REPORT: Data Narrative 2

NAME: MRUGANK PATIL
DEPARTMENT: COMPUTER SCIENCE AND ENGINEERING
ROLL NO: 22110158

I. Overview of the 1st Dataset

A. *The first provided dataset includes statistics on the number of different types of professors employed by a specific college in a specific state, as well as the typical salaries of those professors. This data is helpful for various things such as calculating the pf of the professors , college and faculty statistics comparison, etc.*

II. Overview of the 2nd Dataset

A. *This dataset includes data on universities in the US, such as their FICE code, name, state, whether they are public or private, SAT and ACT scores, enrollment information, tuition expenses, room and board costs, and other details. The dataset can be utilised for a number of tasks, such as trend analysis, college statistics comparison, and data-driven decision making.*

III. SCIENTIFIC QUESTIONS

A. *Question 1*

Determine the PDF and CDF of X, where X is a random variable representing the number of faculty of all ranks at a college, and use this information to calculate the probability of a college having fewer than 20 faculty of all ranks?.

B. *Question 2 / (Hypothesis)*

The probability that a college has more than 10 associate professors should be more than 0.5.

C. *Question 3*

What is the probability of number of colleges depending upon the state with highest average salary for full professor

D. *Question 4*

Suppose we randomly select colleges from the dataset until we find a college that has more than 100 Assistant Professor. What is the expected number of colleges we need to select before finding such a college? (geometrical random distribution)

E. *Question 5*

What is the probability that a randomly selected college from the state of Texas has more than 30 full professors? (binomial random distribution)

F. *Question 6 / (Hypothesis)*

Student prefer to go to public colleges instead of private as private colleges require more fees than public

G. *Question 7*

Compare the probability that a randomly chosen college from the dataset is located in the state of Texas vs New York and has an average math SAT score above 600? (geometrical)

H. *Question 8*

What is the probability that a randomly college selected has a graduation rate of at least 80% with student to faculty ratio is less than 15:1?

I. *Question 9 / (Hypothesis)*

1) The number of colleges in which student have more marks in ACT exam have phd faculty in there college.

J. *Question 10 / (Hypothesis)*

The acceptance rate, tuition, and room and board prices of a college are all related.

IV. DETAILS ON LIBRARIES AND FUNCTION

A. *read_csv(): This function is used to read a CSV file and load it into a pandas DataFrame.*

B. *Pandas : A Python package for data analysis and manipulation is called Pandas[2].*

C. *matplotlib.pyplot : Creating static, animated, and interactive visualisations in Python is made possible by the Matplotlib.pyplot module[3] .*

D. *sum() : gives back the total of all values in the given array or along a particular axis in a multi-dimensional array.*

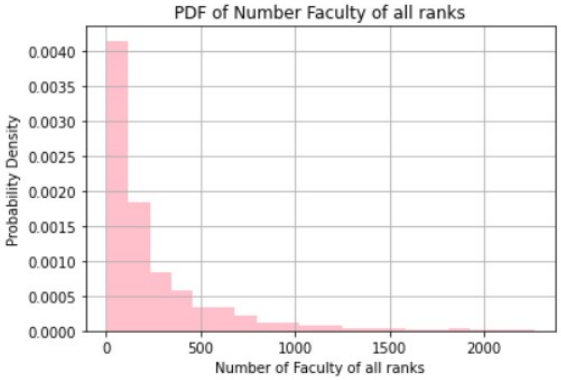
E. *groupby() : Enables the application of aggregate functions to data grouped by one or more columns in a Pandas DataFrame.*

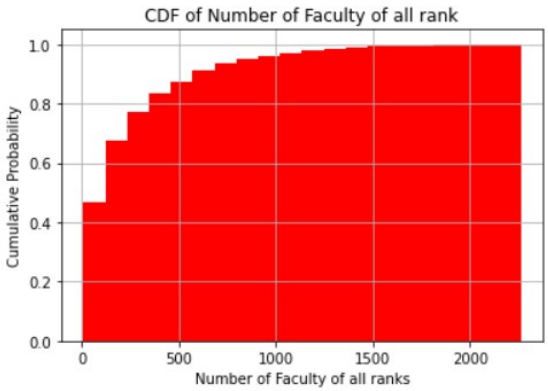
F. *merge() : The Pandas merge() function combines two or more DataFrames based on a common key and supports several join types, including inner, outer, left, and right joins.*

G. *nunique() : Pandas has a function called nunique() that gives the total number of distinct values in a column or series of a DataFrame.*

V. ANSWERS TO THE QUESTIONS

A. ANSWER 1: As we can see that is probability that the college has fewer than 20 faculty of all rank is very less states that it is ther are more than 20 faculty of all rank in the college which can be visualise by the PDF and CDF of the the given questions.

