Data Narrative-2 (US Colleges)

Bhoumik Patidar -22110049, Mechanical Engineering, Prof. Shanmuganathan Raman,IIT Gandhinagar

Abstract-We analyse the higher education system situation in the United States Of America in 1995. For this purpose we use the two datasets aaup.data and usnews.data to obtain results based on various critical parameters.The analysis focuses mainly on the economic aspects of the universities and the perspective of a student looking for higher education opportunities. We also focus the perception aspect giving utmost focus to the visualisation of the data.

I. DATA OVERVIEW[1]

The US News and World Report dataset contains information hundreds of American colleges and universities, including various indicators of their academic quality, tuition fees, graduation rates, faculty size, and student diversity. The data is collected from public and private sources and covers a variety of metrics ranging from financial resources, alumni donations to acceptance rates, and standardized test scores. The dataset ise used to analyze and compare the performance of different institutions in various areas of higher education.

The AAUP dataset for the ASA Statistical Graphics Section's 1995.Data Analysis Exposition contains information on faculty salaries for 1161 American colleges and universities. It provides salary distribution among various ranks of the professors as well as the distribution in various types of American colleges.

II. QUESTIONS/HYPOTHESIS

1. [USNEWS]Question-Which states in the US had a relatively well-developed higher education system during a particular period of time?

2. [USNEWS]Question-What is the variation in the overall perception of books over the time scale?

3. [USNEWS]Question-What is the dependence of ratings of a book on its popularity?

4. [USNEWS]Hypothesis-Books that are more popular have higher probability to be in a persons to read list over the books that have higher

ratings.

5. [USNEWS]Question-Approximation of probability of online availability of the books in forms of audiobooks or ebooks based on the limited information in the dataset.
6. [AAUP]How does the distribution of salaries changes among the different type of universities in the US?
7. [AAUP]What is the distribution of salary in various states by rank?
8. [AAUP]How does the distribution of number of professors varies with the type of college across the states.
9. [AAUP]Analysing the dependence of compensation paid to professors with the state the college is present in.
10. [AAUP]Understanding the impact of the quantity of faculties an institute hires on the relative amount of compensation it is able to pay to its professors.

III. DETAILS OF LIBRARIES AND FUNCTIONS USED

1. PANDAS[2]
Pandas is a data analysis and manipulation library.
Important functions used-
Dataframe creation and corresponding operation on it.All important codes have there snippets in the text below.

2. MATPLOTLIB[3]
This library is used to create visualisations in python.
Important functions used-
The plots of scatter and bar graph along with the additional features included to improve the visualisation

IV. ANSWERS TO THE QUESTIONS/HYPOTHESIS

A. [USNEWS]Question-Which states in the US had a relatively well-developed higher education system during a particular period of time?

• PURPOSE
The idea is to use the available parameters such as instructional expense made by the colleges state wise, the total expenditure a student is burdened with, the average graduation rate of the state and the quality and quantity of the faculty.

- EXECUTION WITH CODE SNIPPETS
  -The total expenditure of a student in calculated using the necessary data-
  For the following criteria, mean of each state is calculated and plot is generated-

1)Instructional expense made by the colleges state wise

```
usnews['TotalExp']=usnews['instate_tuition']+usnews['room_board_cost']+usnews['room_cost']+ usnews['
```

2)Total expenditure a student is burdened with

3)Graduation rate of the state

4) Plot of the number of faculties with a terminal degree is also created

The colleges which are in the 20 top for each criteria are recorded.

Code for the 4th as an example-

```
state_sums.plot(kind='bar', figsize=(12,6))
plt.title('Number of Faculties with a terminal degree by State')
plt.xlabel('State')
```

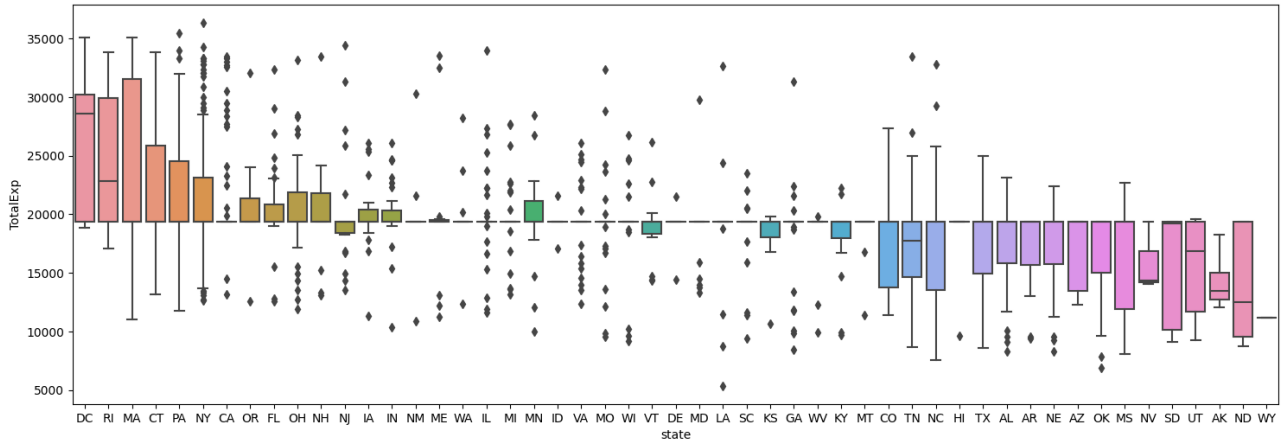
```
state_sums = df.groupby('state').sum()

state_sums = state_sums.sort_values('pct_faculty_terminal_degree', ascending=False)

top_states = state_sums.head(20)
```

- RESULTS

The following graphs are obtained-

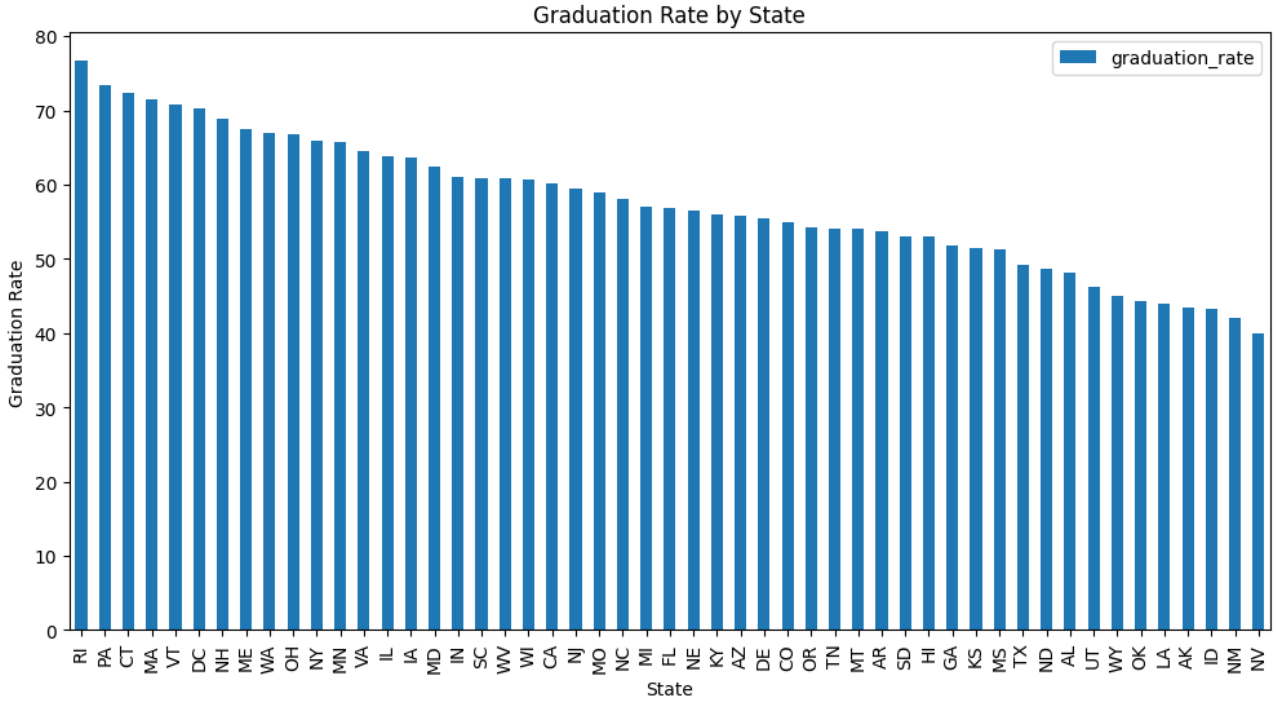
1)The states with the least expenditure are -

[‘WY', 'ND', 'AK', 'UT', 'SD', 'NV', 'MS', 'OK', 'AZ', 'NE', 'AR', 'AL', 'TX', 'HI', 'NC', 'TN', 'CO', 'MT', 'KY', 'WV']



