Analysis of Post-Graduate Admissions

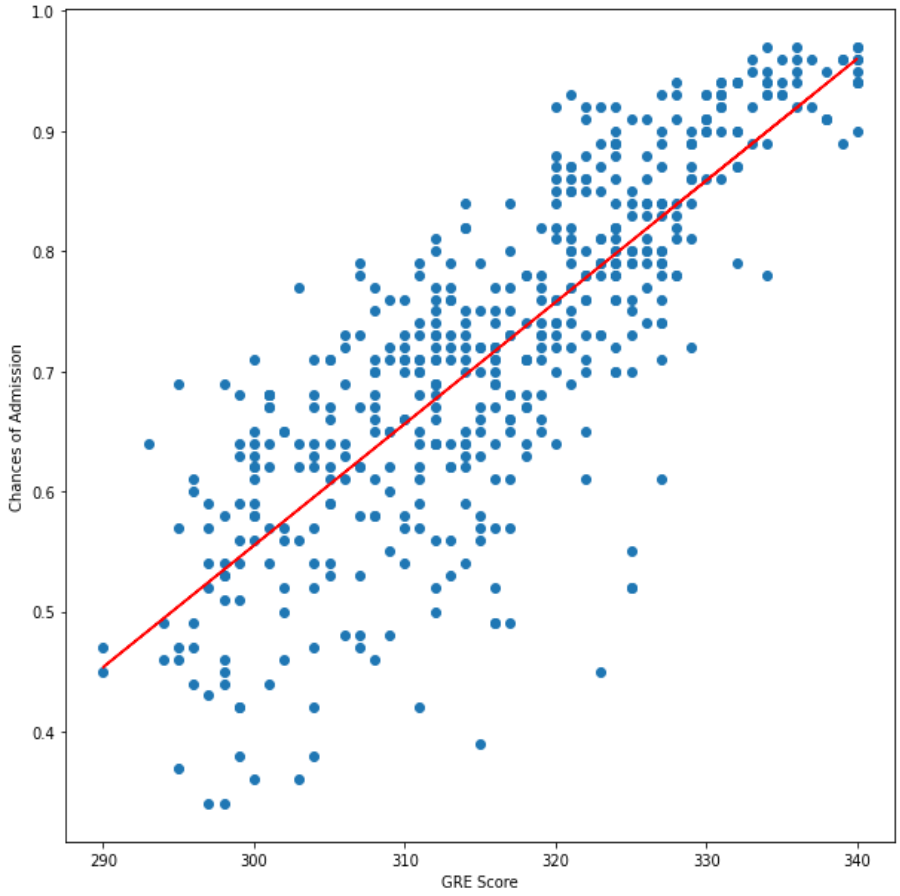
by Tushar Nandy

Introduction

The report takes into account the diversity of the students across various factors. I have taken a reasonable assumption that all students belong to a STEM background since research-experienced students have been compared to their non-experienced fellows in some parts. The report involves a few technical graphs and terms to provide a deep insight but I have made it a point to highlight and explain all such terms. At the same time, since these students are applying to foreign universities, I also expect them to have atleast heard a few terms. I hope you find this report insightful and it actually proves to be of some worth to applicants of Masters program.

Q1: Do GRE & TOEFL scores influence the chance of getting admitted?

1. **Let us first check the relation between GRE scores and chances of admission.**



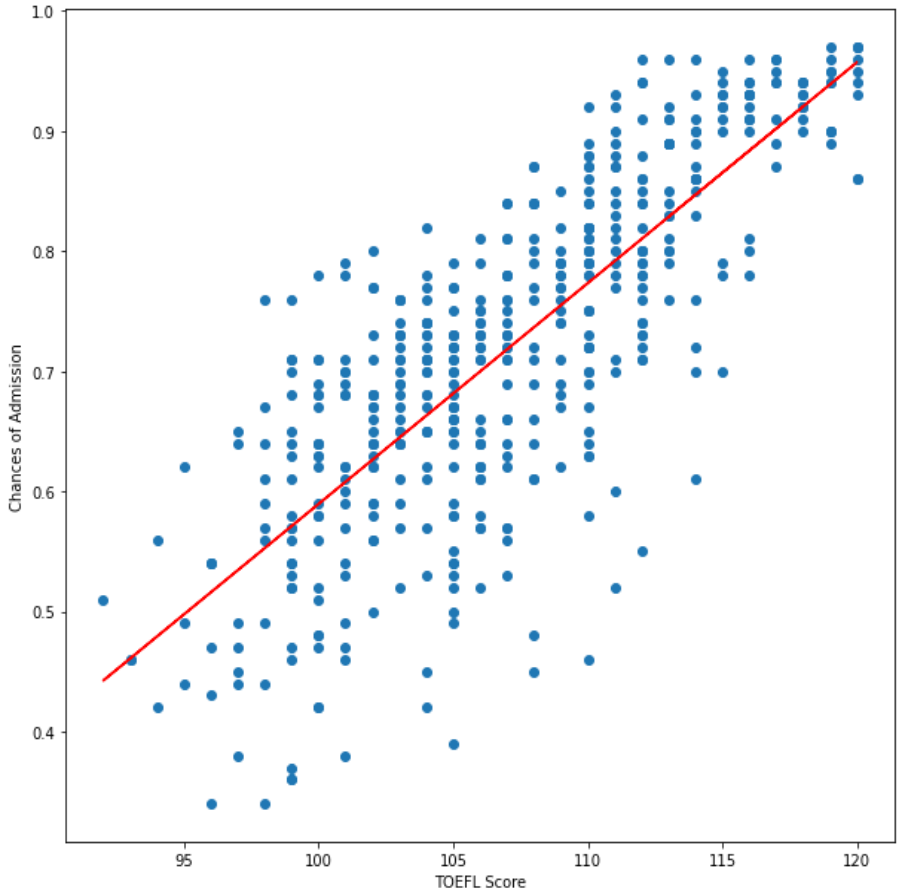
The metrics of the fit are:

- Slope: **0.0101**
- Intercept: **-2.483**
- Linear Correlation: **81.04 %**
- R-squared value: **65.67 %**
- P-value*: **1.088 X 10 $^{-117}$**

> A high p-value is a measure of poor staistical significance and a low p-value indicates high statistical significance between the two quantities. Linear correlation indicates the percentage of a linear format and R^2 tells us the percentage of predictable data for the model we use.

A high linear correlation suggests that the graph nicely obeys a linear format and a low p-value is what tells that GRE score is indeed a strong influencer of someone's chance of admission.

2. **Now lets check the relation between TOEFL Score and Admission Chances:**



The metrics of the fit are:

- Slope: **0.0184**
- Intercept: **-1.25**
- R-squared: **62.76 %**
- Linear Correlation: **79.22 %**
- P-value* : **6.73 X 10^{-109}**

Again, a low 'P-value' suggest that the correlation is quite significant and TOEFL Score also influences the chances of admission.

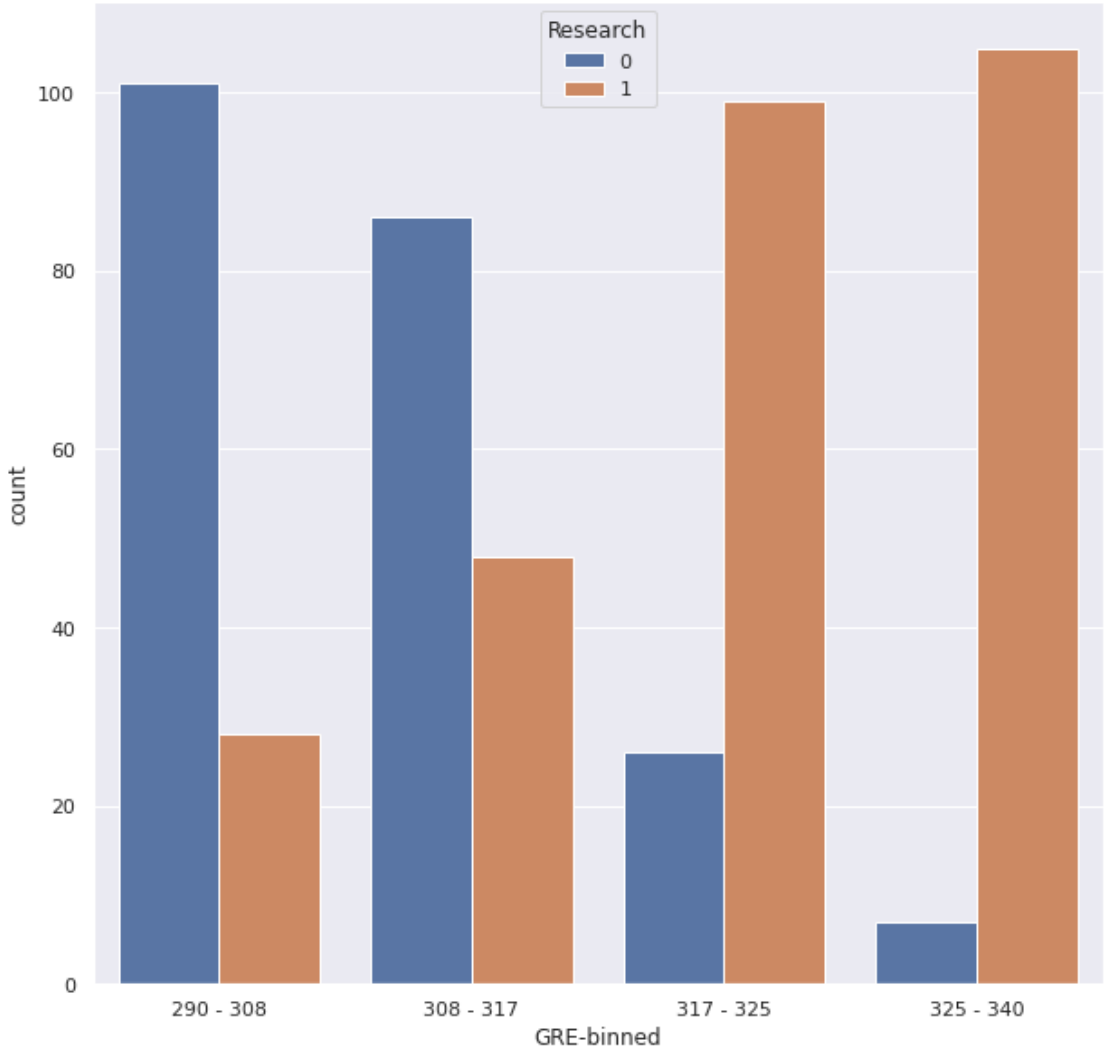
Hence, both these scores are potentially good features to determine a chance of admission

Q2: Does having Research experience make you score well in GRE & TOEFL?

