Data Narrative

Bhavik Patel, Btech Mechanical Engineering
Roll no.-22110047
Prof.- Shanmuga R , IIT Gandhinagar

I. OVERVIEW OF THE DATASET

There are two different datasets Aaup.data and Usnews.data which are available on the website (http://lib.stat.cmu.edu/datasets/colleges/=). Both datasets provide information on a sample of about 1140 colleges in the US. The datasets include variables related to the size, location, type, and salary, as well as variables related to students, faculty, and the student's academic performance and salaries of different faculty.

NOTE: All the numeric values are in the string. So convert them to int dtype before performing any operation on the dataset.

II. SCIENCTIFIC QUESTION/HYPOTHESES

A. *What is the probability that a randomly selected college from the dataset lies in the most popular state in the US for higher studies?*

Approach:

1. First, we must find the most popular states.
   For this, we need to count the number of colleges in each state.
   We will use *value_counts* for this job.
2. Then, we will find the probability of choosing the most popular state.

Equation: Probability=Total no. of favorable outcomes/Total no. of outcomes

Solution:

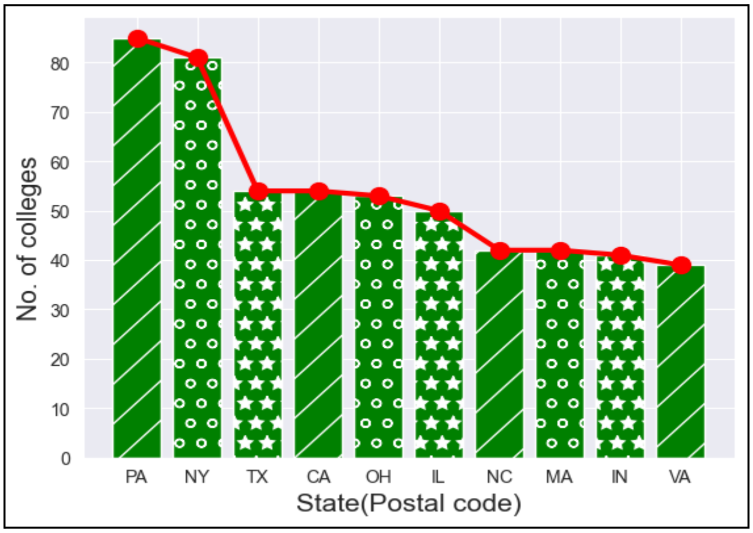


Fig. 1- No. of colleges in each state

From fig. 1 we can say that Pennsylvania(PA) state has the most number of colleges. There is also a great difference in the number of colleges between the rest of the states and the top two states i.e. PA and NY.

The probability of choosing a college that is located in the state of PA is 0.073.

B. *Which type of college pays more to all ranked faculties?*

Approach:

1. We will create a new data frame having the type of college and average salary.
2. Then, we will take the mean of the average salary for each type, i.e., I, IIA, and IIB.

Solution:

- ‘I’ institutions are doctoral-granting universities with a very high level of research activity.
- ‘IIA’ institutions are master’s colleges and universities with larger programs.
- ‘IIB’ institutions are master’s colleges and universities with smaller programs.

The figure below shows that type ‘ I’ colleges pay more average salaries to their faculty than type ‘IIA’ and ‘IIB.’

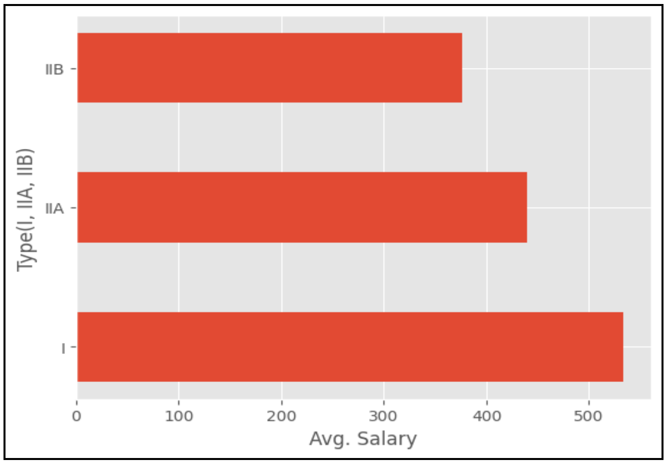


Fig. 2

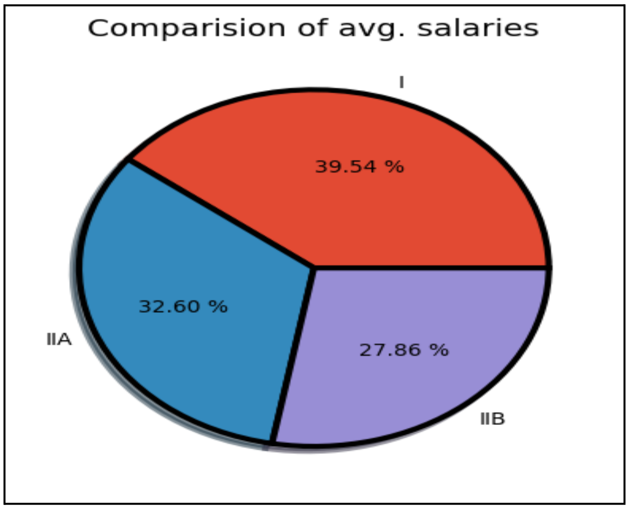


Fig. 3

C. *Does the size of the college affect the average salary of its faculty members?*

Hypothesis: The college with a large number of faculty should pay more to its faculty as it might also have more fees. It could also be assumed that the faculty is highly qualified as it teaches many students and is therefore paid more.

