Report on Data Narratives 2

Pankaj, 22110177, First year btech. Chemical eng, IIT GN

Abstract—**This is report mining of two datasets, first one is aaup.data which contains 1161 US universities data of salaries and number of professors whereas second one usnews.data which contains 1302 US universities data such as about SAT score, fees, student/faculty ratio, graduation . In this report there are 5 scientific questions or hypotheses related to both datasets and answers of those questions with approach to the answers.**

a) Overview of the datasets

First dataset have different columns that contains data of universities such as salary, compensation of full professors, associate professor, assistant professor, instructors and number of these professors whereas the second dataset contains things like average SAT scores, number of applicants applied, excepted, enrolled, their fees, student to faculty ratio, graduation rate etc.

b)Scientific Questions/Hypotheses

First dataset (aaup.data)

First
Scientific Question - Average assistant professors to full professor ratio

Second
Scientific Question - Top 10 US states whose colleges have highest avg. compensation

Third
Scientific Question - What is the probability mass function and cumulative distribution function of average compensation of associate professors? What do you observe in plots ?

Fourth
Scientific Question - Colleges with biggest difference between salaries of full professors and associate professors

Fifth
Scientific Question - What is the probability that a college with number of faculty members greater than 1000 has average salary above 500

Second dataset (usnews.data)

First

Hypotheses- As Instructional expenditure per student increases SAT scores also increase.

Second
Scientific Question - Percentage of public colleges that have more than 80 percent accepted students getting enrolled .

Third
Scientific Question - Compare public colleges and private colleges on the basis of three factors - 1)percentage of faculty with PHDs, 2) Student to faculty ratio and 3) Graduation rate.

Fourth
Scientific question - What is the probability that a private college with more than 50 percent of new students from the top 10 percent of H.S class has a graduation rate above 90.

Fifth
Scientific question - Probability mass functions of SAT scores of public and private college and what are observations .

c) Details of libraries / functions used-

Pandas - Pandas is a library in python that is used to work, manipulate and analyze dataframe and series. It is very good tool to perform different operations on a dataset. Here it is used excessively to answer questions and in analyzing both dataset
.
Matplotlib- It is a library in python which is used to plot different kinds of plots such as bar graph, line graph, pie chart, histograms. Here I used it to better visualize the answers in graphical forms.

NumPy - NumPy is a library in python that helps us to do a wide range of operations on multidimensional arrays and matrices . Here it is used to get unique elements of an array in a question.

np.unique - 'np.unique' is a function provided by numpy library in python, which returns a sorted array of unique elements in a input array

d) Answers to questions/Hypotheses

First dataset (aaup.data)

Answer - 1.353

Approach - First I made an empty list and then I iterated over all the rows of data and found out the ratio of assistant professor to full professors and appended each ratio in the

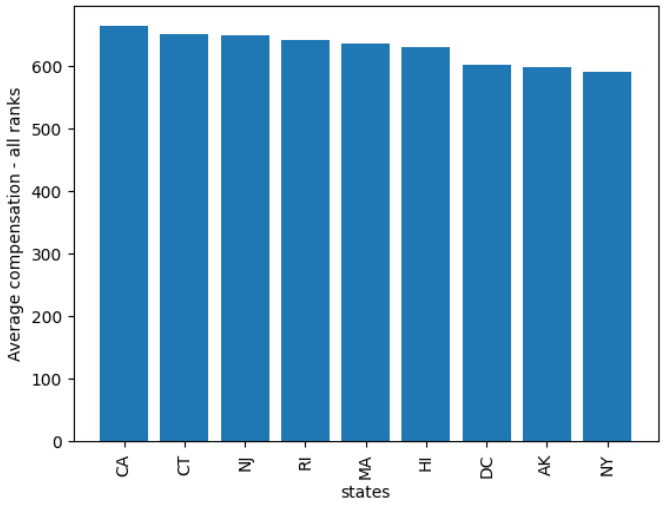
list, then I found the average of elements(ratios) in that list.

Second

Answer - ['CA', 'CT', 'NJ', 'RI', 'MA', 'HI', 'DC', 'AK', 'NY'], are in descending order.

Approach - Firstly I made three empty lists. Then, i iterated over all the rows of data and kept all the unique states in first empty list and corresponding total average compensation of that state in second empty list and correspondingly in third empty list i kept number of universities in that state, then got average compensation list by dividing corresponding elements of second(total compensation) and third lists(number of colleges), then i sorted that list and found top 10 averages and corresponding states.