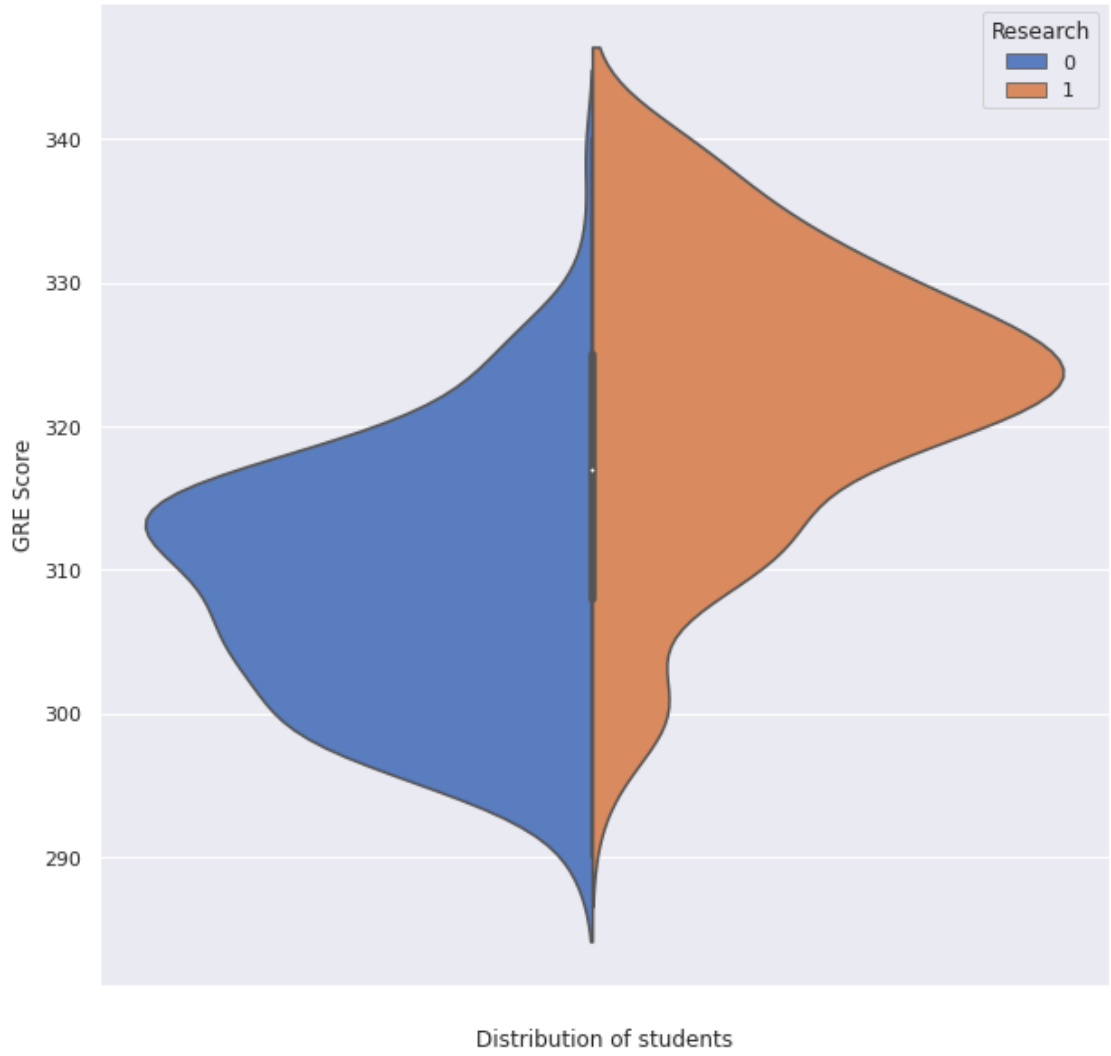
> **Violin Plots**: These are used to show the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared.

1. **Research Experience vs GRE Score**

Research Experience	Mean	Min	25th percentile	50th percentile	75th percentile	Max
No	309.3	290	303	309.5	315	339
Yes	322.1	293	317	323	329	340



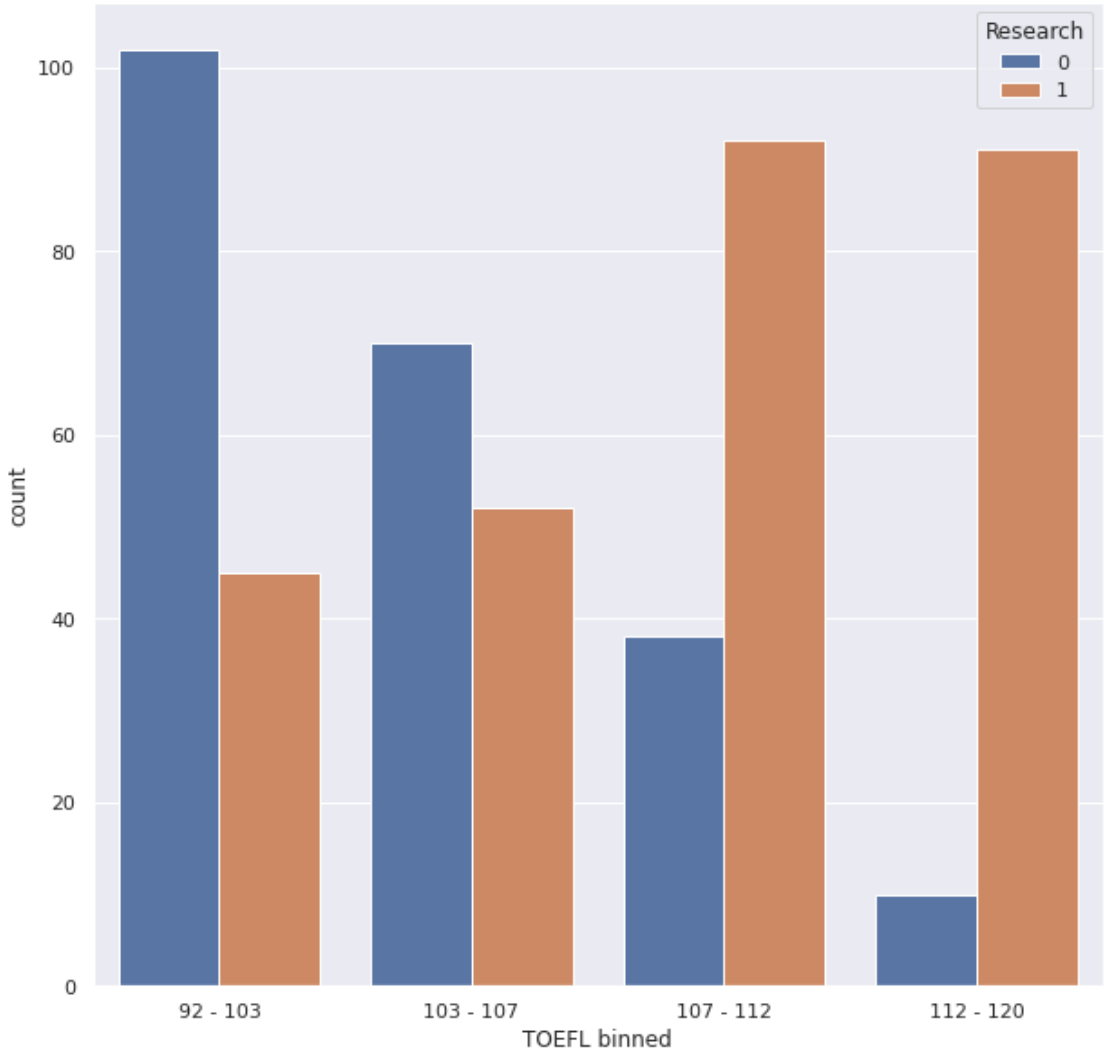
A bar graph showing the count of students in different intervals



