ES114 Probability & Statistic
Data Narrative-2

Archit Dhakar 22110031

***OVERVIEW*— Given dataset consist information regrading different colleges of US. It has the information such that college name, its type, number of faculty , SAT and ACT exam score etc. We can use different factors to suggest some concluding questions from the data set. Using different libraries such as Numpy and Pandas we can analyse the data and can plot some graphs to support our questions and hypothesis statements. Hence it is generally the common dataset for college.**

***Index Terms*— pd: pandas , np: numpy**
matplotlib.pyplot: plt

I. Scientific Questions/ Hypothesis

1. Tier one(Type I) institutions provide higher salaries to its faculty. Use the given data set to demonstrate this point. Also show that high salary is one of the factor for tier one.
2. US businessmen wants to open new a company around best institutes of country(like Gift city near IITGN). Suggest the suitable locations with the help of plot so that he can make profit in his business. Explain why it's suitable?
3. For the randomly selected college from the data set find the division of its faculty and represent with the help of graph. Repeat this activity and state the fact that associate professors are maximum in number in general?
4. A professor decided to join a college on the basis of its average compensation for full professor criteria. Show that it's not good criteria to join with the help of suitable figure.
5. For living in expensive state/location average compensation is given more to professors. On the basis of this criteria find the expensive states/location in the US.
6. Using appropriate plots show that students prefer Public colleges more than private colleges?
7. Show that Math is the scoring subject in SAT exam as as compare to Verbal test. Justify the statement?
8. Draw a Probability density function(PDF) graph to show that a randomly selected college will be having graduation rate atleast 75%.
9. Make a preference list of top 20 colleges for the students who appeared in the SAT and ACT exam on the basis of average SAT and ACT score of different colleges which can be considered as their cutoff score.
10. Using appropriate plot justify the statement that expensiveness of college is directly proportional to its graduation rate. Verify above statement and show that for a less expensive college graduation rate is low.
11. Plot the student/faculty ratio distribution and show that for low student/faculty ratio graduation rate is more. Explain the above statement?

II. Details of Libraries & Functions

Libraries and functions used to visualize the data are:

1. NumPy library: To visualize data more easily
2. Seaborn library: To plot probability density function
3. dropna function to drop '*' values from the dataset
4. Matplotlib library to plot the graphs using data
5. Pandas Library to create the DataFrame from the given csv files to plot the pie charts, bar graphs, scatter plot.
6. Scipy library
7. read_csv function to read the csv files in Python.
8. .head() and .tail() to read some important values of data
9. sorted function to sort the columns of dataframe.

III. Answer of Questions

1. From the line chart plot we can observe that colleges which provides maximum average salary to their professors are of type I. Also colleges which provides minimum average salary to their professors are of type IIA.

