Data Synthesis Narrative

Lavanya (22110130)
Department of Computer Science and Engineering
Indian Institute Of Technology Gandhinagar
Gujarat, India
lavanya.lavanya@iitgn.ac.in

Abstract— This report presents an analysis of datasets containing information about various colleges. The analysis aims to explore patterns and trends in the data and provide insights on the performance of different institutions in different aspects. The report employs data visualization tools , such as scatterplots, histograms, pie charts and box plots to present the findings in a clear and intuitive manner. The analysis also reveals several interesting insights such as the relationship between graduation rate and acceptance rate, state wise distribution of different colleges ,etc.

I. Overview Of The Dataset

The 1995 Data Analysis Exposition, organized by the Statistical Graphics Section of the American Statistical Association, provides a set of datasets for statisticians to analyze and display results at a special session during the Joint Statistical Meetings. The Datasets are derived from two sources, namely the U.S News and World Report's guide to America's Best colleges and the AAUP 1994 Salary survey.

The US News data comprises information on over 1300 colleges, including tuition costs, room/board costs, student/faculty ratios, graduation rates, public/private colleges, instructional expenditure per student and applications received. On the other hand, the AAUP data mainly provides information on faculty salary and ranks.

II. Details of Libraries And Functions

There are various libraries in python, which were of great help to perform different operations on our Dataset.

Libraries used:

A. Pandas (I): Pandas is a python library used for working with datasets. It has functions for analyzing, cleaning, exploring and manipulating data.

Some of the functions I used from Pandas are:

- Read_csv: It is used to load a CSV file into a dataframe.
- Value_counts: It is used to get a series containing counts of unique values.
- Index: It returns the index information of a dataframe.
- Isin: It checks if the dataframe contains the specified values and can also be used to display or extract information common to two dataframes.
- Sort_values: It sorts the dataframe by the specified column name in either ascending or descending order.
- Len: used to return the number of rows in a dataframe.
- Replace: It is used to replace a string, list, dictionary, series, number from a Pandas Dataframe in Python.
- Astype: It is used to cast a Pandas object to a specified datatype.
- Groupby: It is used for grouping the data according to the categories and apply a function to the categories, which helps to aggregate the data efficiently.
- Head: It returns the first n rows from the specified dataframe.
- tolist(): It is used to convert a series type object into a list.
- Reset_index: It allows to reset the index back to the default or original indexes of a dataframe
- Pivot: It is used to reshape a given dataframe organized by given index/column values.

B. Matplotlib (II) : Matplotlib is a visualization library in python for 2D plots of arrays. It is a multi-platform data visualization library built on Numpy arrays

Some of the functions I used from Matplotlib are:

- Plot: It is used to visualize a dataframe by plotting by plotting different columns into a graph.
- Set_xticklabels: It is used to set the xtick labels with a list of string labels.
- Xlabel and ylabel: It is used to provide labels to the x-axis and y-axis respectively.
- Title: It is used to add a title to a plot.
- Scatter: IIt is used to create a scatter plot for the given data values.
- Legend: A legend is an area describing the elements of a graph. The legend function

in Matplotlib is used to place a legend on the axes.

C. Numpy (III) : It is a python library used for working with arrays. It also has functions for working in the domain of linear algebra, fourier transform and matrices.

III. Hypotheses .

A. What is the observed trend in the average compensations of full, associate and assistant professors across multiple colleges situated in different states?

B. Is it possible to assert that the number of full professors would be a crucial determinant in the average compensation- all ranks across various colleges?

C. Is there a correlation in the comparative count of Full, assistant and associate professors in diverse types of colleges?

D. What is the comprehensive analysis of the nationwide distribution of diverse college types, such as I,IIA, IIB, etc. across the United States?

E. What are the differences and similarities in the median salaries of full, associate and assistant professors across the United States?

F. Assuming that Instructional expenditure is the amount spent by each student in the college, is it possible to assert that the greater the percentage of alumni who donate, lesser is the instructional expenditure per student?

G.Could it be a valid assumption that some students who have been granted admission to a college or university do not ultimately enroll due to financial constraints related to the institution's tuition fees?

H. In terms of SAT scores, how do the admission requirements of top public and private colleges compare with each other?

I. Can we make a hypothesis that there is a correlation between the college's acceptance rate and its graduation rate ,such that the lower the acceptance rate the higher the graduation rate?

J. What conclusion can we draw from examining the distribution of public and private colleges across different states in the United States?

