**College Data Analysis**

**Which is the School for me?**

**Team Members:**

**row[0][“Elham Saboori”]**

**row[0][“Nodira Mamatova”]**

**row[0][“Bilwa Buchake”]**

**row[0][“Jayachandran John”]**

**row[0][“Prabhakar Muriki”]**

**Date: 04/05/2018**

**Github: https://github.com/bbuchake/Trend\_Analysis**

**Research Questions:**

1. How does cost of attending a school compare with potential earnings?
2. Is there any correlation between college acceptance rates and cost?
3. Are there any areas in the country where the predominantly expensive/inexpensive schools are located?
4. How do Ivy League schools compare with others in terms of cost of attendance, acceptance criteria and future earnings?
5. Is there any relation between race and being accepted or graduated to the schools
6. How the SAT score could predict the graduation possibility

**Introduction:**

A good education is the foundation of financial success later in life. Finding an affordable school, with a good chance of being accepted and with good earning potential are some of the basic factors when considering a school. In this analysis, we are attempting to find the correlation between these factors.

**Data Sources and Cleanup:**

We used two primary data sources:

1. From the data.gov site, we downloaded College data for the year 2015-16 for over 7000 institutions offering a variety of programs.
2. From <https://www.payscale.com/college-salary-report> , we downloaded salary college data.

Data Cleanup process:

After some deliberation, we decided to focus on those Institutions that predominantly offered Bachelors degrees. We also limited our analysis to Acceptance, Cost and Graduation data. The file we downloaded from data.gov contained over 2000 columns. We used the accompanying Data Dictionary to narrow down the columns we needed. We then removed rows that contained NULL values for the primary measures we were interested in. We ended up with over 2000 institutions in our clean dataset.

Our next data related challenge was to merge this dataset with the salary data. The common field between the two was “Institution Name”. We had to clean up the Institution name in one dataset to match with the second dataset. This included removing special characters, acronyms, spaces etc. to try and find the most number of matches. This exercise resulted in a dataset of 195 records.

**Challenges related to data:**

One of our biggest challenges was finding the “right” dataset for this analysis. There are a lot of datasets available for a similar analysis and the possibilities for research are endless. We found it very challenging to narrow down our scope and weed out the data elements we needed for our analysis. We started off with the IPEDS dataset which is the most popular dataset for College data research. However, downloading the correct data elements with supporting metadata from the IPEDS website was not as easy as we had hoped. After the first few days, we decided to switch to the data.gov dataset (which uses IPEDS as a resource). This caused some delays in working through our analysis. From what we could make out of our dataset, we were not able to easily identify the elements we needed by the Major of study. As a result, all of our analysis is at the Institution level.

We also did not find a dataset comparable in size for the earnings data. Our earnings dataset is much smaller than the College dataset and this has prevented us from analyzing the entire cohort of Institutions we had hoped to analyze.

**Findings and Conclusions:**

1. There are only 9 colleges with an acceptance rate less than 10%.

Among them California Institute of Technology accepts with highest SAT average score: 1544 (ACT average 35.0). That is 47 points higher than Harvard University and 65 points higher that Stanford University!!!

(See: "TopTenByAccRateAndSAT.png)

1. The larger average SAT (ACT) score for acceptance to the college, the more the cost.

(See: "Average SAT vs. Average cost.png”)

1. The States of California, Maine, New Hampshire, Vermont and Florida have colleges with lowest college acceptance rate.

(See: "AcceptanceRateByState.png”)

1. The larger average SAT, the more likely to graduate (See: SATvsGradvsCost.png)
2. Race has no effect on neither being accepted nor graduated from college. (RaceVsAcceptance.png and RaceVsGraduation.png)
3. The most expensive schools are found on the East Coast and the West Coast of the country. (See: College\_Cost\_Distribution.png and Average\_Cost\_Distribution.png)
4. The Salary range has been consistently higher for Ivy League Colleges for Starting Median, Mid-Career Median and Mid-career 90th percentile as compared to other College types. (See: Salary\_Range\_vs\_College.png)
5. The Salary range for Northeastern Colleges has been higher for Starting Median, Mid-Career Median and Mid-career 90th percentile as compared to college in other regions. (See: Salary\_Range\_vs\_Region.png)
6. Major in fields such as Nursing, Nutrition, Education, Forestry, Health Care Administration have seen very low salary figure. (See: Starting\_Median\_Salary\_vs\_Major.png)
7. The ROI (Return on Investment) Scatter plots shows some interesting observations.
   1. The starting median salary data plots are the same for low to high cost of education.
   2. The mid-career median salary data plots are pretty much consistent and stay same for the low to high cost of education
   3. The Mid-career 90th percentile is where we see some good outliers, where we see some very high salary plots for the high cost of education.

(See: Salary\_Range\_vs\_CostOfEducation.png)

The returns on investment from the most expensive schools can be seen more prominently over mid to late career. (See: Salary\_Range\_vs\_CostOfEducation.png)

**Implications of our findings:**

1. Students with higher SAT scores are more likely to graduate.
2. Lower the rate of acceptance, the costlier the school.
3. Starting median salary is not an indicator of ROI. It can be clearly seen that the mid-career salary is greatly influenced by whether or not the education was done in a high cost, low acceptance rate school.

**What we learned:**

1. A good dataset is the most important piece of the puzzle. But it may not always be possible to find the perfect dataset for analysis. It is therefore important to be cognizant of the limitations of your dataset and define the scope of your research accordingly.
2. While using new libraries, it was very important to carefully review the documentation to not have nasty surprises at the end. For example, after finishing the plot using folium, we discovered that there was no easy way to add titles and legends to the plots. Also, the plots could not be saved as PNG files.
3. Advanced libraries such as Seaborn library have been used for our project. We have made use of box , swarm, bar and scatter plots for the data visualization.
4. We have also made use of pandas dataframe and numpy modules.

**The Story:**

One has to be extremely smart and rich to go to a school for extremely smart and rich people to become smarter and get richer!!!