ES 114 - Data Narrative - II

Hrriday Ruparel
22110099
Freshman, Computer Science and Engineering
Indian Institute of Technology
Gandhinagar, India
hrriday.ruparel@iitgn.ac.in

***Abstract*—This is an inferential and statistical data analysis report on a pair of datasets drawn from two sources, U.S. News & World Report's Guide to Americas Best Colleges and the AAUP (American Association of University Professors) 1994 Salary Survey. The data analysis explores the data from a statistical standpoint and draws some conclusions based on the inferences and observations made using data analysis and data visualization tools supported by Python.**

I. Overview of dataset

The dataset includes information of 1300+ US Colleges extracted from U.S. News & World Report's Guide to Americas Best Colleges and the AAUP (American Association of University Professors) 1994 Salary Survey. The AAUP data includes average salary, overall compensation, and number of faculty broken down by full, associate, and assistant professor ranks.

The dataset contains 2 important files relevant for our discussion.

1. **usnews.data** – The U.S. News data contains information on tuition, room & board costs, SAT or ACT scores, application/acceptance rates, graduation rate, student/faculty ratio, spending per student, and a number of other variables for 1300+ schools.
2. **aaup.data** - The AAUP data includes average salary, overall compensation, and number of faculty broken down by full, associate, and assistant professor ranks.

The files that have been predominantly used for this data narrative include **usnews.data** and **aaup.data**.

The datasets had many missing values already present in it. In order to draw some informative results out of the data, the datasets were cleaned at each step and hence, the number of data points used to answer each question varied from question to question.

II. Scientific questions and hypotheses

Some useful scientific questions and hypotheses made using U.S. News data for the data analyses are as follows:

1. *Determine the density of colleges in different regions of U.S. What inferences can you draw from the density plot of colleges?*
2. *Is there any correlation between SAT and ACT scores? What is the general trend of ACT vs. SAT scores for public and private colleges? Are there any outliers that you can spot in the data? Elaborate.*
3. *There is a general hypothesis that a large number of applicants prefer public colleges over private colleges. Is this bias true? If so, demonstrate with a graph?*
4. *What can you say about the average spending per student in the U.S colleges (public and private)? How does it compare with the instructional spending of an institute per student? Are there any outliers?*
5. *What is the general trend of graduation rate from the U.S colleges? Plot a distribution of graduation rate and elaborate. Are there any anomalies in the data?*

Some useful scientific questions and hypotheses made using AAUP 1994 Salary Survey data for the data analyses are as follows:

6. *Describe the relationship between the average salaries of the staff working at the U.S colleges and the type of the colleges.*
7. *Describe the relationship between the average compensation of the staff working at the U.S colleges and the type of the colleges.*
8. *Provide an insight into the relationship between the rank of a professor and their respective salaries.*
9. *Full professors are likely to get more compensation than any other kind of professor. Is this hypothesis true? If so, prove it.*
10. *Mention some institutes where working as a professor is a 'financial boon'. What conclusions can you draw out of this?*

III. Details of libraries and functions

The data narrative has used the various data analysis and data visualization tools supported by Python. These include:

1. **NumPy** – NumPy, also known as Numeric Python, is a Python library provides tools to deal with arrays and large datasets. This library is imported into Python as **np**.
2. **Pandas** – Pandas is Python library that allows the user to create table-like structures, organise data, perform operations on the data and compute various important statistical measures. Pandas has been imported as **pd**.
3. **Matplotlib** – Matplotlib is a Python library that provides the facility to visualize the data by providing tools for plotting graphs. The **pyplot** functionality of this library has been imported as **plt**.
4. **Skimage** – Skimage is a Python library that is useful for performing image operations. Its **io** functionality has been used to import locally stored images for plotting.
5. **SciPy** – SciPy (Scientific Python) is a Python library that is used for scientific scomputing. For the data narrative, the **stats** functionality of the module has been used.

6. **Sklearn** – Sklearn is a Python library that is an extension to the existing SciPy library with the added benefit of applying machine learning concepts in Python. For clustering of data, **KMeans** functionality of **sklearn.cluster** has been imported.
7. **Seaborn** – Seaborn is a Python library that is useful for data visualization, plotting various kinds of graphs and also for cutomising them. It has been imported as **sns**.

Each of these libraries have been extensively used for answering scientific questions stated in the previous section. Some of the notable functions of these libraries used for data analysis and visualization are:

