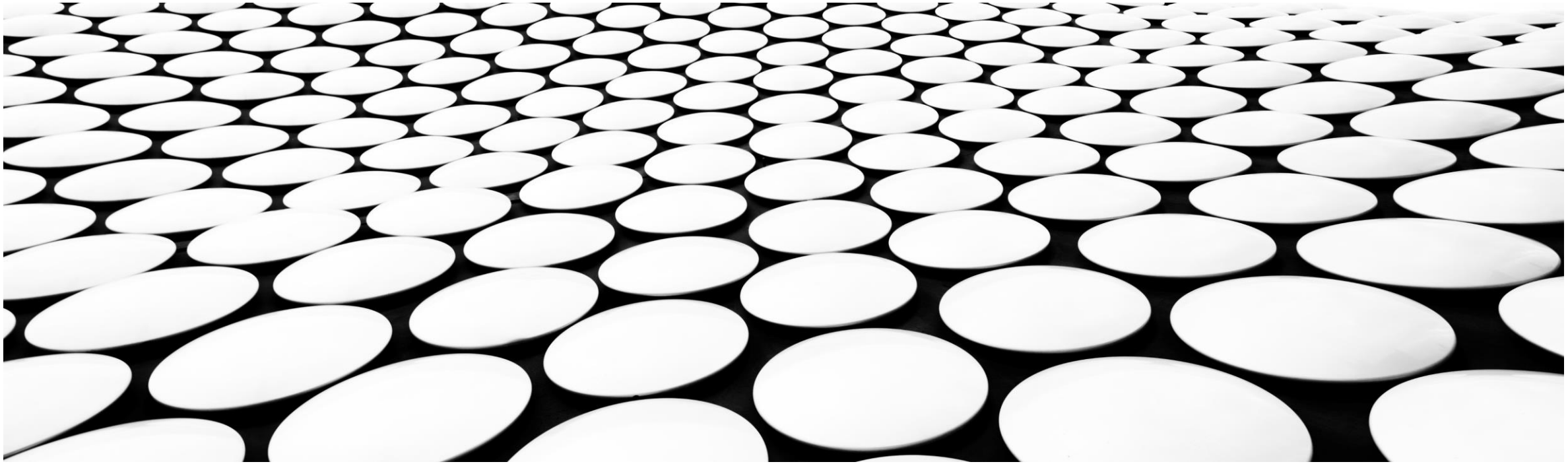
GRADUATE LEVEL ADMISSIONS ANALYSIS

JOHN WACHTER

COMPLETE CODE CAN BE FOUND AT BIT.LY/3HD1VWM



THE DATA

The dataset provides details for 400 students on 9 different features, including one variable that identifies each individual student: Serial No.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65
...	...	...	...	...	...	...	...	...	...
395	396	324	110	3	3.5	3.5	9.04	1	0.82
396	397	325	107	3	3.0	3.5	9.11	1	0.84
397	398	330	116	4	5.0	4.5	9.45	1	0.91
398	399	312	103	3	3.5	4.0	8.78	0	0.67
399	400	333	117	4	5.0	4.0	9.66	1	0.95

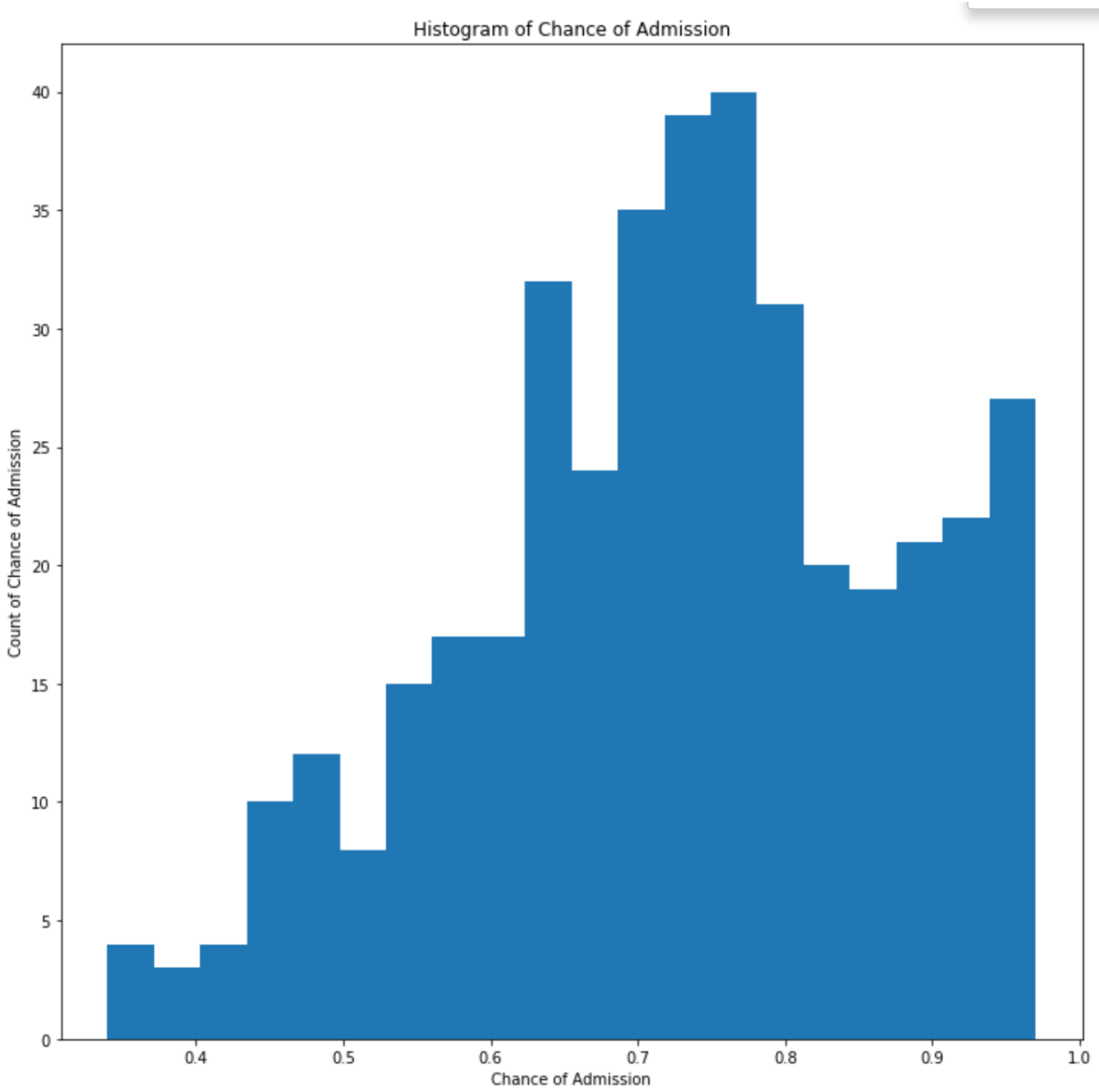
400 rows × 9 columns

VARIABLES

Variable	Detail
Serial No.	Identification Number for each student
GRE Score	Measures scores on the GRE standardized test
TOEFL Score	Score on a test that measures English proficiency
University Rating	Rating for Undergraduate University (higher is better)
SOP	Rating for the Statement of Purpose essay
LOR	Rating for the Letter of Recommendation
CGPA	Undergraduate GPA, on a 1-10 scale
Research	1 indicates student did research as an undergraduate
Chance of Admit	Chances that the student will be admitted