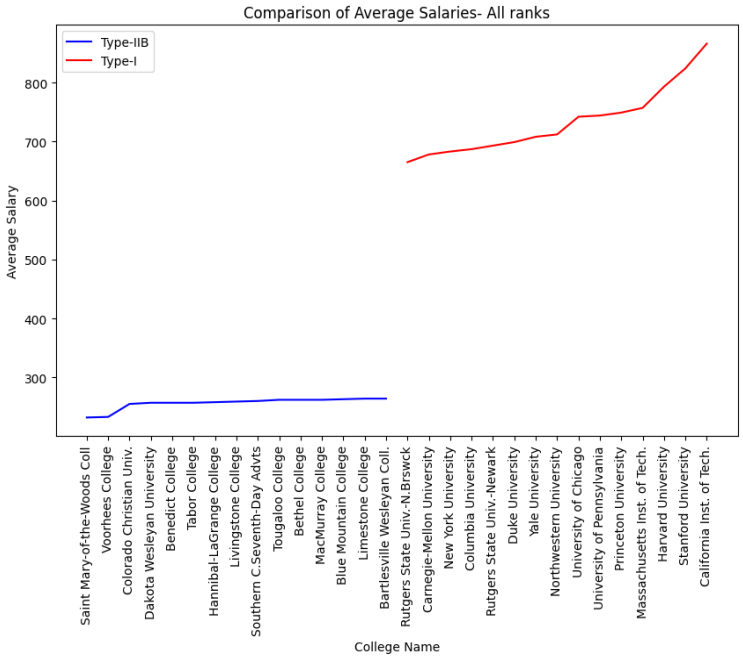

Fig1. Line chart to compare average salaries(100$)- All ranks

From the graph we can see that blue line represent type IIB colleges which are considered as third tier colleges. Red line represents type I colleges which are considered as tier one colleges. Hence from this observation we can conclude that if some college wants to improve their ranking then they have to hier well experienced professors. And also as the quality increases college have to pay more salary. Also tier one colleges are famed due the reason that they have well known professors and researchists.

2. As per the data Type I colleges are tier one colleges. These are colleges with well experienced professors and better in infrastructure. Tier one colleges have student with higher aptitude, so businessmen can hier them to improve his business. Also research work is one of the major factors in tier one colleges , this might help him to grow his company in research field and to get better exposure of researchers.
From the bar graph given below we can observe that:

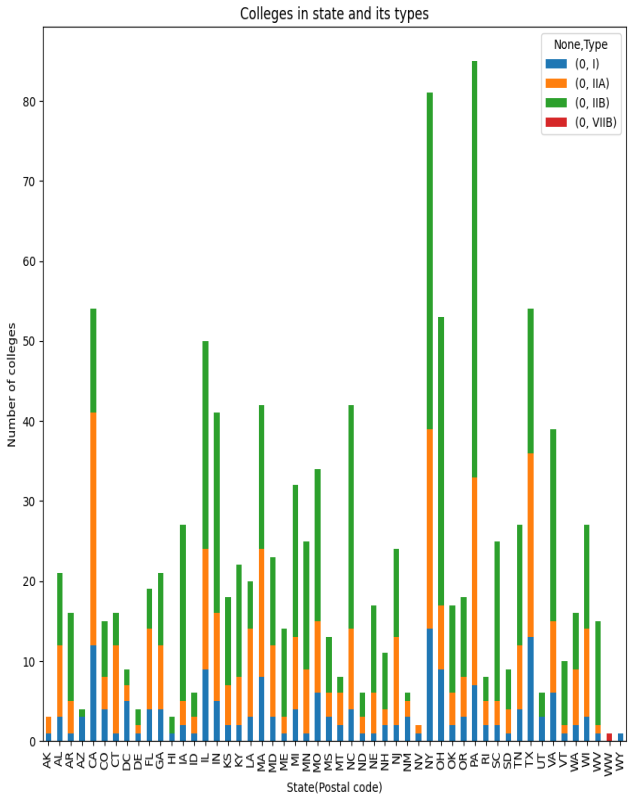

Fig2. Different types of colleges in states of US

States CA,NY,TX and MA etc. have more number of Type I colleges as well as Type IIA colleges which are famous for their research work and education quality. Hence He should open his company around these states.

3. From the given pie chart for a randomly selected college we can observe that number of Associate professors is greater then all other division. After repeating this activity we can observe that in general Associate professors are greater in number then all divisions.

The reason for this observation can be: As in any college professors are assigned on the post of Assistant professors which is generally at the age of 28. After average time period of 14-16 years they are promoted to Associate professors. Hence this is the median phase where most of the professor come in US colleges.