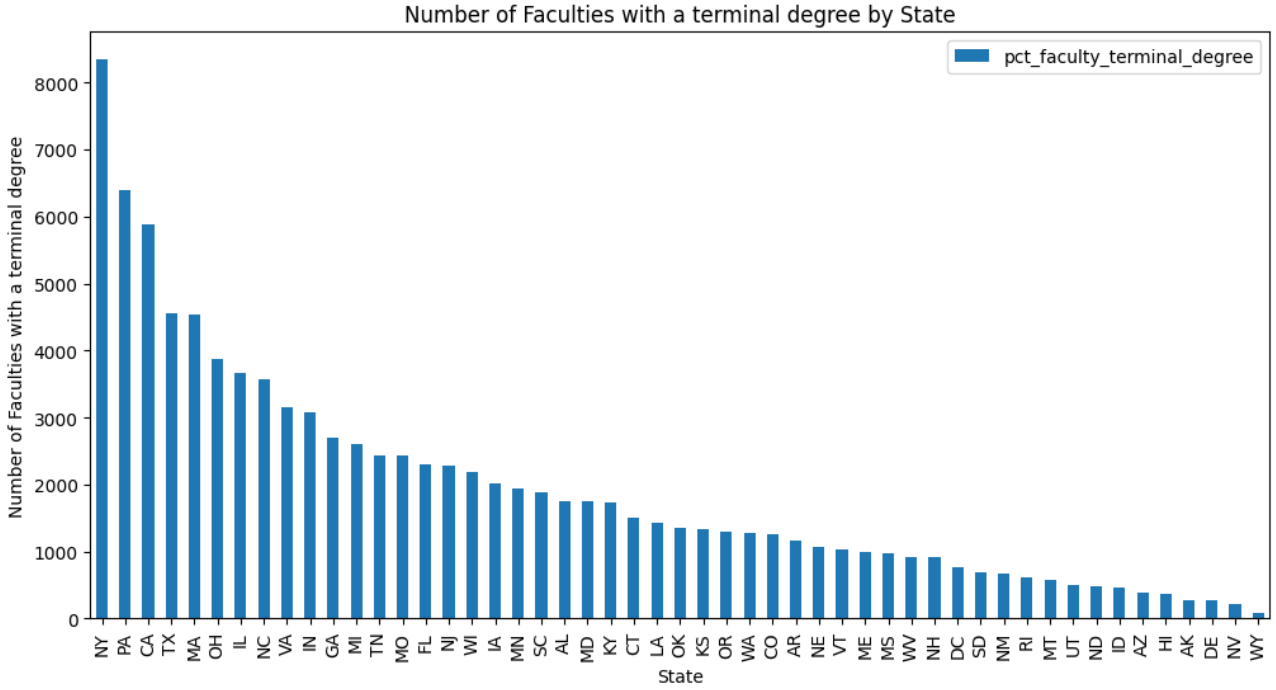
2)The states with the best graduation rates are-

['RI', 'PA', 'CT', 'MA', 'VT', 'DC', 'NH', 'ME', 'WA', 'OH', 'NY', 'MN', 'VA', 'IL', 'IA', 'MD', 'IN', 'SC', 'WV', 'WI']



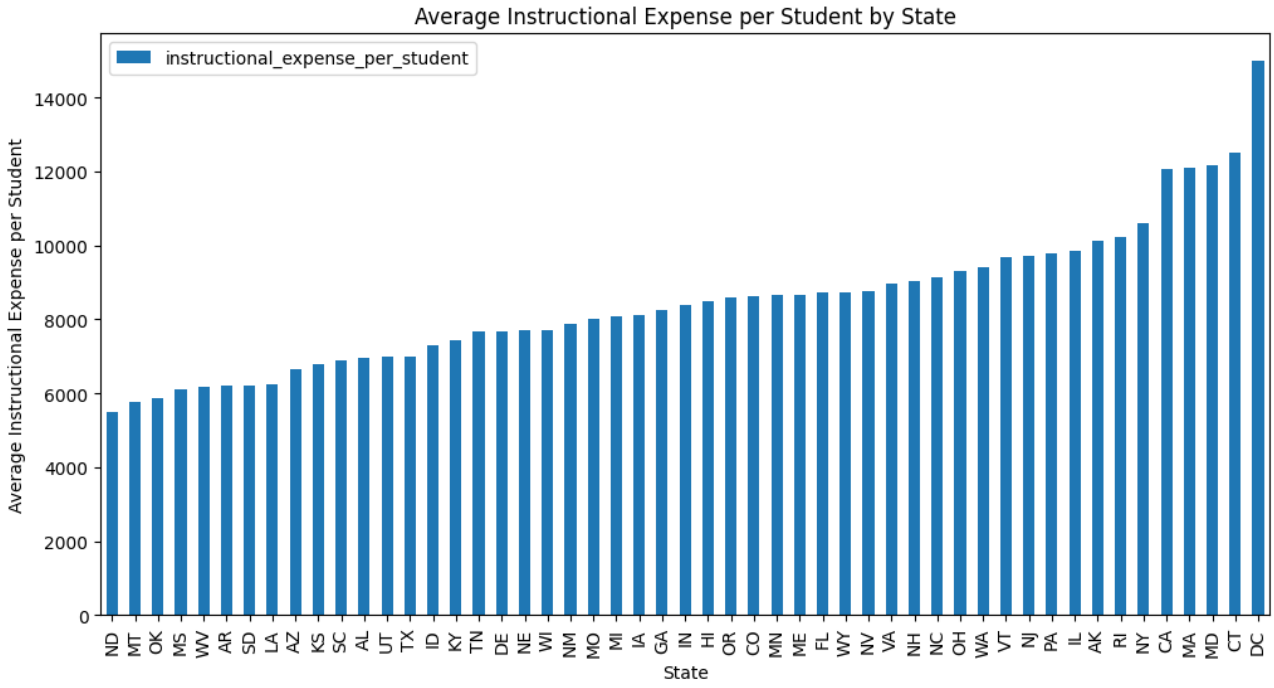
3)The states with the best faculty scenario are-

['NY', 'PA', 'CA', 'TX', 'MA', 'OH', 'IL', 'NC', 'VA', 'IN', 'GA', 'MI', 'TN', 'MO', 'FL', 'NJ', 'WI', 'IA', 'MN', 'SC']



4)The states which spend the largest amount of money on students are-

['DC', 'CT', 'MD', 'MA', 'CA', 'NY', 'RI', 'AK', 'IL', 'PA', 'NJ', 'VT', 'WA', 'OH', 'NC', 'NH', 'VA', 'NV', 'WY', 'FL']



- CONCLUSION,INFERENCES AND USAGE.
  A list of states which were in top 20 in at least 3 of the above criteria is obtained.

Hence, PA, IL, VA, NY, OH, NC and MA are the states with the best education system.

This data is of great importance for students to select a state for higher education based on what factors are most important and relevant to his situation.

B. [USNEWS]Question-What conclusions can be drawn regarding comparison of public and private institutes based on the provided benchmarks?

• PURPOSE
Use the benchmarks of graduation rates, number and strength of the universities and expenditure by the institute to draw conclusions on the state of public v/s private institutes

• EXECUTION WITH CODE SNIPPETS

-A box plot is used to demonstrate the graduation rate in public v/s private institutes.

```
df.boxplot(by='public_private', column='graduation_rate')
plt.title('Graduation Rates of Public and Private Institutes')
```

-A pie chart has been plotted to show the number of public and private institutes in the US.

```
labels = ['Public', 'Private']
sizes = [public_count, private_count]
colors = ['skyblue', 'pink']
plt.pie(sizes, labels=labels, colors=colors, autopct='%1.1f%%')
```

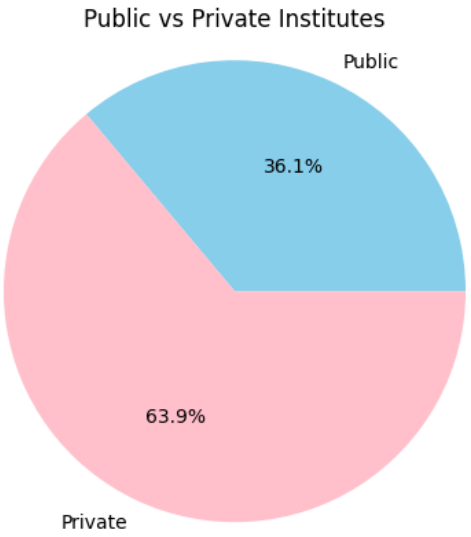
-Bar graphs to represent the undergraduate strength and institutional expenditure on the students.

```
enrollment_data.plot(kind='bar')
plt.title('Total Number of Enrolled Students by Institute Type')
```