Solution:

- The below scatter plot shows the relation between faculty size and average size
- Each point in fig. 4 represents a particular college. We can notice that the density of points at the bottom left of the plot is more, which means there are more colleges with less faculty size and also pays very less to them.
- On the other hand, when the faculty size is larger, the average salary is also relatively much higher.
- From fig. 5, we can see that the average salary increases with an increase in faculty members, but for very large faculty sizes, greater than 1500, there is a slight decrease in the salary, which can be verified by fig. 4, which depicts that most of the highest paying colleges have faculty size around 1000-1500 and not 2000.
- From the above points, we can say that the size of the college affects the average salary, but the hypothesis can't be said to be fully correct as the very high faculty-sized college doesn't pay a very high salary.

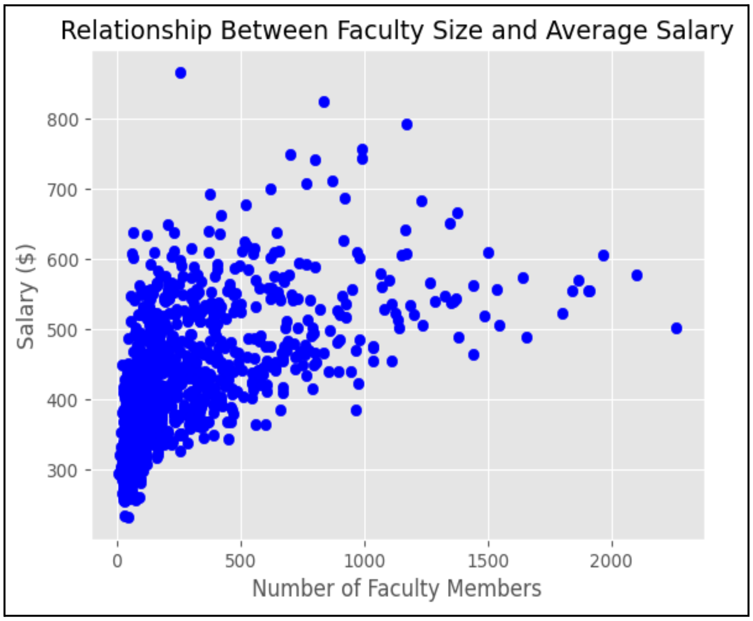


Fig. 4- Scatter plot b/w avg. salary and no. of faculty

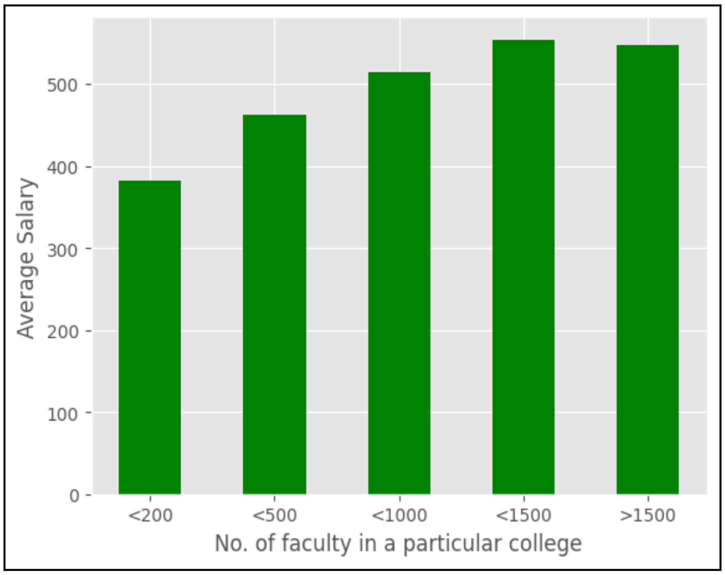


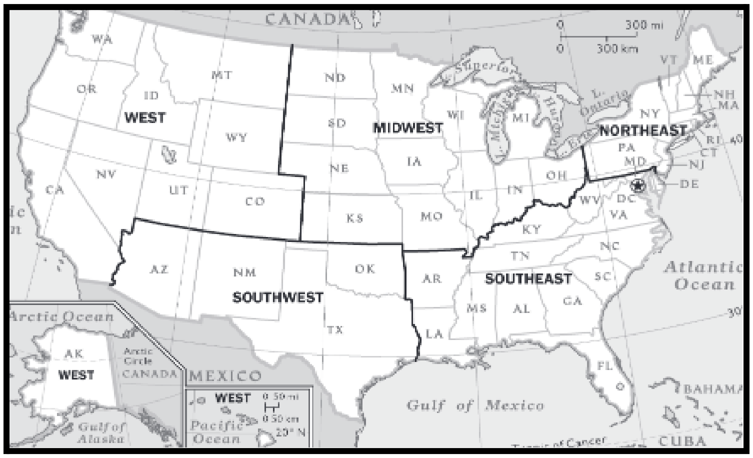
Fig. 5- Bar plot Avg. salary and no. of faculty in classified in different sizes

