Analyzing Universities in the United States

Team 7

Introduction:

Dataset

The dataset for this project is obtained from us-news.

The dataset contains aggregated metrics such as average SAT scores, for universities in the United States. The dataset contains other information such as the tuition and acceptance-rate and the data has been verified not to contain significant outliers.

Research Questions

1. Number of public universities vs private universities per state
   This allows us to find out states which have higher public universities to help students choose from them as they would cost less than private universities.

2. Distribution of Tuition and enrollment in universities across the states.
   The aim is to have an overview of how expensive the tuition of universities is in the United States. Also, we aim to determine the relationship between the tuition fee of a university and the number of enrolled students in the university.

3. Discover which state has universities with the lowest tuition.
   It would be useful to identify the state in the USA that has universities with the lowest tuition, as it would help economical students focus their applications within a particular state.

4. Identify top ranking universities where a good percentage of their students receive aid.
   For economical students with the aspiration to study in top ranking universities, identifying universities with high ranking where a good percentage of their students receive aid would be helpful.

5. Cluster universities into groups based on the average SAT scores of the students.
   The aim here is to create groups where the average SAT scores of the students in the university is what determines the cluster. This would help determine the influence of the SAT scores when compared to another metric like tuition.

Plan

In accomplishing this project, we plan to create tables and graphs to explain the dataset. We will also create a geographic map to visualize the location of the universities. To create the map, we would be using geodata from us-cities. Graphs and tables would be created using python and I would be creating the map using the python folium package.

Additional Dataset used is United States cities databases which has city name and ID, state name, latitude, longitude etc.

Data Analysis:

Data Analysis is the process of applying statistical and/or logical techniques to describe and visualize, and to evaluate the data. Data analysis is an iterative process where it collects data continuously and perform analysis simultaneously.

Exploratory Data Analysis:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

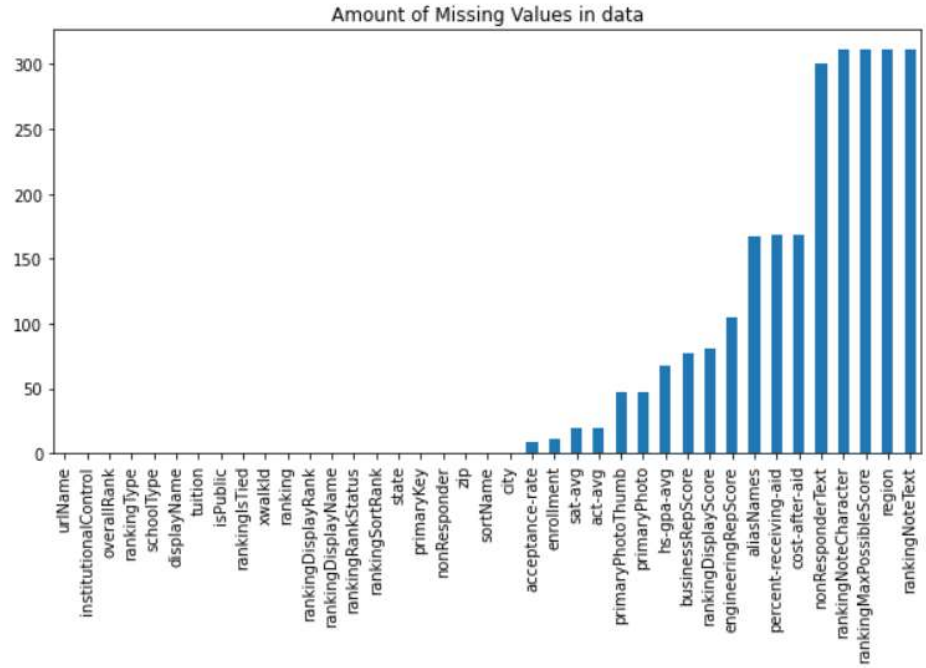
Description of the dataset

Column Variable	Description
primary photo	Photo of the University
primary photo Thumb	Thumbnail of the University
Sat-avg	Average SAT Score of the students in that particular University
Enrollment	Enrollment number for that University
City	City the University is present in
sortName	Name for the University
Zip	Zip code
Acceptance-Rate	Acceptance rate of the university
rankingDisplayScore	rank of the University
Percentage receiving aid	Percentage of students who received aid
cost-after-aid	Tuition fee after students receive the aid
state	State of the University is present in
hs-gpa-avg	High school gpa average of the students present in the University
urlName	University name
rankingDisplayName	Every column has National University
ranking DisplayRank	rank of the University
tuition	Tuition fee for that University
displayName	Display name for that University
schoolType	National University
alias Names	Alias name for that University
ranking type	ranking type of that University
overallRank	Overall rank of the University
institutionControl	Defines whether the university is public, private or proprietary
rankingRankStatus	Defines whether the University is ranked, under ranked or ranking not permitted

