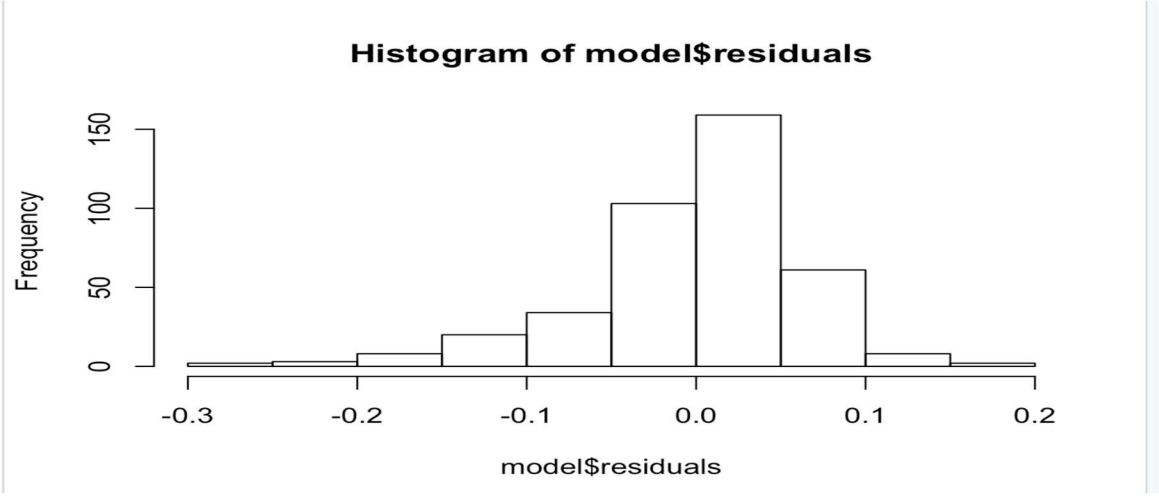
It is quite clear to us that the model can precisely predict the chances of admission between 0.8 and 0.9 (80% - 90%). In contrast, the model does poorly at predicting the chances of admission below 70%. In the diagram below we can also see that the unobserved variables which is known as residuals or predicted errors have a higher impact

on lower chances of admission compared to the higher admissions. The problem of constant variance here known as heteroscedasticity is seen. This implies a misspecification of the model or that we should use some techniques such as logistic regression to transform some variables and go through the process of variance stabilization. This is some advanced statistics that we did not cover in class and hope to cover in upcoming stats courses. However, we believe there are some contextual reasons that may have caused this heteroscedasticity. Note, that the explained by variance of the model is extremely high. One reason for heteroscedasticity is that our dataset and model do not encompass how determined or ambitious the student applying is. For instance, a student who has high scores on all tests and strong Letters of Recommendations is likely to seem to the recruiter as highly ambitious and the recruiter would highly admit the student. The high ambition of the student could have highly contributed to the chances of admission. This shows that there are some unobserved factors that have not been taken into account that contributes highly to the high chances of admission. Moreover, we know that beyond a certain threshold student will be admitted instantly. While students with lower scores will be admitted based on other factors that are only known to the recruiter. Also, this depends on the recruiter himself some recruiters value some factors more than other factors. Therefore, we kind of get a sense that the heteroscedasticity arising might not be from the model misspecification but could be from other factors that are not captured by the dataset.

Graph 7: Plot(model$fittedvalues, model$residuals)

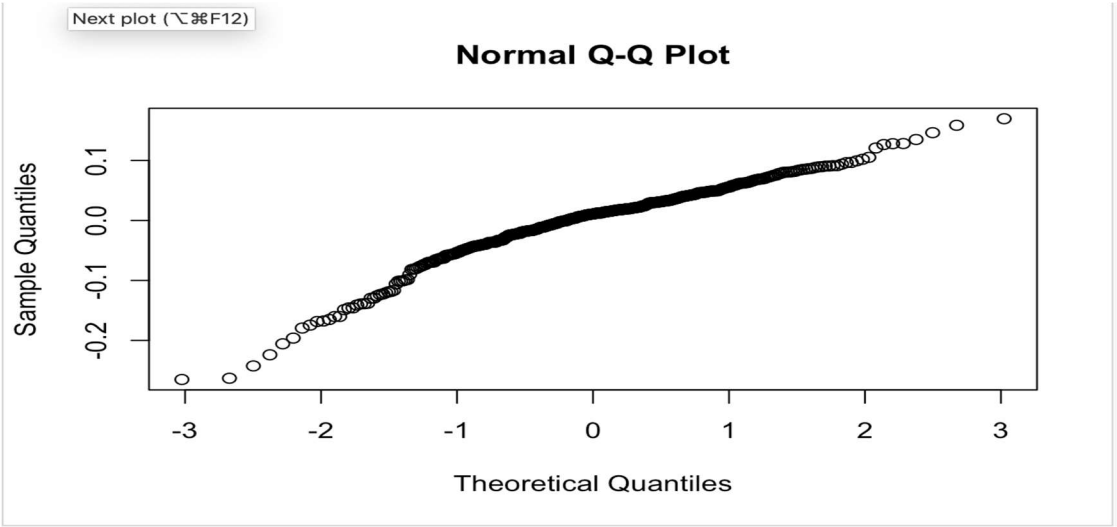


Graph 8: Hist(model$residuals) Checking the assumption of Distribution of estimated errors