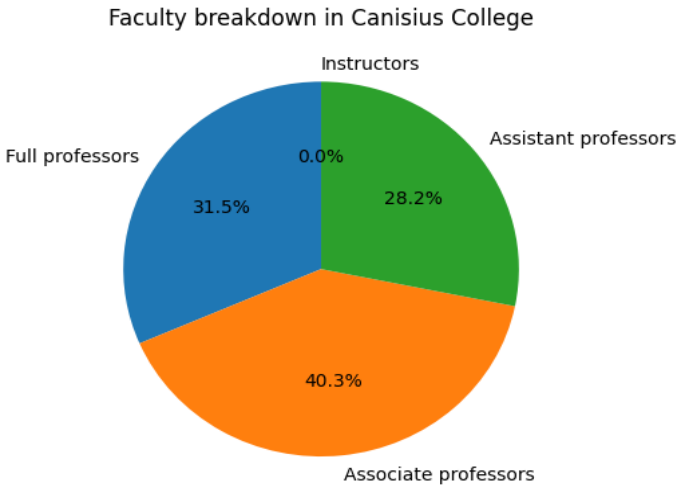

Fig3. Pie chart showing Faculty division in a US college.

4. Using data analysis we can show that average compensation salary for Full Professors paid by US colleges is not the fair criteria to join the colleges. We can observe from the line graph given below:

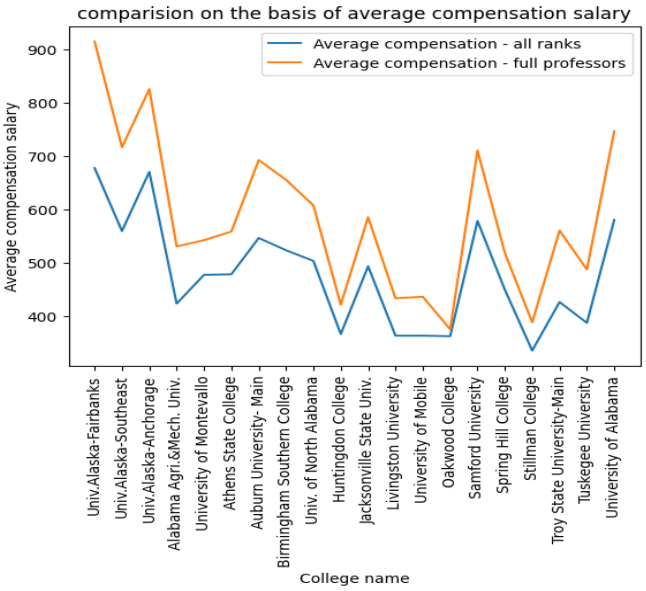

Fig4. Line plot to compare the average Compensation salary(in 100$) of all ranks and full professors.

Average compensation salary for all ranks is much lesser then that of full professors. Hence we should always consider all the parameters while seeing the average . Since, Full professors are more experienced than other categories hence they are given more compensation salary.

5. Colleges which are located in states where living is expensive pays more compensation salary to professors as compare to least expensive cities. From the bar graph given below plotted using the dataset we can conclude that states like CA,NJ,RJ and CT etc. can be considered as most expensive states of US.

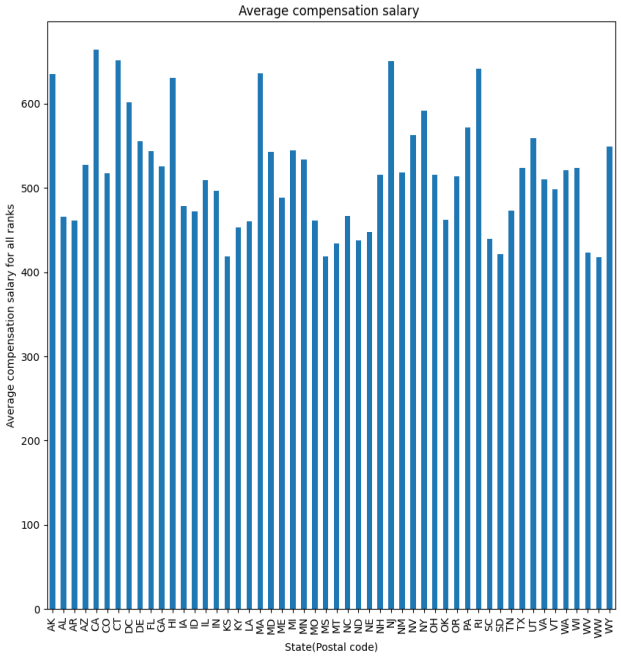

Fig5. Bar graph for Average Compansation salary(in 100$) for States in US.