Exploratory Data Analysis:

For the start in exploratory data analysis, we first checked the missing data and we observed that rankingNotetext, region, rankingMaxPossibleScore, rankingNoteCharacter, nonResponderText has a greater number of missing values in data.

```
data.isna().sum().sort_values().plot(kind='bar', figsize=(10,5), title='Amount of Missing Values in data')
plt.show()
```

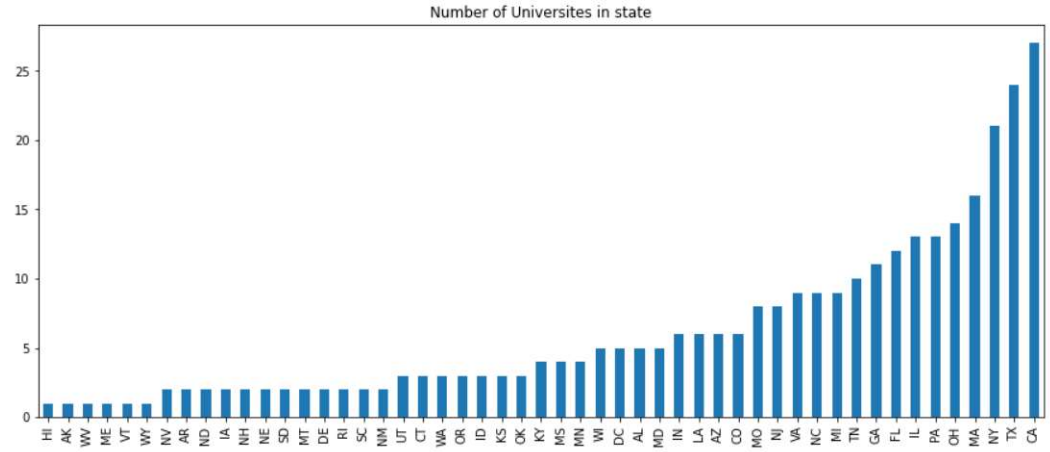


No. of universities in the state:

Then, we visualized the number of universities present in that state and California followed by Texas, New York has a greater number of universities in the state

State

```
In [18]: data.state.value_counts().sort_values().plot(kind='bar', figsize=(15,6), title='Number of Universites in state')
         plt.show()
```

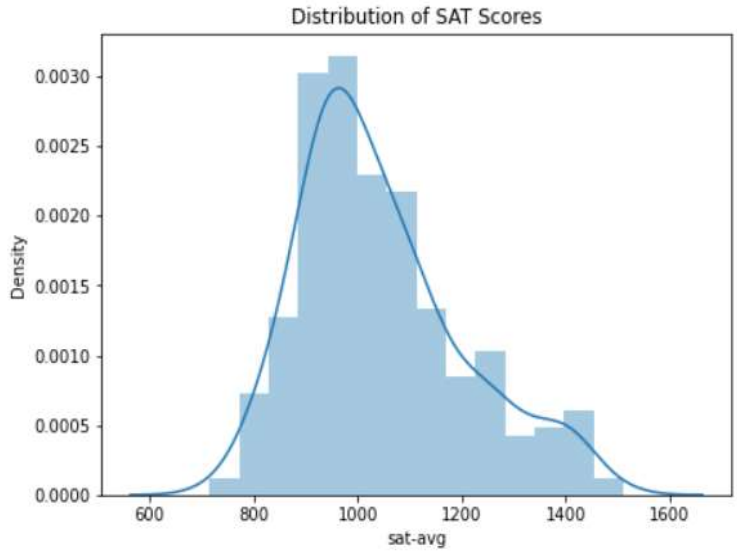


California is the state with the highest number of universities in our dataset

Distribution of the sat scores:

We saw the distribution of SAT Scores, and we observed that SAT Scores are little skewed to the right.

