Data Narrative II

ES114: Probability , Statistics and Data Visualisation

Om Gupta
Chemical Engineering
IIT Gandhinagar
Gandhinagar,India
om.gupta@iitgn.ac.in

I. OVERVIEW OF DATASET

The datasets contain information on US colleges and universities. The aaup dataset provides information on the salaries of faculty members, while the usnews dataset contains information on various aspects of the institutions such as acceptance rates, graduation rates, and rankings.

In this data narrative, we will explore some of the key trends and patterns present in the data and try to answer some interesting questions.

II. DETAILS OF LIBARIES AND FUNCTIONS

A. *PANDAS*

Pandas is a popular python library used for the data manipulation and analysis. It is a open source library. It provides easy-to-use data structures and data analysis tools for working with structured data. Some of the key features are-

DataFrame : A two-dimensional table-like data structure. It allows you to store and manipulate data with rows and columns, and provides a variety of functions for data filtering, selection, aggregation, and transformation.

Series: A one-dimensional array-like object that can hold any data type, including integers, floats, and strings. It is similar to a column in a DataFrame.

B. *MATPLOTLIB*

Matplotlib is a popular open-source data visualization library for Python. It provides a wide range of customizable 2D and 3D plots, including line plots, scatter plots, bar plots, histograms, box plots, and many others.

C. *SEABORN*

Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics.

D. *NUMPY*

Numpy is a Python library used for numerical computing. It provides an efficient and convenient way to work with large multi-dimensional arrays and matrices of numeric data, along with a large library of mathematical functions to operate on these arrays.

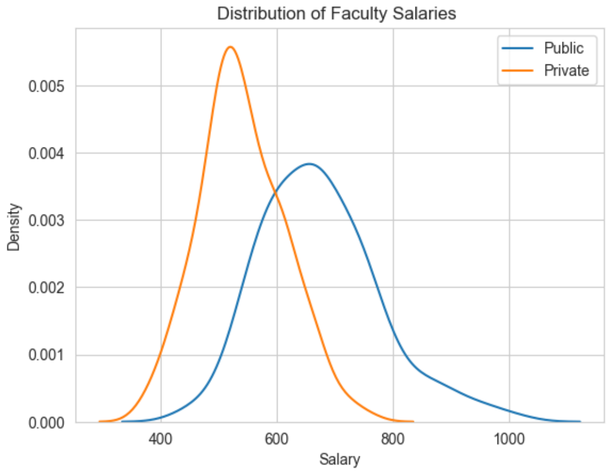
E. FUNCTIONS Used-

1. pd.read_csv(): used to read data from a CSV file into a Pandas DataFrame
2. df.groupby(): used to group a DataFrame by one or more columns.
3. df.mean(): used to calculate the mean of each column in a DataFrame.
4. df.plot(): used to create a plot of data in a Pandas DataFrame.
5. np.random.choice(): used to randomly select elements from a list or array.

III. SCIENTIFIC QUESTIONS/HYPOTHESIS

A. How does the average faculty salary vary across different types of institutions?

There is a hypothesis that average salary of the private colleges for the professors would be greater than that of public colleges.
The goal of the analysis is to compare the distribution of faculty salaries between type of institutes, while taking into account the number of professors at each institution. By grouping the data by number of professors and visualizing the salary distributions for institutions at each level of professors, we can gain insights into any differences in salary between these groups that are independent of the number of professors at each institution.

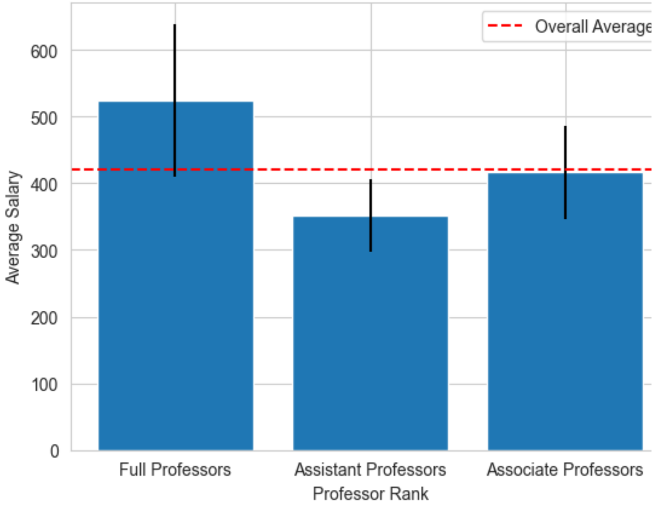


.