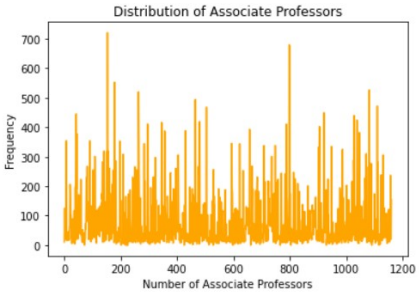





```
Probability that a college has fewer than 20 faculty of all rank: 0.008613264427217916
```

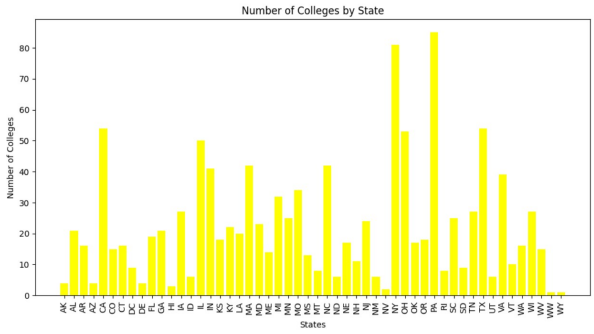
B. ANSWER 2 :As we can see that the probability of a college that has more than 10 associate professors is more than 0.5. Also from the graph we can see that number of associate professor is more in the range from 200 to 800 which implies that there are more associate professors. So our hypothesis is correct.

```
Probability that a college has more than 10 associate professors: 0.91
```



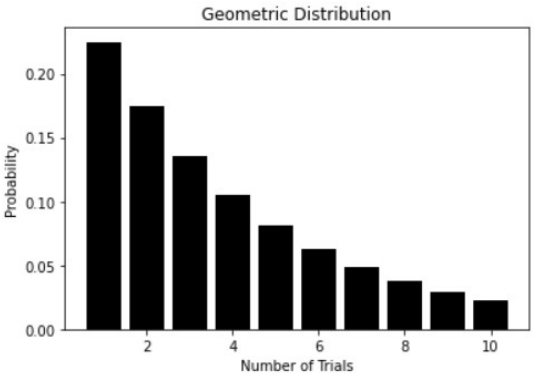
C. ANSWER 3 : The probability of the number of colleges in the state with highest salary is approx 0.046 which implies there are very less number of colleges for a particular state with a high salary. From the graph we can conclude that the there are almost 10 to 30 colleges in a particular state from where we can find only few colleges with high salaries.

```
The probability of the number of colleges in the state with the highest salary is: 0.046511627906976744
```



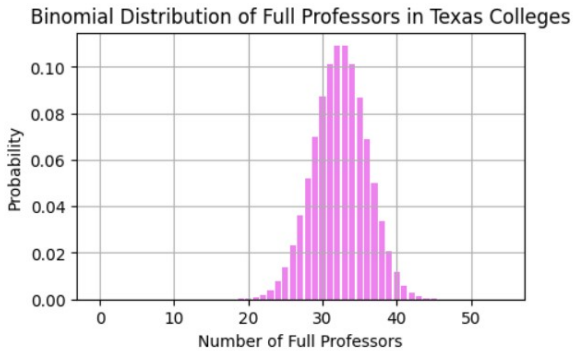
D. ANSWER 4: The expected number of colleges we need to select before finding a college with more than 100 Assistant Professor is 5. This implies that we need to have less trails to find that college. From the graph we can conclude that the probability to find the college is more in the initial trails.

```
The expected number of colleges we need to select before finding a college with more than 100 Assistant Professor is: 5.0
```

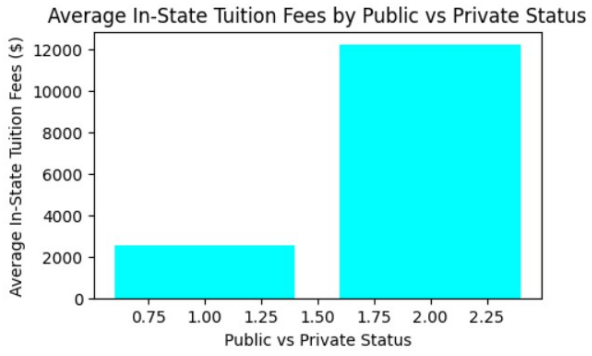


E. ANSWER 5 : The probability that the randomly selected college from Texas having more than 30 full professors is approx 0.7. This means that there is a high percentage to find that type of college easily. This can be visualise from the given graph of binomial distribution given below.

