**CHRIST(Deemed to be University)**

BDS471L – Machine Learning

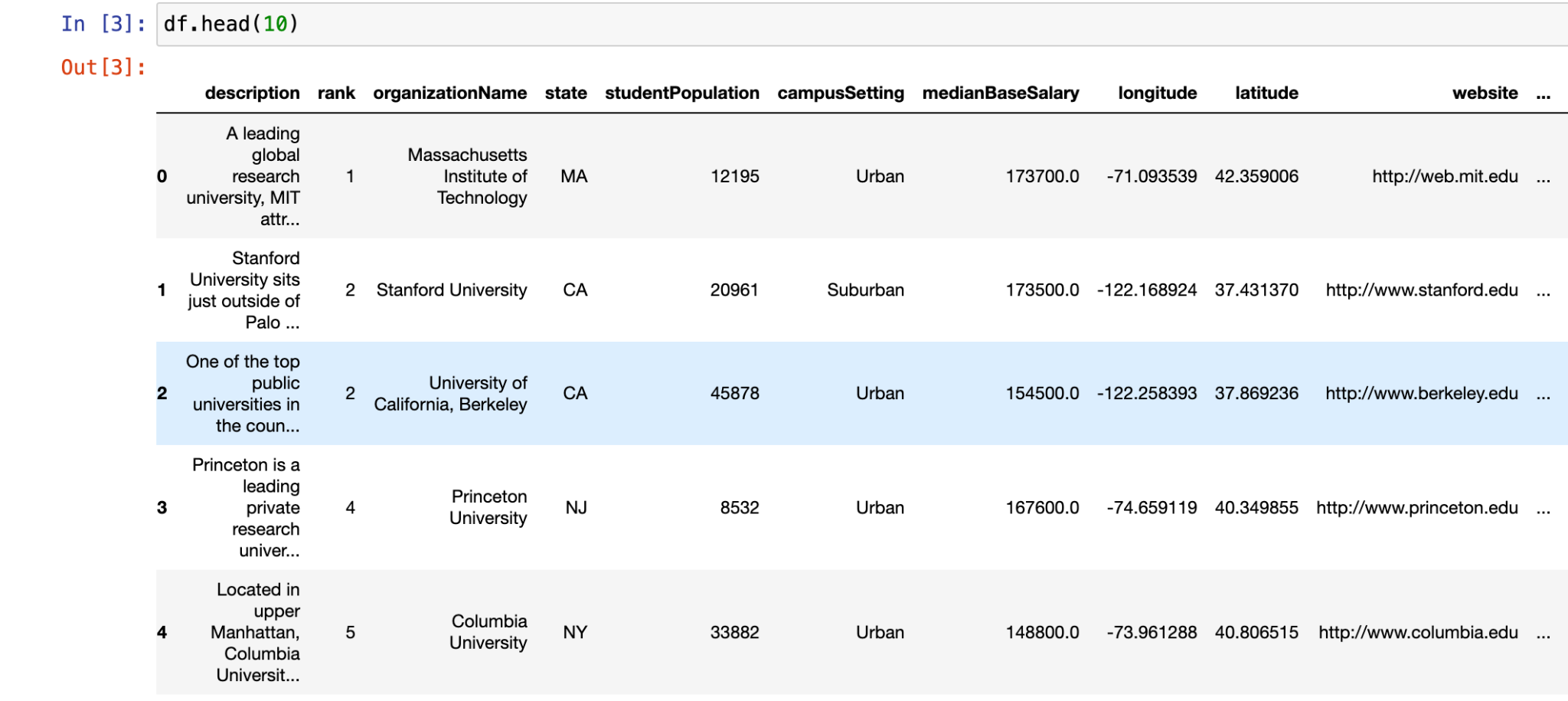
Date: 11/01/2024

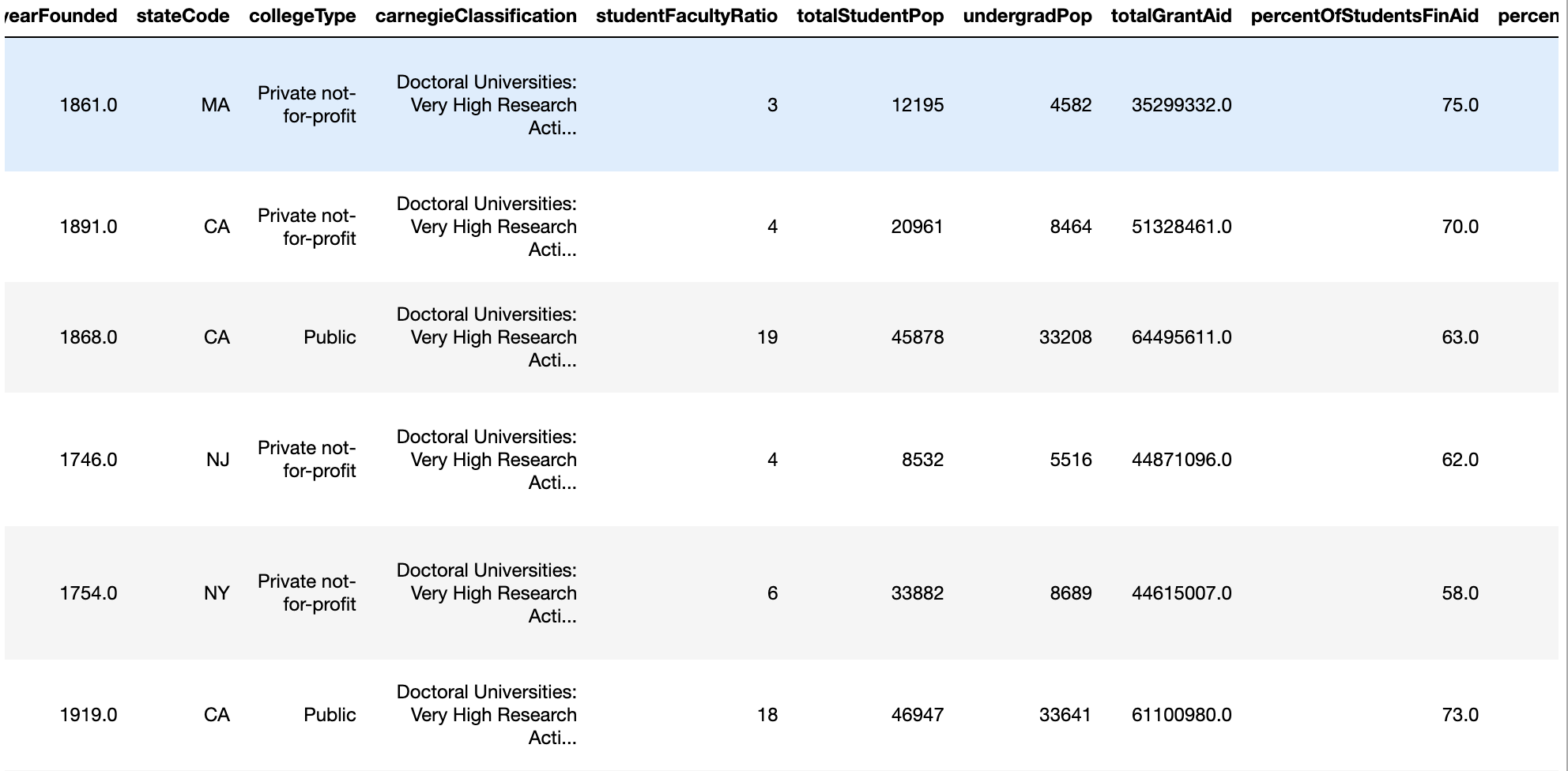
**LAB-1**

Siddharth R Bhardwaj

22112028

1. Dataset: “Top American Colleges 2022”
2. Source: <https://www.kaggle.com/datasets/kabhishm/top-american-colleges-2022>





The dataset talks about the rankings and the factors on which the universities depends on. The dataset has 496 rows and 25 columns.

Nominal data:

1. Description
2. organizationName,
3. State
4. campusSetting
5. City
6. Country
7. State
8. region
9. stateCode
10. college-type
11. carnegieClassification

Ordinal data:

1. rank

Continous data:

1. medianBaseSalary
2. longitude
3. latitude
4. totalStudentPop
5. undergradPop
6. totalGrantAid
7. percentOfStudentsFinAid
8. percentOfStudentsGrant

Discrete data:

1. yearFounded
2. studentFacultyRatio

2) What is the purpose of your analysis?

* The purpose of analysing the USA universities is to compare some key attributes of the universities like ranking, student population, financial aid and geographical location and how they affect the rankings of these universities in the USA.

1. What business problem or question are you trying to address?

* Using the information provided, the educational institutions can know which factors are more important or relevant to have a higher ranking.