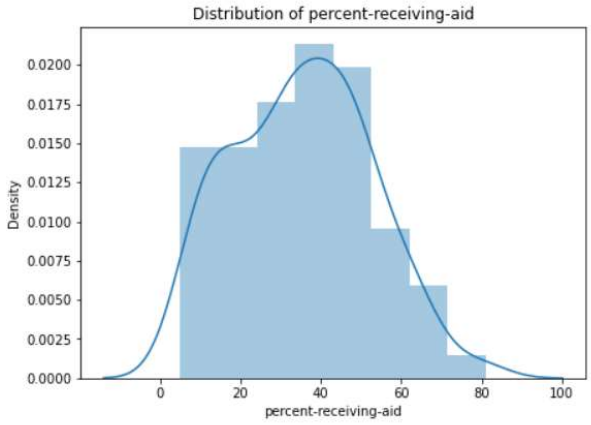
```
plot_dist(data['sat-avg'], 'SAT Scores')
```



Distribution of Percentage Receiving Aid:

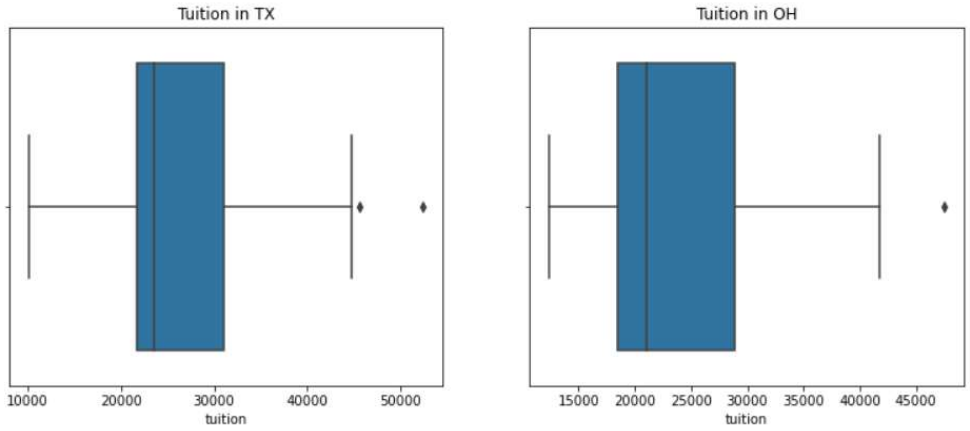
We visualized the distribution of percent receiving aid and we can observe that percentage of university students receiving aid across the states shows a normal distribution.

```
In [23]: plot_dist(data['percent-receiving-aid'], 'percent-receiving-aid')
```



We then compared the tuition fee between two states (Texas and Ohio) and we observed that median tuition for universities in Texas is higher than the median tuition of Universities in Ohio. However, the cheapest and most expensive university of the 2 states is in Texas because the tuition fee in Texas is more widely spread.

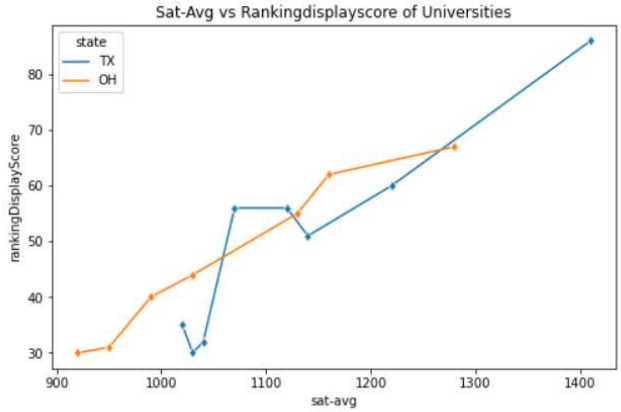
```
In [25]: plot_features(data, y='tuition', kind='boxplot', states=['TX', 'OH'])  # Texas, Ohio
```



We then compared the Sat-Avg vs Ranking Display Score of universities in Texas and Ohio and we observed that there is a positive correlation between the average SAT Scores of University Students and the ranking of the University.

