Data Narrative 2 Report

Name: Mohit Maurya
Discipline:B-tech M-tech Electrical Engineering

Roll number: 22110145

I. Overview of the Dataset

AAUP and USNEWS are the two most common files in the collection. Data for 1161 American colleges and universities are included in the AAUP dataset. The AAUP dataset includes information such as FICE, college name, kind, postal code, average wages for professors at all ranks in dollars, average pay for professors at all ranks in dollars, and the number of academics at various ranks.

Almost 1300 schools and institutions in the United States are represented in the USNEWS statistics. This dataset includes data that is similar to that in the AAUP dataset, such as FICE, college name, and other information like whether the college is public or private, average SAT scores for various subjects, average ACT scores, number of applications received, total number of students accepted, graduation rate, etc.

II. Scientific Question / Hypotheses for USNEWS Dataset

1) What is the likelihood of selecting a public school if I choose 10 school at random from the dataset?
2) How many colleges are there in different area of USA?
3) What is the range of SAT scores accepted by various institution across USA?
4) Do colleges with high Graduation rates make it harder to get accepted?
5) What are the comparison of average Graduation and Acceptance rates, as well as minimum and maximum Graduation and Acceptance rates between Public and Private colleges in USA?

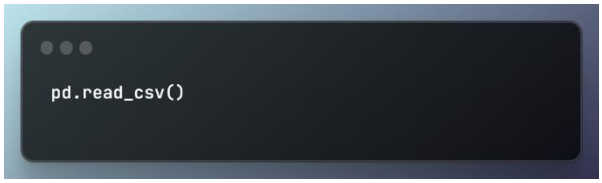
III. Details of Library and Functions

To answer the above scientific question, I have used three libraries: Pandas, Matplotlib, and Numpy.

1) The function used in library Pandas are

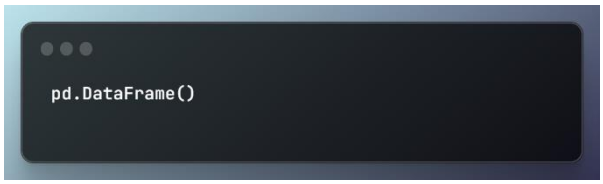
-

The use of this function is to read the excel file,



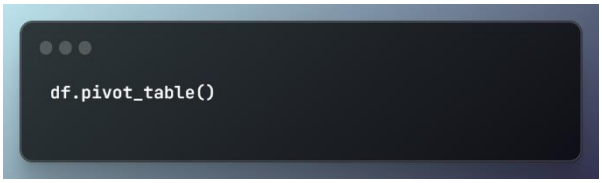
which is in .csv format; and return the dataframe of the information in that excel file

-



This function is used to convert dictionaries, lists, 2-D numpy, or series into dataframes

-



This function takes input like an index and a mathematical function like sum,count, mean, etc., and returns a dataframe whose index is the same as the specified index. That mathematical function is applied to the rest of the columns.

-
```
.head()
```
This function when used without specifying input, give first five rows of the dataframe; when information which is a number(n), is specified, it provides the first n row in the dataframe.

-
```
.sort_values()
```
This function, when provided with the columns of the dataframe it sorts the columns in ascending order along with the dataframe. If ascending =True is passed in the input, and if False is passed instead of True dataframe is sorted in descending order

-
```
.value_count()
```
This function returns the series of frequency/count of the unique values.

-
```
.index
```
It gives the index of the dataframe as an array

-

2) The function in matplotlib are

-
```
fig,ax=plt.subplots()
```
It plots the graph of the two values given as input. Also, it takes other inputs as color to change the color of the bar. Also it can create multiple subplots.

-
```
ax.set_xlabel()
```
Use to label the x axis

-
```
ax.set_title()
```
Use to label y axis

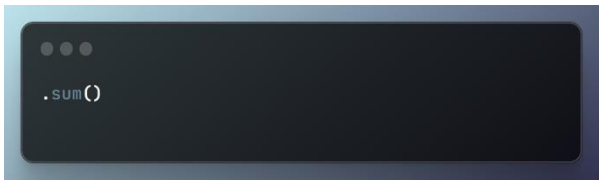
-
```
ax.text()
```
Use to add text on the graph.