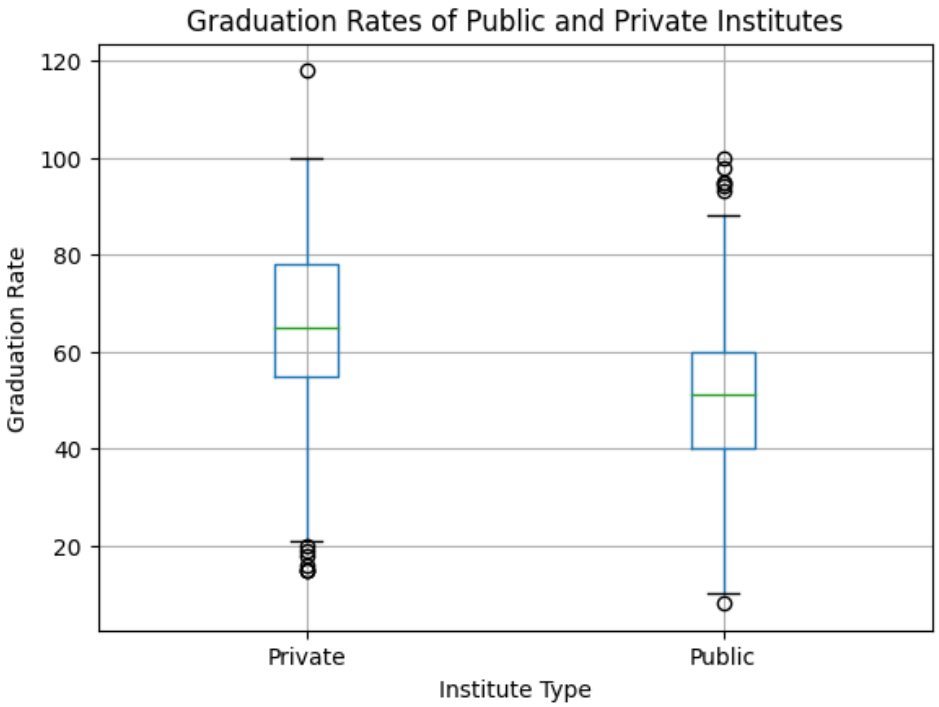
- RESULTS

The following graphs are obtained-

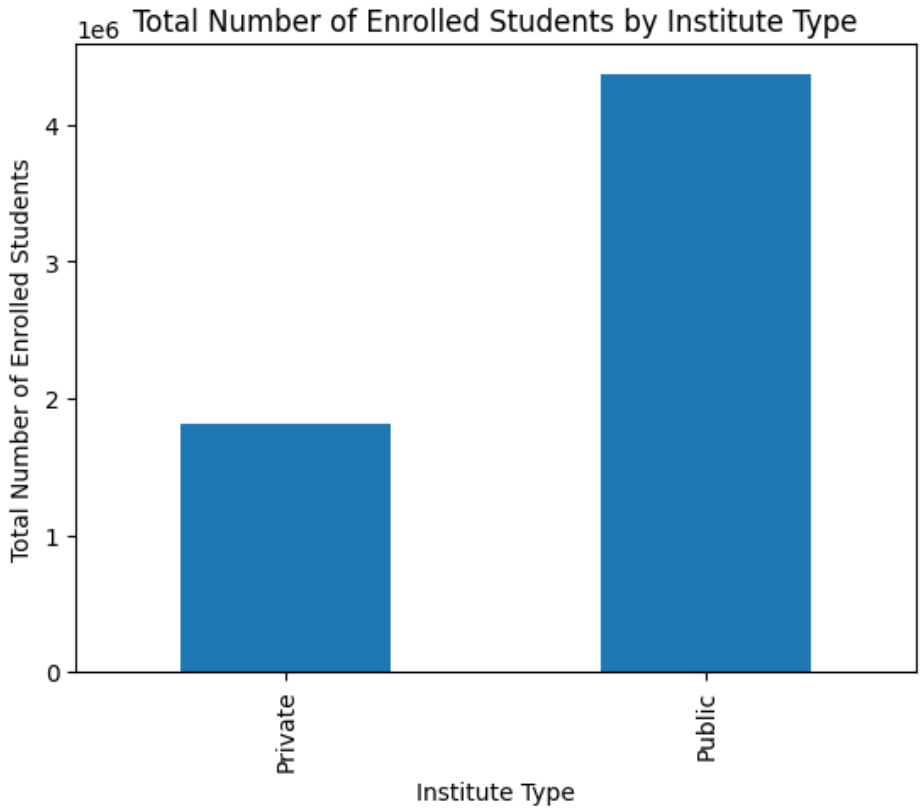
1)Percentage of public and private institutes.



2)Box plot for the graduation rates-

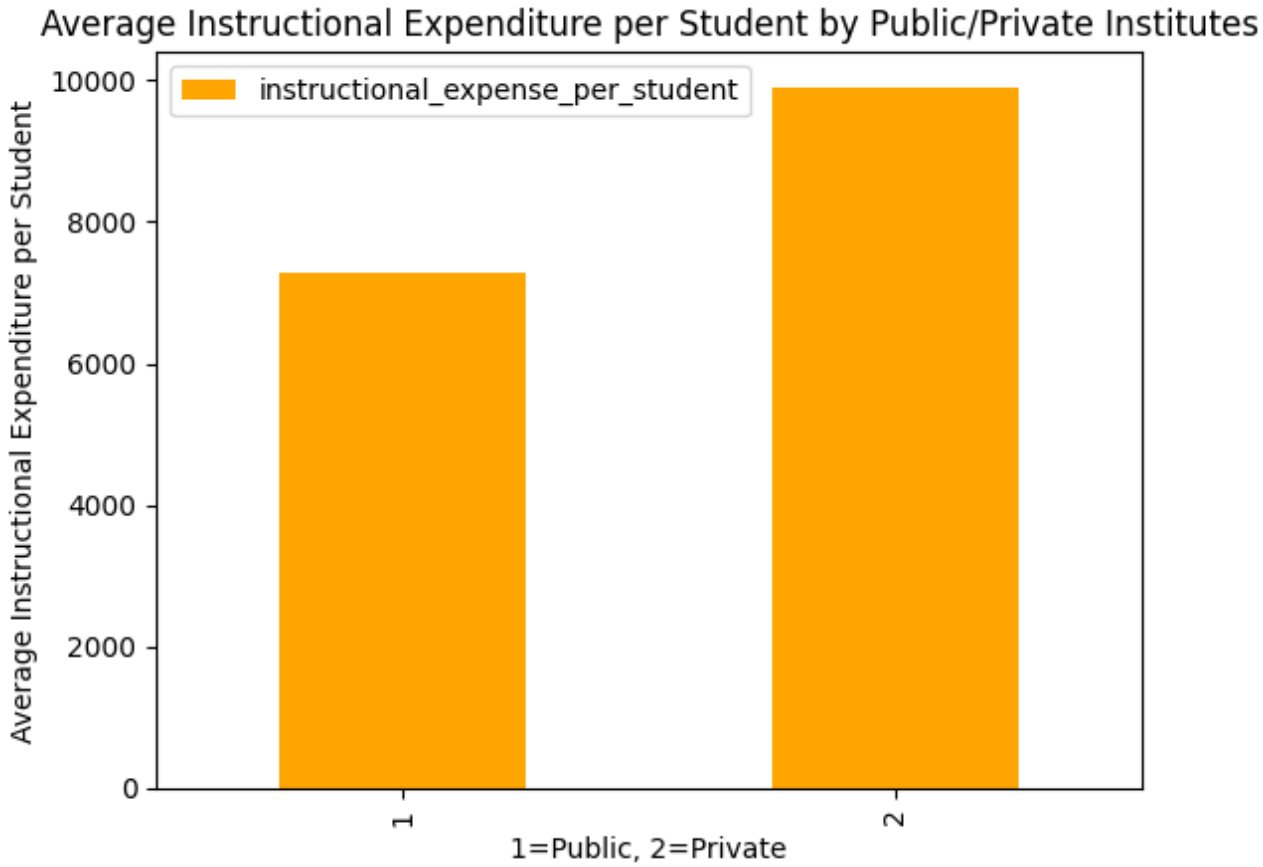


3)Number of undergrads-




```
Institute Type
Private    1820154.0
Public     4370697.0
```

4)Average Instructional expenditure per student-



- CONCLUSION,INFERENCES AND USAGE This concludes some or all of the following points-

a)The number of private institutes is more but the total strength of the public universities is much larger.

b)Graduation rates are better for private institutes.

c)Private institutes spend more on students than public universities

Overall, the quality of private institutes is much better.

C. [USNEWS]Question-Using the data of scores of entrance exams, what conclusion can be drawn regarding the future outcomes in terms of graduation rate?

