DATA NARRATIVE

Kishan Ved
Computer Science Engineering
Indian Institute of Technology Gandhinagar
Roll Number - 22110122

***Abstract*—This data narrative aims to provide a comprehensive insight into the U.S. News and AAUP datasets and analyze the data it contains.**

I. OVERVIEW OF THE DATASET

The U.S. News data contains information on tuition, room & board costs, SAT or ACT scores, application/acceptance rates, graduation rate, student/faculty ratio, spending per student, and a number of other variables for 1300+ schools. The AAUP dataset includes average salary, overall compensation, and the number of faculty broken down by full, associate, and assistant professor ranks.

II. SCIENTIFIC QUESTIONS AND HYPOTHESES

The dataset has been analyzed and researched upon to answer ten scientific questions, five from each dataset. Some questions also involve the use of both datasets.

Scientific Questions and Hypotheses:

1) How can we cluster the universities based on their state-wise distribution?
2) How can we analyze the total expenditure of a university student? Which are the most expensive and the least expensive universities among the ones mentioned in the datasets?
3) Which universities are easy to get admitted to (based on acceptance ratio)?
4) Provide an analysis of the graduation rate of students.
5) How can we model the graduation rate of a university based on the variable–acceptance ratio?
6) ***Hypothesis*** – Professors having a Ph.D. degree are generally paid higher. (The average salary of a university increases with an increase in the number of professors having a Ph.D. degree.)
7) Mine the dataset to find the states having average salary offered in all its universities to be greater than $50,000 a year.
8) ***Hypothesis*** – The average salary of professors in a university increases with an increase in the total university expenditure of a student.
9) ***Hypothesis*** – The universities admitting students with a high SAT score are likely to provide more personalized attention to their students.
10) ***Hypothesis*** – Universities offering higher salaries are more likely to offer lower compensation to the faculty.

III. PYTHON LIBRARIES AND FUNCTIONS USED

The following Python libraries and their mentioned functions were used to analyze the dataset:

- NumPy:
  - np.array() – To generate a NumPy array to operate upon.
  - np.arange() –To get a NumPy array having consecutive values.
- Pandas:
  - pd.read_csv() – To read data from a csv file.
  - pd.DataFrame() – To convert data into a Pandas DataFrame.
  - describe() – To get statistical information about the data.
- Matplotlib:
  - plt.plot() – To plot data.
  - plt.pie() – To display a pie chart of the data.
  - plt.bar() – To plot a bar graph.
- SciPy:
  - linalg.lstsq() – To compute the least square solution of datapoints.

IV. ANSWERS TO THE QUESTIONS

Question 1)
How can we cluster the universities based on their state-wise distribution?

A) Approach:

To cluster universities in the **AAUP** dataset based on their state-wise distribution, we plot a bar graph denoting the states on the x-axis and the number of universities on the y-axis.

Answer - This is achieved by using the value_counts() method available in pandas. The bar graph is plotted using matplotlib's bar() method.

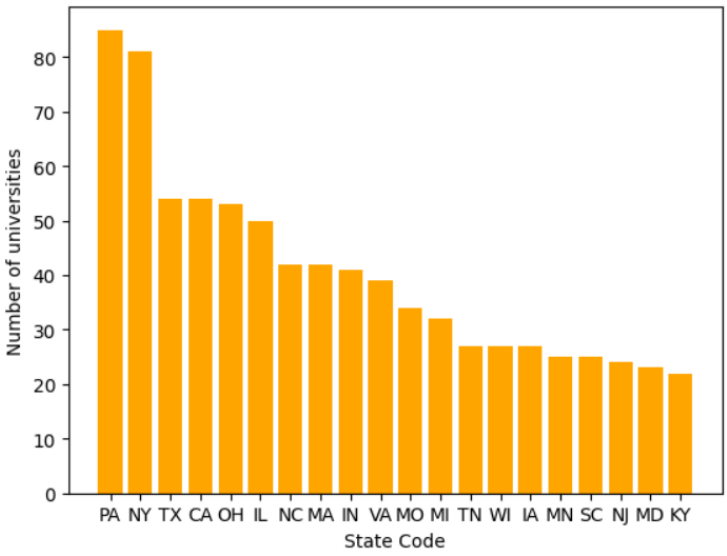


Figure 1: Bar graph showing the state-wise distribution of universities.