3) The Function in Numpy are

-
```
np.where()
```
It provide the array of the values which satisfy the condition sepecified in as an input.

4) Some general function used are

•



Provide the sum of the input.

IV. ANSWER TO THE QUESTIONS FOR USNEWS DATASET.

1) What is the likelihood of selecting a public school if I choose 10 school at random from the dataset?

To respond to this query Using the binomial.pmf function, I calculated the probability of selecting a public or private school before plotting the pmf graph. Among the 10 schools, the probability of choosing one public school was discovered to be about 0.01135. In comparison to other 0.2456, the probability of selecting 3 public schools was higher.

The majority of students, as evidenced by this, prefer private schools to public ones.

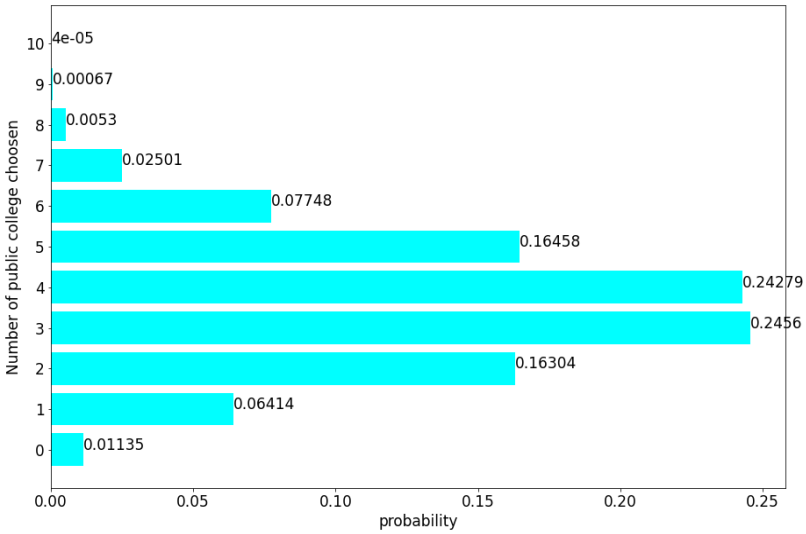


Fig. for question 1

2) How many colleges are there in different area of USA?

I produced the dataframe in order to provide an answer. In the dataset, I counted the number of unique postal codes, followed by the number of colleges in that postal region in the USA, then plotted the bar graph of those two data points.

The graph showed that the 101 institutions are located in postal code NY, which is New York's second-most populous region in the United States after New York is Pennsylvania (PA). Similarly on 3rd position is California (CA).

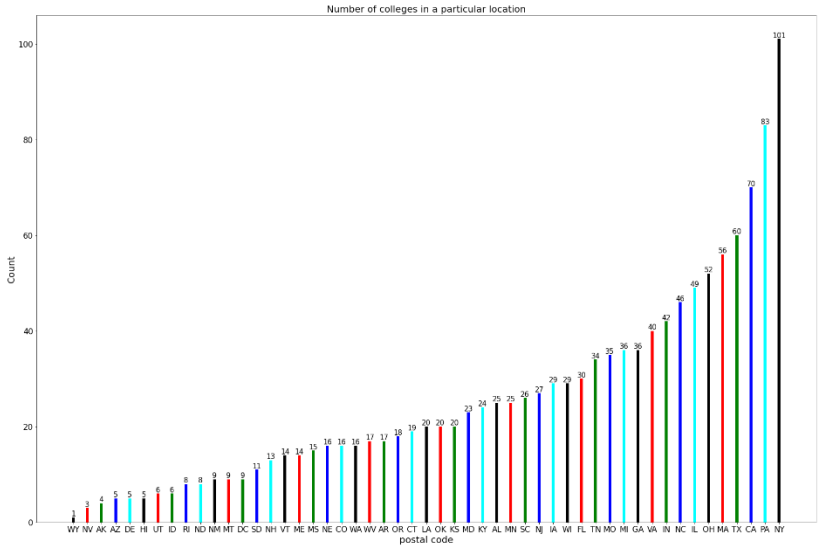


Fig. for question 2

3) What is the range of SAT scores accepted by various institution across USA?

