Data Narrative

Hitesh Kumar, Roll Number - 22110098,

Branch- Computer Science and Engineering,

Professor - Shanmuga Nathan Raman, IITGN.

Aim:

The purpose of making this report is to explore the dataset of colleges, analyse it and make some questions or develop some hypotheses and find the answers to them or try to prove or disprove them. This is for a better understanding of analysing data and getting familiar with python libraries and functions.

Overview:

The‘Colleges’ provides us with two datasets, one ‘usnews.data’ and the other ‘aaup.data’ information about 1302 US colleges. It contains Federal ID numbers, names, postal codes, marks of students on different examinations, and information about students’ expenditures and the other dataset contains information about the number of faculties in a college and their average salary and compensation. We have made a data narrative on the dataset of colleges provided. This report has answered ten scientific questions using the programming language python and some of its libraries like numpy, pandas and matplotlib with proper illustrations. We have also tried explaining the written code attached in the zip file.

Theory:

1) Python: Python is a high-level, general-purpose programming language. Its design philosophy emphasises code readability with the use of significant indentation. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming [1].

2) Libraries:

a) Numpy: It is a software library primarily used to make and analyse arrays and plots and perform different operations.

b) Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

c) Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualisations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication-quality plots.
- Make interactive figures that can zoom, pan, and update.
- Customise visual style and layout.
- Export to many file formats.
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib.

d) Seaborn: Seaborn is a Python data visualisation library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics[2].

Questions and their Answers:

Q1: Is it possible that the private college's average result will be better than the public college's average result in every exam (math, verbal, act)?

Soln: To approach the solution to this problem, we have first read the .csv file using the pd.read_csv() method from the pandas library and then converted the numbers which are stored in the data frame into integer type using the to_numeric() method, then calculated the mean of each column and observed that the mean of average marks in each subject is greater for public than private colleges as shown in Fig 1.

public=1;private=2	Avg.MathSAT	Avg.VerbalSAT	Avg.ACT	1Q_MathSAT	3Q_MathSAT	1Q_VerbalSAT	3Q_VerbalSAT	1Q_ACT	3Q_ACT
1	266.0702127659574	236.53404255319148	11.682978723404256	219.67659574468084	278.63829787234044	193.7872340425532	247.7340425531915	9.217021276595744	11.697872340425532
2	323.02884615384613	297.11538461538464	12.383413461538462	304.80528846153845	383.6911057692308	278.83653846153845	352.25240384615387	10.586538461538462	13.403846153846153

Fig.1

Q2: What is the probability that your application will get accepted with 100% surety if you submit it to a particular US College? (based on previous data)

Soln: Firstly, we have made a new data frame containing names of colleges, the number of applications received and the number of applicants accepted. Then we plot the scatter plot between the indices assigned to each college as per the data frame and the ratio of applicants accepted and applications received using matplotlib.pyplot.
Figure Fig.2 shows that most colleges contain a ratio near one but on calculating the actual probability, we get the answer as 0.027649769585253458.

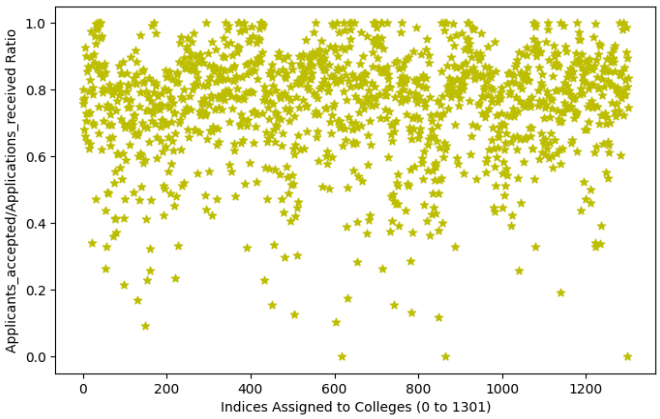


Fig.2

Q3: Is there any connection between the graduation rate and the student/faculty ratio?
Soln: As anybody can think that where the student per faculty is less, the graduation rate would be higher because it would be easy for a professor to teach fewer students better than a huge number of students. We get the same observation when plotting the required information from the dataset as in Fig.3.

