ES114 Data Narrative 2

Pradeep Kumar Meena (22110196)
Chemical Engineering,
IIT Gandhinagar

I. Overview Of The Dataset

The aaup and the usnews data set contains information on faculty salaries and compensation for various institutions of higher education in the United States. The dataset contains the data of about 1160 institutes of the US. The columns provide information on the state where the institution is located, the type of institution (public or private), and the average salary and compensation for full professors, associate professors, and assistant professors. Additionally, the data includes information on the number of full professors, associate professors, and assistant professors, as well as the number of instructors and faculty members overall. This data set could be used to explore various questions related to faculty salaries, compensation, and staffing levels at institutions of higher education in the United States.

The dataset usnews contains data about the colleges and universities of the US. The data includes variables such as the name of the institution, its public or private status, the number of applications and acceptances received in various exams, and other metrics related to academic quality and faculty resources. This data set is a valuable resource for exploring trends and characteristics in higher education in the United States.

II. Scientific Questions or Hypotheses

A. *Is there a relationship between the average compensation of full professors and the average compensation of all faculty ranks combined?*

B. *Does the average salary of faculty members differ between institutions that offer primarily undergraduate degrees versus those that offer graduate and doctoral degrees?*

C. *Which state has the maximum and least number of universities and colleges?*

D. *Is there a difference in salaries and compensation between different states of the United States?*

E. *Do institutions with a higher number of faculty members tend to have higher numbers of instructors as well?*

F. *Do private institutes have higher graduation rates?*

G. *Do states which have higher graduation rates have higher Instructional expenditure per student?*

H. *Which university has the maximum number of applications received?*

I. *Do colleges with higher estimated book costs tend to have higher estimated personal spending?*

J. *What is the probability that a randomly selected institute in the USA has a graduation rate above 80 percent?*

II. Details Of Libraries And Functions

I have used these functions and python libraries in my assignment.

A. *Libraries:*

- Pandas is an open-source library that is made mainly for working with relational or labelled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.[1]
- Matplotlib : Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.[2]

B. *Functions:*

- pandas.read_csv(): To read and store data from csv file to a Pandas dataframe. [1]
- matplotlib.pyplot.scatter(): Makes a scatter plot of x versus y. [3]
- x.corr(y): Correlation between x and y.
- values_count(): It is used to get the occurrence of any list of strings for the top element of a series object.
- hist(): plots the histogram of a Series
- plot(): It is used to plot the graph this function is plt.xlabel() and plt.ylabel() functions are used to label available in both pandas and matplotlib.

bar(): It is the function available in matplotlib to plot bar graph.

III. Answers Of The Questions

A. Is there a relationship between the average compensation of full professors and the average compensation of all faculty ranks combined?

Approach: By reading the file "aaup.csv", plotting the scatter plot and analysing the plot of the average compensation of full professors and the average compensation of all faculty ranks combined. I can find the relationship between the two things.

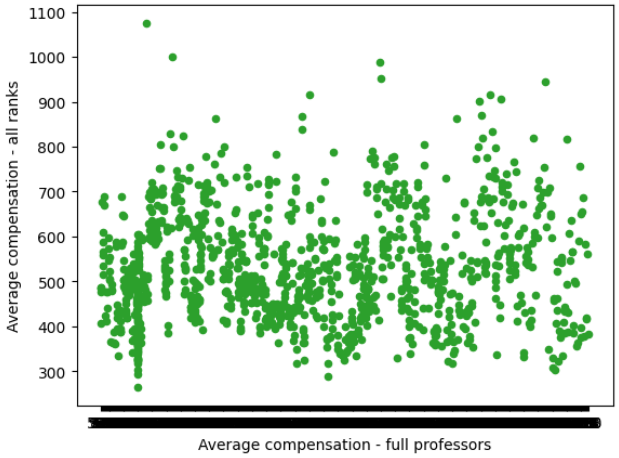

Fig.1 Average compensation - all marks vs Average compensation – full marks

By analysing the Fig1 scatter plot, there is a positive relationship between the average compensation of full professors and the average compensation of all faculty ranks combined. We would see the points on the scatter plot clustered in a roughly linear pattern at the starting sloping upwards from left to right, meaning that as the average compensation of full professors increases correspondingly, the average compensation of all faculty ranks combined is less clustered, which shows that there is a difference in the compensation of the average compensation of full professors and the average compensation of all faculty ranks combined in different universities or colleges.