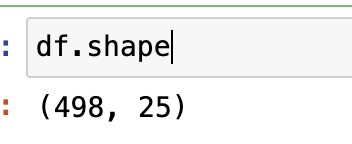
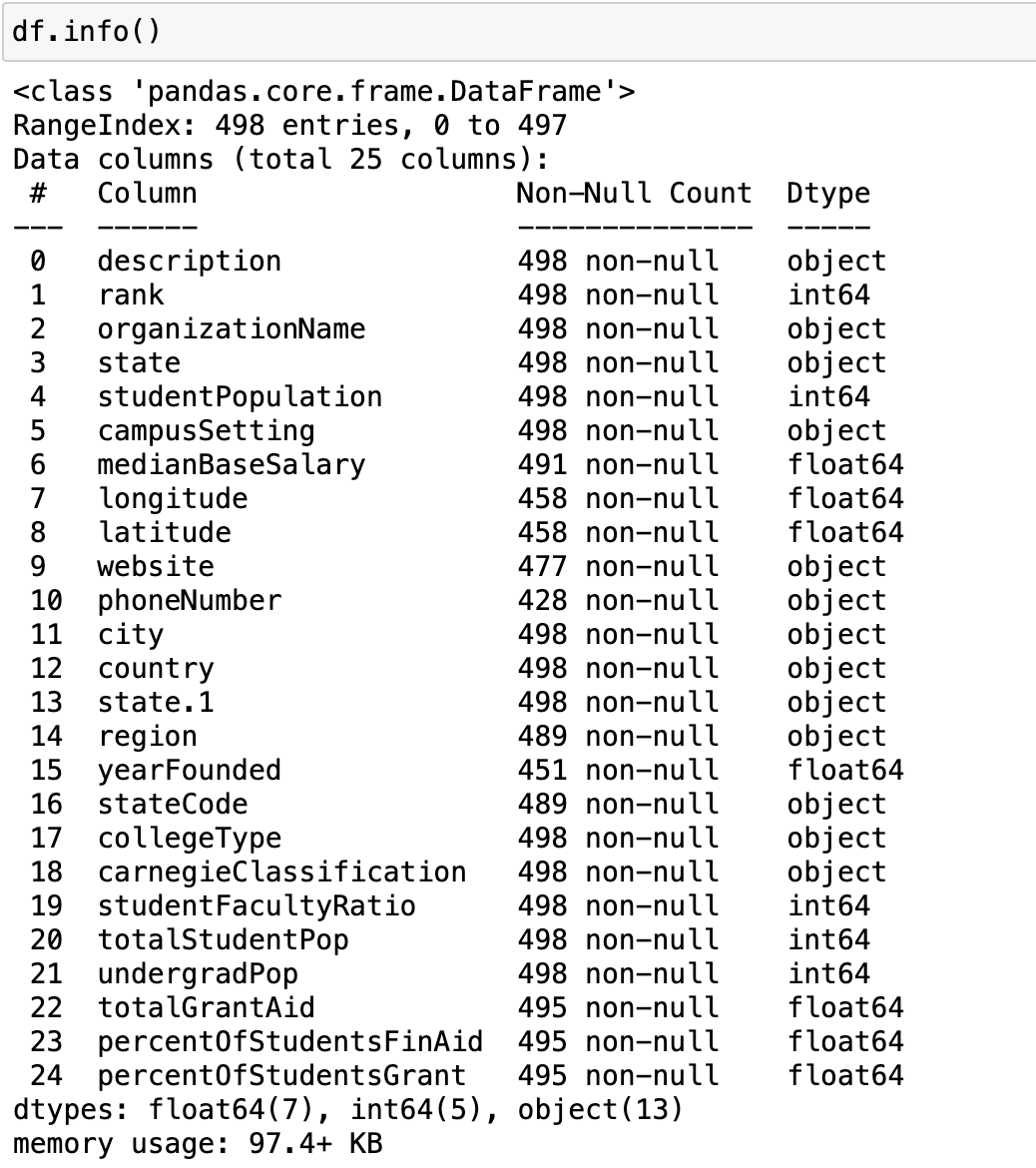
1. What are the goals or objectives of the analysis?

* We can examine the distribution of universities across the nation and how important the location is or where most universities are situated?
* We can examine the ranking distributions of the universities
* We can observe the need for financial aid to be provided to students by the university.
* We can examine the student and faculty ratio in the university.
* We can figure out the factors that make the top-ranked universities stand out from the rest of the universities in America.

1. Are there specific variables or columns that are crucial to your analysis?

* rank :
* organizationName, yearFounded, collegeType
* state and region
* studentPopulation
* campusSetting
* medianBaseSalary
* totalGrantAid, percentOfStudentsFinAid , percentOfStudentsGrant
* studentFacultyRatio

1. What is the structure of the dataset?



1. What do you want to learn or discover from the data?

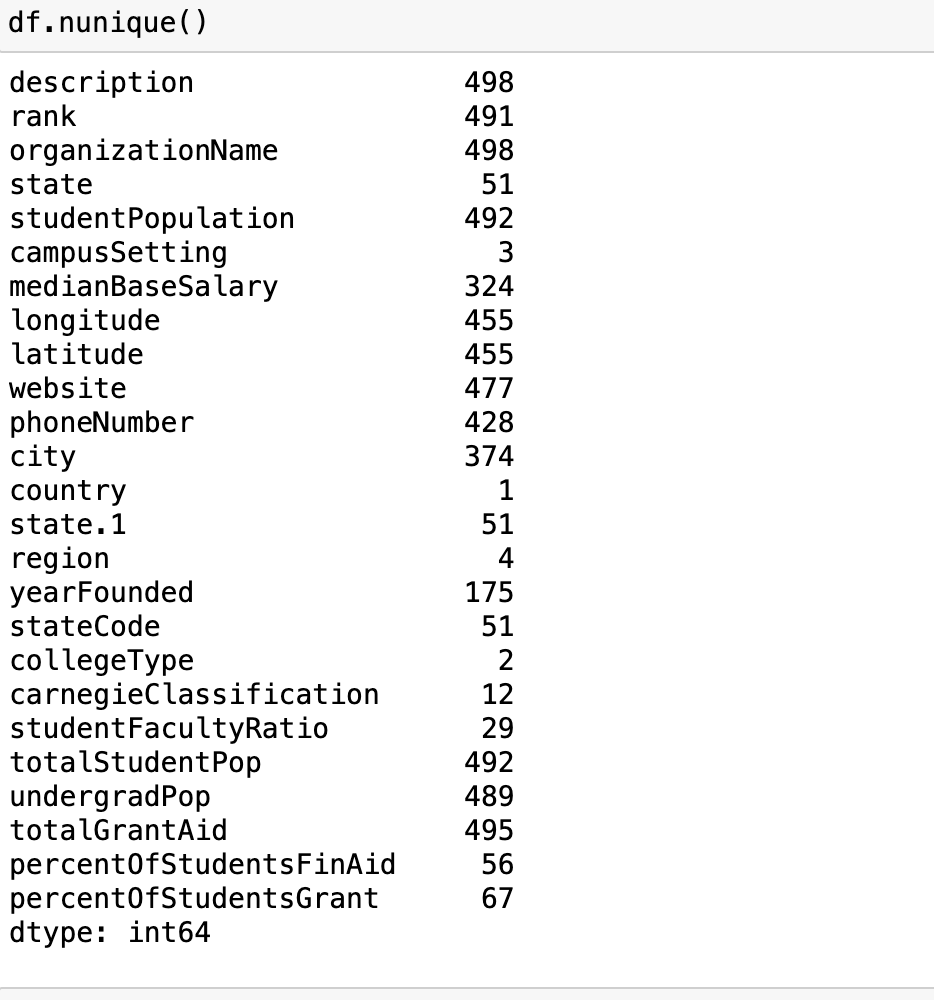
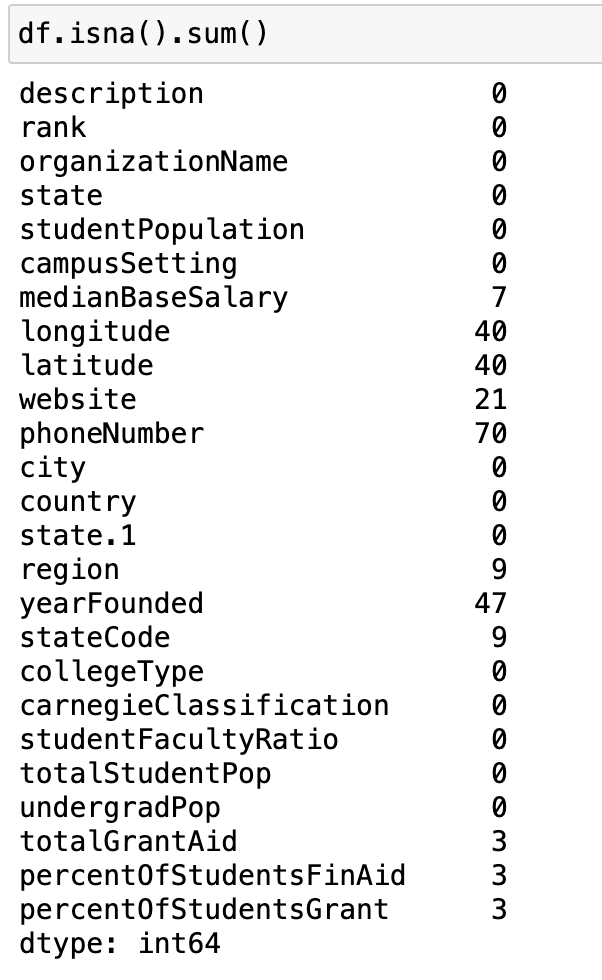
* To learn the impact of the attributes such as, state, longitude, latitude, totalpopulation, Grant, Financial Aid on the ranks of the Universities.

