ES114 Probability, Statistics and Data Visualization

Data Narrative

Astitva Aryan, Mechanical Engineering,
B. Tech'22, Roll Number-22110041,
Indian Institute of Technology
Gandhinagar, astitva.aryan@iitgn.ac.in

***Abstract*— This report is a Data Narrative on AAUP dataset and USNEWS dataset. The report significantly contains scientific questions/ hypotheses on these two datasets, graphs and answers of these questions and summary of the observations on the graph.**

I. Overview of the Datasets

The AAUP dataset provides information about various colleges and universities in the US. It includes the Federal ID number, college name, state, type of college (I, IIA, or IIB), and salary and compensation data for full, associate, and assistant professors. The dataset also includes information on the number of full, associate, and assistant professors, instructors, and faculty members overall for each institution. There are a total of 1,161 rows and 17 columns in this dataset.

The USNEWS dataset also includes information about various colleges and universities across the United States, including their Federal ID number, name, location, public/private indicator, average SAT and ACT scores, number of applications received and number of students enrolled, tuition costs, and other related information such as room and board costs, etc. And this dataset have a total of 1,302 rows and 35 columns.

II. Scientific Questions/Hypotheses

Question 1: *Is there a significant variation in the number of faculty across different states in the United States?*
Question 2: *What is the PMF and CDF of average salaries across all ranks for faculty members?*
Question 3: *Is there a correlation between the average salary and the average compensation of faculty members? For example, do colleges that pay their faculties more also provide higher compensation?*
Question 4: *Does the average compensation of faculty members vary based on the location of the college? For example, do colleges in certain states pay their faculty members more than colleges in other states?*
Question 5: *Is there a significant difference in the median number of faculty members among different types of colleges in the United St*
Question 6: *Is there a statistically significant difference in acceptance and enrollment rates between private and public colleges?*
Question 7: *What is the proportion of colleges by comparison of in-state and out-of-state tuition fees?*
Question 8: *What are the top 50 colleges with the highest average total expenditures by a student other than tuition fees?*
Question 9: *What are the top 50 colleges with the lowest graduation rates in the USA?*
Question 10: *What is the relationship between the acceptance rate and the average SAT score of the top 50 US colleges?*

III. Details of Libraries and Functions

Libraries used:

- *Pandas library:* It is used to analyze Dataset.
- *Matplotlib library:* It is used to create visual representations for the Dataset like bar graphs, pie charts, etc.
- *Numpy library:* It provides fast and efficient tools for working with numerical data.

Functions used:

- *groupby():* It groups rows in a DataFrame based on a specified column or set of columns.
- *sum():* It calculates the sum of values in an array.
- *bar():* It creates a bar chart with the specified data and labels.
- *xlabel():* It sets the label for the x-axis.
- *ylabel():* It sets the label for the y-axis.
- *title():* It sets the title of the chart.
- *xticks():* It sets the labels for the tick marks on the x-axis.
- *rotation():* It rotates the tick labels on the x-axis.
- *agg():* It applies an aggregation function to grouped data.
- *reset_index():* It resets the index of a DataFrame.
- *replace():* It replaces a specified value in a DataFrame with a new value.
- *dropna():* It removes any rows with missing values from a DataFrame.
- *astype():* It converts the data type of a column in a DataFrame to a specified type.
- *unique():* It returns an array of unique values in an array.
- *cumsum():* It calculates the cumulative sum of values in an array.
- *plot():* It creates a line chart with the specified data and labels.
- *scatter():* It creates a scatter plot with the specified data and labels.
- *figure():* It creates a new figure for a plot.
- *mean():* It calculates the mean of values in an array.
- *sort_values():* It sorts a DataFrame by the values in a specified column.
- *pd.read_csv():* It is used to read a CSV file into a pandas DataFrame.
- *b[['column_name']]:* It is used to select one or more columns from a pandas DataFrame.