1. **np.arange()** – This function helps to create an array consisting of whole numbers of desired range.
2. **np.linspace()** – This function helps to equally-spaced-elements array of desired range.
3. **pd.DataFrame()** – This function helps to create a new DataFrame.
4. **pd.read_csv()** – This function helps convert .csv files to Pandas DataFrame.
5. **pd.DataFrame.astype()** – This functions helps to perform typecasting on columns of a DataFrame.
6. **pd.DataFrame.iloc[]** – This property helops to access elements in a DataFrame on the basis of their location indices.
7. **pd.DataFrame.drop()** – This funtion helps to remove rows or columns.
8. **pd.DataFrame.assign()** – This function helps to add new columns to existing DataFrame.
9. **pd.DataFrame.value_counts()** – This function helps to obtain the frequency of unique elements in a column of a DataFrame.
10. **pd.DataFrame.sort_values()** – This function helps to sort the values of a DataFrame on the basis of the arguemnt passed.
11. **pd.DataFrame.describe()** – This function computes various statistical measures of a DataFrame.
12. **pd.DataFrame.groupby()** – This function helps to group columns of a DataFrame on the basis of other columns of the DataFrame.
13. **plt.plot()** – This function helps to create plots of variousb types like bar, hist, scatter, pie, density, etc.
14. **plt.show()** – This function is used to view the plotted graph on the the figure window.
15. **plt.subplots()** – This function helps to work with subplots on a single figure. It provides tools for customisation of subplots and their layouts.
16. **plt.title()** – This function sets the title for the graphs.
17. **plt.xlabel()** & **plt.ylabel()** – These function help to label the x-axis and y-axis respectively.
18. **plt.xticks()** & **plt.yticks()** – These functions help to customise the ticks on the x-axis and y-axis respectively.
19. **plt.xlim()** & **plt.ylim()** – These functions help to set the limit/range of values plotted on the x-axis and y-axis respectively.
20. **plt.grid() –** This function displays the grid on the graphs.
21. **plt.legend()** – This function displays trh legend on the graph.
22. **plt.imshow()** – This fiunction helps to plot an image.
23. **io.imread** – This function helps to load an image from an image file.
24. **stats.zscore** – This function helps to compute the Z-scores of the datapoints of a dataset.
25. **stats.linregress –** This functions helps to compute the linear least-square regression for two sets of measurement.
26. **KMeans.fit –** This function helps to compute k-means clustering on the data provided.
27. **KMeans.predict** – This function helps to compute the closest cluster each datapoint in a dataset belongs to.
28. **sns.jointplot()** – This function helps to plot a graph of two variables. Its plots are more beautiful than what Matplotlib offers.
29. **sns.histplot()** – This function helps to plot a histogram of a data.

These are the major functions and methods used from the the imported Python libraries.

IV. Answers to Scientific Questions and Hypotheses

For the first question, the aim was to plot the density of U.S. colleges in different regions of the U.S. In order to so, I firstly imported an image that depicted the map of the U.S.A labelled by the postal codes used for each state. The coordinates of each state were approximately derived and stored for plotting the density of the colleges, which was obtained by counting the number of colleges corresponding to a unique postal code of the U.S States. This was done to get a complete overview over the dataset from a geographical point of view and understand the geographic distribution of colleges. The following plot was obtained.

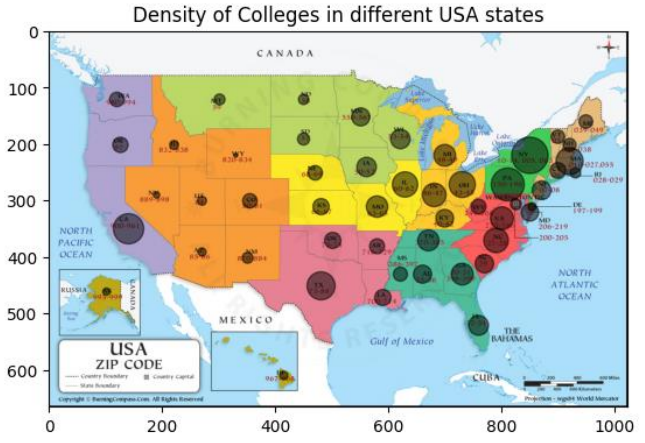

Fig. 1a – Bubble plot depicting the density of colleges in different regions of the U.S.A.

Fig. 1a depicts that the majority of the colleges lie on the eastern and western coastal states of the U.S.A. This is quite relevant from the fact that most of the technological developments and investments are made in these regions.

In order to get a finer perspective on the data, I plotted a more descriptive and detailed graph as follows:

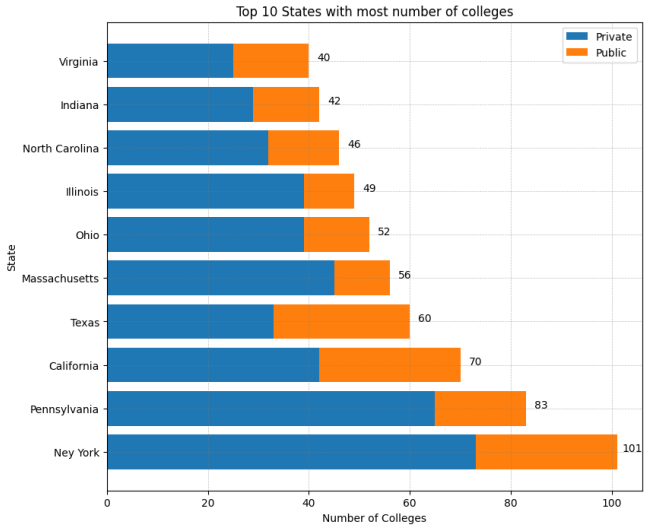

Fig. 1b – Stacked bar graph depicting the regions with most number of the US colleges (both private and public).

