Data Narrative: Students and Professors in American Colleges

Sriram Srinivasan
Electrical Engineering
Indian Institute of Technology, Gandhinagar
Gandhinagar, India
sriram.srinivasan@iitgn.ac.in, Roll No. 22110258

Abstract— **The following report aims to raise significant scientific questions based on two datasets about American colleges. The first dataset contains information on faculty salaries for 1161 American colleges and universities. The second dataset contains information related to SAT scores, per-student-expenditure, student-faculty ratio, enrolment and application of various colleges in the United States. This dataset is taken from the 1995 U.S. News & World Report's Guide to America's Best Colleges.**

The raised questions will be answered by using the following statistical modules in python – NumPy, Pandas and Matplotlib. All the statements will be backed by evidence using figures, tables and plots.

I. Overview of the Dataset

The dataset contains well-compiled information about American colleges classified state-wise, including salary given to the various ranks of professors, average SAT score, per-student-expenditure, student-faculty ratio, enrolment and application of various colleges. The AAUP dataset was taken from the March-April 1994 issue of Academe. Maryse Eymonerie, a consultant to AAUP, was instrumental in supplying the data. The AAUP and the USNEWS datasets are available in the following repository: http://lib.stat.cmu.edu/datasets/colleges/.

II. Scientific Questions/Hypotheses

A. *Which state should an academician prefer to relocate to based on the median of the average incomes of his/her rank across states (full, associate, assistant)?*

B. *Find the top 10 states to spend the highest amount on the salary and compensation of faculty (all ranks)? This analysis is made to understand the educational inclination of different states.*

C. *As a follow-up to the purpose of Q2, find the total number of professors (all ranks) state by state.*

D. *Universities on the east coast of the U.S. have strong competition with the ones on the west coast. They can be compared based on the proportion of full professors amongst all ranks of professors considering only the top 5 colleges on each coast. The top 5 colleges are decided based on the amount spent on the faculty.*

E. *To find the top 5 colleges offering doctoral degrees (I), master's degrees (IIA) and baccalaureate degrees (IIB) respectively. The top 5 colleges are classified based on the amount spent on the faculty.*

F. *Does the quality of students entering a college directly affect the graduation rate? The quality of students entering a college can be estimated based on their SAT scores.*

G. *Consider Harvard University, Massachusetts Institute of Technology (MIT), Stanford University, California Institute of Technology (Caltech) and Princeton University, which are among the top 10 colleges in the US. Compare the acceptance and joining rates of these colleges.*

H. *What is the trend in the average difference between the in-state tuition and the out-of-state tuition across the states?*

I. *What is the trend in the average of the total amount spent by students across the states? This analysis is made to help high school students from other countries decide a suitable state in the US to pursue further education based on their budget.*

J. *Does alumni donation depend on whether the college is public or private?*

III. Details of Libraries and Functions

The following libraries will be used to analyse the given dataset and answer the above questions:

A. *Pandas*

Pandas provides powerful data structures such as Series and DataFrame that allow for the manipulation and analysis of structured data. It has a variety of functions for reshaping, grouping, merging and plotting data that increase the efficiency and performance of data analysis tasks.

pandas.read_csv(): Reads data from a CSV file and creates a Pandas DataFrame for easy manipulation.

pandas.DataFrame.group_by(): Groups data based on one or more columns and allows for applying aggregate functions to each group.

B. *NumPy*

Numpy is a library used for numerical computing in Python. It provides a powerful array object that enables one to perform mathematical operations on large datasets. Numpy has a wide range of mathematical functions for operations such as linear algebra, Fourier transforms, and random number generation.

np.mean(): It is used to calculate the average or mean of an array or a portion of an array along a specified axis.

np.median(): It is used to compute the median value of an array or a portion of an array along a specified axis. The median is the middle value of a sorted array, or the average of the two middle values if the array has an even number of elements.