- *apply(pd.to_numeric, errors='coerce'):* It is used to convert the data type of a column to numeric values. The errors='coerce' parameter tells pandas to convert any non-numeric values to NaN.

- *plt.stem():* It is used to plot a stem plot.

- *plt.ylim():* It is used to set the limits for the y-axis of a plot.

- *enumerate():* It is used to iterate over a sequence and provide an index to each element.

- *plt.annotate():* It is used to add text annotations to a plot.

- *plt.show():* It is used to display a plot.

- *plt.pie():* It is used to create a pie chart.

- *plt.legend():* It is used to add a legend to a plot.

- *pd.to_numeric():* It is used to convert the data type of a column to numeric values.

- *plt.subplots_adjust():* It is used to adjust the spacing between subplots in a figure. It can also be used to fine-tune the position of subplots.

IV. Answers to the Questions

1. The first bar plot shows the total number of faculty for each state, with each state represented by a bar. The second bar plot shows the standard deviation of the number of faculty for each state. Looking at the first bar plot, we can see that California has the highest number of faculty, followed by New York and Pennsylvania. On the other hand, some states such as Wyoming, Alaska, and Vermont have a relatively low number of faculty. The second bar plot shows the variation in the number of faculty across states, with a higher standard deviation indicating a greater variation. We can see that some states such as Arizona, Hawaii, and Florida have a relatively high standard deviation, indicating a greater variation in the number of faculty among universities within those states. Overall, these plots suggest that there is significant variation in the number of faculty across different states in the United States, with some states having a much larger number of faculty than others.

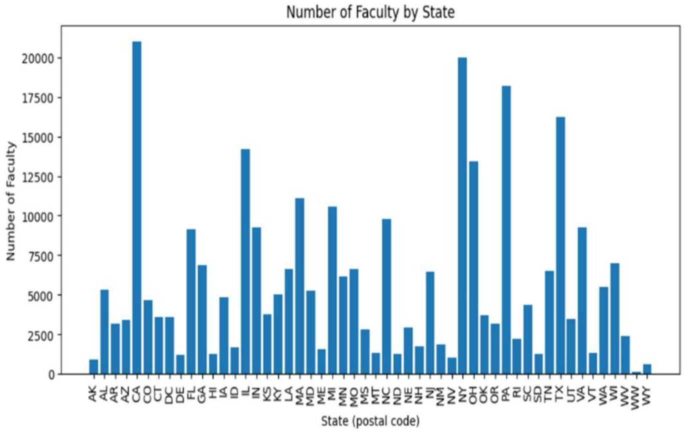

Fig 1 – Number of Faculty by State

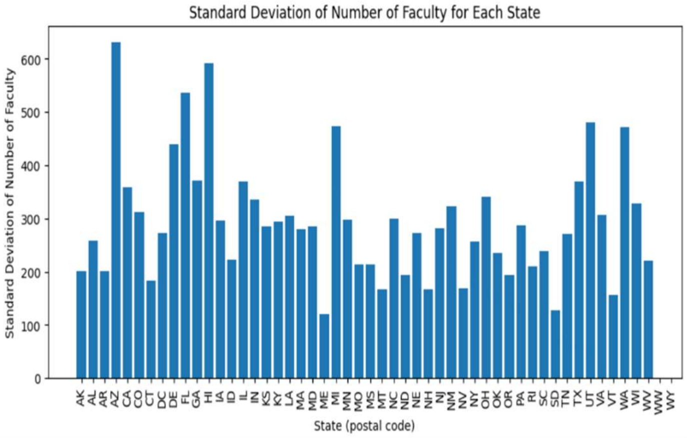

Fig 2 - Standard Deviation of Number of Faculty by each State

2. The PMF (probability mass function) and CDF (cumulative distribution function) plots are used here to visualize the distribution of average salaries across all ranks for faculty members. The PMF plot shows the probability of each possible value of average salary across all ranks. The CDF plot shows the cumulative probability of observing a value less than or equal to a given salary. Together, these plots provide a comprehensive view of the distribution of average salaries across all ranks for faculty members. The PMF plot shows the probability of observing a specific salary, while the CDF plot shows the probability of observing a salary less than or equal to a given value.