EXAMINING THE DISTRIBUTION OF CHANCE OF ADMISSION

The Chance of Admission variable is my main variable of interest. The data do not follow a normal distribution. Very few students who apply have a low chance of admission; this makes sense, as typically motivated and accomplished individuals apply for graduate school, and so this self-selection (needing to take the TOEFL and GRE) results in good candidates. Most candidates have a chance of admission around .7, but there is another cluster of students at around 95%. This is a bi-modal distribution, but with the centers shifted to the right.



MEAN AND STANDARD DEVIATION OF THE CHANCE OF ADMISSION

Let's get a little more technical and examine the mean and standard deviation to better understand the chance of admission variable. The mean chance of admission, is .72, and the standard deviation is about .14 (the image shows two ways to calculate standard deviation). Standard deviation is a measure of the spread of the data.

Part 6 - Calculate Mean and Standard Deviation of the chance of admission

In [529]:
```
# Calculate mean of chance of admission
data['Chance of Admit'].mean()
```

Out[529]: 0.7243499999999996

In [530]:
```
# Calculate standard deviation of chance of admision - Option 1
np.sqrt(sum([(i - np.mean(data['Chance of Admit']))**2 for i in data['Chance of Admit']])/(len(data
```

Out[530]: 0.1424309569580995

In [531]:
```
# Calculate standard deviation of chance of admission - Option 2
np.std(data['Chance of Admit'])
```

Out[531]: 0.1424309569580995

GRE SCORES FOR HIGH-CHANCE ADMISSIONS

The average GRE scores for students with a chance of admission above 85% was approximately 331, nearly 15 points above the average GRE for all students.

Part 2 - Calculate mean GRE score of students with a chance of admission above 85%

In [564]:
```
# Option 1
gre_85 = data[data['Chance of Admit'] > .85].mean()['GRE Score']

# Option 2
gre_85_2 = data[data['Chance of Admit'] >.85]['GRE Score'].mean()
```

In [565]:
```
# Print both to make sure they are the same value
gre_85, gre_85_2
```

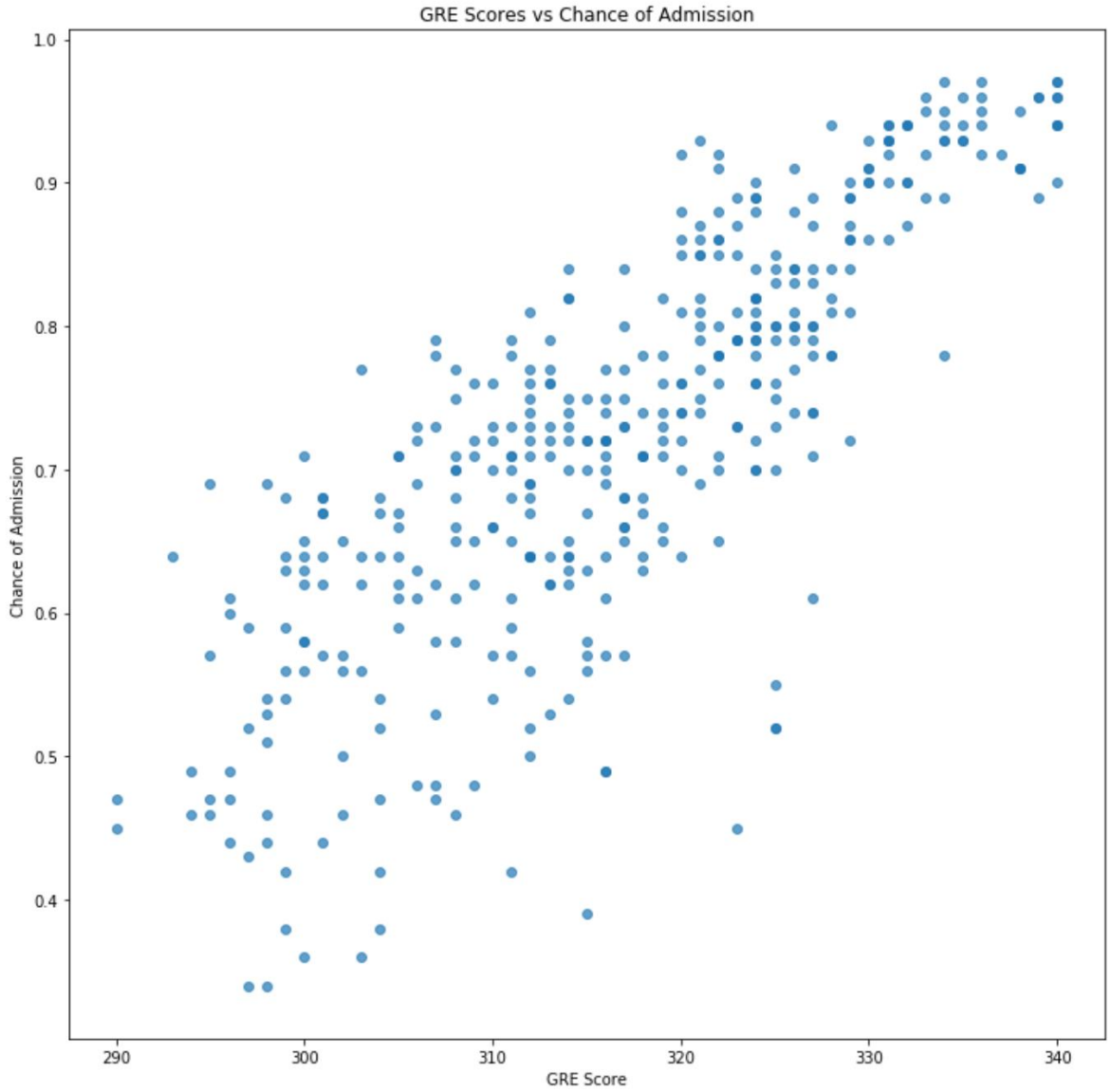
Out[565]: (331.144578313253, 331.144578313253)