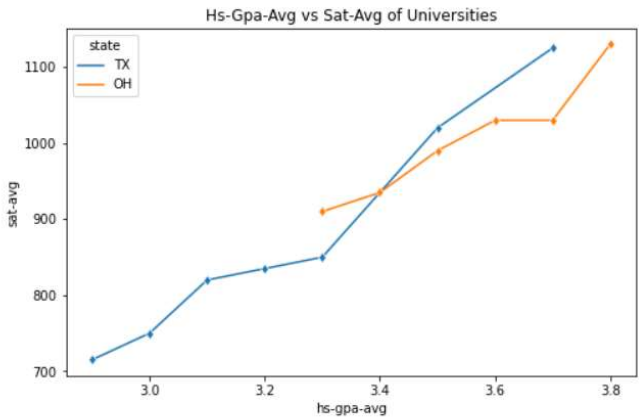
Multivariate Analysis

```
In [26]: plot_features(data, x='sat-avg', y='rankingDisplayScore', kind='lineplot', states=['TX', 'OH']) # Texas, Ohio
```



Comparing the GPA-Avg score and Sat-Avg of Universities and there is also a positive correlation between the average GPA scores of students in their high school and their average SAT Scores.

```
In [27]: plot_features(data, x='hs-gpa-avg', y='sat-avg', kind='lineplot', states=['TX', 'OH']) # Texas,  Ohio
```



Modelling and results:

Research questions and answers:

1) **Number of public universities vs private universities per state.**
   This allows us to find out states which have higher public universities to help students choose from them as they would cost less than private universities.

   From the analysis, Texas has more public universities summing to 19 and New York has 16 private universities which is the highest among other states.

```
In [11]: #Public, Private and proproetary Universities per each state

         univ
```

Out[11]:

state	institutionalControl	
AK	public	1
AL	public	5
AR	public	2
AZ	proprietary	3
	public	3
CA	private	12
	proprietary	2
	public	13
CO	private	1
	public	5
CT	private	2
	public	1
DC	private	5
DE	private	1
	public	1
FL	private	4
	public	8
GA	private	3
	public	8
HI	public	1
IA	public	2
ID	public	3
IL	private	8
	public	5
IN	private	1
	public	5
KS	public	3
KY	private	2
	public	2
LA	private	1
	public	5

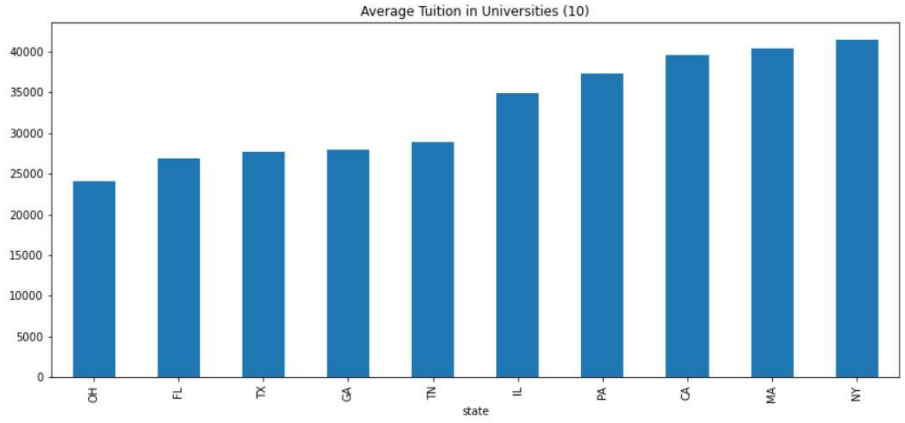
MA	private	12
	public	4
MD	private	1
	public	4
ME	public	1
MI	private	1
	public	8
MN	private	1
	proprietary	2
	public	1
MO	private	4
	public	4
MS	public	4
MT	public	2
NC	private	3
	public	6
ND	public	2
NE	public	2
NH	private	1
	public	1
NJ	private	3
	public	5
NM	public	2
NV	public	2
NY	private	16
	public	5
OH	private	4
	public	10
OK	private	1
	public	2
OR	public	3
PA	private	9
	public	4
RI	private	1
	public	1
SC	public	2
SD	public	2
TN	private	4
	public	6
TX	private	5
	public	19
UT	private	1
	public	2
VA	private	3
	public	6
VT	public	1
WA	private	1
	public	2
WI	private	3
	public	2
WV	public	1
WY	public	1

2) **Distribution of Tuition and enrollment in universities across the states.**

The aim is to have an overview of how expensive the tuition of universities is in the United States. Also, we aim to determine the relationship between the tuition fee of a university and the number of enrolled students in the university.