B. Does the average salary of faculty members differ between institutions that offer primarily undergraduate degrees versus those that offer graduate and doctoral degrees?

Approach: I assume that the type of institutes IIA, IIB, and VIIB are the non-research institutions and type "I" is the research institute and plot the graph of the Average salary of all rank (Mean salary) of Professors vs the type of all four institutes.

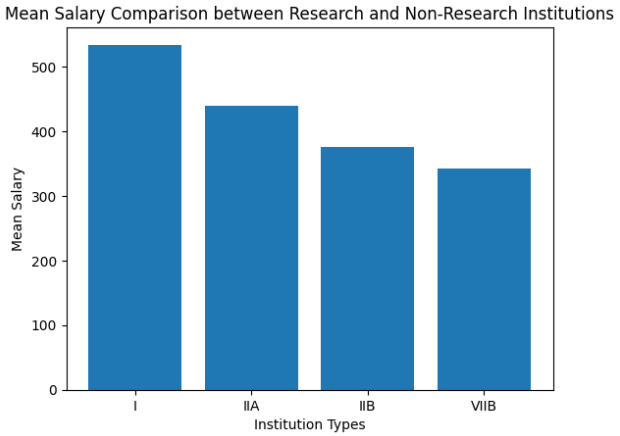

Fig.2 Mean salary vs Institute types

Fig.2 shows four bars, for each type of institution, with the height of the bar indicating the mean salary. I note that the average salary at Research institutions(I) is higher than that of Non-Research institutions(IIA, IIB, and VIIB), I would expect the Research institution bar to be taller compared to the Non-Research institution bars. So, in conclusion we can say that the salary of professors of research institutes is higher than the professors of non-research institutes.

C. Which state has the maximum and least number of universities and colleges?

Approach: Calculate the number of universities for each state and plot a line plot.

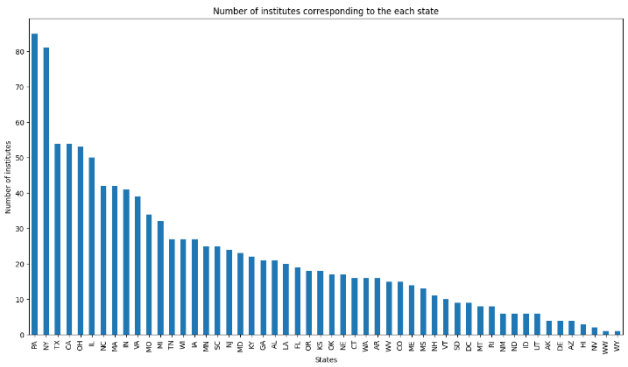

Fig.3. Number of institutes corresponding to each state

By looking at the Fig.3, we can easily say that the state PA(Pennsylvania) has 85 institutes which is the maximum number of institutes and the state WY (Wyoming) has only one institute.

D. Is there a difference in salaries and compensation between different states of the United States?

Approach: Calculate the mean salaries and compensations for each state by grouping the mean of the "Average salary - all ranks" and "Average compensation - all ranks columns" and plot the scatter plot.

