Data Narrative-2

ShauryKumar Patel
Mechanical engineering Dept.
Indian Institue of Technology Gandhinagar
Roll No-22110241
shaurykumar.patel@iitgn.ac.in

Prof. Shanmuganathan Raman
Computer Science and Engineering Dept.
(jointly Electrical engineering)
Indian Institue of Technology Gandhinagar
Jibaben Patel Chair Associate Professor
shanmuga@iitgn.ac.in

***Abstract*—The American Association of University Professors (AAUP) and US News World Report datasets are used to examine the relationship between teacher salaries, student-to-faculty ratios, and university rankings in the United States. The analysis uses various statistical techniques to find patterns and trends in the data, including linear regression, cluster analysis, and visualisation tools. In addition to identifying the universities that excel in these areas, the narrative seeks to shed light on the factors that contribute to high faculty salaries and top university rankings. Policymakers, university administrators, as well as students and their families who are thinking about higher education options, may find the study's results useful.**

I. Overview Of Dataset

- The aaup dataset (aaup.data) includes details on the characteristics and salaries of teachers for the academic year 1987–88. The American Association of University Professors (AAUP) gathered the data, which details 1,210 public and private schools. The dataset contains variables such as school type, faculty rank, salary, and years of expertise.
- The usnews dataset (usnews.data) includes information from the U.S. News World Report's 1995 rankings of colleges and universities in the United States. Data from 1,469 institutions are included in the dataset, which also contains information on the institution's type, admissions rate, SAT scores, graduation rate, and staff resources.

II. Scientific Questions/Hypotheses on AAUP

A. ***What is the probability that a randomly selected institution from this dataset has an average salary for full professors above $70,000?***

B. ***Distribution of Average salary of all ranks and mark mean and variance in graph.***

C. ***Distribution of no of colleges on the basis of state type***

D. ***How does no.of instructors vary as type of colleges ?***

E. ***How Can I define cost of living in the area based compensation ?***

III. Scientific Questions/Hypotheses on USNEWS

A. ***How does student/faculty ratio affect graduation rate ?.***

B. ***Is there a correlation between the mean MAT score and the mean SAT Math score among the colleges in the dataset?***

C. ***How can we determine quality of education of all colleges using distribution of student-faculty ratio ?***

D. ***How can we determine campus life of all colleges using distribution of room and board costs ?***

IV. Details of Libraries and Functions

Python and several of its helpful, built-in libraries, including Numpy, Pandas, MatPlotLib, and Seaborn, were used to analyse the given dataset. Among the methods from these libraries that are most frequently used are: Data from a CSV File was read onto a DataFrame using Pandas'.read csv() method.

- The dataset's unique data points can be counted using the Pandas.value counts() tool.
- The main tool for plotting the plots was MatPlotLib's.pyplot() function.
- The N biggest and smallest data points in the DataFrame were removed using Pandas'.nlargest() and.nsmallest() functions.
- The Kernel Density Estimate plot of different distributions was created using Seaborn's.kdeplot() function. It results in a cleaner

V. Answers to the Questions of AAUP

A. Answer of Q1

The question involves the use of simple probability. Here let A be the event that counts the number of college has n average salary for full professors above $70,000 and B is the number of total colleges.

$$P(A/B) = \frac{\text{A}}{\text{B}}$$

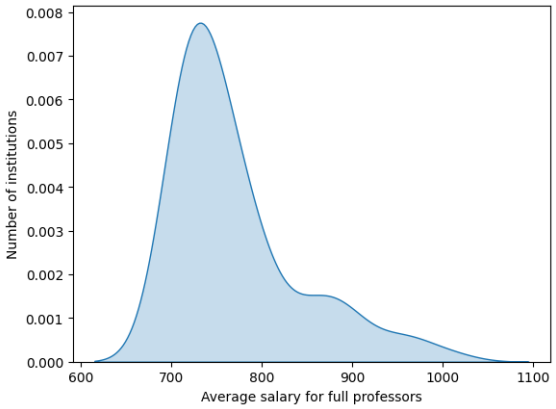

Fig. 1. Q1

B. Answer of Q2

The density plots of the distribution of the salaries(smooth histogram scaled down) shows that all these follow Gaussian/Normal Distribution since the mean is mostly at the median and major part(ideally 68%) lies with one standard deviation of the mean.