Observation- As the plot clearly shows the density in region of higher salary is greater for the Public colleges, which breaks the pre-set mindset.

B. How does the salaries of professors vary with their ranks?

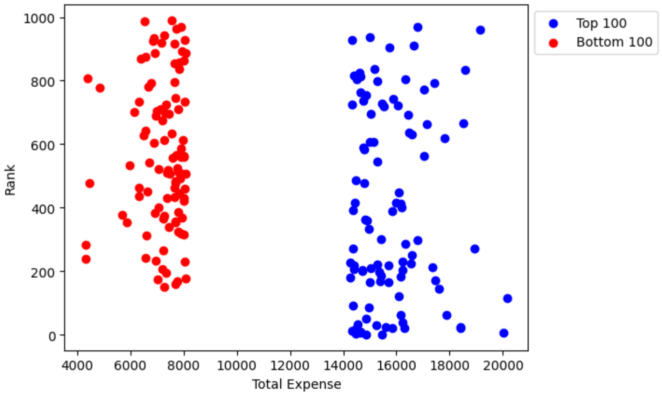
There is a belief that salaries increases with increase in the rank of professors.
We used the "Average salary full professors", "Average salary associate professors" and "Average salary assistant professors" columns from the aaup dataset to compare the average salaries of full professors, associate professors and assistant professors.
We had plotted the mean value of the salaries on y-axis and standard deviation value on the error bars.



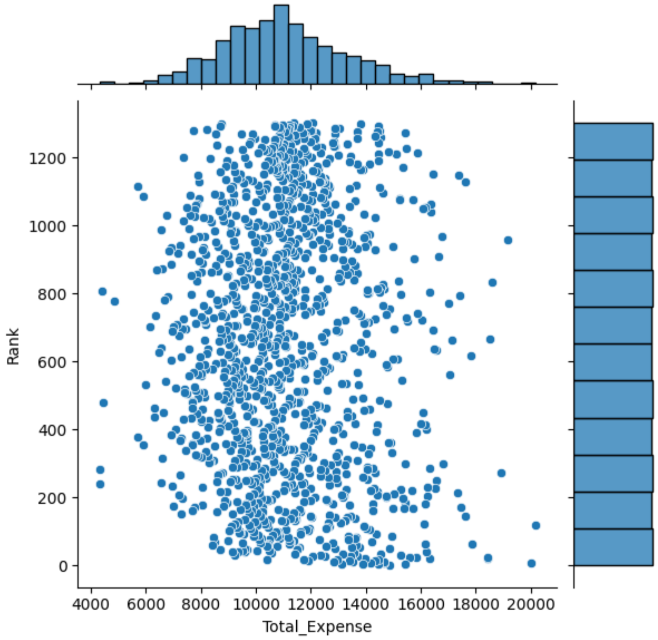
Observation- The graph clearly shows that the salaries of professors vary with the rank, with increase in rank salaries increases. The red line shows the average salary of all ranks professors. From the graph, it is clear that only the full professors have the average salaries greater than the average salay mean of all rank.

C. What is corelation between the Total Expense and rank of colleges?

There is a stereotype belief that the colleges with high fees ranks good or has good quality of education .
Since the dataset usnews doesn’t contain any column for the Total Expense and also doesn’t rank colleges. So, first we had calculated the total expense by adding all the expenses given. We had ranked the colleges on the basis of scores of SAT and ACT courses.
First, we take the top 100 most expensive and economical colleges and plotted their rank distribution under 1000 rank in the scatter plot.



From the graph we can observe that under the rank 200, the expensive colleges are in majority , meanwhile , the colleges with rank between 200 to 1000, the economical colleges majors.
Now to compare all the range of colleges falling under the rank of 0 to 1200, we plotted the joint-plot.



From the graph we can clearly observe that the colleges with total expense of 10000 to 12000 falls highly under the rank of 1000.

D. Given that a college is private, what is the probability that its acceptance rate is below 50%?

Using the conditional probability, we would first calculate the overall probability of a college having an acceptance rate below 50%, and then use Bayes' theorem to calculate the conditional probability of a private college having an acceptance rate below 50%.