IV. Answer to the Hypothesis

A. What is the observed trend in the average compensations of full, associate and assistant professors across multiple colleges situated in different states?

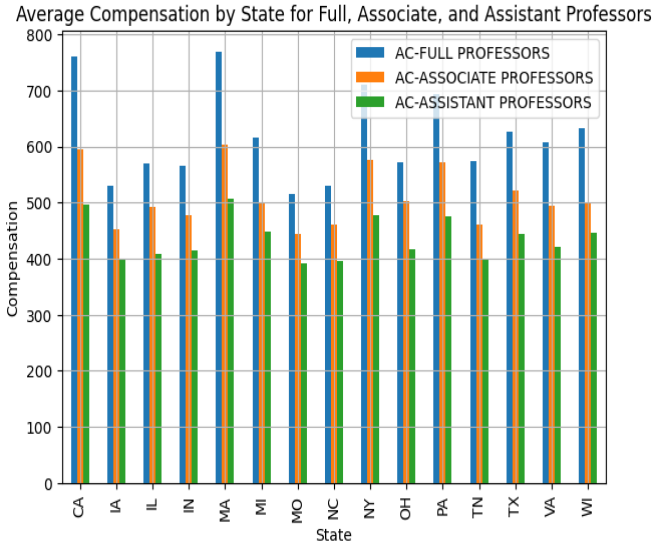

Figure 1.

The graph in fig1. is plotted for top 10 states having the highest number of colleges among all other states. From the graph, it is evident that the average compensation for each category decreases in the order of full, associate and assistant professors. Based on this trend, it is reasonable to conclude that full professors generally receive higher average compensation compared to their counterparts in associate and assistant professor roles.

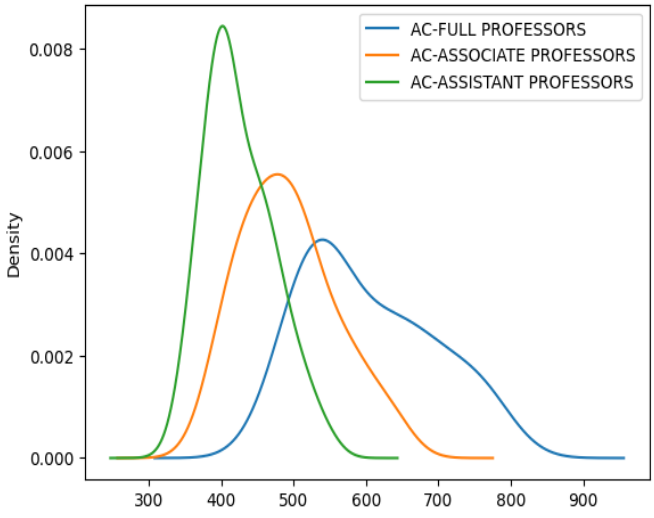

Figure 2..

The same conclusion can also be drawn from the density graph in figure 2., where the peak for the average compensation of Full professors is shifted towards right, indicating a higher concentration of data points with higher average compensation values.

B.Is it possible to assert that the number of full professors would be a crucial determinant in the average compensation- all ranks across various colleges?

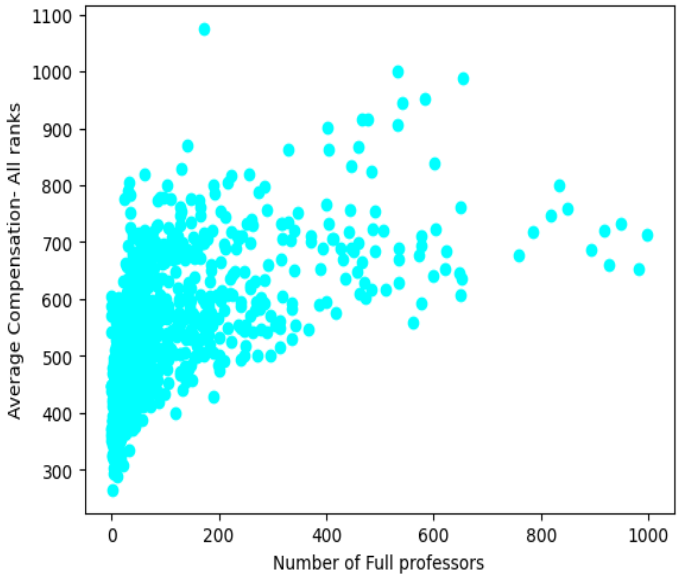

Figure 3.