Average Tuition In Universities:

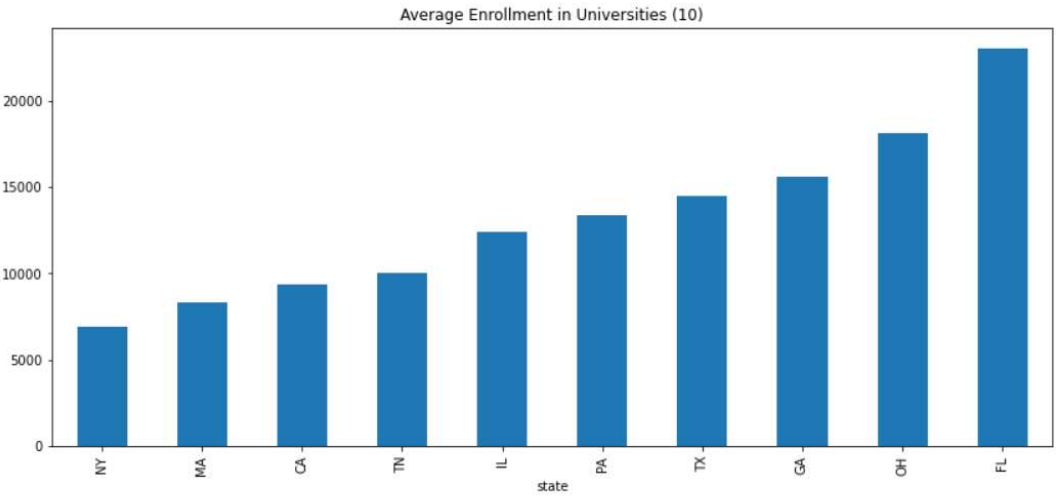
```
In [35]: ten_uni_df.groupby(by=['state'])['tuition'].mean().sort_values().plot(kind='bar', figsize=(14,6))
         plt.title('Average Tuition in Universities (10)')
         plt.show()
```



From the plot above, we can observe that Universities in New York has the highest average(4000) of Tuition in the United States from our balanced sample data

Average Enrollment in Universities:

```
In [36]: ten_uni_df.groupby(by=['state'])['enrollment'].mean().sort_values().plot(kind='bar', figsize=(14,6))
         plt.title('Average Enrollment in Universities (10)')
         plt.show()
```

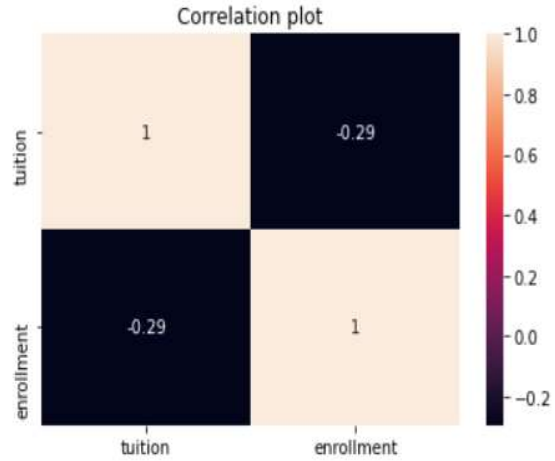


How the tables have turned. New York state universities have the lowest average enrollment, while Florida has the highest average enrollment of university students.

Heatmap between tuition and enrollment:

Heatmap below has confirmed the negative correlation between tuition and enrollment, however it's a weak negative correlation of -0.3

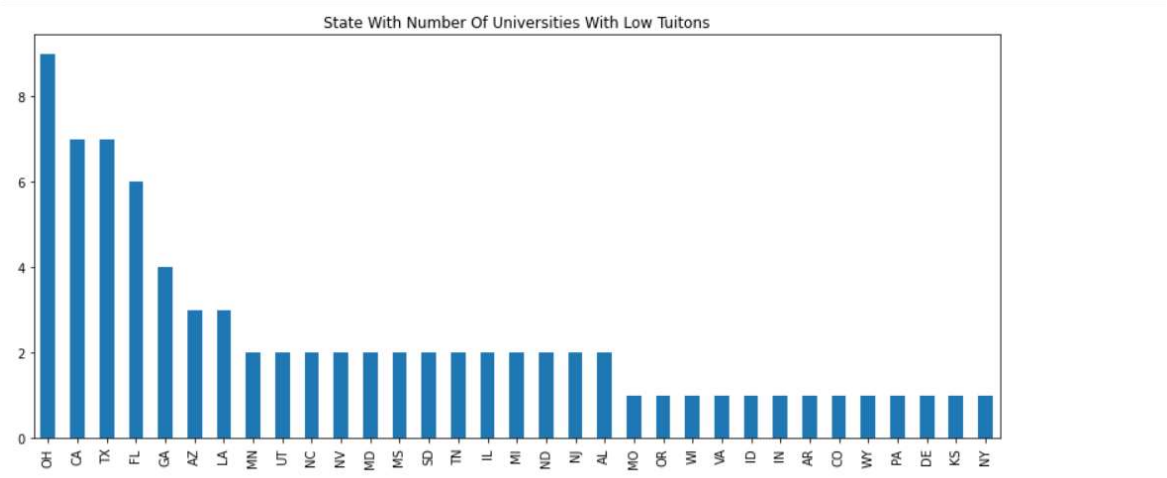
```
In [37]: sns.heatmap(ten_uni_df[['tuition', 'enrollment']].corr(), annot=True)
         plt.title('Correlation plot')
         plt.show()
```