C. *Matplotlib*

Matplotlib is a plotting library in Python that provides various functions for creating visualizations such as line charts, scatter plots, bar graphs and histograms.

xticks(): It is used to set or get the x-axis tick locations and labels of a plot.

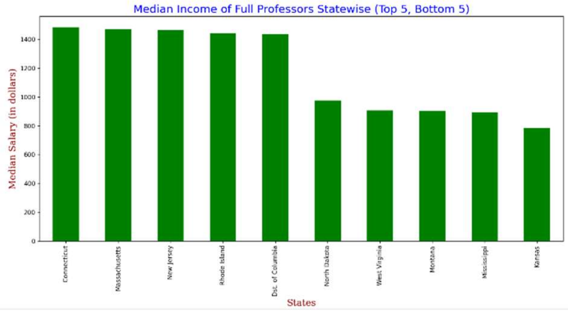
twinx(): It can create a twin y-axis that shares the same x-axis as the original plot.

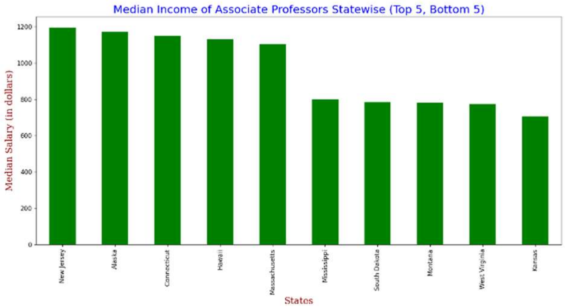
gca(): It is used to get the current axes instance of the current figure.

IV. Answers to the Questions (with Appropriate Illustrations)

A. *Question 1: Which state should an academician prefer to relocate to based on the median of the average incomes of his/her rank across states (full, associate, assistant)?*

Solution: All the colleges are grouped together by state and the average salary and average compensations for full, associate and assistant professors are added together independently. And then, the median of the sums for colleges in a particular state are calculated and plotted against the states as a bar graph using matplotlib.





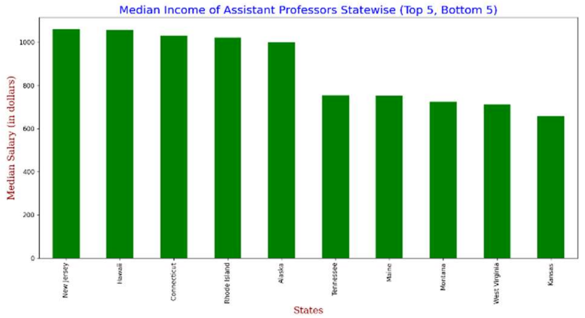


Figure 1: Median Income of Full, Associate and Assistant Professors of the top 5 and bottom 5 states

B. *Question 2 Find the top 10 states to spend the highest amount on the salary and compensation of the faculty (all ranks). This analysis is made to understand the educational inclination of different states.*

Solution: All the colleges in the same state are grouped together, and the total expenditure on the faculty is calculated by adding the average salary and average compensation and multiplying it by9 the number of faculty. The total expenditure is plotted against the top 10 states.

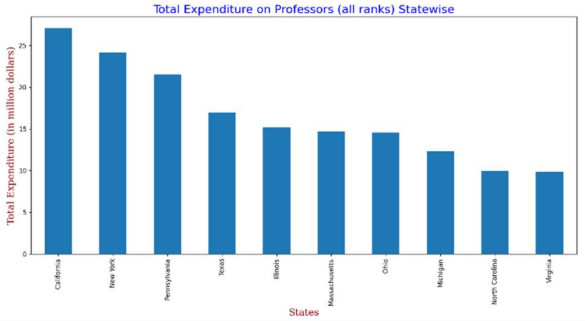
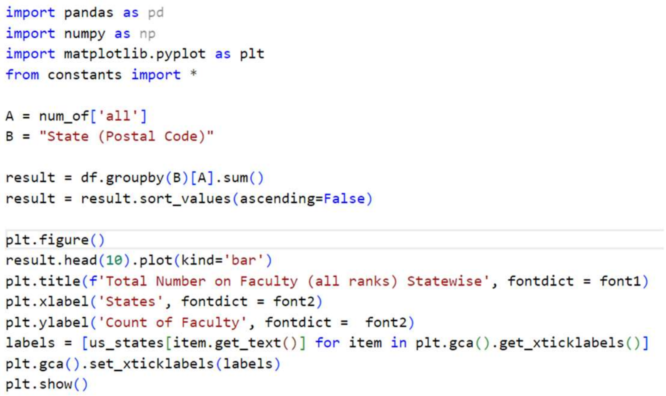