In this I use the previous question data and took out the top 15 region which has most number of colleges and bottom 15 region which has least number of colleges, then calculated the maximum and minimum SAT scores required to get into any of the colleges situated in the region.

I founded that there is more variation in scores in region where there more colleges. Maximum score is almost twice the minimum score in such region. Where as there was not much variation in scores in the region which has least number of colleges.

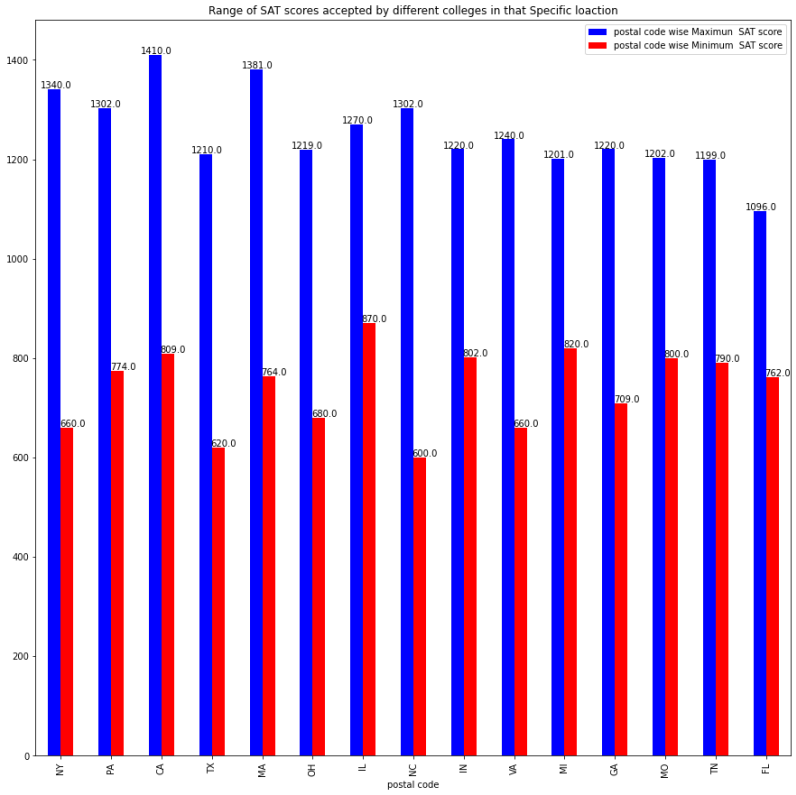


Fig. Top 15 colleges

