James Avery **3/12/19**

Admissions Study

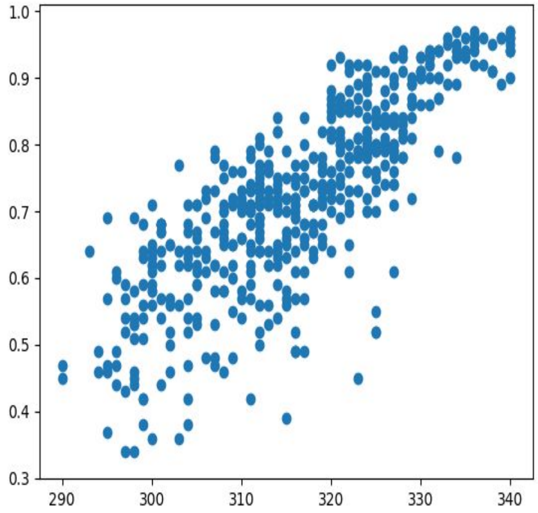
This will be an examination and comparison of multiple data analysis models in order to better understand each one and how they function, and to also learn which model is the most effective for this project.

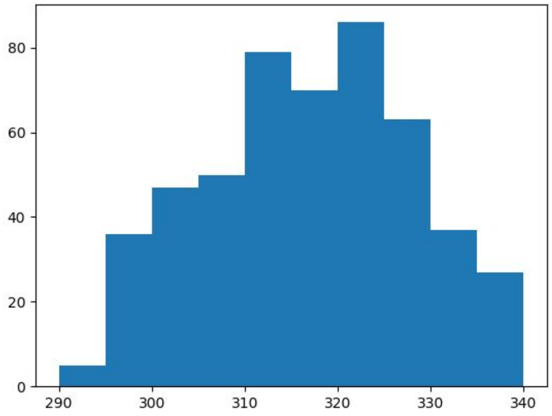
1. Statistical Model

This is a purely statistical model, meaning that it's not leveraging any machine/ deep learning just statistical methods, mainly linear regression.

Our goal with this model is to predict the chance of admission based on factors of test scores and cumulative GPA. Of course there's more than just those two factors of admission, especially now with universities implementing "holistic" review processes. Nevertheless a lot of college admissions are still centered around the numbers. Our goal here is to also shed some light on whether test scores or GPA is the more important factor in admissions.

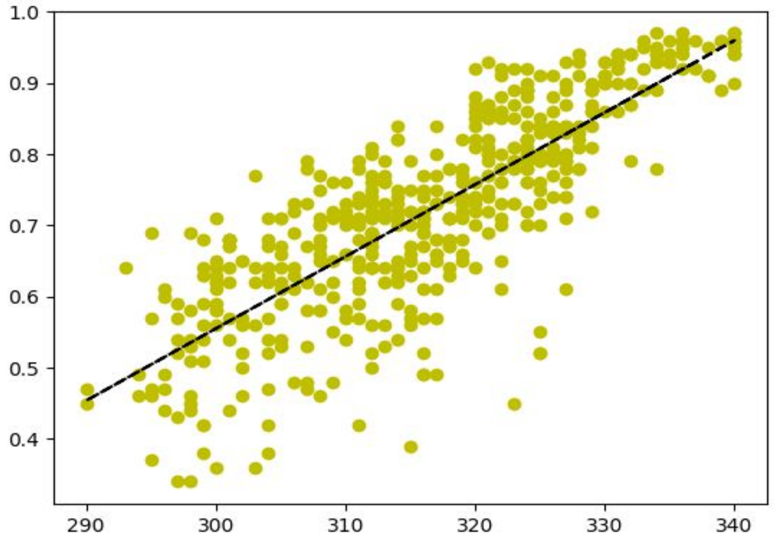
A good place to start is by looking at the data of Test Scores (GRE) vs Chance of Admit:





Immediately we can see there's a good positive trend which is what we would expect, as your Test Score increases the chances of you being admitted also increase. Also we can see that the data is approximately symmetric around a single peak, this makes sense as the population is known to be Normally distributed. It's important to note here that this is for all colleges, low and high ranked

