DataNarrative 2

Ushasree Thumma
Roll No. :22110272
Computer Science and Engineering

***Abstract*—This report is about analyzing 2 datasets (USNEWS and AAUP) using python programming including libraries like matplotlib, pandas,numpy.**

Keywords—Quartile, In-state,Out-state,terminal degree,Average Compensation,Graduation ratePython, Libraries, Functions, Pandas, Matplotlib,Numpy.

I. OVERVIEW OF THE DATASET

The USNEWS dataset contains information about 1300 American colleges and universities. It has data about

- Type of institution
- Their average math,verbal,act SAT score
- Above mentioned exams score in the first quartile and third quartile
- No. of applications received
- No. of applicants accepted
- No. of new students enrolled
- Pct. of new students from 10% and 25% of H.S.class
- No. of full time and part time undergraduates
- In-state and out-state tuition fees and various expenditures
- Pct. of faculty with Ph.D.'s and terminal degree
- Pct. of alumni who donate
- Graduation rate

The AAUP dataset contains information on

- Average of faculty salaries
- Average compensation of faculty
- No. of faculty (of different position)
- Ranks of average salary, average compensation and no. of faculty

In these datasets, there are some missing values which are denoted with *. All salary and compensation figures are yearly in $100's.

II. SCIENTIFIC QUESTIONS OR HYPOTHESIS

Based on USNEWS dataset:

1. Which state has a higher education focus?
2. What is the probability that a randomly chosen college is private (or public)?
3. Is there any relation between undergraduate enrollment and Graduation rate?
4. What is the probability that a randomly chosen college has a graduation rate greater than 90%?
5. Plot the distribution of student-faculty ratio by college type.

Based on AAUP dataset:

1. Is there a significant difference in the average salaries of faculty members in different academic ranks in Alaska?
2. How does the average salary for full professors vary across different types of institution (i.e. I,IIA,IIB,VIIB)?
3. What are the top 10 colleges who have the maximum number of instructors?
4. How does the no. of faculty members in different academic ranks vary across different types of institutions (i.e. across I,IIA,IIB,VIIB)?
5. PLot the distribution of type I colleges in states of the US using pie chart.

III. DETAILS OF LIBRARIES AND FUNCTIONS

Libraries:

1. *Pandas* [1]*:*

   It is used for working with datasets. It can analyze any large data. We can perform any operations on the dataset. It can also be used for visualizing the dataset, though it's large.

2. *Matplotlib* [2]*:*

   It is used for plotting the data. We can set the parameters and can plot the graph in whichever way we want to by choosing colors,markers,etc for plotting. It is the best way to visualize the data.

3. *Numpy* [3]*:*

   It is used for numerical computing in Python. It provides a wide range of mathematical functions for performing operations on arrays.

Functions:

1. .value_counts():

   It gives the frequency of data in descending order.

2. .drop(index):

   It drops the rows of mentioned index.

3. .head(n):

   It prints the first n rows.

4. .astype():

It converts data from one type to another.

5. numpy.arange(start,stop,end):

   It gives the array from start to stop with step size equal to end. But here step size should be integer.

6. .groupby():

   "A groupby operation involves some combination of splitting the object, applying a function, and combining the results[1]."

7. .mean():

   It calculates the mean.

8. .sort_values(ascending=bool):

   If ascending=True, it sorts in ascending order and if ascending=False, it sorts in descending order.

IV.ANSWERS TO THE QUESTIONS (WITH APPROPRIATE ILLUSTRATIONS)

Based on USNEWS dataset:

1.

Input:

Here is the code:

```
plt.figure(figsize=(16,6))
usnews["State (postal code)"].value_counts().plot(kind="bar",color="m")
plt.title("No. of colleges in state",size=16,color="b")
plt.xlabel("States",size=16,color="#FF4500")
plt.ylabel("No.of colleges",size=16,color="#FF4500")
plt.show()
```

Approach:

We say a state has a higher education focus when it makes education available to more people. In other words, it promotes more no. of colleges. So in this code, I am counting the number of colleges in each state and plotted the bar graph for it.