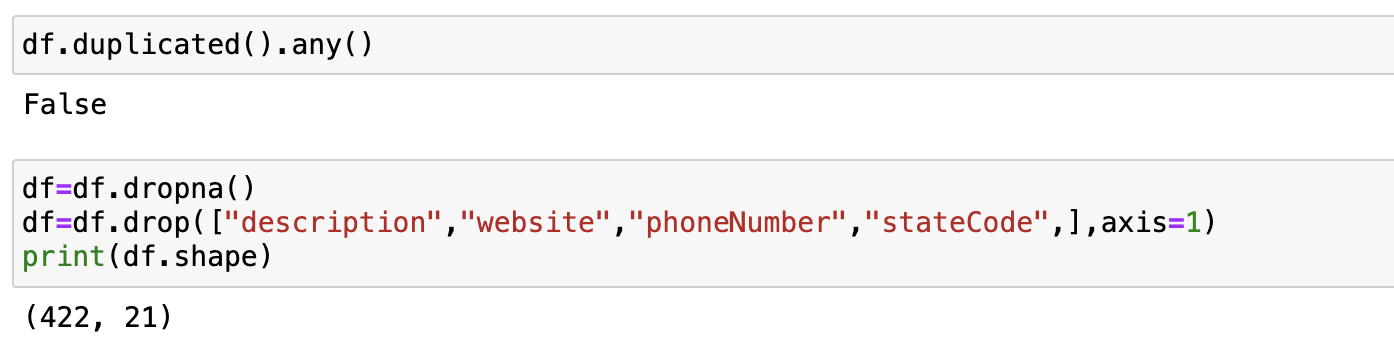
1. Break down your main question into smaller 5 specific questions.
2. To sort the colleges based on the rankings.
3. To sort ranking of colleges based on the median salary.
4. To sort top organisations by student population and medianBaseSalary.
5. To sort the ranking of colleges based on Financial Aid.
6. To find Grant ratio per college.
7. What is the correlation between student population and total grant aid?
8. What is the median\_Salary based on the ranking and organisation's name?
9. What is the count of the organisation in each state?
10. What is the count of organisations in each city?
11. What is the distribution of universities across different campus settings?
12. How does the median base salary vary across different states?
13. How does the student-faculty ratio vary among different university types?

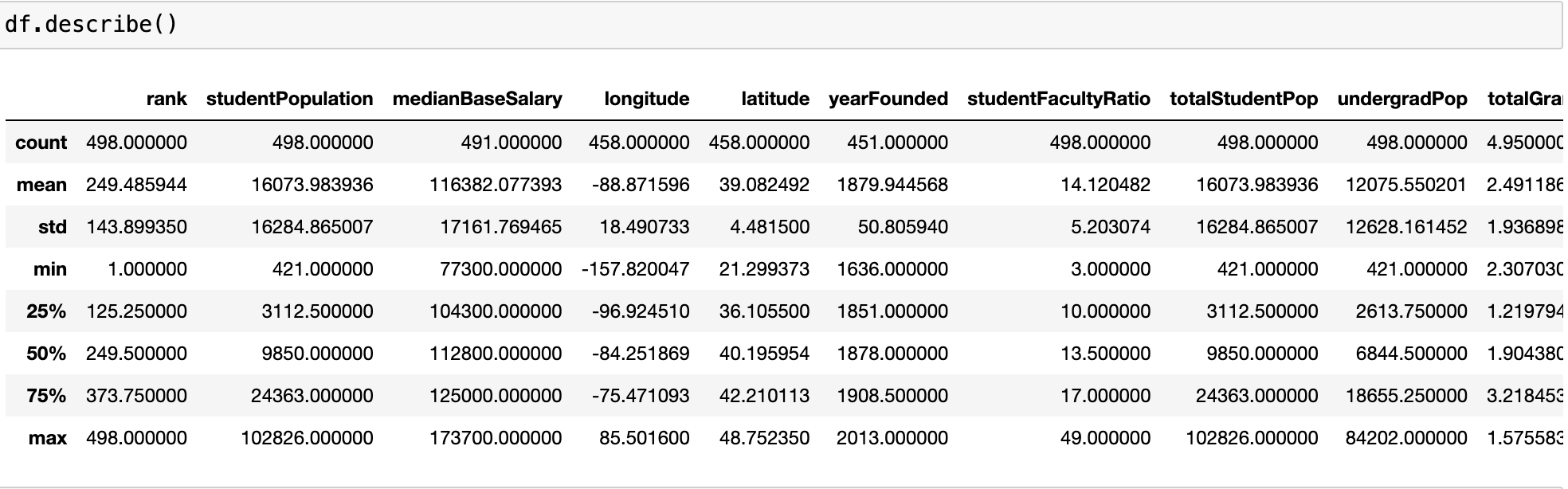
9) Depending on your goals, use exploratory data analysis (EDA) for all the

questions and display the findings.

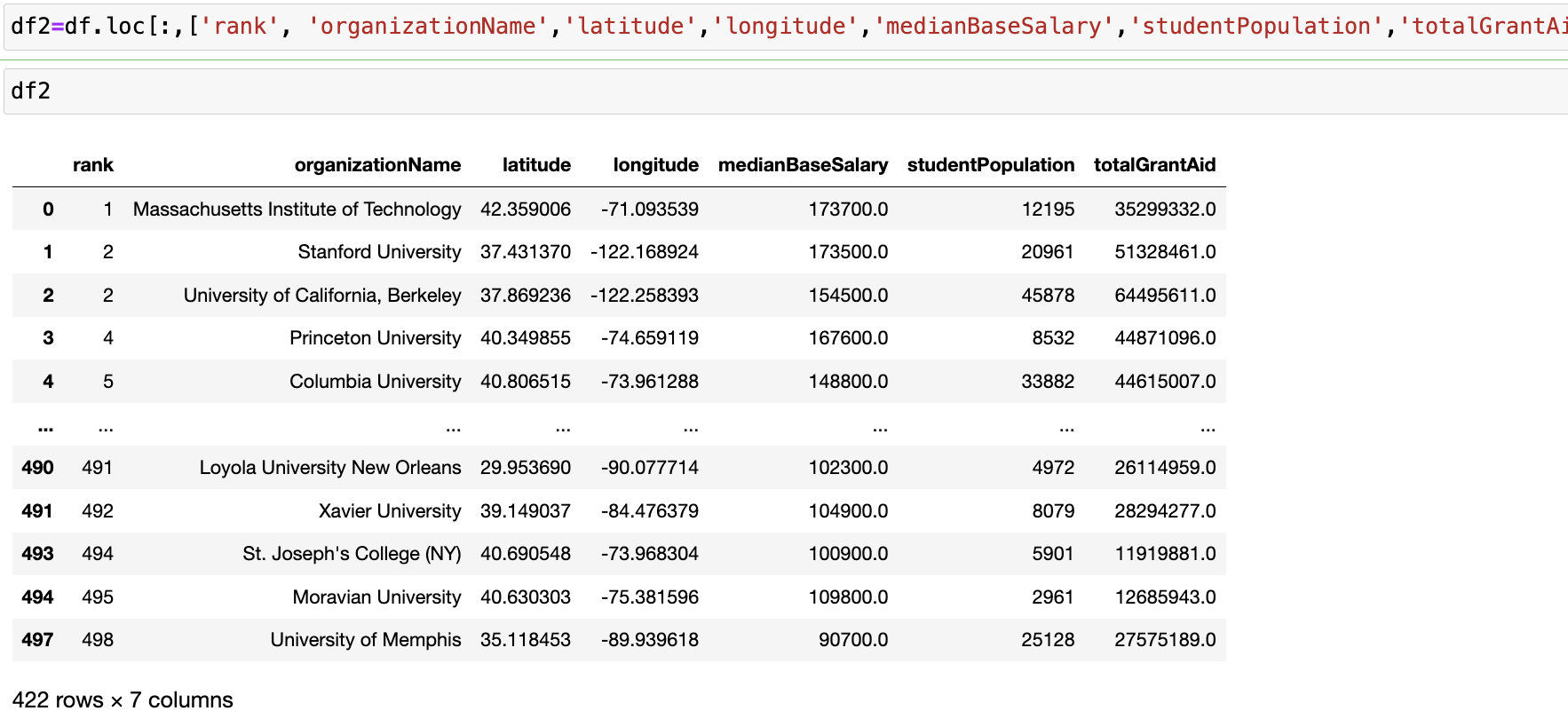
* These are the steps of data preprocessing being showcased below:

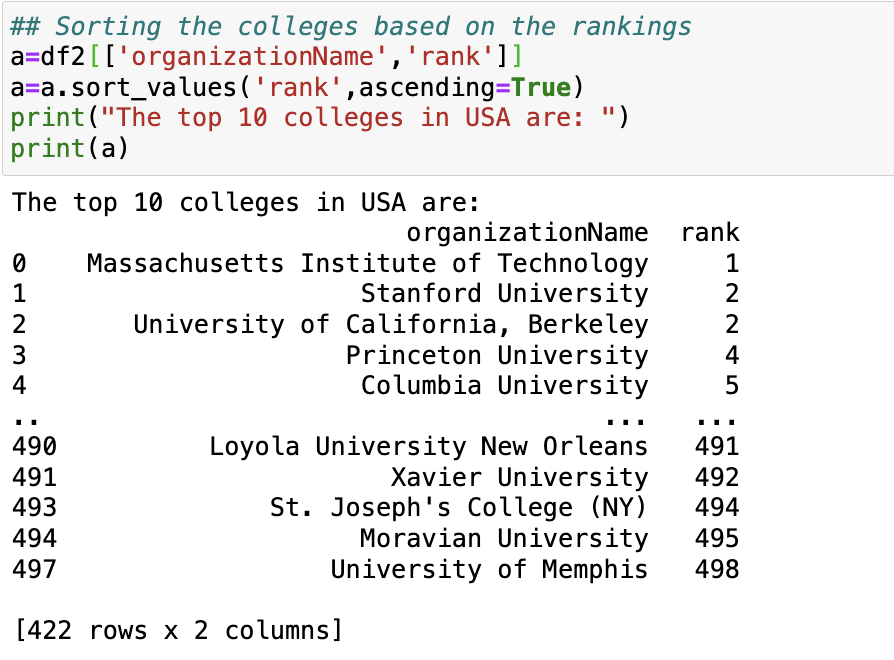




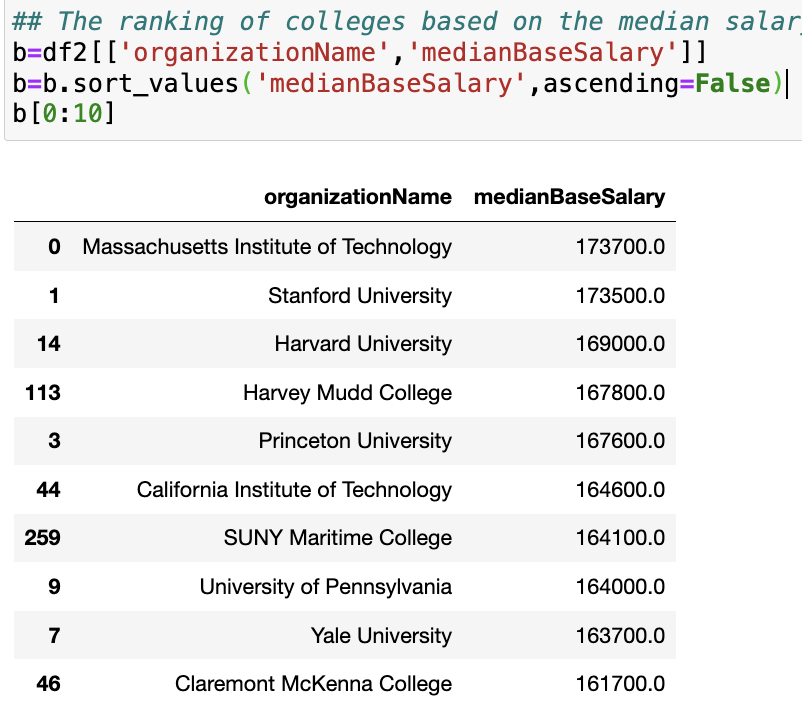


* df2 is created by fixing some of crucial attributes of the data and by using df2, insights will be as the size of the dataset has been decreased so we know which columns are beneficial for drawing inferences.

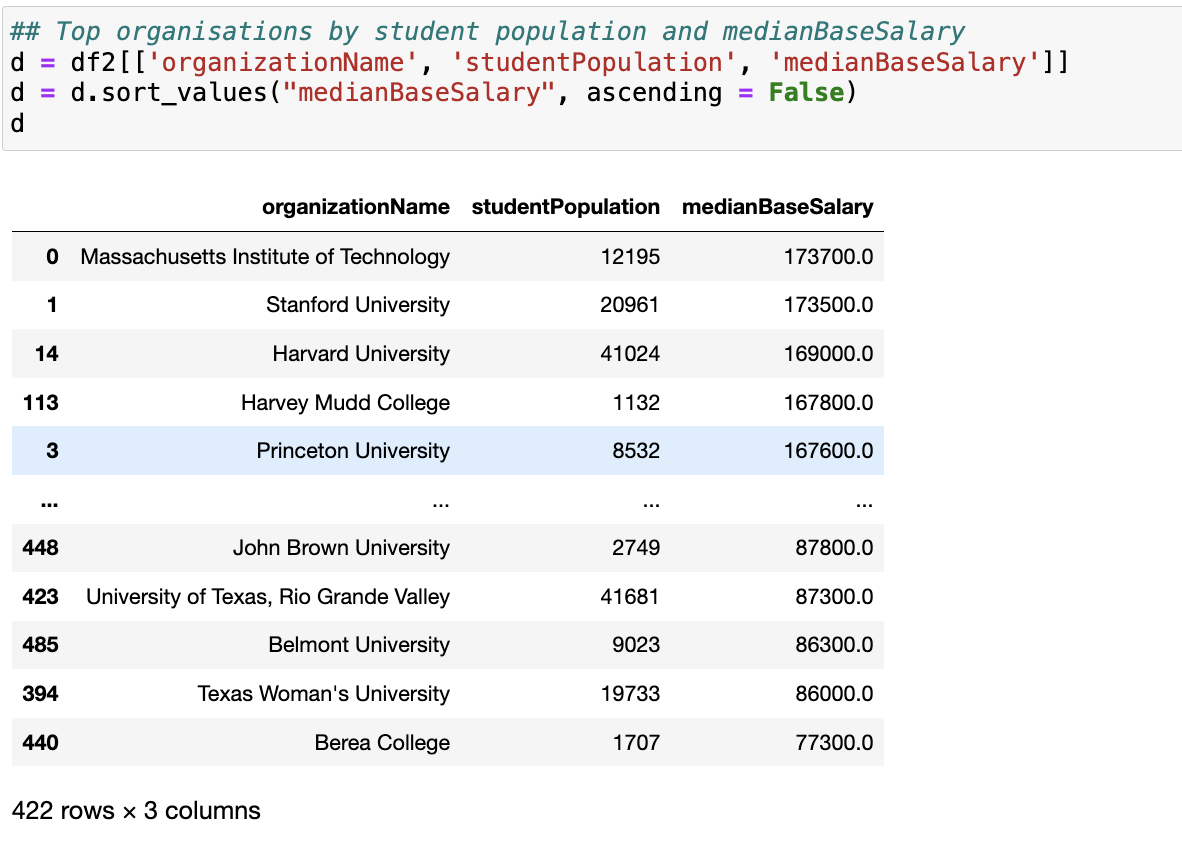




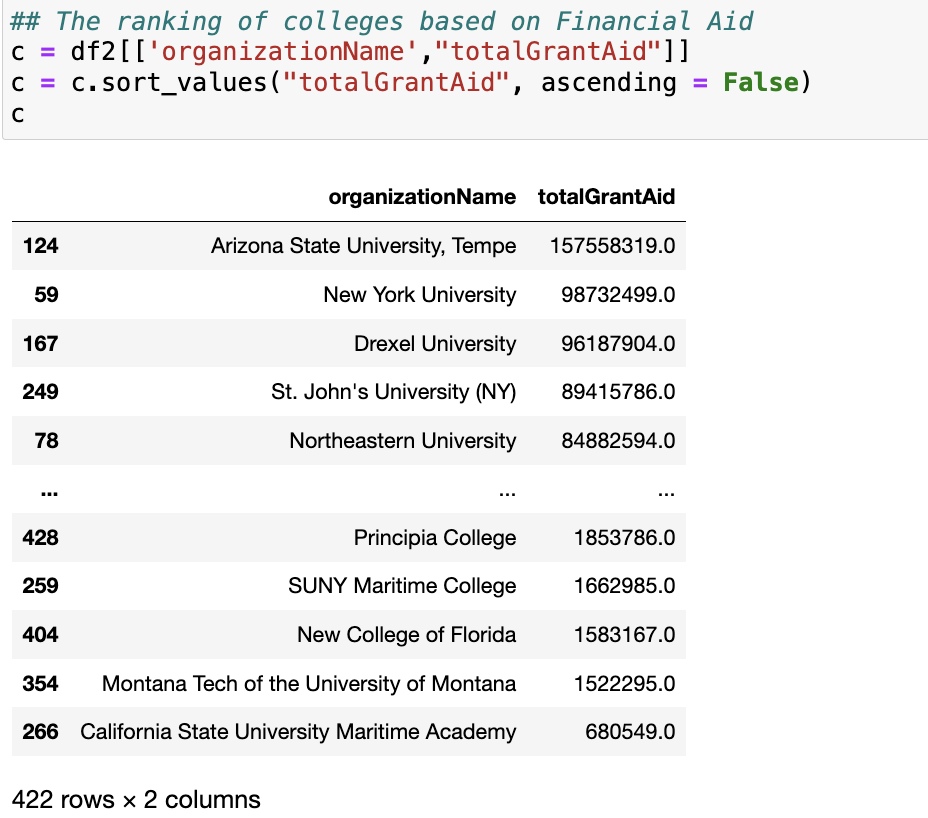
* To sort the colleges based on the rankings we get MIT to be at 1st rank and University of Memphis to be at 497th.