3) Discover which state has universities with the lowest tuition.

It would be useful to identify the state in the USA that has universities with the lowest tuition, as it would help economical students focus their applications within a particular state.

The 25% percentile of tuitions is the threshold we chose to determine if a university's tuition is low.

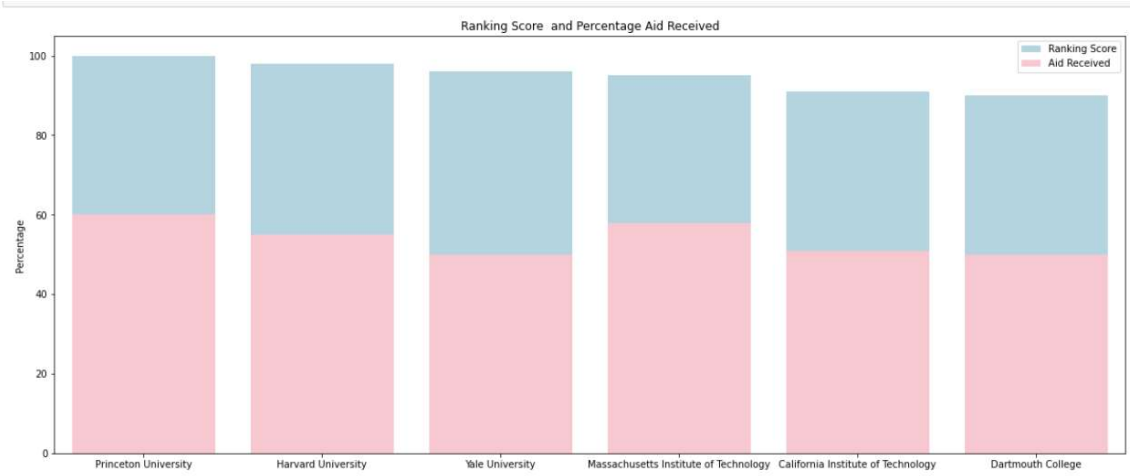


From the graph above, Ohio state has the highest number of universities with 'low' tuition.

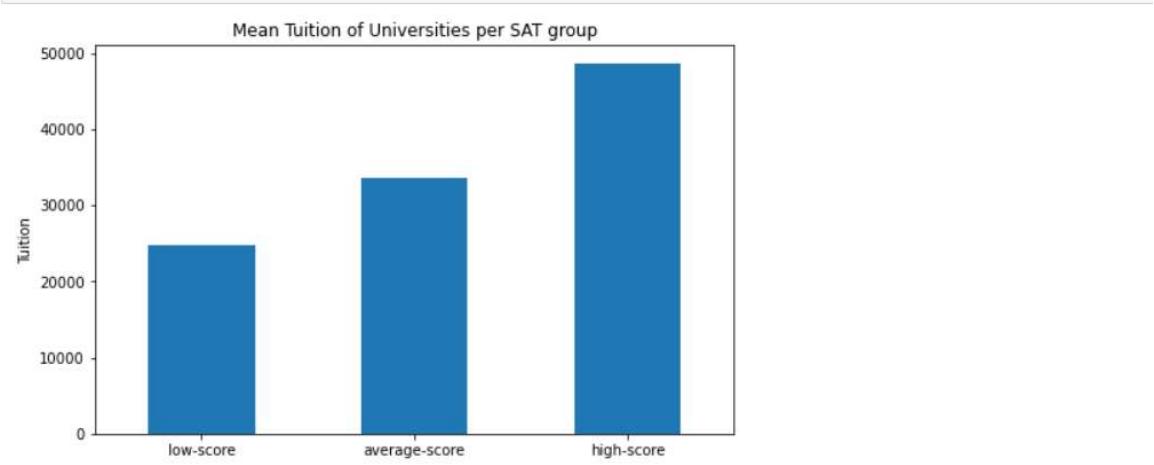
There are a total of 9 Universities that have tuition below the 25th percentile of all tuition fee of Universities across United States

4) **Identify top ranking universities where a good percentage of their students receive aid**.
For economical students with the aspiration to study in top ranking universities, identifying universities with high ranking where a good percentage of their students receive aid would be helpful.