From the graph we can easily conclude the most expensive states where professors are paid more amount. This data set can alse be used by a businessmen to set a company around this area for rapid growth.

6. From the pie chart we can easily observe that number of application received in Public colleges is almost 60% while those of private colleges is 40 %. Reason for these numbers is students trust public colleges as they have better job securities, high infrastructures and great research work. Here Pubilc colleges(1) are shown with Blue and Private(2) are shown with Orange.

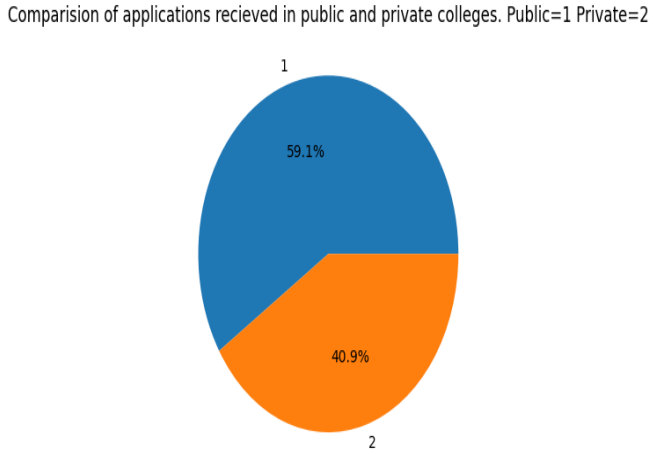

Fig6. Pie chard representing No of application received in Public and private colleges.

7. Generally Math is considered as difficult subject but in SAT exam using double bar graph I have observed that maths is more scoring subject then verbal.

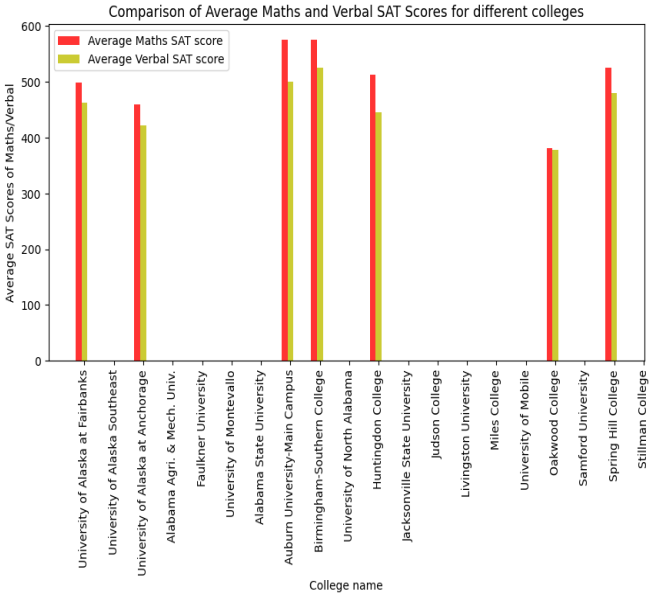

Fig7. Double bar graph to compare math and verbal score in SAT exam.

From the graph we can observe that for those colleges whose complete data was present, math score is higher than verbal score. Like in Spring Hill College Average Math score is approx. 510 while in verbal is about 470. We can also observe in every college that score difference is fairly enough. We can use this conclusion to suggest future SAT aspirants that they should focus on Maths as its easily scorable.

