ES114 Probability,Statistics and Data Visualization Report: Data Narrative 2

Keshav Sobania, *B TECH 22,Civil Engineering, ROLL NO. 22110118*

Dataset 1

A. Overview of The Dataset:-

The dataset includes statistics on faculty members' salaries and numbers for various institutions. The dataset was used to compare the salary and benefits offered by multiple establishment types and the number of faculty members at different ranks.

The state and postal code for each institution is contained in the "State(Postal code)" variable, and a particular institution is public or private is indicated by the "Type" field.

Each institution's unique identification, known as the FICE variable

B. Scientific Questions/Hypothesis:-

Question 1: what is the probability that a randomly chosen college has average salary of a full professor greater than 500 and given that it is a type IIB college.

Question 2: Given that college type is I or IIA then what is the probability that a randomly chosen college has an average salary of all ranks per number of professors of all ranks greater than 5?

Question 3: Make a pie chart on distribution of colleges of their type.

Question 4: Make a Dataframe that shows how many unique states(postal code) are in each type and show them in a bar graph.

Question 5: Find the different college name which has the highest “Number of full professor”, “Number of associate professors”, “Number of assistant professors” respectively.

C) Details of libraries and functions:-

a)libraries:-

- **Numpy-**Python Library used for working with arrays containing various functions for working out on matrices,linear algebra and many such topics.

- **Pandas-**it is used in analyzing data and help in make dataframe similar to spreadsheet

- **Matplotlib**-Very helpful in visualizing data by plotting it in various ways such as histograms,bar graphs,pie charts etc

b)Functions:-

- Loops, indexes and if else functions have been used by me for deducing the answers.

D) Answers to the Questions:-

1. Let ,

P(A/B) = the probability of college has an average salary of a full professor greater than 500 and given that it is a type IIB college
where ,
A = number of colleges having an average salary of a full professor greater than 500.
B=number of college has type IIB
P(A/B) = P(B and A) / P(B)
By code , we found that
P(A/B) = 0.6726094003241491

2.
Let ,
P(B/A) = probability of college having an average salary of all ranks per number of professors of all ranks greater than 5 and given college type is I or IIA.

where ,
A =number of college having an average salary of all ranks per number of professors of all ranks greater than 5
B = number of college having type I or IIA
P(A/B) = P(B and A) / P(B)
By code , we found that
P(A/B) = 0.055248618784530384

3. We will plot a pie on the distribution of colleges of their type.

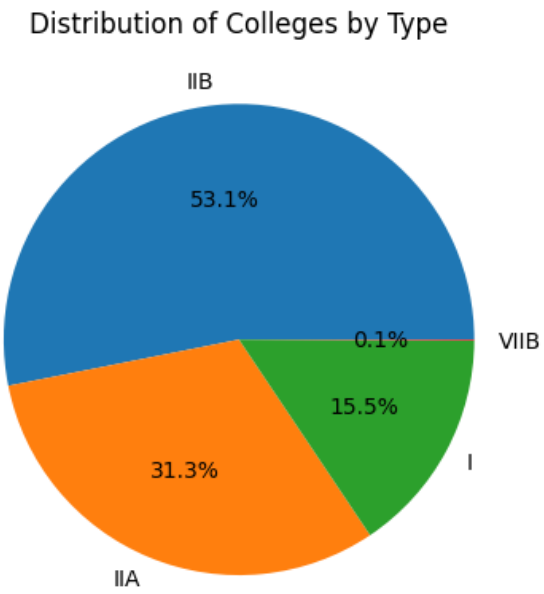
So first we count the number of colleges of different types by function count in pandas and make a data frame.

IIB 617

IIA 363

I 180

VIIB 1

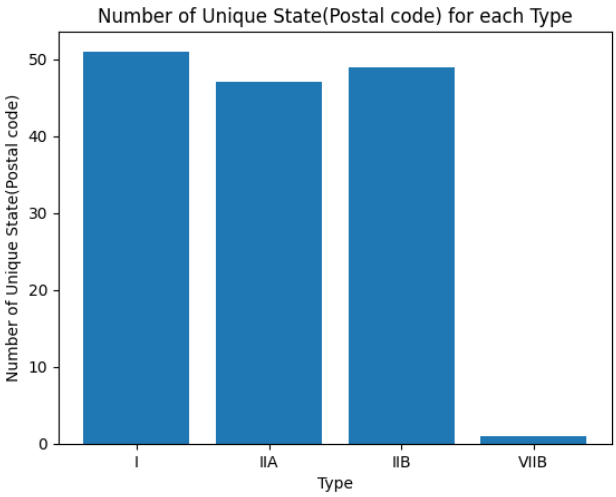


By help of code , we plotted the pie chart

4.
Dataframe on number of unique states in each type .
Number of unique states in each type

Type	
I	51
IIA	47
IIB	49
VIIB	1

Plotted a bar chart by help of code on above dataframe



5.

College with the highest number of full professors: Univ. of Texas at Austin

College with the highest number of associate professors: University of Florida

College with the highest number of assistant professors: University of Florida

College with the highest number of instructors: Louisiana St.Univ.and A&M C.

Dataset 2

A. Overview of The Dataset:-

The dataset offers details on colleges, focusing on enrolment and admissions figures.

This dataset can be used to analyze enrollment and enrollment patterns at various colleges and universities and explore the relationship between admission criteria and research outcomes. It can also be used to compare the characteristics of public and private institutions and identify patterns and trends in enrollment and enrollment over time.

B. **ScientificQuestions/Hypothesis:-**

Question 1:What is the probability that a randomly chosen college is in a state that starts with the letter "A" or "W"?

Question 2: Make a line graph that represents how many colleges are in a particular state.

