(Data Narrative-Assignment 2

Probability, Statistics, Data Visualisation)

- Kaushik
- IIT Gandhinagar
- Chemical Engineering
- Kaushik.kaushik@iitgn.ac.in

Abstract—This document comprises different results from the dataset based on us universities, where they are located condition of their of their staff and professors and the whole data being compressed to some graph which makes reading the data easy. Bar graph, scatter graph etc. has been used to present data in

OVERVIEW OF THE DATASET

In this assignment we have been two dataset on American universities which included type of universities, where they are located, at what sat score students took admission , how many got rejected. What is the condition of their staff, how is future prospects of working their or studying. We have simplifies the data in form of various graphs such as bar graph, scatter along with its marking of titles and subtitles to help understand, know the data better and take decision.

QUESTION ON DATASET

Question 1) What is the Probability of Type I has more salary for associate professor then for full professors. Plot the graph for distribution of full professors vs associate professor.

Question 2) What is the probability that type I has more then 50 percent of professors as full time professor. Plot the distribution graph for the same.

Question 3) What is the probability has that random selected college in this dataset is a type II institute in the state of California (CA) and has an average salary for all ranks off over $500,000. Plot the graph for average salary

Question 4)What is the probability that a college chosen at random has an average salary for assistant professors is higher than the average salary for associate professor, given that college has the least 50 faculty members .

Question 5) What is the probability that college chosen random in California has more associate professor then assistant professor. Plot the graph for average salary at all ranks.

Question 6) What is the probability that chosen at random is private or public?

Question 7) What is the probability students have SAT scores for than 1000 and they are from 25 percent of their high school class?

Question 8) What is the relationship between number of students applied for college and number of students applied?

Question 9) What the Ten college with highest acceptance rate and have highest number of new enrollment of students?

Question 10) What is the probability that the college in California has average SAT score of minimum 700? Also plot the graph of top 10 colleges in America according to SAT scores.

DETAILS OF LIBRARIES USED IN PROGRAM

The most comman libraries used in our representation of data is pandas, Numpy, matplotlib there are thousands of functions that we can use and there are hundreds of libraries which make our work easy and some of them are used in this assignment. These librariers helped us to visualize data very clearly and we also discuss in details about uses of these data. Some of the most used are mentioned below:

A. Pandas (Most used library in this assignment)
   - "Pandas[1] is a Python package used for working with data sets," is how you define it. It includes tools for cleaning, examining, altering, and analysing data. Why is it employed? It is used to reduce large amounts of data into a statistical form that is simple to understand. functions employed by the "Panda" library.
   - ‘pd.read_csv’ is used to read the file from the given link and saves it in a variable which we declare and help us to print.
   - “df.head()”used to define till what values we have to take the data.
   - “df.tail()”used to define data from starting from end to maximum value of tail.
   - “values-count()” is used to count how many time a element is present there it returns us a value in series.
   - “mean()” is used to find the mean Equations
   - “id.max()” is used to find the max from the present .

□ □ □

B. matplotlib.pyplot[2](Second most used library in this assignment)
C. Seaborn (third most used library in the assignment)

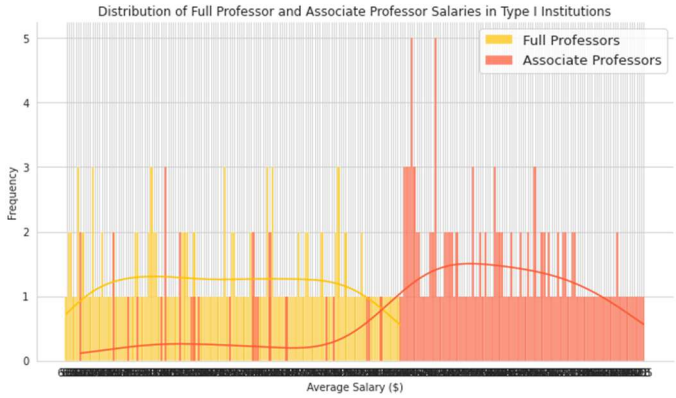
Functions used :

- “plot()” is used to plot the graph between x and y .
- The “ show()” is used to show the plotted graph.
- “title()” is used to print the title at the bottom of the graph.
- “x-label()”is used to print the x label on the graph .
- “y-label()” is used to print the label at y axis of the graph.