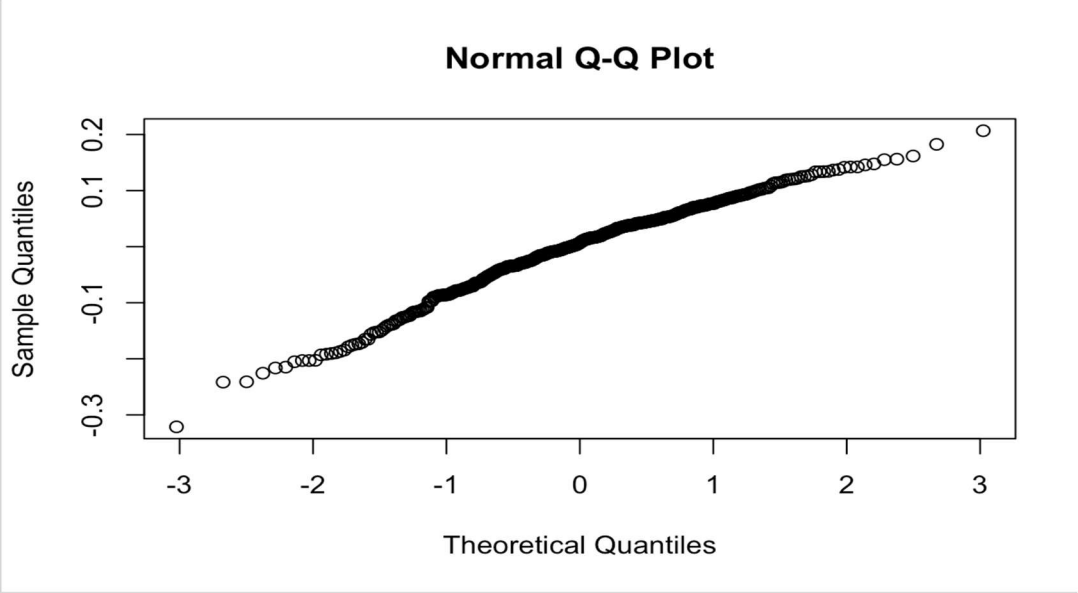
Here checking the distribution of the model residuals, we can see that it is fairly negatively skewed, and it is not satisfying the model assumption of normality. On the next page, we used the q-q norm function and we see a slight curvature which shows that it doesn't satisfy the model assumption that errors are normally distributed ***(Graph 9).*** A QQ plot combines a normally distributed set of quantiles and if they are both normally distributed that means we should see a straight line. In our case we see a slight curvature which further enhances that the model assumption of the errors being normally distributed is violated

Graph 9: QQnorm Checking the assumption of Distribution of estimated errors



After some research we decided to transform some terms to try to satisfy the model assumption. Therefore, our new model is model2 = lm (`Chance of Admit`~ 0 + LOR + CGPA +log (`GRE Score`*`University Rating`)). By adding the term log () to the interaction term we see in **Graph 10** is that the curvature is slightly fixed. Also, all the p-values for the variables and the global F – test are significant with a slightly lower adjusted R-square of .98. This means that our model explains 98% of the observed variations in the chances of predictions of admitting students.

Graph 10: QQnorm with the log() term



Conclusion and Future Directions

Based on the above findings, or the analysis that we have done/presented, we can infer the below mentioned info:

- We are 95% confident that the mean of all chance of admits lies between 70.63% and 73.37% i.e., many more than half of the students who apply get an admit.
- Using ANOVA, we concluded that higher the GRE Score, more are the chances of a student getting admitted to a high university ranking!
- CGPA, GRE Score, LOR, and Research comes out to be the most important factors among all the variables on which the chances of admission depend.
- Since all of these have a positive relationship with the Chance of Admission, higher the independent factors, more will be the Chance of Admission.

Our report is open to use for future research as the data may change and we may find more data features with all of them that were given to us. We would suggest to collect data that involves some of the following:

- The dataset didn't include the circumstances in which a student applied in. For instance, a student can be privileged in terms of financial resources. He may have used his family reputation or wealth in getting an admit. This isn't captured in the dataset which could be a reason for inaccuracy percent of the model.
- In addition, two universities with the same rating can have a totally different process of student selection. Some universities focus more on the CGPA of the candidate and others emphasize on the GRE score. So,

having the University name and selection criteria for each record in the dataset can truly help us in further digging into the patterns and going more granular.

References

- 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
- Jaggia, Sanjiv, and Alison Kelly. 2019. Business statistics: communicating with numbers, 3rd edition.
- Statistics for Managers Using Microsoft Excel, 8th edition, by David M. Levine, David F. Stephan, and Kathryn A. Szabat.