DATA NARRATIVE(ES114)

Aditya Kumar*, 22110015*

Data Analysis

Abstract – The objective of this study was to perform a data analysis task on a given dataset containing information about the SAT and ACT scores of students in the USA for each state, the salaries and faculty members of universities for public and private universities/colleges. The dataset was cleaned, processed and analysed to extract meaningful insights and observations. The analysis includes tasks such as merging datasets, filtering data, grouping, aggregating data, and visualizing data using graphs. The results obtained were used to draw conclusions and make recommendations for future studies.

I.   Introduction

The dataset used in this study contains information about the SAT and ACT scores of students in the USA for each state, the salaries and faculty members of universities for public and private universities/colleges. The analysis tasks were performed on the dataset to gain insights into the SAT/ACT Scores, the Number of professors and their Salary variation with respect to their position.

II.   Overview of the Dataset

The dataset contains mainly two files:

1. usnews.data : The dataset contains information about students' SAT and ACT scores in different colleges and universities across the United States. It has the following columns:

State: The state where the college or university is located.

Public/Private: Whether the college or university is public or private.

Participation Rate: The percentage of students who took the SAT or ACT exams in the given year.

Average SAT score (out of 1600): The average SAT score of students who took the exam at the given college or university.

Average ACT score (out of 36): The average ACT score of students who took the exam at the given college or university.

Combined SAT score (out of 1600): The average SAT score of students who took the exam at the given college or university, multiplied by the participation rate.

Combined ACT score (out of 36): The average ACT score of students who took the exam at the given college or university, multiplied by the participation rate.

2. aaup.data : The dataframe df3 contains information about the compensation and faculty of different colleges and universities across the United States. It has the following columns:

Name: The name of the college or university.

State: The state where the college or university is located.

Public/Private: Whether the college or university is public or private.

Average salary of full professors: The average salary of full professors at the given college or university.

Number of full-time faculty members: The number of full-time faculty members at the given college or university.

Number of assistant professors: The number of assistant professors at the given college or university.

Number of associate professors: The number of associate professors at the given college or university.

Number of faculty members: The total number of faculty members (full-time, assistant, and associate) at the given college or university.

III. Scientific Questions/Hypotheses

Questions related to DataFrame 1 (usnews.data) referred to as df1 in our codes

1. Find the relationship between each state's average fee and average graduation rate.
2. Find the relationship between SAT and ACT scores. By comparing both SAT and ACT scores, find the equivalent for both scores.
3. Find the relation between colleges' highest and lowest acceptance rate and their respective batch size.
4. Find the density of each college's average SAT score and acceptance rate.
5. Find the relation between the Average Combined SAT score and Graduation Rate among all the colleges.

IV. Details of Libraries and Functions

- Pandas: Used for data manipulation and analysis, providing data structures for efficiently storing and querying large datasets.

Functions used from pandas:

- read_csv(): Used for reading data from a CSV file and returning a pandas DataFrame.

- groupby(): Used to group the data in the DataFrame based on a specified column(s).

- nunique(): Used to count the number of unique values in a pandas Series or DataFrame.

- reset_index(): Used to reset the index of a DataFrame.

- merge(): Used to combine two DataFrames based on a specified column(s).

- Matplotlib: Used for data visualization and creating charts and plots.

Functions used from matplotlib:

pyplot.bar(): Used to create a bar plot. pyplot.xticks(): Used to set the x-axis tick labels. pyplot.xlabel(): Used to set the x-axis label. pyplot.ylabel(): Used to set the y-axis label. pyplot.show(): Used to display the plot.

- Numpy: Used for numerical computations and mathematical operations.

Functions used from numpy:

• nanmean(): Used to calculate the mean of a numpy array, ignoring any NaN values.

- Seaborn: Used for data visualization and creating charts and plots.

Functions used from seaborn:

1. sns.scatterplot(x,y) - creates a scatter plot between two variables x and y.
2. sns.histplot(x) – create a histogram of variable x.