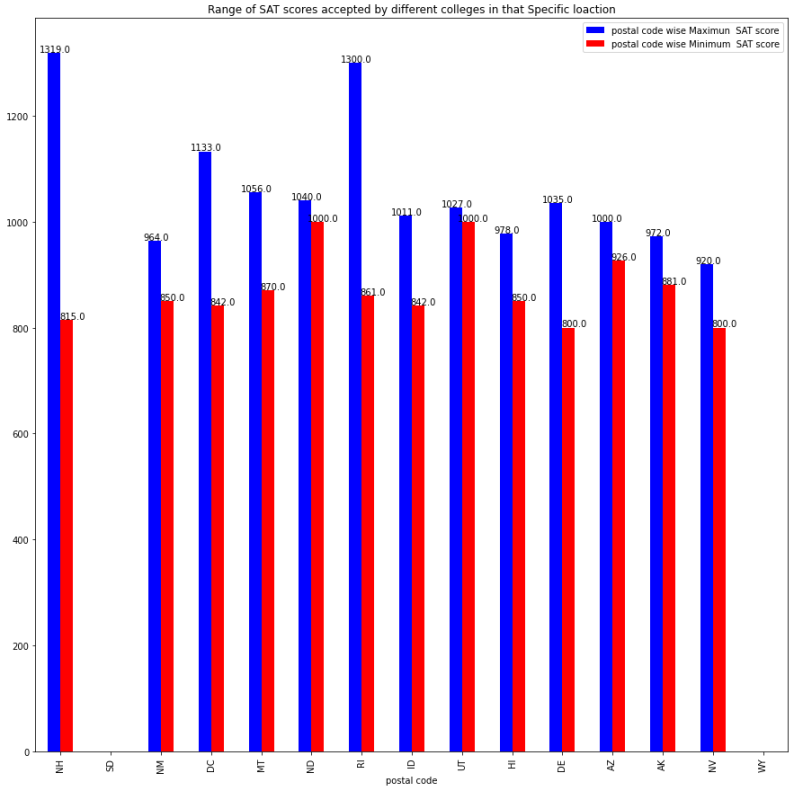


Fig. Bottom 15 colleges

4) Do colleges with high Graduation rates make it harder to get accepted?

In order to answer this question I calculated the acceptance rated of colleges by dividing the number of application accepted by colleges with number of application received . I created 2 plots of acceptance rate plotted against graduation rate but one according to the graduation rates for top 20 colleges and 2nd according to the acceptance rate for top 20 colleges.

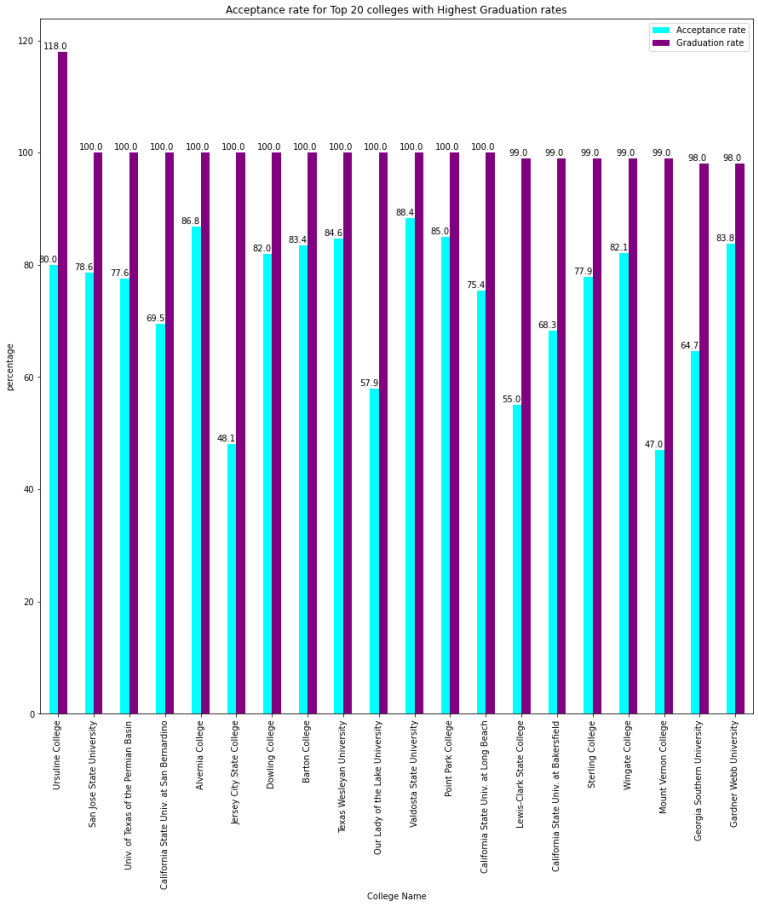
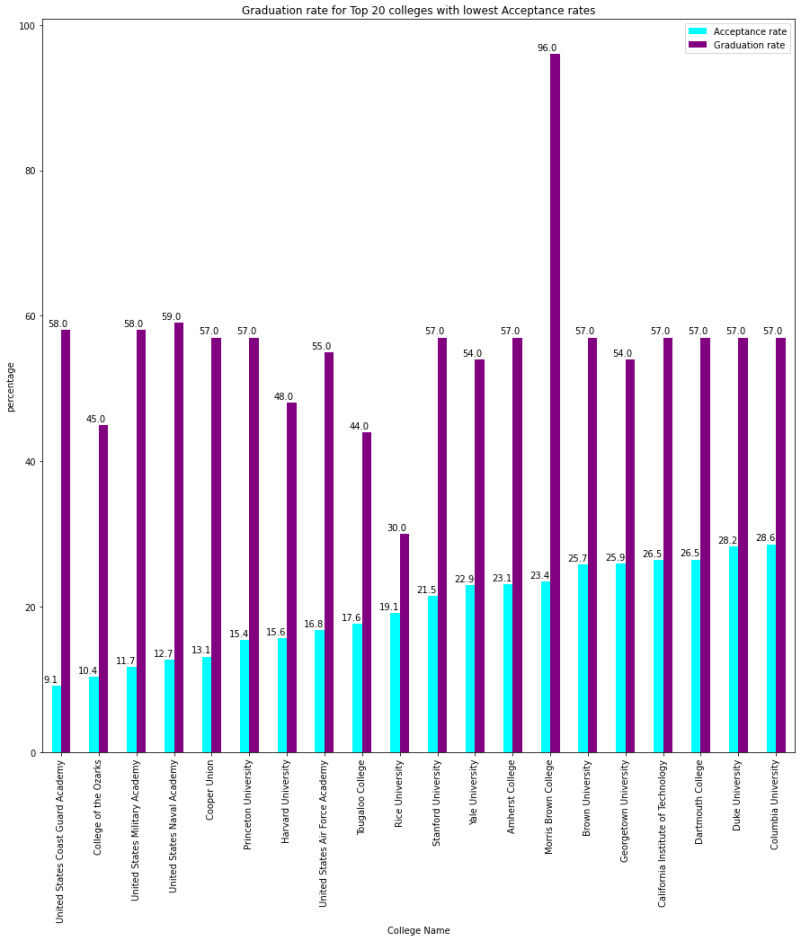