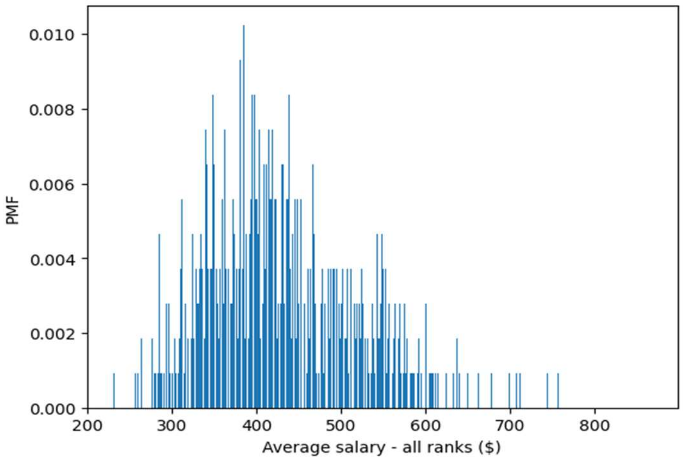

Fig 3 – PMF of Average Salary of all faculty members

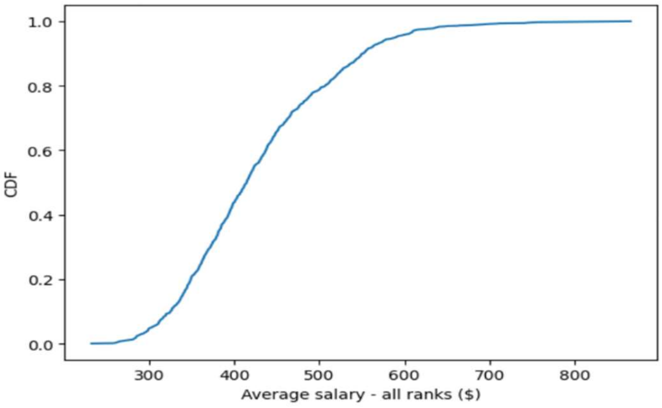

Fig 4 – CDF of Average Salary of all faculty members

3. The scatter plot shows a relationship between the average salary and average compensation of faculty members across all ranks. Each point on the plot represents a college or university, with the x-axis representing the average salary of faculty members and the y-axis representing the average compensation of faculty members. From the plot, it appears that there is a positive correlation between the average salary and average compensation of faculty members. This suggests that colleges and universities that pay their faculties more also provide higher compensation, although the relationship is not perfect and there are some outliers.

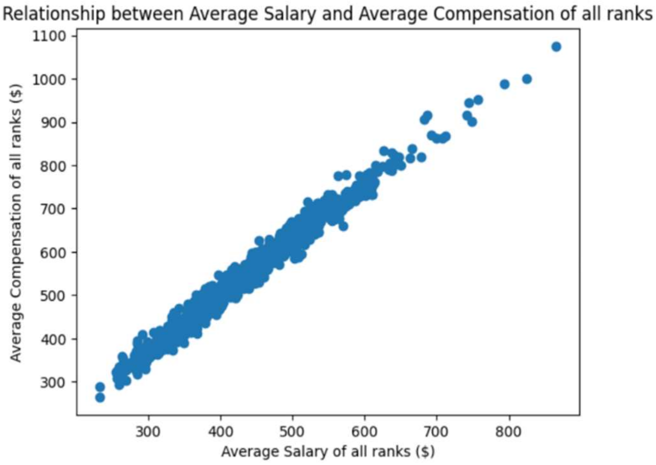

Fig 5 – Relationship between the Average Compensation of all faculty members

4. The line plot shows the average compensation of faculty members for colleges in each state of the US. The plot reveals that there are significant variations in the average compensation provided to faculty members across different states. Colleges in states such as California, Connecticut and New Jersey offer the highest average compensation, while colleges in states such as Mississippi and Kansas offer the lowest average compensation. Overall, the plot suggests that there is a relationship between the location of a college and the average compensation offered to faculty members, indicating that colleges in certain states pay their faculty members more than colleges in other states.