D. *Are there significant differences in the faculty salaries at colleges in different geographical regions of the US?*

Hypothesis: The regions containing the most famous states for education in the US, like Massachusetts(MA) and Colorado(CO) have very highly qualified faculty, and those faculty are paid more than other region faculty.

Unique Approach:

1. We need to add a new column in the dataset containing different geographical regions for different states. We will use the below image for the same.



Source: National Geography

2. Create a boxplot of 'Avg. Salary' for different regions with the help of the seaborn library.

Solution:

- ➢ The below box plot graph shows the distribution of average faculty salaries across different geographic regions in the US. Each box represents the data's interquartile range (IQR), with the central line indicating the median.
- ➢ From the graph, we can see that the regions with the highest median salaries are "West" and "Northeast," while the regions with the lowest median salaries are "The Southeast" and "Plains."
- ➢ However, it's important to note that there is a lot of variability within each region, with some regions having a wide spread of salaries and others having a narrower spread.
- ➢ Overall, the box plot graph provides a helpful visualization of the distribution of salaries across regions.
- ➢ The box plot clearly proves our hypothesis, with 'West' and 'Northeast' having the highest average salaries.

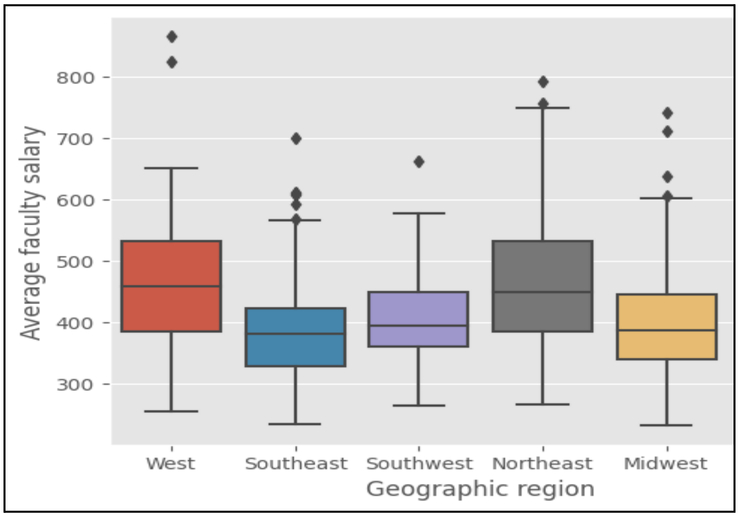


Fig. 6

E. How does the number of faculty members vary with the type of college?

Approach:

1. First, make a new data frame containing only the 'type' and 'No. of all ranked faculty' columns.
2. Then, create a box plot of the above-made data frame using the seaborn library.

Solution:

- ➢ From the box plot, we can see that the median number of faculty members is highest for Type I colleges, followed by Type IIA and then by Type IIB.
- ➢ Type I colleges also have the broadest range of faculty members, with several outliers above the upper whisker.
- ➢ Type IIB colleges have the smallest range of faculty members and no outliers.

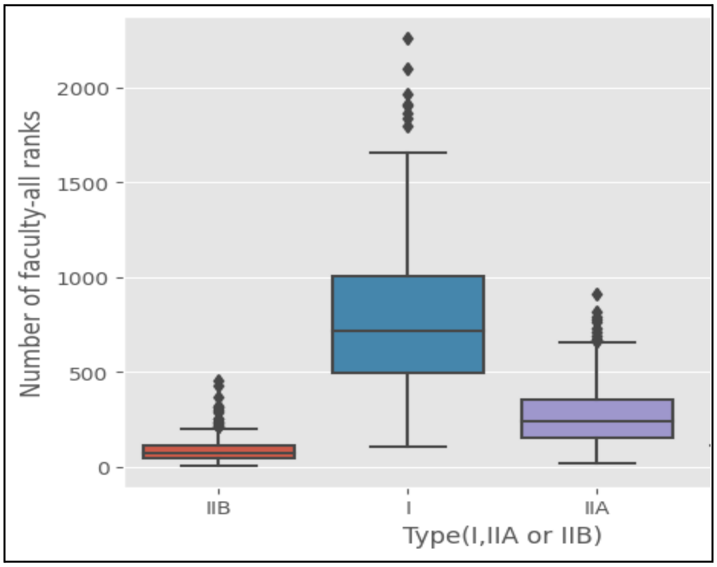


Fig. 7

US news Dataset

A. *Is there a correlation between the median SAT score for incoming freshmen and the graduation rate for a given college or university?*

Hypothesis: Colleges with higher median SAT scores for incoming freshmen have higher graduation rates.

