Collegiate Rankings:
An Exploratory Data Analysis

Kevin Black

June 25, 2023

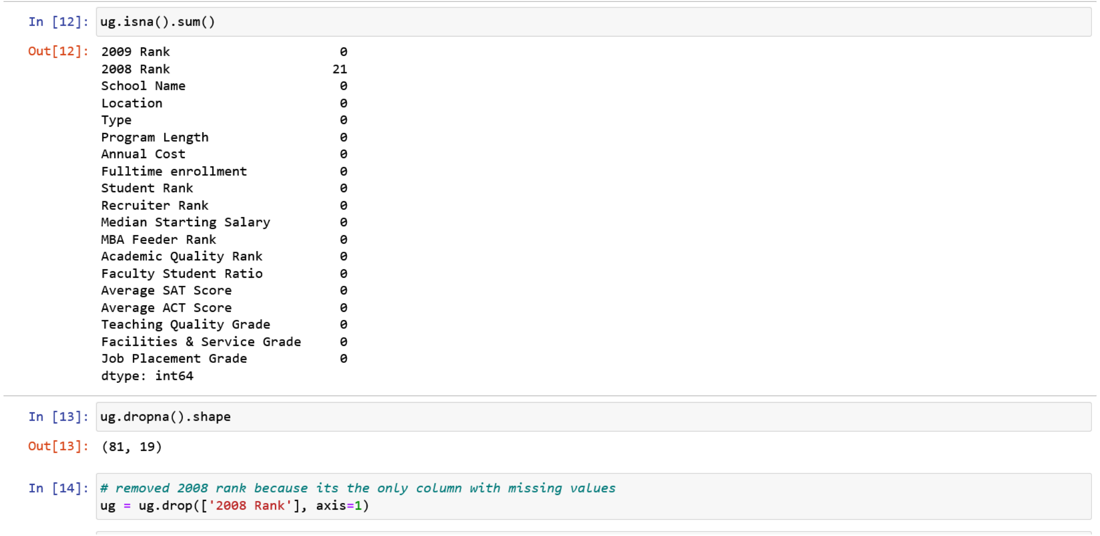
1 Learning About the Data

In order to begin this exploratory data analysis, the first step was to open up the data itself and take a look at it. I opted to load the data into Python right from the start and read it into a pandas dataframe. Upon first looking at the data, I quickly realized there were two separate excel sheets. One sheet contained graduate MBA data, while the other contained undergraduate data. Some of the columns were the same, such as "School Name" and "Type". In addition, they both took into consideration school ranking systems, although these ranking bodies appeared to be different between the datasets. They also both contained variables for starting salaries. However, the MBA data listed the average starting base salary, while the undergraduate data, listed the median starting salary. These are just a few of the similarities and differences I noticed between these two datasets on first glance.

2 Cleaning the Data

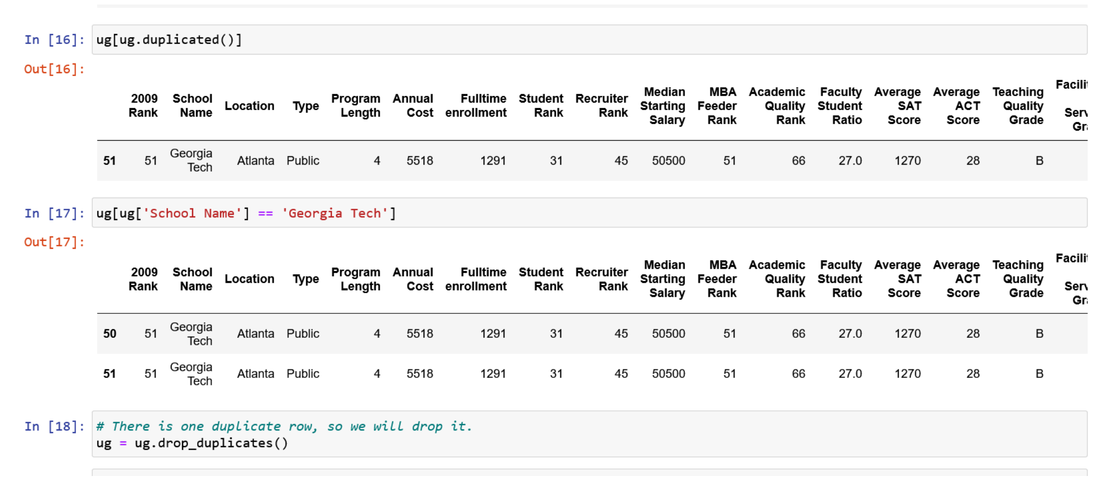
The next step I took was to look into the number of missing values in each dataset. The MBA dataset contained 70 rows and 13 columns. However, upon looking into the missing values, I noticed that there were a decent amount spread fairly evenly across the different variables, so removing a single column wouldn't help. If I removed all rows containing missing values, I would end up with 60 rows, which means I would lose 1/7 of my dataset. After realizing this, I decided to look into the undergraduate dataset to see if I could retain more of my data after cleaning.