Output:

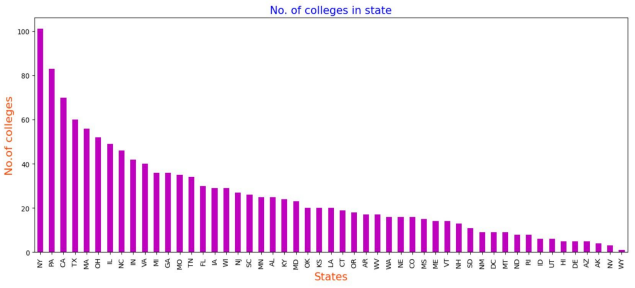

Fig. 1. No. of colleges in each state of US

Conclusion:

It is clear that the state with postal code NY(i.e. NewYork[4]) has more no. of colleges. Therefore, it has the highest education focus among other states of the US..

2.

Input:

Here is the code:

```
private=usnews[usnews["Public/private indicator (public=1, private=2)"]==2]
prob_private=len(private)/len(usnews)

# bar plot of probability that chosen college is public or private
plt.figure(figsize=(2,2))
x=plt.bar(["private","public"],[prob_private,1-prob_private],color="#FFFF00")
plt.bar_label(x,labels=[round(prob_private,2),round(1-prob_private,2)],label_type="center",color="r")
plt.title("Probability of chosen college being public or private",size=9,color="b")
plt.xlabel("institution",size=9,color="g")
plt.ylabel("probability",size=9,color="green")
plt.show()

print("Probability that chosen college is private :",round(prob_private,2))
print("Probability that chosen college is public :",round(1-prob_private,2))
```

Approach:

I want to know what is the probability for the selected college to be private. For that I have calculated the no. of colleges which are private. Then, I have calculated the probability by dividing no. of private colleges with total no. of colleges. Then the probability of selected college to be public will be 1- probability of private college (as sum of all probabilities is equal to 1).

I've written code to plot both the probabilities in the graph.

Output:

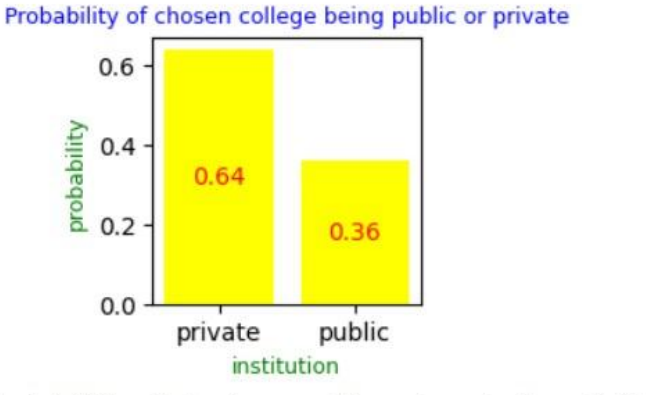

Fig. 2. Probability that college to be private(or public)

Answer:

Probability that chosen college is private=0.64

Probability that chosen college is public=0.36

3.

Input:

```
usnews=usnews.drop(usnews[usnews["Graduation rate"]=="*"].index)
y1=usnews["Graduation rate"].head(5)
y2=usnews["Number of fulltime undergraduates"].head(5)
y3=usnews["Number of parttime undergraduates"].head(5)
x=usnews["College name"].head(5)
y2=y2.astype({'Number of fulltime undergraduates':'int'})
y3=y3.astype({'Number of parttime undergraduates':'int'})
y1=y1.astype({'Graduation rate':'int'})
x_ax=np.arange(len(x))
plt.figure(figsize=(20,15))
plt.subplot(2,1,1)
plt.bar(x_ax,y3,0.2,label="No. of parttime undergraduates",color="#00BFFF")
plt.bar(x_ax+0.2,y2,0.2,label="No. of fulltime undergraduates",color="#1E90FF")
plt.title("No. of parttime and fulltime undergraduates in 5 colleges",size=16,color="#000080")
plt.xlabel("College",size=14,color="r")
plt.ylabel("No. of undergraduates",size=14,color="r")
plt.xticks(x_ax+0.1,["Alaska Pacific University","University of Alaska Southeast","Alabama Agri. & Mech. Univ.","Faulkner University","University of Montevallo"])
plt.legend()

plt.subplot(2,1,2)
plt.bar(x,y1,color="#808080")
plt.title("Graduation rate",size=16,color="#000080")
plt.xlabel("College",size=14,color="g")
plt.ylabel("Graduation rate",size=14,color="g")
plt.show()
```

Approach:

I want to know the relationship between undergraduate enrollment and graduation rate,if any. For the sample,I've chosen 5 colleges and plotted the no. of part-time,full-time undergraduates in one figure and graduation rates of those colleges in another figure.