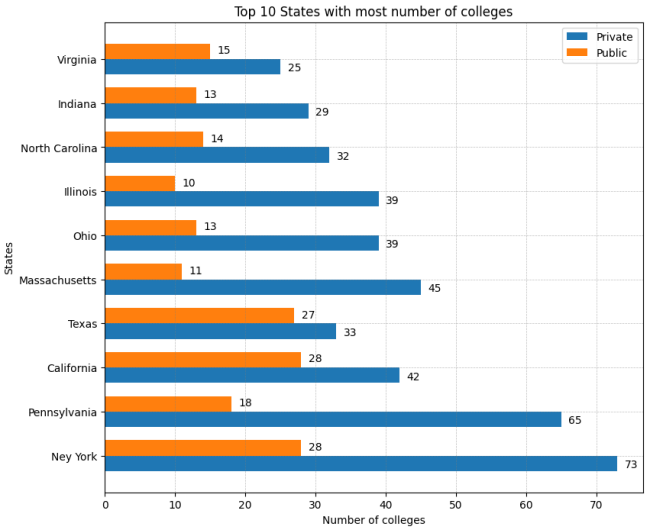

Fig. 1c – Bar graph depicting the regions with most number of the US colleges (private and public separated).

Fig. 1b and Fig. 1c are obtained by simply sorting the dataset. The graph shows that a maximum of 101 colleges lie in the state of New York (NY). Not only that, the graph also shows that the number of colleges among these states is majorly dominated by the private sector.

For the second question, the aim was to plot the average ACT scores (out of 32) against the average SAT scores (out of 1600) of colleges provided in the data set and draw inferences from it. This was included as a point for discussion because whenever someone mentions about colleges and their comparison, the human mind easily diverts to the average ACT and SAT scores.

The following plots were obtained when ACT scores were plotted against SAT scores of each and every college (public and private separated):

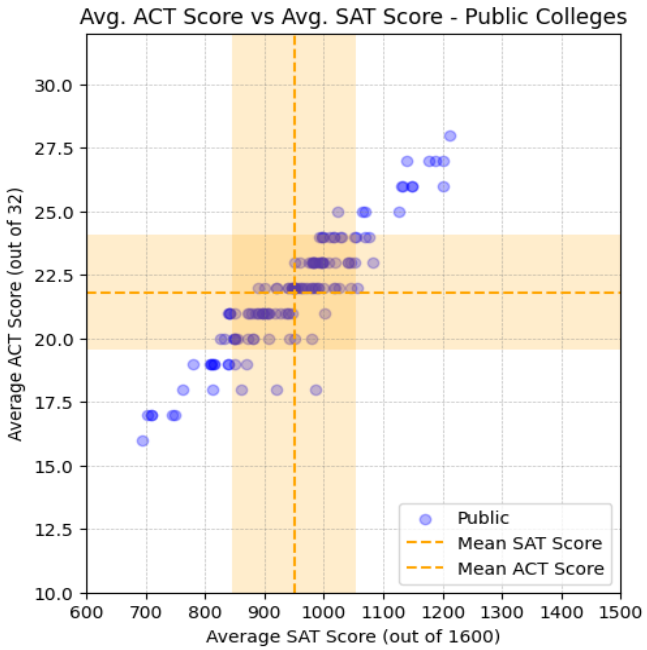

Fig. 2a – Scatter plot depicting the mean ACT scores plotted against mean SAT scores for US colleges (public).

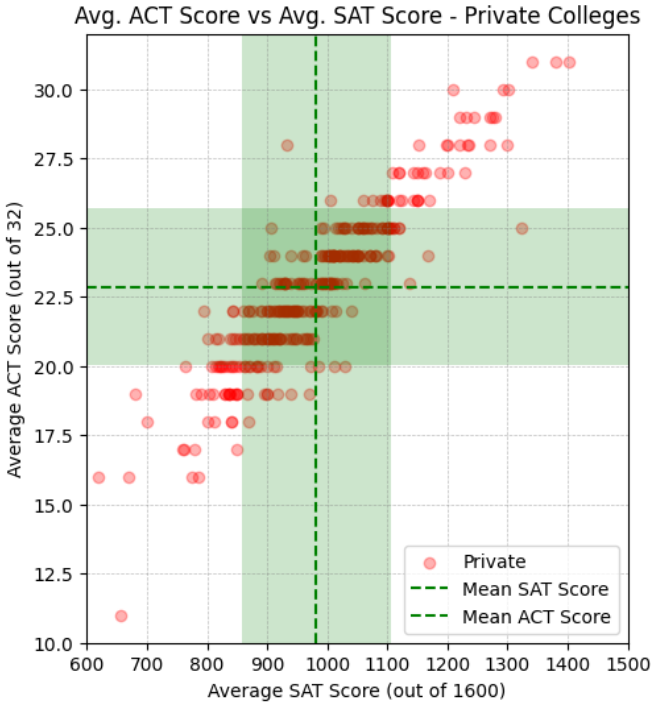

Fig. 2b – Scatter plot depicting the mean ACT scores plotted against mean SAT scores for US colleges (private).

In Fig. 2a and Fig. 2b, the colour coding has been done in such a way that the blue dots (and red dots in Fig. 2b) represent each college. The blue dots (and red dots) are translucent and when many of these points overlap, the blue (and red) colour become(s) intense, indicating high density. So, the plot also provides an idea about the density distribution. The two orange (and green) dotted lines represent the mean of mean ACT and SAT scores among all the colleges. The orange (and green) shaded region represents the range from $(\mu - \sigma)$ to $(\mu + \sigma)$, where μ represents the mean of mean ACT or SAT scores and σ represents the standard deviation of mean ACT or SAT scores among all the colleges.