In [563]:
```
# Mean score of GRE of
data['GRE Score'].mean()
```

Out[563]: 316.8075

GRE VS CHANCE OF ADMISSION

Let's see how the two variables we have been discussing appear when plotted together. On the x-axis, the GRE score is shown for all students, and their chance of admission is shown on the y-axis. Each dot represents a student, and because some students have the exact same GRE and Chance of Admission, their dots will overlap and appear darker. There is a strong positive association between the two: higher GRE scores are associated with a higher chance of admission. In fact, their correlation coefficient is .8.




```
data['GRE Score'].corr(data['Chance of Admit'])
```

```
0.8026104595903503
```

MEAN CHANCE OF ATTENDANCE FOR STUDENTS WITH RESEARCH EXPERIENCE

The mean chance of attendance for students with research experience is nearly .8, while those without research experience have a mean chance of attendance of only .63. The mean chance of attendance for all students is roughly .72 for a point of comparison.

Part 3 - Calculate mean chance of attendance for students with Research experience

In [524]:
```
# Option 1
data[data['Research'] == 1]['Chance of Admit'].mean()
```

Out[524]:
```
0.7959817351598172
```

In [525]:
```
# Option 2
data.groupby('Research')['Chance of Admit'].mean()
```

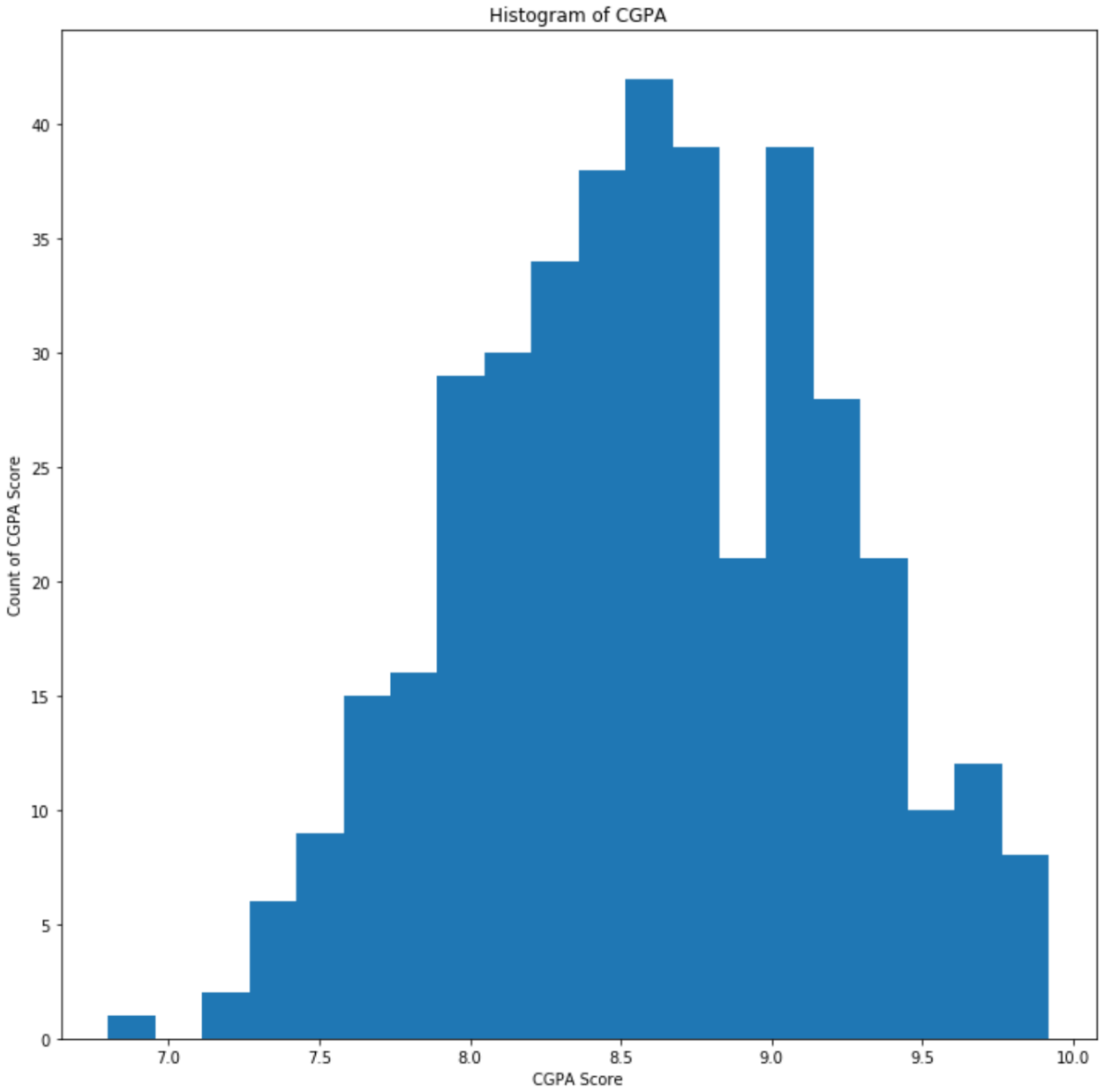
Out[525]:
```
Research
0    0.637680
1    0.795982
Name: Chance of Admit, dtype: float64
```

In [566]:
```
data['Chance of Admit'].mean()
```

Out[566]:
```
0.7243499999999996
```

EXAMINING THE DISTRIBUTION OF CGPA

The CGPA data seem to follow a normal distribution, except for a large gap around 8.8. It makes me wonder if the conversion somehow makes the value less likely, or if this is just a quirk of the small number of observations we have. This is a type of data (similar to test scores) that is typically normally distributed.



MEAN AND STANDARD DEVIATION OF CGPA

Let's get a little more technical and examine the mean and standard deviation to better understand the CGPA variable. The mean CGPA, is .85, and the standard deviation is about .59 (the image shows two ways to calculate standard deviation). Standard deviation is a measure of the spread of the data.