The undergraduate dataset contained 102 rows and 19 columns initially. Upon looking at missing values, I noticed that there were 21. If I removed all of the rows with missing values, I would be left with 81 rows, which means I would be losing approximately 1/5 of my dataset. However, I noticed that all of these 21 missing values were in the same column, entitled "2008 Rank". I decided the best course of action was the remove this column, thereby allowing me to retain all 102 rows of my dataset. Losing this column wasn't a big loss because we have an updated 2009 Ranking in the dataset already, so unless I wanted to compare rankings between 2008 and 2009, this column wasn't really necessary.



```
In [12]:  ug.isna().sum()

Out[12]:  2009 Rank                      0
          2008 Rank                     21
          School Name                    0
          Location                       0
          Type                           0
          Program Length                 0
          Annual Cost                    0
          Fulltime enrollment            0
          Student Rank                   0
          Recruiter Rank                 0
          Median Starting Salary         0
          MBA Feeder Rank                0
          Academic Quality Rank          0
          Faculty Student Ratio          0
          Average SAT Score              0
          Average ACT Score              0
          Teaching Quality Grade         0
          Facilities & Service Grade     0
          Job Placement Grade            0
          dtype: int64

In [13]:  ug.dropna().shape

Out[13]:  (81, 19)

In [14]:  # removed 2008 rank because its the only column with missing values
          ug = ug.drop(['2008 Rank'], axis=1)
```

Once I was finished with missing values, I decided to look for duplicate values as well. In doing so, I found that there was one duplicate row for the school "Georgia Tech". Once I determined that all values for the duplicate rows were identical, I removed the duplicate from the dataset.



```
In [16]: ug[ug.duplicated()]
```

Out[16]:

	2009 Rank	School Name	Location	Type	Program Length	Annual Cost	Fulltime enrollment	Student Rank	Recruiter Rank	Median Starting Salary	MBA Feeder Rank	Academic Quality Rank	Faculty Student Ratio	Average SAT Score	Average ACT Score	Teaching Quality Grade	Facili Serv Gr
51	51	Georgia Tech	Atlanta	Public	4	5518	1291	31	45	50500	51	66	27.0	1270	28	B	

```
In [17]: ug[ug['School Name'] == 'Georgia Tech']
```

Out[17]:

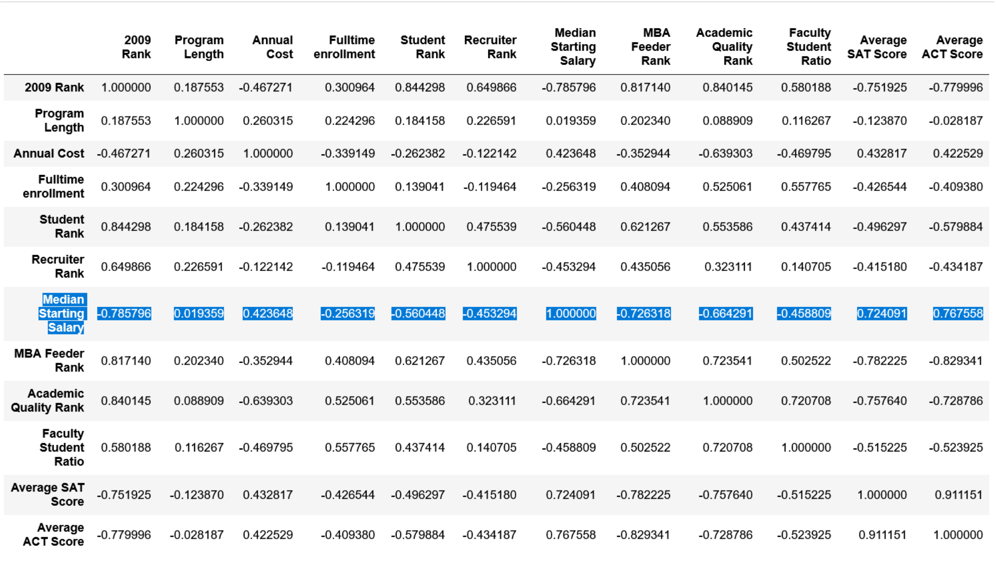
	2009 Rank	School Name	Location	Type	Program Length	Annual Cost	Fulltime enrollment	Student Rank	Recruiter Rank	Median Starting Salary	MBA Feeder Rank	Academic Quality Rank	Faculty Student Ratio	Average SAT Score	Average ACT Score	Teaching Quality Grade	Facili Serv Gr
50	51	Georgia Tech	Atlanta	Public	4	5518	1291	31	45	50500	51	66	27.0	1270	28	B	
51	51	Georgia Tech	Atlanta	Public	4	5518	1291	31	45	50500	51	66	27.0	1270	28	B	

```
In [18]: # There is one duplicate row, so we will drop it.
         ug = ug.drop_duplicates()
```

