Data Narrative -2

ES-114

Name : Aryan Sahu

Roll no. : 22110038

I. Overview of the Dataset

The given data is similar to an online catalogue for different kinds of colleges and schools. It contains details of numerous universities and specifications of the respective institutes. This dataset is helpful in guiding students to decide what institute they want to enrol in.

II. Scientific Questions/Hypotheses

We can understand many inferences by analysing the dataset. Following are some valuable inferences we would want to get from the given data:

A. *It is believed that more faculty results in more graduation. Check for the statement's correctness using the given data.*

B. *Can working at an educational institution be considered to be a well-paying job? Find out using the average salaries of workers in such institutions.*

C. *Using this data, find states with high average salaries in the institutes. What does this indicate about these states?*

D. *What is the average workforce division inside the institutes, and what is the inference we get from this?*

E. *What is the relation between the average salary and average compensation provided to the professors? What does the data indicate?*

F. *Highly reputed colleges don't differentiate between students from their state or other states. Show some evidence to support this argument.*

G. *What are the average expenses of a student studying in US institutes? Plot a graph to support your answer.*

H. *Throw light on the difference in monetary requirements of studying in a private college and a public college.*

I. *Colleges with higher math SAT scores tend to have higher verbal SAT scores as well, and vice versa. Prove this using the given dataset.*

J. *Find the colleges having the highest donation culture in their alumni. What does this indicate about these institutes?*

III. Important Libraries and Functions Used

Following are some of the libraries and functions used to mine the dataset effectively and thoroughly:

A. *Pandas*

Pandas is a library offered by the python programming language, which helps analyse tabular data. It offers to create data types consisting of rows and columns and helps analyse large amounts of data.

B. *Read CSV*

In the Pandas library, we are using the read_csv function to import the CSV file to read the data from and analyse it for our benefit.

C. *Dataframe*

We are creating dataframes as a medium to represent the data inside the python program. Dataframes provide various in-built functions which are useful to derive important inferences from the given dataset.

D. *Replace and Dropna function*

We have used the Replace function to replace the empty entries, represented by '*' in our data, with the NA entity. We further used the Dropna function to drop all unrequired blank data for our convenience.

E. *Matplotlib*

It is a library that offers plotting methods and data visualisation options. We have used this library to plot the diverse data and visualise it with the help of graphs.

IV. Solutions to the Questions

We are using the above-specified tools to mine the dataset and obtain various inferences. Following are the answers got from this process.

A. It is believed that more faculty results in more graduation. Check for the statement's correctness using the given data.

We can plot the data of the number of faculty and the corresponding data of graduation rate in that institution. This would give us an understanding of the statement's validity.

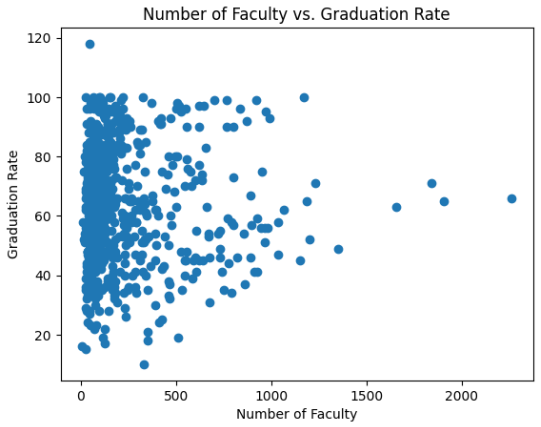

Fig. 1 Scatter plot of the Number of faculty vs Graduation rate.

From this data, we can observe that all sorts of graduation rates are present equally for any number of faculty. It is not the case that the colleges with more faculty have high graduation rates.

B. Can working at an educational institution be considered to be a well-paying job? Find out using the average salaries of workers in such institutions.

We plot the data of the various institutions about the average salaries and their frequencies to figure out the most probable salary an average person could get by working at an educational institution.