Part 8 - Calculate the mean and standard deviation of the CGPA

In [533]:
```
# Calculate mean of chance of admission
data['CGPA'].mean()
```

Out[533]: 8.598924999999998

In [534]:
```
# Calculate standard deviation of chance of admision - Option 1
np.sqrt(sum([(i - np.mean(data.CGPA))**2 for i in data.CGPA])/(len(data.CGPA)))
```

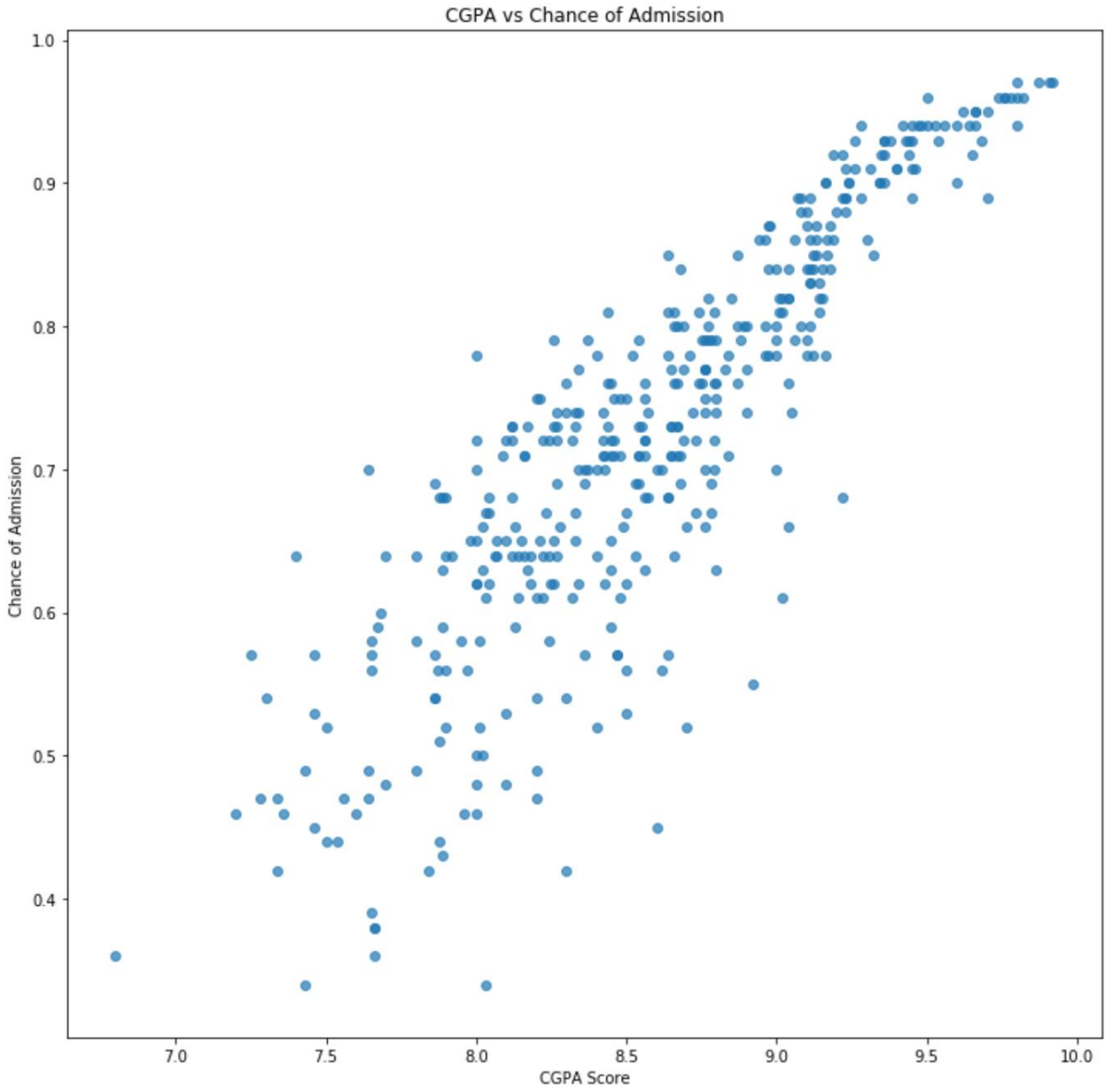
Out[534]: 0.5955712336698271

In [535]:
```
# Calculate standard deviation of chance of admission - Option 2
np.std(data.CGPA)
```

Out[535]: 0.5955712336698271

CGPA VS CHANCE OF ADMISSION

Let's see how CGPA and Chance of Admission appear when plotted together. On the x-axis, CGPA is shown for all students, and their chance of admission is shown on the y-axis. Each dot represents a student, and because some students have the exact same measurements, their dots will overlap and appear darker. There is a strong positive association between the two: higher CGPA is associated with a higher chance of admission. In fact, their correlation coefficient is .87. The association between the two is especially strong when CGPA is very high




```
data['CGPA'].corr(data['Chance of Admit'])
```

```
0.8732890993553001
```

CORRELATION COEFFICIENT BETWEEN CGPA AND CHANCE OF ADMISSION

The correlation coefficient between CGPA and Chance of Admission is .87, indicating that the variables are strongly positively correlated (increasing one variable is associated with an increase in the other)

Part 4 - Calculate the correlation coefficient between CGPA and Chance of Admission

In [526]:
```
# Option 1 - by hand
xy = data.CGPA * data['Chance of Admit']
x2 = data.CGPA**2
y2 = data['Chance of Admit']**2

((len(xy)*sum(xy)) - sum(data.CGPA)*sum(data['Chance of Admit'])) / np.sqrt( (len(x2)*sum(x2) - sum(data.CGPA)**2) * (len(y2)
```