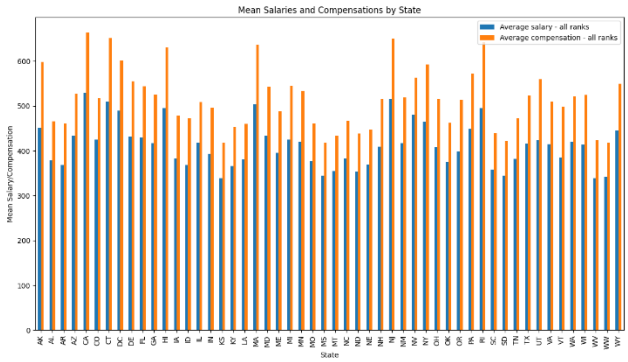

Fig.4 Bar graph of mean salaries and compensations by state

Fig.4 shows two bars in the above plots for each one for the mean salary and one for the mean compensation. By analysing the plot, we can conclude that in general there is no large differences in the mean compensations and salary in the different states, but there are some states where this difference is about 20k dollars for example:- In CA(California) the mean compensation and the mean salary is about 70k dollars and 55k dollars respectively while in AL(Alabama) the mean compensation and salary is only about 45k and 37.5k which shows that there is large difference in the average compensation and salary in the different states of the United States.

E. Do institutions with a higher number of faculty members tend to have higher numbers of instructors as well?

Approach: We have given the data of the number of faculty members and number of the instructors just for the different institutions. We just directly plot the scatter plot and observe the trend.

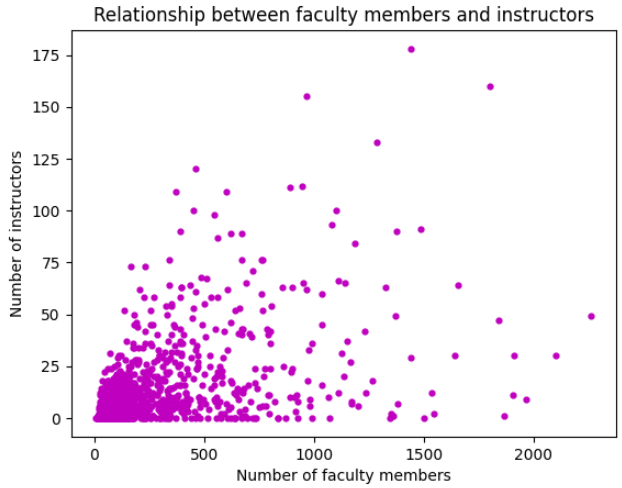

Fig.5 Relationship between faculty members and instructors

Fig.5 clearly shows that institutions which have higher number of faculty members have a higher number of instructors. Well It depends on various factors. If the size of the institute is large, they will require more faculty members and instructors to support their programs.

F. Do institutions with a higher number of faculty members tend to have higher numbers of instructors as well?

Approach: first calculate those institutes which are private or public and store them in two variables and then find the graduation rate for both private and public institutes. And plot the line plot for Graduation rate for both private and public institutes.

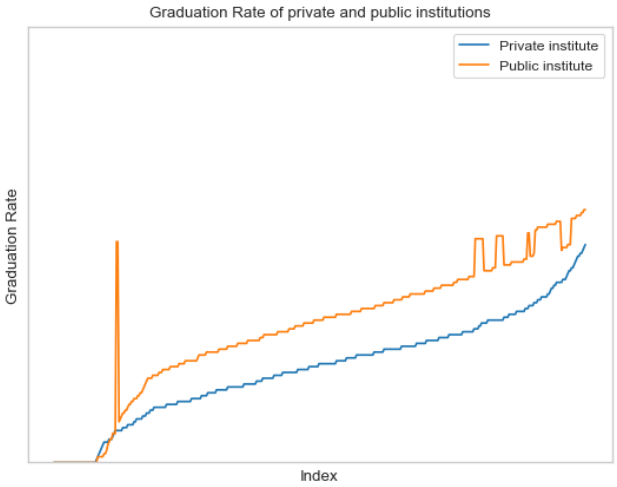

Fig.6 Graduation rate of public and private institutions

By looking at the Fig.6 graph we can conclude that my hypothesis is wrong. Instead, I found that the Graduation rate for public institutes is more than the graduation rate of private institutes. By looking at the plot, can also say that in general the graduation rate of public institute is approx double of the graduation rate of the private institutes.