Figure 2: Total Expenditure on Professors State wise

C. *Question 3: As a follow-up to the purpose of Q2, find the total number of professors (all ranks) state by state.*

Solution: All the colleges in the same state are grouped together, and the total number of faculty in all the colleges in a given state is plotted against the state names as a bar graph.



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from constants import *

A = num_of['all']
B = "State (Postal Code)"

result = df.groupby(B)[A].sum()
result = result.sort_values(ascending=False)

plt.figure()
result.head(10).plot(kind='bar')
plt.title(f'Total Number on Faculty (all ranks) Statewise', fontdict = font1)
plt.xlabel('States', fontdict = font2)
plt.ylabel('Count of Faculty', fontdict =  font2)
labels = [us_states[item.get_text()] for item in plt.gca().get_xticklabels()]
plt.gca().set_xticklabels(labels)
plt.show()
```

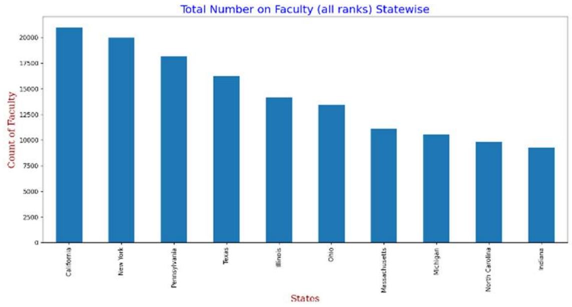


Figure 3: Total Number of Faculty State wise

D. *Question 4: Universities on the east coast of the U.S. have strong competition with the ones on the west coast. They can be compared based on the proportion of full professors amongst all ranks of professors considering only the top 5 colleges on each coast. The top 5 colleges are decided based on the amount spent on the faculty.*

Solution: The west coast and the east coast states are first clearly identified. The amount spent on the faculty is calculated for all the colleges present in the west coast and east coast states separately. The top 5 colleges from the west coast and the top 5 colleges from the east coast are considered. For each of the 10 colleges, proportion of full professors is calculated and plotted against the respective college name as a lollipop graph.

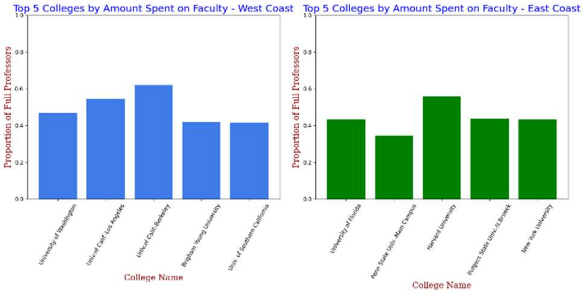
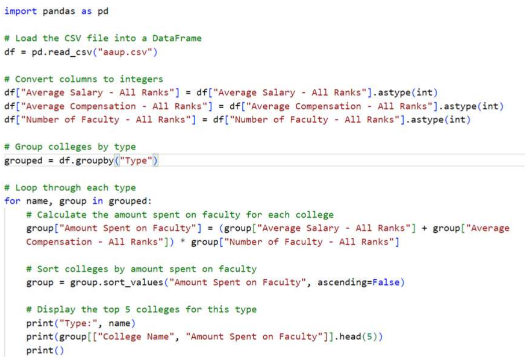


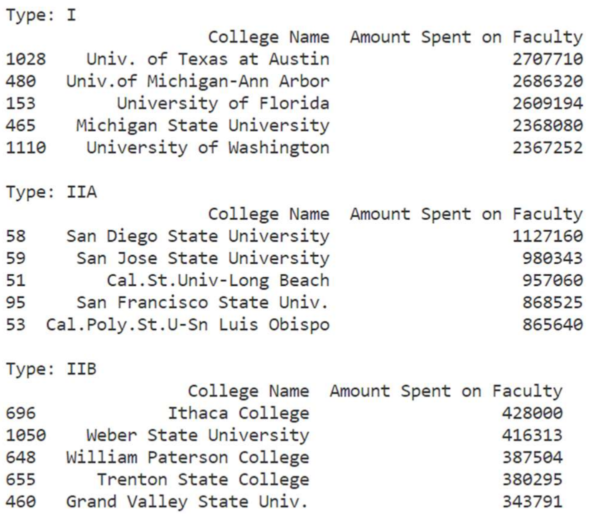
Figure 4: Proportion of Full Professors in the Top 5 colleges

E. *Question 5: To find the top 5 colleges offering doctoral degrees (I), master's degrees (IIA) and baccalaureate degrees (IIB) respectively. The top 5 colleges are classified based on the amount spent on the faculty.*

Solution: All the colleges were first grouped together based on the type (I, IIA, IIB) and the top 5 colleges in each category were chosen based on the amount spent on faculty (= (average salary + average compensation) * number of faculty). The data is then finally displayed as a table.