Fig. Colleges sorted according to graduation rates

Fig. Colleges sorted according to Acceptance rates



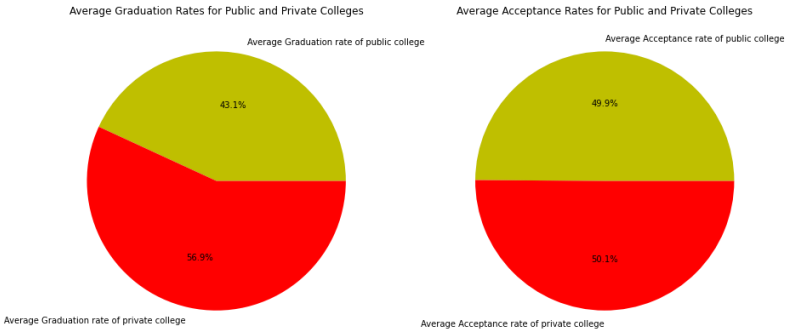


Fig. Pie chart of average Graduation and Acceptance rates

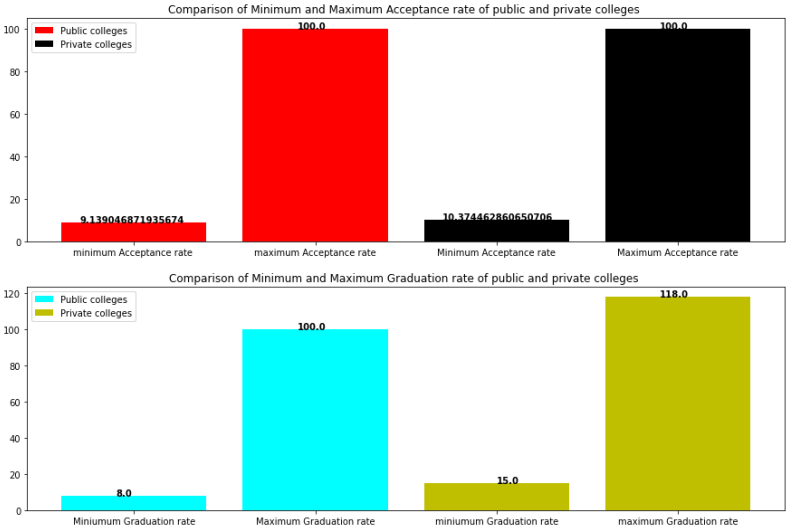


Fig. Bar graph of minimum and maximum Graduation and Acceptance rates

First of all 1st graph shows that according to data its not very tough to get admission in colleges which has high Graduation rates as average acceptance rate is about 80%.