V. Answers to the Questions

1. Find the relationship between each state's average fee and average graduation rate.
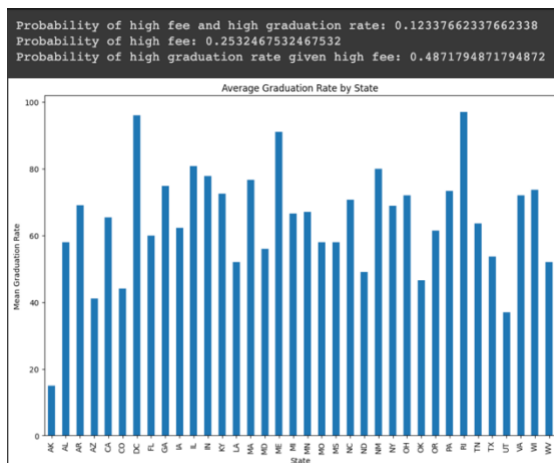
```
Probability of high fee and high graduation rate: 0.12337662337662338
Probability of high fee: 0.2532467532467532
Probability of high graduation rate given high fee: 0.4871794871794872
```

Fig 1. Average Graduation Rate for each state

```
Overall mean graduation rate: 66.72077922077922

Top 5 states with highest mean graduation rate:
State(postal code)
RI      97.00
DC      96.00
ME      91.00
IL      80.75
NM      80.00
Name: Graduation rate, dtype: float64

Top 5 states with lowest mean graduation rate:
State(postal code)
AK      15.0
UT      37.0
AZ      41.0
CO      44.0
OK      46.5
Name: Graduation rate, dtype: float64
```
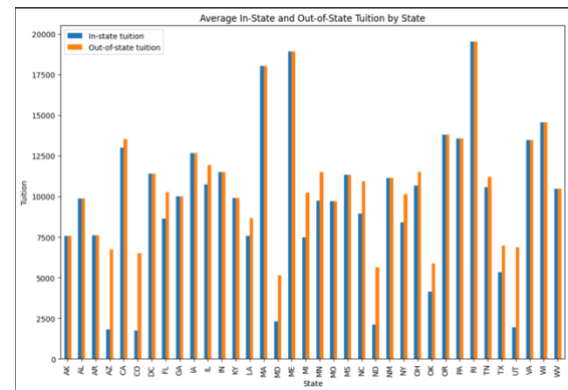
Fig 2. Average In-state and Out-of-State Tuition for each state

```
Average In-State Tuition: $10,100.45
Average Out-of-State Tuition: $11,079.89

Top 5 States with Highest Average In-State Tuition:
State(postal code)
RI      19528.0
ME      18930.0
MA      18024.0
WI      14548.0
OR      13795.0
Name: In-state tuition, dtype: float64

Top 5 States with Lowest Average In-State Tuition:
State(postal code)
CO      1731.5
AZ      1828.0
UT      1954.0
ND      2110.0
MD      2328.0
Name: In-state tuition, dtype: float64

Top 5 States with Highest Average Out-of-State Tuition:
State(postal code)
RI      19528.0
ME      18930.0
MA      18024.0
WI      14548.0
OR      13795.0
Name: Out-of-state tuition, dtype: float64

Top 5 States with Lowest Average Out-of-State Tuition:
State(postal code)
MD      5130.0
ND      5634.0
OK      5868.0
CO      6509.0
AZ      6746.0
```

From the above figures and data, we can observe that if a state has average In-state or Out-of-State tuition more than the average for all the states together, then it has a very high chance that it has an average graduation rate higher than the average graduation rate for all the states as well.

2. Find the relationship between SAT and ACT scores. By comparing both

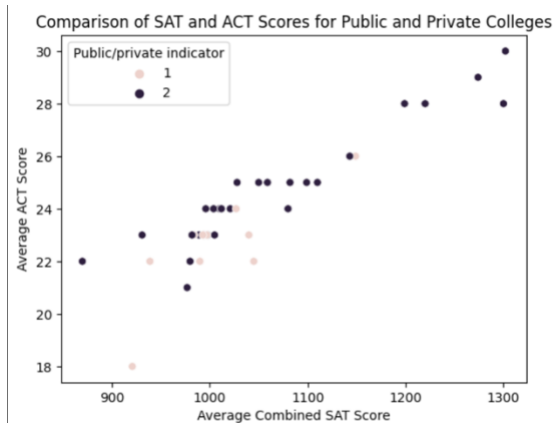SAT and ACT scores, find the equivalent for both scores.