G. Do states which have higher graduation rates have higher Instructional expenditure per student?

Approach: Plot the scatter plot between the Graduation rate and Instructional expenditure per student.

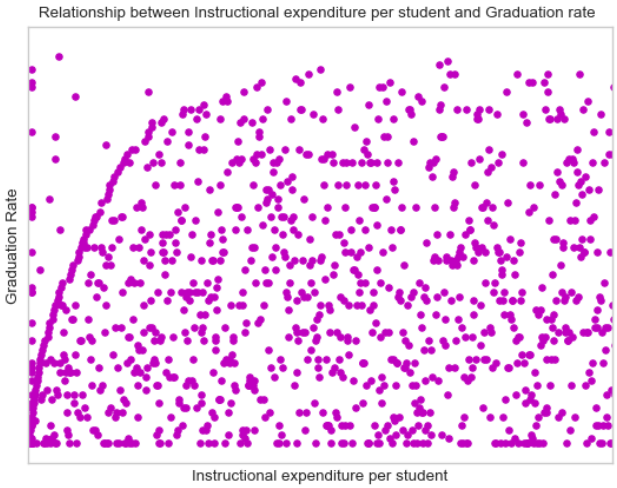

Fig.7 Relationship between instructional expenditure per student and Graduation rate

By analysing the Fig.7 graph we can say that as the instructional expenditure per student increases the

graduation rate also increases for a certain limit but after a certain point as the instructional expenditure per student increases change in the Graduation rate is distributed all over the range. So we can conclude that after a certain point the graduation rate no longer depends on the instructional expenditure per student.

H. Which university has the maximum number of applications received?

Approach: To find out the university with the maximum number of applications received from the given dataset of us news, we can sort the dataframe in descending order based on the column "Number of applications received" and select the top row.

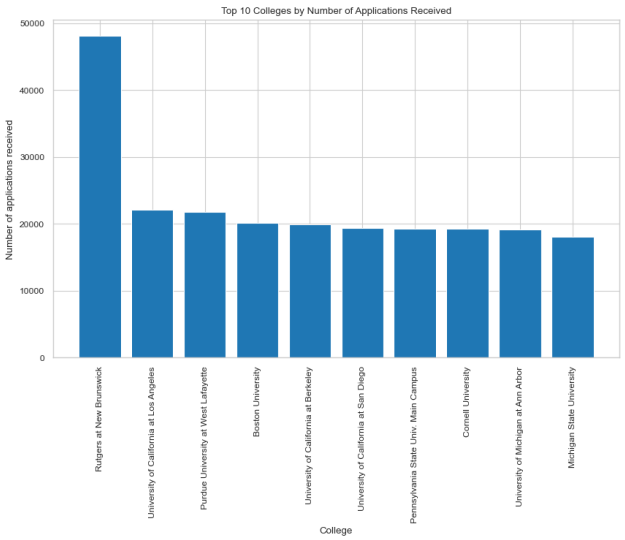

Fig.8 Top 10 institutes by number of Applications received

By analysing Fig.8 we can see that the university Rutgers at New Brunswick is located in the state of New Jersey has the highest number of applications received among all the universities which is about 48000. In comparison to other institutes this institute has a very high number of applications received so we can say that this institute is the most famous institute in the USA and has high demand.

I. Do colleges with higher estimated book costs tend to have higher estimated personal spending?

Approach:
We can check the above hypothesis is wrong or right by simply using the scatter plot between the Estimated personal spending and estimated book costs for different universities.