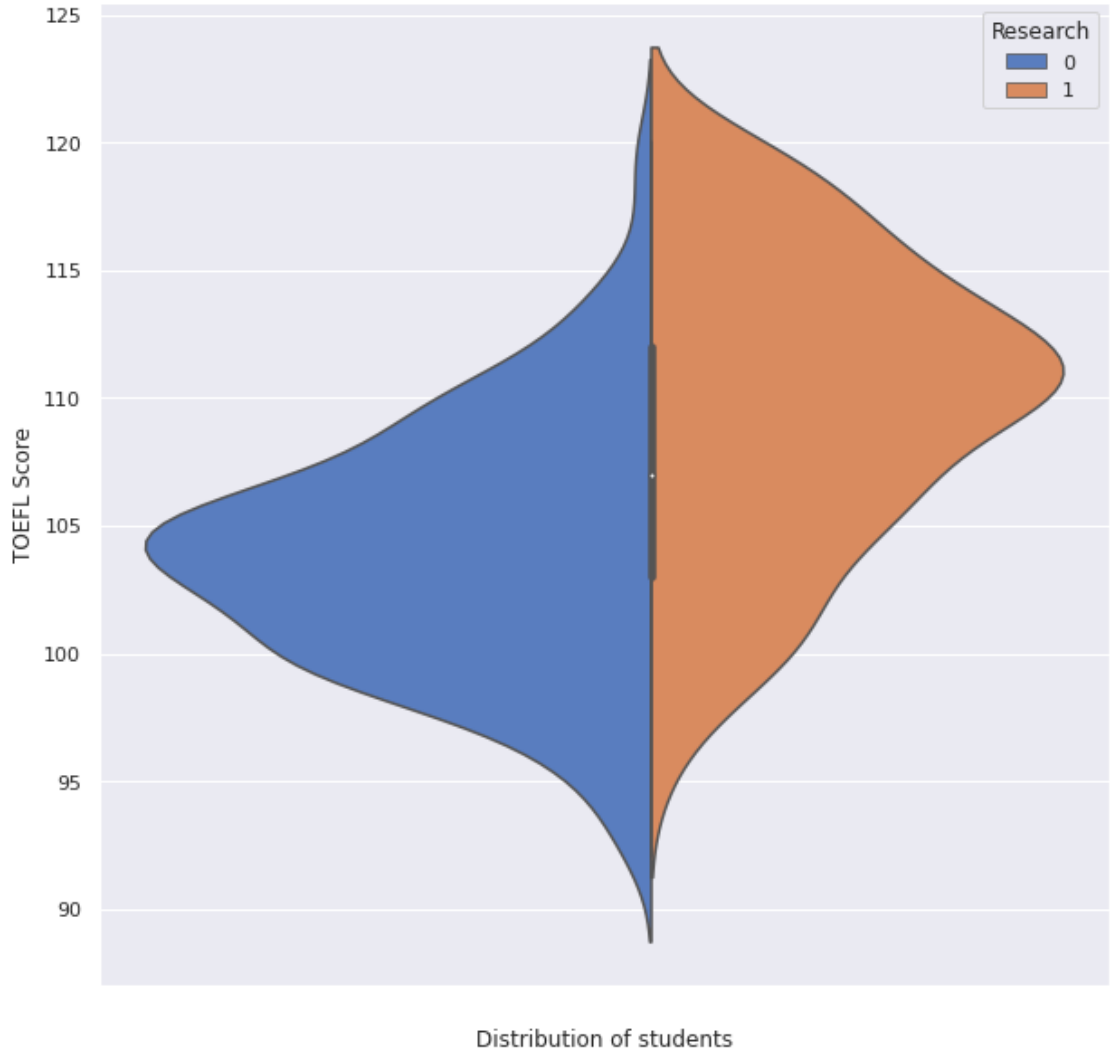
A violin plot showing the distribution of students across the scores

2. **Research Experience vs TOEFL Score**

Research Experience	Mean	Min	25th percentile	50th percentile	75th percentile	Max
No	103.99	92	100	104	107	120
Yes	109.71	95	106	110	114	120



A bar graph showing the count of students in different intervals

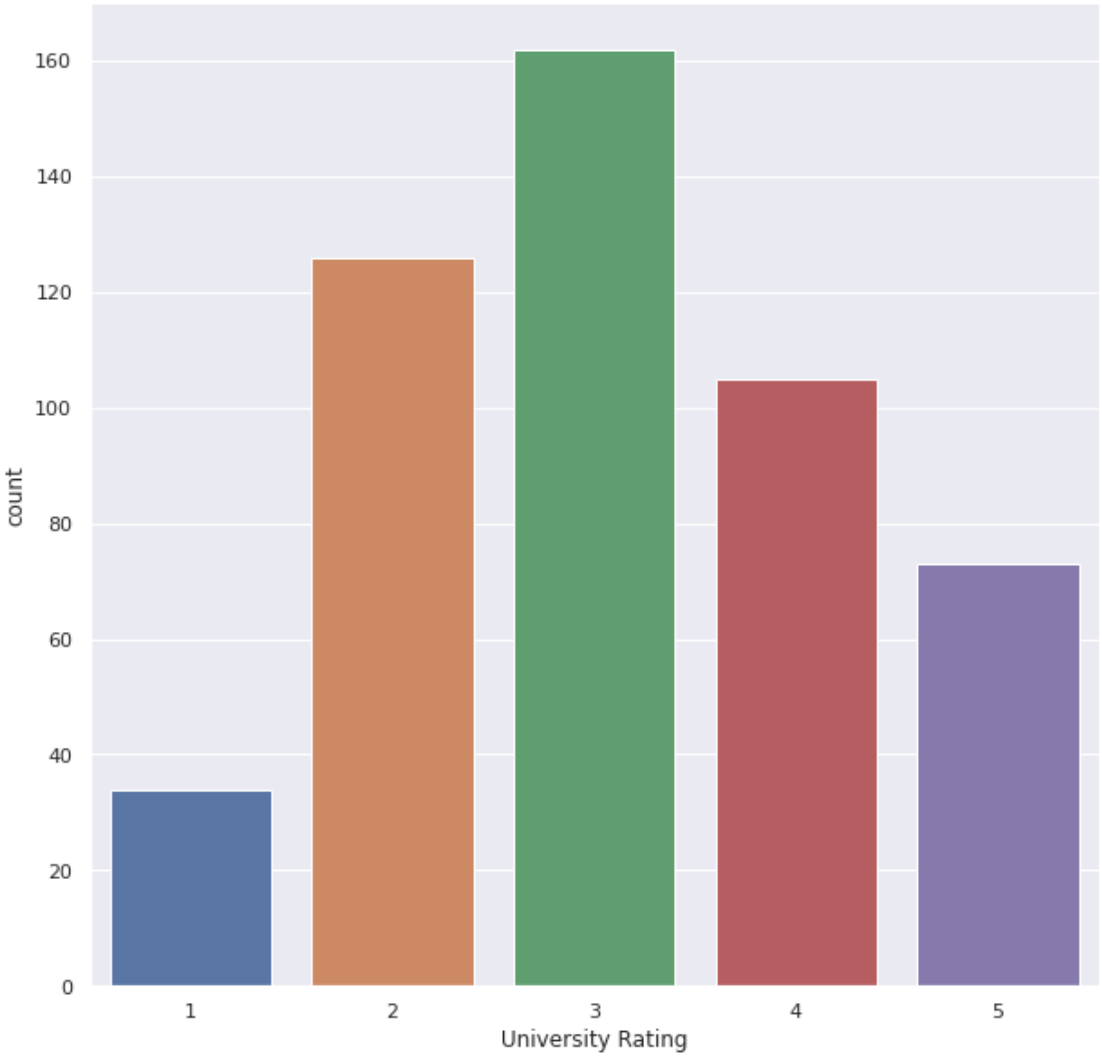


A violin plot showing the distribution of students across the scores

- The bar graphs convey that as the range of scores keeps getting higher, the proportion of research-experienced students increases while that of non-experienced students decreases.
- The violin plots also display a high density of research students in higher range of scores and in both graphs, there is a visible difference with respect to the position of peaks on both the sides.
- The numbers also convey that the scores of research experienced students is higher on each quartile (except the max score in TOEFL where it is equal).
- The mean score of research experienced students in GRE is **3.76%** higher and that in TOEFL is **4.76%** higher.
- So, whether it is the analytical skills gained via research experience, a systematic way of approaching the problems or perhaps just an aspiration for better reseach opportunities at foreign universities, research experienced students seem to have an edge over their non-experienced fellows.

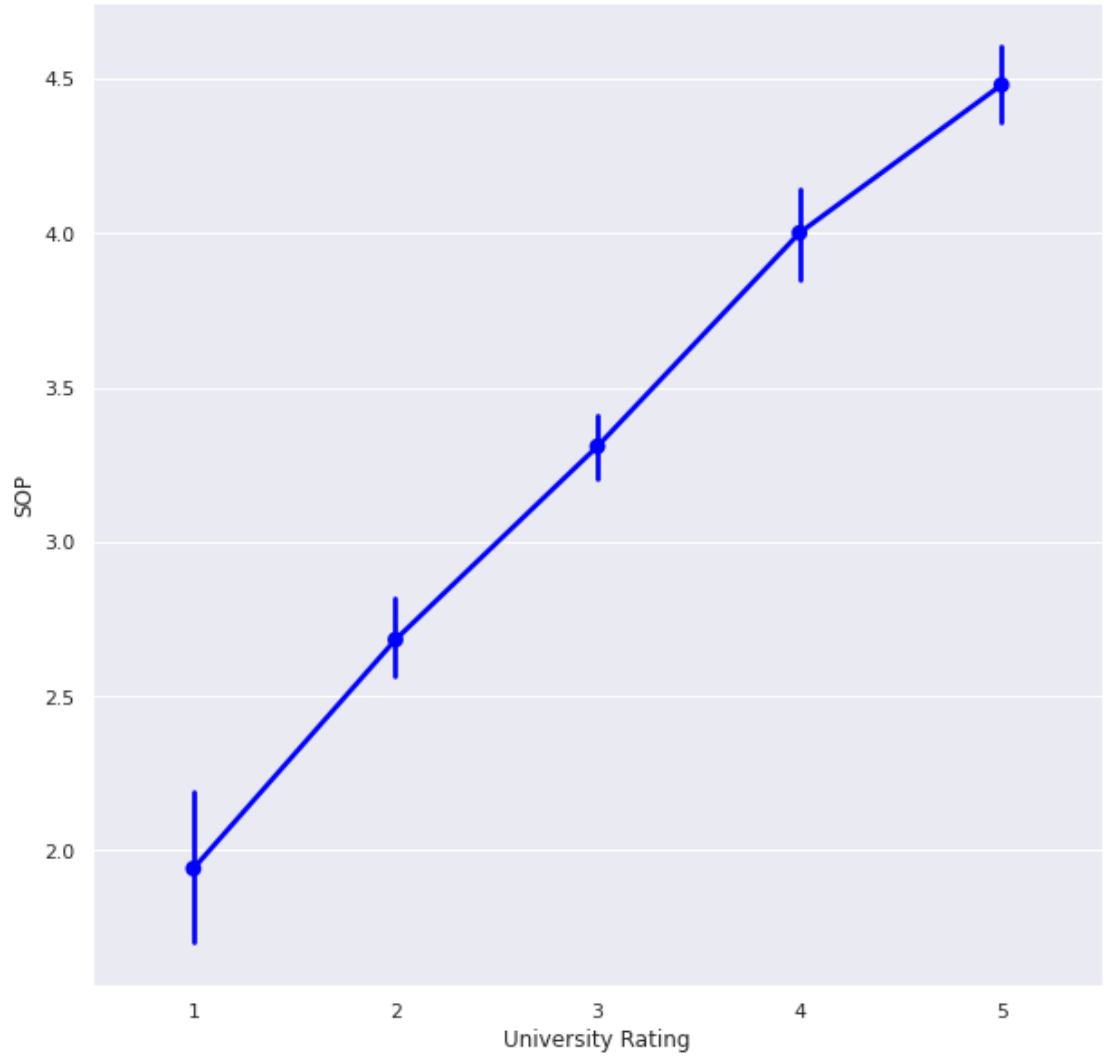
Q3: Does the University Rating influence my SOP and LOR Rating?