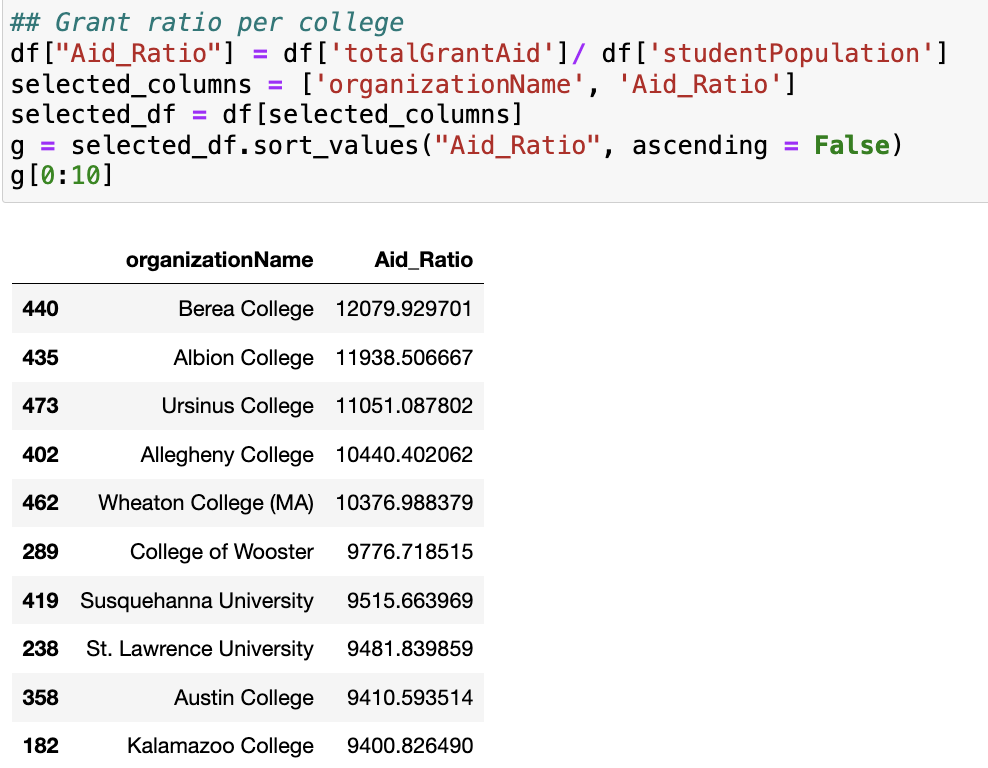
* MIT has the highest Median Base Salary among all the American Colleges whereas Claremont McKenna College lies on the 10th position.



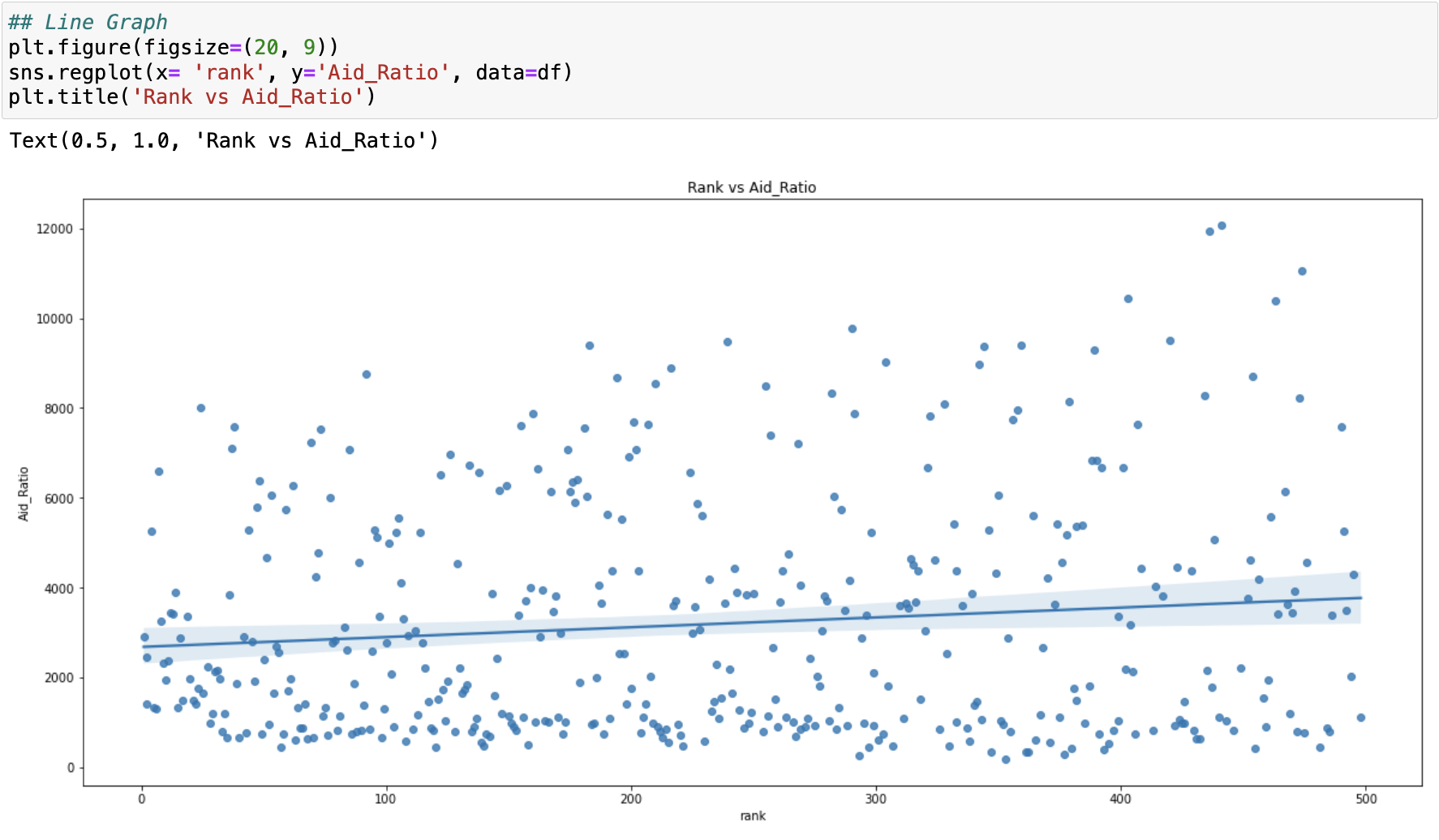
* MIT has the highest median salary based on the population of student and Berea College lies on the last in the terms of student population and the median Salary.



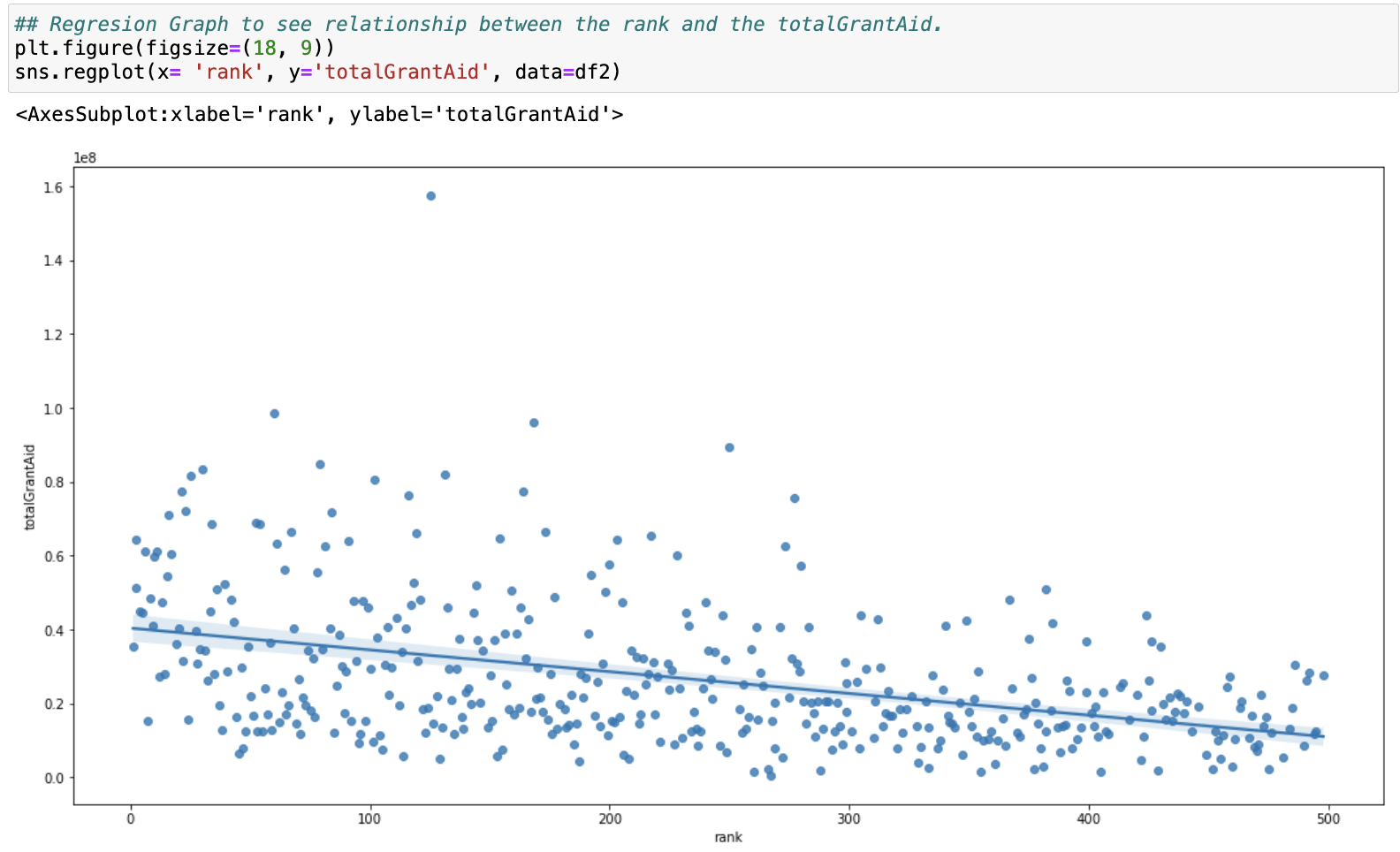
* Arizona State University has the highest Grant\_Aid in American University, as i was going through the GrantAid and the total student population, I came up with a new column named “Aid\_Ratio”.
* It is the ratio of = totalGrantAid / studentPopulation
* It gives us the ratio of the grant provided by the oragnisation by the population, which is the ratio of grant each student can avail in that organisation.