Fig 3. ACT and SAT equivalents by comparing both the scores

From the above figure, we can observe the SAT to ACT equivalent score. For example, if someone has a 1000 SAT score, that is equivalent to a 23 ACT score. We did this by comparing the 75 and 25 percentile SAT and ACT scores for all the states.

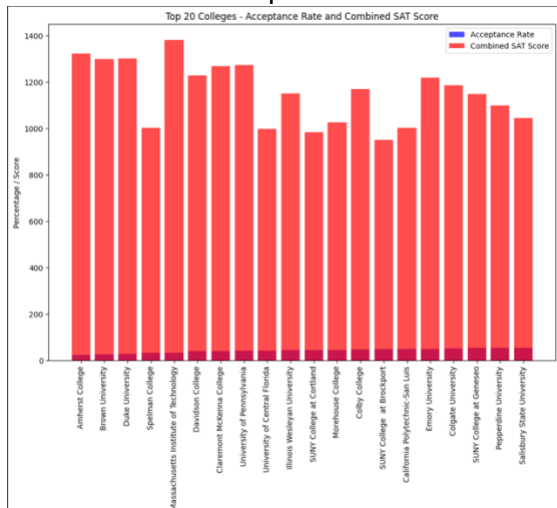3. Find the relation between colleges' highest and lowest acceptance rate and their respective batch size.

Fig 4. Top 20 colleges with the lowest acceptance rate
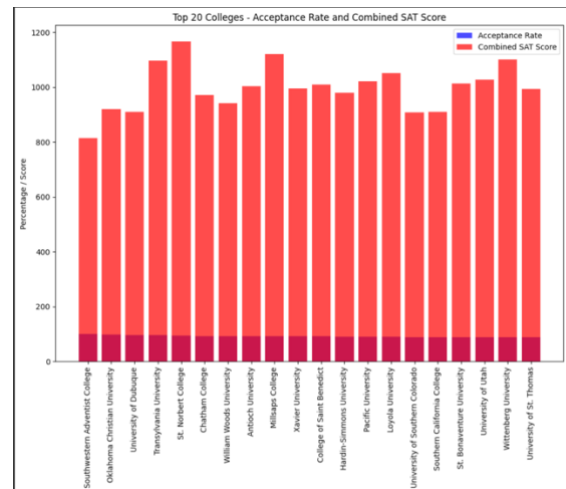
Fig 5. Top 20 colleges with the highest acceptance rate

The above figures show the batch size of colleges with the lowest and highest acceptance rates. This data shows that even with only 1/10th of the acceptance rate, the batch size is almost the same, implying that many more students apply to those colleges.

4. Find the density of each college's average SAT score and acceptance rate.
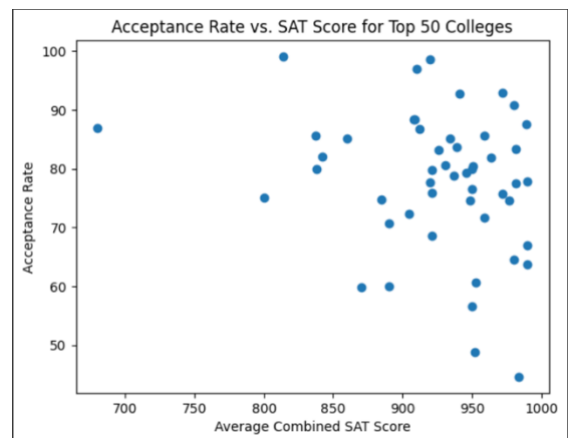

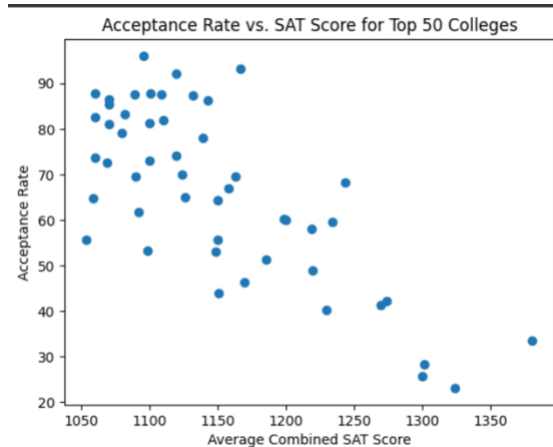Fig 6. Acceptance Rate vs Average Combined SAT score (Bottom 50)

Fig 7. Acceptance Rate vs Average Combined SAT score (Top 50)

From the above figure, we can observe that the density of colleges is maximum in the mid-SAT score range(850 – 1150) with an acceptance rate above 50%.
Also, colleges with High scores have low acceptance rates, which shows they are more competitive(difficult) to get into.