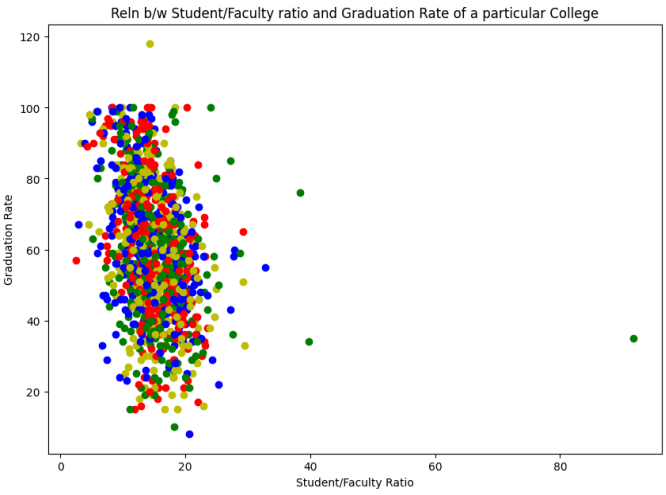


Fig. 3

Q4: Does the estimated book cost depends on the state they are studying, or is it the same in every state of the US?

Soln: For this question, we used .groupby() method to group according to the state postal

codes and stored the mean value against them, then we plotted the estimated book cost and the mean values, and we observed something like this in Fig. 4