```
import pandas as pd

# Load the CSV file into a DataFrame
df = pd.read_csv("aaup.csv")

# Convert columns to integers
df["Average Salary - All Ranks"] = df["Average Salary - All Ranks"].astype(int)
df["Average Compensation - All Ranks"] = df["Average Compensation - All Ranks"].astype(int)
df["Number of Faculty - All Ranks"] = df["Number of Faculty - All Ranks"].astype(int)

# Group colleges by type
grouped = df.groupby("Type")

# Loop through each type
for name, group in grouped:
    # Calculate the amount spent on faculty for each college
    group["Amount Spent on Faculty"] = (group["Average Salary - All Ranks"] + group["Average
    Compensation - All Ranks"]) * group["Number of Faculty - All Ranks"]

    # Sort colleges by amount spent on faculty
    group = group.sort_values("Amount Spent on Faculty", ascending=False)

    # Display the top 5 colleges for this type
    print("Type:", name)
    print(group[["College Name", "Amount Spent on Faculty"]].head(5))
    print()
```



```
Type: I
                    College Name  Amount Spent on Faculty
1028    Univ. of Texas at Austin                  2707710
480   Univ.of Michigan-Ann Arbor                  2686320
153        University of Florida                  2609194
465     Michigan State University                 2368080
1110     University of Washington                 2367252

Type: IIA
                    College Name  Amount Spent on Faculty
58     San Diego State University                 1127160
59      San Jose State University                  980343
51         Cal.St.Univ-Long Beach                  957060
95      San Francisco State Univ.                  868525
53  Cal.Poly.St.U-Sn Luis Obispo                  865640