In order to compare the measure between the two sectors of colleges (public and private), a combined graph was plotted:

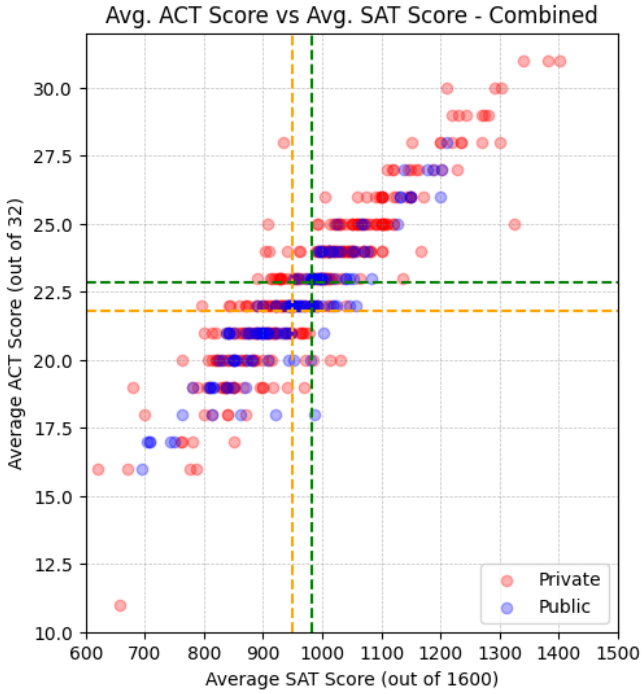

Fig. 2c – Combined plot of Fig. 2a and Fig. 2b.

The Fig. 2c depicts that the mean of mean SAT and ACT scores in case of private colleges (mean SAT score = 984.784884; mean ACT score = 22.854651) is more than that of public colleges (mean SAT score = 950.090278; mean ACT score = 21.812500). The figures Fig. 2a, Fig. 2b and Fig. 2c clearly show that there is a positive correlation between the mean ACT and SAT scores of US colleges, whether it be private or public college.

This plot inspired the idea of utilising the k-means clustering algorithm to segregate the colleges on the basis of the mean SAT and ACT scores of US colleges. Hence, after applying the k-means algorithm, the following plots were obtained:

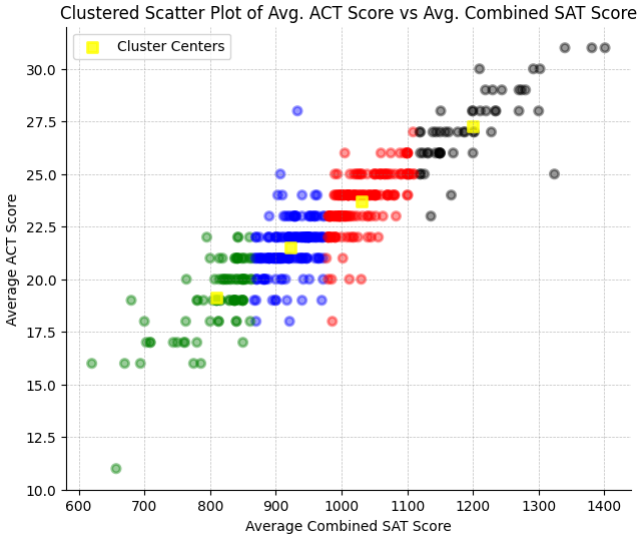

Fig. 2d – Clustered scatter plot of mean ACT score vs. mean Combined SAT score of every US college.

Clustering the data with 4 cluster centers provides a good way of ranking colleges on the basis of the accepted mean ACT and SAT scores. The black-coloured cluster represents those US colleges that accept students having very high SAT and ACT scores, implying higher-ranked colleges. The red-coloured and blue-coloured clusters represents those US colleges that accept students who have scored moderately well in ACT and SAT examinations. As observed before, the mean of the two variables occur in these two clusters. Hence, these two clusters represent the averagely-ranked colleges. The green cluster represents the colleges that may be ranked as 'below-average'.

While looking closely at Fig. 2d, one can observe that there exist many colleges that lie far from the bulk of the data points. These data points, called outliers, can be visibly evident as well as mathematically derived and be tagged as outliers.