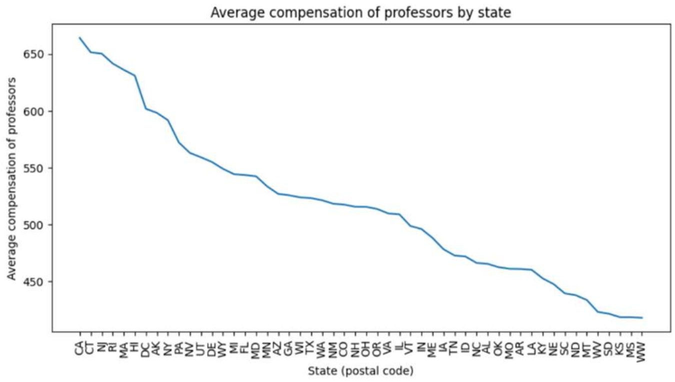

Fig 6 – Average Compensation of Professors by state

5. The bar plot shows the median number of faculty members among different types of colleges in the United States. The x-axis represents the college types (I, IIA or IIB), and the y-axis represents the median number of faculty members. The plot shows that colleges of type IA have the highest median number of faculty members, followed by colleges of type IIA, while colleges of type IIB have the lowest median number of faculty members. Therefore, there seems to be a significant difference in the median number of faculty members among different types of colleges in the United States.

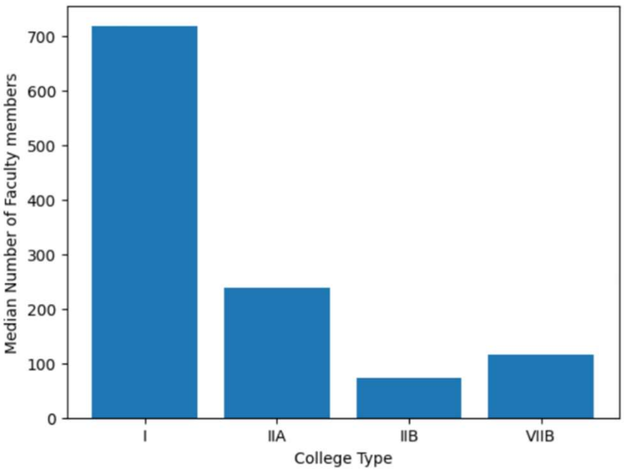

Fig 7 – Median Number of Faculty members vs College type

The bar graph shows College type VIIB which makes no sense as according to given dataset College type can be IA, IIA or IIB which means there is an error in the dataset of AAUP.

6. The stem plots show the average acceptance rates and enrollment rates for public and private colleges in the US. The first stem plot shows that the acceptance rate for public colleges (75.3%) is nearly equal to private colleges (75.6%). The second stem plot shows the mean enrollment rates for public and private colleges in the US. The enrollment rate is higher for public colleges (51.1%) compared to private colleges (around 42.3%). This difference in enrollment rates suggests that a higher proportion of accepted students enroll in public colleges compared to private colleges. However, it's important to note that there may be other factors at play that influence these rates, such as location, reputation, selectivity, etc.