I would first like to present the distribution of applicants to universities of different ratings:



1. **Impact on SOP**:

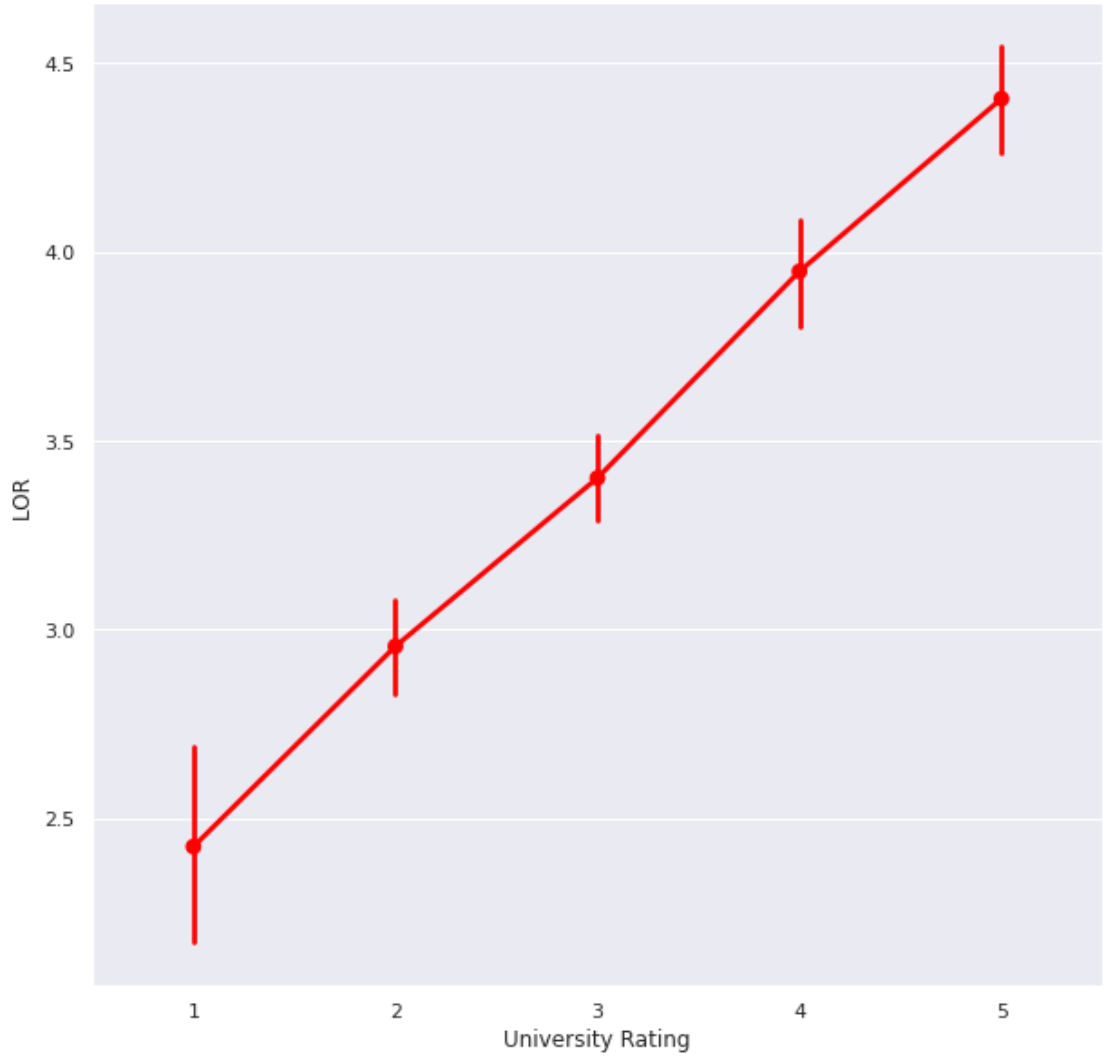
University Rating	count	Mean	Min	25th percentile	50th percentile	75th percentile	Max
1	34	1.94	1.0	1.5	2.0	2.38	3.5
2	126	2.68	1.0	2.13	2.5	3.0	4.5
3	162	3.30	2.0	3.0	3.5	3.5	5.0
4	105	4.00	1.5	3.5	4.0	4.5	5.0
5	73	4.48	3.0	4.0	4.5	5.0	5.0



The vertical lines denote range of variation and the points are estimates of central tendency

2. **Impact on LOR**:

University Rating	count	Mean	Min	25th percentile	50th percentile	75th percentile	Max
1	34	2.43	1.0	2.0	2.25	3.0	4.0
2	126	2.96	1.5	2.5	3.0	3.5	4.5
3	162	3.40	2.5	3.0	3.5	4.0	5.0
4	105	3.95	2.0	3.5	4.0	4.5	5.0
5	73	4.40	3.0	4.0	4.5	5.0	5.0



The vertical lines denote range of variation and the points are estimates of central tendency

- The points in these graphs are the average SOP/LOR scores for each category of university and the vertiacl lines, as mentioned, are the range of variation. Also note that these vertical lines do not show overlapping while covering their respective range of scores.
- Increasing trends combined with no overlap convey that the rating and quality of these documents increases as students approach higher ranked univeristies.
- Intersetingly, the tables for each part display that the maximum score are same for the universties with rankings 3, 4 & 5. The fact that the vertical bars are nowhere close to these scores indicates that most of the data points cluster in a close range near the center.
- We also notice a sharp decrease in the no. of applicants to universities rated from 3 to 5 while the scores increase, suggesting that only very few students with extremely good SOPs and LORs apply for top universities.

We can, therefore, say that the most probable ratings of SOPs and LORs are influenced by the rating of the university they are sent to.

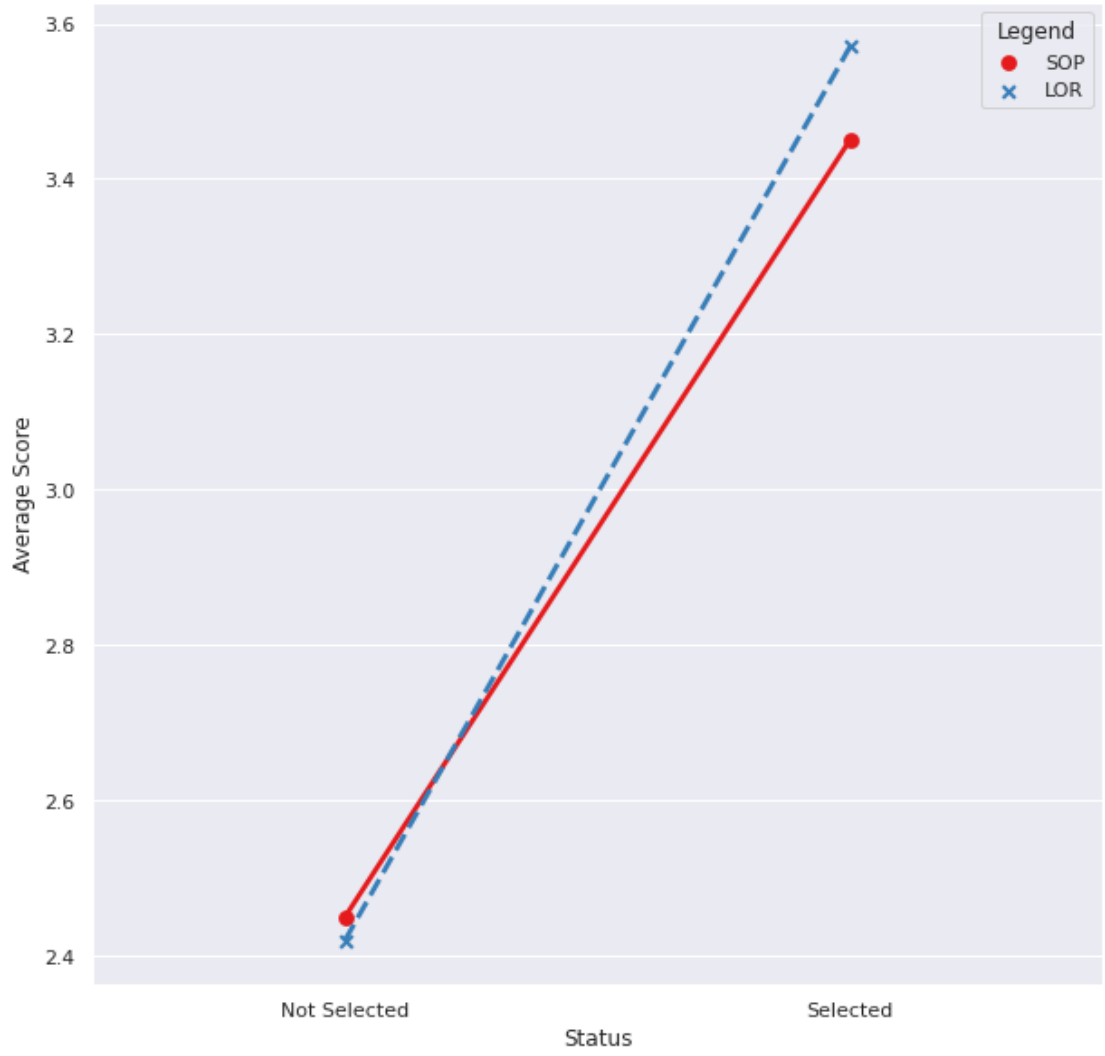
> **A technical detour**: This graph must not be confused for a linear regression plot. Since we are comparing discrete variables each having a small range of interval, the scatter plots form

> vertical lines over each university rating. The fact that the linear correlation of both the graphs are **53%** and **37%** respectively, a linear regression would not have been a good model for representation.

Q4: What is the average SOP & LOR Rating of students who got admitted?