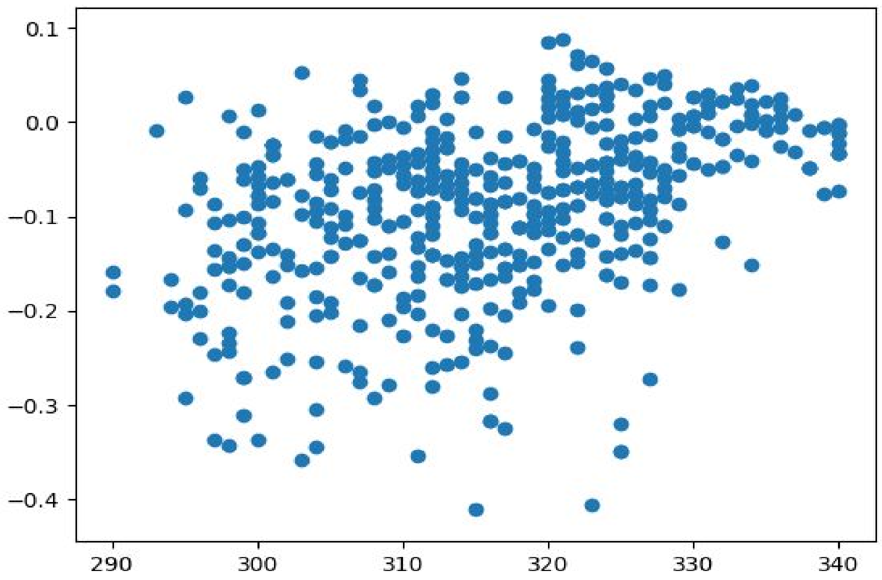
Next let’s examine an LSRL and correlation for the data.



r = 0.810

A correlation of 0.81 is moderately strong and is in line with what we’ve seen and would expect in this scenario. The coefficient of determination, r^2 = 0.66, which means our regression line accounts for roughly ⅔ of the variation in the data.

Another good thing to do is create a residuals plot of our data which will confirm if we should be using a linear model.

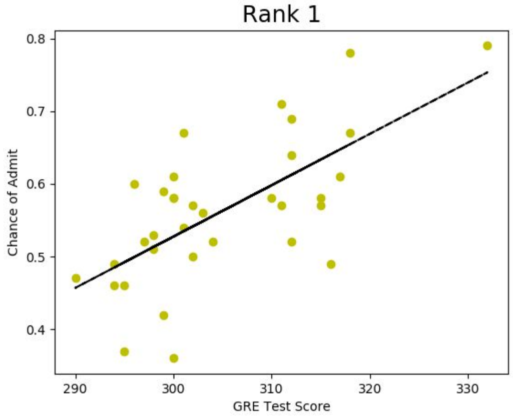


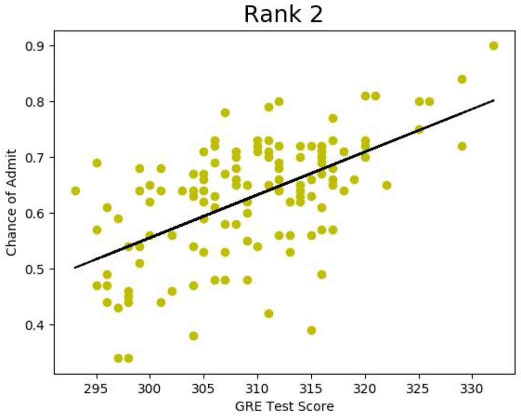
The residuals plot shows no obvious pattern, like a curve or something, therefore we know that a linear model is the most appropriate here.