I began with a dataset of 102 rows and 19 columns. After cleaning, I was left with 101 rows and 18 columns. Therefore, I decided to focus my exploratory analysis on only the undergraduate dataset.

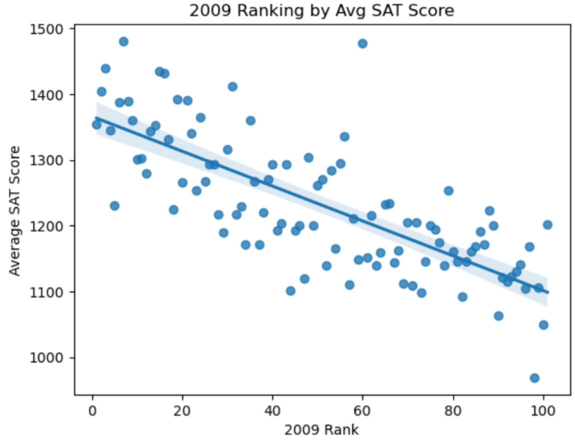
3 Exploring the Data

Now that we have cleaned the data and decided to focus on just the undergraduate dataset, it's time to begin exploring the data. After learning a little bit about the data itself, a idea came to mind. This data could be used to help a high school student who is looking to attend college and hopes to make a certain level of pay upon graduation. In order to help guide a student in this situation, the first thing to do was determine what qualities in a school correlate to a higher median starting salary. My first course of action was to create a correlation plot in order to see which variables, if any, had a noticeable effect on the median starting salary.



