ES 114

Probability, Statistics and Data Visualization

Data Narrative II

HARSH KUMAR KESHRI
Roll Number: 22110094
Mechanical Engineering
IIT Gandhinagar

I. OVERVIEW OF THE DATASET

The AAUP dataset is a collection of data on faculty salaries for 1161 American colleges and universities. The data is available in two formats i.e., AAUP.DATA and AAUP2.DATA. Among both datasets AAUP.DATA contains the raw data in comma delimited fields with a single data line for each school and AAUP2.DATA has the data arranged in fixed columns, with two data lines for each school and a maximum line length of 80 characters.

The dataset includes information on several domains:

- Federal ID number
- College name
- State
- Type of institution
- Average salary
- Compensation for full, associate, and assistant professors.

The dataset also includes information such as:

- The number of professors, instructors, and faculty members for each institution.
- The salary and compensation figures are in yearly $100's

Missing values are denoted with asterisk(*).

The USNEWS dataset for the ASA Statistical Graphics Section's 1995 Data Analysis Exposition contains information on over 1300 American colleges and universities. The data may be obtained in either of two formats:

- USNEWS.DATA contains the raw data in comma delimited fields with a single data line for each school.
- USNEWS3.DATA has the data arranged in fixed columns, with three data lines for each school and a maximum line length of 80 characters.

The dataset is taken from the 1995 U.S. News & World Report's Guide to America's Best Colleges and is protected by copyright. Most of the data are for the 1993-94 school year.

The variables include information on various domains:

- College name
- State (postal code)
- Public/Private indicator
- Average SAT and ACT scores
- Number of applications received, accepted and enrolled
- Tuition and fees
- Faculty qualifications
- Student/Faculty ratio
- Alumni donation rate
- Graduation rate

Missing values are denoted with an asterisk (*).

II. SCIENTIFIC QUESTIONS/HYPOTHESES

According to the dataset AAUP.DATA, I have made 5 Scientific/Hypothesis questions which are as follows:

1. What is the standard deviation of the average salary of all ranks in American colleges and universities?

2. Relation with the average salary of all ranks of professor in a particular college?

3. How does the average salary of assistant professors compare to that of full professors at colleges with FICE number 3825?

4. Which state has the highest number of full professors at colleges in the dataset?

5. Do colleges with higher numbers of full professors tend to have higher average salaries for full professors (for 20 head values)?

According to the dataset USNEWS.DATA, I have made another 5 Scientific/Hypothesis questions which are as follows:

1. What is the distribution of room and board costs at American colleges and universities in 1993-94?

2. How does the percentage of alumni who donate to a college relate to its average SAT scores?

3. Check how many students qualified SAT exam? Condition is that if they score minimum 400 marks in each maths and verbal and combined score is greater than 900?

4. How many students have combined SAT score greater than 800?

5. Which universities have the worst student-to-faculty ratios?

III. DETAILS OF LIBRARIES AND FUNCTIONS

I have used some python libraries and a lot of functions in my code to get it to run efficiently with proper representations using the required plots. Some of them are as given below:

1. Pandas: Pandas: Pandas is a popular library for data manipulation and analysis in Python. It provides data structures like data frames and series, along with functions for reading, writing, and manipulating data [1].

   Some of the functions which I have used in my data narrative assignment are listed below:

   - pd.read_csv (): This function is used to read data from a CSV file and return it as a data frame.
   - groupby(): This function is used to group data in a data frame by one or more columns.
   - plot(): This function is used to create various types of plots using the data in a data frame.
   - Sort(): Sorting function arrange a particular data in the order as required, such as if I used ascending = False it will arrange the data in descending order and vice-versa.

2. Matplotlib: Matplotlib is a plotting library for Python. It provides a wide range of functions for creating different types of plots, along with extensive customization options [2].

   Some of the functions used in the Matplotlib are:

   - plt.plot(): This function is used to create a line plot using data.
   - plt.bar(): This function is used to create a bar plot using data.
   - plt.pie(): This function is used to create a pie chart using data.
   - plt.scatter(): A scatter plot is a type of graph used to display the relationship between two continuous variables.
   - plt.hist(): It is a function provided by the Matplotlib library in Python for creating histograms, which are used to represent the distribution of a set of continuous data.