In order to identify these outliers, the concept of Z-score used in probability estimations was used and the following colleges were found out to be outliers in the dataset:

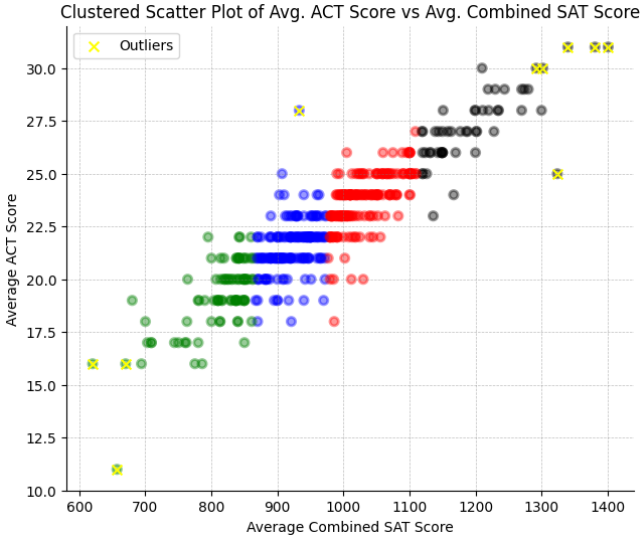

Fig. 2e – Clustered scatter plot of mean ACT and combined SAT score of US colleges along with outliers

college name	state(postal code)	avg combined sat	avg act
Pomona College	CA	1340	31
Stanford University	CA	1401	31
Massachusetts Institute of Technology	MA	1381	31
Johns Hopkins University	MD	1292	30
Duke University	NC	1302	30
Voorhees College	SC	670	16
Jarvis Christian College	TX	657	11
Texas College	TX	620	16

Fig. 2f – Table representing outliers in the data calculated mathematically using Z-scores.

	college name	state(postal code)	avg combined sat	avg act
432	Amherst College	MA	1324	25

	college name	state(postal code)	avg combined sat	avg act
846	Utica College of Syracuse University	NY	933	28

Fig. 2g – Table representing outliers in the data tagged visually using Fig. 2e

The figure Fig. 2f shows that there exist some outliers like Pomona College, Stanford University, Massachusetts Institute of Technology, John Hopkins University and Duke University that are highly ranked, hence, come under the category of outliers. Some colleges like Voorhees College, Jarvis Christian College and Texas College are outliers because they are ranked way below the average.

The figure Fig. 2g shows that there also exist colleges like Amherst College which accept students having very high combined SAT score but moderate ACT scores. On the other hand, colleges like Utica College of Syracuse University accept students having high ACT scores but moderate SAT scores.

For the third question, I had to test the hypothesis that whether there existed a bias among the students to prefer public US colleges over private US colleges. In order to test this hypothesis, the number of applications received, number of applicants accepted and number of new students enrolled were plotted for each sector – public and private, and the following graph was obtained.

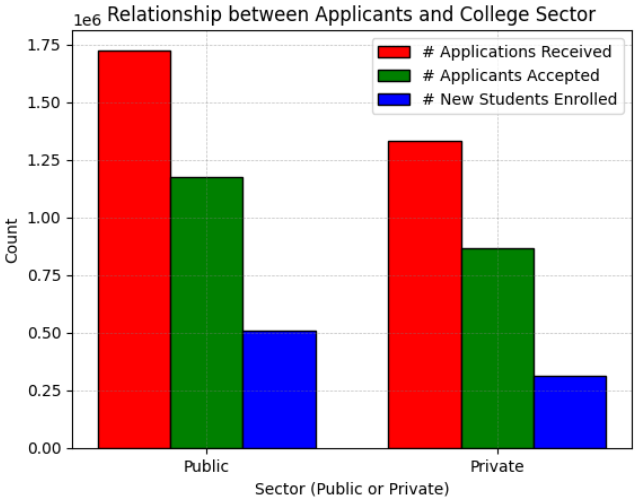

Fig. 3 – Column graphs depicting the number of applications received, number of applicants accepted and number of new students enrolled plotted against each sector – public and private

Fig. 3 clearly displays that the hypothesis is true and it holds for all the cases. The reasons for this trend are explored further down the narrative.

For the fourth question, the main aim is to compute the average spending per student in US colleges (both public and private) and draw conclusions from it. From the dataset, the types of costs that a student incurs is broken down into – In-state tuition, out-of-state tuition, room-board costs, additional fees, book costs and personal miscellaneous personal spending. All these types of costs incurred per student were plotted against the type of college-public and private and the following graph is obtained:

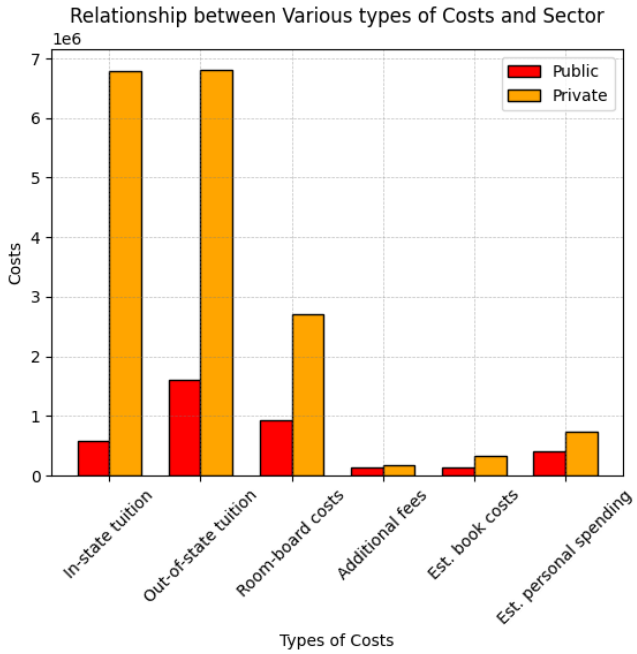

Fig. 4a – Column graph depicting the relationship between different types of costs incurred per student in public and private US colleges.