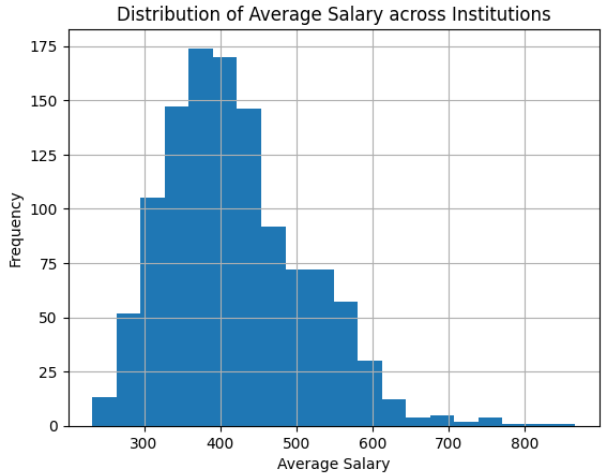

Fig. 2 Average salary of workers (in $100s) vs their frequency across various institutions.

We can see that the most frequent salary in an institution is around $40,000. This is considerably less and insufficient to fulfil the needs of an average person living in the urban areas where these institutions are built. Thus the workers working in most institutions don't get paid enough. This could indicate that the ones earning the most profit through the institutions are the owners and not the employees.

C. Using this data, find states with high average salaries in the institutes. What does this indicate about these states?

The following plot displays the number of institutes present in each state having an average salary greater than $60,000.

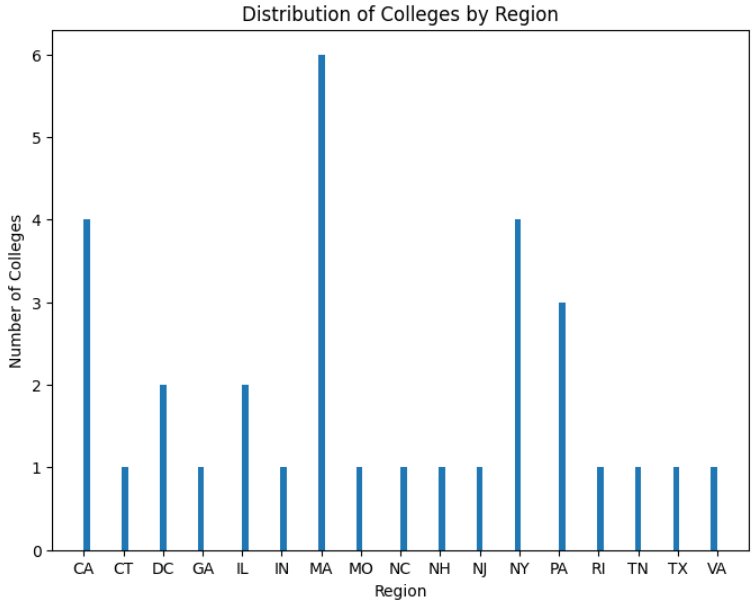

Fig. 3 No. of institutes present in each state having average salaries greater than $60,000.

As we can observe, the state with code MA has 6 institutes having average salaries greater than $60,000. This indicates that these states have a high standard of living, and it is expensive to make a living in these states.

D. What is the average workforce division inside the institutes, and what is the inference we get from this?

We can plot the average numbers of different kinds of professors and instructors.

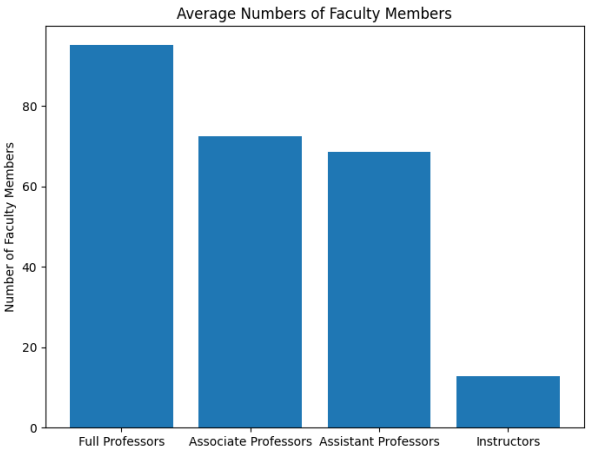

Fig. 4 An average division of the workforce inside the institutes.

We can see that the number of full professors is the most while the number of associate and assistant professors is approximately the same. But the number of instructors is the least. This indicates which posts are more important and play an important role in maintaining that institute. Also, along with full professors, the other professors combined are required to support all the professors, and thus there are almost two other professors for each full professor.