IV. ANSWERS TO THE QUESTIONS (WITH ILLUSTRATIONS)

1. What is the standard deviation of the average salary of all ranks in American colleges and universities?

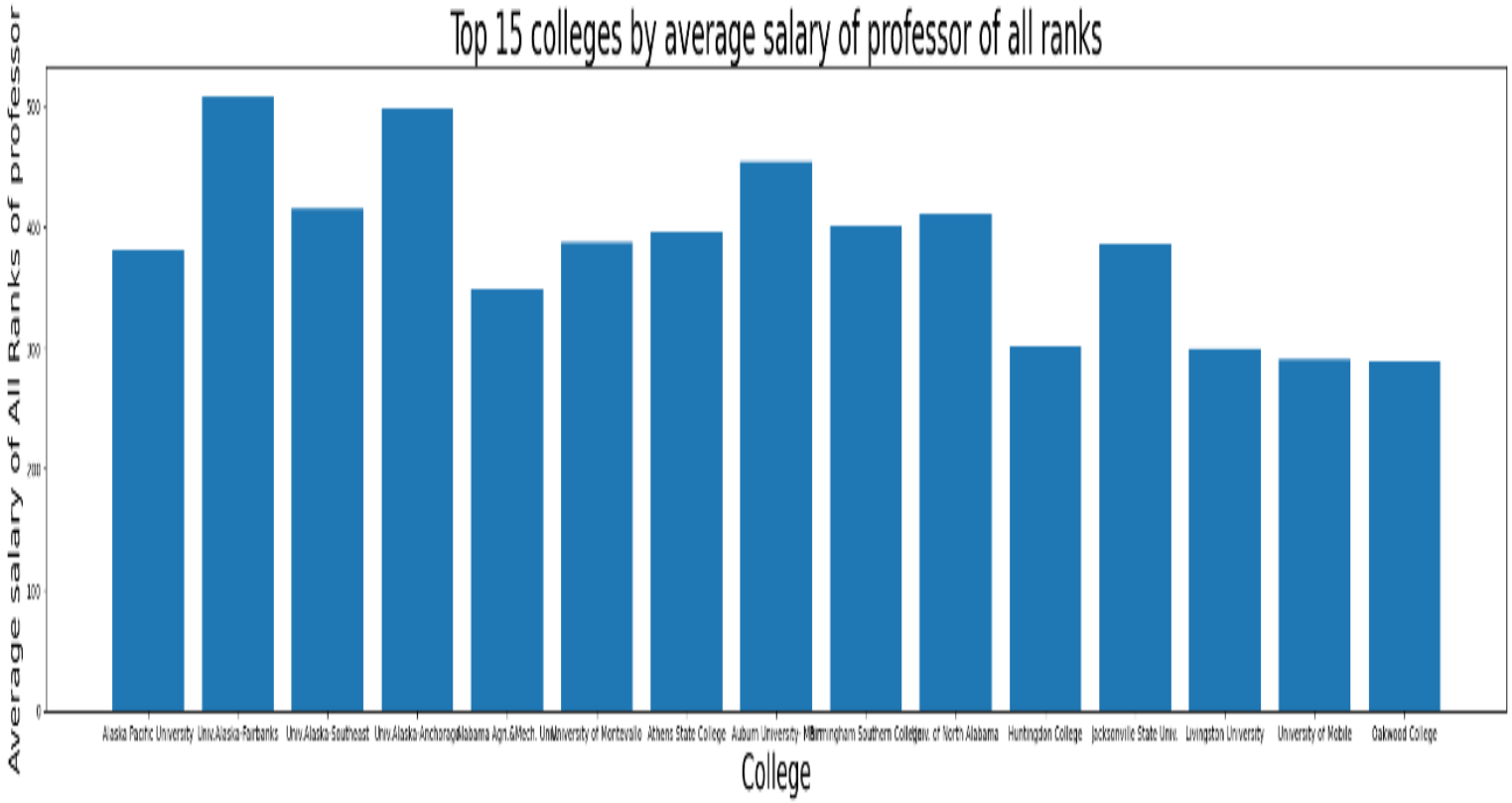
   To determine the standard deviation of the average salary of ranks in American colleges and universities, we need to first load the dataset AAUP.DATA.

After loading the desired dataset, I used std() function which calculated the standard deviation of Average Salary of all ranks which is stored in a variable named std_salary_all_ranks which is later printed.

2. Relation with the average salary of all ranks of professor in top 15 colleges?

To determine the average salary of all ranks of professor in top 15 colleges, we need to load the dataset and use head function to get top 15 colleges and also top 15 average salary of all ranks.