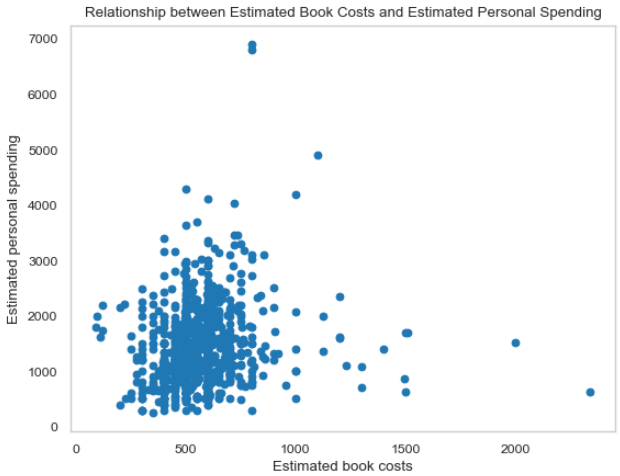

Fig.9 Estimated book costs and estimated personal spending

Fig.9 shows the relationship between the Estimated personal spending and estimated book costs. By analysing the Fig.9 we can see that our hypothesis is wrong that institutes with higher estimated book costs tend to have higher estimated personal spending. In the graph we can see that there is no such relationship between these data. Instead, the above plot shows that the average estimated personal spending is about 2000$.

J. What is the probability that a randomly selected institute in the USA has a graduation rate above 80 percent?

Approach: First calculate the proportion of colleges in the dataset with a graduation rate above 80 percent, which will give us an estimate of the probability of selecting a college with such a high graduation rate. We can then use a bar plot to visualise the distribution of graduation rates in the dataset and highlight the proportion of colleges with a graduation rate above 80 percent.

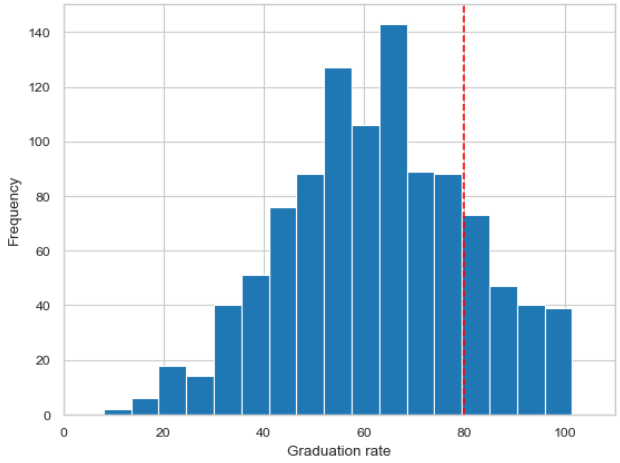

Fig.10 Frequency vs Graduation rate

Probability = Favourable outcomes / Total outcomes

By using the dataset I found that the probability of a randomly selected college in the USA having a graduation rate above 80% is 0.175.

Fig.10 shows that the frequency of the Graduation rate for different universities in the USA. and the bins right to the red line shows the frequency of the graduation rate greater than 80 percent.

V. Summary Of Observations

- It is observed that institute with higher graduation rate is located in developed states.
- It is found that institutes have less graduation rate.
- The graduation rate of public institutes is higher than the private institutes

VI. References

[1] "Introduction to pandas in Python," *GeeksforGeeks*, 09-Feb-2023. [Online]. Available: https://www.geeksforgeeks.org/introduction-to-pandas-in-python/.

[2] *Matplotlib bars*. [Online]. Available: https://www.w3schools.com/python/matplotlib_bars.asp.

[3] *Matplotlib bars*. [Online]. Available: https://www.w3schools.com/python/matplotlib_bars.asp.

VII. Acknowledgements

Iwant to say thank you to statlib@lib.stat.cmu.edu for conducting such a survey and collect the data from all the universities of the USA. I am also grateful to Prof. Shanmuga Nathan for giving the opportunity to work on this dataset. I am also thankful to Mrs Seema for helping me throughout this project.