College Analytics

A Visual storytelling of Colleges in US

Data Analysis Capstone Project

-Shilpa Rao

Checkpoint 1 : Selecting your business issue and dataset

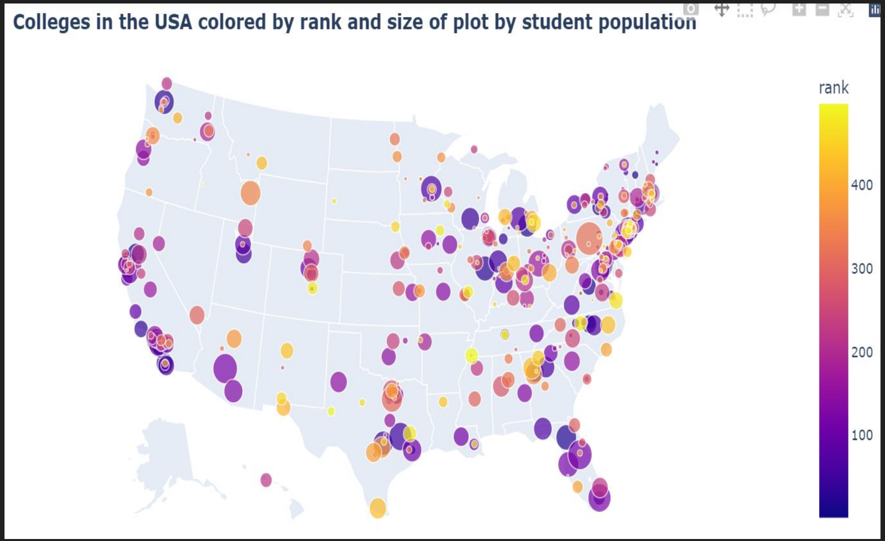
The data set I am using gives a detailed listing of top American colleges. This project will allow prospective students and families to get a better understanding of the opportunities in front of them. College searching can be very difficult to navigate especially for immigrant families or those who have not gone through this process before. This visual storytelling is an attempt to make this process easier for all
Problem : Navigating search for Colleges in US

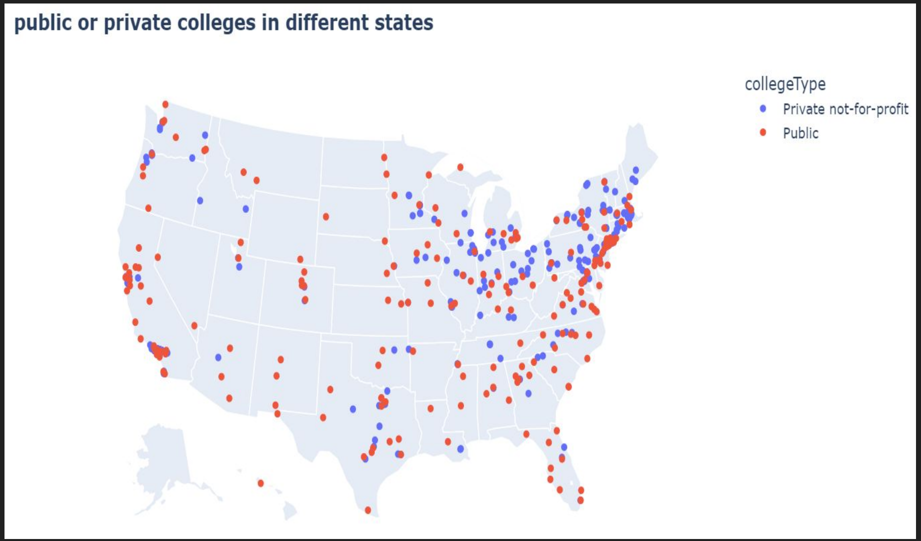
- Solution : Provide a visual analysis of Colleges in US for future students
- Feature : Give a comparison of rank of colleges to median base salary
- Feature : Give a visual map of public and private schools
- Feature: Show the opportunity of grants and financial aids
- Full Project Tech stack: Python, Tableau, Jupyter notebook

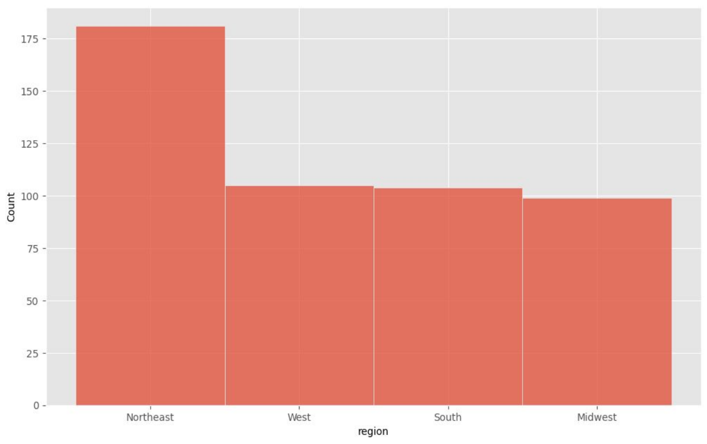
Dataset I used : https://www.kaggle.com/datasets/kabhishm/top-american-colleges-2022