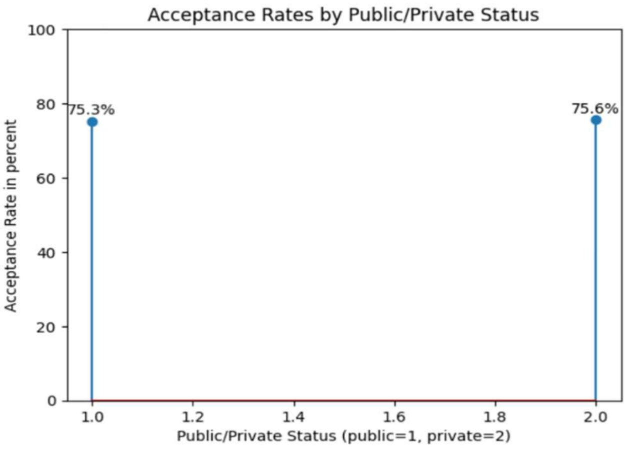

Fig 8 – Acceptance Rates by Public/Private Status

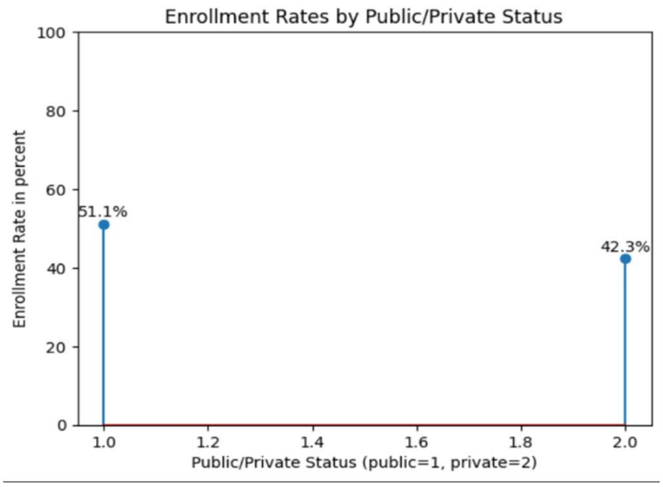

Fig 9 – Enrollment Rates by Public/Private Status

7. The pie chart shows the proportion of colleges based on the comparison of in-state and out-of-state tuition fees. The chart is divided into three sections: "In-state tuition fees > Out-of-state tuition fees," "In-state tuition fees < Out-of-state tuition fees," and "In-state tuition fees = Out-of-state tuition fees." Each section shows the percentage of colleges that fall into that category. The chart indicates that the majority of colleges (64.1%) charge the same tuition for both in-state and out-of-state students, while some colleges (31.3%) charge higher tuition fees for out-of-state students than for in-state students and a smaller proportion of colleges (4.7%) charge less for out-of-state students. Overall, the pie chart provides a clear representation of the distribution of colleges by comparison of in-state and out-of-state tuition fees in the dataset.

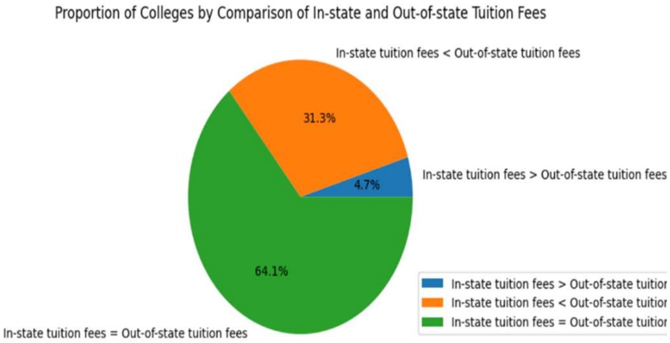

Fig 10 – Proportion of colleges by Comparison of in-state and out-of-state Tuition fees

8. The bar chart shows the top 50 colleges in the US with the highest average total expenditures by a student, which include room and board costs, additional fees, estimated book costs, and estimated personal spending, other than tuition fees. The chart displays the total expenditure of each college on the y-axis, and the college name on the x-axis. The bars represent the total expenditure of each college, and they are sorted in descending order from left to right. The chart helps to identify the colleges that spend the most on each student's needs, apart from tuition fees. University of California at San Diego, University of California at Santa Barbara and University of California at Davis are top 3 US colleges with the highest average total expenditures by a student which is over 12,000 USD.