8. This is the most interesting plot using which we can easily guess that for a randomly selected college atleast how much % of student graduated. The plot is plotted using the concept of PDF.

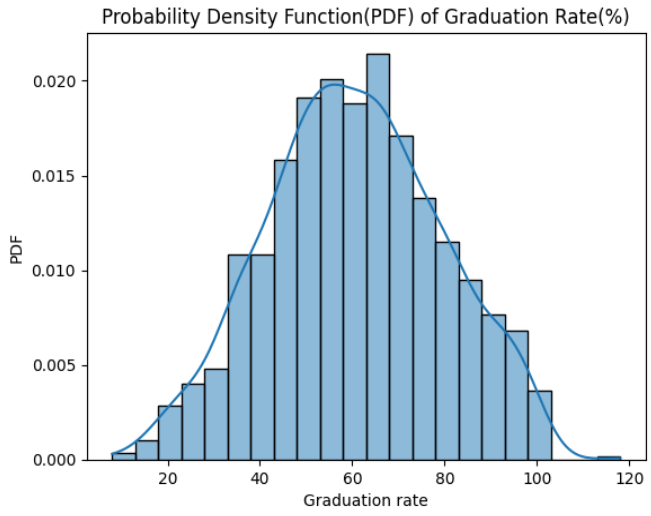

Fig8. PDF plot for graduation rate

Suppose we want to see that for a randomly selected college have atleast 75% of graduation rate. Hence we can observe the area of graph for more then 75, which is approx. 0.25 from 1. So the probability is 0.25 .

9. Using the data set I have prepared the preference list for the SAT aspirants of average score of Top 20 colleges in US.

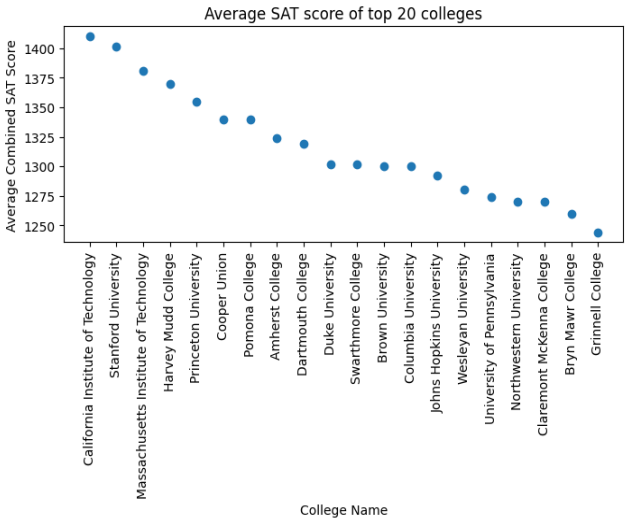

Fig9(a). Scatter plot showing Top 20 colleges SAT score.

Hence we can observe that in order to take admission in California Institute of Technology a student have to score approx. 1425 marks which is highest in all.

Also for those aspirants who give ACT exam:

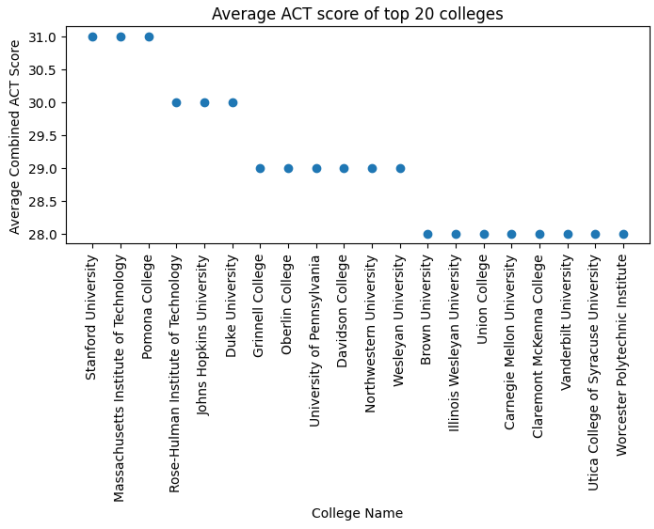

Fig9(b). Scatter plot showing Top 20 colleges ACT score.