From fig. 3, it is clearly observed that the data points in the scatter plot are moving in an upward direction from left to right, indicating that the Average Compensation- All ranks increases with the increase in the number of Full Professors. The relation observed is also practical, since Full professors tend to have more experience and higher qualifications and thus may command higher salaries than others. Hence, the assertion is correct.

C Is there a correlation in the comparative count of Full, assistant and associate professors in diverse types of colleges?

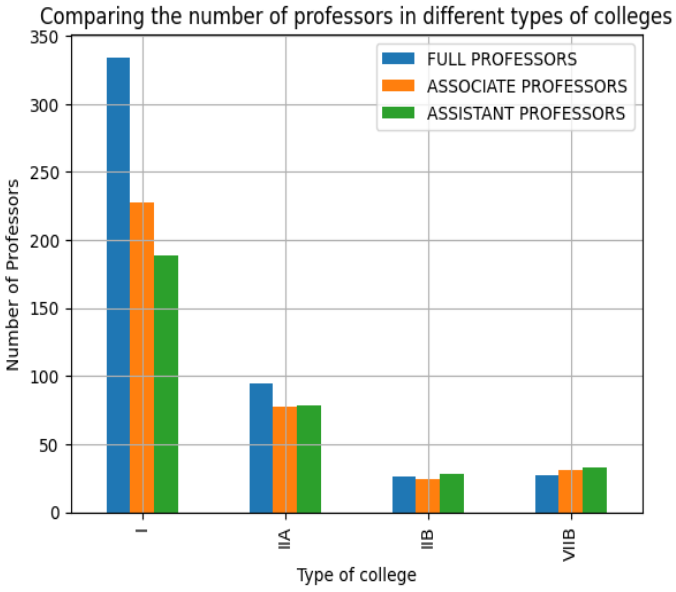

Figure 4.

From Fig.4, we can observe that the number of professors in

colleges tend to decrease in the order of I, IIA, IIB and VIIB, indicating that colleges with a better reputation tend to have more professors. Furthermore, it can be observed that, for each type of college, the number of full professors is usually higher than the number of associate and assistant professors in most cases. This also suggests that full professors may be more valuable to academic institutions due to their prior experience and higher qualifications.

D. What is the comprehensive analysis of the nationwide distribution of diverse college types, such as I,IIA, IIB, etc. across the United States?

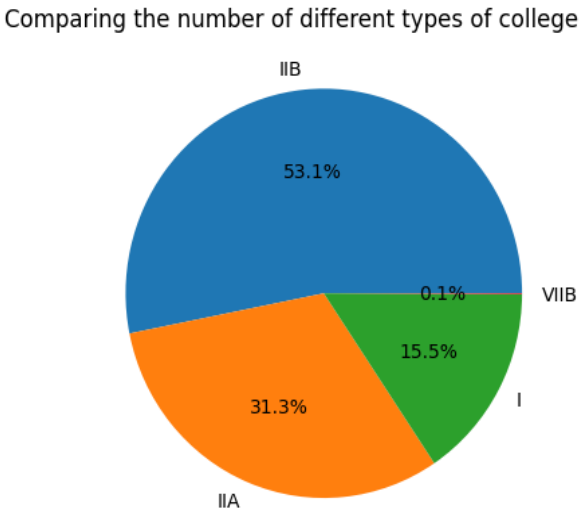

Figure 5.

From fig.5, we can observe that there is a gradual decrease in the number of colleges and universities in the United States as we move from the categories of IIB, IIA, I to VIIB. This indicates that the availability of prestigious higher education universities is relatively less in comparison to other colleges, therefore there is a higher level of competitiveness to secure admissions in these institutions.

E. What are the differences and similarities in the median salaries of full, associate and assistant professors across the United States?