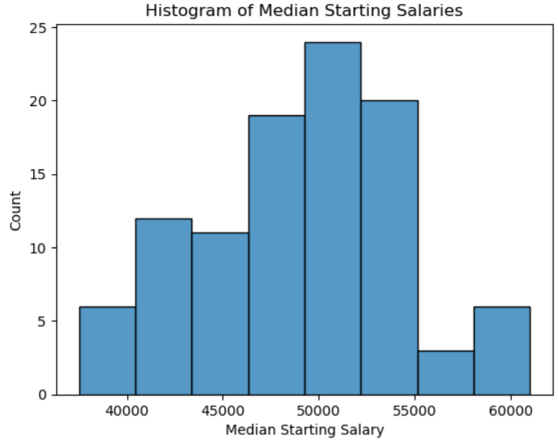
	2009 Rank	Program Length	Annual Cost	Fulltime enrollment	Student Rank	Recruiter Rank	Median Starting Salary	MBA Feeder Rank	Academic Quality Rank	Faculty Student Ratio	Average SAT Score	Average ACT Score
2009 Rank	1.000000	0.187553	-0.467271	0.300964	0.844298	0.649866	-0.785796	0.817140	0.840145	0.580188	-0.751925	-0.779996
Program Length	0.187553	1.000000	0.260315	0.224296	0.184158	0.226591	0.019359	0.202340	0.088909	0.116267	-0.123870	-0.028187
Annual Cost	-0.467271	0.260315	1.000000	-0.339149	-0.262382	-0.122142	0.423648	-0.352944	-0.639303	-0.469795	0.432817	0.422529
Fulltime enrollment	0.300964	0.224296	-0.339149	1.000000	0.139041	-0.119464	-0.256319	0.408094	0.525061	0.557765	-0.426544	-0.409380
Student Rank	0.844298	0.184158	-0.262382	0.139041	1.000000	0.475539	-0.560448	0.621267	0.553586	0.437414	-0.496297	-0.579884
Recruiter Rank	0.649866	0.226591	-0.122142	-0.119464	0.475539	1.000000	-0.453294	0.435056	0.323111	0.140705	-0.415180	-0.434187
Median Starting Salary	-0.785796	0.019359	0.423648	-0.256319	-0.560448	-0.453294	1.000000	-0.726318	-0.664291	-0.458809	0.724091	0.767558
MBA Feeder Rank	0.817140	0.202340	-0.352944	0.408094	0.621267	0.435056	-0.726318	1.000000	0.723541	0.502522	-0.782225	-0.829341
Academic Quality Rank	0.840145	0.088909	-0.639303	0.525061	0.553586	0.323111	-0.664291	0.723541	1.000000	0.720708	-0.757640	-0.728786
Faculty Student Ratio	0.580188	0.116267	-0.469795	0.557765	0.437414	0.140705	-0.458809	0.502522	0.720708	1.000000	-0.515225	-0.523925
Average SAT Score	-0.751925	-0.123870	0.432817	-0.426544	-0.496297	-0.415180	0.724091	-0.782225	-0.757640	-0.515225	1.000000	0.911151
Average ACT Score	-0.779996	-0.028187	0.422529	-0.409380	-0.579884	-0.434187	0.767558	-0.829341	-0.728786	-0.523925	0.911151	1.000000

From this correlation chart, we can see which variables have a positive correlation and which ones have negatives correlations. However, one thing to keep in mind is that we need to understand what our data means, otherwise these numbers can be deceiving. For example, we see a negative correlation between '2009 Rank' and 'Average SAT Score'. This would seem odd at first glance because we would expect the SAT scores to impact 2009 rank in a positive way. We need to remember that our data is numerical and therefore python reads this as rank 100 being higher than rank 1, which we know is not true in real-world context. If we graph it, we can see the relationship more clearly.

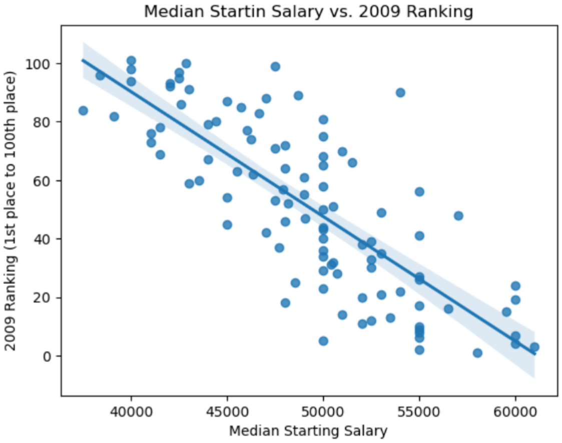


We can see through seaborn's regplot that, although the relationship appears to be negative, if we read the graph correctly, this is not true. When we read the graph, we can see that a 2009 rank of 1 has a much higher Average SAT Score than a rank of 100. This exemplifies just how important it is to pay attention to context and understand the data before jumping to conclusions.

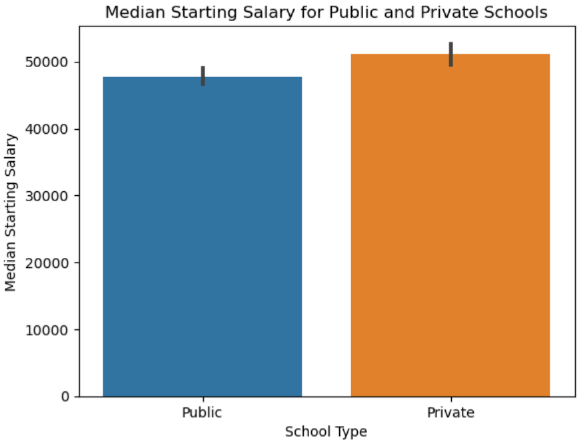
Now that we understand the importance of context, we can continue our analysis. From the correlation chart above, we can see that the largest correlation with 'Median Starting Salary' appears to be with '2009 Rank'. Therefore, we created some visualizations to look at this more closely. First, lets look at the distribution of median starting salaries from schools.