Question 2)
How can we analyze the total expenditure of a university student? Which are the most expensive and the least expensive universities among the ones mentioned in the datasets?

A) Approach:

Mine the **U.S. News** dataset and find the estimated total expenditure by adding the room and board costs, estimated book cost, and estimated personal spending of the student.

Use this data to find statistical elements like the mean, standard deviation etc. and also the most and the least expensive universities.

Answer – Statistical elements are obtained using the describe() function.

```
mean      6622.413242
std       1478.296849
min       2588.000000
25%       5568.000000
50%       6468.000000
75%       7441.500000
max      13300.000000
```

Thus, a student's mean estimated total expenditure is $6,622, with a standard deviation of $1,478. The maximum value is $13,300 and the minimum value is $2,588.

The top 5 most expensive universities are:

University Name	Room and Board Cost	Additional Fees	Est. Book Cost	Est. Personal Exp.	Total Expenditure
University of California at San Diego	6607.0	4128.0	630.0	1935.0	13300.0
University of California at Santa Barbara	5990.0	4372.0	630.0	2045.0	13037.0
University of California at Davis	5285.0	4374.0	841.0	2374.0	12874.0
Saint Louis University	4730.0	80.0	800.0	6800.0	12410.0
University of California at Santa Cruz	6081.0	4110.0	633.0	1191.0	12015.0

Thus, we observe that the chain of universities of The University of California are the most expensive universities.

The 5 least expensive universities are:

University Name	Room and Board Cost	Additional Fees	Est. Book Cost	Est. Personal Exp.	Total Expenditure
Shaw University	1738.0	100.0	250.0	500.0	2588.0
LeMoyne-Owen College	1700.0	65.0	500.0	500.0	2765.0
Lees-McRae College	1560.0	458.0	200.0	625.0	2843.0
Grambling State University	1306.0	12.0	500.0	1200.0	3018.0
College of the Ozarks	1900.0	120.0	600.0	500.0	3120.0

Question 3)
Which universities are easy to get admitted to (based on acceptance ratio)?

Approach:

It is easy to get admitted to a university having a high acceptance ratio. The acceptance ratio of a

university is the number of applications received upon the number of students accepted.

Answer – The U.S. News dataset is used to find the acceptance ratio as mentioned above and then find the top 10 values after sorting.

The 10 universities in which one can get admission easily are-

University Name	Acceptance Ratio
California Lutheran University	1.000000
Texas College	1.000000
Northern State University	1.000000
Brewton-Parker College	0.978827
Oral Roberts University	0.972308
College of the Ozarks	0.964497
Peru State College	0.914172
Lee College	0.909639
Wayland Baptist University	0.907407
University of Arkansas at Little Rock	0.903177

Question 4)

Provide an analysis of the graduation rate of students.

Approach: Find statistical elements like the mean graduation rate and its standard deviation. Plot a pie chart denoting the graduation rate approximated to the nearest tens. Find the top 20 universities with the highest graduation rate.

Answer – The statistical elements related to graduation rate are:

	Graduation Rate
count	1203.000000
mean	60.357440
std	18.823692
min	8.000000
25%	47.000000
50%	60.000000
75%	74.000000
max	100.000000

The pie chart denoting the graduation rate approximated to nearest tens is as follows:

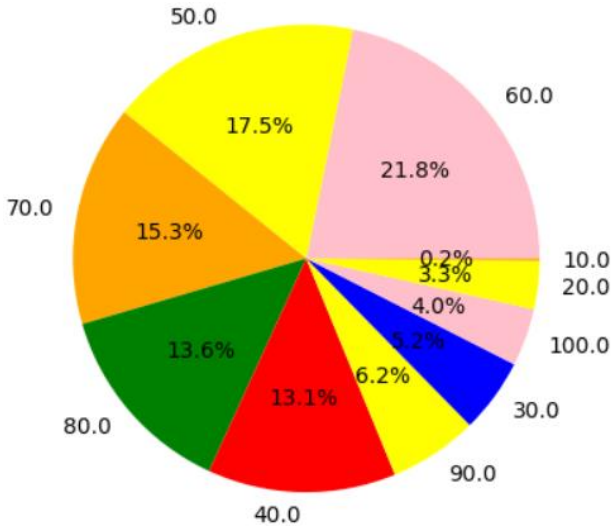


Figure 2: Pie Chart denoting the graduation rate

Hence, most of the universities have a graduation rate of nearly 60%.