Approach:

1. First, make a new data frame containing only the ‘SAT score’ and ‘Graduation Rate’ columns.
2. Then, create a scatter plot of the above-made data frame using the seaborn library.

Solution:

➢ We can see from the below plot that there is less graduation rate when the SAT score is less.

➢ While very few colleges have very high SAT scores (greater than 1200), those who have also have very high graduation rates as well.

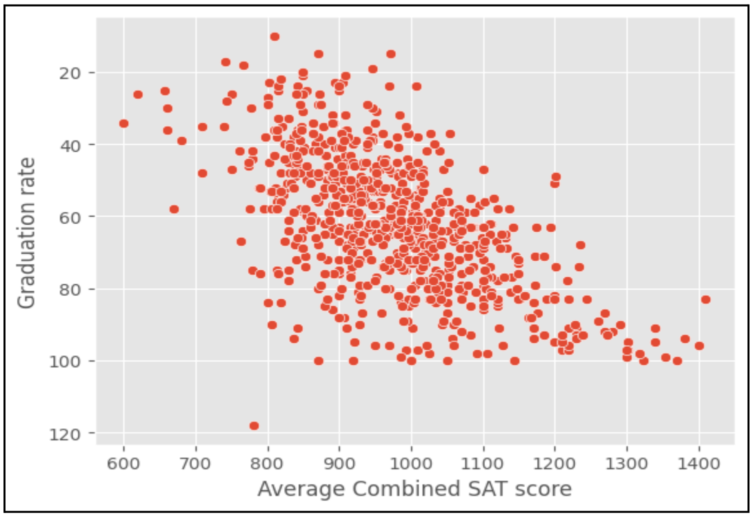


Fig. 8- Relation b/w SAT score and Graduation

B. *Is it harder to get into a public college in the US or a private college?*

Approach:

1. Load the data frame containing Public/Private indicators and combined SAT score and remove the rows having ‘*.’
2. Then, create a box plot of the above-made data frame using the seaborn library.
3. Load the data frame containing verbal and Math SAT scores.
4. Then, create a scatter plot of the above-made data frame using the seaborn library.

Solution:

❖ From the boxplot in fig. 9 shows that the median SAT score required for admission is almost the same for both public and private colleges

❖ The distribution of combined SAT scores is almost the same for both but is slightly higher in private colleges.

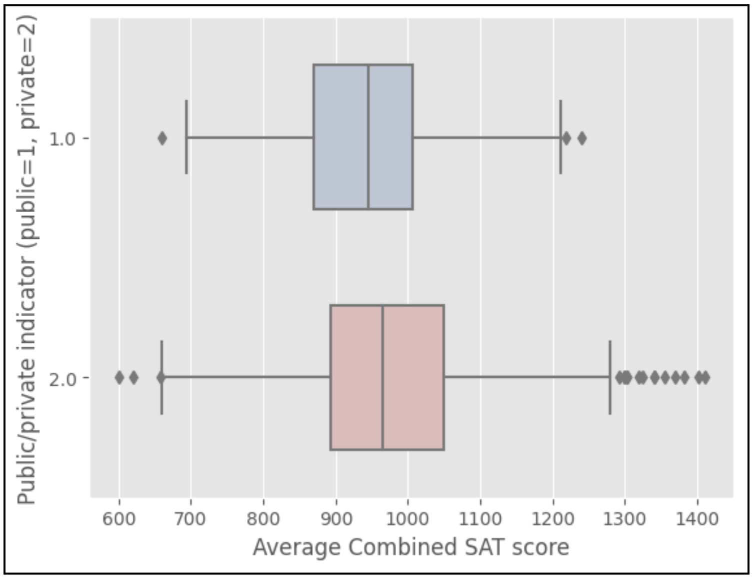


Fig. 9

❖ From fig. 10 and 11, it is quite clear that private colleges are very large in numbers compared to public colleges

❖ From the kernel density plot that both the distribution of SAT scores is the same just plot of private colleges is a scaled version of public colleges due significant difference in the number of public and private colleges.