Answers

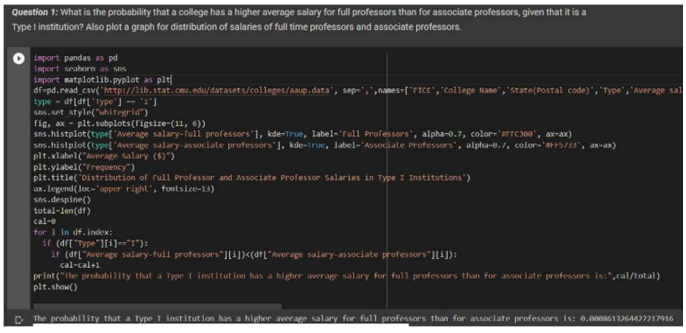
1) Probability of Type I has more salary for associate professor then for full professors is **0.000861326442721916**.

Chances of your salary being a associate professor be higher than that of full professor is quite less. Specially if the college if type I is chances get further reduced. Data of salaries of full professors and associate professor is clearly visualized here which will help you to know distributions of salaries of professor in various situations.



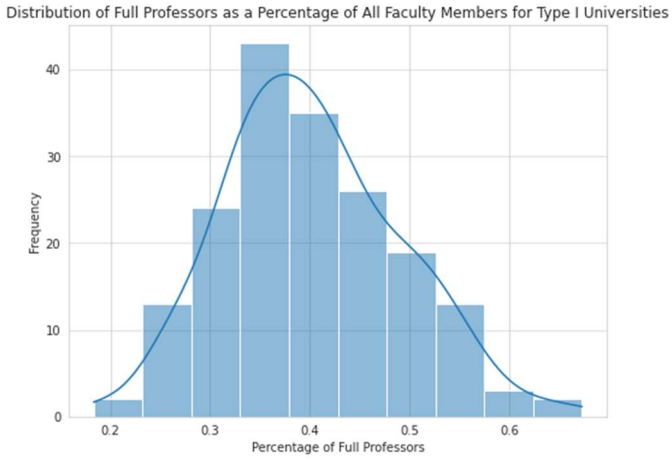
The graph tell more distribution in favour of full professors compared to associate professors.

The code for the following:

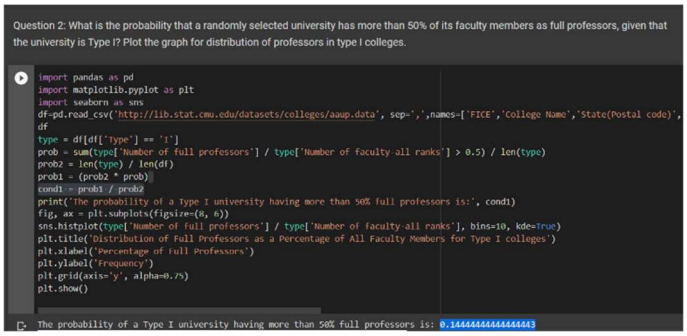


2) The probability that the Type I college has more then 50 percent of professors as full time professors is 0.144444444443.