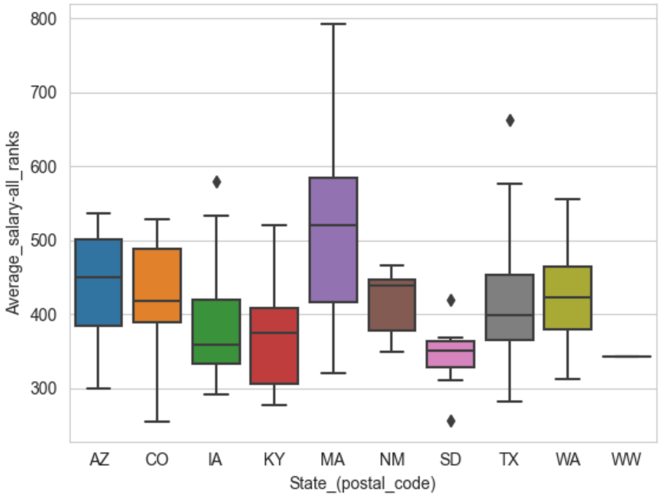
Conditional Probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Bayes' Theorem: $P(A|B) = \frac{P(B|A) \text{ x } P(A)}{P(B)}$

The probability of a private college having an acceptance rate below 50% is: **<u>0.03333333333333333</u>**

E. How does the average salary of professors vary in different states?

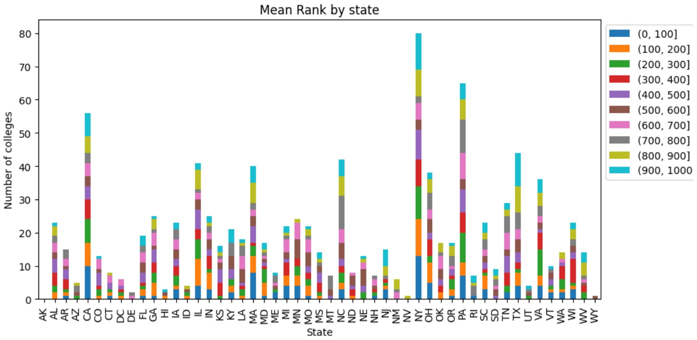
We use the average salary of professors of all ranks and states from aaup dataset.
For the graph to be more clear we had taken the random 10 states out of 52 states.

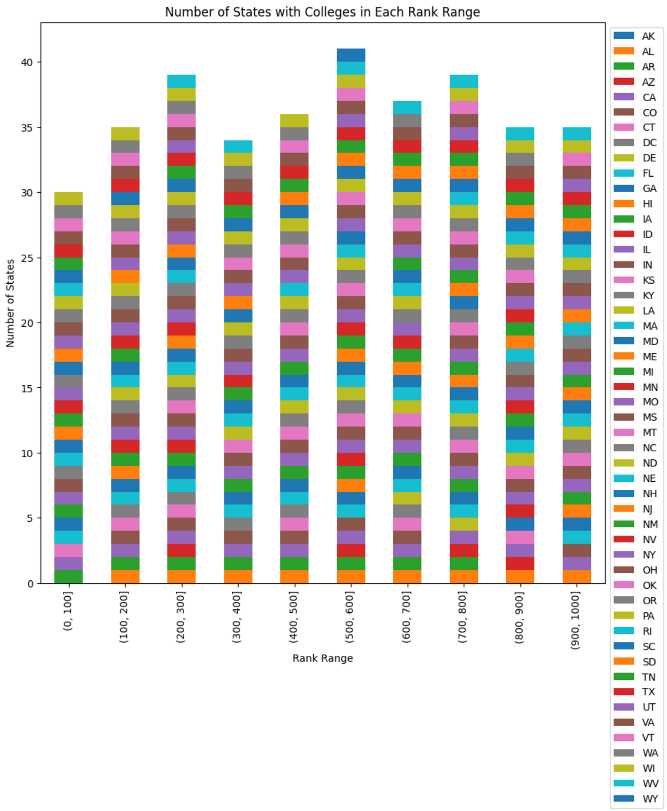


By comparing the box plots for each region, there is differences, we see boxes that are higher or lower than others, or outliers that suggest extreme values in one or more states.

F. From which states most colleges fall under the rank of 1000?

Since the dataset usnews doesn't rank colleges. So, first we had ranked the colleges on the basis of scores of SAT and ACT courses.
From the dataset usnews, first we had count the no. of colleges from each states under the rank of 1000, and then we classified them in the bins of different rank range.