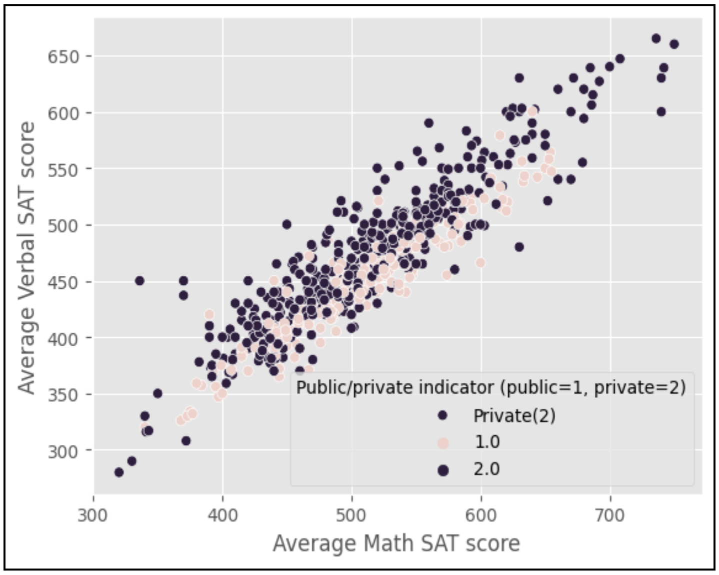


Fig. 10

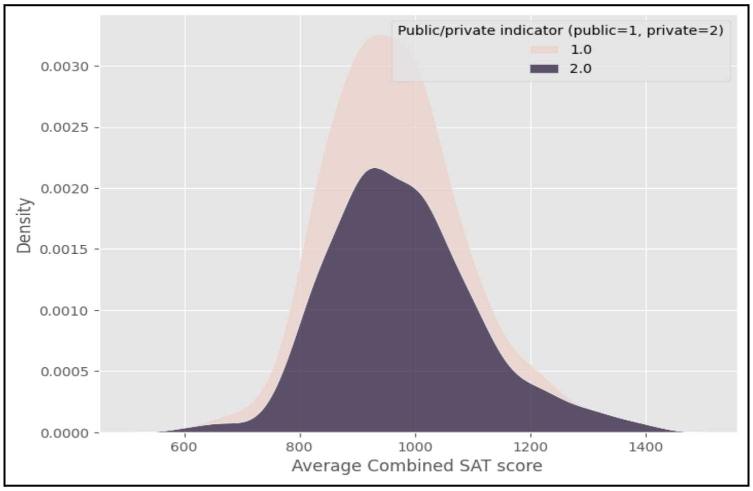


Fig.11 KDE plot of Combined SAT scores

C. *Is there a relationship between the percentage of faculty with PhDs and the percentage of new students joining college from the top 10% of H.S. classes?*

Hypothesis: Most students from the top 10% of H.S. tend to get admission to colleges with a higher % of Ph.D. faculties.

Approach:

1. Load data containing 10% new students and pct. Faculty with Ph.D.
2. Then make a scatter plot using matplotlib.pyplot.
3. Add a new column to the loaded data named 'type.'
4. With the help of this column, we will segregate colleges into two types, one having more than 70% of the faculty with PhDs and the other with 70% or less.
5. Then plot the box plot using the seaborn library.
6. Then plot kernel density plot using the kdeplot method in the seaborn library.

Solution:

- From the scatter plot in fig.12, we can see that pct. toppers are increasing with an increase in pct. of faculty with Ph.D. in colleges.
- Even some of the colleges having almost 100% of faculties with Ph.D. has all the new students from the top 10% of H.S. class.

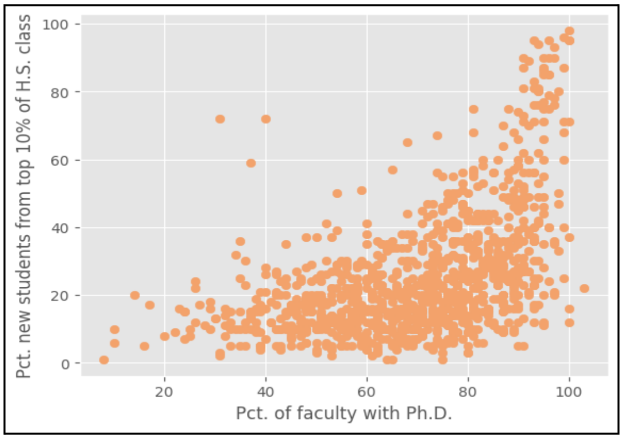


Fig. 12

- Fig. 13 gives a clearer picture, dividing colleges into two types based on pct. of Ph.. faculty.
- The median pct. of toppers is significantly large in colleges with a higher pct. of Ph.D. faculty than lower ones.
- Through the density plot in fig.14, we can see that the red line is always higher than the blue line, and the difference between them increases as the pct. of toppers increases.
- From the density plot, it can also be seen that almost no college with having Ph.D. faculty of less than 70% has more than 50% of the new students from the top 10% of H.S. class.
- Hence, we can say that our hypothesis is precisely correct. Most students from the top 10% of H.S. tend to get admission to colleges with a higher % of Ph.D. faculties.