From this histogram, we can see that most starting salaries are between approximately $46,000 and $55,000, with a large drop off after the $55,000 mark. Next, lets visualize the relationship between 'Median Starting Salary' and the variable with the highest correlation, '2009 Rank'.

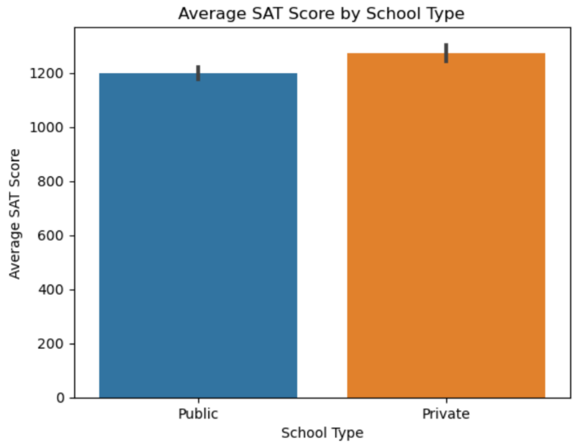


From this seaborn regplot, we can see that students who graduated from schools that had a higher ranking in 2009 tended to demand a higher starting salary. Therefore, if we were trying to guide a high school student looking for a higher end starting salary after college, we would recommend they attend a higher ranked college. In addition to this, the student might wonder if it is best they attend a public school or a private school in order to obtain a higher starting salary. We can look into that as well.



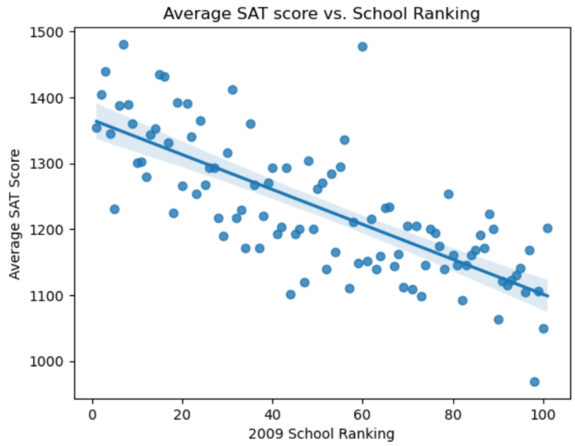
Using the information provided by this bar graph, we would explain to the student that attending a private school tends to result in a higher median starting salary. However, the gap between the two appears to be only a $4,000 to $5,000 per year difference.

Now that we know the importance of attending a higher ranked college, and to a lesser degree a private college, we can help provide the student guidance on how to achieve these goals. One thing they can control is how hard they study for their SAT exam. Therefore, it is important that we look at how a student's SAT scores relate to these goals. First we'll look at how SAT scores relate to attending a private school.



We can see in the above graph that if the student wants to attend a private college, they should strive to have SAT scores above 1200, but if they intend to attend a public college, a score between 1100 and 1200 should be sufficient. In our case, since attending a private school does appear to give a slight edge on starting salary, we would explain to the student that they should do their best to achieve an SAT score greater than 1200.

Next, we'll look at the relationship between SAT scores and attending a top ranked school.



We can see from this plot that a higher SAT score will give the student a better chance at attending a higher ranked college. Although there are some outliers, in general, students who attended a top 20 ranked college had SAT scores above 1300. Therefore, we would suggest that the student aim even higher than we originally thought and aim for an SAT score close to 1300 or above.

4 Conclusion

In completing this project, we were able to take a database of new data, import it into Python, and clean the data to prepare it for analysis. From there, we performed a basic exploratory analysis of the data. However, exploring data provides no value if there's no story to tell. By using using our analysis of the data to help a high school student who may be deciding how to best prepare him/herself to make a high starting salary after college graduation. In doing so, we were able to tell a story through the data and provide value from our exploratory analysis. In the end, our findings were that the student should strive to get into a college that had a higher 2009 ranking. In addition, a private school may also provide a slightly better salary. In order to get accepted into these schools, they should strive for an SAT score close to 1300, and no lower than 1200.