• PURPOSE
We can use the data of scores of ACT and SAT and compare it with the graduation rate of the respective college to get insights into there relationship.

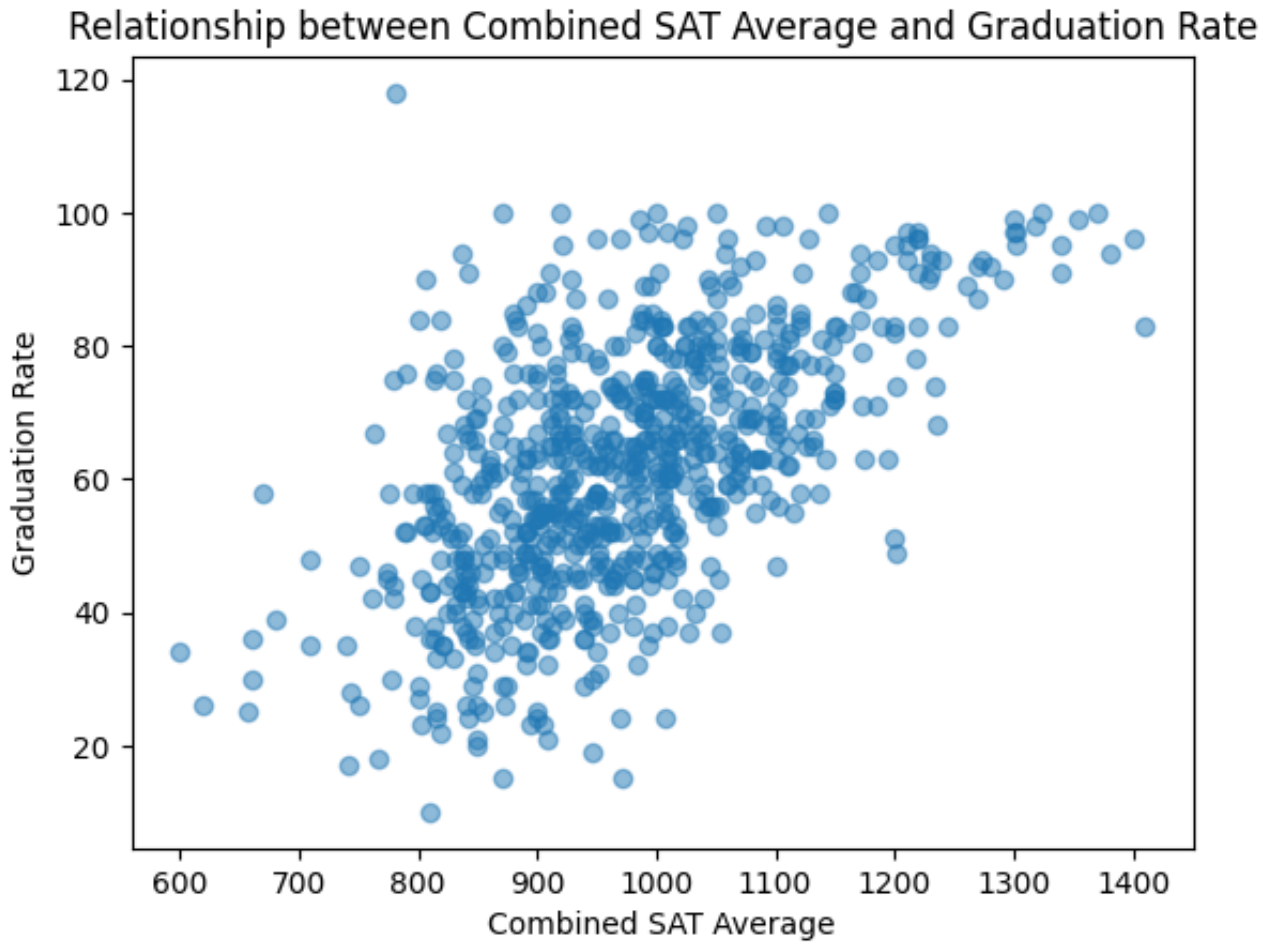
•

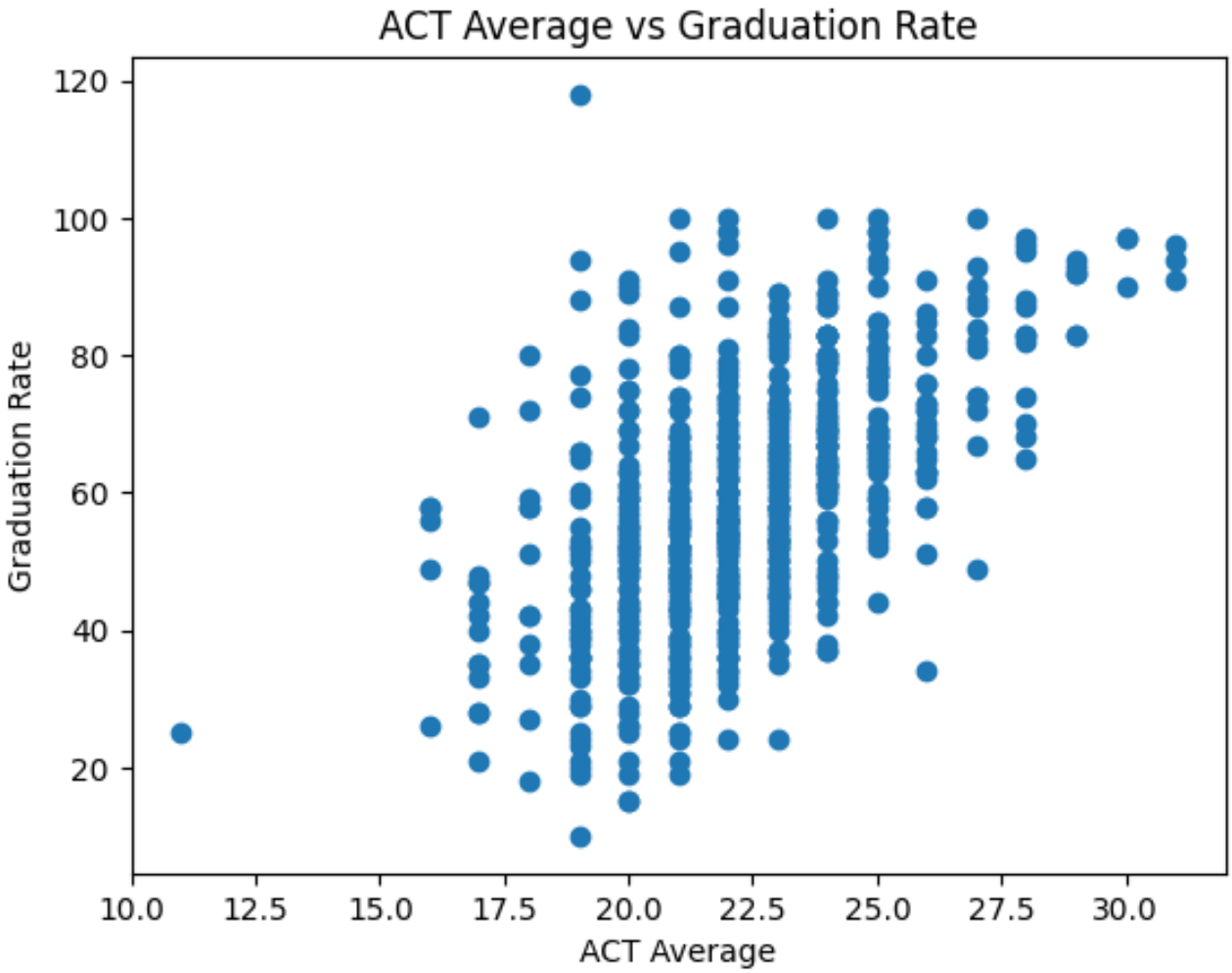
-
-
- EXECUTION WITH CODE SNIPPETS
  -We draw a scatter plot to represent the relationship of combined SAT score and ACT scores of the students v/s the graduation rate college wise.

```
plt.scatter(df['combined_sat_avg'], df['graduation_rate'], alpha=0.5)
plt.title('Relationship between Combined SAT Average and Graduation Rate')
```

- RESULTS
  -We obtain the following graphs-

Clearly colleges with higher entrance scores end up having good graduation rates.





- CONCLUSION,INFERENCES AND USAGE
  -We see that as the scores in the entrance exams increases, the percentage of successful graduation also improves.

This observation is very useful for universities to understand whether the current entrance system is working correctly or not.

D. [USNEWS]Question-How impactful is a good student/faculty ratio on the success of the university and creating a positive perspective for the college?

• PURPOSE
We will see the relationship of the colleges with good student/faculty ratio and there graduation rates as well as the percent alumni donations to note the impact of a good ratio.We also analyse how this leads to a surge in the institutes expenses.