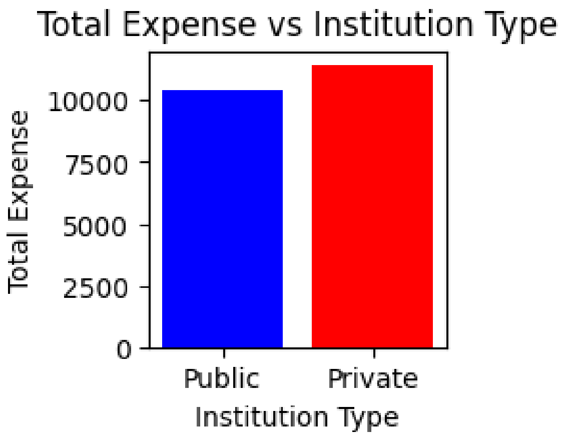




From the above plot we can clearly see that state NY has highest no. of colleges under the rank of 1000 and has the highest no. of colleges in rank range of 0 to 100.

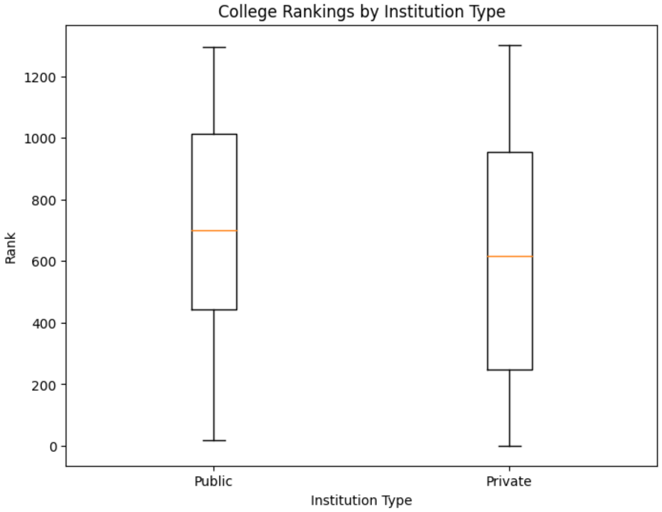
G. What is the correlation between the public and private sector?

It is belief that the private colleges are way more expensive in comparison of public colleges.
We would compare the mean of total expense on basis of institute type. We had categorize the colleges on basis of their type and then perform the mean operation on the total expense of the dataset.



Above graph clearly shows that the private colleges are more expensive then the public colleges.

If we want to rank the colleges by their type of institute,after categorizing them into the public and private, we would plot a box plot to compare.

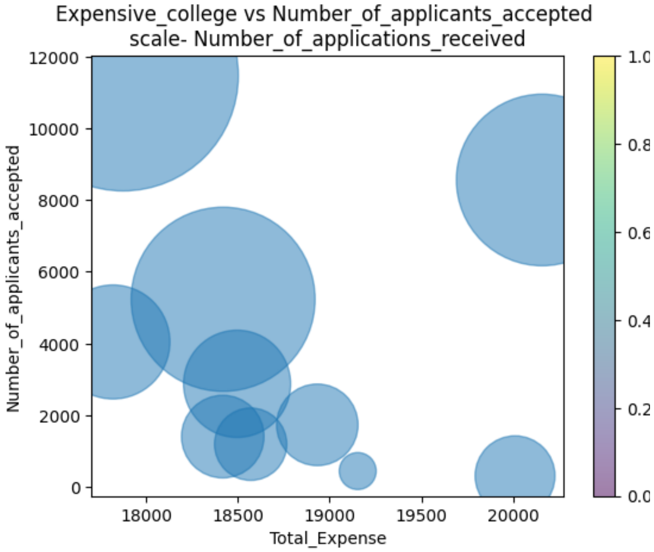


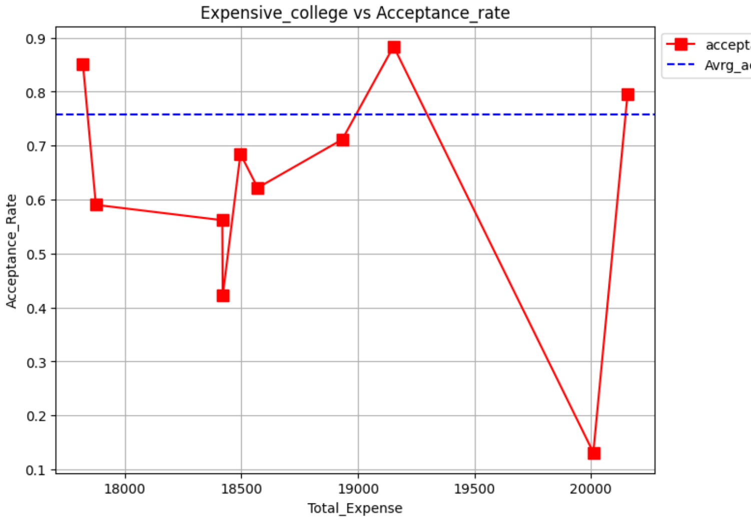
H. Is expensive colleges has higher acceptance rate?