Question 3: If we randomly select two colleges from the dataset then What is the probability that the first college is private and the second college is public?

Question 4: Plot a bar graph of top 20 colleges which have the highest average act score.

Question 5: Find the different college name which has the highest 'Number of applications received','Number of applicants accepted','Number of new students enrolled','Number of full time undergraduates'.

C) Details of libraries and functions:-

a)libraries:-

- **Numpy-**Python Library used for working with arrays containing various functions for working out on matrices,linear algebra and many such topics.

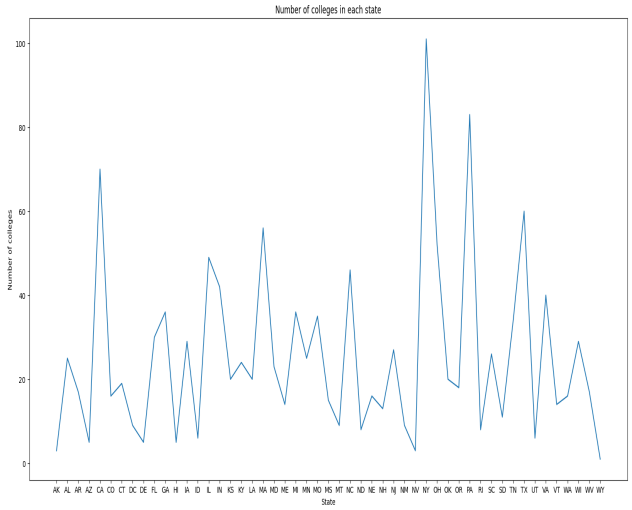
- **Pandas-**it is used in analyzing

data and help in make dataframe similar to spreadsheet

● **Matplotlib**-Very helpful in visualizing data by plotting it in various ways such as histograms,bar graphs,pie charts etc

b)Functions:-

● Loops, indexes and if else functions have been used by me for deducing the answers.



D) Answers to the Questions:-

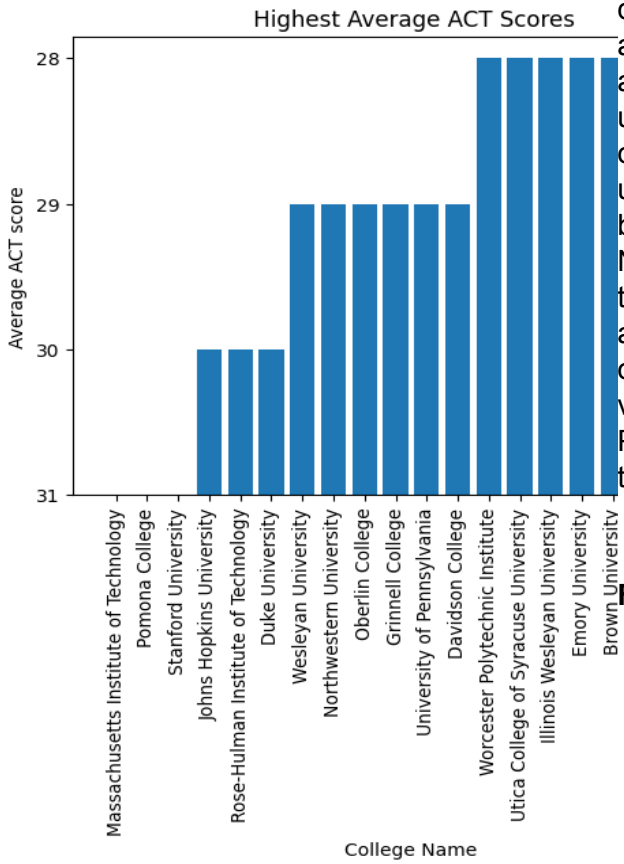
1.A=number of total college
B=number of college in state start with A and W
By code ,we found that
A=1301
B=113
So,probability =A/B
0.08685626441199078

2.

Line graphs that represent how many colleges are in a particular state.

3.A=No of total college
B=No. of private colleges
C=No. of public colleges
First college is private and second is public
p1=B/A
p2=C/A
prob=p1*p2
By code ,we found
A=1301
B=831
C=470
prob=0.2307513702284236

4. Bar graph of top 20 colleges which have the highest average act score by using matplot and sort the array by sort_values() function in pandas and select top 20



5.College with the highest number of application received: Westminster College

College with the highest number of application accepted: Southern Illinois University at Edwardsville

College with the highest number of students enrolled: Hampshire College

College with the highest number of full time undergraduates: University of Arkansas at Fayetteville

E)Summary to the observation:-

I've come to the conclusion that a Python library can be used to read CSV files like pandas containing very large amounts of data. You can also create a data frame and draw a chart using this data. You can also perform many operations on this data using pandas functions. Also, there is an open source library called matplotlib that is used to plot graphs such as histograms, bar charts, pie charts, scatter plots. Numpy is also an open-source library for the Python programming language that allows working with matrices and one-dimensional arrays. Seaborn is also a very popular open source library in Python. Used for data visualization. Used to create informative statistical charts.

F) References:-

- https://www.w3schools.com/python/pandas/pandas_dataframes.asp
- https://www.w3schools.com/python/pandas/pandas_csv.asp
- https://www.w3schools.com/python/pandas/pandas_plotting.asp
- https://www.w3schools.com/python/pandas/pandas_ref_dataframe.asp
- https://www.geeksforgeeks.org/how-to-get-column-names-in-pandas-dataframe/
- https://www.geeksforgeeks.org/how-to-get-column-names-in-pandas-dataframe/
- https://www.geeksforgeeks.org/python-pandas-dataframe-series-head-method/?ref=lbp

F)Acknowledgements:-

I would like to thank my prof. Shanmuga for providing guidance.