The top 20 universities based on graduation rate are:

University Name	Graduation Rate
Harvard University	100.0
College of Mount St. Joseph	100.0
Lindenwood College	100.0
University of Richmond	100.0
Missouri Southern State College	100.0
Harvey Mudd College	100.0
Amherst College	100.0
Heritage College	100.0
Goddard College	100.0
Siena College	100.0
Grove City College	100.0
Santa Clara University	100.0
Williams College	99.0
Columbia University	99.0
York College of Pennsylvania	99.0
Yale University	99.0
Princeton University	99.0
Salem-Teikyo University	98.0
Saint Mary's College	98.0
James Madison University	98.0

Question 5)

How can we model the graduation rate of a university based on the variable–acceptance ratio?

Approach -

Plot the data points corresponding to (acceptance ratio, graduation rate) and then perform a linear regression using scipy's linalg() method, which plots the least square solution for the given set of data points to analyze the trend.

Answer –

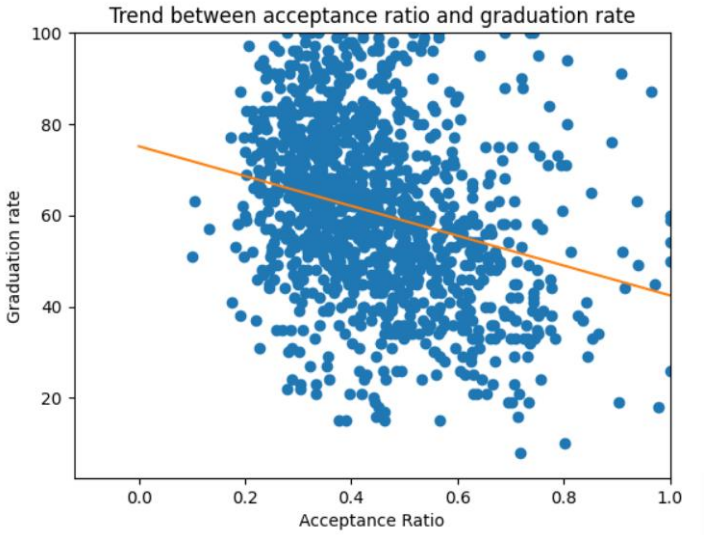


Figure 3: Trend between acceptance ratio and graduation rate.

The following graph clearly indicates that the graduation rate of a university increases with a decrease in the acceptance rate. Thus, a university in which it is difficult to get admitted is more likely to have a high graduation rate.

Question 6)

Hypothesis – Professors having a Ph.D. degree are generally paid higher. (The average salary of a university increases with an increase in the number of professors having a Ph.D. degree.)

Approach –

Plot the data points corresponding to (Number of Ph.D. faculty, Average salary) and then perform a linear regression using scipy's linalg() method, which plots the least square solution for the given set of data points.

Answer -

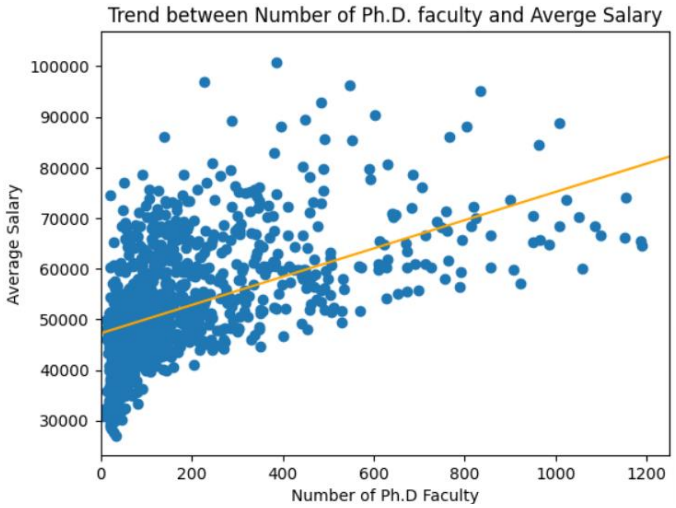


Figure 4: Trend between number of Ph.D. faculty and average salary.