We Plotted top Ranking Universities and Percentage Aid Received.
Form the plot below, we see that the Universities that have at least 50% of students receiving aid and have well above 80 ranking score.
Princeton University has 100 rank score and 60% of its students receiving aid.



5) **Cluster universities into groups based on the average SAT scores of the students.**
The aim here is to create groups where the average SAT scores of the students in the university is what determines the cluster. This would help determine the influence of the SAT scores when compared to another metric like tuition.



From the chart above, we see that students with high SAT scores are likely to go to universities with high tuition

From the plot above, we observed that the students with high SAT scores are likely to go to universities with high tuition.

Map Visualization

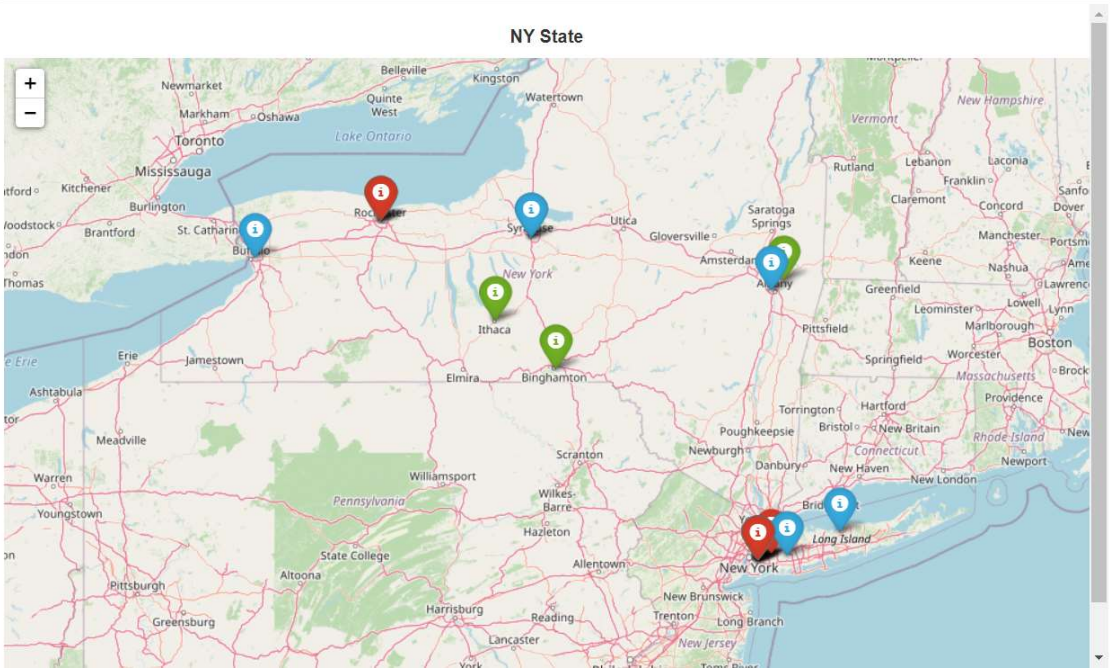
Green represents universities that fell into the **high** sat-scores group.

Blue represents universities that fell into the **average** sat-scores group.

Red represents universities that fell into the **low** sat-scores group.

The above map shows the locations of Universities in the United State.