But 2nd graph shows that there many colleges which have low acceptance rate too don't have very high graduation rates

5) What are the comparison of average Graduation and Acceptance rates, as well as minimum and maximum Graduation and Acceptance rates between Public and Private colleges in USA?

For this I different function and plotted pie chart for average graduation rate, acceptance rates of public and private colleges I also plotted bar graph for maximum and minimum graduation rates and acceptance rates for public and private.

V. Scientific Question / Hypotheses for AAUP dataset

1) What is the liklihood of selecting certain typpes of colleges (IIB,IIA,I)?

2) Which kind of colleges pay their faculty the highest aveage salaries and compensation?

3) What is the comparasion of the professor's salaries and compensation's at public and private colleges?

4) What are the highest and lowest faculty salaries, benefits, and staffing levels at public and private colleges?

5) What is the distribution of the entire faculty according to salary range?

VI. Answer to the questions for AAUP dataset.

1) What is the liklihood of selecting certain typpes of colleges (IIB,IIA,I)?

In order to answer this question I calculated probability of colleges of type IIB,IIA,I by dividing the number IIB,IIA,I colleges by total number of colleges and then I plotted the probability

According to graph IIB type college is more probable to get ramdomly selected than IIA type college and I type of colleges

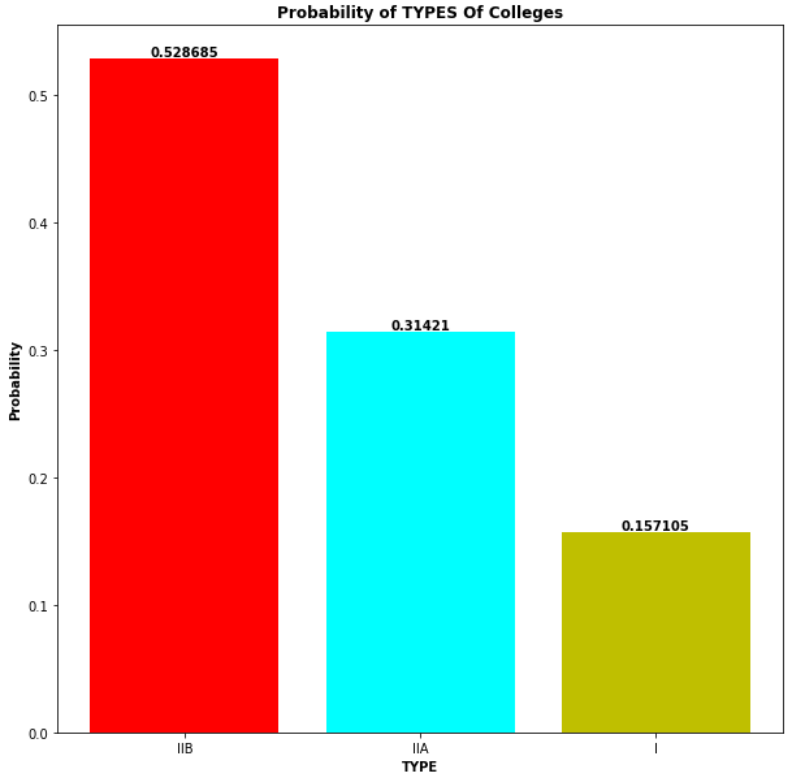


Fig. for question 1

2) Which kind of colleges pay their faculty the highest aveage salaries and compensation?

To answer above question I created a datafrsme where I separately calculated the professor salaries and compensation and then plotted it.

From graph it was clear that professor form colleges type IIB are highly paid followed by IIA type colleges and lastly I type college professor.