Type: IIB
                  College Name  Amount Spent on Faculty
696             Ithaca College                   428000
1050      Weber State University                 416313
648     William Paterson College                 387504
655        Trenton State College                 380295
460     Grand Valley State Univ.                 343791
```

Figure 5: Top 5 Colleges based on Amount Spent on Faculty

Question 6: Does the quality of students entering a college directly affect the graduation rate? The quality of students entering college can be estimated based on their SAT scores.

Solution: 10 colleges are selected randomly from the dataset, and the combined average SAT score and the graduation rates are plotted on two y-axes against the college names on the x-axis. The lollipop graph is plotted using matplotlib.

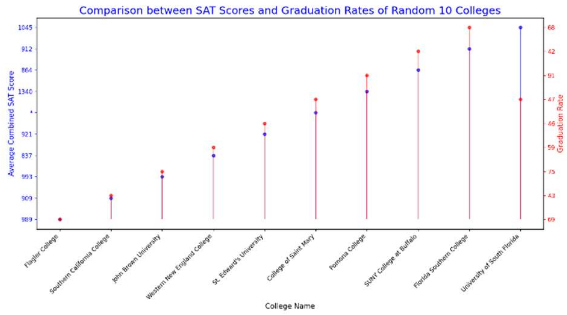


Figure 6: Comparison between SAT Scores and Graduation Rates

F. *Question 7: Consider Harvard University, Massachusetts Institute of Technology (MIT), Stanford University, California Institute of Technology (Caltech) and Princeton University, which are among the top 10 colleges in the US. Compare the acceptance and joining rates of these colleges.*

Solution: Firstly, the acceptance rate (= number of applicants accepted/number of applications received) and the enrollment rate (= number of new students enrolled/number of applications received) are calculated for the given colleges. The two new quantities are plotted against the college names as a line graph.

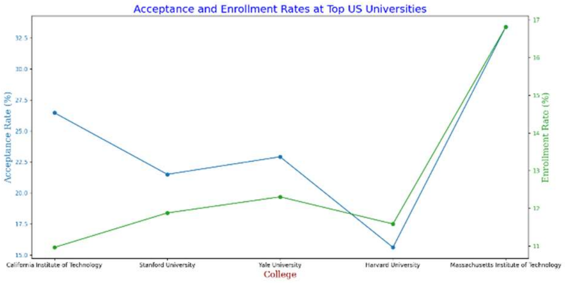


Figure 7: Comparison between Acceptance and Enrolment Rates

G. Question 8: What is the trend in the average difference between in-state tuition and out-of-state tuition across the states?

Solution: All the colleges are grouped together statewise and the mean in-state tuition fee and out-of-state tuition fee is calculated. The difference in the average in-state and out-of-state tuition fee is calculated for each state and represented as a lollipop graph in increasing order.

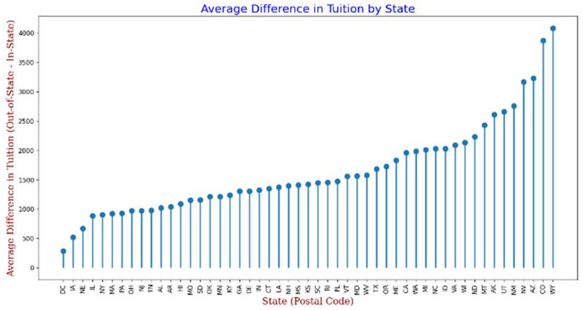


Figure 8: Average Difference between In-state and Out-of-state Tuition Fee across States

H. Question 9: What is the trend in the average of the total amount spent by students across the states? This analysis is made to help high school students from other countries decide a suitable state in the US to pursue further education based on their budget.

Solution: All the colleges are grouped together state wise, and the total expenditure of the students is calculated by adding out-of-state tuition fee, room and board costs, additional fees, estimated book costs and estimated personal spending. The average of the total expenditure is taken across the states and this quantity is then plotted against the state names.

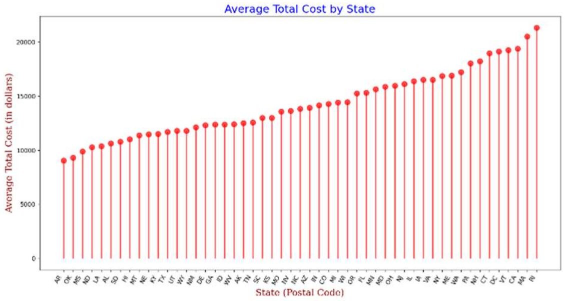


Figure 9: Average Total Expenditure State wise

I. Question 10: Does alumni donation depend on whether the college is public or private?

Solution: All the colleges are first grouped together based on the type (1 = public, 2 = private) and then the average of the percentage of alumni who donate is calculated. The data is then finally displayed as a table with the average percentage of alumni donations for public and private.

Public/Private	Avg. Pct. Of Alumni Donations
Private	24.582873
Public	13.449438

Figure 10: Average of the Percentage of Alumni Donations for Public and Private Universities

V. Summary of the Observation

A. Median Income of Professors Statewise

For the role of full professors, academicians should move into states like Connecticut, Massachusetts and New Jersey. New Jersey and Alaska are favourable ones for associate professors. New Jersey and Hawaii are good choices for assistant professors. Kansas should be avoided for the roles of full, associate and assistant professorship as it provides the lowest median salary to all 3 ranks.