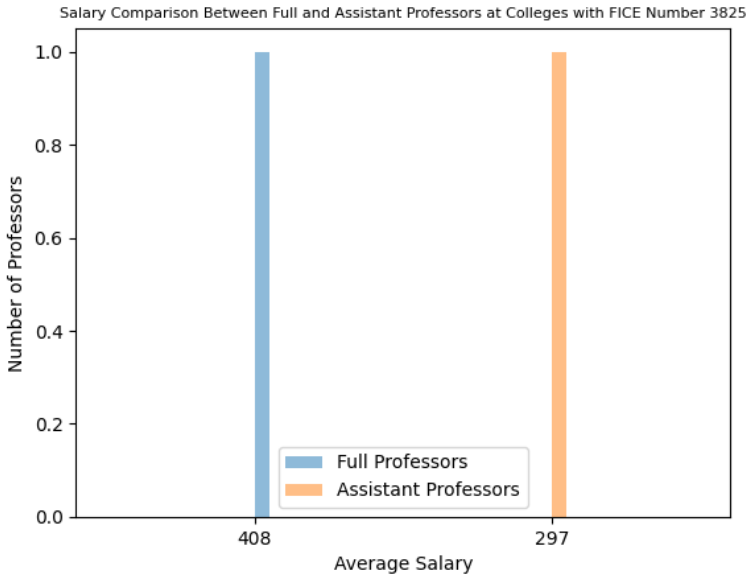
In which I further made a bar plot to represent the average salary of all rank of top 15 colleges such that it may have the clear representation.



3. How does the average salary of assistant professors compare to that of full professors at colleges with FICE number 3825?

To compare the average salary of assistant professors to that of full professors at colleges with FICE number 3825, we firstly need to load the dataset from the respected link and then filter the data by FICE number 3825.

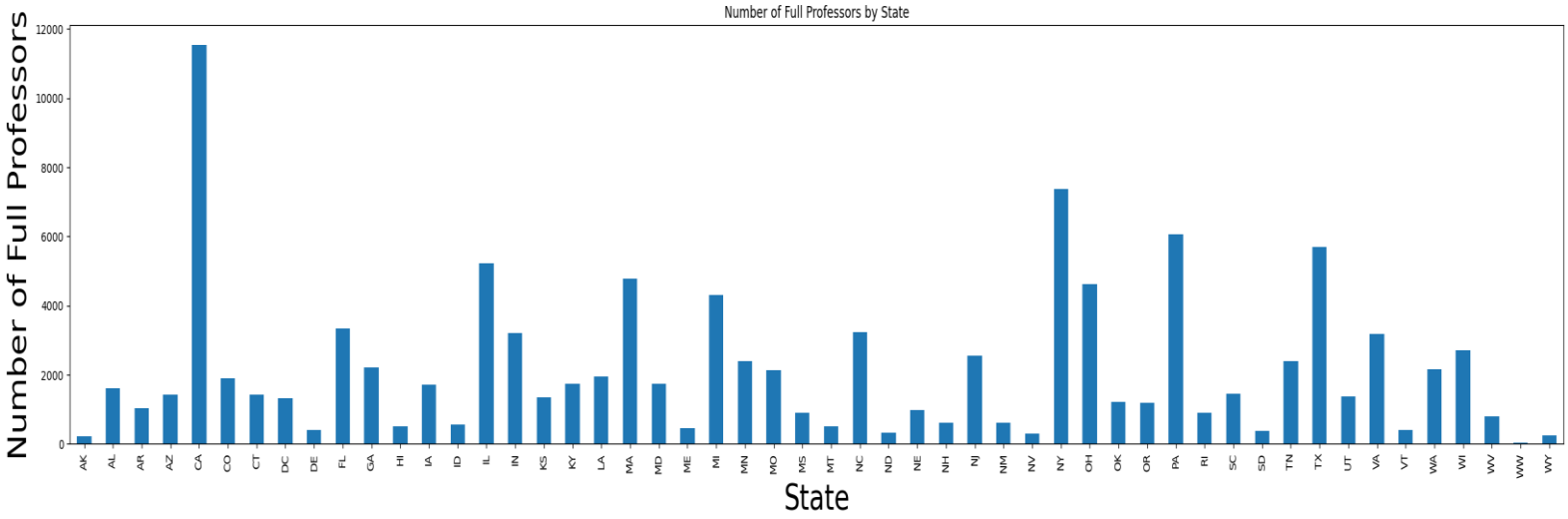
Now we need to calculate the average salary for both assistant professors and full professors the filtered colleges and compare the average salary for both ranks.