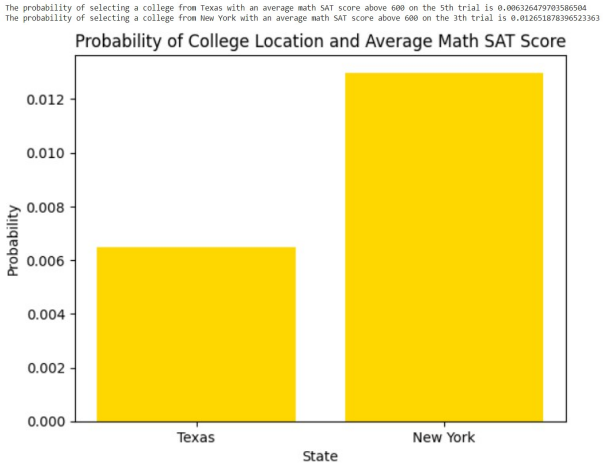
```
The probability of a randomly selected college from Texas having more than 30 full professors is: 0.7033006384501113
```



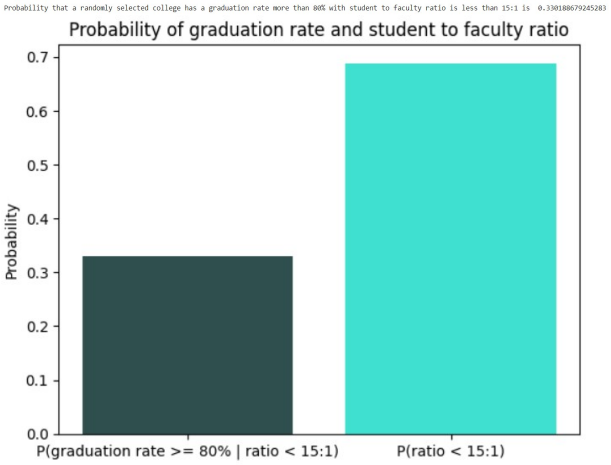
F. *ANSWER 6 : From the graph we can conclude that student prefer to go to public colleges instead of private as public college require less fees. The first bar represent the average tution fees of the public college and the second bar represent fees of the private college which is more than the public. So our hypothesis is correct.*



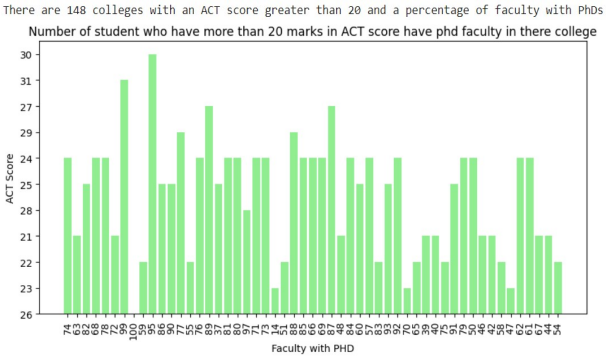
G. *ANSWER 7 : Probability of college location and average math SAT score between Texas and New York can be seen from the graph. It states that texas has less probability of math SAT score to be more than 600 than New York. Also the probability of selecting a college from Texas with avg math SAT score to be more than 600 on the 5th trail is approx 0.0063 and probability of selecting a college from New York with score more than 600 on its 3rd trail is 0.0126.*



