Data Narrative – Assignment 2

Parag Sarvoday Sahu
Electrical Engineering
(Roll No. - 22110179)
IIT Gandhinagar
Gandhinagar, Gujarat
parag.sahu@iitgn.ac.in

***Abstract*—To mine the given Datasets containing information on US colleges and report the observations.**

Keywords—python, pandas, data-analysis, aaup, usnews

I. Overview of the Dataset

The description of the two datasets is as follows:

1. USNEWS dataset – This dataset was created for a data analysis exposition held in the year 1995. It contains general information about 1302 US colleges on subjects like SAT scores, fees, no. of applications received and accepted, no. of students, student to faculty ratio etc.
2. AAUP dataset – This dataset was also part of the same 1995 exposition. It contains information on faculty salaries for 1074 US colleges.

II. Procedure employed for analysis

The dataset was first cleaned by removing rows which contained empty cells, merging the columns which contained similar information and converting the numerical information stored as 'str' data-type to 'int' data-type. Thereafter, the dataset was analyzed by asking questions from the user's point of view to retrieve meaningful information.

```
college_prof[lst] = college_prof[lst].apply(pd.to_numeric)
college.drop(college[college["Out-of-state tuition"] == "*"].index,
inplace=True)
college_prof["Avg_salary_full_prof"] = college_prof["Avg_salary_full_prof"] +
college_prof["Avg. compensation (full prof.)"]
```

Code snippet illustrating the cleaning process

III. Scientific Questions/Hypotheses

A. *Which states in USA are more intelectually dense i.e. what is distribution of colleges and universities amongst various US states?*

B. *Hypotheses: Universities that are more selective in their selection process necessarily acquire the brightest students.*

C. *Is it true that the universities which charge high fees from the students, also spend an equally high amount of money in instructing the students?*

D. *Is there a relationship between the graduation rate and the acceptance rate?*

E. *Are there any colleges that strike the ideal balance between the fees charged and the services offered?*

F. *If for a given college, a full professor earns a given amount of money then does it mean that the associate professor earns just as much?*

G. *What all are the colleges where the full-professors rae most generously paid?*

H. *Do these high paying colleges also hire a lot of faculties to get on top of rankings? Is it that they are able to offer these high salaries because they hire less faculties?*

IV. Details of Libraries and Functions

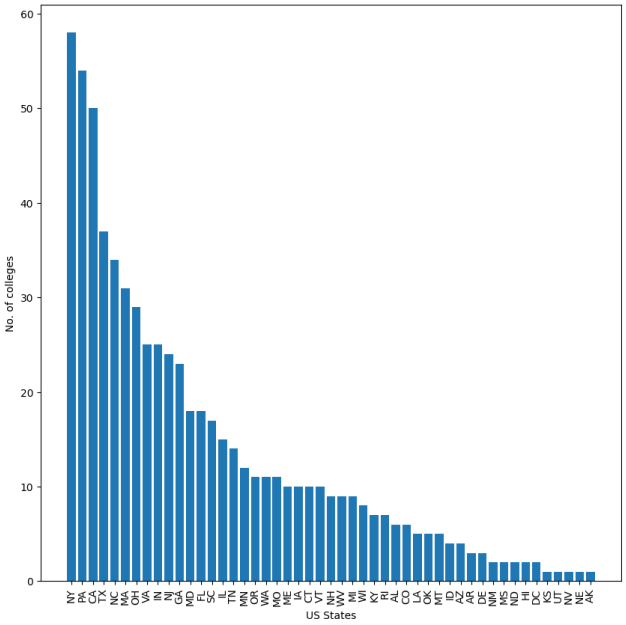
The various libraries used in this analysis are as follows:

1. Pandas – This library is used to process and analyse large datasets.
2. Matplotlib – This library is used to get various types of plots using python language.
3. Numpy – This library is used to perform quick and efficient calculations using arrays, in python.
4. Seaborn – This library is used to get more detailed and modern looking plots than Matplotlib using python language.
5. Plotly – This library is used to make more advanced interactive graphs using python language.
6. Scikit-learn – This library is used to perform predictive data analysis. It was used to find correlation between the data of two columns.