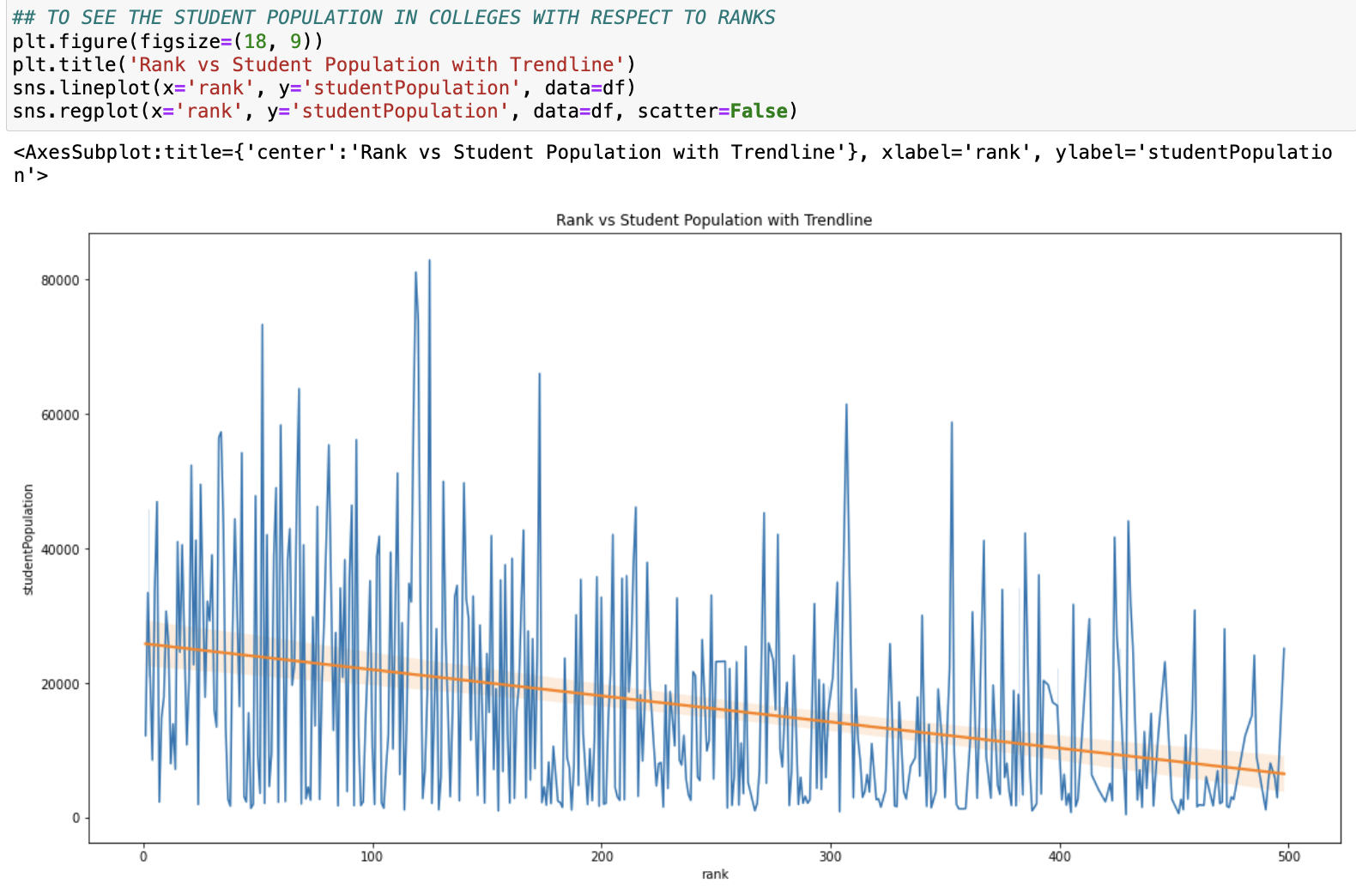
* Berea College which has the least median based salary provides with the highest ratio of grant per student.



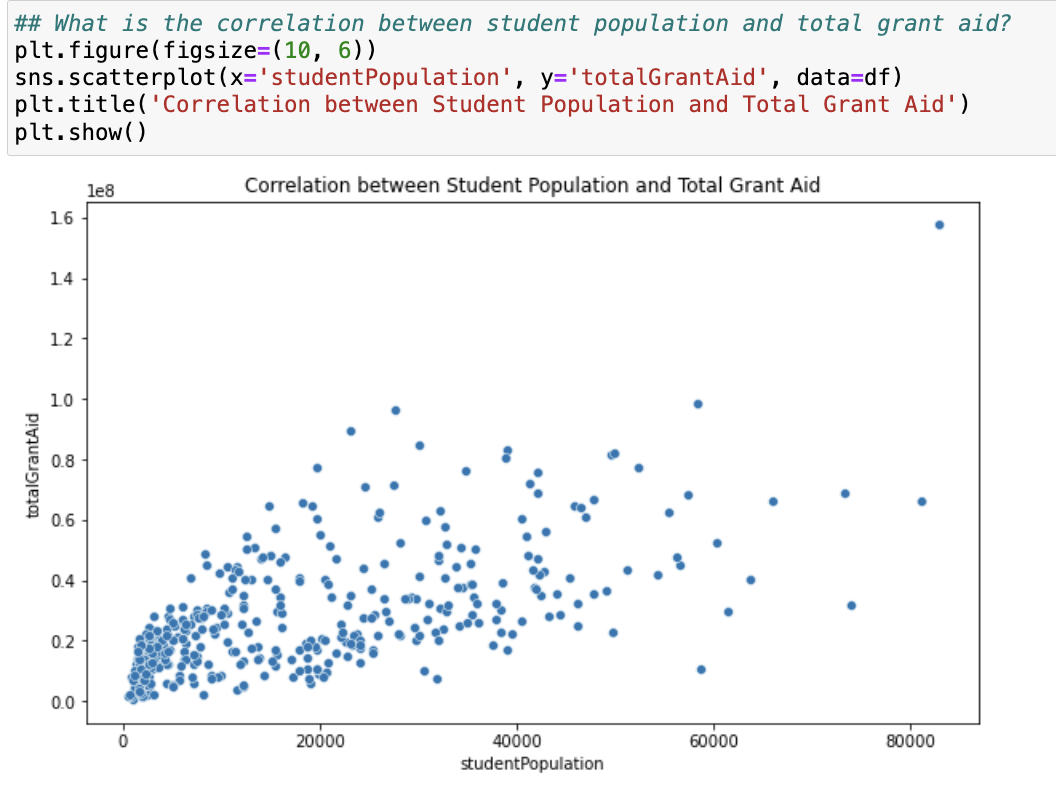
* This shows that as the rank increases the Aid\_Ratio also increases.