E. What is the relation between the average salary and average compensation provided to the professors? What does the data indicate?

We can plot the average salaries vs average compensation as follows:

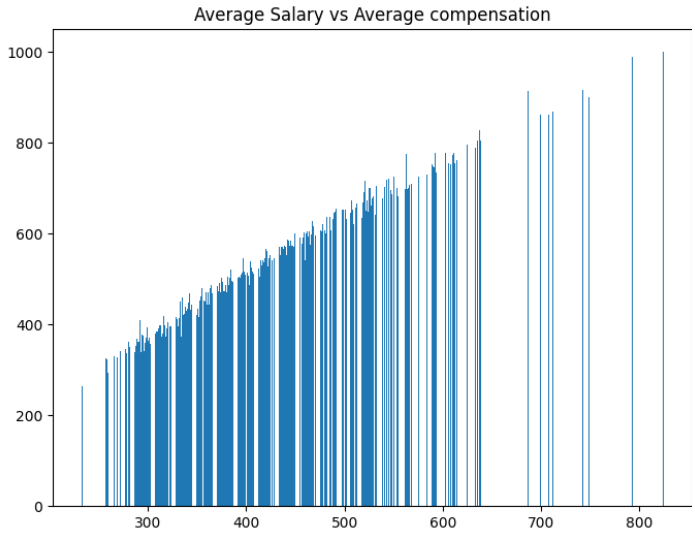

Fig. 5 Average salary($100s) vs Average compensation($100s)

Observing the graph, we can interpret that there is a linear relation between the average salary and compensation received by professors. This indicates that even though the institutes provide high salaries to someone, they require equally high compensation to match the living expenses in that area. The same is true for lower salaries. This means that even though they have a high salary, they might still have the same standard of living.

F. Highly reputed colleges don't differentiate between students from their state or other states. Show some evidence to support this argument.

We can plot a graph of in-state and out-of-state tuition fees demanded at those reputed institutes to indicate this.

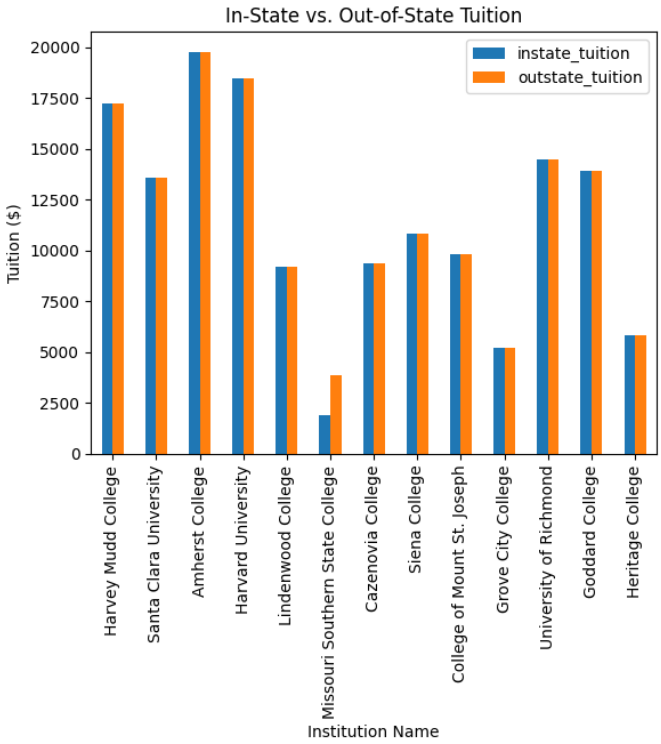

Fig. 6 Institutes with high graduation rates and in and out-state tuition.

These are the institutes having graduation rates higher than 99%. As we can observe, the in-state and out-of-state tuitions in these institutes are exactly the same for most of them. This proves that reputed colleges and schools do not differentiate based on students' states.

G. What are the average expenses of a student studying in US institutes? Plot a graph to support your answer.

We can plot the total expenses of a student in the institutes, including tuition fees, book costs, room costs, additional fees, and personal expenses, of each institute.