Plot -

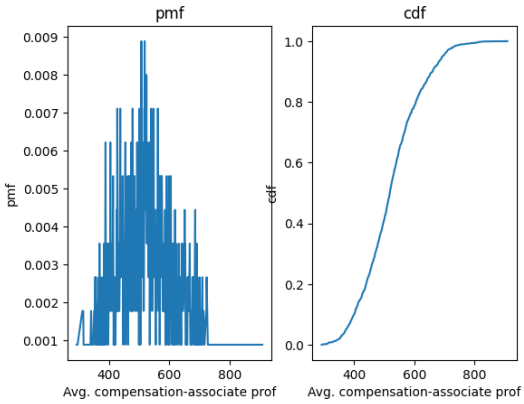


Third

Approach - Firstly i made an array of the average compensations all colleges and then i used numpy to got two arrays, one is all the unique average compensation and other array named "count" is their corresponding count using this -
"unique_nos, count = np.unique(data, return_counts=True)"

and then I found an array of pmf by dividing each element in count by number of total universities using -
"pmf = count / len(data)"
and got cdf by -
"cdf = np.cumsum(pmf)"

Plot -



Fourth

Answer - ['University of Wyoming', 'Wheeling Jesuit College', 'West Virginia Wesleyan College.', 'West Virginia University'] are in descending order

Approach - First I made two arrays one of all colleges and other of their correspond difference of salaries of full professors and associate professors, then I sorted the difference array and corresponding university array using bubble sort and then reversed the sorted arrays using -
"def sort(arr1,arr2):

```
n = len(avg)
for i in range(n):
    for j in range(n-i-1):
        if arr1[j] > arr1[j+1]:
            arr1[j], arr1[j+1] = arr1[j+1], arr1[j]
            arr2[j], arr2[j+1] = arr2[j+1], arr2[j]
```

sort(diff,uni)

diff.reverse()

uni.reverse()"

and then I took the first 5 colleges of this array and got required colleges.

Fifth

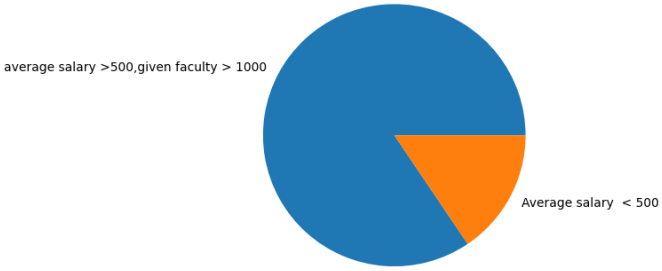
Answer - Probability: 0.84

Approach - Firstly i read and converted dataset into dataframe named df using pandas and then i made another dataframe out of that in which number of faculty in colleges are more than 1000 named df1 after that i made a dataframe from df1 in which average salary in colleges is more than 500 using code -

```
"df1=df[df["Number of faculty - all ranks"]>1000]
df2=df1[df1["Average salary - all ranks"]>500]"
```

then for probability i divided len(df2) by len(df1)

plot -



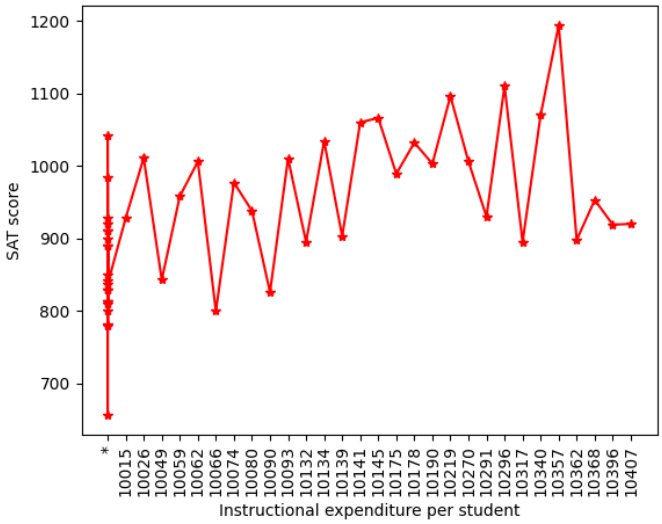
Second dataset (usnews.data)

First

Answer - My hypothesis is wrong .

Approach - First I sorted the dataset according to Instructional expenditure per student then I took the top 50 rows and plotted a plot between SAT scores and Instructional expenditure per student and observed the plot and found out that there is no such direct link between Instructional expenditure per student and SAT score .

Plot -



Second

Answer - 3.829

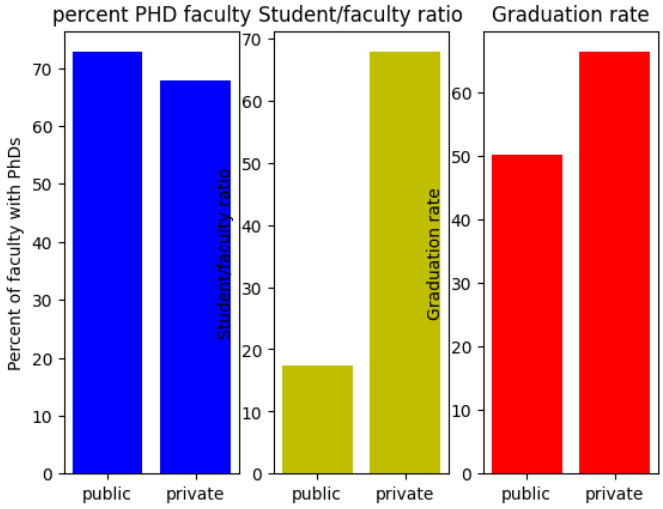
Approach- First i made a dataframe of all public colleges out of whole dataset and then i iterated over this dataframe using a for loop and started a counter for colleges that have more than 80 percent accepted students getting enrolled , then i divided that counter value by number of public colleges and got final required percentage.