- EXECUTION WITH CODE SNIPPETS
  -We plot the graphs-

1)Student/Faculty Ratio vs. Graduation Rate

```
plt.scatter(df['student_faculty_ratio'], df['graduation_rate'], s=10)
plt.title('Student/Faculty Ratio vs. Graduation Rate')
```

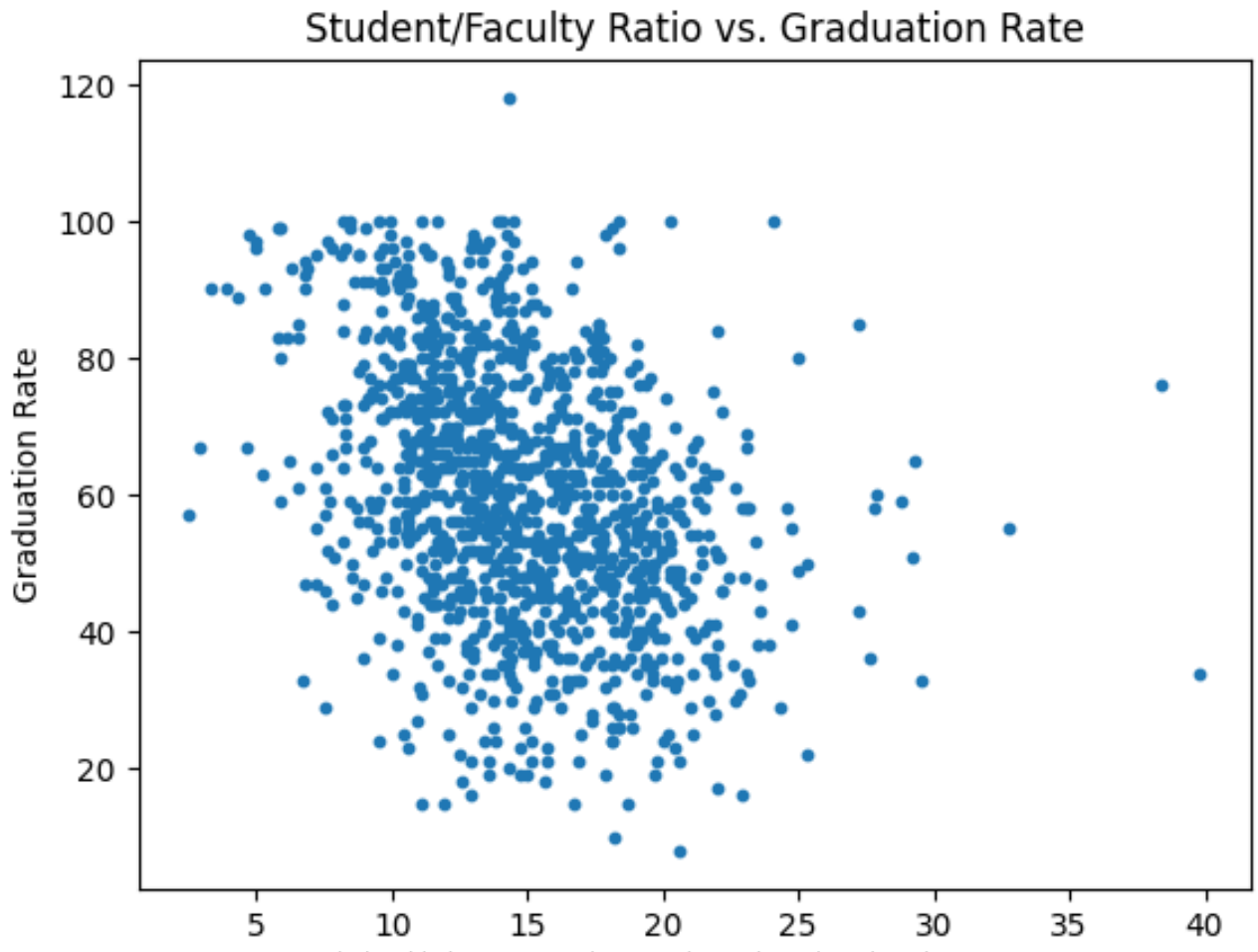
2)Student/Faculty Ratio and % Alumni Donate

3)Instructional Expense and Student/Faculty Ratio

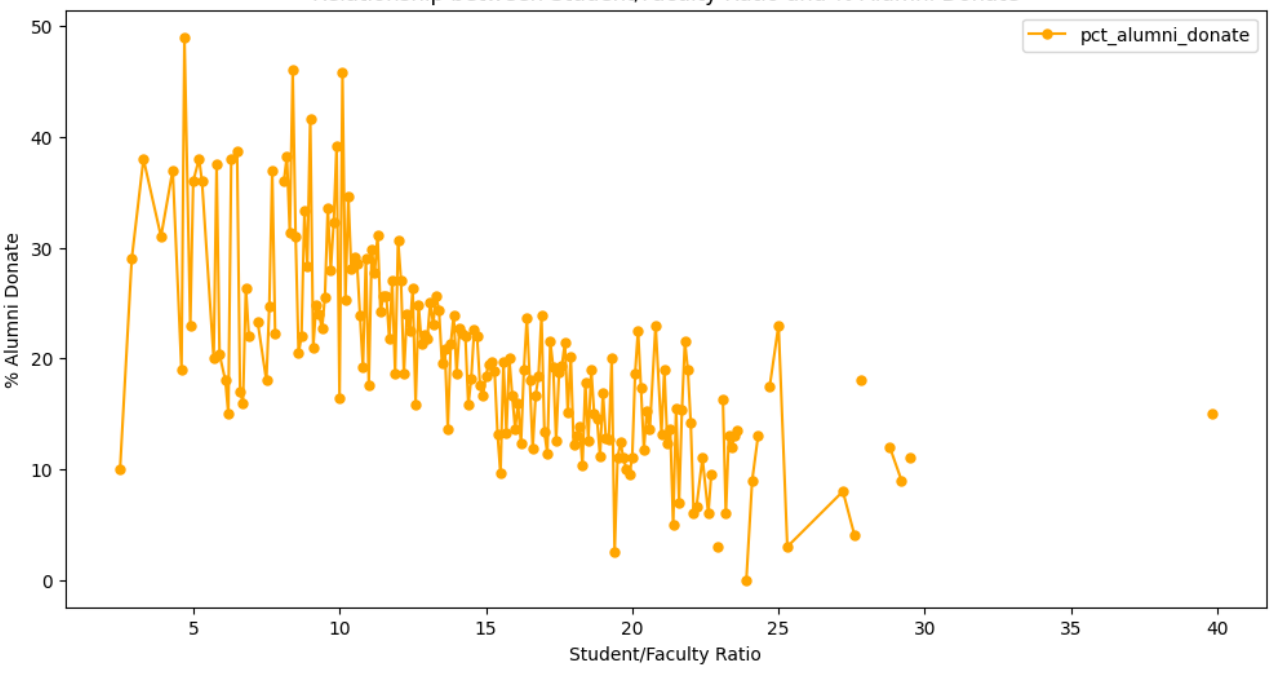
We limited the student/faculty ration to 45 to remove extreme data.