- The selection chance is considered to be **0.50**. Anybody with a chance equal to and above this is considered to be selected. The following are the average SOP and LOR ratings:

Status	Count	Average of SOP	Average of LOR
Not Selected	39	2.45	2.42
Selected	461	3.45	3.57

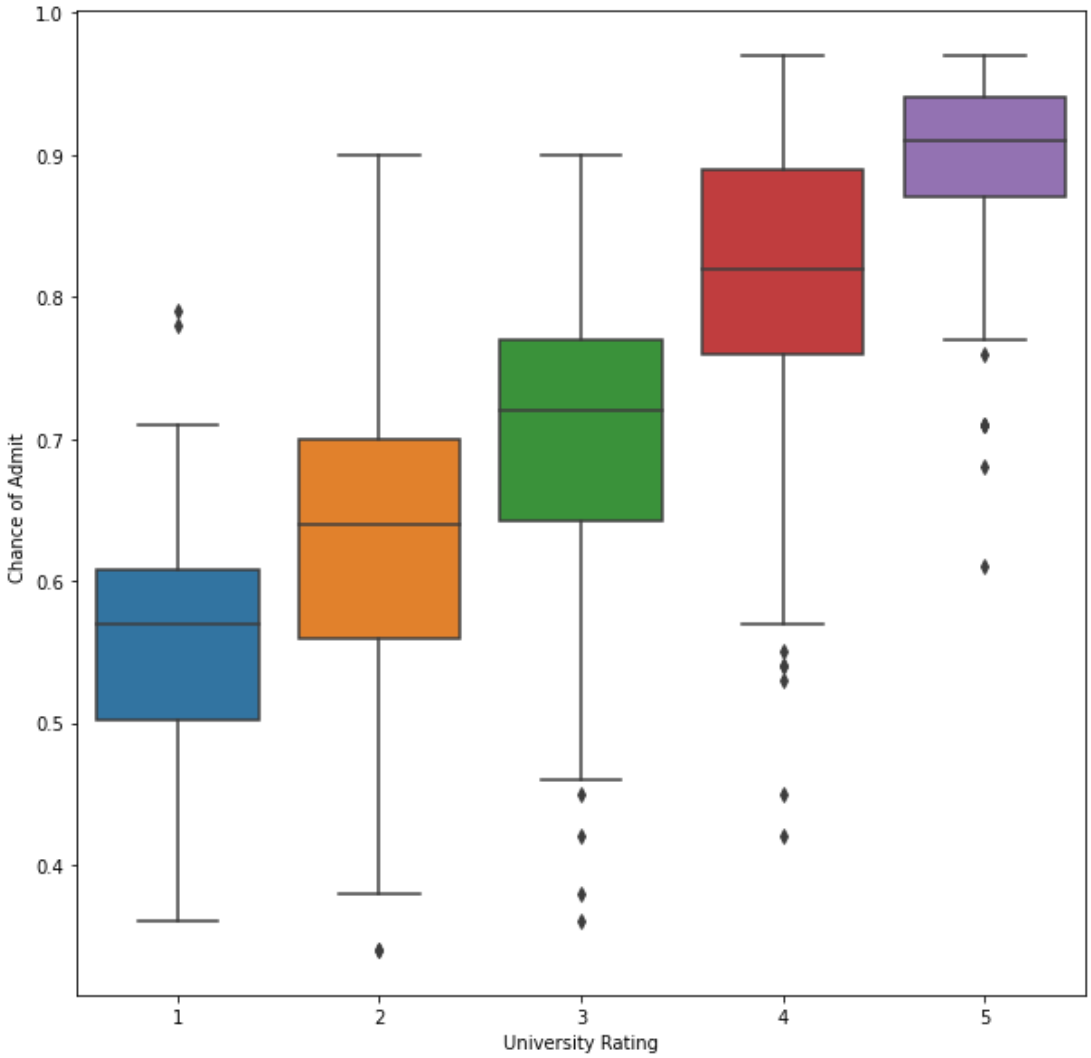


Q5: How does the University Rating improve the chance of getting admitted?

> **Box and whisker plots/Box plots**: A boxplot is a standardized way of displaying the distribution of data based on a five number summary ('minimum', first quartile (Q1), median,

> third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if your data is skewed.

- First, let’s look at the box plot displaying the distribution of students:



The plot suggests that the number of students getting rejected by the lowest ranked universitites is close to **25%** of their total applicants. The number of rejections keep decreasing, with finally all applicants being selected to top universities.

The question is tricky as it could be interpreted as: ***"If I apply to a better ranked university, do I have a better chance of admission ?"*** This is surely not the case because a university rating can by no means influence a student’s chance of admission as it depends on his/her own achievements and performance.
Infact it is the other way round. Students with only very good chance of admission are confident enough to apply for high ranked universities and also end up getting selected because most of them are bright.
Whereas, students with lesser chances of admission apply to low ranked universities but all do not get selected. Probably because these universities want to improve their ranking and don’t select academically very poor students.

Allow me to present a different opinion. It could also be possible that universities with different rankings come up with their own cutoffs with respect to different scores. This might also end up in a situation where students with chances around **0.8** get rejected from highly ranked universities but students with chances around **0.4** or even **0.35** make it to low ranked universities. Since our data set comprisies of only **500** students applying over **5**

categories of universities, that too with unequal proportions, I do not place much confidence into this theory. We would require more information to analyze this.

Q6: What should be your Scores for 90 % Chance of Admission?

- Let’s look at the various combination of scores achieved by students with **90%** admission chances:

GRE	TOEFL	SOP	LOR	CGPA
330	115	4.5	3.0	9.34
340	114	4.0	4.0	9.60
332	119	5.0	4.5	9.24
332	118	4.5	3.5	9.36
329	119	4.5	4.5	9.16
324	111	4.5	4.0	9.16
331	119	5.0	4.5	9.34
330	114	4.5	4.5	9.24
327	116	4.0	4.5	9.48

- The average for each category is:

GRE	TOEFL	SOP	LOR	CGPA
330.56	116.11	4.5	4.11	9.32

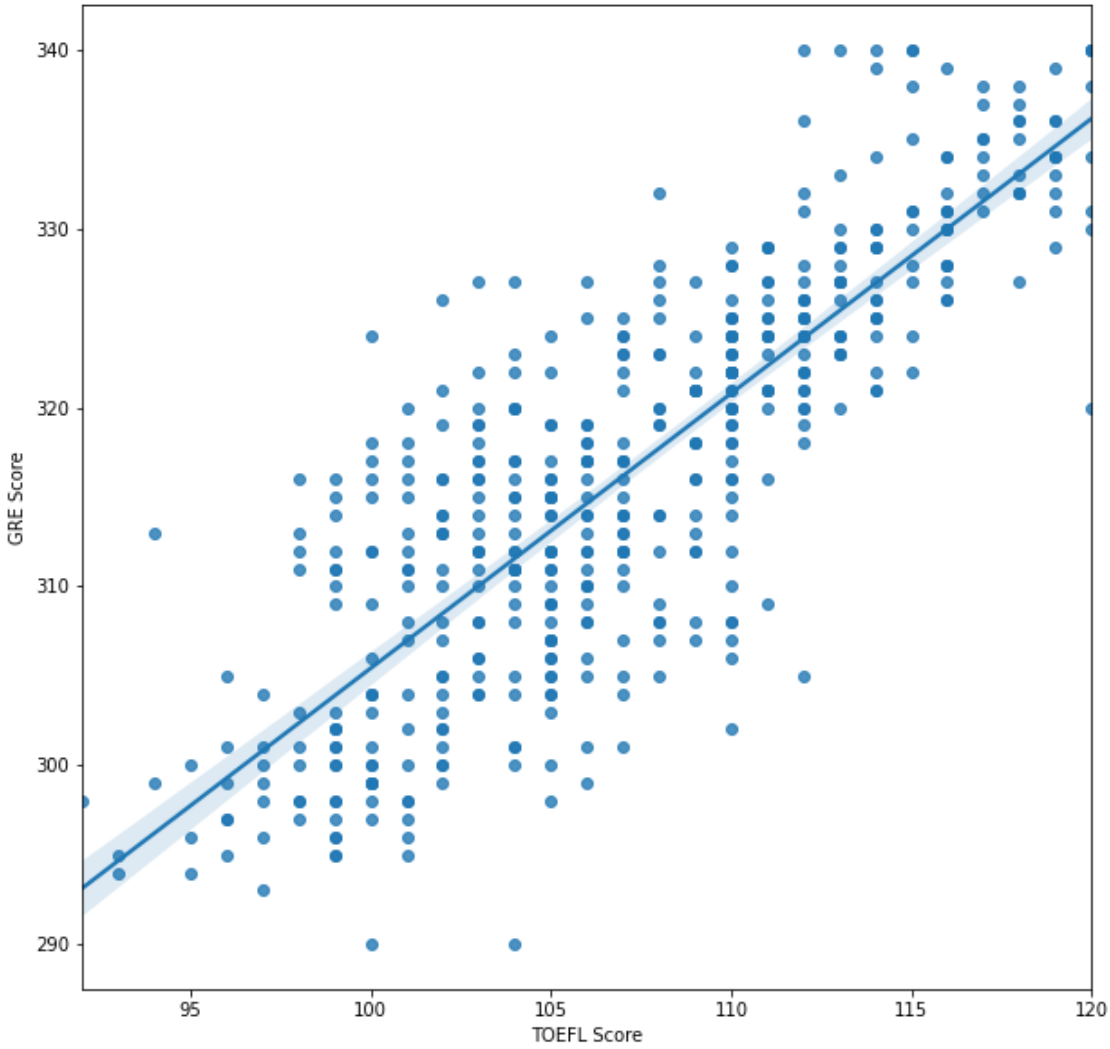
- Based on these the safest combination of scores, which will definitley ensure a **90%** entrance is:

GRE	TOEFL	SOP	LOR	CGPA
334	117	4.5	4.5	9.4

> This slightly differs from the averge because an average is computed independently for each category whereas the scores in this combination are extracted with respect to one another, to ensure a 90% chance.

Q7: Can we predict the GRE Score of a student, given the TOEFL Score ?

A relation between these two test scores might seem very significant to some while others might debate that they are quite independent. However, the statistics tell us that there is a good linearity and statistical interdependence. It might indeed be possible to predict one score from another.