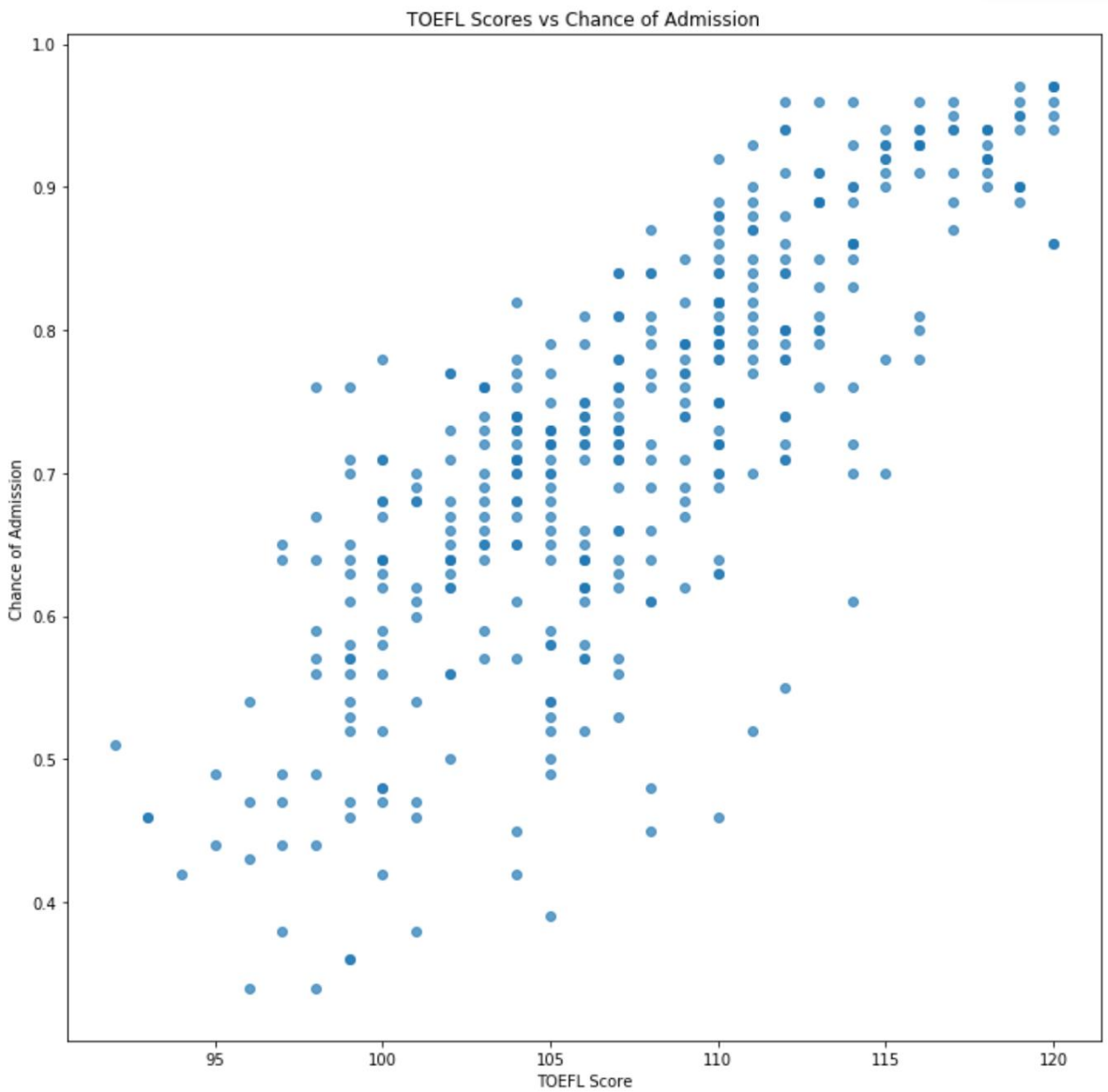
Out[526]: 0.8732890993552173

In [527]:
```
# Option 2 - Use panda's library
data.CGPA.corr(data['Chance of Admit'])
```

Out[527]: 0.8732890993553001

TOEFL SCORE VS CHANCE OF ADMISSION

Let's see how TOEFL Score and Chance of Admission appear when plotted together. On the x-axis, TOEFL score is shown for all students, and their chance of admission is shown on the y-axis. Each dot represents a student, and because some students have the exact same measurements, their dots will overlap and appear darker. There is a strong positive association between the two: higher TOEFL scores are associated with a higher chance of admission. In fact, their correlation coefficient is .79




```
data['TOEFL Score'].corr(data['Chance of Admit'])
```

```
0.7915939869351044
```

CORRELATION COEFFICIENT MATRIX

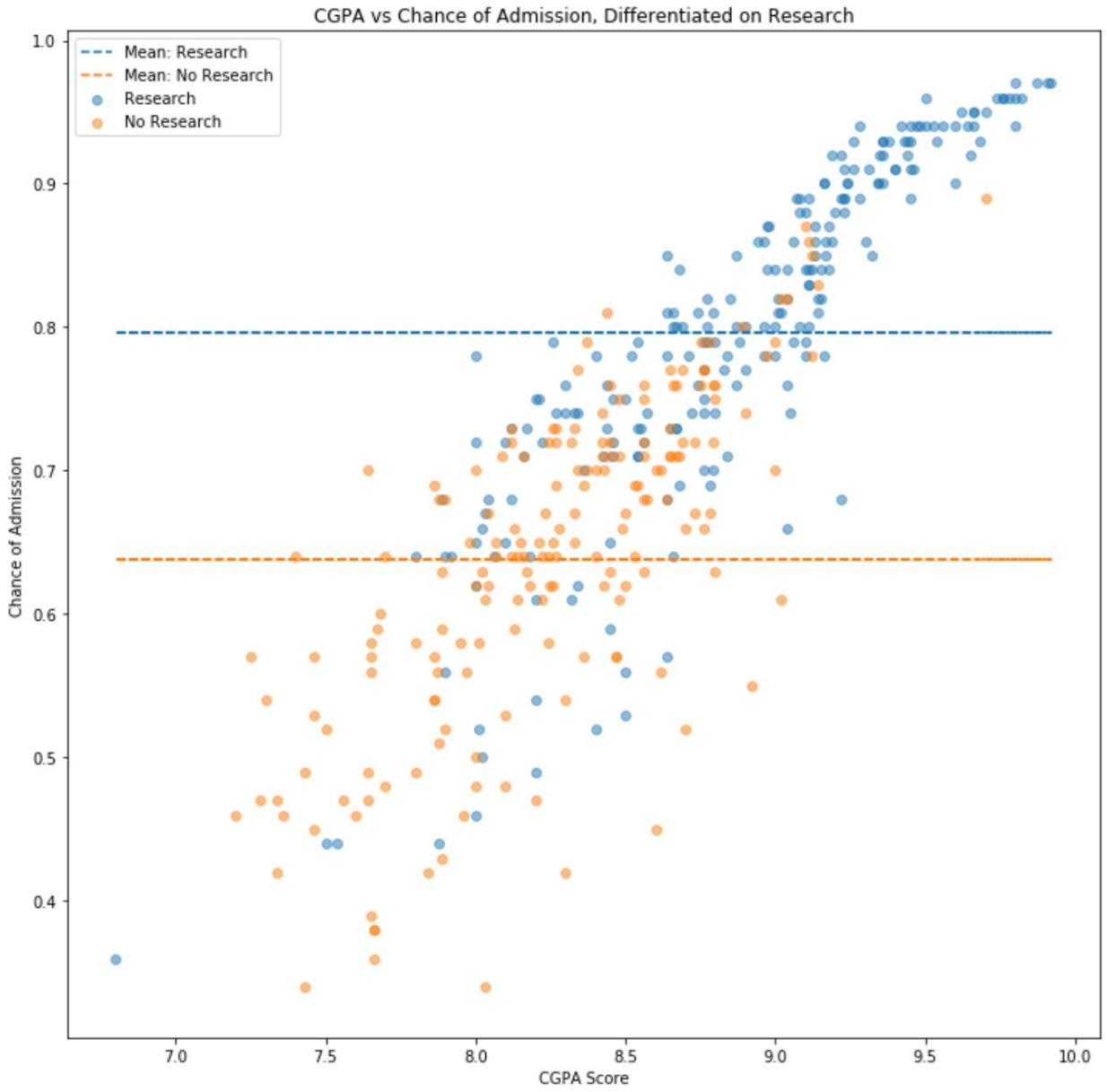
This correlation coefficient matrix shows the correlation coefficients for all variables in the data set. The diagonal of 1s shows each variable's correlation with itself. Negative correlations are indicated in blue, positive correlations in red (darker the hue, stronger the association).