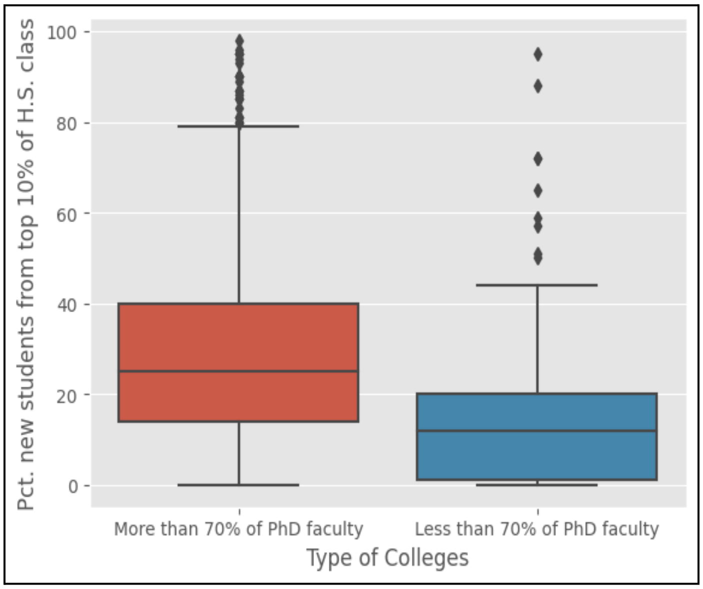


Fig. 13

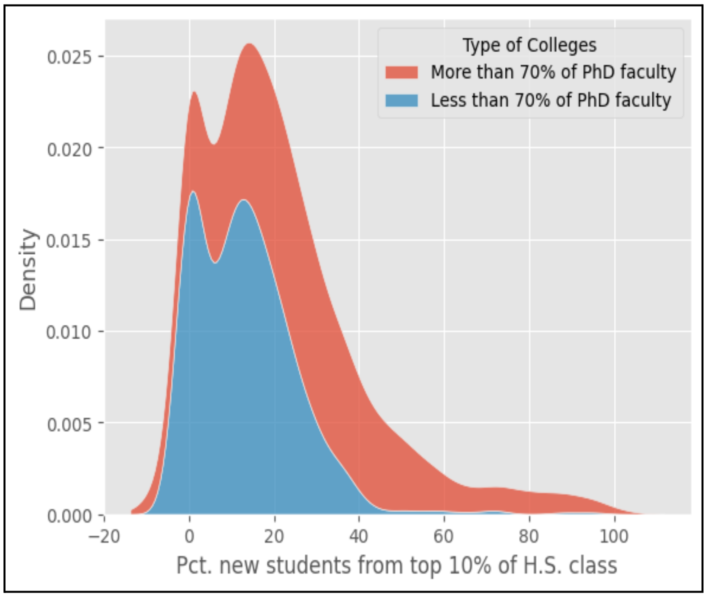


Fig. 14

D. *Which type of college, public or private, has a higher acceptance rate of applicants? Is there any correlation between the acceptance rate and the percentage of new students from the top 10% of H.S. classes joining the college?*

Hypothesis: The colleges with low acceptance rates should have high pct. of new students from the top 10% of H.S. class.

Approach:

1. Create a new column of acceptance rate in %, which is the ratio of no. of applications accepted to no. of applications received.
2. Plot kernel density plot of the acceptation using the seaborn library.
3. For the second question, use a scatter plot between the pct. of new students from the top 10% of H.S. class and the acceptance rate%.

Solution:

- ➢ Fig. 15 is the density plot of the acceptance % of private and public colleges.
- ➢ It is evident from the figure that public colleges have high acceptance rate than private colleges. So it can be said that going to get admission to public colleges is relatively easier than going to private colleges.
- ➢ One common thing that can be seen in the graph is that most of the colleges, private or public, have around an 80% acceptance rate since there is a very high peak at an 80% acceptance rate in the graph.

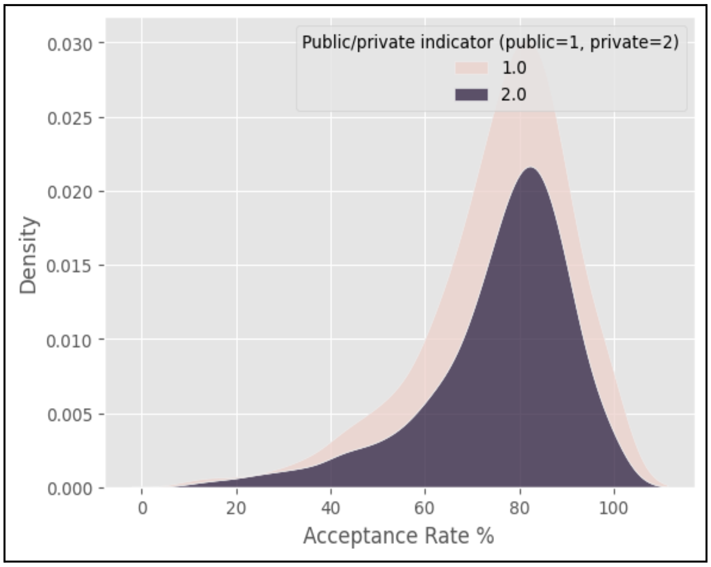


Fig. 15

- ➢ For the second part of the question, we used a scatter plot through which we can see that when the acceptance rate is very high, the pct. Of new students from the top 10% of H.S. class is very low, around 20% to 30%.
- ➢ While the colleges with a very low acceptance rate tend to have high pct. of new students from the top 10% of H.S. schools.
- ➢ Hence we can say that our hypothesis is correct, and it can be said that only toppers of H.S. class are able to make it up to the colleges having very lower acceptance rates.