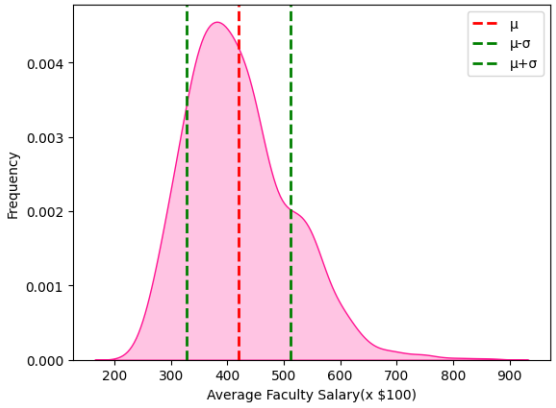

Fig. 2. Q2

C. Answer of Q3

Figure.2 shows the bar plot of Number of Colleges vs. the States. It can be observed that Pennsylvania and New York are the top two states with high number of colleges, which implies that the North-Eastern coast of US provides the best demo-graphical and geographical conditions for setting up new education institutions.

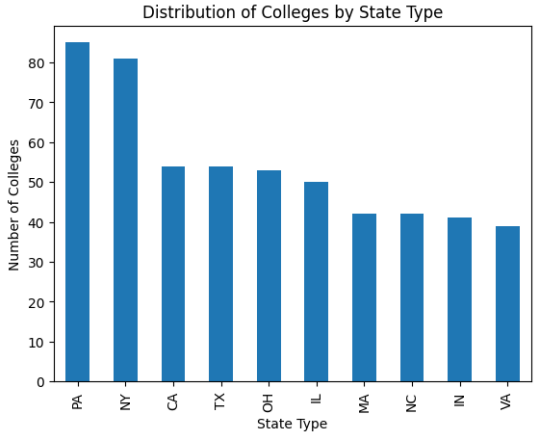

Fig. 3. Q3

D. Answer of Q4

This code plots a bar chart that shows the mean number of instructors for each type of college. The x-axis shows the type of college and the y-axis shows the mean number of instructors. The title of the chart is 'Variation of Number of Instructors by College Type'. from fig we can see the proper distribution of instructors over type I,IIA,IIB and VIIB

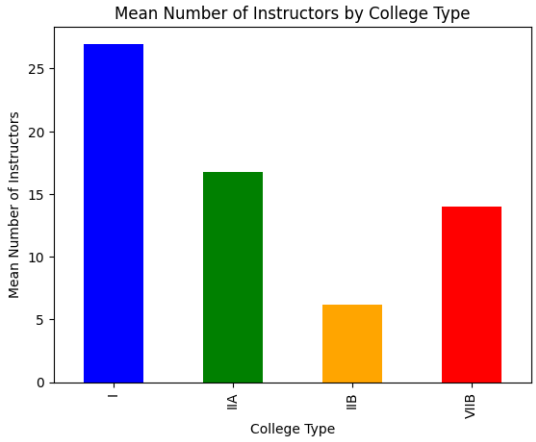

Fig. 4. Q4

E. Answer of Q5

We can use average compensation of all professors to define the cost of living as it is understandable that more the compensation, more the cost of living in that particular area.