- Slope: **1.536**
- Intercept: **151.80**
- Linear Correlation: **82.72%**
- R^2 : **68.42%**
- P Value: **9.29×10^{-127}**
- RMS Error: **6.34**

Lets look at these four predicted GRE scores:

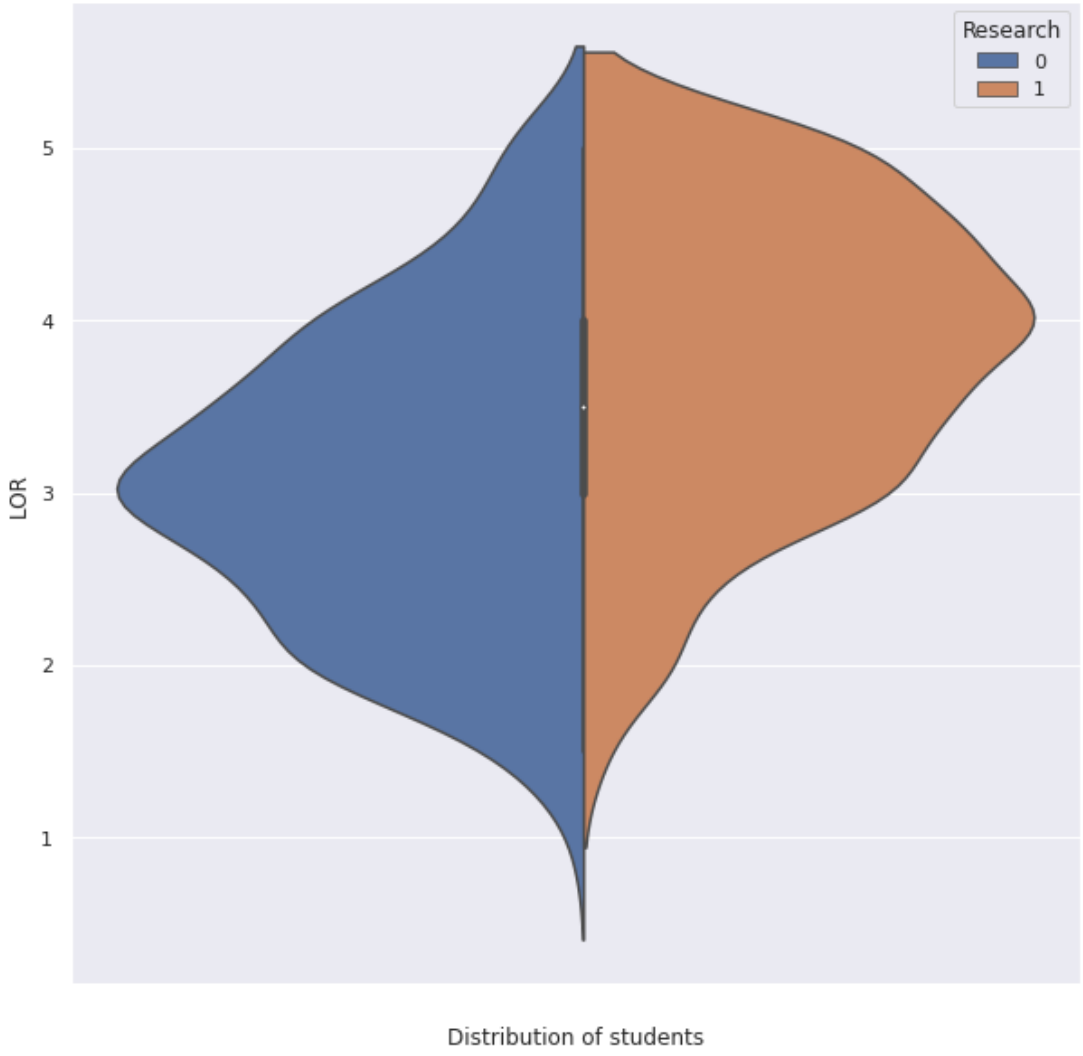
> Please note that there is a lot of clustering on certain TOEFL scores. In that case, the actual scores are an average* of all GRE Scores, rounded off to the nearest integer. The predicted scores too are rounded off to the nearest integer. This is due to the fact that these scores are integers only.

TOEFL scores	Actual GRE Score	Predicted GRE Score	Absolute Error	Error %
92	298	293	5	1.47%
105	311*	313	2	0.59%
113	327*	325	2	0.59%
120	335	336	1	0.29%

Q8: Does Research experience influence LOR rating ?

As we discussed earlier, the LOR is in the hands of someone under whom a student has worked and is experienced enough to pass a good judgement about the student.

Let us now look whether having a research experience influences LORs or not with the help of a violin plot:



Distribution of students

	Research experienced	Not experienced
Below 3.0	32 students	76 students
Above 4.0	92 students	21 students
Most probable score	4.0	3.0

- **Students with no experience**: The plot for these students bulges around **3.0** and tapers near both the ends. Visually, this suggests that even though most of the students have an LOR rated around **3.0** while there are lesser number of students distributed in the good and poor range of LOR rating.
- **Students with research experience**: Non uniform tapering and a bulge near **4.0** suggests that most of the LORs belong to higher ratings and very few are below **3.0**. The number of students below **3.0** is lesser than half and students above **4.0** is more than **4**-folds compared to that of the non-experienced category.

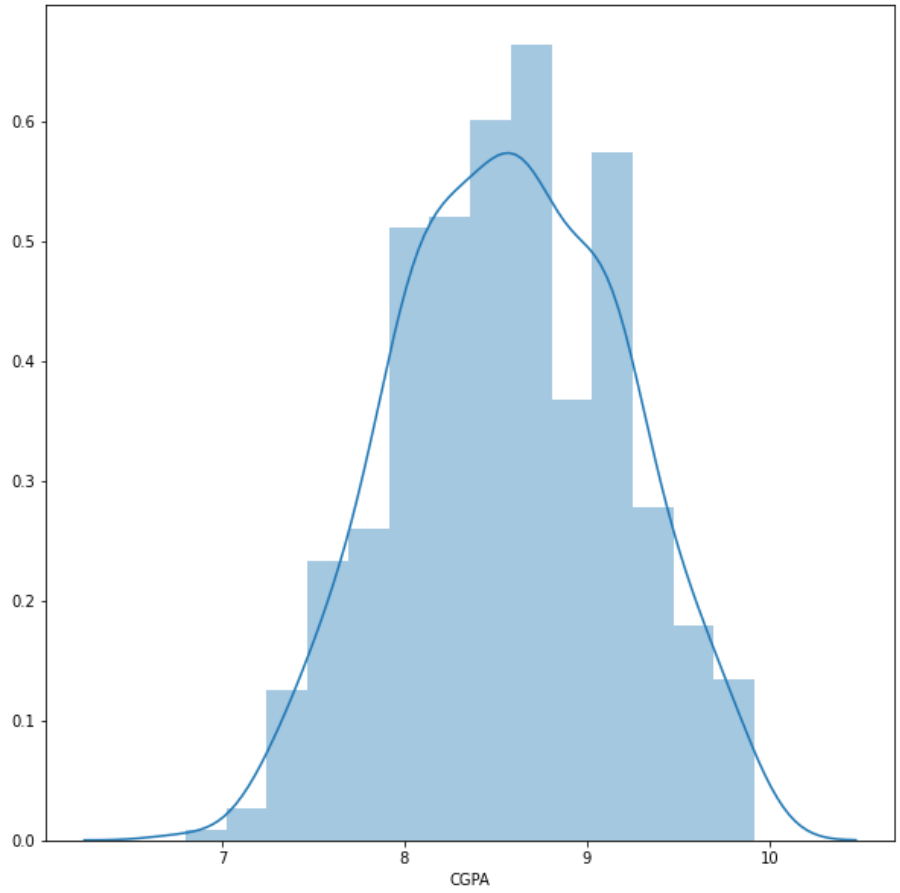
But again, not to forget that the violin plots are an extension of the box plots and the overlapping must not be overseen. We can, on the basis of the numbers stated, say the least that chances of making it above **4.0** for each

category are:

- **Research experienced**: $\frac{92}{280} \times 100 = 32.86\%$
- **Non-experienced**: $\frac{21}{220} \times 100 = 9.55\%$

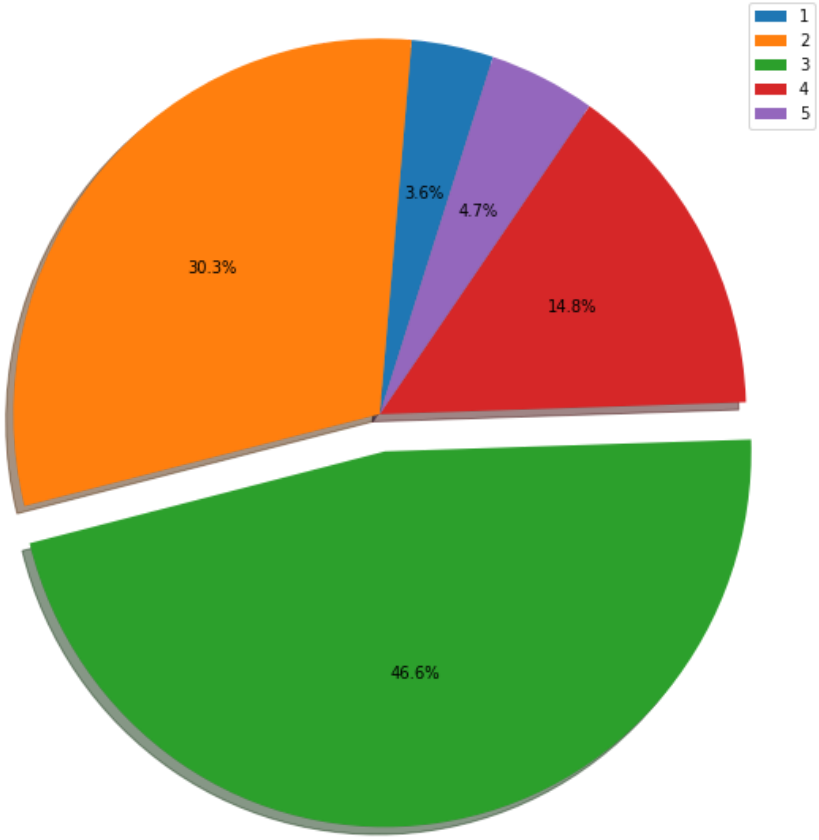
Q9: What is the distribution over different types of universities, for students with CGPA between 8.0 and 9.0 (inclusive) ?