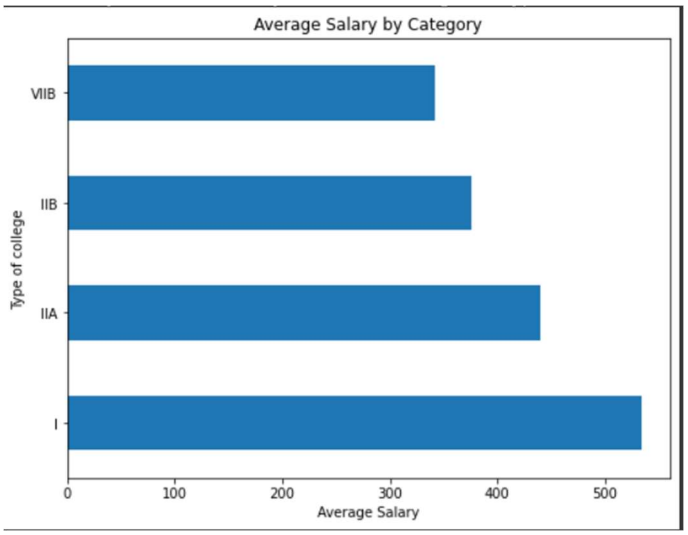
Looking at the graph frequency of full-time professor is quite less. It has height in the middle that’s it. The graph clearly indicates that does keep all the full time but just like corporate in different positions.



The code for the following graph:



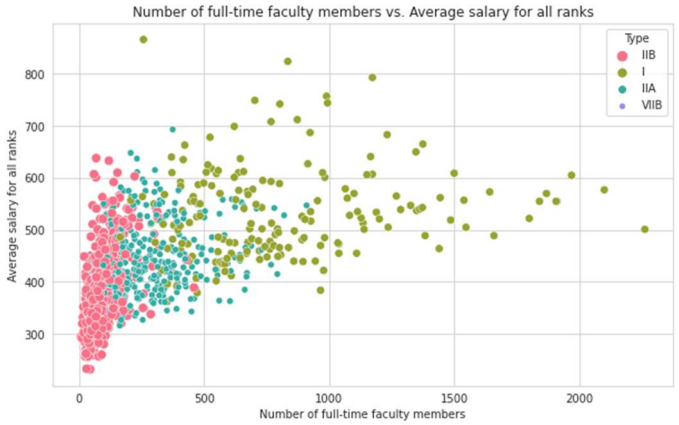
3) The probability has that random selected college in this dataset is a type II institute in the state of California (CA) and has an average salary for all ranks off over $500,000 is **0.020671834625323**



The graph shows that average salary is quite high in Type I institute compared to other institute.

There are two type of institute in Type II(A,B) from the graph it can be make clearly IIA institute has more average salary.

4) The probability that a college chosen at random has an average salary for assistant professors is higher than the average salary for associate professor, given that college has the least 50 faculty members is **0.00516795865633075**.



The code for the above graph is below:

Question 4: What is the probability that a college chosen at random has an average salary for assistant professors is higher than the average salary for associate professors, given that the college has at least 50 faculty members? Plot the graph for faculty members and their salaries.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
df=pd.read_csv('http://lib.stat.cmu.edu/datasets/colleges/aaup.data', sep=',',names=['FICE','College Name','State(Postal code)','Type','A
df=pd.DataFrame(df)
count=0
for i in df.index:
  if df["Number of faculty all ranks"][i]>=50:
    if ((df['Average salary-assistant professors'][i])>(df['Average salary-associate professors'][i])):
        count=count+1
p=count/len(df)
print("Probability that salary for assistant professors high than salary for associate professors, given that the college has at least 50
palette = sns.color_palette("husl", len(df['Type'].unique()))
plt.figure(figsize=(10,6))
sns.scatterplot(data=df, x="Number of faculty all ranks", y="Average salary all ranks", hue="Type", palette=palette, size="Type")
plt.title("Number of faculty members all ranks vs. Average salary for all ranks")
plt.xlabel("Number of faculty members all ranks")
plt.ylabel("Average salary for all ranks")
plt.show()
```

5) The probability that college chosen random in California has more associate professor then assistant professor is **0.02756244617097326**.

No. of colleges in California is 54 according to this dataset and probability of choosing college in California is **0.046511627906976744**.

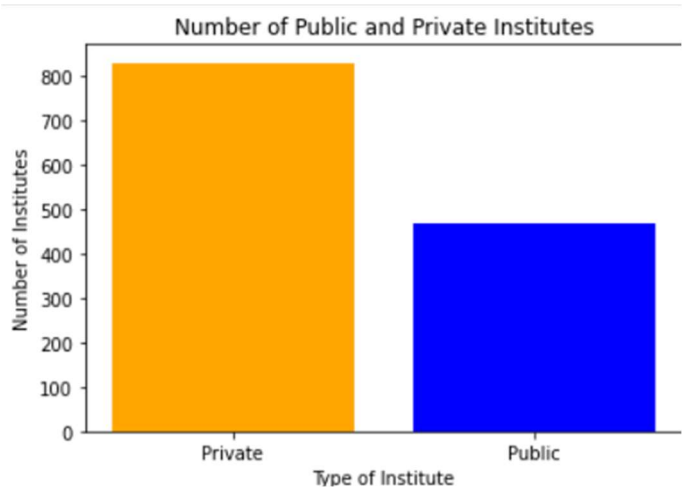
Question 5: What is the probability if college choosen at random in california has more associate professor than assistance professor?