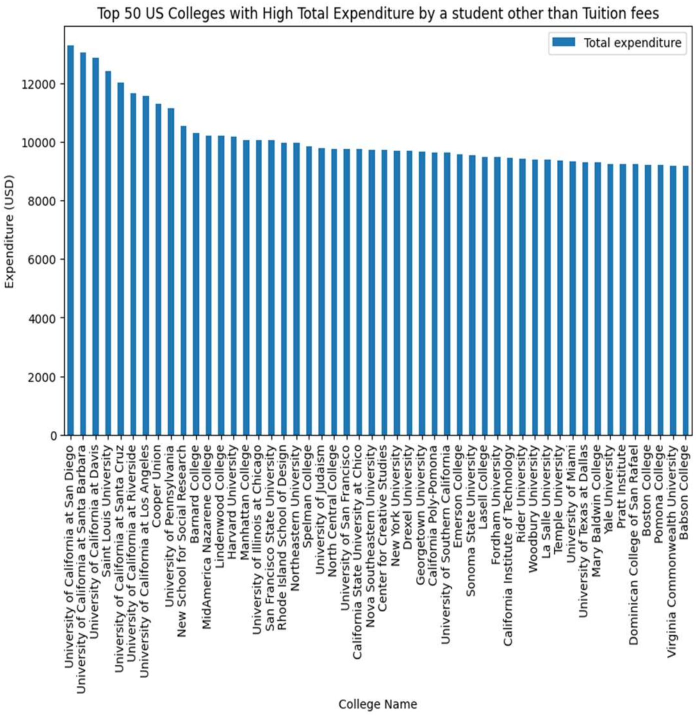

Fig 11 – Top 50 US Colleges with High Total Expenditure by a student other than Tuition fees

9. The bar chart shows the top 50 US colleges with the lowest graduation rates. The x-axis shows the names of the colleges and the y-axis shows the graduation rate as a percentage. The chart is sorted in ascending order, meaning the colleges with the lowest graduation rates are shown at the top of the chart. The chart is helpful in comparing the graduation rates of different colleges and identifying the colleges with the lowest rates. University of Houston - Downtown and Texas Southern University have the lowest graduation rates in US which is less than or equal to 10%.