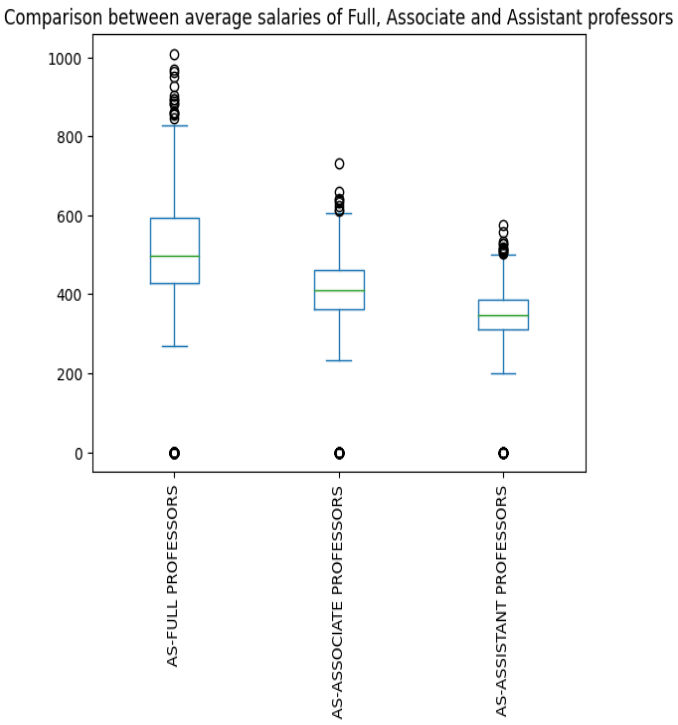

Figure 6.

From fig. 6, it can be observed that the median of average salaries for professors decreases in the order of Full, Associate and Assistant professors. Also, the range of salaries which encompasses the minimum and maximum salary values, is greatest for full professors.

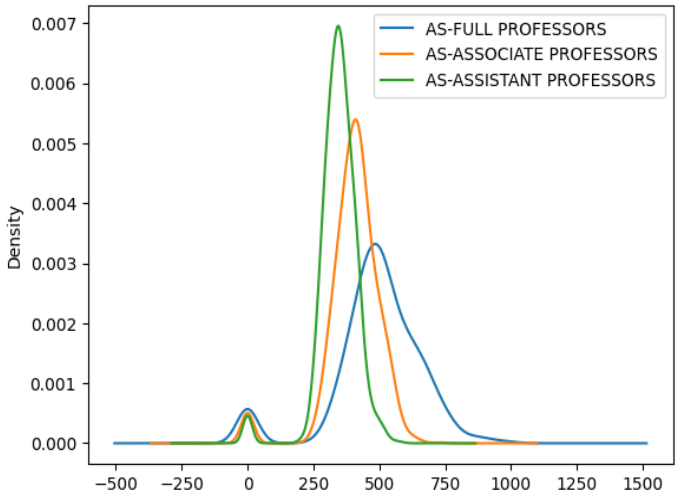

Figure 7.

A similar conclusion can be drawn from the density graph in fig.7. , where the peak for average salaries of Full Professors is shifted towards right, indicating a higher concentration of data points with higher average salary values.

Since a similar trend was observed for average compensations of professors, we can conclude that there is a direct relationship between a professor's salary and their compensation. Specifically, professors receiving higher salaries are expected to receive higher compensations.

F. Assuming that Instructional expenditure is the amount spent by each student in the college, is it possible to assert that the greater the percentage of alumni who donate, lesser is the instructional expenditure per student?

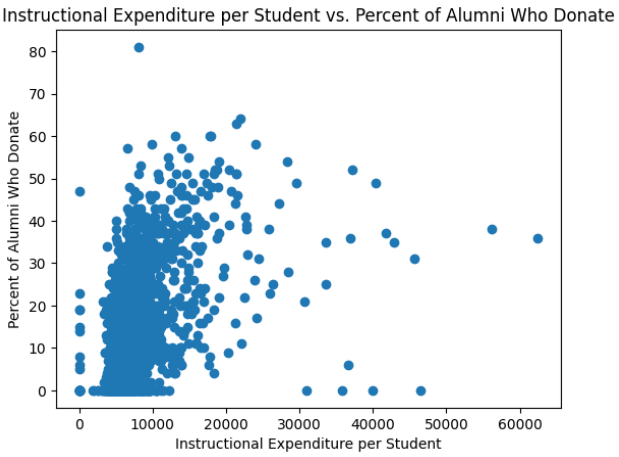

Figure 8.

Initially, it was expected that with the increase in percentage of alumni who donate, the institution would receive a greater amount of funds, leading to a corresponding decrease in the instructional expenditure per student, assuming that instructional expenditure is the amount spent by each student in the college. From fig.8, we can observe that this expected trend is not observed across many colleges. It also suggests that there must be other factors influencing the instructional expenditure per student other than the amount of funds college receives.

G. Could it be a valid assumption that some students who have been granted admission to a college or university do not ultimately enroll due to financial constraints related to the institution's tuition fees?