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1	-0.1	-0.1	-0.2	-0.2	-0.09	-0.05	-0.06	0.04
GRE Score	-0.1	1	0.8	0.7	0.6	0.6	0.8	0.6	0.8
TOEFL Score	-0.1	0.8	1	0.7	0.7	0.6	0.8	0.5	0.8
University Rating	-0.2	0.7	0.7	1	0.7	0.7	0.7	0.4	0.7
SOP	-0.2	0.6	0.7	0.7	1	0.7	0.7	0.4	0.7
LOR	-0.09	0.6	0.6	0.7	0.7	1	0.7	0.4	0.7
CGPA	-0.05	0.8	0.8	0.7	0.7	0.7	1	0.5	0.9
Research	-0.06	0.6	0.5	0.4	0.4	0.4	0.5	1	0.6
Chance of Admit	0.04	0.8	0.8	0.7	0.7	0.7	0.9	0.6	1

All variables (excluding Serial No. which is just an ID) are positively correlated with the Chance of Admission. This means that all of the information the admissions officers collect are positive attributes - they are not tracking things like the number of truancys, or number of arrests, which would likely have very little correlation or negative correlation. In other words, if a hypothetical applicant scored high on all of the variables, they would have a very high chance of getting in.

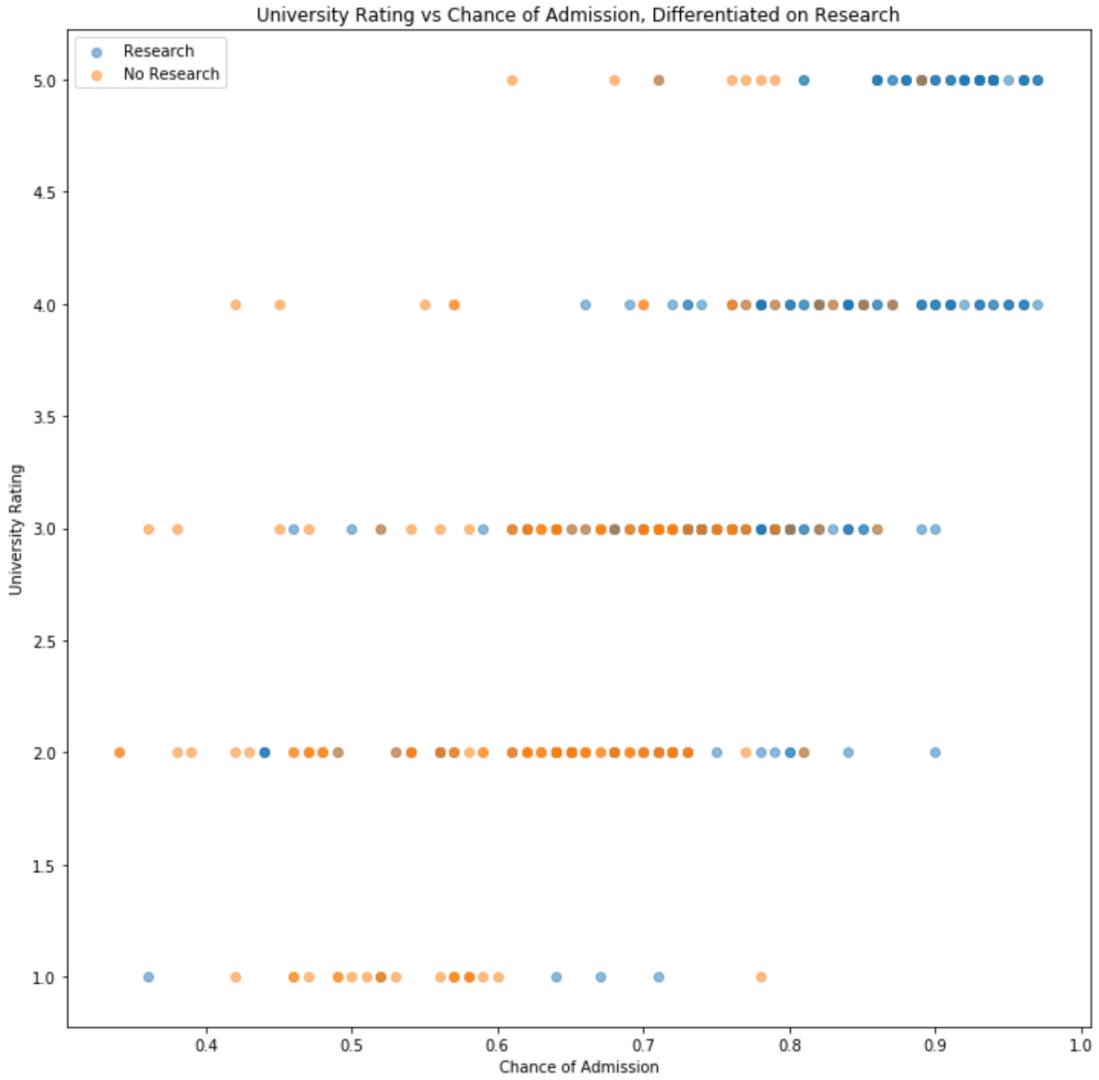
CGPA VS CHANCE OF ADMISSION, DIFFERENTIATED ON RESEARCH

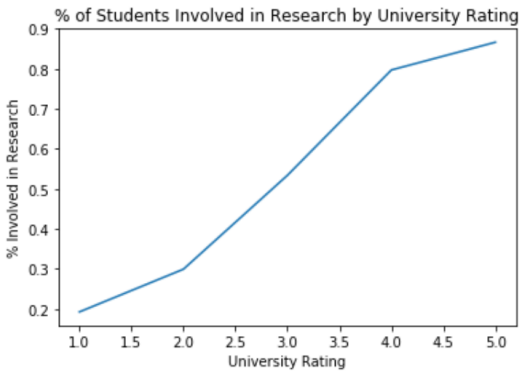
Of all the variables, CGPA is most strongly associated with Chance of Admission, while Research is the least strongly associated. In fact, Research is weakly correlated with most of the other variables: .6 is the strongest correlation. This chart shows a couple things: first, a stronger CGPA score is positively associated with a Chance of Admission, which we already know. Second, this chart shows that students who did not do research tended to have a lower CGPA.