Output:

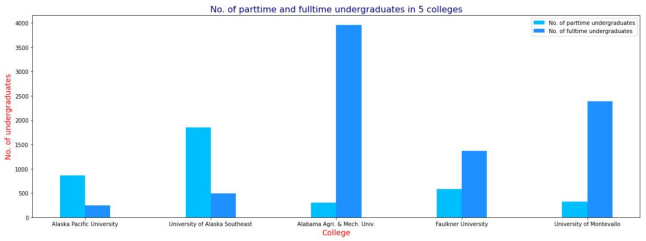

Fig. 3. Undergraduate enrollment in 5 colleges

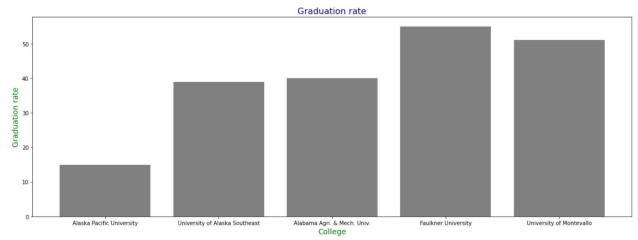

Fig. 4. Graduation rate in those 5 colleges

Answer:

The colleges with the highest number of full-time undergraduates (University of Montevallo and University of Alaska Southeast) have relatively low graduation rates compared to the other colleges in the top five. This suggests that having a high number of full-time undergraduates does not necessarily guarantee higher graduation rates and that other factors may also play a role in determining graduation rates.

4.

Input:

```
usnews=usnews.drop(usnews[usnews["Graduation rate"]=="*"].index)
usnews=usnews.astype({'Graduation rate':'int'})
grad_gt_90=usnews[usnews["Graduation rate"]>90]
prob_grad_gt_90=len(grad_gt_90)/len(usnews)

print("Probability that randomly chosen college has graduation rate greater than 90% is",round(prob_grad_gt_90,2))
```

Approach:

To calculate this probability, I've calculated the no. of colleges which have a graduation rate greater than 90%. Then I have divided this no. with the total number of colleges.
Thereby, we get the probability.

Answer:

Probability that the randomly chosen college has graduation rate greater than 90% is 0.07

5.

Input:

```
private_sf = usnews[usnews['Public/private indicator (public=1, private=2)'] == 2]['Student/faculty ratio']
public_sf = usnews[usnews['Public/private indicator (public=1, private=2)'] == 1]['Student/faculty ratio']

plt.figure(figsize=(10,6))
plt.hist(private_sf,label="Private",color="#FFFF00")
plt.hist(public_sf,label="Public",color="#FF4500")
plt.title("Distribution of Student-Faculty Ratio by College Type",color="m",size=14)
plt.xlabel("Student/Faculty ratio",color="b",size=12)
plt.ylabel("Frequency",color="b",size=12)
plt.ylim(0,550)
plt.xlim(0,80)
plt.legend()
plt.show()
```

Approach:

I want to know the student/faculty ratio in private and public colleges. For this, I'm storing the student/faculty ratio of private colleges in variable private_sf and the other in the variable public_sf.
I am plotting a histogram of both using matplotlib.

Output:

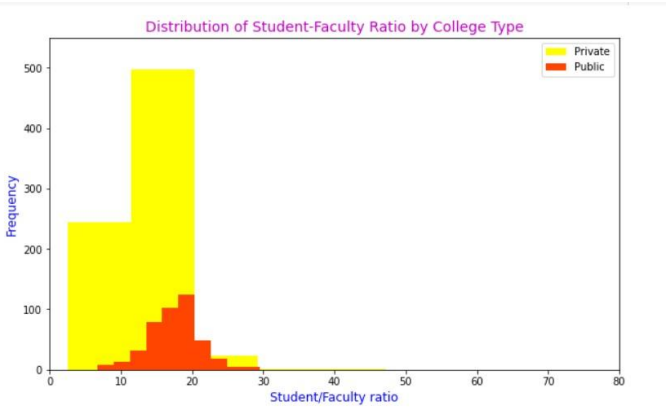

Fig. 5. Distribution of Student/Faculty ratio by college type

Answer:

From the above graph, we can see that the distribution of student/faculty ratios for private colleges and universities is generally more concentrated around lower ratios than the distribution for public colleges and universities.This suggests that private colleges tend to have smaller class sizes and a higher level of individual attention of students.
However, it's important to note that the distributions do overlap quite a bit, and there is significant variation within each group.This suggests that there is no clear-cut difference between public and private colleges and universities in terms of student-faculty ratio, and other factors may also play a role.

Based on AAUP dataset:

1.

Input:

Here is the code:

```
alaska=aaup[aaup["State (postal code)"]=="AK"]
y1=alaska["Average salary - full professors"]
y2=alaska["Average salary - associate professors"]
y3=alaska["Average salary - assistant professors"]
x=alaska["College name"]
y2=y2.astype({'Average salary - associate professors':'int'})
y3=y3.astype({'Average salary - assistant professors':'int'})
y1=y1.astype({'Average salary - full professors':'int'})
x_ax=np.arange(len(x))
plt.figure(figsize=(20,8))
plt.bar(x_ax,y3,0.2,label="Assistant professor",color="#FFFF00")
plt.bar(x_ax+0.2,y2,0.2,label="Associate professor",color="#FFA500")
plt.bar(x_ax+0.4,y1,0.2,label="Full professor",color="#FF4500")
plt.title("Average salary of faculty in Alaska",color="b")
plt.xlabel("College",size=14,color="r")
plt.ylabel("Average salary (in $100's)",size=14,color="r")
plt.xticks(x_ax+0.2,["Alaska Pacific University","Univ.Alaska-Fairbanks","Univ.Alaska-Southeast","Univ.Alaska-Anchorage"])
plt.ylim(bottom=0)
plt.legend()
plt.show()
```

Approach:

The state with postal code "AK" is Alaska[4].
No. of colleges in Alaska is 4. I've stored the average salary of faculty at different academic ranks in each college in variables y1,y2,y3 and plotted this using multiple barchart.

Output:

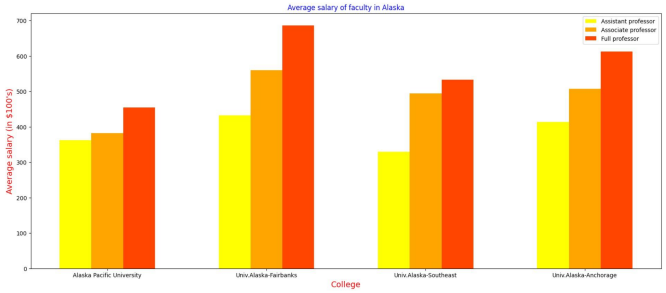

Fig. 6. Average salary of faculty in Alaska

2.

Input:

Here is the code:

```
aaup=aaup.drop(aaup[aaup["Average salary - full professors"]=="*"].index)
aaup=aaup.astype({"Average salary - full professors":"int"})
mean_salary=aaup.groupby("Type (I, IIA, or IIB)")["Average salary - full professors"].mean()
mean_salary.plot(kind="bar",color="#DEB887")
plt.title("Mean salary for full professors by institution type",color="r")
plt.xlabel("Institution type",color="orange",size=12)
plt.ylabel("Mean salary(in $100's)",color="orange",size=12)
plt.show()
```

Approach:

I want to know the salary of full professors in different types of institutions. For that I've used .groupby() function which groups a column according to different values of another column. So,I've grouped the average salary of full professors according to type of institution and calculated mean for them as mean uses all the values in the dataset to give us an average.

Then I've plotted that mean for different types of institutes using matplotlib.