From this plot we can conclude that a student have to score approx. 31 marks in order to get admission in top colleges such as Stanford University. We can also conclude the ranking of colleges on the basis of these average exam in the test scores.

10. After analysing the data set I concluded that colleges which are expensive means total cost paid by a student to study in that college is higher than all have better graduation rate then those whose fees cost is least.

We can analyse these facts using data plots.

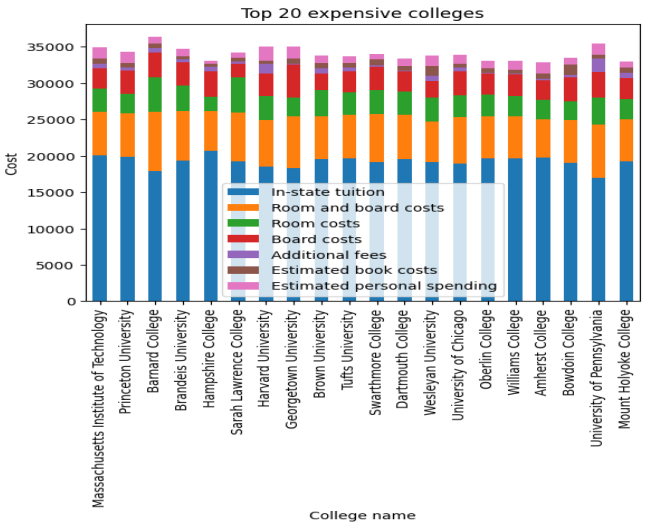

Fig10(a). Stack bar graph for Top 20 most expensive colleges.

From this graph we can observe different factors contributing to the total cost. As these colleges are located in expensive cities so Boarding cost is high. Like in Harvard University In state tuition cost is fairly high.

Now the supporting plot that these have better graduation rate.

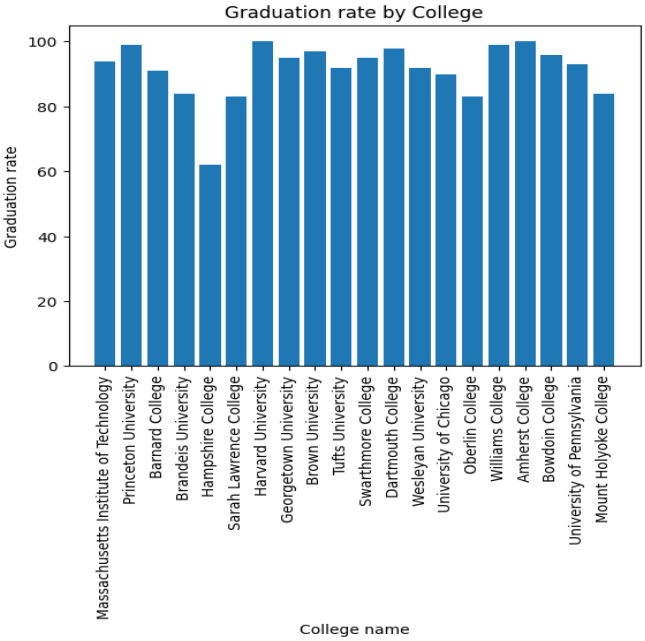

Fig10(b). Bar graph showing the graduation rate for Top 20 colleges.

From this we can conclude that almost all have 90% of graduation rate on an average.

Now the analysis for Least expensive colleges.