UNIVERSITY RATING VS CHANCE OF ADMISSION, DIFFERENTIATED ON RESEARCH

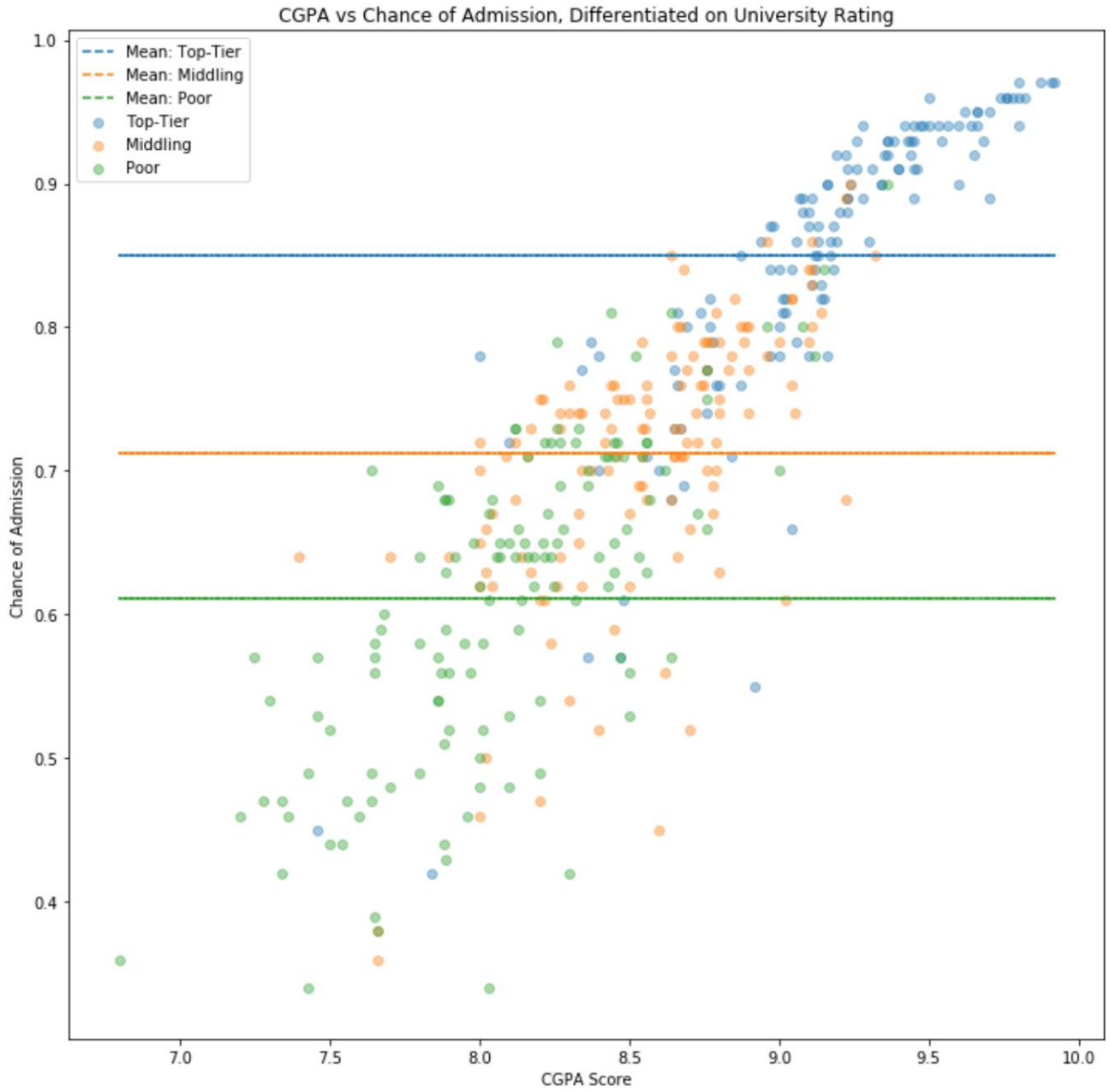
We can see from this chart, where Chance of Admission is now on the X-axis, that highly rated Universities had more people doing research. I would guess this is because highly rated Universities tend to have more funding, be research Universities, and be able to attract faculty that can help coordinate and supervise research activities. The overlapping dots make it difficult to assess how many students were involved in research at each University type, which is shown in the line graph. Clearly, students at better Universities were more involved in research.