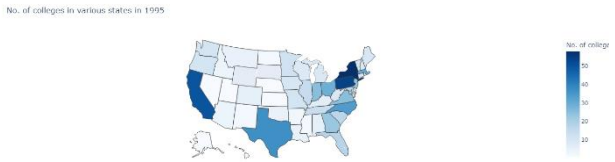
V. Answers to the Questions

A. *State-wise distribution of colleges*

New York state has the greatest number of colleges followed by Pennsylvania and California. It was observed that the number of colleges amongst various states decreases almost exponentially. Also, the states that are deep inside the country have very few colleges as compared to the ones which are located at the borders.



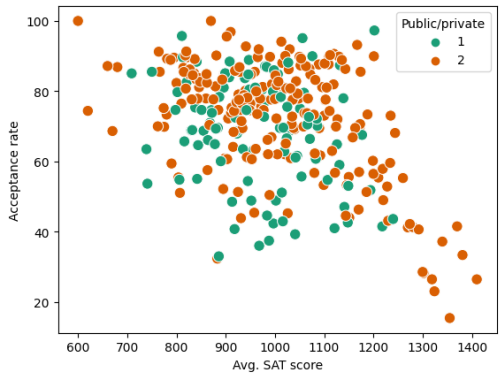
Graph depicting the number of colleges in various states



Pictorial representation of number of colleges in various states

B. *Universities more selective in their selection process acquire the brightest students*

The colleges which have low acceptance rate only admit the brightest students. An interesting observation is that only private universities are highly selective in their selection process and hence only they get to have the best students.

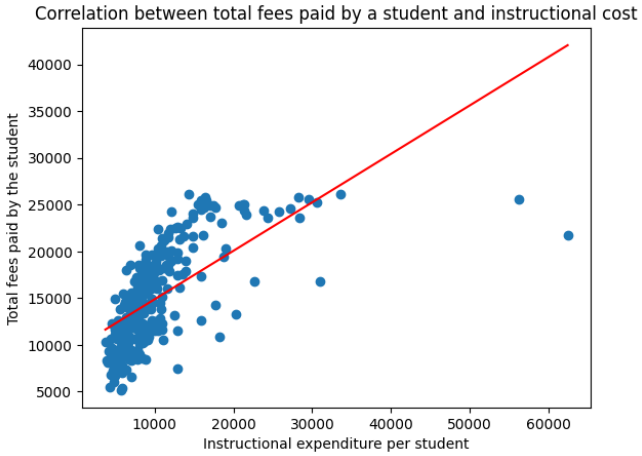


Scatter-plot depicting the relation of acceptance rate with average SAT score.

Assumption: The quality of the students is gauged using the average SAT score though it may not be the absolute measure it.

C. *Relatioship between instructional expenditure per student and the total fees paid by the student*

- There exists a positive correlation (~0.64) between the fees paid by the students and the value services received from the universities. But some universities deliver services worth more than was the student had paid for. One reason for this behaviour can be the generous donations received by the universities.



A scatter plot depicting the relationship between the instructional cost and fees paid by the student.

```
# plotting the data
plt.scatter(df1["Instructional expenditure per student"], df1["Total ex-
penses"])

# This will fit the best line into the graph
plt.plot(np.unique(df1["Instructional expenditure per student"]),
np.poly1d(np.polyfit(df1["Instructional expenditure per student"], df1["Total
expenses"], 1))
         (np.unique(df1["Instructional expenditure per student"])),
color='red')

plt.title('Correlation between total fees paid by a student and instructional
cost')
plt.xlabel('Instructional expenditure per student')
plt.ylabel('Total fees paid by the student')
```