B. Total Expenditure on Faculty Statewise

California is the most educationally inclined state as it spends over 25 million dollars on education. It is closely followed by New York and Pennsylvania which spend a fairly good amount of over 20 dollars.

C. Total Number of Faculty Statewise

Similar to the observations of the previous question, California, New York and Pennsylvania are once again the states to have the maximum faculty strength. This inference is definitely expected as the 3 states house some of the most reputed colleges like California Institute of Technology (Caltech), Cornell University and Carnegie Mellon University respectively.

D. Proportion of Full Professors on the West coast and East coast

Comparing the graphs of the west and east coasts, the proportion of full professors seems to be marginally higher in the colleges on the west coast. Univ. of Calif-Berkeley and Univ. of Calif.-Los Angeles clearly assert their dominance. On the east coast, Harvard University clearly stands out.

E. Top 5 Colleges of each type (I, IIA, IIB)

Univ. of Texas at Austin spends the highest amount on the faculty among the research-oriented universities. San Diego State University comes at the top amongst post-graduate universities and Ithaca College holds the first place when it comes to undergraduate degrees. The comparisons were based on the amount spent on faculty.

F. Comparison between SAT Scores and Graduation Rates

From the graph, it is clearly visible that there is a direct relation between the SAT of the incoming students of a college and the graduation rates of the outgoing students. But at the same time, there will also be cases where the motivation levels of a student might change after joining some colleges and might result in a decline in the graduation rate.

G. Comparison between Acceptance and Enrolment Rate

From the line graph, it is seen that MIT has the highest acceptance as well as enrollment rates. On the other hand, Harvard University is the exact opposite which has low acceptance and enrollment rates. There is a considerable gap between the two rates for the California Institute of Technology. This graph makes it clear that it is comparatively easier for a student to get into MIT as compared to Harvard University.

H. Average difference in Tuition Fee by State

District of Columbia D.C. has the lowest difference in the in-state and out-of-state tuition fees. This is probably done to encourage more abroad students to join colleges in DC. On the other hand, Wyoming makes a clear distinction between in-state and out-of-state students. It discourages a fraction of possible candidates.

I. Average Total Expenditure Statewise

Students planning to move to Rhode Island and Massachusetts should be wary of the high expenditure rates. On the other hand, Arkansas and Oklahoma have reasonably lower expenditure rates, which students should consider while choosing a college for higher studies.

J. Alumni Donation dependency on Public/Private

Private universities receive a donation from about 24% of the alumni and for public universities, it comes to around 13.5%. This trend is as expected because public universities already receive grants from the government, making the alumni less likely to think of donations.

ACKNOWLEDGMENT

I would like to thank Professor Shanmuga Raman for giving me the opportunity to work on the dataset provided and use my data analysis skills to extract valuable information and trends from the source.

REFERENCES

[1] "API Reference — Pandas 1.5.3 Documentation," n.d. https://pandas.pydata.org/docs/reference/index.html#api.

[2] "NumPy Documentation," n.d. https://numpy.org/doc/.

[3] "API Reference — Matplotlib 3.7.0 Documentation," n.d. https://matplotlib.org/stable/api/index.html#.