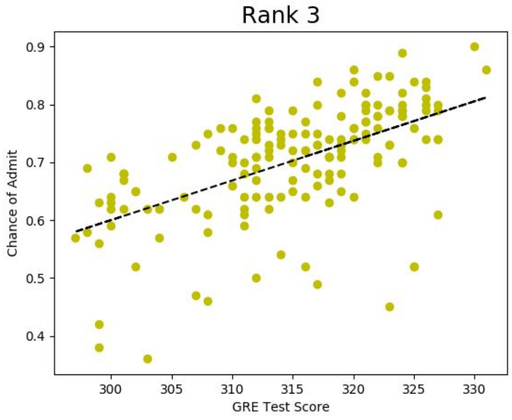
The next thing we should consider is the university rank. The notion is that when applying to a highly ranked competitive university the applicant pool will have a high average test score thus applying with such a score that would be considered high percentile at a lower ranked university, you will end up just being average at the top ranked university.

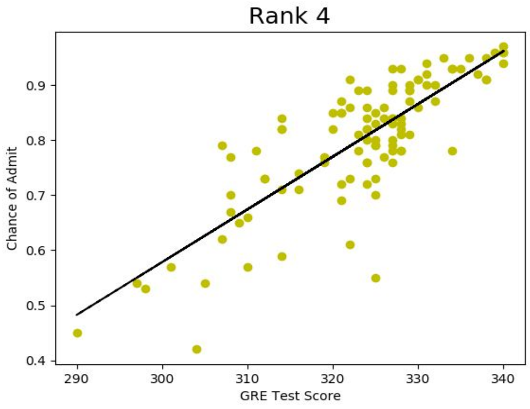
How do you exactly factor in the rank of a university?

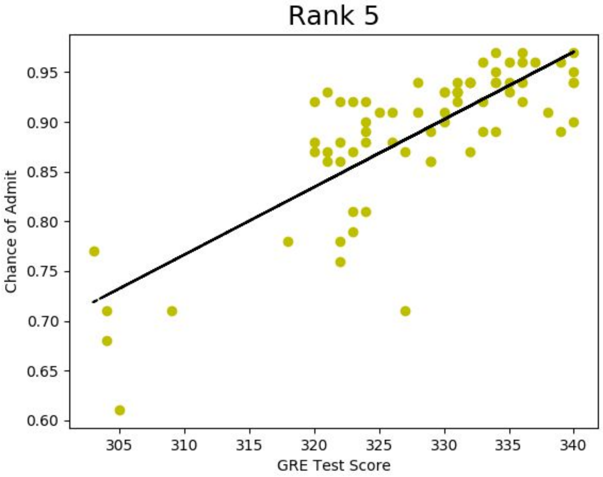
Let’s start by dividing our dataset into subsets of people applying to each rank (1-5). Then we will perform a regression analysis on each rank subset and in addition we will perform a 95% confidence interval of the test score mean. This will give you an idea on what kind of students are applying to each rank, also we want to avoid extrapolation as much as possible since the predictions can quickly become awry outside the sample interval.