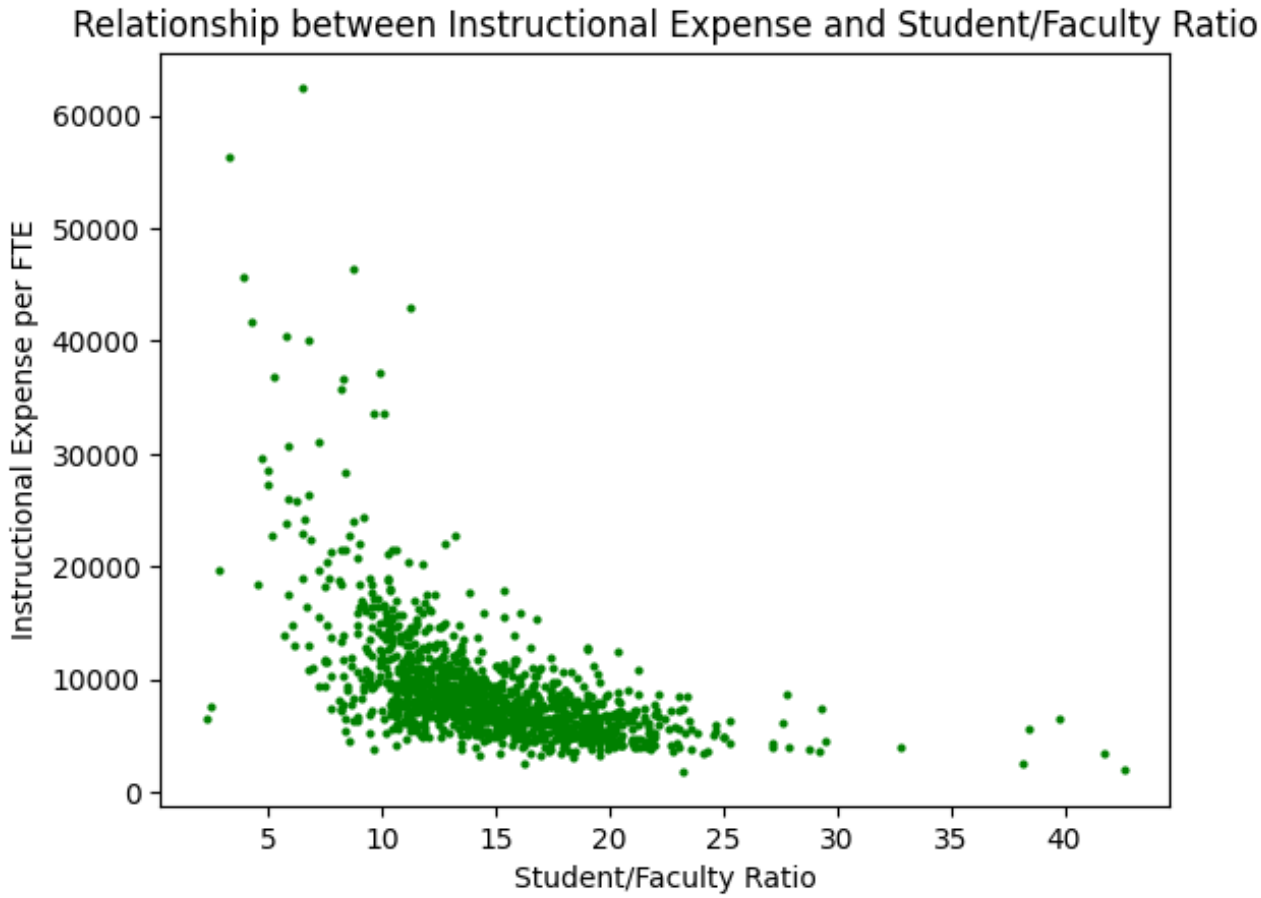
- RESULTS

We obtain the following graphs-



Relationship between Student/Faculty Ratio and % Alumni Donate





In all of the graphs we can see that as the student/faculty ratio becomes better(its value decreases, the y axis quantity increases.

- CONCLUSION,INFERENCES AND USAGE
  1)Student/Faculty Ratio vs. Graduation Rate - A low Student/Faculty Ratio has a very positive impact on the graduation rate.

2)Student/Faculty Ratio and % Alumni Donate - Colleges with a good student/faculty ratio get more donations from alumni, meaning that the alumni feel the importance of there institute in there success.

3)Instructional Expense and Student/Faculty Ratio - The expense does increase as the ratio improves.

Overall, though the expense increases, but the impact on the perception and the graduation rate overloads it. Hence, a good student/faculty ratio seems to be essential for a college's success.

E.[USNEWS]What is the difference in various states of the US as well as the public and private universities in terms of there tuition fee for instate and outstate students?

• PURPOSE

We will plot the ratio of outstate tuition/instate tuition for all the states with public and private universities separately for each state.

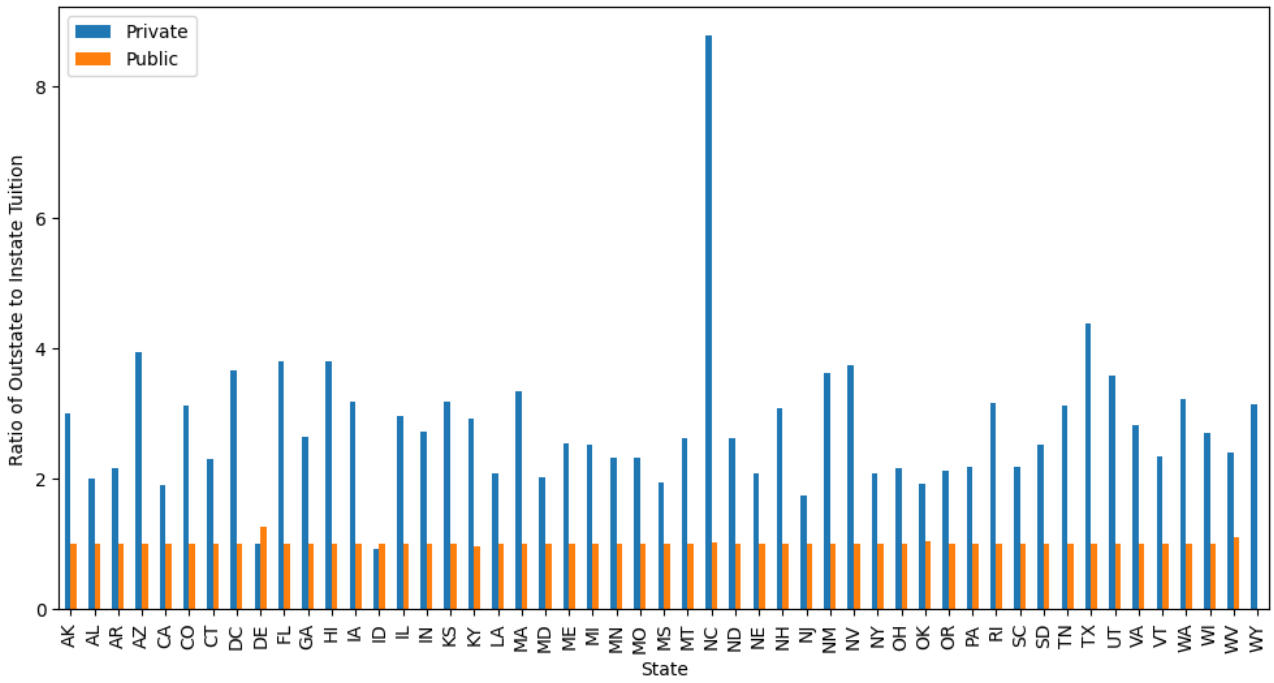
• EXECUTION WITH CODE SNIPPETS

Two bar graphs representing public and private institutes in every state are plotted.

```
ratios = grouped['outstate_tuition', 'instate_tuition'].mean()
ratios['ratio'] = ratios['outstate_tuition'] / ratios['instate_tuition']

ratios = ratios.reset_index().pivot(index='state', columns='public_private', values='ratio')
```

• RESULTS

The following interesting graph is obtained-



• CONCLUSION,INFERENCES AND USAGE

Almost all private institutes in America do not differentiate between instate and outstate students in terms of fee structure.This is very important for students who want to decide a college.

A few states - NC, TX, AZ, HI have very high outstate/instate ratio.

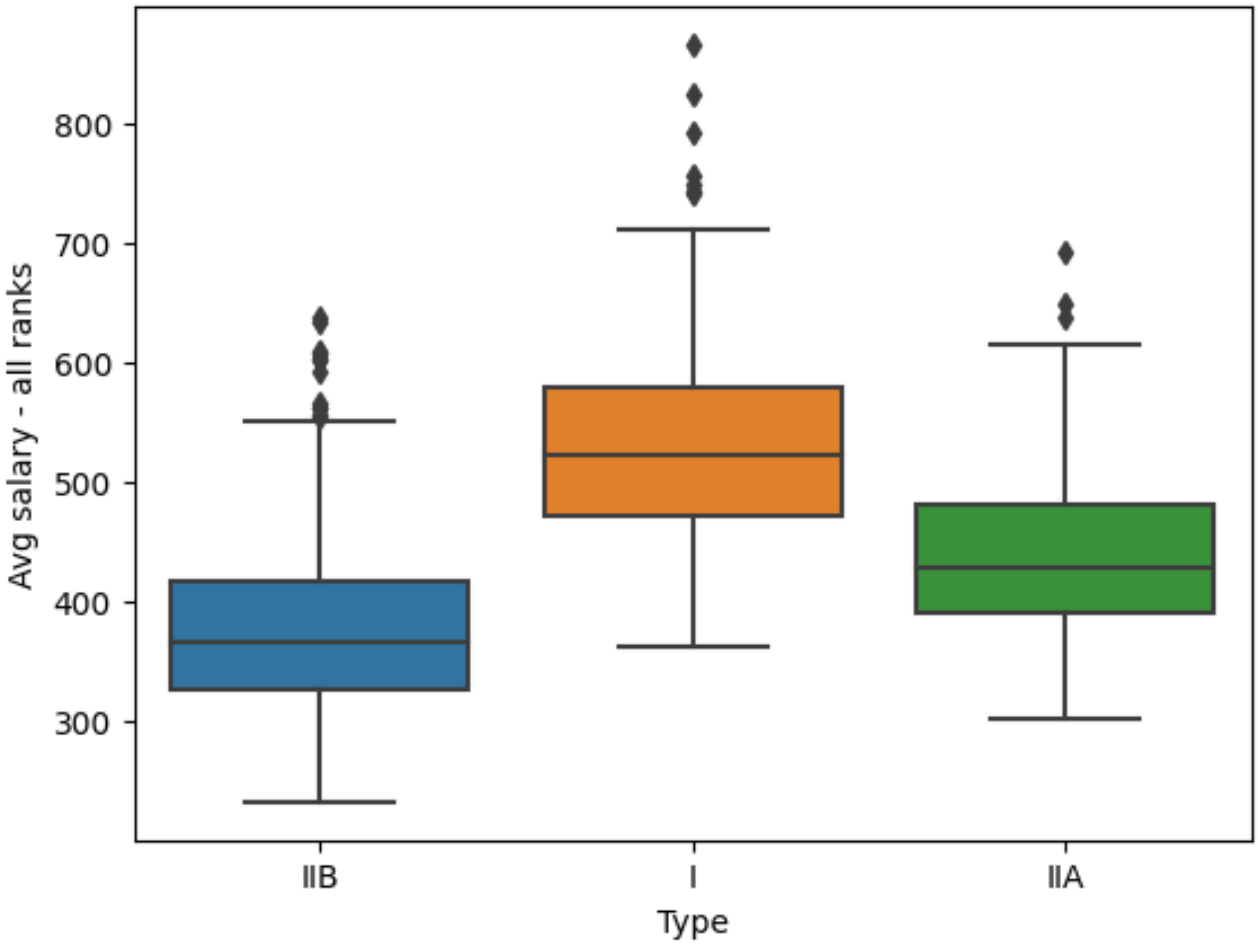
F.[AAUP]How does the distribution of salaries changes among the different type of universities in the US?