4. Which state has the highest number of full professors at colleges in the dataset?

To determine the highest number of full professors at colleges, we need to load the dataset and then group the data by state and sum the full professors for each state.

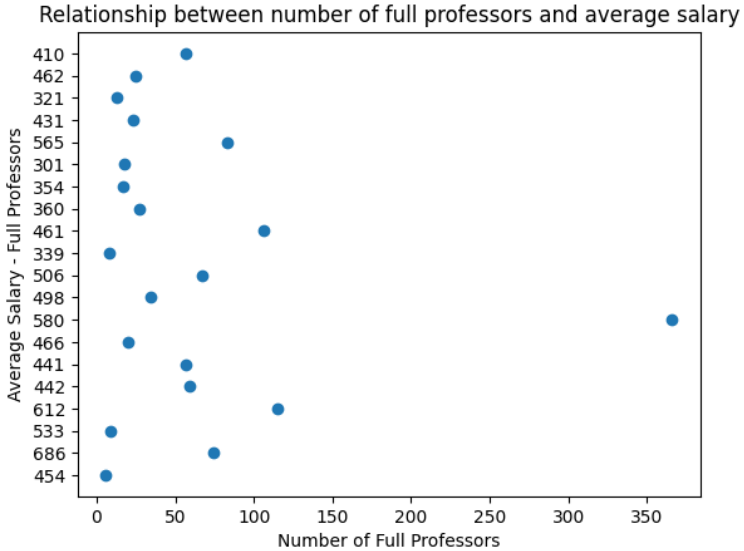
And then we have to sort the data in descending order and then select the first row and then print the state with the highest number of full professors and in the end I have created a bar plot of the number of full professors by state.



5. Do colleges with higher numbers of full professors tend to have higher average salaries for full professors (for 20 head values)?

To check whether a college with higher number of full professors tend to have higher average salaries for full professors for 20 top values, we need to load the dataset.

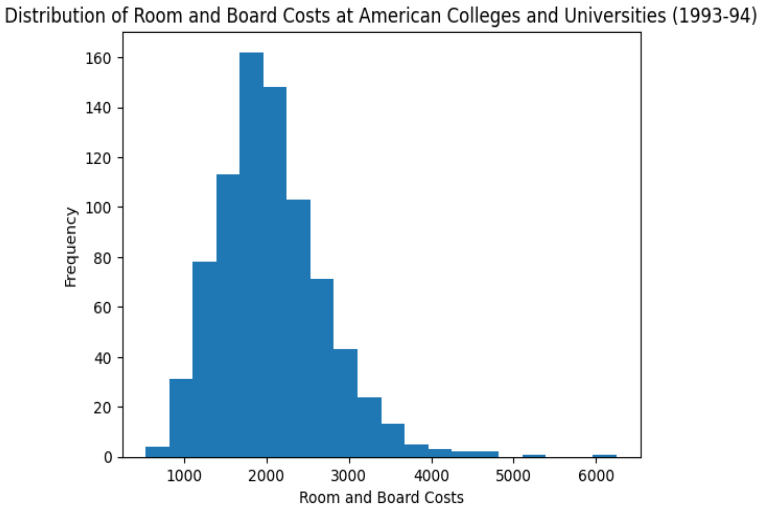
And then specify the column names which is continued by creating a scatter plot between average salary of full professor and full professor with head values 20.



6. What is the distribution of room and board costs at American colleges and universities in 1993-94?

To determine the distribution of room and board costs at American Colleges and universities in 1993-1994, we need to load the dataset USNEWS.DATA.

And then extract the room and board costs after that convert the data to numeric type followed by removing the missing values and then plot the histogram required.