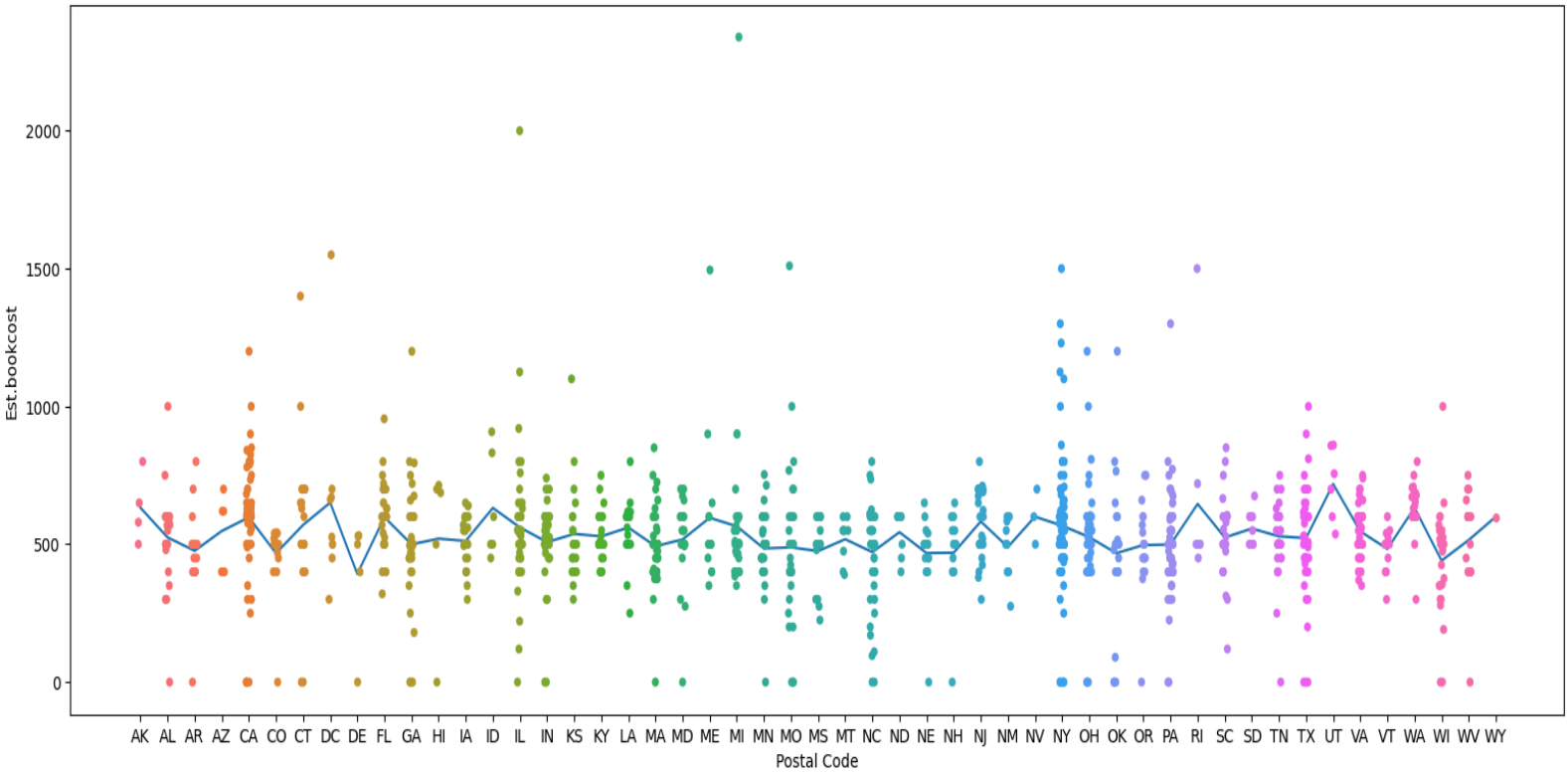


Fig.4

Q5: Which are the most popular US colleges among students whose applications are accepted?

Soln: We can find this particular solution by using the ratio of the number of students enrolled to the number of applicants accepted. This will give us the data that if a person gets accepted into a particular college, and the probability of them joining it or not. You can also refer to the figure Fig.5.

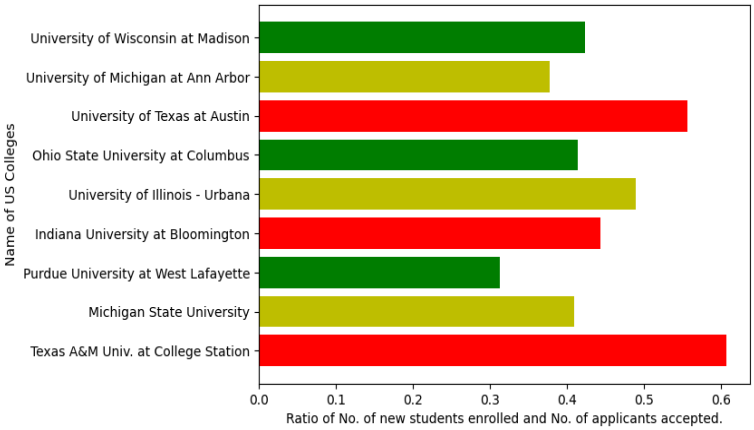


Fig.5

Q6: Should the salary of professors in IIA and IIB types of Colleges be the same? (Hypothesis)

Soln: We made the new dataset in which we took only the 'IIA' and 'IIB' types of colleges and took the mean of the data by grouping it

among the types. Then we plot the output on a pie chart (Fig.6).

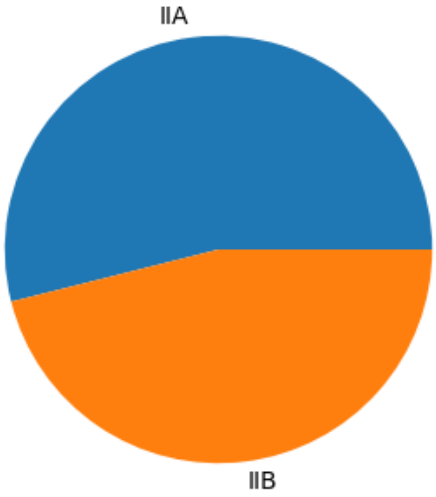
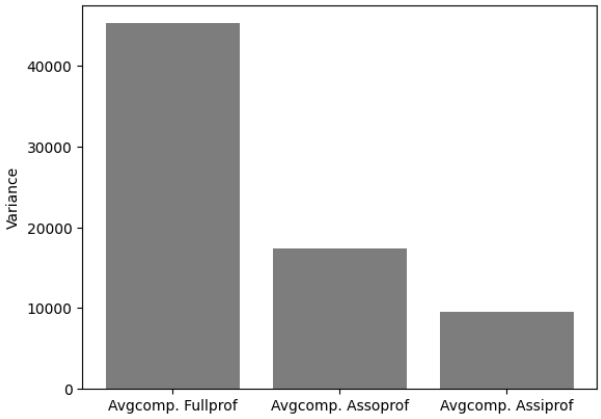


Fig.6

Q7: Calculate the mean of variances in compensation salaries of the different US colleges' professors and plot them using a bar graph.