Third

answer - percent of faculty with PHDs:
public- 72.74704491725768
private- 67.8315649867374
Student/faculty ratio:
public- 17.45744680851064
private- 13.244562334217505
Graduation rate:
public- 50.03782505910166
private- 66.25994694960212

approach - Firstly I read and converted the dataset into a dataframe named DF and then I made two dataframe DF1 and DF2, out of which DF1 is a data frame which contains all the public colleges whereas DF2 contains all the private colleges. Then I found the average of columns "Pct. of faculty with Ph.D. 's", "Student/faculty ratio" and "Graduation rate" in both DF1 and Df2 . And found out that in terms of PHDs and student/faculty ratio public colleges are better than private colleges but when it comes to graduation rate private colleges are ahead of public colleges

plot -



Fourth

Answer - 0.461

Approach -
Firstly I read and converted the dataset into a dataframe and then I made another dataframe DF out of it which contains only private colleges , then I made DF2 which contains all the rows in which "Pct. new students from top 10% of H.S. class"]>50, then i proceeded by making DF3 out of dF2 which have "Graduation rate"]>90, then for probability i divided len(DF3) by len(DF3) and got my required probability.

Fifth -

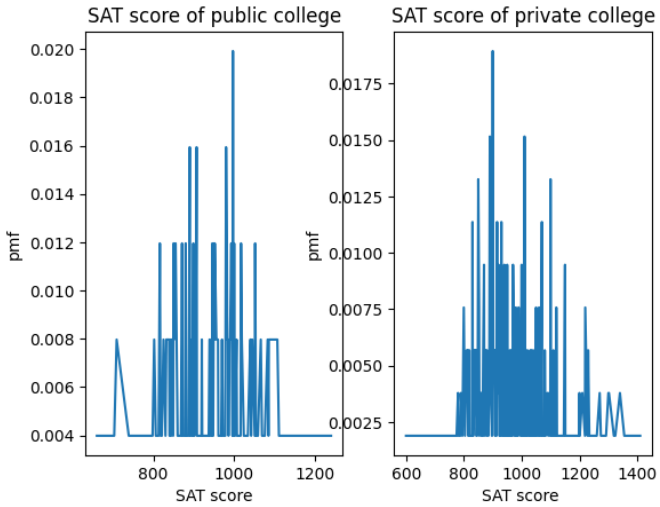
Approach -
Firstly I read and converted the dataset into a dataframe named DF and then I made two dataframe DF1 and DF2, out of which DF1 is a data frame which contains all the public colleges whereas DF2 contains all the private colleges. and

then i made two arrays of SAT scores of both DF1 and DF2 and found their pmf using - “unique1, count1 = np.unique(score1, return_counts=True)
unique2, count2 = np.unique(score2, return_counts=True)

pmf1 = count1/len(score1)

pmf2 = count2/len(score2)”

plot -



observation - SAT scores of public colleges seem more distributed as compared to SAT scores of private colleges which are more towards low to average. Also by a peak we can observe there are more students on the low side as compared to low side of public colleges and similarly there are more students on the high side of private colleges as compared to high side of private colleges

e) Summary of observations-

So, summarizing my observations i would like to say in aaup.data dataset the average assistant professors to full professor ratio is 1.35, top US state with highest average compensation in colleges id “CA”, college with biggest difference in salary of assistant and associate professor is University of Wyoming', and the probability that a college with a number of faculty members greater than 1000 has average salary above 500 is 0.84. And in the usnews.data i found that As Instructional expenditure per student increases SAT scores need not to necessarily increase, Percentage of public colleges that have more than 80 percent accepted students getting enrolled is 3.829, the probability that a private college with more than 50 percent of new students from the top 10 percent of H.S class more than 50 has a graduation rate above 90 is 0.46 and while comparing public and private colleges i find out that percent of faculty with PHDs:
public- 72.74704491725768
private- 67.8315649867374
Student/faculty ratio:
public- 17.45744680851064
private- 13.244562334217505
Graduation rate:
public- 50.03782505910166
private- 66.25994694960212

g) References

“Pandas - Python Data Analysis Library,” n.d. https://pandas.pydata.org/.

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

“NumPy,” n.d. https://numpy.org/.

h) Acknowledgements

I am really thankful to the professor and teaching assistant for this learning opportunity. I learned so many things while making this report and this could not have been done without the help of my teaching assistant and professor. Their guidance was instrumental in helping me to complete this report Once again thank you.

.