The graph apparently displays the divide among the public and private US colleges in terms of spending per student. It clearly shows why many students might want to opt for public colleges because the personal spending is significantly low in case of public US colleges as compared to private ones.

Now, the question arises whether the spending per student is worth the outcome or return that one may expect from an educational institute. In order to test this, the instructional expenditure of the institute per student was plotted against the spending of the students, which produced the following results:

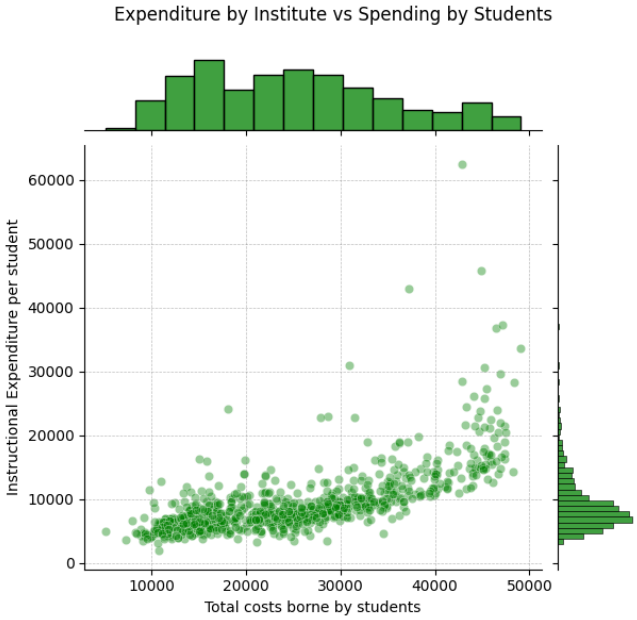

Fig. 4b – Joint plot showing the relation between expenditure of the institute on education per student vs. the spending per student.

The Fig. 4b helps us to draw the basic conclusion that there is a positive correlation between the monetary investment of the institute on education and the amount of the money that each student has to pay to experience such facilities.

From Fig. 4b, one could draw many data points that exist as outliers. Among these outliers, only a few of the outliers that showcase exceptionally high instructional spending per student are as follows:

college name	pub(1)-pvt(2)	instructional expd. per student
California Institute of Technology	2	62469
Washington University	2	45702
Antioch University	2	42926
Harvard University	2	37219
University of Chicago	2	36854
Massachusetts Institute of Technology	2	33541
University of Judaism	2	31012
Columbia University	2	30639
Dartmouth College	2	29619
Emory University	2	28457
Princeton University	2	28320
Duke University	2	27206
University of Rochester	2	26037
University of Pennsylvania	2	25765
Carnegie Mellon University	2	24386

Fig. 4c – Table showing the various colleges that have very high instructional spending per student.

The Fig. 4c provides a good estimate on the quality of education imparted in the US colleges. From the table, one can observe that educational institutions like California Institute of Technology, Washington University, Harvard University, Massachusetts Institute of Technology, Princeton University, Duke University, University of Pennsylvania and Carnegie Mellon University, which are some of the renowned institutes, are also included in the table. This also proves the idea that good educational institutes invest heavily on educational resources. One interesting observation made over here is that all the institutes come under private sector, thus, strengthening our prior analysis about the divide between the public sector and private sector applicants.

For the fifth question, the main aim is to plot the distribution of the graduation rate from different US colleges. The distribution plot of graduation rates is as follows:

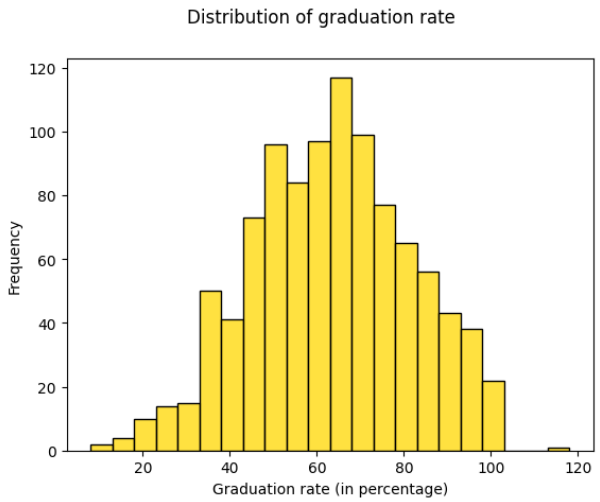

Fig. 5a – Distribution plot of graduation rate from US colleges.