```
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_csv('http://lib.stat.cmu.edu/datasets/colleges/aaup.data', sep=',',names=['FICE','College Name','State(Postal code)',
df=pd.DataFrame(df)
c=(df["State(Postal code)"]=="CA").sum()
total=len(df)
print("Total collegs in California:",c)
ca=c/total
print("Probability of a college choosen at random is in California:",ca)
count=0
for i in df.index:
  if (df["State(Postal code)"][i]=="CA") and (df["Number of associate professors"][i]>df["Number of assistant professors"][i]):
      count=count+1
t=count/c
a=ca*t
print("Probaility of college choosen in california has more asssociate professors then assistant professors",a)
```

```
Total collegs in California: 54
Probability of a college choosen at random is in California: 0.046511627906976744
Probaility of college choosen in california has more asssociate professors then assistant professors 0.027562446167097326
```

6) Probability of choosing college at random is private is **0.64.** The probability of choosing college at random is public is **0.36.**

America is capitalist country so it very much obvious that that private college is public colleges. But do you most ivy league are college are private owned for example Harvard, Stand ford university.



The code for the above graph below:

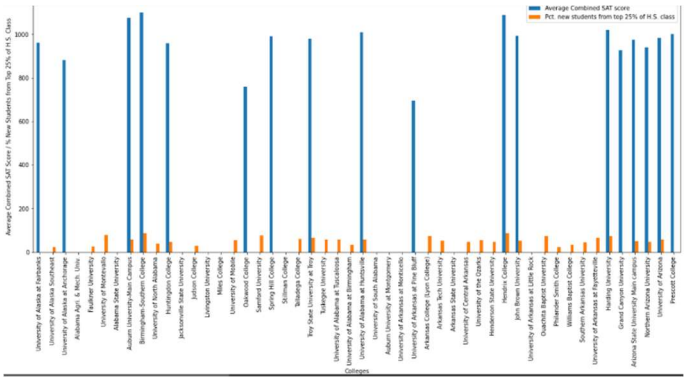
Question 6: What is the probability college chosen at random is private insitute or public institute?

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('http://lib.stat.cmu.edu/datasets/colleges/usnews.data')
new_df = df.iloc[:, :20]
new_df.columns = ['FICE (Federal ID number)','College Name','State(Postal code)','Pub
counts = new_df['Public/private indicator (public=1, private=2)'].value_counts()
print("Probability of a college chosen at random being a private institute: 0.64")
print("Probability of a college chosen at random being a public institute: 0.36")
labels = ['Private', 'Public']
colors = ['orange', 'blue']
plt.bar(labels, counts, color=colors)
plt.title('Number of Public and Private Institutes')
plt.xlabel('Type of Institute')
plt.ylabel('Number of Institutes')
plt.show()
```

```
Probability of a college chosen at random being a private institute: 0.64
Probability of a college chosen at random being a public institute: 0.36
```

7) The probability students in have SAT scores for than 1000 and they are from 25 percent of their class is **0.003843197540353574**.