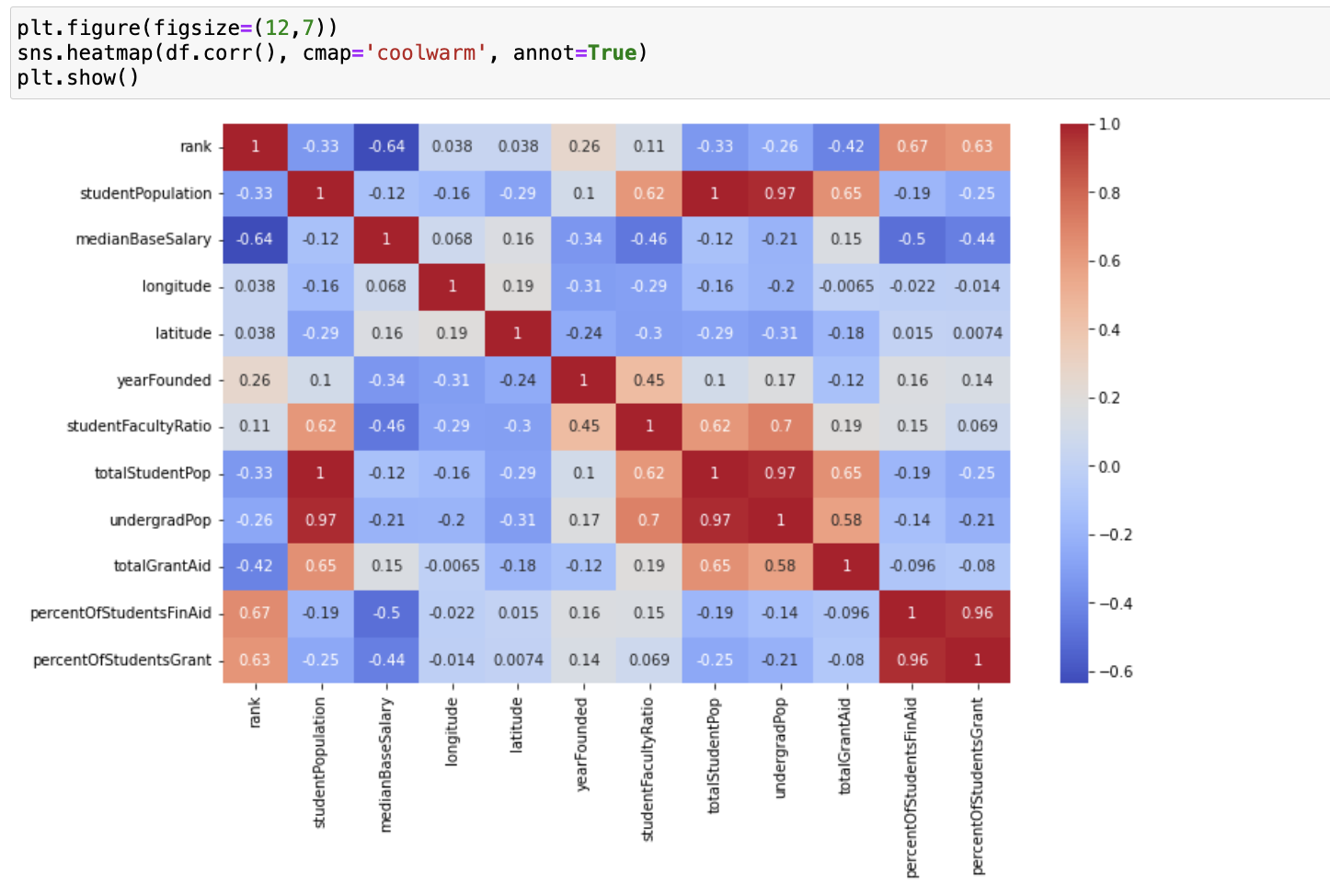
* This plot shows that as the Rank of the colleges increases the grantaid provided also decreases.



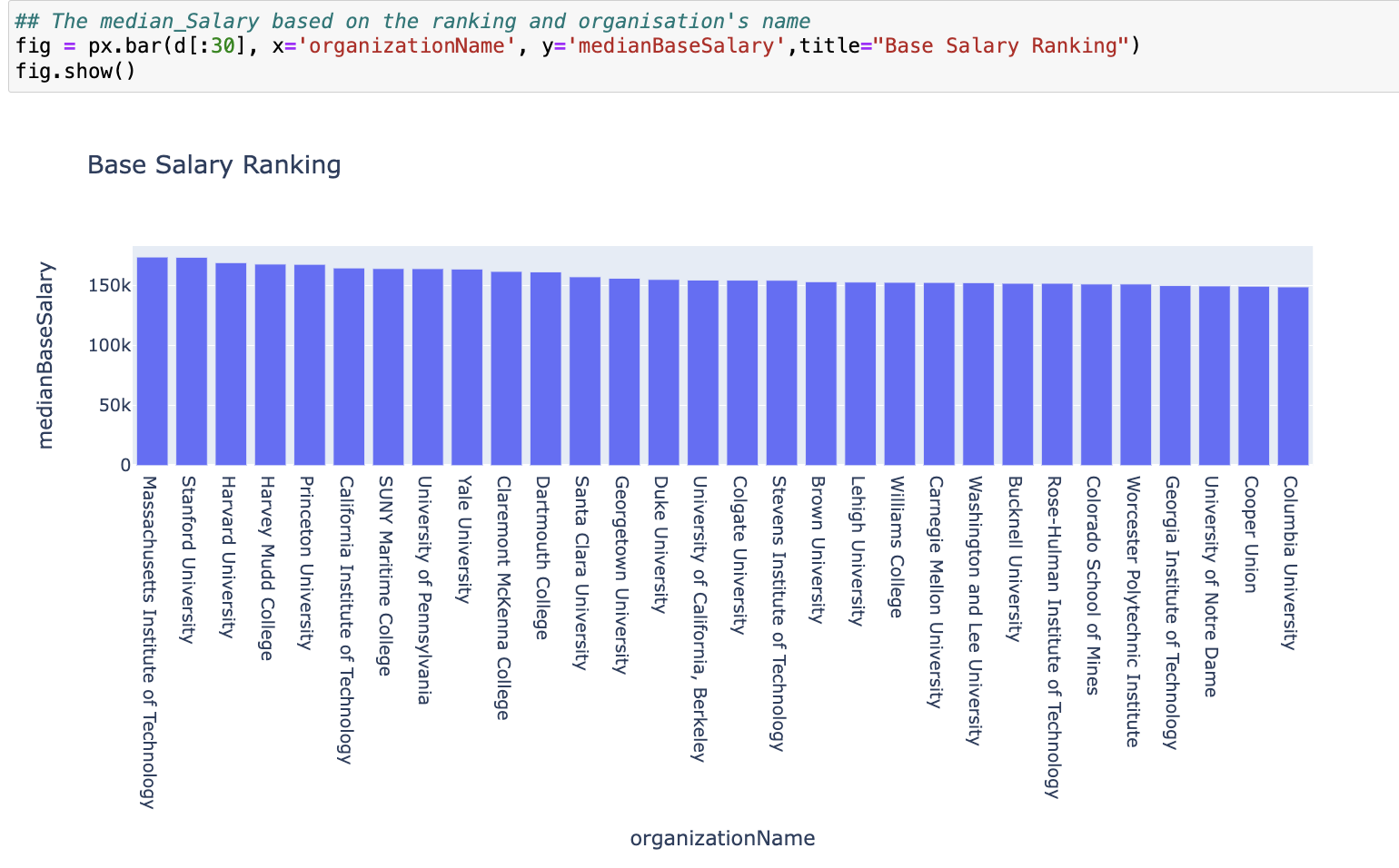
* The population of the students also tend to decrease as the ranking of the college decreases, the students tend to get enrolled in the college with better ranks.



* The datapoints are dense in the region of 0-20000 students population, which mean as the population increases the totalGrantAid decreases for the colleges and maybe others are the outliers.



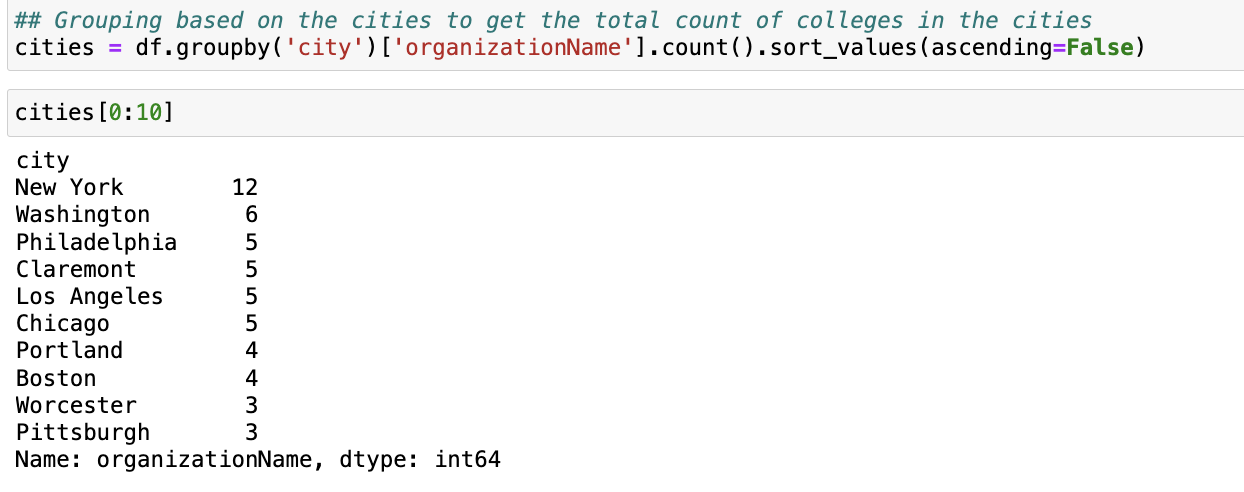
* This is a heatmap, it shows that there is strong positive correlation between:
* There is negative correlation between rank and medianBasesalary which tells that as the rank of a university gets better the university gives a higher salary.
* We see as the rank gets better the amount of financial aid decreases.
* From the correaltipn heat map we can also confirm that longitude and latitude play no role in effecting any other variable.