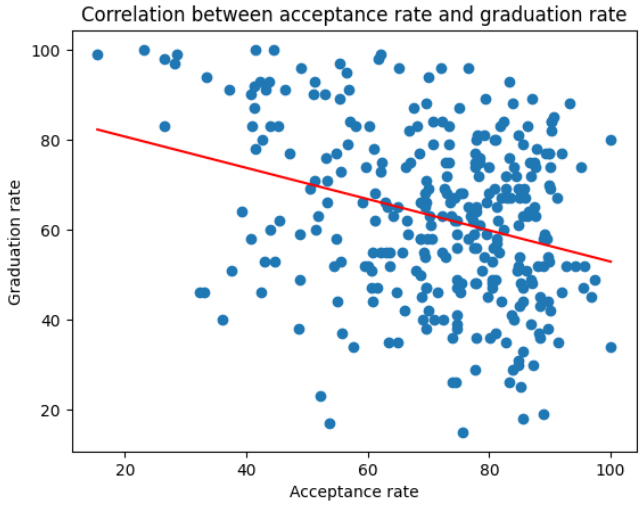
Code used to plot the graph

Assumption: The total fees include the fees paid to the college as well as the boarding costs and personal expenditures.

D. *Relationship between graduation rate and acceptance rate*

The graduation and acceptance rates are negatively corelated (~0.22) which means lower the acceptance rate, higher the graduation. The universities which are more selective in their admission process are successful in picking out the right talent. These students sustain their performance which results in high graduation rate. This

result can be summarised very well as ‘pick wisely, reap profusely’.



Scatter plot depicting the relation between graduation and acceptance rate for various colleges

E. *Colleges that stike an ideal balance between fees charged and services offered*

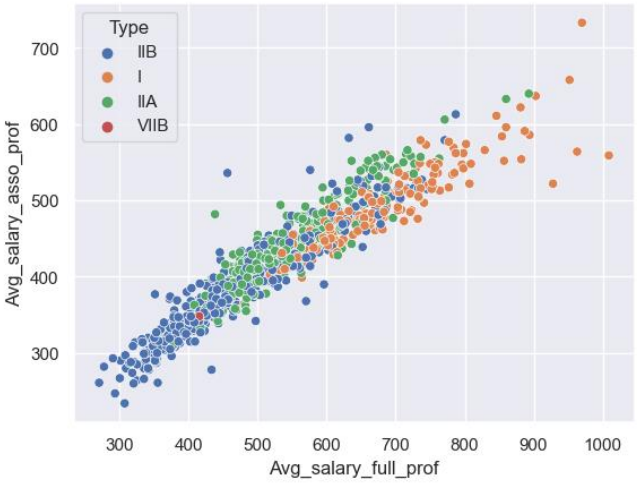
Various parameters were kept above the mean their value and the following colleges were found to meet the criterions:

S. No.	College Name	Total expenses ($)	Acceptance rate	Average SAT score	Instructional expenditure per student	Percentage of faculty with PhD	Student to faculty ratio
1	Alaska Pacific University	11680	75.6	972	10922	76	11.9
2	Hendrix College	11995	87.6	1089	8588	82	13.1
3	University of California at Riverside	13129	77.2	982	12528	98	13.4
4	LaGrange College	10572	73.3	780	9067	77	12.5
5	Piedmont College	9260	84.7	951	7309	89	13.2
6	University of Hawaii at Manoa	7498	72.7	978	12833	85	11.8
7	Huntington College	13870	95.5	902	9466	76	11.8
8	University of Missouri at Columbia	12542	70.5	1066	10145	87	12.7
9	North Carolina Wesleyan College	12472	84.8	826	10090	77	12.7
10	University of Utah	10832	88.1	1027	9275	89	12.8
11	Virginia Military Institute	12820	75.5	1025	10709	84	11.1
12	Christendom College	12330	88.8	1136	10922	92	9.3

Table containing the list of colleges that strike an ideal balance between the fees charged and the services offered

F. *Relation between salaries of full-professors and associate-professors for a given college*

The salaries of associate professors do increase as the salaries of full-professors increases. But given the hierocracy of the two posts, the salary of an associate professor is always on the lower side. An interesting observation is that professors earn the most in type-I colleges followed by type-IIA and type-IIB respectively.