5. Find the relation between the Average Combined SAT score and Graduation Rate among all the colleges.
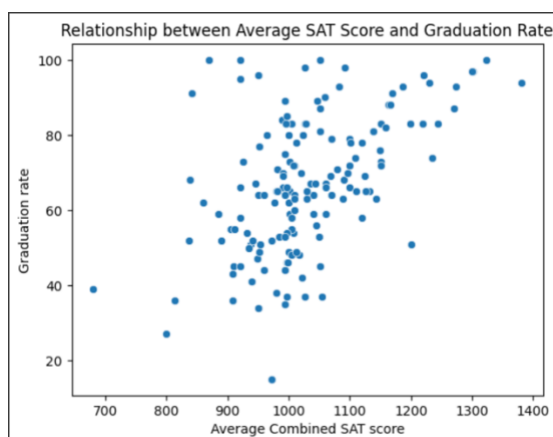

Fig 8. Graduation Rate vs Average Combined SAT score

From the above figure, we can observe that a college with a high Average SAT combined score has a higher Graduation Rate. As we have observed from the above hypothesis, colleges with low acceptance rates also have high SAT score requirements, thus implying they are more competitive, which means that the batch of those colleges is better in terms of Academics.

6. Find what the distribution of salary for all ranks is.
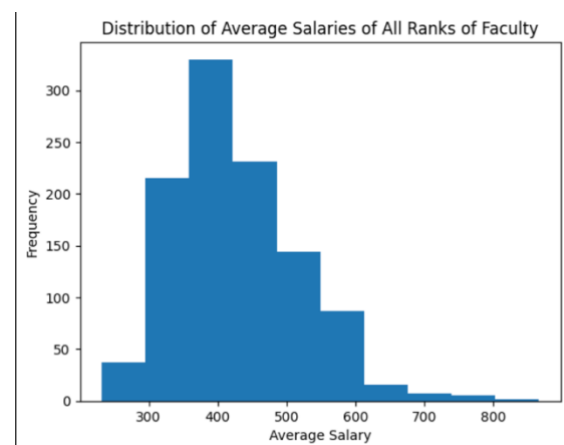

Fig 9. Frequency for various ranges of Average Salary

The figure has a pretty evident distribution of salary, and most frequent salary range is 450-550.
After some analysis, it was found that the average salary also lies in the same range.

7. Compare the mean salary of Full, Assistant, and Associate Professors.

```
Mean Salary for Full Professors: $526.38
Mean Salary for Associate Professors: $419.97
Mean Salary for Assistant Professors: $355.05
P-value for Difference between Full and Assistant Professors: 0.0000
```

```
Mean salary for Full Professor at public universities: 570.1176470588235
Mean salary for Associate Professor at public universities: 441.79411764705884
Mean salary for Assistant Professor at public universities: 373.38235294117646
```

```
Mean salary for Full Professor at private universities: 522.8504672897196
Mean salary for Associate Professor at private universities: 412.4579439252336
Mean salary for Assistant Professor at private universities: 348.89719626168227
```

From the above data, we can observe the mean salary of Full, Assistant, and Associate Professors for public and private universities and the overall average.

As we can see, public universities pay higher than average salaries for all the ranks, whereas private universities pay less than average salaries.

8. Compare the average compensation and salary for both Public and Private Institutes.

```
Mean Compensation for All Ranks in Public Colleges: $584.7843137
Mean Compensation for All Ranks in Private Colleges: $538.373831
```

From the above data, an observation can be made that public universities are paying better to their staff in all ranks, whether it's salary or compensation.