- First of all, let us see why should we consider this range significant:



This probability distribution plot clearly suggest that there are maximum students (277, to be precise) within this range and we could have more students in future who land up here. We must also keep in mind that this also a good range of performance and we can expect great competition amongst these students while applying for universities.

So, the distribution of students to various universities with different ratings is:

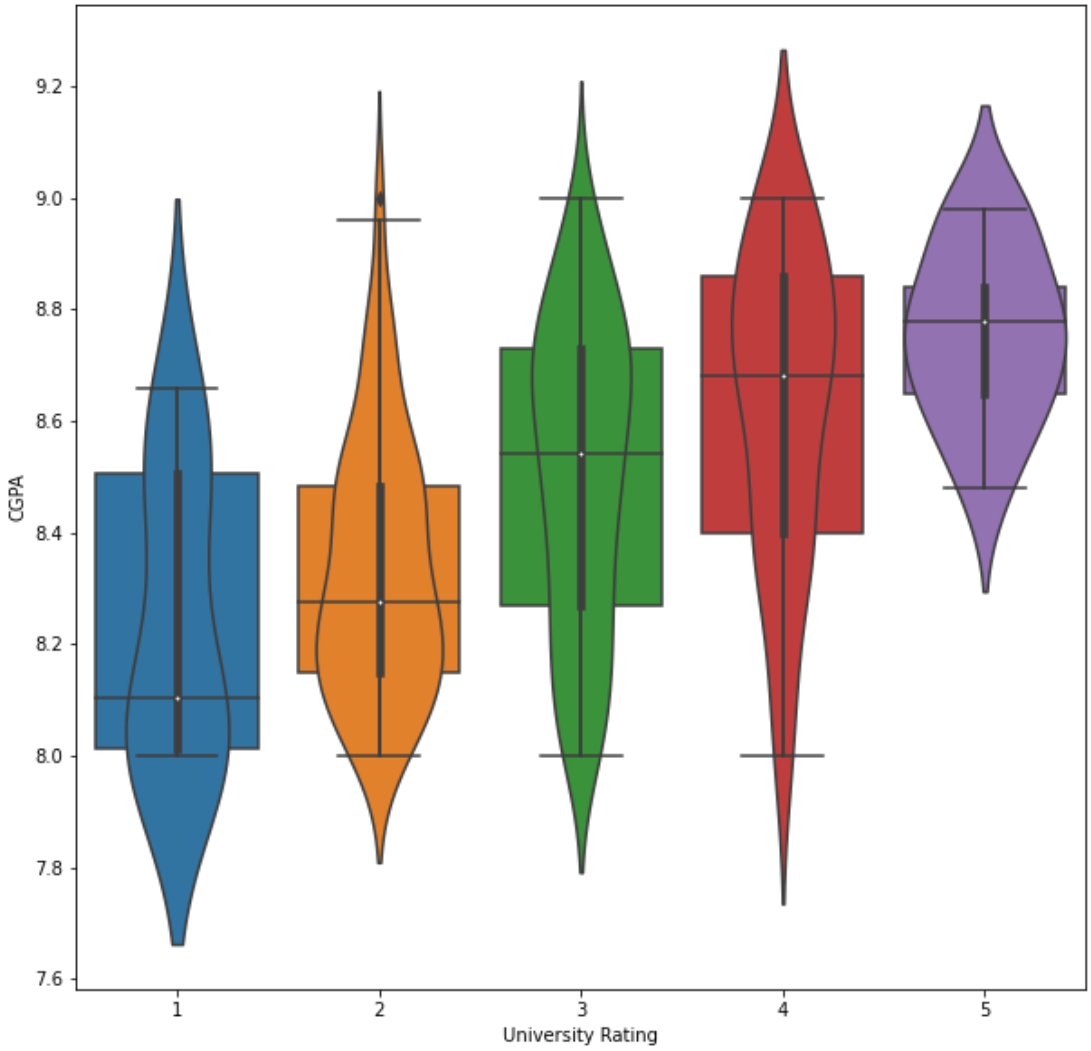


We can clearly see that most of these students get admitted to 3-rated universities, followed by universities with rating 2 and quite surprisingly, very few make it to the top colleges and some even have to take admissions in poorly rated universities.

Q10: What sets students apart in the above distribution?

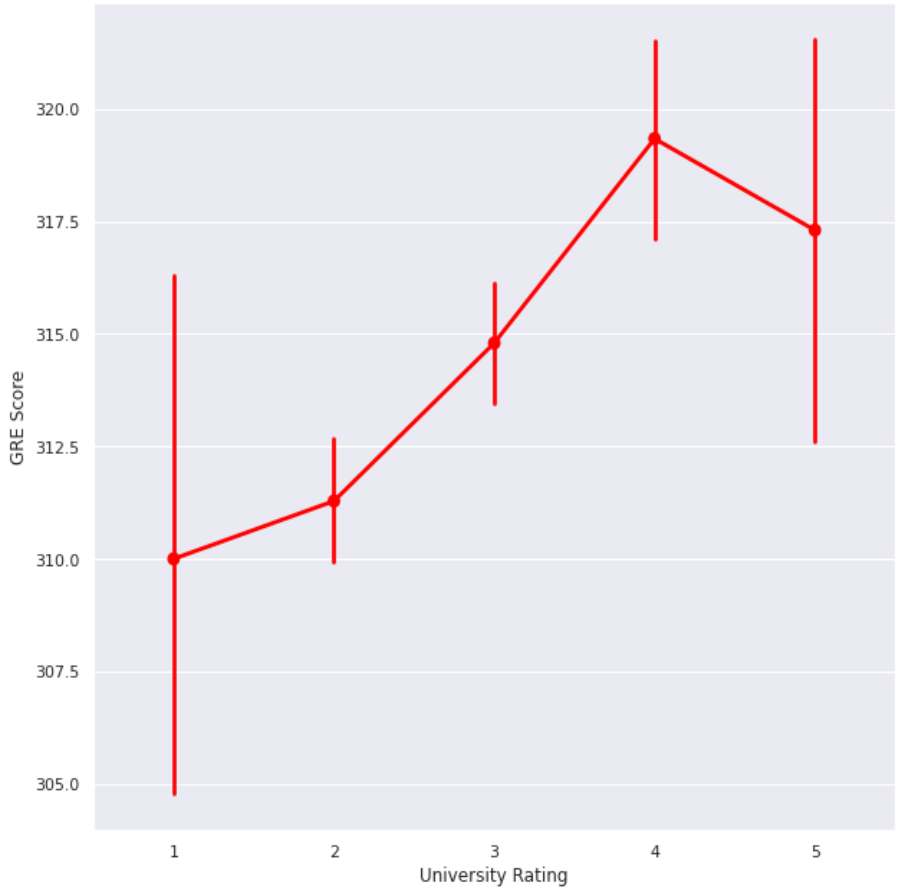
- As you might expect, this is a necessary follow up question to the previous one. The answer to this has been divided into small sub-sections, each shedding light on various factors.

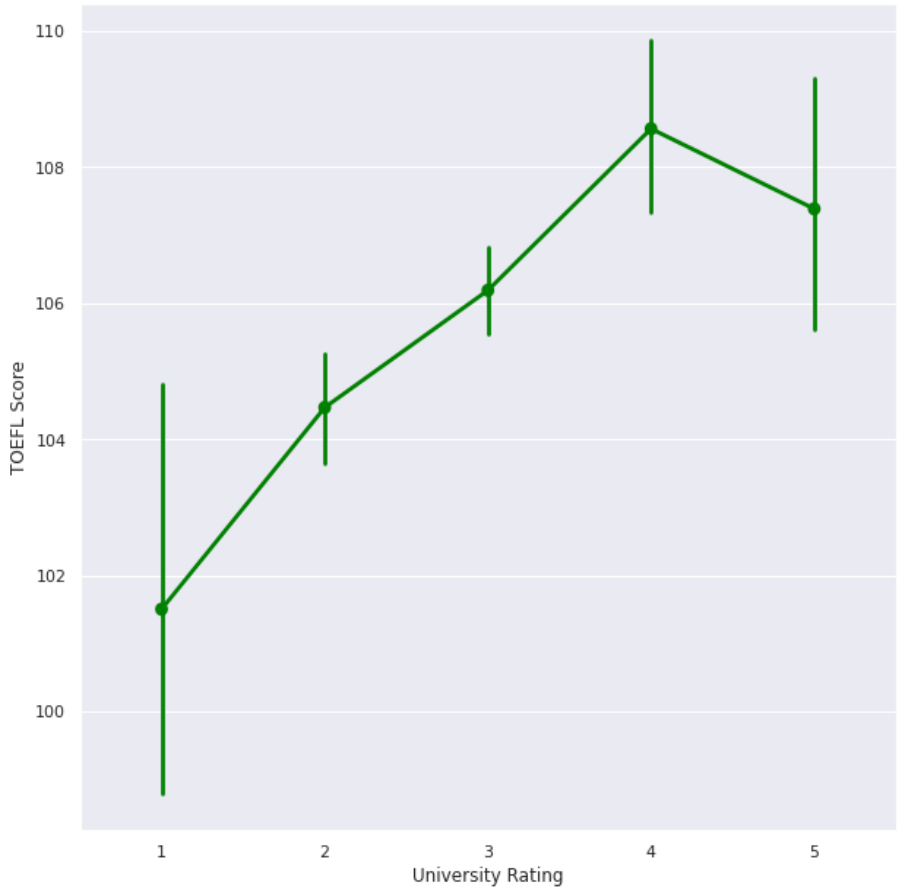
1. **Granularity in CGPA**:
    - Since this small range of CGPA contains **277** students, it is important to again check for any granularity in CGPA vs University distribution.