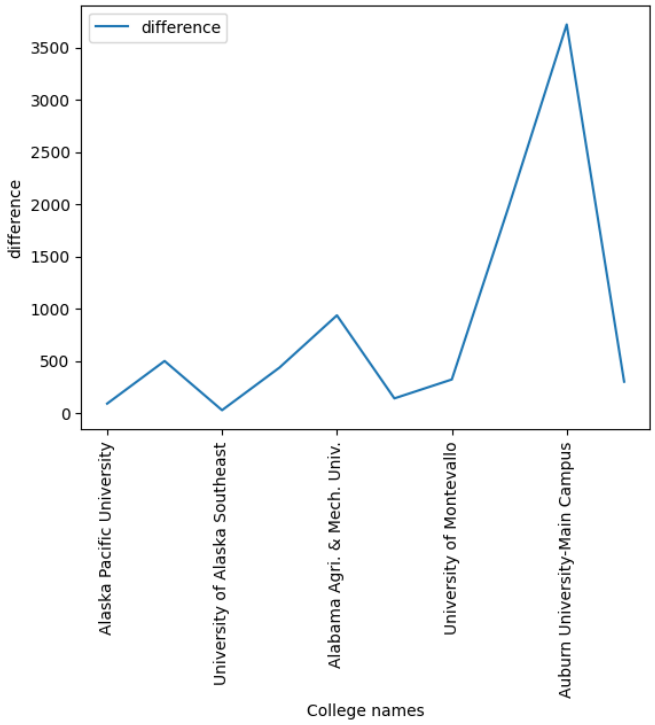

Figure 9.

In order to evaluate the validity of the assumption taken, a sample of ten colleges, based on having the highest Out-of-State tuition fees, was selected for analysis(only seven are displayed in the graph). Specifically, the difference between the number of students accepted and the number of students who actually enrolled in these colleges was observed. Figure 9 displays the corresponding results, which indicate that there exists considerable variation in this difference across the sample of ten colleges, with no particular trend observed between this difference and the cost of college. This finding suggests that students may take into account several other factors in addition to tuition fees before deciding to enroll in a particular college. Some of these factors could include student-faculty ratio, graduation rate, and other related considerations.

H. In terms of SAT scores, how do the admission requirements of top public and private colleges compare with each other?

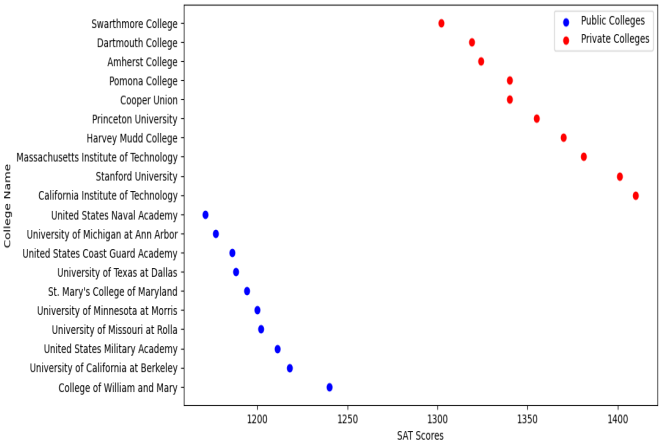

Figure 10.

To compare the SAT scores of public and private colleges, we have considered ten public and private colleges, based on having the highest SAT score requirement for admission. Upon examination of the corresponding scatter plot in fig.10, it can be evident that the SAT scores of the top private colleges are higher in comparison to those of top ranked public colleges. It also indicates a difference in competitiveness in selection for admission in public and private colleges.

I. Can we make a hypothesis that there is a correlation between the college's acceptance rate and its graduation rate ,such that the lower the acceptance rate the higher the graduation rate?