Relation between SAT scores and percentage from top 25% of class.



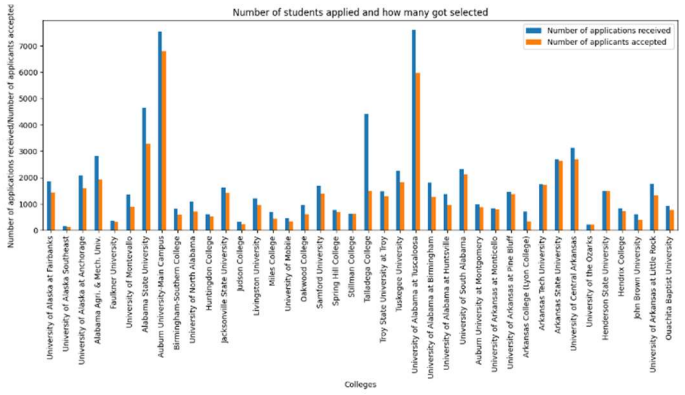
The code for the above graph:

Question 7: What is the relationship between SAT scores and percentage of new students from the top 25 percent of H.S class? What the probability of college has students have SAT scores atleast 1000 and they are in 25 percent from high school?

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('http://lib.stat.cmu.edu/datasets/colleges/usnews.data')
df = df.iloc[:, :20]
df.columns = ['FICE (Federal ID number)','College Name','State(Postal code)','Public/private indicator (public-1, private-2)','Average Ma
df=pd.DataFrame(df)
a=len(df.index)
df=df.head(50)
df = df.replace('*', float('nan'))
df['Average Combined SAT score'] = df['Average Combined SAT score'].astype(float)
df['Pct. new students from top 25% of H.S. class'] = df['Pct. new students from top 25% of H.S. class'].astype(float)
df.plot(kind='bar', x='College Name', y=['Average Combined SAT score', 'Pct. new students from top 25% of H.S. class'], figsize=(20,8))
plt.title('SAT Scores and Percentage of New Students from Top 25% of H.S. Class by College')
plt.xlabel('Colleges')
plt.ylabel('Average Combined SAT Score / % New Students from Top 25% of H.S. Class')
df.dropna(subset=['Number of new students enrolled'], inplace=True)
c=0
for i in df.index:
  if df["Average Combined SAT score"][i]>1000:
    if df["Pct. new students from top 25% of H.S. class"][i]>25:
      c=c+1
print("he probability of college has students have SAT scores atleast 1000 and they are in 25 percent from high school is ",c/a)
plt.show()
```