- For the first 2 categories, we see that the maximum kernel density (marked by the bulges) remain biased towards lower CGPAs for the first two and towards higher side for universities with rating 3 and 4.
- Specifically for universities with rating 1, the graph ends much earlier, indicating that students with CGPAs on the higher side do not target them and the best these universities got is around 8.7.
- However, there is also a considerable overlap across a wide range for universties with rating 2, 3 and 4. The thickness of the bulges is also quite low. Indicating that not much can be infered for these.
- Coming to the top universities, the plot is quite compact and the kernel density is more as compared to the previous ones. This indicates that one may not get an admission with a CGPA lower than 8.5 .

2. **GRE and TOEFL Combined**:

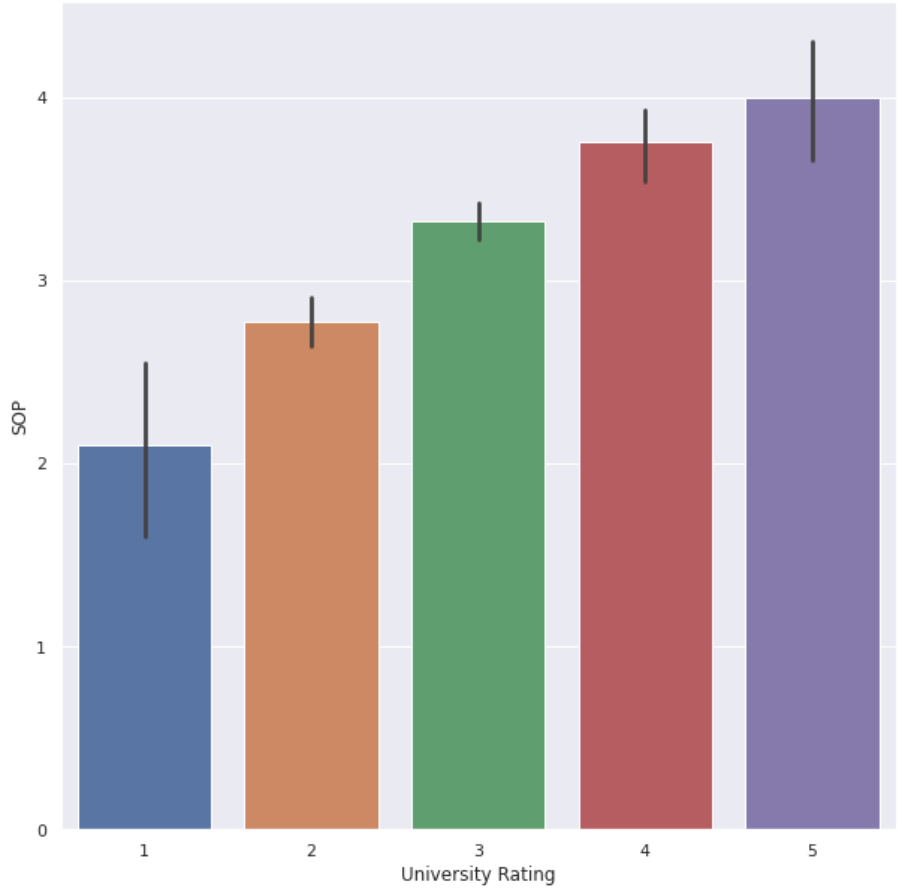


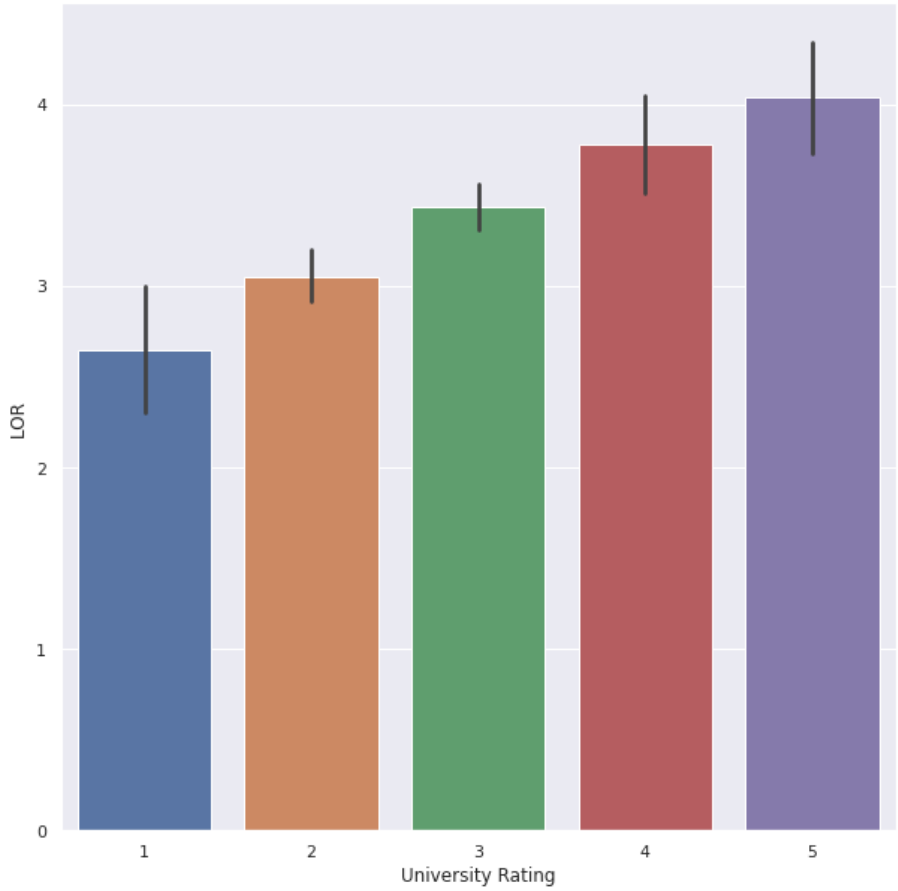


-The vertical lines denote range of variation and the points are estimates of central tendency-

- Average scores increase with rating of universities and drops slightly from 4 to 5 but the average for 5 is still higher than those for 1, 2 and 3.
- We can infer that good scores indeed determine your fate especially when you compare universities with rating 1 and 3.
- Premier universities require you to have scores above a certain level, and a good level infact, but once you're beyond that, there might be other factors at play to get you admission.

3. **SOPs and LORs**:

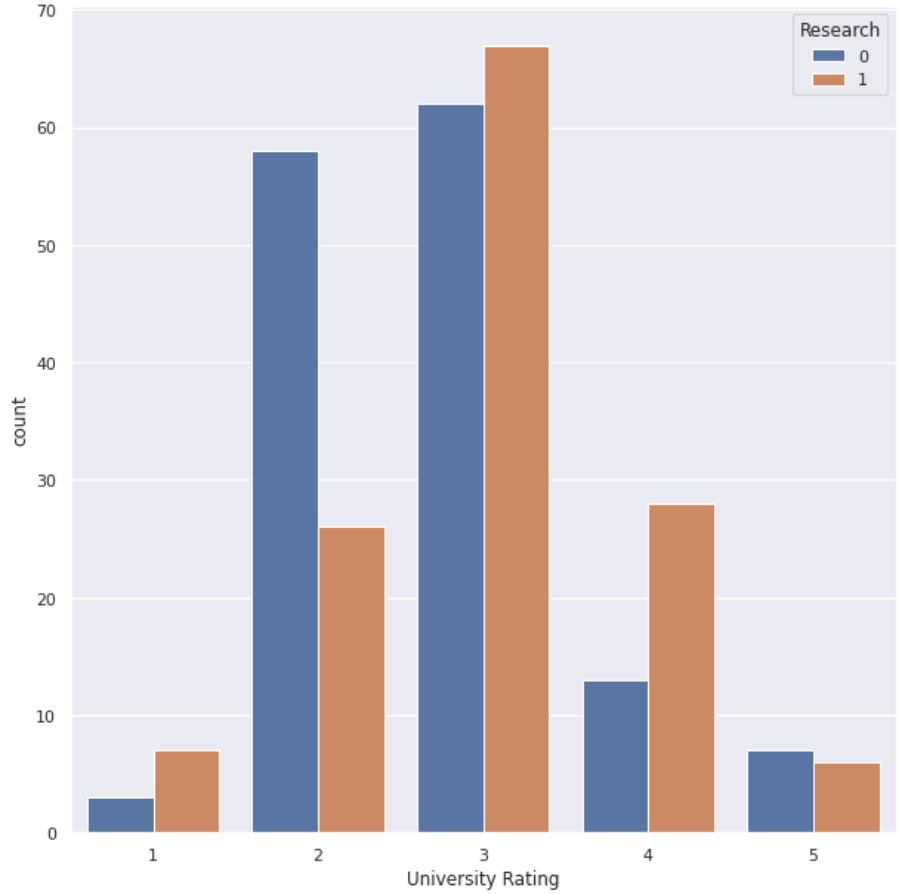




- I believe that we have, more or less, found those "*factors at play*". A constant rise in the LOR and SOP qualtites vs University Rating indicate that for students in the range we're analyzing (CGPA b/w 8.0 and 9.0), SOPs and LORs also make a good distinction, especially when we compare universities with rating 4 and 5.

4. **Research matters (or maybe not)** :

> An aside: There two things to be kept in mind before analyzing this graph. One, we are just representing students within a particular CGPA range so this may not be the actual distribution of research students. Two, all that matters here is the comparison of blue with orange in a particular category and not cross comparison of orange with orange.



- We note that for universties with rating 1, 3 and 5 (which is really **odd** :p), there isn't much difference in heights whereas the graphs for universities with rating 2 and 4 are biased towards different sides.
- The difference in heights between blue and orange across different categories displays a rather astonishing reality that maybe, for this range CGPA, reseach might not be a good determining factor

Ending Note

It was a wonderful experience to analyze this dataset and it was very insightful for me too. I hope you had the same experience. This report might seem very very basic to you but a tremendous amount of effort has been put in from my side, considering that I haven't even completed my CS101 course.

I know that I am competing against internship-experienced data analysts and my chances of making it in are quite dim. I just want to request you to kindly give me a good feedback so that I can be better prepared for selections next year.

THANK YOU