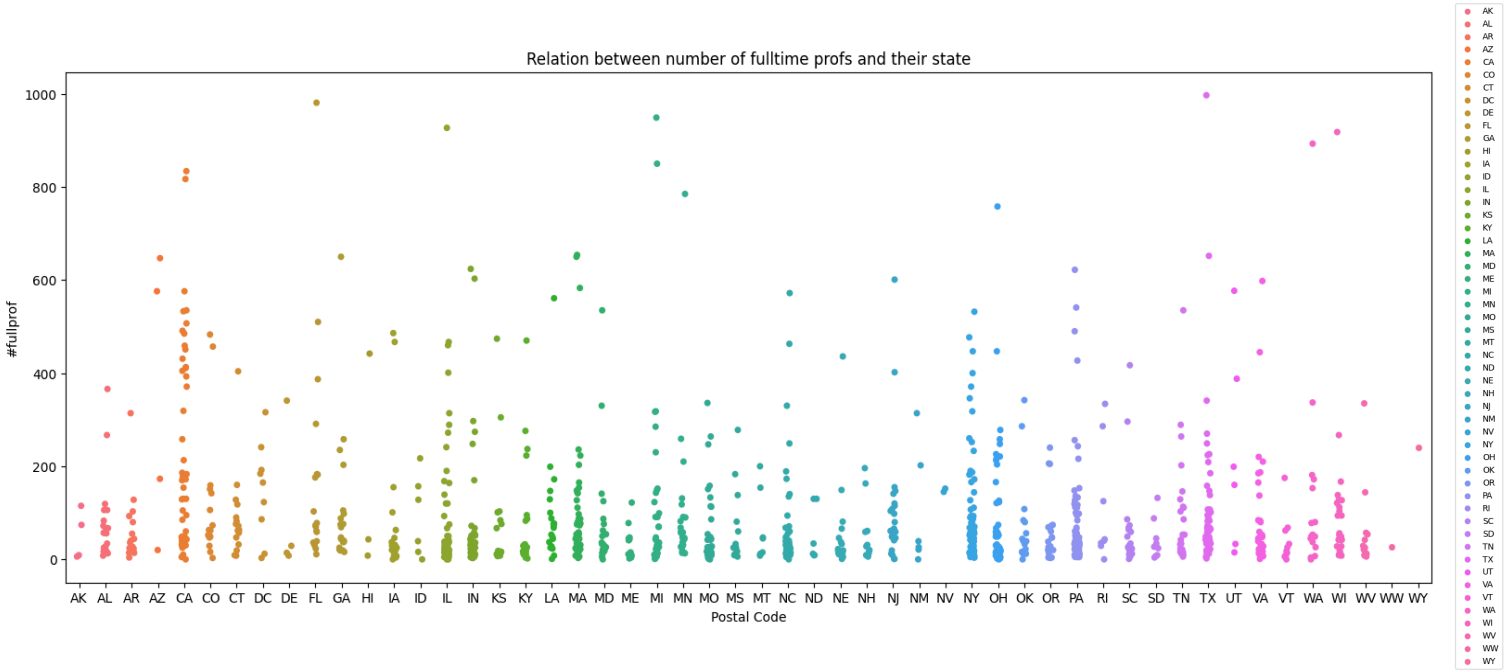
Soln: First, we calculated the variance in different types of salaries, then on taking the mean of that, we got the output as 24055.948800083162.



Q8: If a particular College has the maximum number of faculty, then what is the probability that the College is of type I?

Soln: We can simply calculate it by taking the product of probabilities of the cases when 'I' type is coming as output and when the maximum number of faculties is coming as output.

Q9) Do full-time professors prefer to join the College in a particular state?