```
he probability of college has students have SAT scores atleast 1000 and they are in 25 percent from high school is  0.003843197540353574
```

8) According to the graph, Talladega college rejected so many students followed by university of Alabama. But, Henderson University accepted almost the students who applied there followed by University of Central Arkansas.



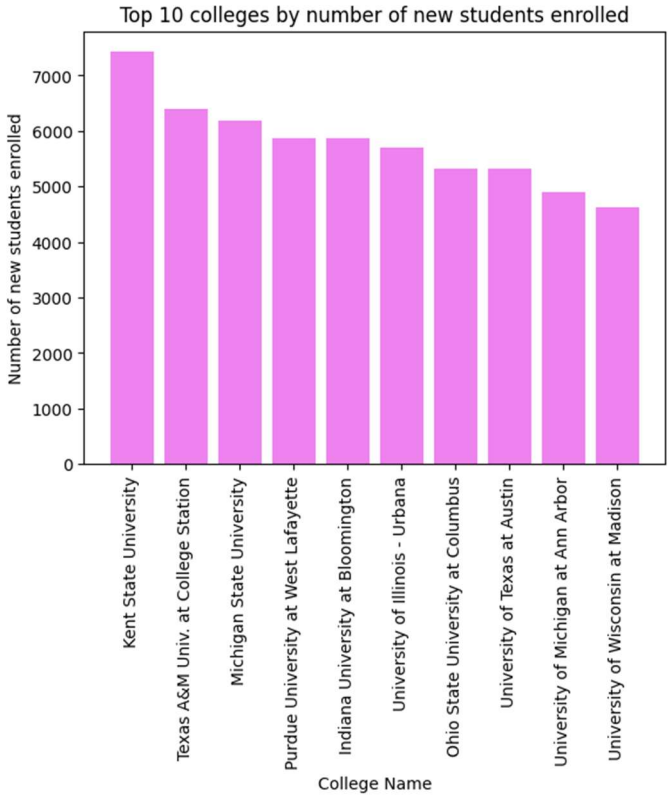
The code for below graph:

Question 8: What is the relation between students applied and got accepted?

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('http://lib.stat.cmu.edu/datasets/colleges/usnews.data')
df = df.iloc[:, :20]
df.columns = ['FICE (Federal ID number)','College Name','State(Postal code)','Public/private indicator (public-1, private-2)','
df = df.head(40)
df = df.replace('*', float('nan'))
df['Number of applications received'] = df['Number of applications received'].astype(float)
df['Number of applicants accepted'] = df['Number of applicants accepted'].astype(float)
df.plot(kind='bar', x='College Name', y=['Number of applications received', 'Number of applicants accepted'], figsize=(15,5))
plt.title('Number of students applied and how many got selected')
plt.xlabel('Colleges')
plt.ylabel('Number of applications received/Number of applicants accepted')
plt.show()
```

9) Ten college with highest acceptance rate and have highest number of new enrollment of students are as follows:

Kent state university has highest acceptance rate as it has enrolled in 7425 students in a year.



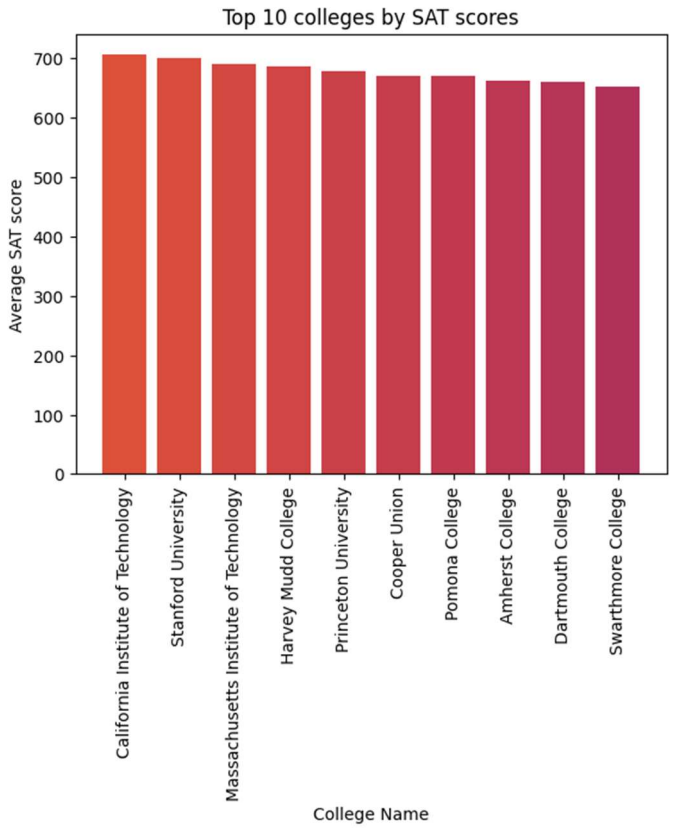
Question 9: Which colleges have highest acceptance rate?

```
import matplotlib.pyplot as plt
import pandas as pd
df = pd.read_csv('http://lib.stat.cmu.edu/datasets/colleges/usnews.data')
df = df.iloc[:, :20]
df.columns = ['FICE (Federal ID number)','College Name','State(Postal code)','Public/private indicator (public=1,
df['Number of new students enrolled'] = pd.to_numeric(df['Number of new students enrolled'], errors='coerce')
df.dropna(subset=['Number of new students enrolled'], inplace=True)
top = df.nlargest(10, 'Number of new students enrolled').reset_index(drop=True)
print(top[['College Name', 'Number of new students enrolled']])
plt.bar(top['College Name'], top['Number of new students enrolled'],color="violet")
plt.xticks(rotation=90)
plt.xlabel('College Name')
plt.ylabel('Number of new students enrolled')
plt.title('Top 10 colleges by number of new students enrolled')
plt.show()
```

```
                          College Name  Number of new students enrolled
0                Kent State University                           7425.0
1  Texas A&M Univ. at College Station                           6392.0
2           Michigan State University                           6180.0
3  Purdue University at West Lafayette                          5874.0
4    Indiana University at Bloomington                          5873.0
5      University of Illinois - Urbana                          5705.0
6    Ohio State University at Columbus                          5330.0
7        University of Texas at Austin                          5329.0
8  University of Michigan at Ann Arbor                          4893.0
9   University of Wisconsin at Madison                          4631.0
```