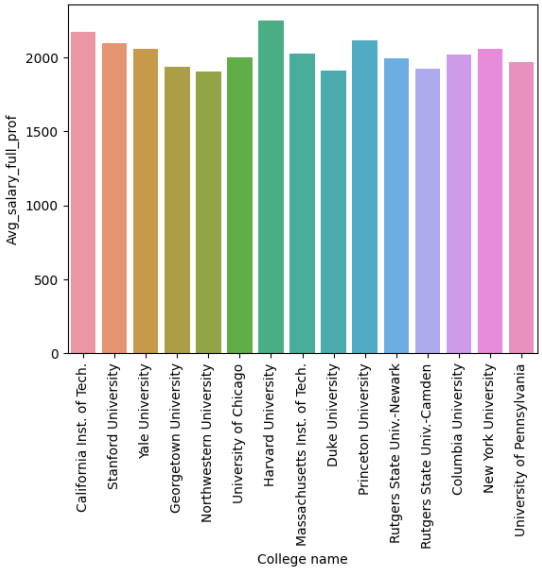


Relation between salaries of full-professors and associate-professors for a given college

Assumption: Since the compensation part includes research grants, extra bonuses etc., the average compensation was added to their respective average salaries before analysis.

G. *Colleges where full-professors enjoy a good salary*

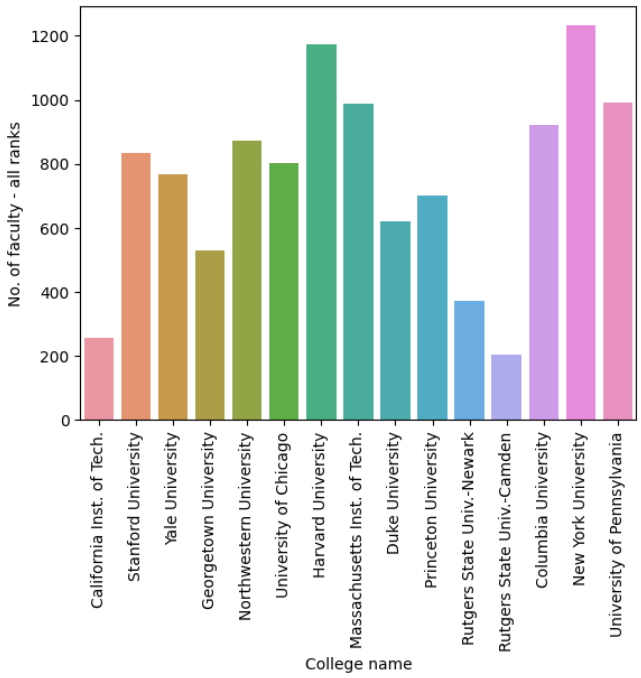
The colleges which pay a high salary to its full-professors are the ones which score well in various rankings. It shows that these colleges have put in the required amount of funds to get to the place they are at.



Bar graph depicting the salaries of full-professors in top colleges

H. No. of faculty in the highly paying colleges

It cannot be said that these colleges hire more or less faculties just by looking at the salaries because there is no trend observed here. The reason for this situation can be the different number of enrolled students in these universities.



Bar graph depicting the total number of faculties in the highly paying universities

VI. Summary of the Observations

The private colleges in USA enjoy a clear advantage over the public colleges in the terms of the quality of students and the resources available. Hence, they are also able to charge a significantly higher fees than their public counterparts. This shows the divide in economical divide that exists in a capitalist country like USA.

Acknowledgment

I would like thank Prof. Shanmuganathan Raman for providing us with this opportunity of experiencing Data Analysis in our 1st year of BTech program. I also thank the creators of all the free resources available on the internet.

References

- Pandas user guide .March 25, 2023. https://pandas.pydata.org/docs/user_guide/index.html
- Matplotlib user guide. March 25, 2023

  https://matplotlib.org/stable/users/index.html
- Seaborn user guide. March 26, 2023. https://seaborn.pydata.org/tutorial.html
- Data Analysis with Python Course. March 26, 2023.

  https://jovian.com/learn/data-analysis-with-python-zero-to-pandas