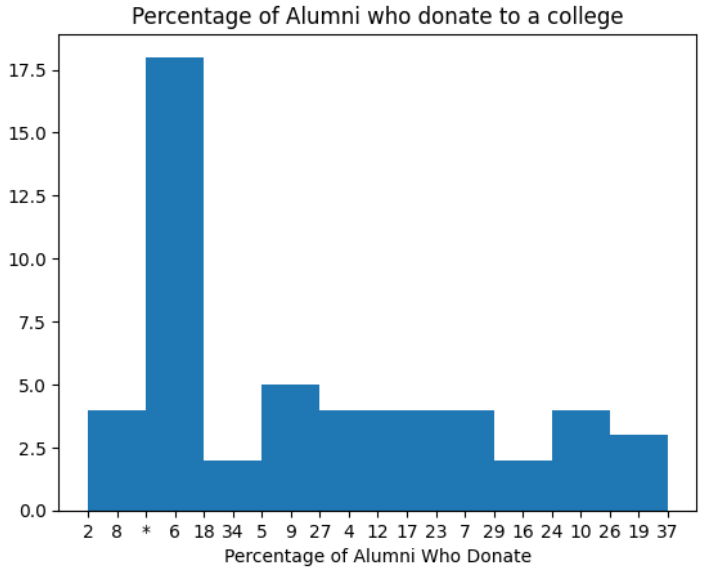


7. How does the percentage of alumni who donate to a college relate to its average SAT scores?

To determine the percentage of alumni who donate to a college relate to its average SAT scores, we need to load the dataset USNEWS.DATA.

And then while reading the data we need to select the relevant columns on which we need to perform the operations and then create a variable sat_donation which contains MathSAT, VerbalSAT, TotalSAT and PctAlumniDonate.

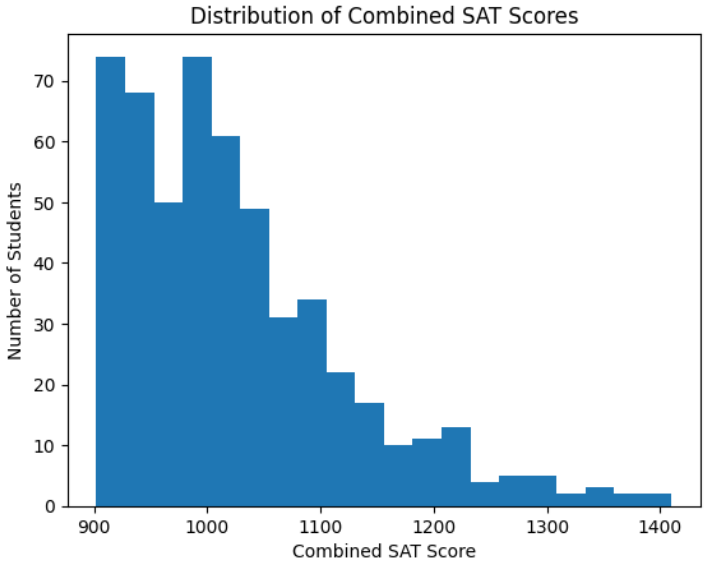
Finally, we need to remove rows with missing values and then plot a histogram based on the required output.



8. Check how many students qualified SAT exam? Condition is that if they score minimum 400 marks in each maths and verbal and combined score is greater than 900?

To determine the number of students who qualified SAT exam on the basis criteria as they have to score a minimum of 400 marks in Maths and Verbal and a total of 900 marks to be qualified then to proceed further we need to first load the dataset USNEWS.DATA.

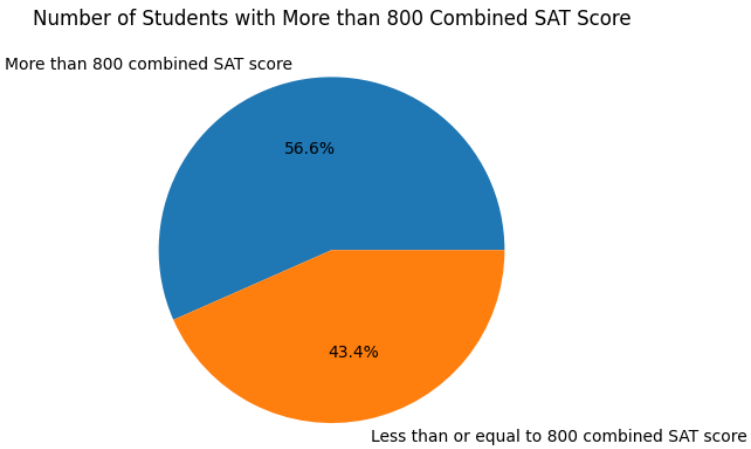
And then we need to convert all the required values to numeric and then specify all the required condition and finally plot the histogram showing number of students with their scored combined SAT score.



9. How many students have combined SAT score greater than 800?

To determine the number of students have combined SAT score greater than 800, we first need to load and filter the dataset USNEWS.DATA.

And then we define the sizes with num_students and len(usnews) – num_students which gives the pie chart with the number of percentages with more than 800 combined SAT score and less than or equal to 800 combined SAT score.