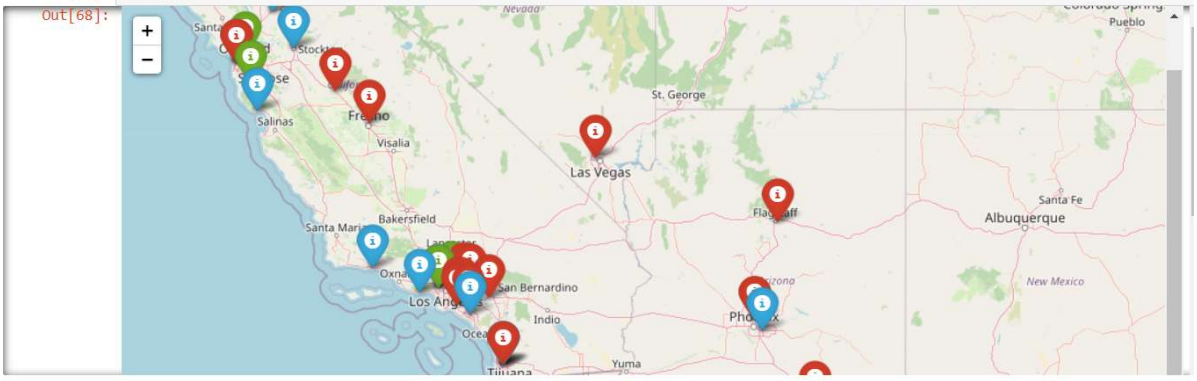
```
In [65]: new_york_map
Out[65]:
```



Above map shows the location of universities present in the New York State

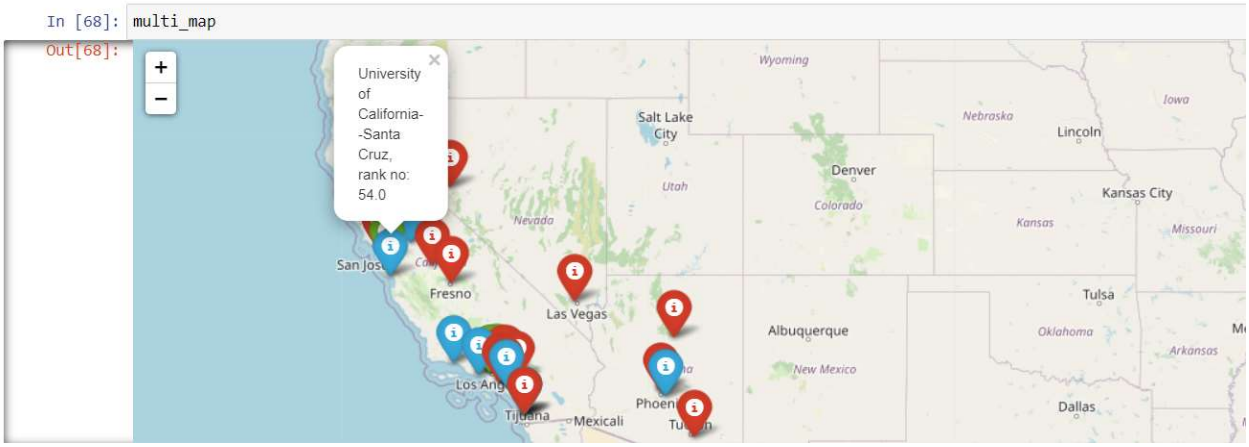
Map below shoes the location of Universities in California, Arizona, and Nevada States.

```
In [66]: multi_map = folium.Map(location=[40.4, -74.7], zoom_start=11)
In [67]: plot_map(geo_df, multi_map, states=['CA', 'AZ', 'NV'])  # ['California', 'Arizona', 'Nevada']
In [68]: multi_map
Out[68]:
```



When we click on the point, we can see that there will be a pop of university name and its rank displayed. Below is an example of that.

University of California- Santa Cruz and its rank is 54.



Discussion and future work:

As we performed different data analysis on the dataset, we dropped few columns at the beginning which has 60% or more missing values. These values won't be useful for the analysis. We also filled the null values for a field with mean of its data. One interesting observation we had is New York has the highest average tuition in universities and in turn New York has the lowest enrollment in their universities. Florida has the highest average enrollment to their universities, and this can be because the population of Florida is high.

For the future work, we can build a Machine Learning model using Logistic Regression which predicts whether the University is a public or private University.

We can also develop a model using 2-Principal of Component Analysis to predict whether the University is a private or public University.

References:

US Cities Database | Simplemaps.com

What is Data Wrangling?

Data Wrangling: What it is and why it's important

Let's Understand All About Data Wrangling

Data Analysis

Best Data Wrangling Tools

Pandas Data Wrangling Data Sheet

Python Data Wrangling Tutorial

What is Exploratory Data Analysis? | by Prasad Patil | Towards Data Science

chapter4.pdf (cmu.edu)

2d Histogram (plotly.com)

matplotlib Tutorial => Basic Plots (riptutorial.com)

Matplotlib — Visualization with Python

seaborn: statistical data visualization — seaborn 0.11.2 documentation (pydata.org)

Folium — Folium 0.12.1 documentation (python-visualization.github.io)

Leaflet - a JavaScript library for interactive maps (leafletjs.com)