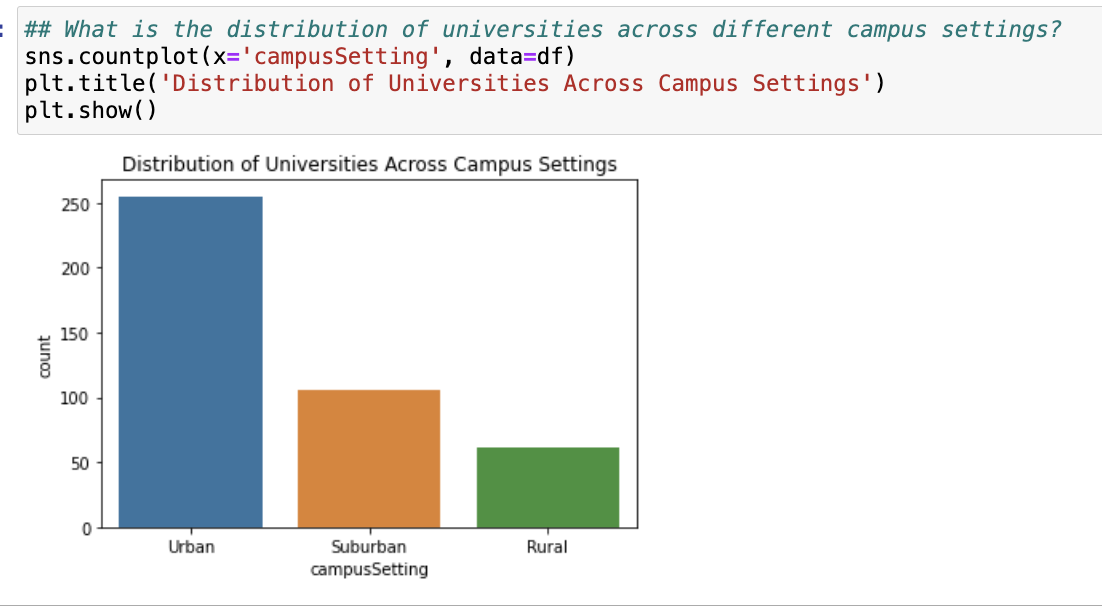
* The average salary decreases with the ranks of the college, being MIT at the top.



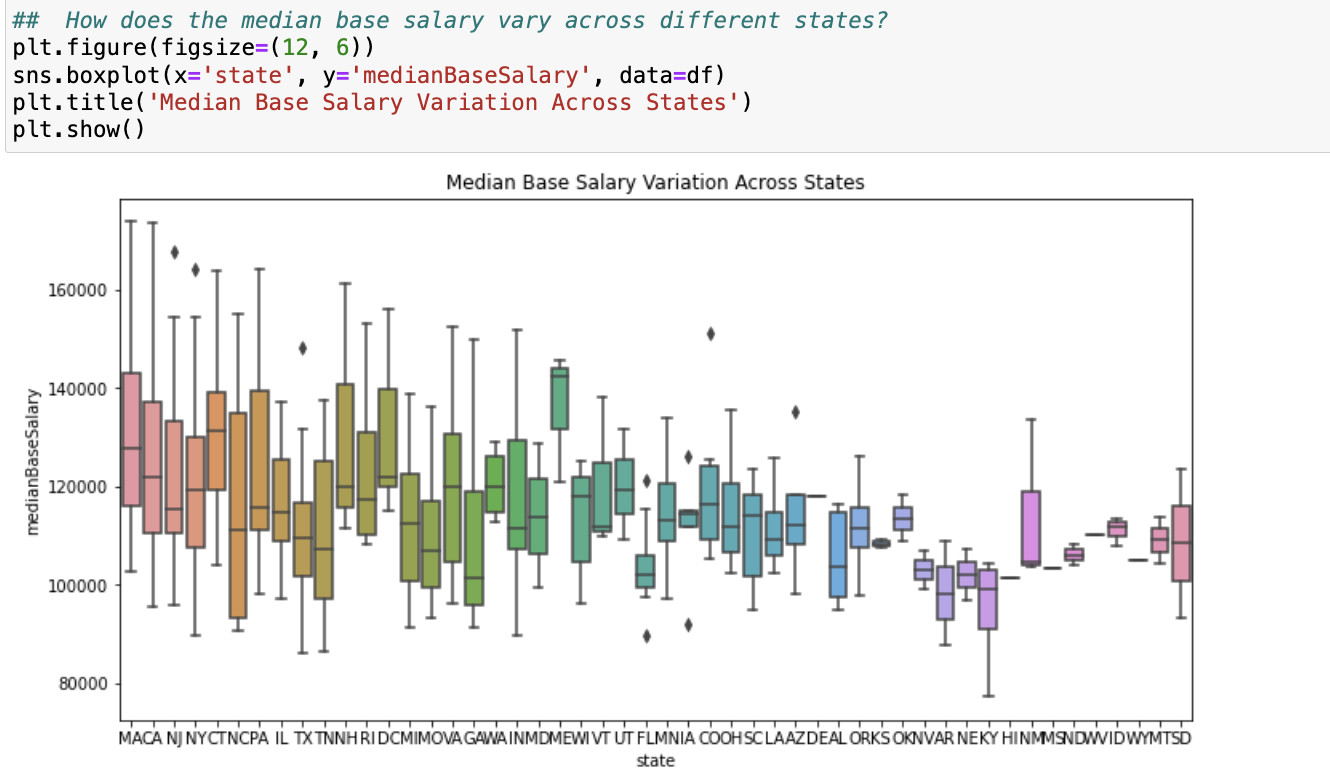
* This tells us the number of colleges/organisations in top 10 states of USA.

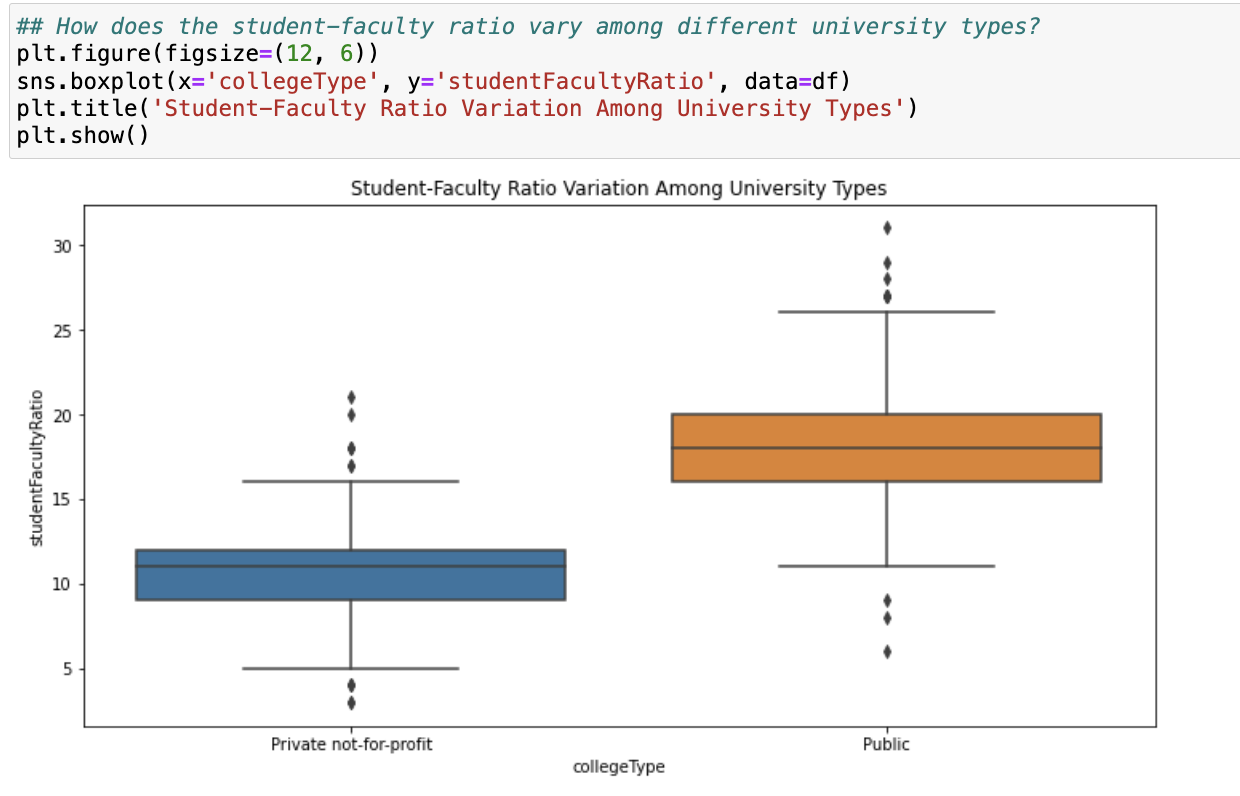


* This gives us the count of the colleges/organisations situated in top 10 cities of USA.



* There are three categories of campus settings available in the USA, Urban have most of the educational organisations while Rural having the least number of organisations.



* This plot showcases the distribution of the median base salary across the states, there are some states which have more median base salary(outliers). MA state having the highest along with ME.
* 
* According to the plot we can say the student faculty ratio is better in Public universities rather than Private non-Profit univerities