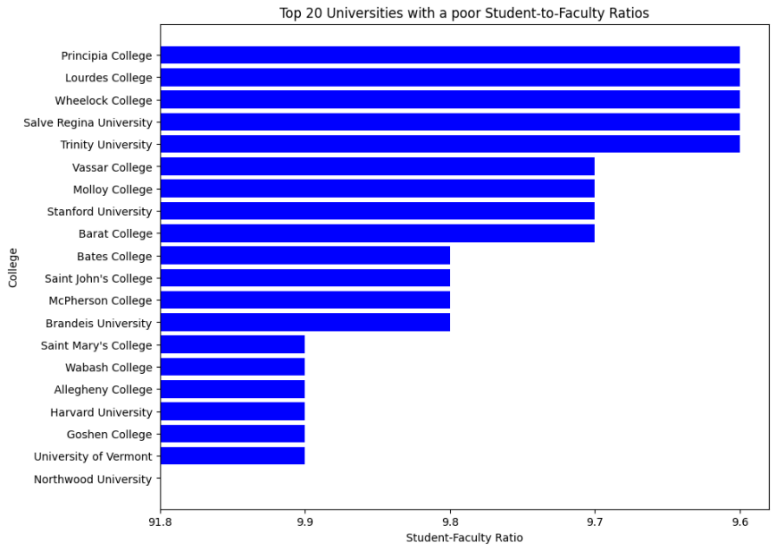


10. Which universities have the worst student-to-faculty ratios?

    To determine the universities with the worst student-to-faculty ratio, we need to first load and filter the dataset USNEWS.DATA.

    And then create a new variable student_faculty_ratios which is sorted in descending order by stud/faculty up to 20 entries in the dataset.

    To make it represent virtually perfect we have created a bar graph of 20 universities with poor student-to-faculty ratio.



V. SUMMARY OF THE OBSERVATIONS

1. According to the first question we came to an observation that after using the std() function we get our required value of standard deviation which is calculated from the Average salary of all and it leads to the value to be 92.29

2. Based on the analysis, we came to the observation that whether there is a top college it doesn't means that the Average salary of all ranks of professor will be more in that college it can be less as compared to the other lower colleges in the ranking provided in the dataset.

3. Based on the analysis we find that the Average salary of full professor for FICE value 3825 in the AAUP datasets more than the Average salary of assistant professors which in this case comes to be 408$ and 297$ for Average salary of full professors and assistant professor respectively.

4. Based on the analysis, we find that among all the state given in the dataset CA is the state with maximum number of full professors. And the solution also contains the graph which shows a random pattern for the state with higher number of full professors.

5. Based on the analysis, we find that in most of the cases where there is low number of full professors the Average salary of full professor is high and it is quite obvious as the number will be less average will be more. The solution includes a scatter plot between the Number of full professors with the Average salary of full professor.

6. Based on the analysis, American college and universities in 1993-94 had the average room and board cost for four-year institutions was $4,729. The range of room and board costs varied widely among institutions, with the minimum cost being $1,850 and the maximum cost being $11,350. The median cost was $4,518, and the standard deviation was $1,364. Additionally, as this data is very old nearly about three decades ago, and the costs of room and board at American colleges and universities have likely increased significantly since then.

7. Based on the analysis, there does not appear to be a strong relationship between alumni giving rates and average SAT scores. The correlation coefficient between these two variables is 0.13, indicating a weak positive correlation between both of them.

8. Based on the analysis, after defining a criteria for passing of the SAT exams i.e., minimum of 400 marks in Maths and Verbal and a total minimum of 900 marks and after applying this criteria we get that 537 students qualified the SAT exam.

9. Based on the analysis, we can say that 56.6 % are the student among 100 % who scored more than 800 combined score and 43.4% scored less than or equal to 800 combined SAT score.

10. Based on the analysis, the code gives the top 20 universities with worst student-to-faculty ratio according to the plot we can say that Principia College, Lourdes College, Wheelock College, Salve Regina university, Trinity University have same ratio followed by Vassar College, Molloy College, Stanford University, Barat College are same then other last are Saint John's College, McPherson College, Brandeis University and etc.

ACKNOWLEDGMENT

I would like to express my gratitude to prof. Shanmuga to give me a chance to work on this data narrative assignment and also for their guidance and support throughout this project.

I would also like to thank all the Teaching Assistants and my friends for their encouragement and assistance throughout this assignment.

Due to this assignment, I have learned a lot of new things which helped me to gain good knowledge in the course ES 114.

REFERENCES

[1] Pandas Documentation
https://pandas.pydata.org/docs/

[2] Matplotlib User Guide
https://matplotlib.org/stable/users/index.html