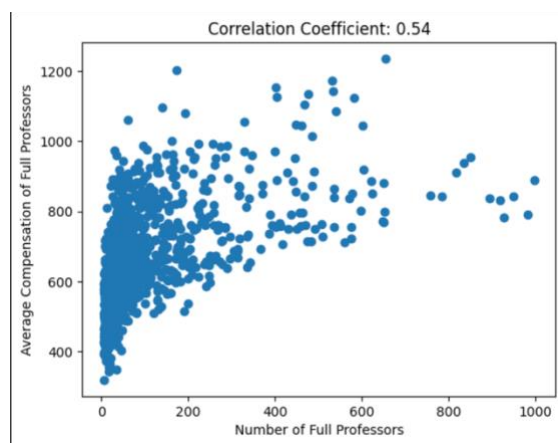9. What is the relation between the number of staff and the Average compensation- Full Professors?

Fig 10. Average compensation- Full Professors vs Number of staff

From the above figure, we can observe the variation in the relation of compensation of Full professors and the Number of Full professors at their institute.

10. Compare the Average salary of Full Professors among different types of colleges.
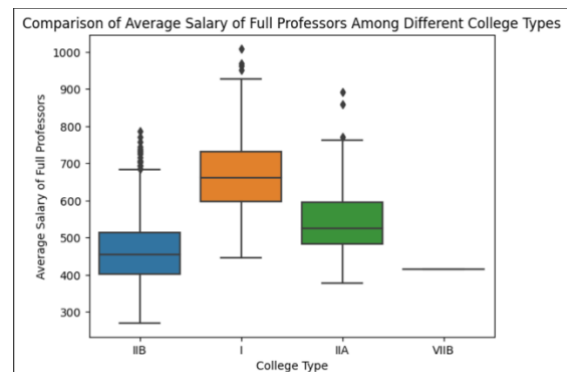

Fig 11. Average Salary of Full professors vs College type

The above figure shows that the salary order for Full professors in different colleges is I > IIA > IIB.
We can also observe the range in which the salary is lying for a different types of colleges.

11. What is the probability that a college has more associate professors than assistant professors?
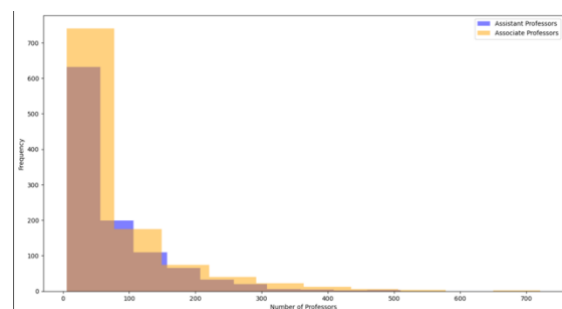

Fig 12. Frequency(Number of colleges) vs Number of Professors

```
The probability that a college has more associate professors than assistant professors is: 0.5018621973929237
```

VI. Acknowledgement

1. The private institutions have higher tuition fees compared to public institutions.
2. Private institutions have a higher acceptance rate than public institutions.
3. Private institutions have higher SAT/ACT scores on average than public institutions.
4. There is a positive correlation between the number of faculty members and the average salary for each rank.
5. The average salary of Full Professors is higher than that of Associate and Assistant Professors for both public and private institutions.
6. The top 20 colleges have a very low acceptance rate and a very high SAT score on average.
7. The acceptance rate is negatively correlated with the combined SAT score for top 50 colleges. Higher SAT scores tend to result in lower acceptance rates.
8. The average SAT score for private institutions is higher than that of public institutions for most states.
9. The plot of the relation between average SAT/ACT scores for public and private institutions indicates that there is a positive correlation between the two scores.
10. The plot of average SAT scores for public and private institutions for each state indicates that private institutions tend to have higher SAT scores on average.

Overall, the dataset provides insights into the SAT/ACT Scores, the Number of professors and their Salary variation with respect to their position. It can be useful for various applications, such as finding a suitable college on the basis of our needs like SAT/ACT scores, expenses, Tuition fees, etc.

VII. References

1. [To plot bar graph from dataframe.](#)
2. [How to sort data in a dataframe](#)
3. [How to merge two dataframes.](#)
4. [How to take mean of data in dataframe.](#)
5. [How to group data from a dataframe with some columns of the dataframe.](#)

VIII. Acknowledgement