From the Fig. 5a, one can immediately observe an anomaly in the graph where graduation rate in more than 100%. It turns out that the college was Cazenovia College which had a graduation rate of 118%. Assuming that the anomaly existed, I plotted a refined version of the distribution of the graduation rates as follows:

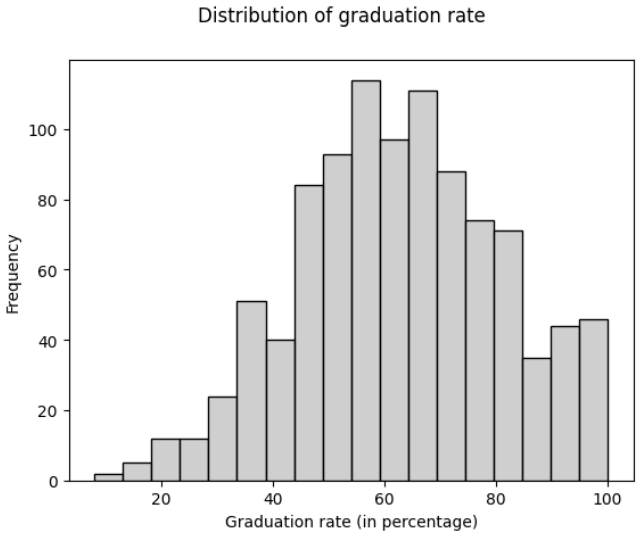

Fig. 5b – Refined distribution plot of graduation rate of US colleges.

Fig. 5b and its corresponding values depict that the mean graduation rate from US colleges is around 62.688933 with a standard deviation of 18.395330.

For the sixth question, the main aim was plot the relationship between the average salaries of professors (full, associate and assistant) against the type of college and it yielded the following results:

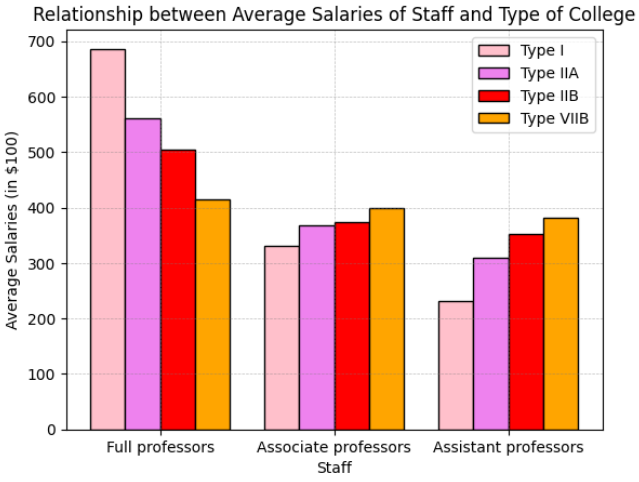

Fig. 6 – Relationship between average salaries of staff and type of college.

One interesting observation that can be made over here, is that though the average salaries for the professors keeps decreasing from full professors to associate professors to assistant professors irrespective of the type of college, but in case of associate professors and assistant professors, the average salaries increase as the type of colleges change from Type I to Type IIA to Type IIB to finally Type VIIB. The case for full professors is exactly the opposite.

For the seventh question, the main aim was plot the relationship between the average compensation of professors (full, associate and assistant) against the type of college and it yielded the following results:

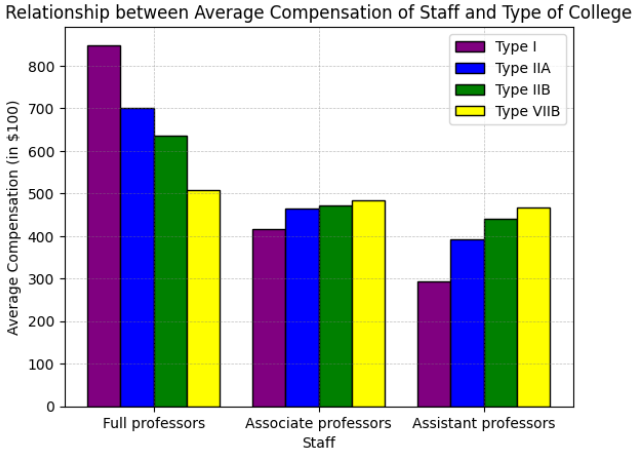

Fig. 7 – Relationship between average compensation of staff and type of college.

On close inspection, the column charts in the figures Fig. 6 and Fig. 7 turn to be very similar to each other (isomorphic). The same kind of trend follows in case of compensation as in salaries. This shows that, the average salaries of professors is strongly correlated to average compensation offered.

For the eighth question, the main goal was to plot a distribution of average salaries obtained by different professors and draw inferences from it. A plot was obtained as follows:

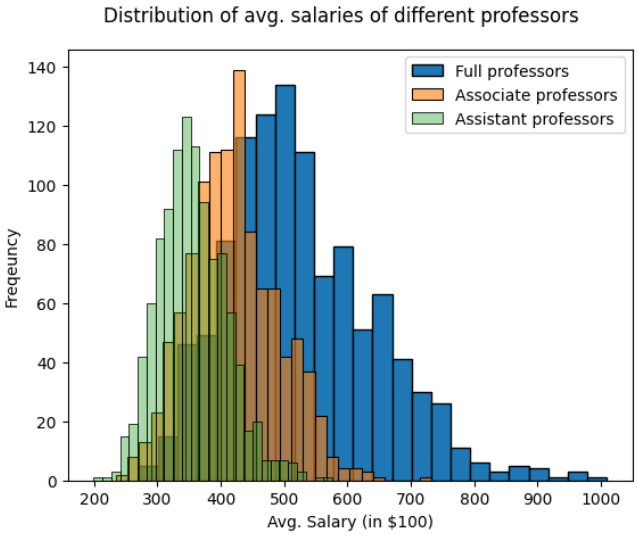

Fig. 8 – Distribution plot of various professors showing the average salaries obtained.