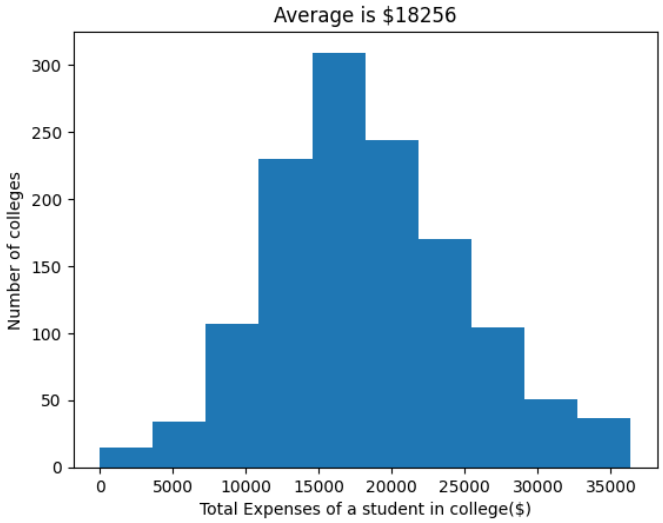

Fig. 7 Expenses of a student studying in US colleges

We can see that the highest peak is between $15000 and $20000. From this data, we get the average total expenses to be nearly $18000. This shows that supporting a student in college would require a major portion of their salaries.

H. Throw light on the difference in monetary requirements of studying in a private college and a public college.

We plot the separate data of expenses in a private and public college from the given data.

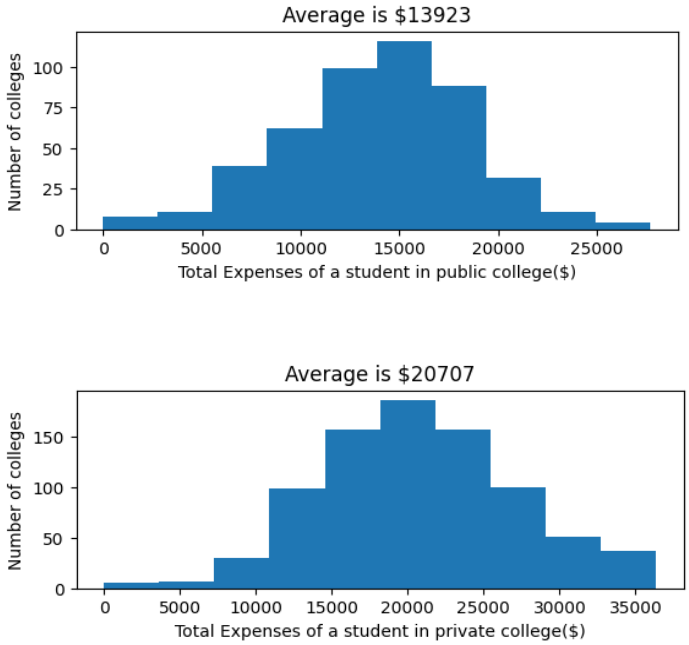

Fig. 8 Total expenses of students in different kinds of institutes, namely public and private.

We can clearly see that there is a significant rise in the students' expenses as we shift from public to private institutes. The average requirement for public colleges is nearly $13000, while it is around $20000 for private ones. This is because government institutes provide various subsidies to students, which lower the fee requirements of students and thus create the above-seen difference.

I. Colleges with higher math SAT scores tend to have higher verbal SAT scores as well, and vice versa. Prove this using the given dataset.

We plot the graph of average verbal SAT scores vs math SAT scores for each college.

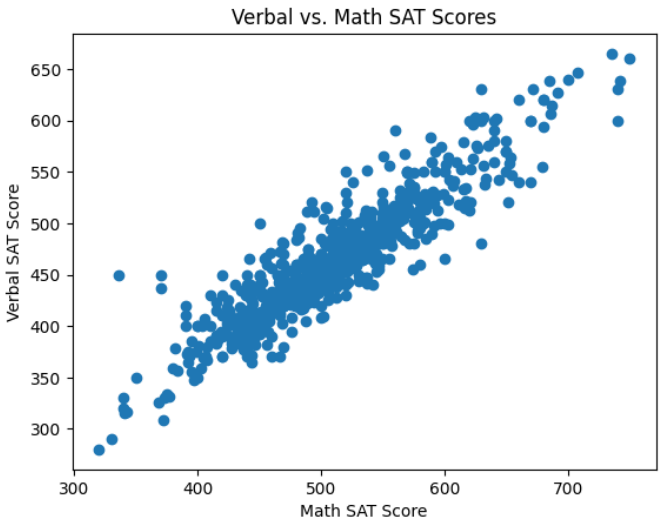

Fig. 9 Marh and Verbal SAT scores of different institutes.