• PURPOSE
We will plot the average salaries of professors of all ranks for the university types I, IIA and IIB.

• EXECUTION WITH CODE SNIPPETS

We plot the box plot for each category of university - I, IIA and IIB.

• RESULTS
The following graph is obtained-



• CONCLUSION,INFERENCES AND USAGE

We see that the professor in type I colleges are highest paid followed by IIA and then IIB.

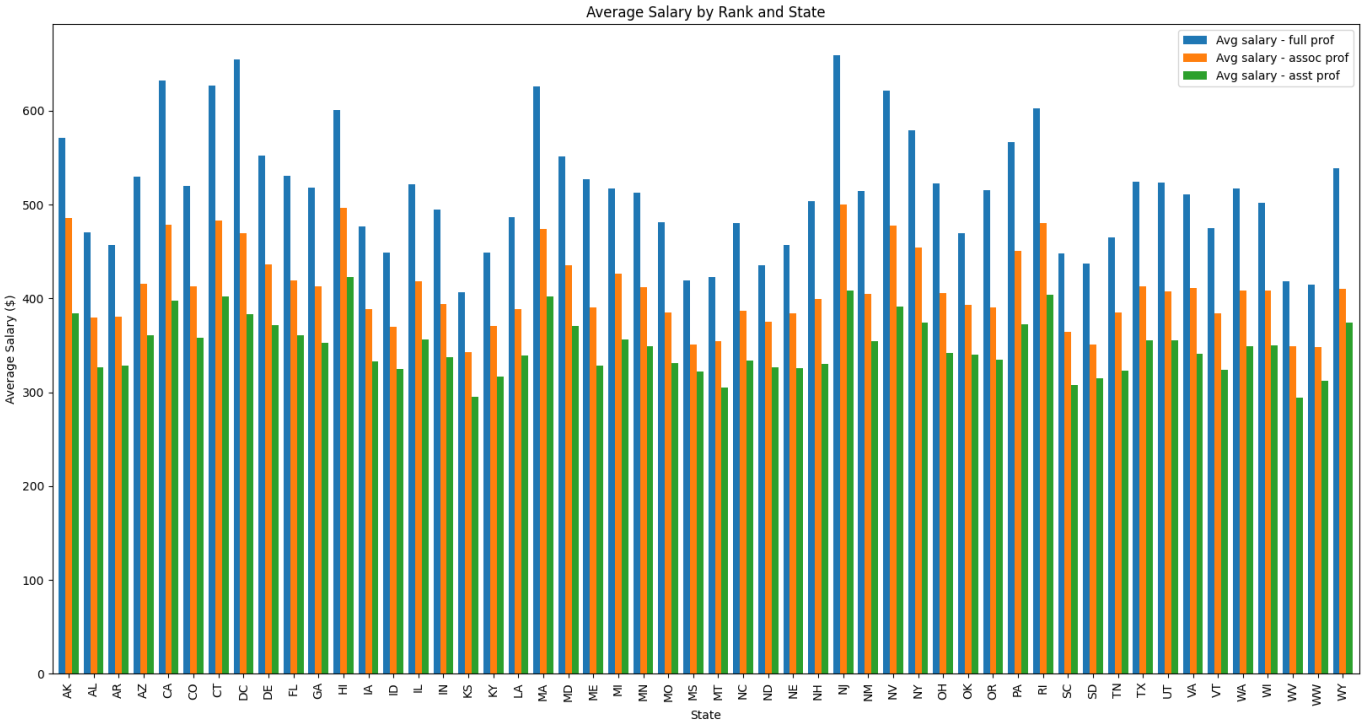
G.[AAUP]What is the distribution of salary in various states by rank?

• PURPOSE
We plot 3 bars for each of the state with 1 bar for each rank of the professors.

• EXECUTION WITH CODE SNIPPETS

```
state_salaries = df.groupby('State').mean()

state_salaries.plot(kind='bar', width=0.8, figsize=(20,10))
plt.title("Average Salary by Rank and State")
```

• RESULTS
The following graph is obtained-



• CONCLUSION,INFERENCES AND USAGE

We see that AZ, CT, DC, MA, NJ, RJ pay the professors very good.

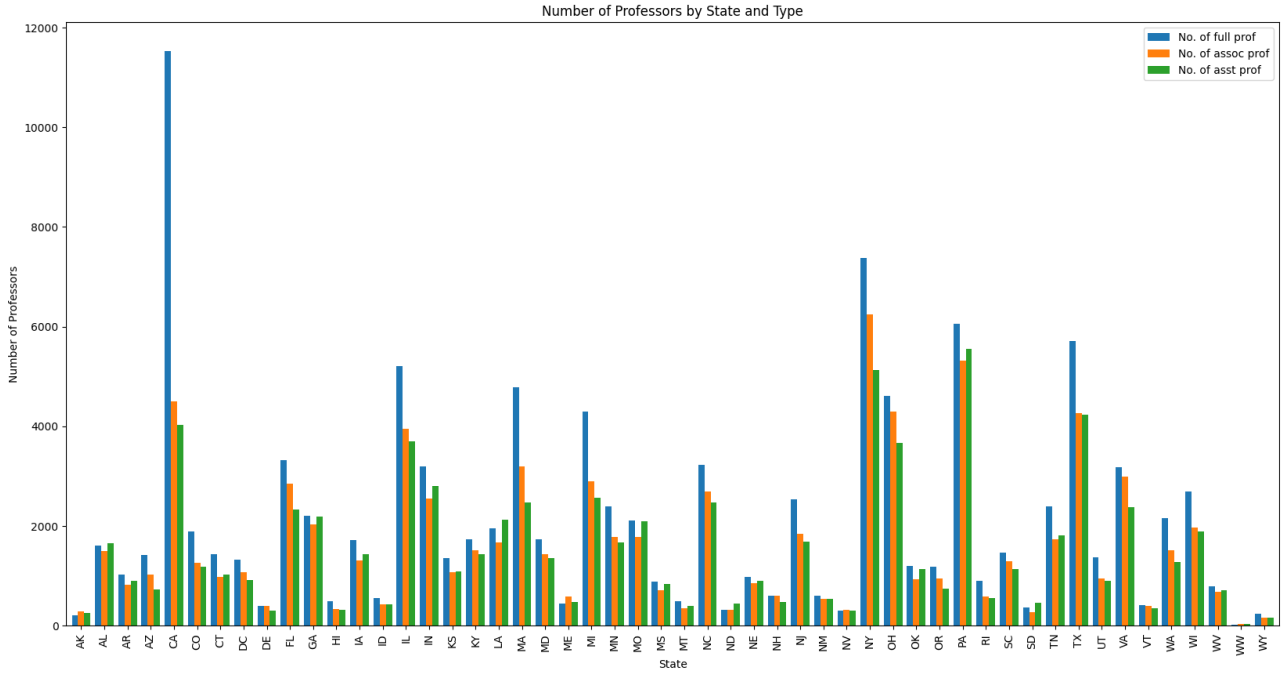
H.[AAUP]How does the distribution of number of professors varies with the type of college across the states.