From the Fig. 8, it is evident that full professors acquire larger amount of salary than any of its kinds, which seems quite reasonable as they are working as a full-time professor, implying more workload.

For the ninth question, the main aim is to test whether the hypothesis that full professors would have higher compensation that its kinds holds true or not in case of US colleges. In order to answer this question, I plotted the distribution of the average compensation of the different kinds of professors as follows:

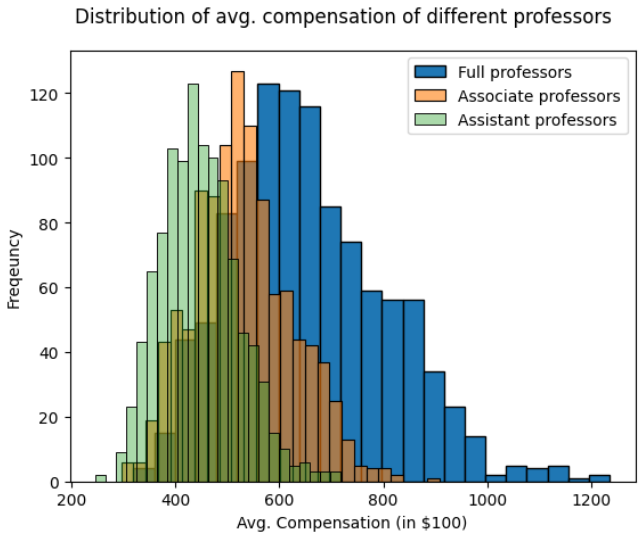

Fig 9 – Distribution plot of various professors showing the average compensation offered.

From the Fig. 9 and from prior analysis on a similar argument, I came to know that the average compensation for the full professors is indeed the highest amongst its kinds. This is strengthened by the prior observation that there is a strong positive correlation between average salaries of professors and average compensation.

For the tenth and the final question, our main goal is to find such educational institutes which invest large amounts of money on the professors' salaries and compensations, typically meaning 'financial boon; for the professors. For this, I computed the average institutional spending per professor and sorted the data according to it. The following plot was obtained:

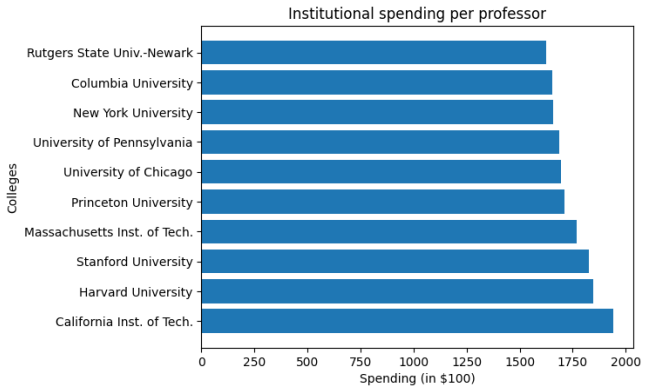

Fig. 10 – Bar graph showing the top 10 colleges that invest heavily on their professors.

From the Fig. 10, one can observe that educational institutions that are highly ranked liked California Institute of Technology, Harvard University, Massachusetts Institute of Technology, Princeton University, University of Pennsylvania and New York University, are on the top of the list, suggesting that good institutions invest heavily on the quality of the professors and staff.

V. SUMMARY OF OBSERVATIONS AND LEARNINGS

The data sets provided an insight into the various variables that are associated with the US colleges and helped develop an understanding how these variables are interconnected and correlated to each other. The US colleges that are located on the eastern and western coasts of the U.S.A have the most

number of colleges, and also have the best colleges, that accept only a handful of highly talented students having high ACT and SAT scores. These colleges are usually private institutions, charging hefty amounts from students for education but at the same time providing facilities and quality education, backed by well-versed professors earning in tens and hundreds of thousands of dollars.

VI. Unaswerable Questions

1. What does 'type' of colleges actually signify?
2. Why most of the US colleges are located on the eastern and western coasts of the U.S.A?

VII. Acknowledgement

This is to thank Professor Shanmuganathan Raman for providing this opportunity to work on a dataset and perform data analysis and visualization in order to extract some useful information.

:

VIII. References

1. 'NumPy documentation,' accessed Mar 30, 2023, https://numpy.org/doc/1.24/

2. 'Pandas documentation,' accessed Mar 30, 2023, https://pandas.pydata.org/docs/

3. 'SciPy Documentation' accessed Mar 30, 2023, https://docs.scipy.org/doc//scipy/index.html

4. 'Sklearn Documentation' accessed Mar 30 2023, https://scikit-learn.org/stable/index.html

5. Seaborn Documentation' accessed Mar 30, 2023, https://seaborn.pydata.org/

6. 'Matplotlib documentation,' accessed Mar 30, 2023, https://matplotlib.org/stable/index.html

7. Siya Sivarajah, 'Most Important Pandas Functions-Full Tuorial,' accessed Mar 30, 2023, https://towardsdatascience.com/pandas-full-tutorial-on-a-single-dataset-4aa43461e1e2