10) The probability that the college in California has average SAT score of minimum 700 is **0.0015727901616141429**7.

Top 10 hardest colleges in America to get into according to their acceptance SAT scores are as follows:



The code for the graph:

```
Question 10: What are the top 10 hardest colleges in America to get into? What is probability that a college in California has SAT score more than 700?

color = cm.inferno_r(np.linspace(.4, .8, 30))
color
df = pd.read_csv('http://lib.stat.cmu.edu/datasets/colleges/usnews.data')
df = df.iloc[:, :20]
df.columns = ['FICE (Federal ID number)','College Name','State(Postal code)','Public/private indicator (public-1, private-2)','Aver
df = df[(df['Average Combined SAT score'] != '*') &
        (df['Average Math SAT score'] != '*') &
        (df['Average Verbal SAT score'] != '*')]
df['Average Combined SAT score'] = pd.to_numeric(df['Average Combined SAT score'])
df['Average Math SAT score'] = pd.to_numeric(df['Average Math SAT score'])
df['Average Verbal SAT score'] = pd.to_numeric(df['Average Verbal SAT score'])
df['Average SAT score'] = (df['Average Math SAT score'] + df['Average Verbal SAT score']) / 2
topcollege = df.sort_values('Average SAT score', ascending=False).head(10).reset_index(drop=True)
print(topcollege[['College Name', 'Average SAT score',"State(Postal code)"]])
plt.bar(topcollege['College Name'], topcollege['Average SAT score'], color=color)
plt.xticks(rotation=90)
plt.xlabel('College Name')
plt.ylabel('Average SAT score')
plt.title('Top 10 colleges by SAT scores')
c=0
for i in df.index:
  if df['State(Postal code)'][i]=="CA":
    if df["Average SAT score"][i]>700.0:
      c=c+1
print("The probaility that a college in California has SAT score more than 700",c/1301)
plt.show()
```

Summary

The dataset is about the salaries and average salaries about on Full professor, assistant and associate professors and about no. of teaching staff. It will you learn about working of various level of teaching department, with many levels just like a corporate ladder. Next it is about average SAT scores of students and how many applied and how many got selected. Basically everything about one needs to know for admission.

Acknowledgement

I would like to thank to Professor Shanmuga R sir as due to him we learnt so much about python and different libraries like pandas, numpy, scipy, and matlpotlib I am very grateful to you to assign us such a good and useful assignment.

Refrences

[1] "Pandas Tutorial." n.d. Www.w3schools.com. https://www.w3schools.com/python/pandas/default.asp.

[2] "Getting Started Tutorials — Pandas 1.5.3 Documentation." n.d. Pandas.pydata.org. Accessed March 30, 2023. https://pandas.pydata.org/docs/getting_started/intro_tutorials/index.html.

[3] "Pandas Tutorial." 2020. GeeksforGeeks. February 28, 2020. https://www.geeksforgeeks.org/pandas-tutorial/.