Rank 1


Rank 2


Rank 3

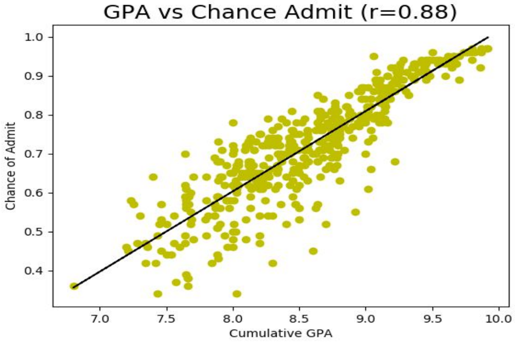

Rank 4



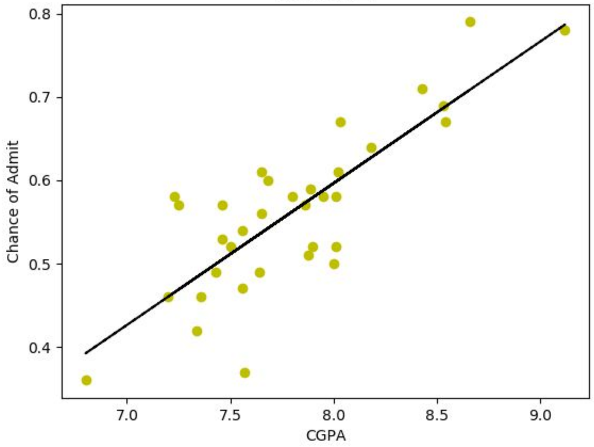
Rank	x	Sx	n	95% T int
1	304.911	9.38	34 (6.8%)	**(301.64, 308.18)**
2	309.135	8.124	126 (25.2%)	**(307.7, 310.57)**
3	315.03	8.062	162 (32.4%)	**(313.78, 316.28)**
4	323.3	9.95	105 (21%)	**(321.37, 325.23)**
5	327.89	8.66	73 (14.6%)	**(325.87, 329.91)**

So we can see that the mean score consistently increases with each rank level, as we would expect. Another thing to pay attention to is the 95% T Interval which also consistently increases nicely as we would also expect. I think this gives you a good idea of what kind of scores are in each rank. The test prep company, Kaplan, reports that the top 10% of test takers are 328 or higher, while 300 or lower is considered below average and doesn't give you good chances in admissions to most graduate schools. The Kaplan claims are in line with our confidence intervals almost spot on. Another thing that jumps out at me is the 'n' column; I've went ahead and also added the proportion of the dataset each subgroup makes up, and we can see that the data is fairly symmetrical with the most being rank 3 applicants.

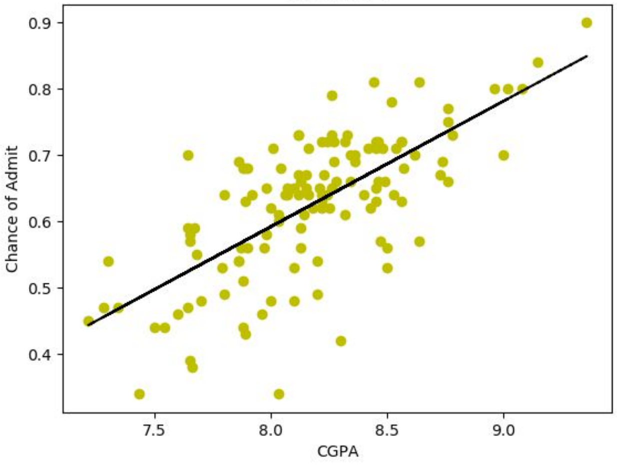
Now let's move on to Cumulative GPA.