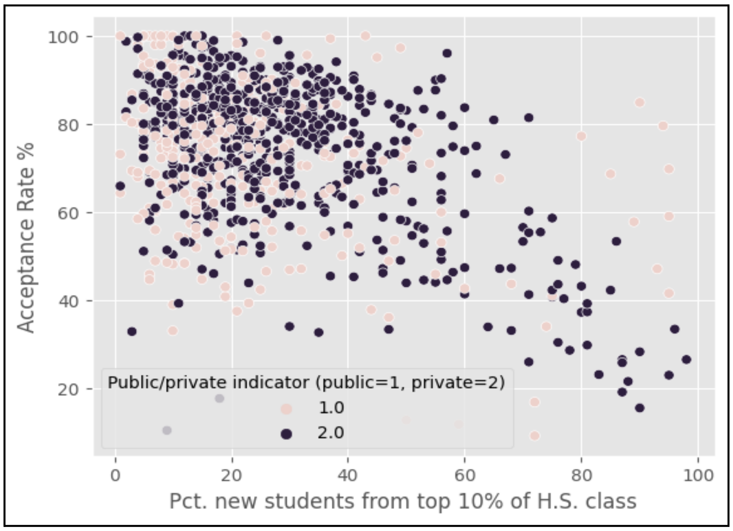


Fig.16

E. *Does the number of applications a college receives depend upon the total cost a student must pay to the college? Which college, public or private, is expensive to students?*

Approach:

1. First, calculate the total cost for a student, which is the sum of tuition fees, room, and board costs, and book costs.
2. Since the tuition fee differs for in-state and out-state students, we will take the average of both to proceed with the question.
3. Create a scatter plot between total cost $ and no. of applications received.
4. Also, create a density plot of the total cost for public and private colleges.

Solution:

- ➢ The distribution of points, i.e., colleges in fig. 17, seems normal. The number of students applying to different colleges is almost the same irrespective of the total cost.
- ➢ But there is a clear disparity between public and private colleges in case of the total cost a student needs to pay.

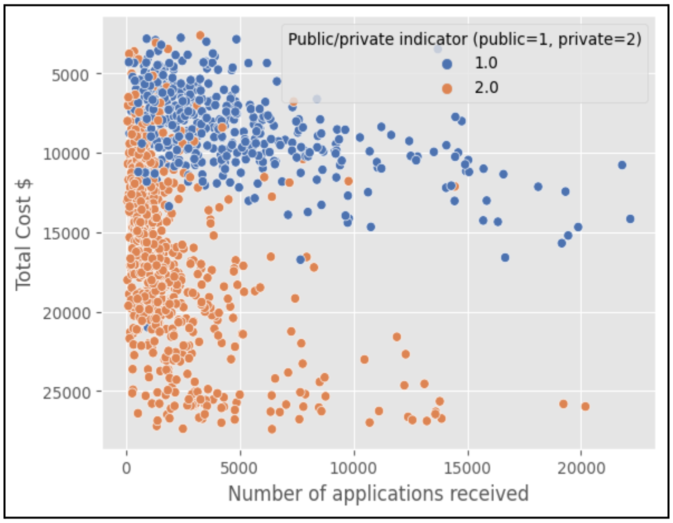


Fig. 17

- Fig. 18 gives a straightforward view of this disparity. The green area denotes the public college with a peak of around 7K $ - 8k $, while the purple plot is for private colleges, whose peak is around 13K $ - 15K $.
- Hence private colleges are more expensive than public colleges.

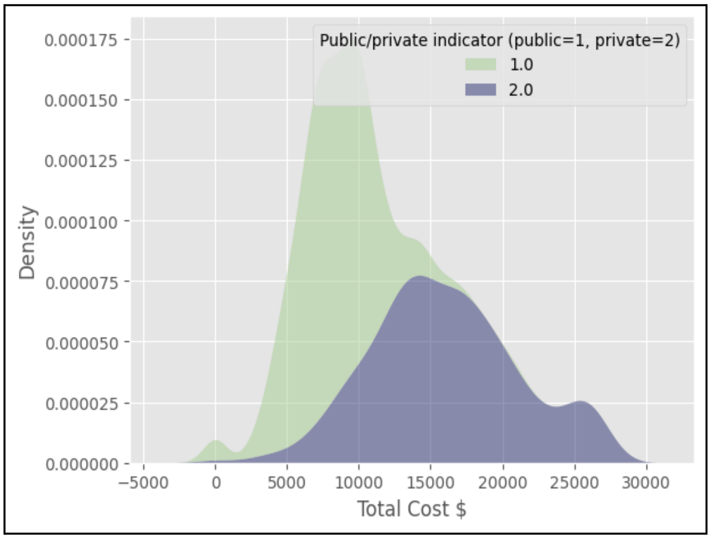


Fig. 18

III. DETAILS OF LIBRARIES

1. PANDAS: Pandas is a popular open-source Python library for data manipulation, analysis, and preparation. It provides fast and flexible data structures for structured data, such as tables, time series, and matrices, and offers tools for data cleaning, exploration, and visualization. The two primary data structures in pandas are Series and DataFrame.[2]

2. MATPLOTLIB: Matplotlib is a well-known open-source Python library for creating high-quality data visualizations. Line plots, scatter plots, bar charts, histograms, and other plot types are available. Matplotlib is built to work with NumPy arrays and can create static and interactive visualizations. Some major features of this library are easy customization supports, multiple output formats, etc.