The following graph clearly indicates that the average faculty salary in a university increases with an increase in the number of Ph.D. faculty. Thus, it is preferred to get a job at a university with a large number of Ph.D. faculty as it is likely to offer a higher salary.

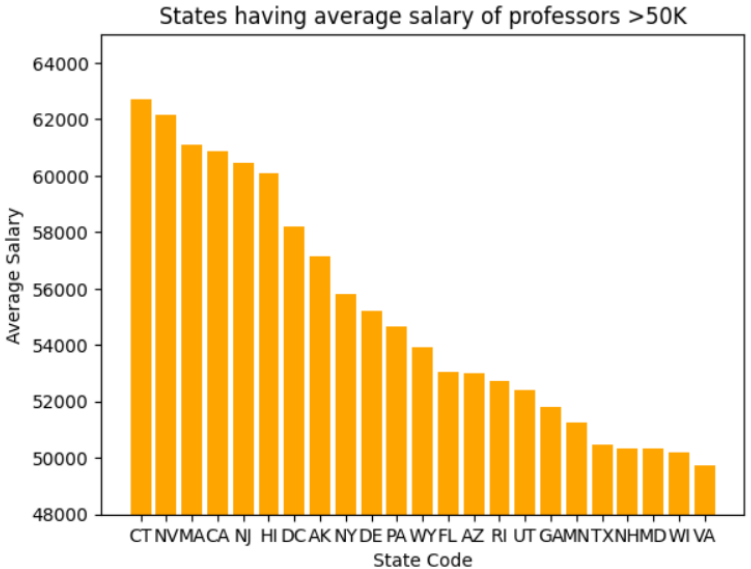
Question 7)

Mine the dataset to find the states having average salary offered in its universities to be greater than $50,000 a year.

Approach-

Cluster the data state-wise. Then calculate the mean of average salaries offered in universities in each state. This gives the average salary offered state-wise. Sorting this and filtering necessary data gives the required answer.

Answer –



Thus, universities in Connecticut (state code CT) offer the highest salary to faculty. There are a total of 23 states having the mean of average salaries greater than $50,000 a year.

Question 8)

Hypothesis - The average salary of professors in a university increases with an increase in the total university expenditure of a student.

Approach - Plot the data points corresponding to (Total expenditure, Average salary) and then perform a linear regression using scipy's linalg() method, which plots the least square solution for the given set of data points to analyze the trend.

Answer –

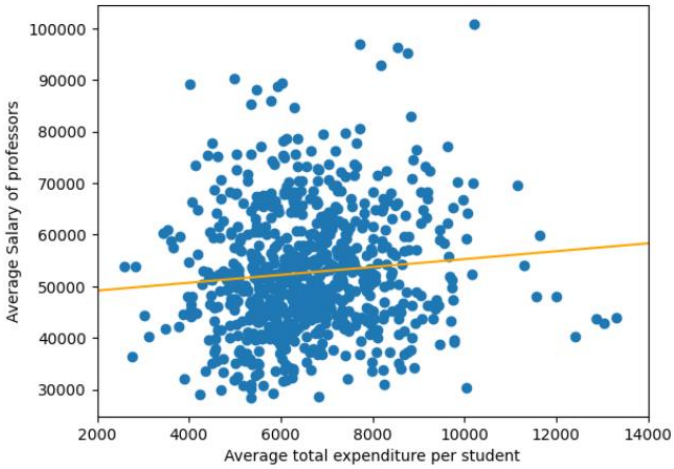


Figure 5: Trend between total expenditure and average salary.

This shows that no significant trend exists between the average total expenditure per student in his university and the average salary of professors. However, as the line has a positive slope, there is a slight increase in the salary with an increase in the total expenditure per student.