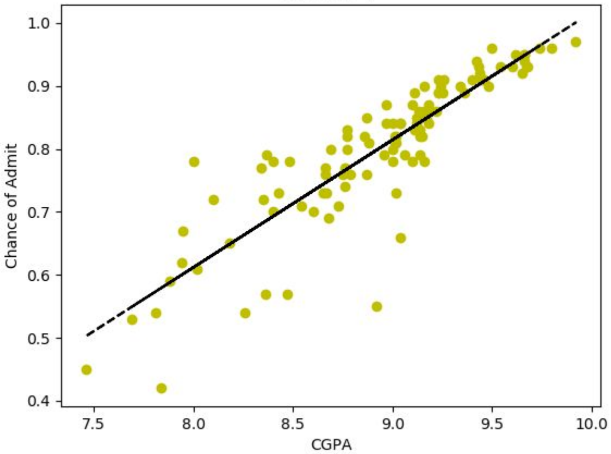
Rank 1



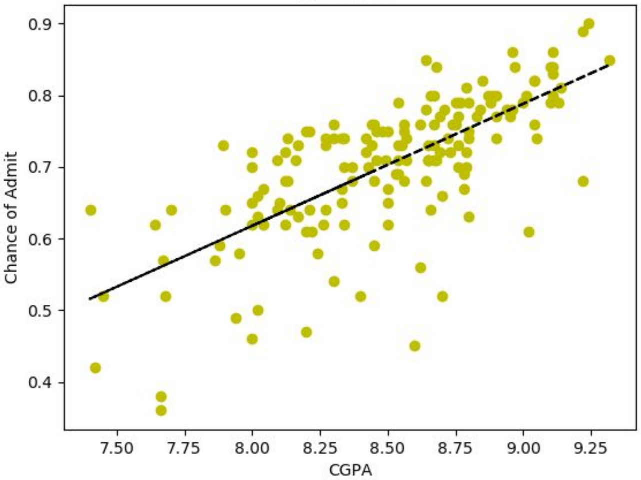
Rank 2



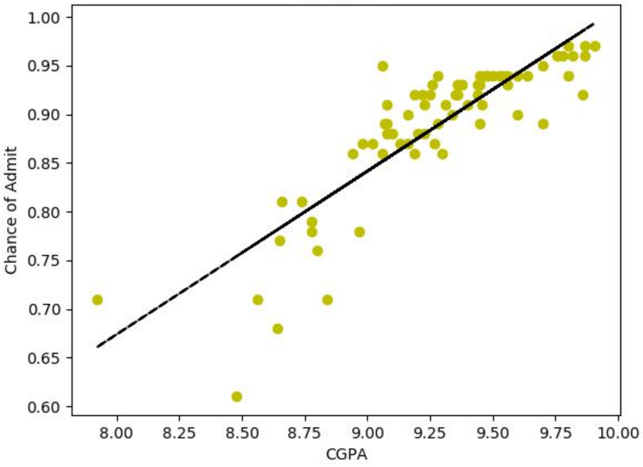
Rank 4



Rank 3

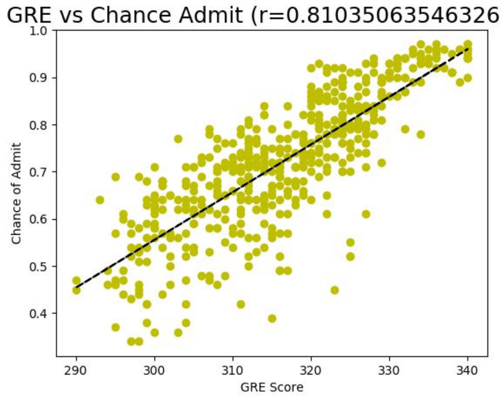


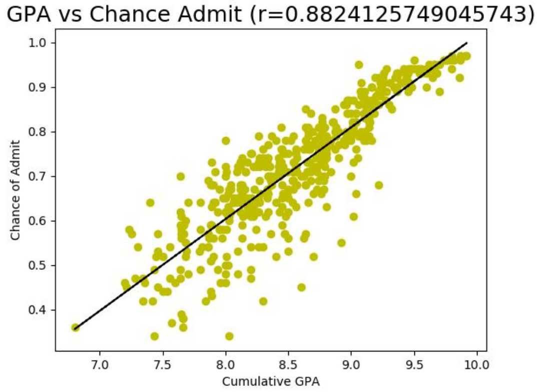
Rank 5



Rank	x	Sx	n	95% T int	r
1	7.799	0.426	34	**(7.633, 7.965)**	0.81
2	8.178	0.399	126	**(8.108, 8.249)**	0.70
3	8.500	0.406	162	**(8.437, 8.563)**	0.70
4	8.937	0.51	105	**(8.838, 9.036)**	0.88
5	9.279	0.381	73	**(9.189, 9.367)**	0.86

We can see that overall our linear regression model and the use of a 95% T Interval sheds some light on the total population of students applying and also gives a rough predictor on chance of admit.





Also we can see here that the overall shape of both distributions are very similar, however because the correlation coefficient for GPA is higher this means GPA is a better predictor of Chance Admit than GRE Score.

In conclusion, I would say that this purely statistical approach gives you a good idea of the population of indian applying graduate students as well as a good estimator on what to expect in the admissions process. Although this analysis doesn’t give you more precise estimates as more advanced machine learning models would, it’s a quick and simple way of describing and then making inferences of a population based on this sample of 500 college students. For practical purposes you could compare an arbitrary GRE and GPA to the confidence interval or use the regression line to get an approximate idea on your chances of admission.