3. NUMPY: NumPy is a widely used open-source Python library for numerical computation. It provides a quick and efficient way to work with numeric data arrays and matrices, as well as various tools for mathematical operations, random number generation, linear algebra, and more. Some key features are multidimensional array support, broadcasting, mathematical operations, and random number generation.

4. SEABORN: Seaborn is a popular open-source Python data visualization library. It is based on Matplotlib and offers a high-level interface for creating informative and appealing statistical graphics. Seaborn is built to work well with Pandas data structures and can easily handle large datasets.

IV. DETAILS OF FUNCTIONS

1. *read_csv*: The read_csv function is a method provided by the Pandas library in Python for reading and parsing data from a comma-separated values (CSV) file. It is a flexible function that can handle a wide range of input data formats and options.

2. *plt.plot:* The plt.plot function is a Python method provided by the Matplotlib library for creating line plots. It is a simple and flexible function that can handle a wide range of input data formats and options. The plt.plot function takes one or more arrays of x and y values and creates a data line plot. It provides various options for customizing the appearance of the plot, such as setting the color, line style, marker style, and label.

3. *value_counts():* The value_counts() function is a Python method the Pandas library provides for calculating the frequency of unique values in a Pandas Series. It returns a new Series object that contains the count of each unique value in the original Series.

4. *drop_duplicates():* The drop_duplicates() function is a Python method the Pandas library provides for removing duplicate rows from a data frame. It is a powerful tool for cleaning and preprocessing data and is commonly used in data analysis and exploration tasks.

5. *merge():* The merge() function is a method provided by the Pandas library in Python for combining two or more data frames based on a common set of columns. It is a powerful tool for merging and joining data from different sources and is commonly used in data analysis and exploration tasks.

6. *Scatterplot:* A scatter plot is a type of data visualization that displays the relationship between two variables in a dataset. It is a graph that uses dots or markers to represent individual data points, with the position of each dot representing the value of the two variables it corresponds to.

7. *Boxplot:* A box plot, also known as a box and whisker plot, is a type of graphical representation of data that displays the distribution of a dataset through its quartiles. It is a standardized way of visualizing data distribution based on five key summary statistics: the minimum value, the first quartile (Q1), the median, the third quartile (Q3), and the maximum value. In a box plot, a rectangular box is drawn from Q1 to Q3, with a line inside the box indicating the median value. Lines, or "whiskers," extend from the box to the minimum and maximum values, excluding outliers plotted as individual points beyond the whiskers.

8. *Kernel density plot:* A kernel density plot, also known as a density plot, is a graph showing the probability density function of a continuous variable. It is used to visualize the distribution of a dataset by estimating the probability density function of the underlying variable. A smooth curve is drawn over the data points in a kernel density plot, with the area under the curve representing the variable's probability density at each point. The shape of the curve depends on the bandwidth, or smoothing parameter, used in the kernel density estimation.

V. REFERENCE LIST

1. “Index of /datasets/colleges.” http://lib.stat.cmu.edu/datasets/colleges/

2. “United States Regions.” https://education.nationalgeographic.org/resource/united-states-regions/

3. “Pandas Documentation — Pandas 1.5.3 Documentation,” n.d. https://pandas.pydata.org/docs/

4. “Matplotlib — Visualization with Python,” n.d. https://matplotlib.org/.

5. “NumPy Documentation — NumPy v1.24 Manual,” n.d. https://numpy.org/doc/stable/

6. “Seaborn: Statistical Data Visualization — Seaborn 0.12.2 Documentation,” n.d. https://seaborn.pydata.org

VI. SUMMARY OF THE OBSERVATION

- In the Aaup dataset, first, we saw that Pennsylvania(PA) and New York had the most no. of colleges.
- After that, we compared the average salaries of the I, IIA, and IIB types of colleges, and the majority were the type I colleges.
- We also saw how the average salary increased with an increase in the number of faculty.
- In the US dataset, we compared the average SAT scores of public and private colleges. It found that getting into public colleges was relatively difficult since there are few colleges compared to private colleges.
- We also found the cost of each college which a student needs to pay.
- We also found that in which type of colleges students from the top 10% of H.S. classes prefer to go to. And it seemed that they preferred colleges having more Ph.D. faculty.

VII. ACKNOWLEDGEMENT

I sincerely thank Professor Shanmuga R. for his invaluable assistance completing this assignment. His guidance and support were instrumental in enabling me to develop a deeper understanding of the course material. His willingness to share his time, knowledge, and expertise with me was inspiring.