Output:

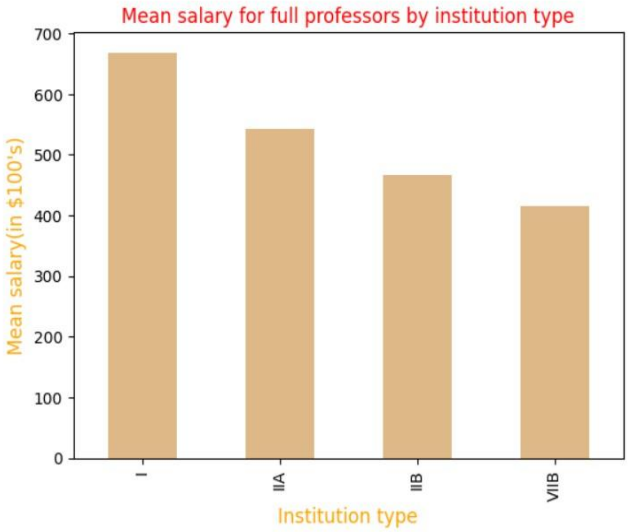

Fig. 7. Average salary for full professors by institution type

3.

Input:

```
x=pd.read_excel("/content/aaup.xlsx",usecols=["College name","Number of instructors"])
x=x.sort_values(by="Number of instructors",ascending=False)
print(x.head(10))
```

Approach:

To know the top 10 colleges who are having maximum no. of instructors, sort the dataset based on no. of instructors in descending order.The top listed 10 colleges will be the answer.

Answer:

TABLE I

College name	No. of instructors
Louisiana ST.Univ. and A&M C.	178
Penn State Univ. -Main Campus	160
Ball State University	155
Univ. of Maryland-Coll. Park	133
Idhao State University	120
Northern Illinois University	112
Howard University	111
Georgia Southern University	109
Univ. of Texas -Pan Amer.Univ	109
Univ. of Hawaii at Manoa	100

4.

Input:

```
aaup=aaup.astype({"Number of full professors":"int"})
aaup=aaup.astype({"Number of associate professors":"int"})
aaup=aaup.astype({"Number of assistant professors":"int"})
mean_fullp=aaup.groupby("Type (I, IIA, or IIB)")["Number of full professors"].mean()
mean_assocp=aaup.groupby("Type (I, IIA, or IIB)")["Number of associate professors"].mean()
mean_assisp=aaup.groupby("Type (I, IIA, or IIB)")["Number of assistant professors"].mean()

x_ax=np.arange(4)
plt.figure(figsize=(20,8))
plt.bar(x_ax,mean_assisp,0.2,label="Assistant professor",color="#FFFF00")
plt.bar(x_ax+0.2,mean_assocp,0.2,label="Associate professor",color="#FFA500")
plt.bar(x_ax+0.4,mean_fullp,0.2,label="Full professor",color="#FF4500")
plt.title("Mean no. of professors by institution type",size=16,color="#000080")
plt.xlabel("Type",size=16,color="r")
plt.ylabel("Number of professors",size=16,color="r")
plt.xticks(x_ax+0.2,["I","IIA","IIB","VIIB"])
plt.legend()
plt.show()
```

Approach:

I want to know the no. of faculty members in different types of institutions. For that I've used .groupby() function which groups a column according to different values of another column. So,I've grouped the no. of faculty of different ranks according to type of institution and calculated mean for them.

Then I've plotted that mean for different types of institutes using matplotlib.

Output:

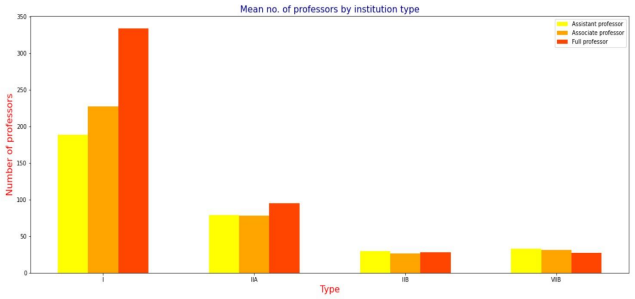

Fig. 8. Average no. of faculty by institution type

5.