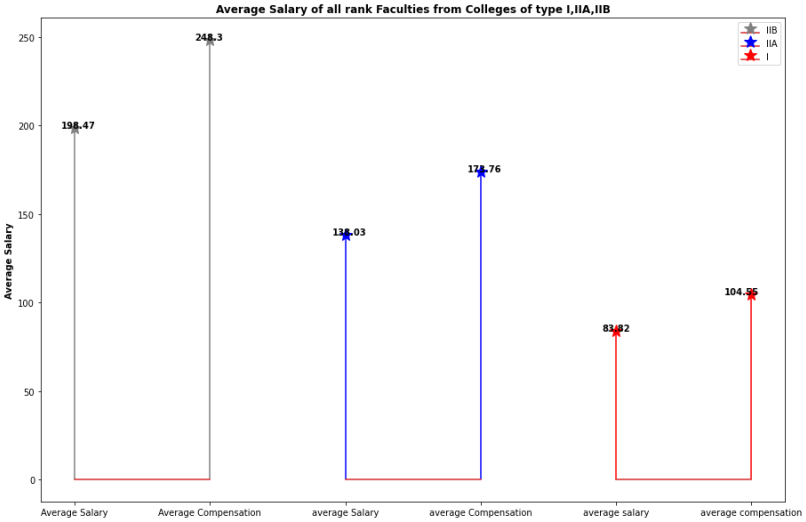


Fig. for question 2

3) What is the comparasion of the professor's salaries and compensation's at public and private colleges?

By comparision of salaries and compensation between public and private college through a bar graph it was found that public colleges pay more to faculties as compared to private colleges also they give more compensation to them

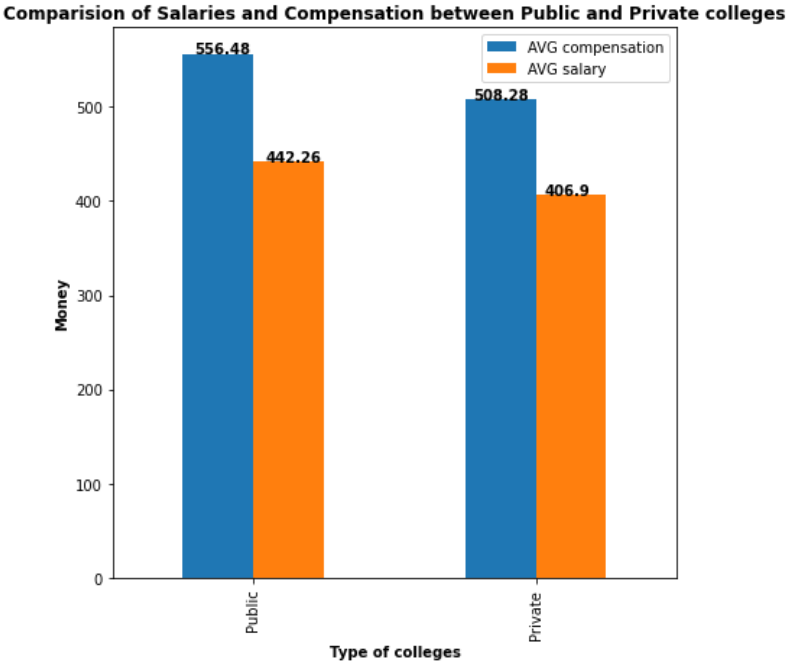


Fig. for question 3

4) What are the highest and lowest faculty salaries, benefits, and staffing levels at public and private colleges?

To answer this question I plotted joint bar graph of minimum and maximum salaries, compensation of professor number of professors.
And from the plot it was clear that maximum salaries are offered in private colleges same goes for minimum to and also point is to note that the have less number of professor as compared to public colleges then to the have less average salaries as pervoisuly answered which means that there is a lot of variation in salaries of professors which is not the case for public college.