VI. Answers to the Questions of USNEWS

A. Answer of Q1

As the student/faculty ratio increases, the graduation rate tends to decrease, indicating a negative correlation between the

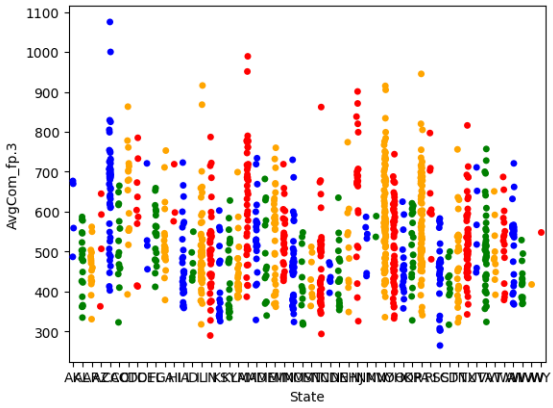

Fig. 5. Q5

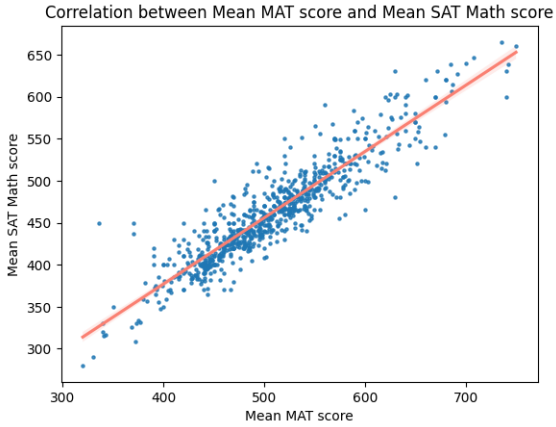

Fig. 7. Q2

two variables. The regression line suggests that there is a moderately strong negative correlation between the student/faculty ratio and graduation rate. The scatterplot shows a concentration of data points in the lower left corner, indicating that many colleges have low student/faculty ratios and high graduation rates, while fewer colleges have high student/faculty ratios and high graduation rates.

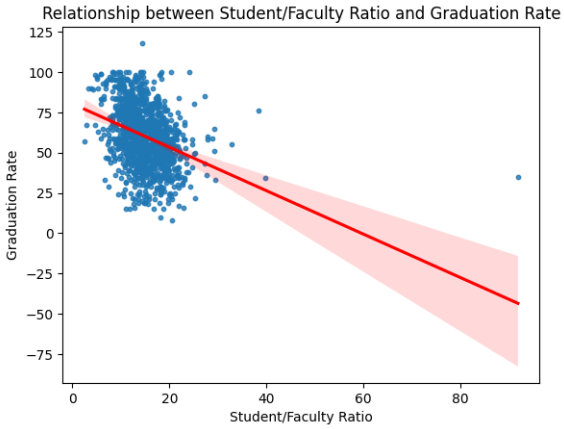

Fig. 6. Q1

B. *Answer of Q2*

The regression line suggests that there is a positive correlation between the two variables, meaning that as the Mean MAT score increases, the Mean SAT Math score tends to increase as well. The scatterplot shows the distribution of data points around the regression line, with some variation observed in the data. The correlation coefficient value of 0.91 indicates a strong positive correlation between the two variables, confirming the observations made from the scatterplot and regression line. Overall, this plot provides valuable insights into the relationship between Mean MAT score and Mean SAT Math score among colleges in the dataset.

C. *Answer of Q3*

The kernel density plot provides a smoother representation of the distribution of student-faculty ratio than a histogram. It shows that the majority of colleges have a student-faculty ratio between 10 and 20. It also shows that there are very few colleges with a ratio greater than 30. The plot highlights the distribution more clearly than a histogram and allows for more detailed analysis of the data. Additionally, the use of color and a filled area adds an aesthetically pleasing aspect to the plot.