Input:

```
z=pd.read_excel("/content/aaup.xlsx",usecols=["State (postal code)","Type (I, IIA, or IIB)"])
z=z[z["Type (I, IIA, or IIB)"]=="I"]
z=z["State (postal code)"].value_counts()
x=["NY","TX","CA","IL","OH","MA","PA","MO","VA","DC","IN","TN","NC","MI","CO","GA","FL","AL",
   "OR","AZ","NM","WI","MS","UT","LA","MD","NJ","KY","OK","SC","RI","IA","WA","KS","MT","NH",
   "SD","WV","VT","AK","NV","NE","ND","MN","ME","ID","HI","DE","CT","AR","WY"]

#Plotting pie graph
plt.figure(figsize=(25,15))
plt.pie(z,labels=x)
plt.show()
```

Approach:

As I need only Type I institutions , I've considered the rows which are under type I. Then I've calculated how many type I colleges are there in each state using .value_counts() and depicted this in pie chart.

Output:

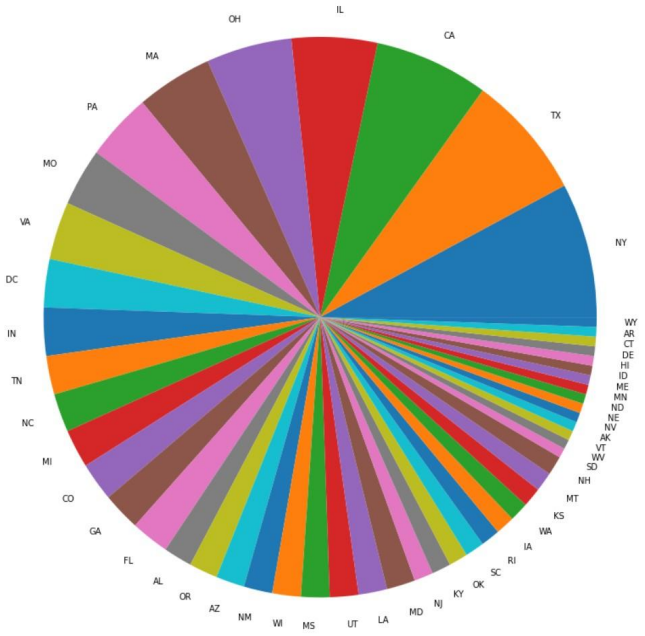

Fig. 9. No. of Type I colleges in each state

V.SUMMARY OF THE OBSERVATIONS

Based on USNEWS dataset:

1. Newyork has higher education focus.
2. In US, there are more private colleges and universities compared to those of public.
3. Having high number of full-time undergraduates does not necessarily guarantee higher graduation rates.This suggests that factors other than undergraduate enrollment, such as academic rigor, student support services, and institutional culture, may also contribute to higher graduation rates.
4. There are very less no. of colleges who are having graduation rate greater than 90%.
5. Most of the private colleges tend to have smaller class sizes and a higher level of individual attention of students. But whereas in public colleges, it is implemented in less no. of colleges.

Based on AAUP dataset:

1. In general,
   Academic ranks order: Full-time>Associate>Assistant
   Their salary order is also: Full-time>Associate>Assistant
2. Average salary for full professors:
   I > IIA > IIB > VIIB
3. Louisiana ST.Univ. and A&M College has maximum no. of instructors.
4. Average no. of faculty of rank
   full professor: I > IIA > IIB > VIIB
   Associate professor: I > IIA > VIIB > IIB
   Assistant professor: I > IIA > IIB > VIIB
   No. of professors are more in type I colleges.
5. Newyork has more no. of type I colleges.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my ES114 instructor, Dr.R.Shanmuga for giving me the opportunity to analyze 2 datasets. I am extremely grateful for the knowledge he offered me.

References

[1] McKinney, Wes. "pandas: a foundational Python library for data analysis and statistics." *Python for high performance and scientific computing* 14, no. 9 (2011): 1-9.

[2] Hunt, John, and John Hunt. "Introduction to Matplotlib." *Advanced Guide to Python 3 Programming* (2019): 35-42.

[3] Idris, Ivan. *NumPy: Beginner's Guide*. Packt Publishing Ltd, 2015.

[4] "Appendix D - USPS State Abbreviations and FIPS Codes." U.S. Bureau of Labor Statistics. U.S. Bureau of Labor Statistics, September 27, 2005. https://www.bls.gov/respondents/mwr/electronic-data-interchange/appendix-d-usps-state-abbreviations-and-fips-codes.htm.