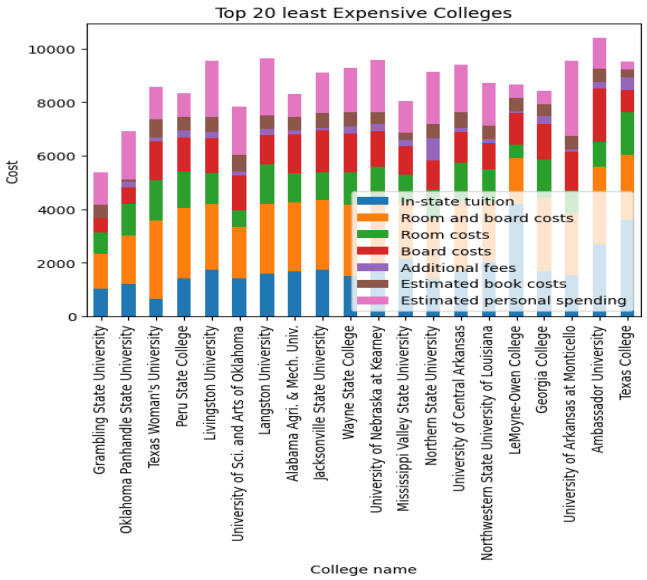

Fig10(c). Stack bar graph for Top 20 least expensive colleges.

We can easily observe that total cost to study in these colleges is almost ¼ the then that for most expensive.

Now supporting plot that these have very low graduation rates.

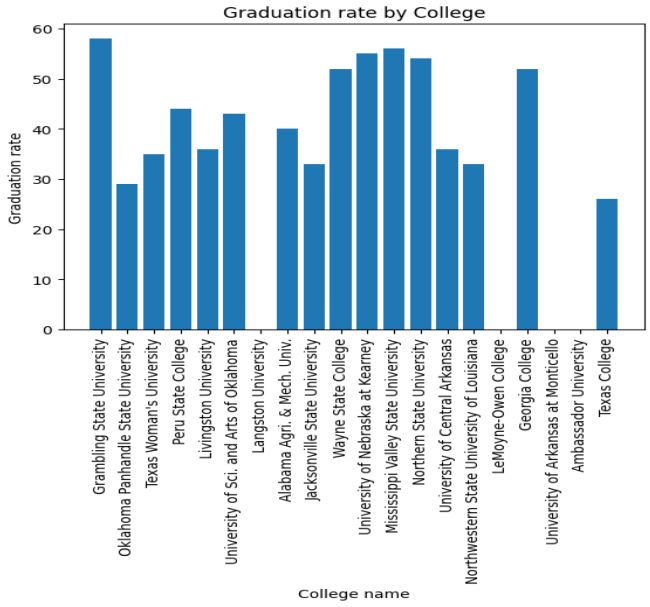

Fig10(d). Bar graph for graduation rates.

By observing data we can easily conclude that these colleges gave 50% of graduation rate on an average which is not goof for students. Hence students can consider these all concluding points before applying for any of the college after SAT and ACT exams.

11. Using the data set I have observed that colleges with low student/faculty ratio have better graduation rates. The reason for this statement is that if the colleges have more no of faculty than student can get their concepts clear and can get personal mentorship and more.

Here are the plots to support these facts.