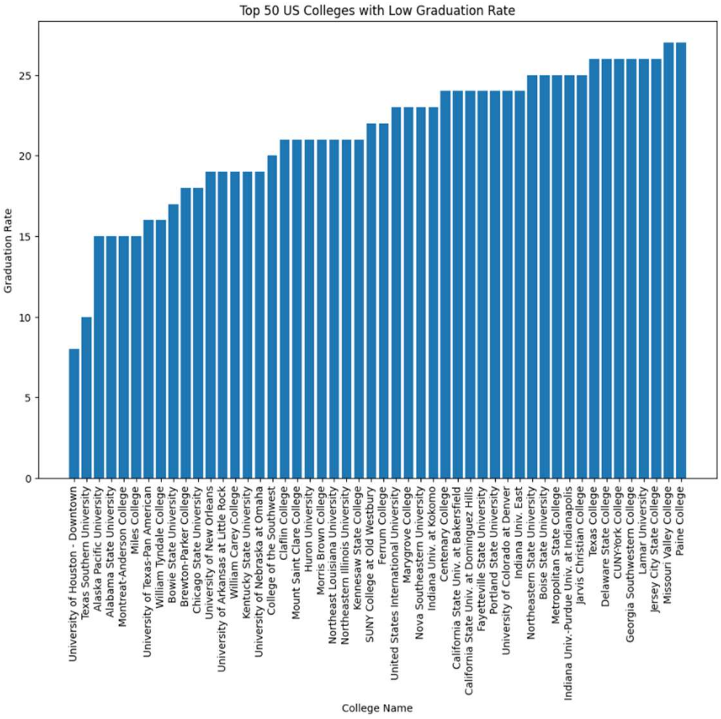

Fig 12 – Top 50 US Colleges with Low Graduation rate

10. The scatter chart shows the relationship between the acceptance rate and the average combined SAT score of the top 50 US colleges. Each dot represents a college, with the x-axis indicating the average combined SAT score and the y-axis indicating the acceptance rate. The chart shows that there is a negative correlation between the two variables, meaning that as the average SAT score increases, the acceptance rate tends to decrease. This indicates that higher-ranked colleges tend to have a more selective admissions process, accepting fewer applicants and admitting students with higher SAT scores. Conversely, lower-ranked colleges tend to have a higher acceptance rate and admit students with lower SAT scores. However, it's important to note that this relationship may not hold true for all colleges and there may be other factors that impact acceptance rates besides SAT scores like ACT scores, etc.

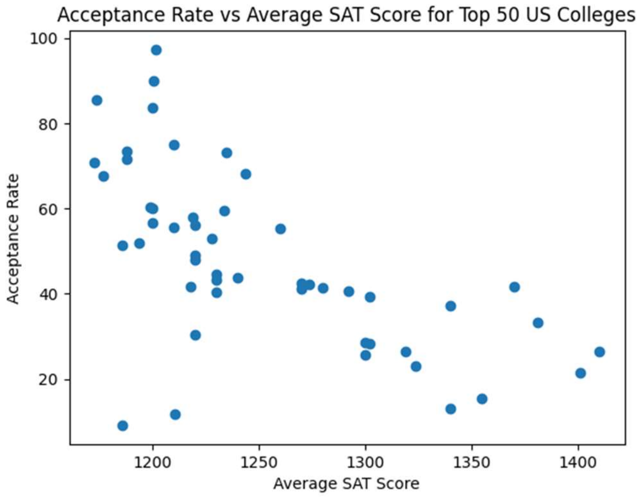

Fig 13 – Acceptance rate vs Average SAT score for Top 50 US Colleges

V. Summary of the Observations

Based on the above answers and graphs, conclusions can be drawn about the state of higher education in the US that there is significant variation in the number of faculty across different states in the US. There is a positive correlation between the average salary and average compensation of faculty members. There is significant variation in the average compensation provided to faculty members across different states. There is a significant difference in the median number of faculty members among different types of colleges in the US. Public colleges have a higher enrollment rate compared to private colleges. The majority of colleges charge the same tuition for both in-state and out-of-state students. Some colleges spend significantly more on each student's needs, apart from tuition fees. Overall, these observations suggest that there are significant differences in the state of higher education in the US, depending on factors such as state location, type of college, and public versus private funding. However, it is important to note that these conclusions are based on the limited data provided in the observations, and further analysis may be needed to draw more comprehensive and precise conclusions.

VI. Unanswerable Questions

Question 1: Can we calculate actual acceptance rate from given datasets?

Question 2: What are the other faculty members apart from full, assistant and associate professors?

VII. References

1) 'Pandas documentation,' accessed 30 Mar, 2023, https://pandas.pydata.org/docs/

2) 'Matplotlib documentation,' accessed 30 Mar, 2023, https://matplotlib.org/stable/plot_types/index.html

3) 'aaup.data,' accessed 30 Mar, 2023, http://lib.stat.cmu.edu/datasets/colleges/aaup.data

4) 'GeeksforGeeks,' accessed 30 Mar, 2023, https://www.geeksforgeeks.org/

5) 'Numpy documentation,' accessed 30 Mar, 2023, https://numpy.org/doc/

6) 'usnews.data,' accessed 30 Mar, 2023, http://lib.stat.cmu.edu/datasets/colleges/usnews.data

VIII. Acknowledgements

I would like to express my sincere gratitude to Professor Shanmuganathan Raman for their valuable guidance. Their expert advice and feedback helped me in developing my skills and improving the quality of this report.