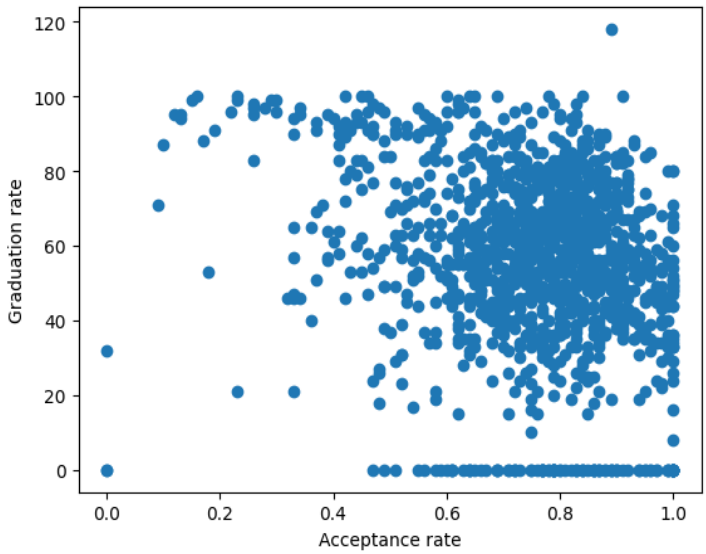

Figure 11.

The initial assumption was that colleges with lower acceptance rates may have higher graduation rates, since they are admitting a more selective and academically capable student body. On the other hand, colleges with higher acceptance rates may have lower graduation rates, since they may be admitting students who are less prepared or motivated to complete their degrees.
From the scatter plot in fig.11, it is observed that although many colleges follow the assumed trend, there are a lot of exceptions to the trend. This observation helps us conclude that there are many factors apart from acceptance rate that influence a college's graduation rate. Some of the determining factors may be quality of education of college, resources available to students, etc.

J. What conclusion can we draw from examining the distribution of public and private colleges across different states in the United States?

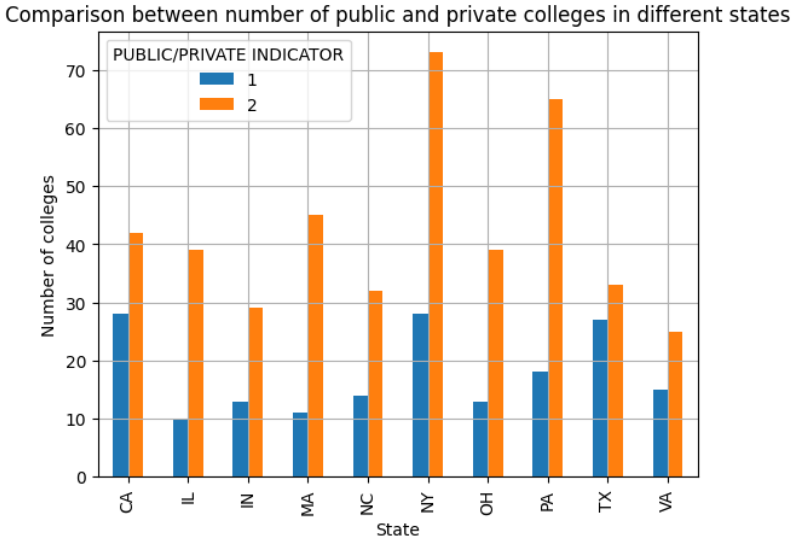

Figure 12.

To draw the aforementioned comparison, we have considered the top ten states having the highest number of colleges. From the graph in fig.12, it can be observed that for each of the considered states, the number of public colleges (indicated by 1) is less than the number of private colleges (indicated by 2).

V. SUMMARY OF THE OBSERVATIONS

The analysis of the data sets has yielded several interesting observations.. It appears that full professors tend to receive higher average compensation compared to associate and assistant professors due to their greater experience and qualifications. Moreover, colleges with a better reputation tend to have more professors, with full professors being more valuable to academic institutions. There is also a direct relationship between a professor's salary and their compensation. Additionally, the availability of prestigious higher education universities in the United States is relatively low, resulting in greater competitiveness for admissions. Also, the expected trend of instructional expenditure per student decreasing with an increase in alumni donations is not always observed. Furthermore, private colleges tend to have higher SAT scores as compared to public colleges. Finally, although there is a perceived correlation between acceptance rate and graduation rate, many other factors can influence a college's graduation rate.

VI. Acknowledgement

I thank Professor Shanmuganathan Raman, department of Computer Science and Engineering and Electrical Engineering, Indian Institute of Technology, Gandhinagar for your invaluable guidance and support. Your dedication to teaching and research has been a source of inspiration to me.

VII. References

(I) "Pandas Introduction," n.d.
https://www.w3schools.com/python/pandas/pandas_intro.asp.
(II) GeeksforGeeks. "Python Introduction to Matplotlib," May 14, 2018.
https://www.geeksforgeeks.org/python-introduction-matplotlib/.
(III) "Introduction to NumPy," n.d.
https://www.w3schools.com/python/numpy/numpy_intro.asp.