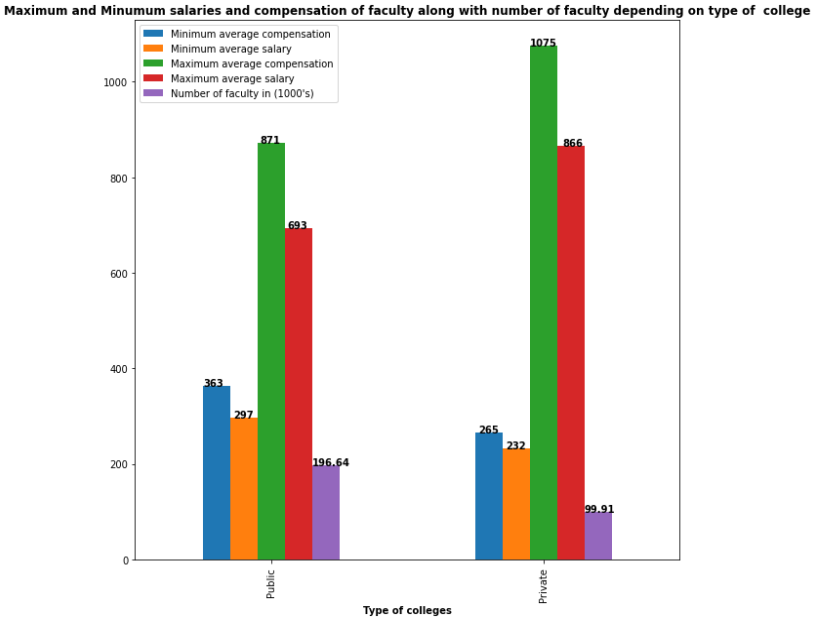


Fig. for question 4

5) What is the distribution of the entire faculty according to salary range?

To find the range in which majority of the professor get paid I first decided the range of the intervals when counted the number of professor which get paid in that particular range

I founded that majority of the professor get paid between 50000 to 55000 dollar's yearly according to the table provided below

salary range	Number of faculty
(200, 250]	74
(250, 300]	3765
(300, 350]	16873
(350, 400]	36950
(400, 450]	56901
(450, 500]	53733
(500, 550]	57984
(550, 600]	38224
(600, 650]	19198
(650, 700]	5467
(700, 750]	4136
(750, 800]	2161
(800, 850]	834
(850, 900]	257

Table of ranges of the salaries of faculty

VII. Summary of the Observations

In conclusion, the AAUP and USNEWS datasets offer insight on American colleges and universities. With the help of tools like Numpy, Pandas, Scipy and Matplotlib, we were able to provide answers to a number of scientific questions and hypotheses.

we discovered that there was a low possibility of choosing a public school and a higher likely of choosing a private school. Also New York, Pennsylvania, and California were the US states having the most colleges per capita. In addition, we discovered that locations with more colleges have more variation in the SAT scores required for admission than regions with fewer colleges.

There was no as such corelation between Acceptance rate and Graduation rate. Overall, this investigation offers insightful knowledge into the American higher education landscape.

VIII. References

1) "Matplotlib 3.7.1 Documentation#." Matplotlib documentation - Matplotlib 3.7.1 documentation. Accessed March 27, 2023. https://matplotlib.org/stable/index.html

2) NumPy documentation. Accessed March 27, 2023. https://numpy.org/doc/

3) "Pandas Documentation#." pandas documentation - pandas 1.5.3 documentation. Accessed March 27, 2023. https://pandas.pydata.org/docs/

4) "Pandas Tutorial." GeeksforGeeks. GeeksforGeeks, March 17, 2023. https://www.geeksforgeeks.org/pandas-tutorial/

5) "Scipy Documentation#." SciPy documentation - SciPy v1.10.1 Manual. Accessed March 27, 2023. https://docs.scipy.org/doc/scipy/

6) "W3Schools Free Online Web Tutorials." W3Schools Online Web Tutorials. Accessed March 27, 2023. https://www.w3schools.com/

IX. Acknowledgements

1) *Thanks to all TA's for hepling me in DATA NARRATIVE.*
2) *Thanks to Piyush singh.*
3) *Thanks to Hari balaji.*
4) *Special thanks to Mithlesh tandon.*
5) *Special thanls to Famida sayyed.*