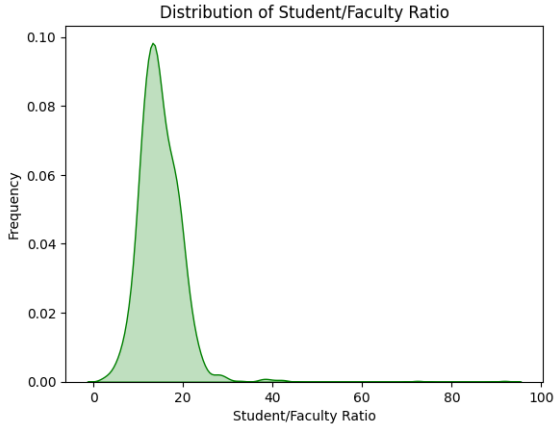

Fig. 8. Q3

D. *Answer of Q4*

This plot shows the distribution of room and board costs in colleges. The plot indicates that the majority of colleges have a room and board cost between $5000 and 10000$. Additionally, the plot is skewed to the right, indicating that there are a few colleges with very high room and board costs. Overall, this plot provides an idea of the range of room and board costs that students can expect to pay while attending college.

VII. Summary of the Observations on AAUP

- The probability that a randomly selected institution from this dataset has an average salary for full professors above

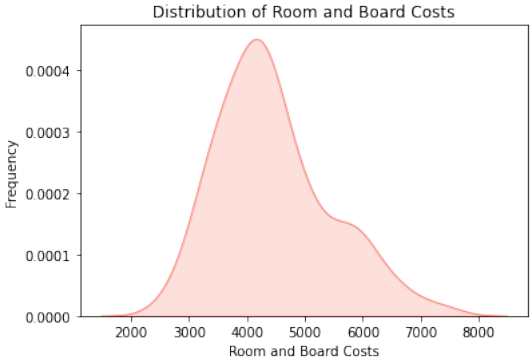

Fig. 9. Q4

$70,000:

0.07838070628768304

- The distribution of average salary of all ranks can be visualized using a histogram or a kernel density plot. The mean and variance can be marked on the graph using vertical lines or annotations.
- States with the most schools include Pennsylvania and New York, demonstrating the suitability of the Eastern Coast as a location for educational institutions.
- from this whole large big dataset we got 4 types of colleges i.e I,IIA,IIB,VIIB ,among these I type colleges have maximum number of instructors.

VIII. Summary of the Observations on USNEWS

- There was a negative correlation between student/faculty ratio and graduation rate, indicating that lower student/faculty ratios were associated with higher graduation rates.
- There was a positive correlation between the mean MAT score and the mean SAT Math score among the colleges in the dataset.
- A lower student-faculty ratio indicates a higher quality of education. Thus, we can use the distribution of student-faculty ratios to determine the quality of education of all colleges.
- The distribution of room and board costs can provide insights into the affordability and quality of campus life for students at different colleges. Higher room and board costs may indicate better facilities and amenities on campus, but could also make it more difficult for students to afford attending the college.

IX. Acknowledgement

- The American Association of University Professors (AAUP) and US News World Report deserve our sincere appreciation for supplying the datasets used in this data narrative. This study would not have been feasible without their efforts in gathering and compiling this important data.
- I also want to express my gratitude to the Carnegie Mellon University Library for opening up these databases to the general public. We highly value their dedication to promoting open access to data.
- Last but not least, I would like to express my gratitude to Prof. Shanmuganathan Raman, without whose direction and oversight this data narrative would not have met the standards of data mining necessary for such a valuable data collection.

References

[1] Index of /datasets/colleges. "Index of /Datasets/Colleges," n.d. http://lib.stat.cmu.edu/datasets/colleges/.

[2] pandas documentation — pandas 1.5.3 documentation. "Pandas Documentation — Pandas 1.5.3 Documentation," n.d. https://pandas.pydata.org/docs/.

[3] NumPy Documentation. "NumPy Documentation," n.d. https://numpy.org/doc/.

[4] seaborn: statistical data visualization — seaborn 0.12.2 documentation. "Seaborn: Statistical Data Visualization — Seaborn 0.12.2 Documentation," n.d. https://seaborn.pydata.org/.

[5] Matplotlib documentation — Matplotlib 3.7.0 documentation. "Matplotlib Documentation — Matplotlib 3.7.0 Documentation," n.d. https://matplotlib.org/stable/index.html.