• PURPOSE
We plot the bar graphs for the number of full, associate and assistant professors for each state.

- EXECUTION WITH CODE SNIPPETS

```
ax = df_prof.plot(kind='bar', figsize=(20,10), width=0.8)

ax.set_title('Number of Professors by State and Type')
ax.set_xlabel('State')
ax.set_ylabel('Number of Professors')
```

- RESULTS
  The following graph is obtained-



- CONCLUSION,INFERENCES AND USAGE

The states of CA, IL, NY, OH, PA, TX have a large pool of professors.

I.[AAUP]Analysing the dependence of compensation paid to professors with the state the college is present in.

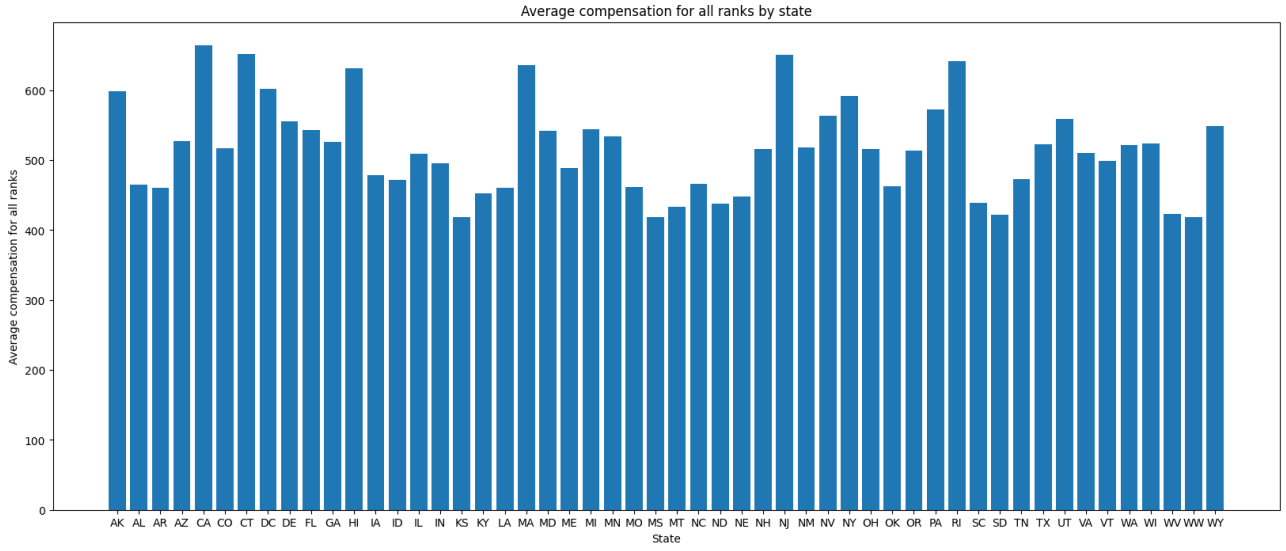
- PURPOSE

We plot the graph showcasing the average compensation paid to the professors statewise.

- EXECUTION WITH CODE SNIPPETS

```
plt.figure(figsize=(20,8))
plt.bar(state_compensation.index, state_compensation.values)
plt.xlabel('State')
plt.ylabel('Average compensation for all ranks')
plt.title('Average compensation for all ranks by state')
plt.show()
```

- RESULTS
  The following graph is obtained-



- CONCLUSION,INFERENCES AND USAGE

We see that the compensation paid to a professor does depend on the state the college is present in.

J.[AAUP]Understanding the impact of the quantity of faculties an institute hires on the relative amount of compensation it is able to pay to its professors.

• PURPOSE

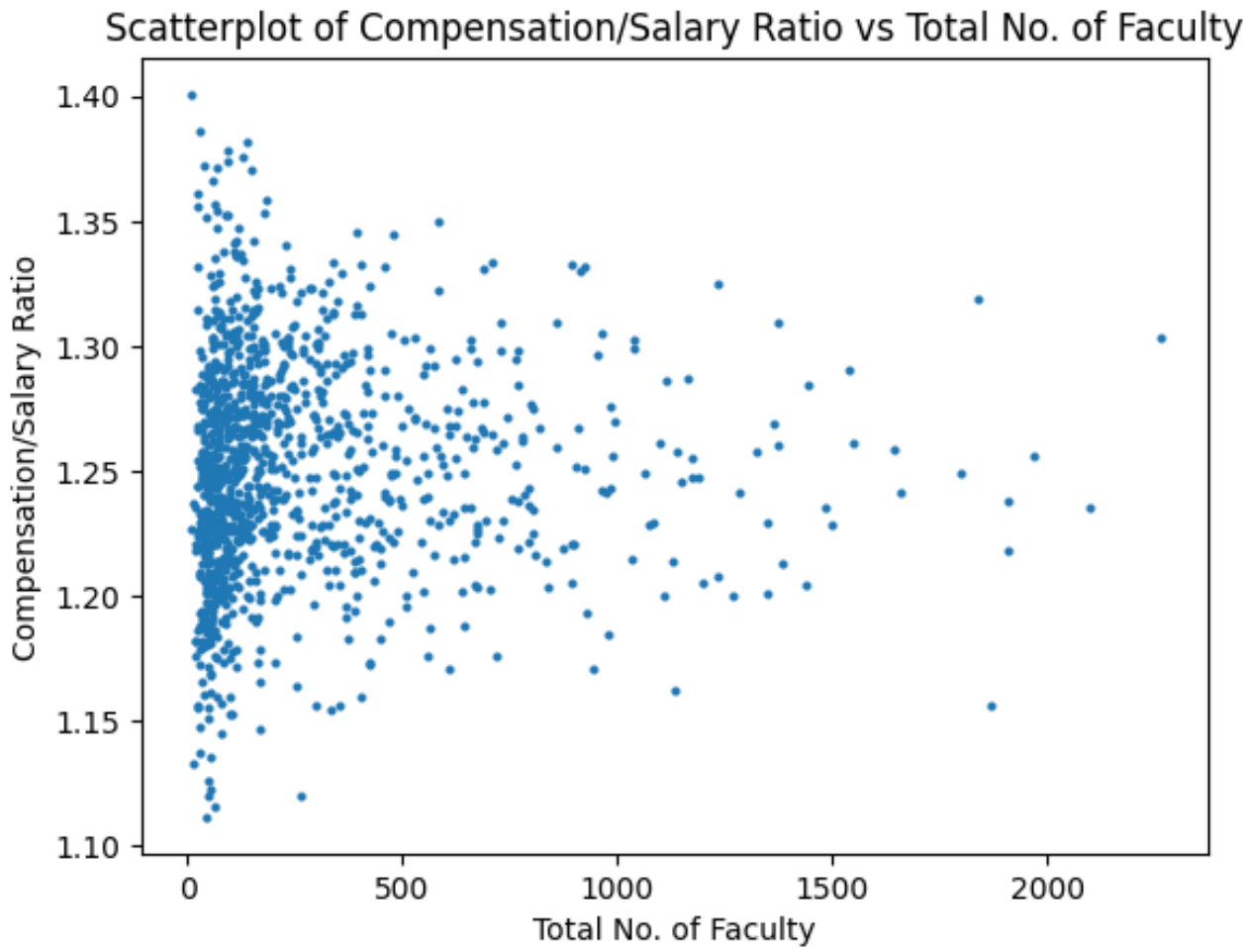
We plot a scatterplot of Compensation/Salary Ratio vs Total No. of Faculty to observe the dependence.

• EXECUTION WITH CODE SNIPPETS

```
import matplotlib.pyplot as plt
aaup["Compensation/Salary Ratio"] = aaup["Avg compensation - all ranks"] / aaup["Avg salary - all ranks"]

plt.scatter(aaup["No. of faculty - all ranks"], aaup["Compensation/Salary Ratio"],s=4)
```

• RESULTS

The following graph is obtained-



• CONCLUSION,INFERENCES AND USAGE

We see that as the number of faculty increases, the ratio tends to become less by a small amount. Hence, we can conclude that having a large pool of professors does impact the relative compensation a college is able to pay to the professors.

V. OVERALL SUMMARY/OBSERVATION

We were able to obtain critical insights into how the American higher education system worked in the mid-90s. We understood the distribution of institutes and expenditure among states, type of institutes in terms of quality and quantity.This data narrative has valuable insights for anyone looking to choose a US college for higher studies.

VII. REFERENCES

[1] http://lib.stat.cmu.edu/datasets/colleges/

[2] https://pandas.pydata.org/

[3] https://matplotlib.org/

VIII. ACKNOWLEDGEMENTS

Prof. Shanmuganathan Raman,IIT Gandhinagar
The team of teaching assistants for the course ES114
The creator of the dataset