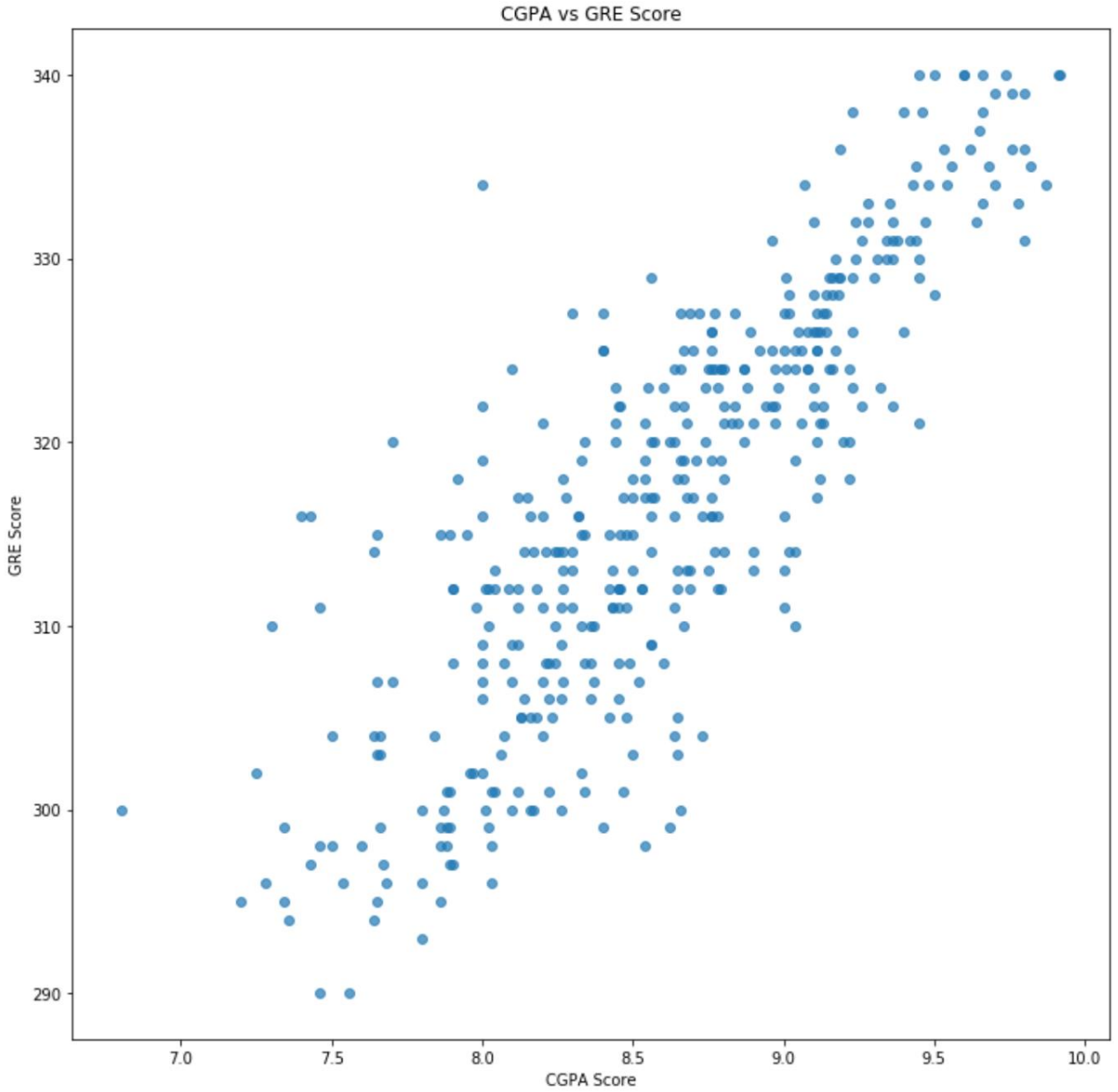
CGPA VS CHANCE OF ADMISSION, DIFFERENTIATED ON UNIVERSITY RATING

We have already seen that higher University Rating is positively correlated with Chance of Admission, but let's take a closer look. To make the view simpler, I will categorize university rating as 'Top-Tier', 'Middling' and 'Poor'. On average, students at poor Universities are clustered in the bottom-left of the chart which corresponds to low CGPA score and low chance of admission. As the University rating gets better, the average student moves up and to the right: better CGPA score and higher chance of admission.



RELATIONSHIP BETWEEN GRE SCORE AND CGPA

Let's this relationship to see if students with good GPAs scored well on the GRE – if true, it could be that these students are 'smarter' or simply better test takers, among other plausible explanations. For the most part, students with high GPAs scored well on the GRE.



LOW CGPA, HIGH GRE SCORE STUDENTS

In the last chart, we saw the that most students with high CGPAs had high test scores, but there were some outliers. Let's look at those students separately and see if there is anything interesting about them. The table on the right shows average score for the Low CGPA, High GRE score students compared to the overall average. Nothing seems striking about these low-CGPA, high-GRE students. The difference between their averages and overall averages are very small. The table above makes me think that the 'Chance of Admit' is formulaic, depending on some combination of the other variables.

	Low CGPA High GRE Average	Overall Average	Difference
GRE Score	322.600	316.807500	5.792500
TOEFL Score	108.200	107.410000	0.790000
University Rating	3.000	3.087500	-0.087500
SOP	3.600	3.400000	0.200000
LOR	3.300	3.452500	-0.152500
CGPA	7.924	8.598925	-0.674925
Research	0.600	0.547500	0.052500
Chance of Admit	0.682	0.724350	-0.042350

CHANCE OF ADMISSION AS A LINEAR COMBINATION OF THE OTHER FEATURES?

As a way to test the hypothesis that the Chance of Admit is some linear combination of the other variables, I find some students with the exact same scores on all other variables and check if they have the same 'Chance of Admit' score. First, I group students by GRE Score, TOEFL Score, University Rating, Research, SOP, and LOR (there are no students that have all the same scores for all variables including CGPA, so CGPA was left out). The table shows unique groupings of students with the same scores, and then the count of the students in each group. For example, two students scored 314 on the GRE, 107 on the TOEFL, went to a 2 rated University, did not do research, had a 2.5 SOP score, and a 4.0 LOR Score (first row).

GRE Score	TOEFL Score	University Rating	Research	SOP	LOR	Serial No.	CGPA	Chance of Admit
314	107	2	0	2.5	4.0	2	2	2
315	105	3	0	2.0	2.5	2	2	2
321	109	4	1	4.0	4.0	2	2	2
326	112	3	1	3.5	3.0	2	2	2
329	111	4	1	4.5	4.0	2	2	2
332	118	5	1	5.0	5.0	2	2	2
335	117	5	1	5.0	5.0	2	2	2
340	120	5	1	4.5	4.5	2	2	2

CHANCE OF ADMISSION AS A LINEAR COMBINATION OF THE OTHER FEATURES? CONTINUED

I randomly select three groupings from the previous table, and explore the different CGPA and Chance of Admission values.

The first pair of students has all the same scores except one has a 9.6 CGPA while the other has a 9.91, and the different in their Chance of Admission is also .3 Perhaps this is evidence of a formula?

The next pair of students has a difference of .29 in the CGPA, but the person with the lower CGPA has a higher Chance of Admission. This seems counterintuitive. Either we do not have complete information, or the Chance of Admission is not formulaic.

The third pair of students has a very small difference in their CGPA (.05) but a difference of .1 in their Chance of Admissions.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
25	26	340	120	5	4.5	4.5	9.60	1	0.94
202	203	340	120	5	4.5	4.5	9.91	1	0.97

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
244	245	314	107	2	2.5	4.0	8.56	0	0.63
322	323	314	107	2	2.5	4.0	8.27	0	0.72

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
104	105	326	112	3	3.5	3.0	9.05	1	0.74
128	129	326	112	3	3.5	3.0	9.10	1	0.84