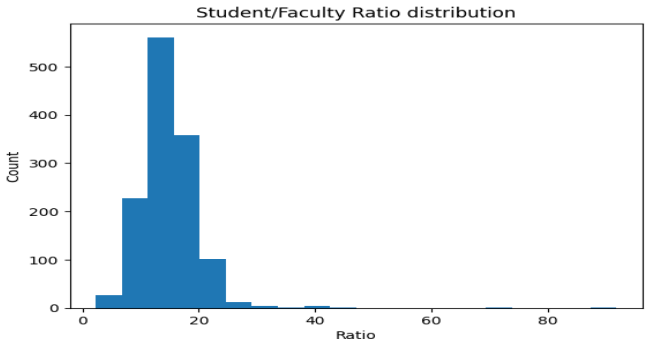

Fig11(a). Histogram for Student/faculty ration distribution

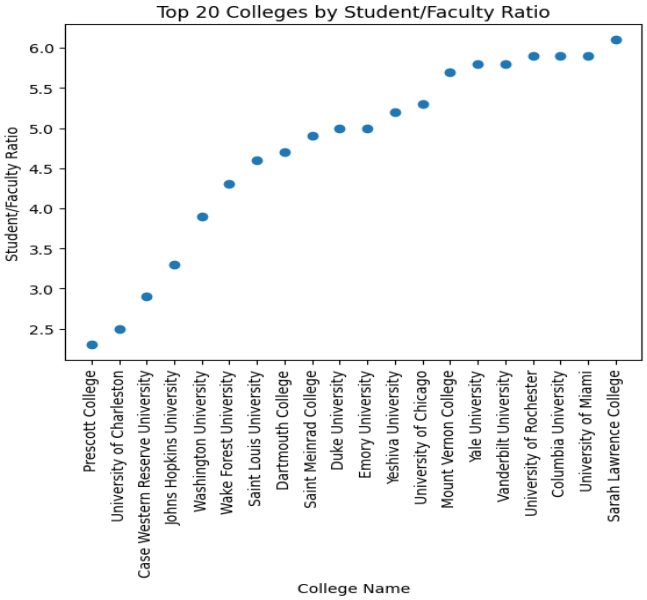

Fig11(b). Scatter plot for top 20 colleges with less ratio value

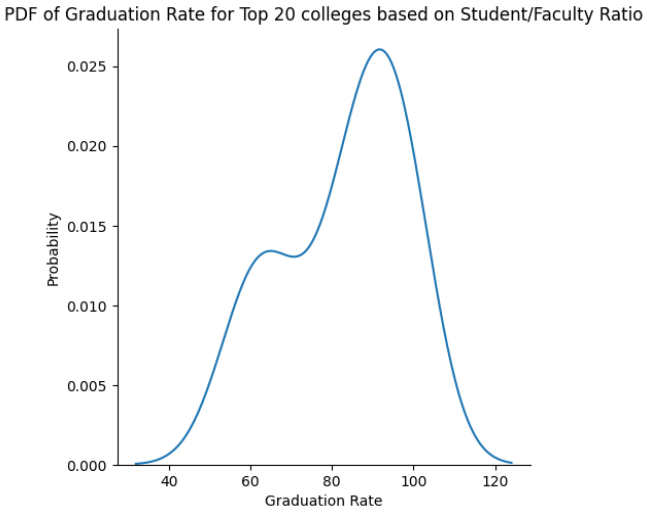

Fig11(c). PDF plot for graduation rate.

Hence we can conclude from these graphs that such colleges have higher graduation rate as area of curve is more towards higher graduation rate.

IV. Summary of Observations

In this whole dataset we have observe that how we can help the students to get better choices. Also using different PDF plots we were able to guess some of the facts that graduation rate will be higher in a good institute. Also Colleges with top ranks have more amount of fees. We also seen that Type I colleges are colleges with better infrastructure, research work and well experienced professors. Also these institute pay more money to their professors. We also observed that which states in US are more expensive and more developed. And in what location/states best colleges are located. We have also seen the faculty division in different colleges and how the average is not the good parameter to take our important decisions.
We also prepared a preference list for students to help them to choose best colleges based on their SAT and ACT score.
Hence the dataset was about to conclude important parameter regarding colleges.

V. References

[1] To create bar graph using pandas.
[2] Plot pie chart using pandas.
[3] Sort the column of DataFrame in pandas
[4] To open the csv file using dataset url
[5] To make the line plot using matplotlib
[6] To drop the '*' values from the dataset

VI. Acknowledgment

I would like to express my special thanks of gratitude to Professor Shanmuga for his guidance and support to evoke critical thinking.

Date: 29-03-2023

Name: Archit Dhakar