We can observe that the graph is very close to being linear. This indicates that the colleges having high math SAT scores also have high verbal SAT scores, and vice-versa. The scatter plot is dense around the centre, emphasising that most colleges have an average SAT score of around 500-600; for both math and verbal SAT.

J. Find the colleges having the highest donation culture in their alumni. What does this indicate about these institutes?

We can find this by plotting the top ten colleges having a high percentage of alumni who donate to their institutes.

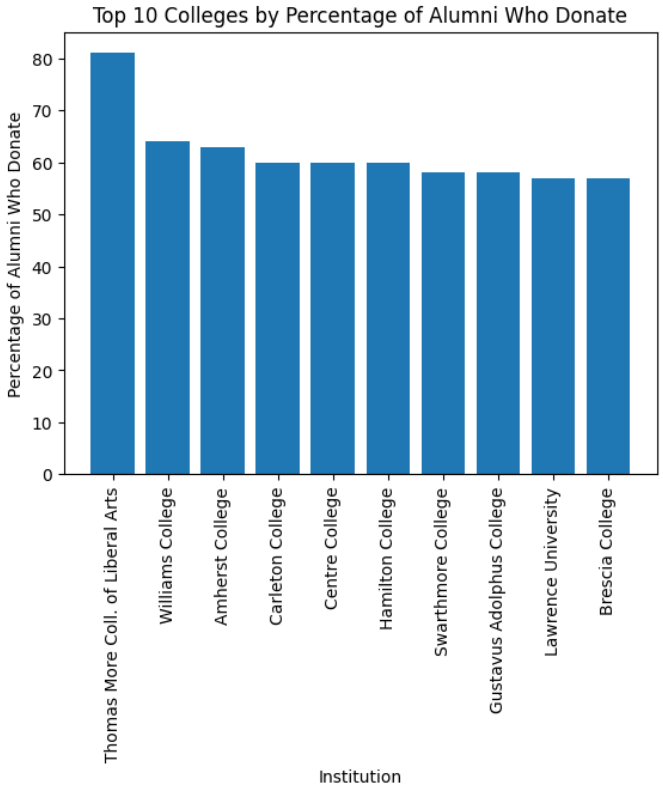

Fig. 10 Top 10 colleges with respect to donation culture among their alumni.

We can see that except for the Thomas More College of Liberal Arts, all the other top colleges have almost 60% of their alumni donating to the institute, while Thomas More had 80% of their alumni donating. This shows that these colleges were able to build and establish strong connections in the minds of their students, such that they felt to thank them via their donations.

V. Summary of the Observations

After analysing the data from the two files 'aaup.data' and 'usnews.data', we can make several observations:

1. The 'aaup.data' file contains information about the number of faculty members and their average salaries at different institutions. The data suggests that the number of faculty members generally increases with the size of the institution. From this data, we found various interpretations, such as the distribution of work in the institution. We also found about the living standards in different states using this data.
2. The 'usnews.data' file contains information about various characteristics of colleges and universities, including admission test scores, tuition fees, graduation rates, and alumni donations. We found out about the expenses in colleges, the difference between public and private colleges and also about SAT scores.

Overall, the data provided useful information for students and educators who are interested in comparing and selecting colleges and universities. By examining different characteristics of the institutions, such as admission test scores, tuition fees, faculty salaries, and graduation rates, students can make informed decisions about which institutions to apply to and attend.

VI. Acknowledgment

I would like to express my hearty gratitude to our course instructor, Prof. Shanmuga, for allowing me to work on this project and to have this wonderful experience. I also thank all the Teaching Assistants for their guidance across this project and for helping me grasp all the learnings.

Aryan Sahu
Roll no. 22110038

VII. References

This narrative has used the following references to gain knowledge of using the libraries and functions:

1. Pandas Documentation-

   https://pandas.pydata.org/docs/

2. Stack Overflow-

   https://stackoverflow.com/

3. Geeks for Geeks-

   https://www.geeksforgeeks.org/

4. McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56)

5. J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.