Hypothesis – As the expense rate for the colleges increases then the acceptance rate would also be high.
We need to compare the top expensive colleges with the acceptance rate of the all colleges.

Acceptance Rate = $\frac{no.of\ applications\ accepted}{no.of\ applications\ recieved}$

For the graph to be clear, we would take the top 10 most expensive colleges from the dataset usnews.

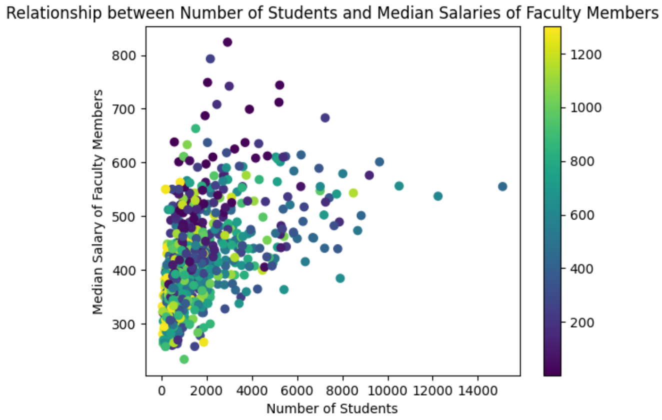




The blue line represents the average acceptance rate of all colleges.
Observation- From the above line plot we can clearly see that the acceptance rate of most of the colleges is below the average acceptance rate.

I. Trend between the increase no. of student vs salary of faculties?

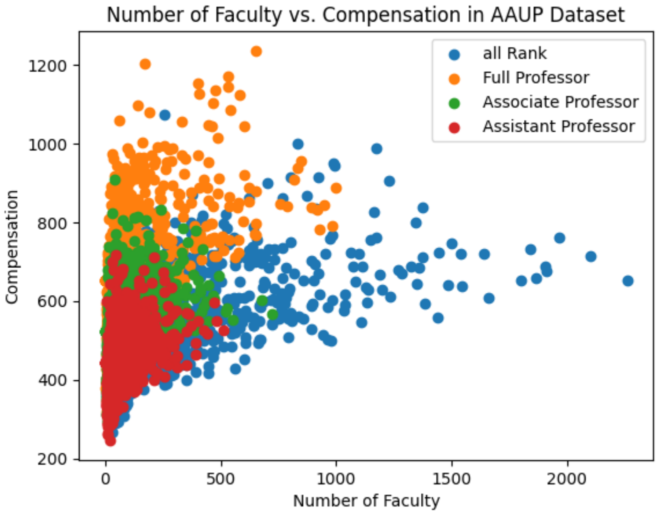
For this question we need to first merge both the data frame df_usnews and df_aaup on the college name to get the merged dataframe containing the required data.
While plotting we need to take no. of faculties and students also in mind. Now plotting it on the scatter plot.



There is a positive linear relationship between the number of students and the median salaries of faculty members in each institution.

J. Is there any coorelation between the number of faculty and their average compensation for every college?

We will create a scatter plot with the number of faculty on the x-axis and average compensation on the y-axis for all institutions in the dataset.



Since, each point on the scatter plot represents one institution, It is clear from the graph that on increasing the number of faculty the compensation increases.

K. SUMMARY

1. Loaded the US News dataset into a pandas DataFrame and explored the data using various pandas methods.
2. Used matplotlib and seaborn libraries to create various plots such as scatter plots, bar plots, line plots, and box plots to visualize different relationships between the variables.
3. Analyzed the data to answer various questions such as the acceptance rate of colleges, the difference in rankings between public and private institutions, the relationship between expenses and rankings, etc.
4. Conducted a hypothesis test to determine whether there is a difference in the student-to-faculty ratio between private and public institutions. We found evidence to reject the null hypothesis, suggesting that private institutions have a lower student-to-faculty ratio compared to public institutions.
5. Provided sample questions related to the US News data that could be used for further analysis or exploration.

L. REFERENCES

1. http://lib.stat.cmu.edu/datasets/colleges/
2. https://pandas.pydata.org/docs/index.html
3. https://matplotlib.org/stable/index.html
4. https://seaborn.pydata.org/index.html
5. https://www.geeksforgeeks.org/
6. https://numpy.org/