Question 9)

Hypothesis – The universities admitting students with a high SAT score are likely to provide more personalized attention to their students.

Approach - Plot the data points corresponding to (Average SAT score, Student to faculty ratio) and then perform a linear regression using scipy's linalg() method, which plots the least square solution for the given set of data points to analyze the trend.

Answer –

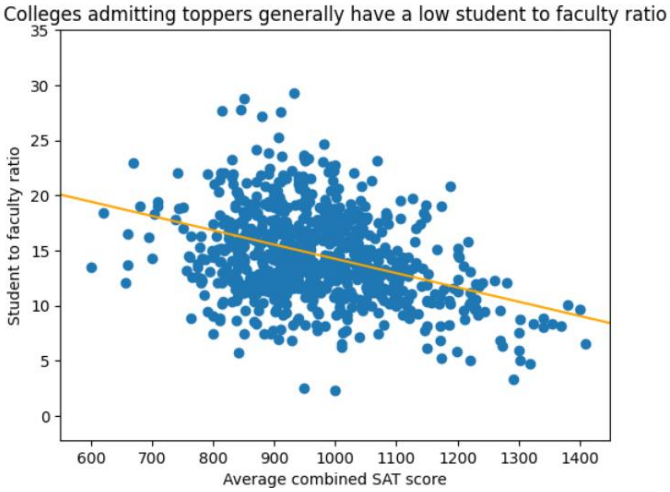


Figure 6: Trend between average SAT score and student to faculty ratio.

As the combined SAT score to enter the university increases, the student to faculty ratio decreases. It is clearly evident from the above graph that a university having students with a higher combined SAT score is more focused in imparting quality education by ensuring personalized attention, which can be judged by the university's student to faculty ratio.

Question 10)

Hypothesis - Universities offering higher salaries are more likely to offer lower compensation to the faculty.

Approach –

Plot the data points corresponding to (Average salary, Average compensation) and then perform a linear regression using scipy's linalg() method, which plots the least square solution for the given set of data points to analyze the trend.

Answer – The following graph shows the trend between the two variables, average salary and average compensation (in $).

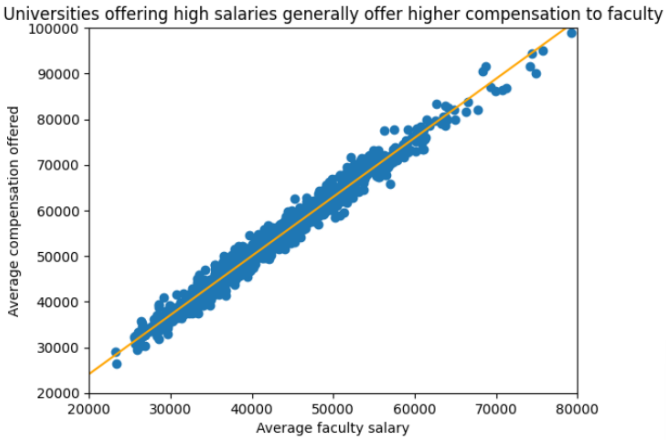


Figure 7: Trend between average salary and average compensation.

As the line has a positive slope, and it is also evident from the data points, a university that offers a high salary to faculty is likely to offer a high compensation as well. Hence, the hypothesis has been disproved.

V. SUMMARY OF THE OBSERVATIONS

The U.S. News and AAUP datasets have been mined in the data narrative to answer 10 questions which include 4 hypothesis. The overview is –

- Thus, a student's mean estimated total expenditure is $6,622, with a standard deviation of $1,478. The maximum value is $13,300 and the minimum value is $2,588.
- California Lutheran University, Texas College and Norther State University are the easiest to get admission in, based on their 100% acceptance policy.
- Most of the universities have a graduation rate of nearly 60%.
- A university in which it is difficult to get admitted is more likely to have a high graduation rate.
- Professors having a Ph.D. degree are generally paid higher.
- Universities in Connecticut (state code CT) offer the highest salary to faculty. There are a total of 23 states having the mean of average salaries greater than $50,000 a year.
- The universities admitting students with a high SAT score are likely to provide more personalized attention to their students.
- Universities offering higher salaries are more likely to offer higher compensation to the faculty.