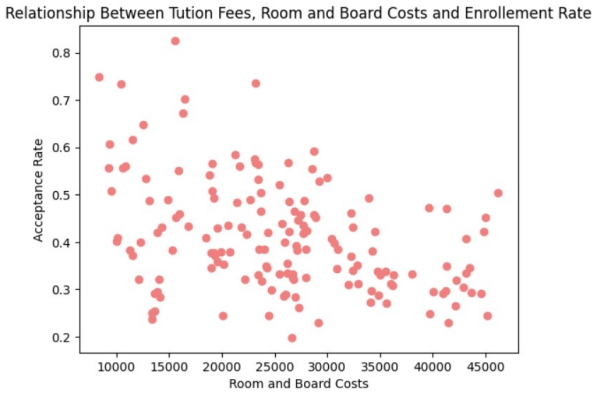
H. *ANSWER 8 : The probability that randomly selected collge has a graduation rate more than 80% with student to faculty ratio less than 15:1 is 0.33. Thus it conclude that more the interation between the student and the faculty is their more will be the understanding of the concept of the syllabus.*



I. *ANSWER 9 : The number of colleges in which student have more marks in ACT exam have PHD faculty in their college are 148. This could be visualise from the graph given below. This conclude that more the experienced PHD faculty teaches the student, more the better understanding of the topic. Hence our hypothesis is correct.*



J. *ANSWER 10 : From the graph we observe that from the range of 10k to 35k the scatter points are more which conclude that majority of student enroll more in that college in which the college expenditure is less for them. So our hypothesis is correct.*



VI. SUMMARY OF THE OBSERVATION

A. *OBSERVATION 1: The fact that probability of colleges having more than 20 all rank faculty is very less state that the the faculty is more in other fields.*

B. *OBSERVATION 2: The probability that a college has more than 10 associate professors is more than 0.5 means that number of such type of professor is large in colleges.*

C. *OBSERVATION 3: There are extremely few schools for a given state with a high income, according to the likelihood of the number of colleges in the state with the highest salary, which is approximately 0.046. According to the graph, there are roughly 10 to 30 institutions in a given state, but only a small number of them offer significant incomes.*

D. *OBSERVATION 4: We anticipate that we will need to choose five institutions before discovering one with more than 100 assistant professors. This suggests that there should be fewer trails required to locate that college. We can infer from the graph that the early trails have a higher chance of leading to the college. This establishes that there will be more assistant professors in more colleges.*

E. *OBSERVATION 5: There is a high chance of finding a college in California with more than 30 full professors. So, it should not be difficult to find this type of college in California.*

F. *OBSERVATION 6: Student prefer public colleges over private as it require less fees. This conclude than student think about their finacial status and wanted to save money.*

G. *OBSERVATION 7: The math SAT score in Texas is less than the state of New York means that students in New York states are good in studies.*

H. *OBSERVATION 8: From observing the graph we conclude that their should be less student ot faculty ratio in the school and colleges as their will be better understanding of the concept among the students because faculty could pay attention to each student in a better way.*

I. *OBSERVATION 9: There are over 140 colleges in which students who have performed well on the ACT are taught by faculty members holding a PhD. This implies that having faculty members with extensive knowledge and experience in their respective fields can help students gain a deeper understanding of the concepts being taught.*

J. *OBSERVATION 10: The financial status of the students is important for choosing college. As very less student will choose college with high fees and other expenditure. This can be observed from the graph given.*

VII. REFERENCES

A. *[2] "Pandas Documentation." pandas documentation - pandas 1.5.3 documentation. Accessed February 23, 2023. https://pandas.pydata.org/docs/.*

B. *[3] "Matplotlib 3.7.0 Documentation." Matplotlib documentation -Matplotlib 3.7.0 documentation. Accessed February 23, 2023.https://matplotlib.org/stable/index.html.*

C. *NumPy documentation. Accessed February 23, 2023.https://numpy.org/doc/*

VIII. Acknowlegement

A. *The Python Software Foundation, which is responsible for the creation and upkeep of the Python programming language.*

B. *The designers and collaborators of the Pandas, Matplotlib, and Numpy libraries for their support of the Python ecosystem.*

C. *Thank you to each and every one of my teachers for teaching me and assisting me in reaching my goals. Thank you to my friends and other students for creating moments with me that I will always treasure.*

D. *I would like to thank GeeksforGeeks and w3schools for offering such a fantastic environment for learning and developing programming skills. That is very fantastic that you are so committed to make computer science available to everyone.*

E. *I want to express my gratitude to my mother for her unwavering love, unending help, and constant presence in my life. I appreciate everything you have done for me and for being my life's inspiration.*