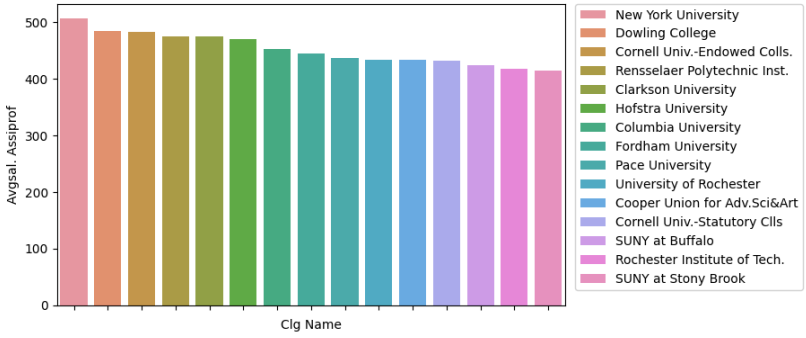


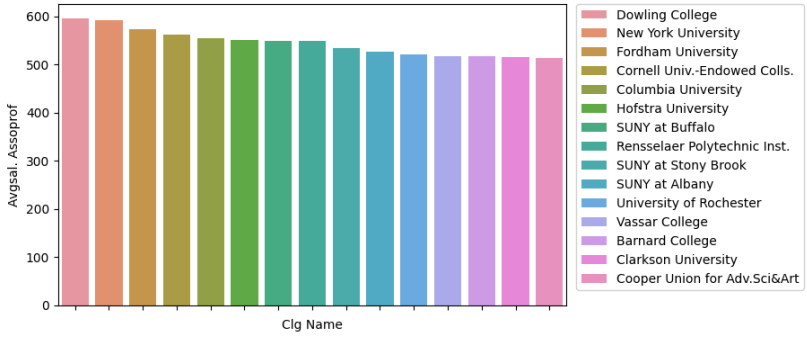
Soln: For this question, we found the number of full-time professors in a particular state and then plotted the respective strip plot using

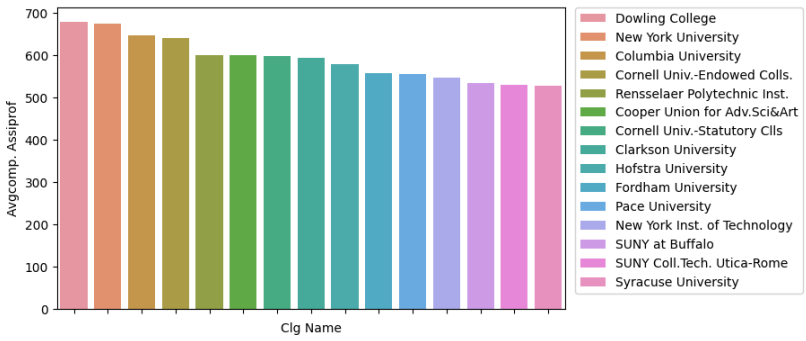
Seaborn. Then we calculate which state has the maximum number of professor count.

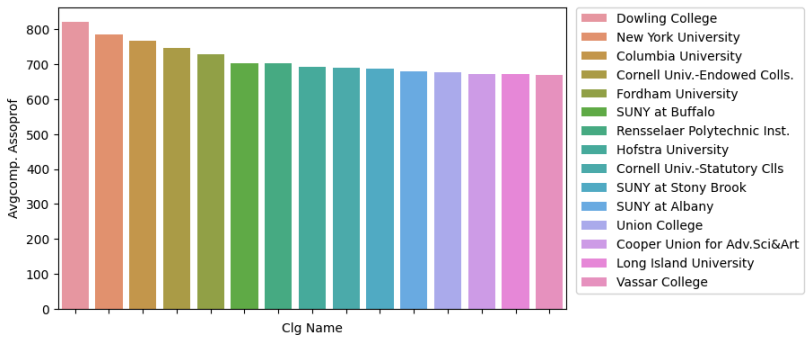
Q10) What would you prefer if you want to join as an assistant or associate professor in New York?

Soln: Before joining any College, I will see the average and compensation salary. If I change Colleges time by time, I will prefer the Colleges with a higher average salary; for the long term, I will go for a high compensation salary.
So, on observing the graphs, anyone can predict that the professor will join Dowling College.









These were the ten questions/hypotheses based on the datasets and their answers. We hope that readers will understand these concepts and the approach very well.

Unanswerable Questions:

Some questions cannot be answered from the data given in the dataset provided, such as-

- Questions related to the age, gender, etc. of students and professors.
-

Acknowledgements:

I would like to thank my teachers and batchmates who helped me understand the question, what we have to write in a Data Narrative to make it better and how to cite in Chicago style.

Reference:

[1] *Python (programming language)* (2023) *Wikipedia*. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/Python_(programming_language) (Accessed: March 30, 2023).

[2] Waskom, M.L. (2021) *Seaborn: Statistical data visualization*, *Journal of Open Source Software*. Available at: https://joss.theoj.org/papers/10.21105/joss.03021 (Accessed: March 30, 2023).