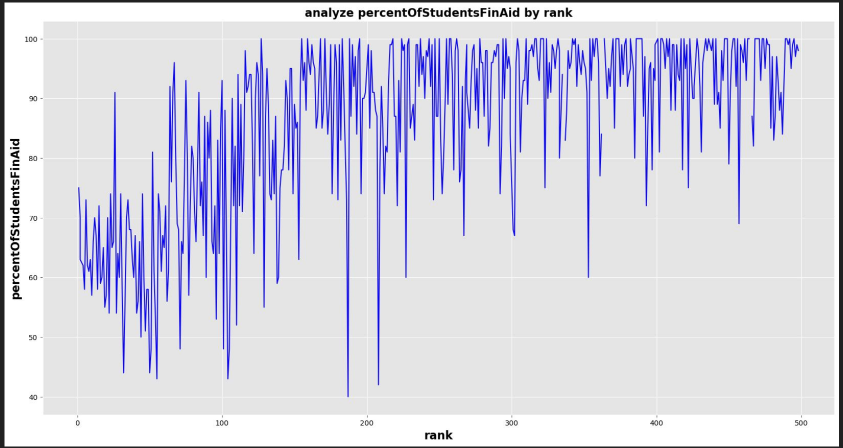
Checkpoint 2 : Exploratory Data Analysis(EDA)

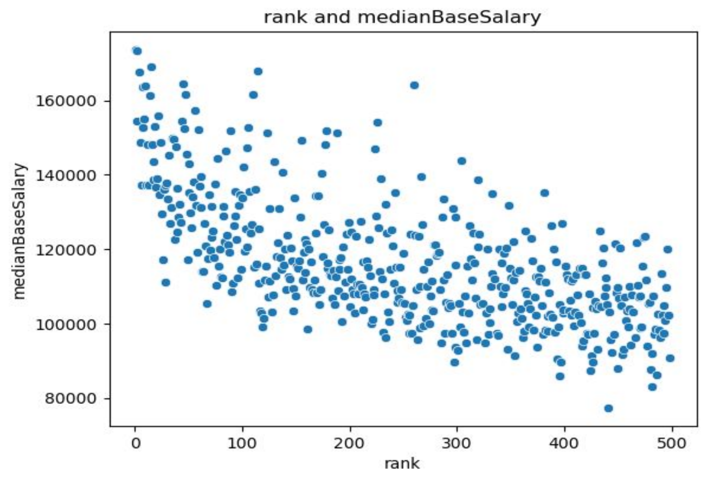
- Understand the data set by checking info, description, first few rows, shape..
- Created multiple visualizations to understand the dataset


Colleges in the USA colored by rank and size of plot by student population
rank
400
300
200
100







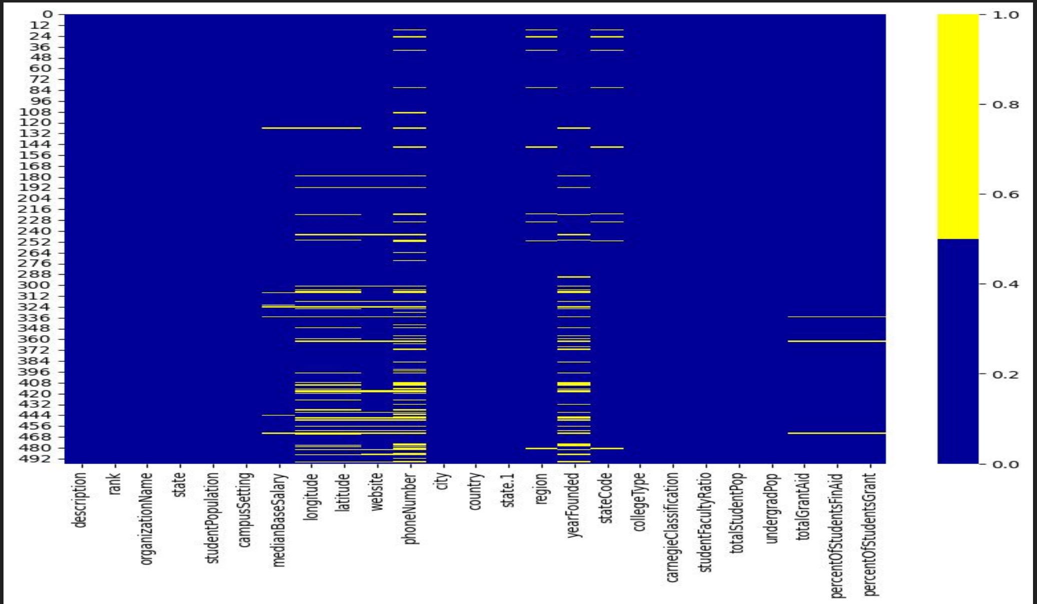


- colleges are concentrated in coastal locations(especially northeast) and lesser in midwest
- looks like higher ranked and higher population schools are connected and are higher in number along the coast
- missing values, and repeated columns need handling
- unnecessary columns such as website, phone number, and country name(since all colleges here are in US) can be removed

Checkpoint 3 : Cleaning Data

- Identified missing data and visualize it
- Dropped website, phone Number - lots of missing data and irrelevant
- Filled missing medianBaseSalaray, totalGrantAid with mean
- Imputed the missing latitude and longitude values - not the best options
- Drop studentPopulation as it has same values as totalStudentPop - unnecessary data

Graph of missing data



Checkpoint 4 : Manipulate, Interpret, and Visualize data

Tableau story reflects the findings from EDA after the Cleaning process

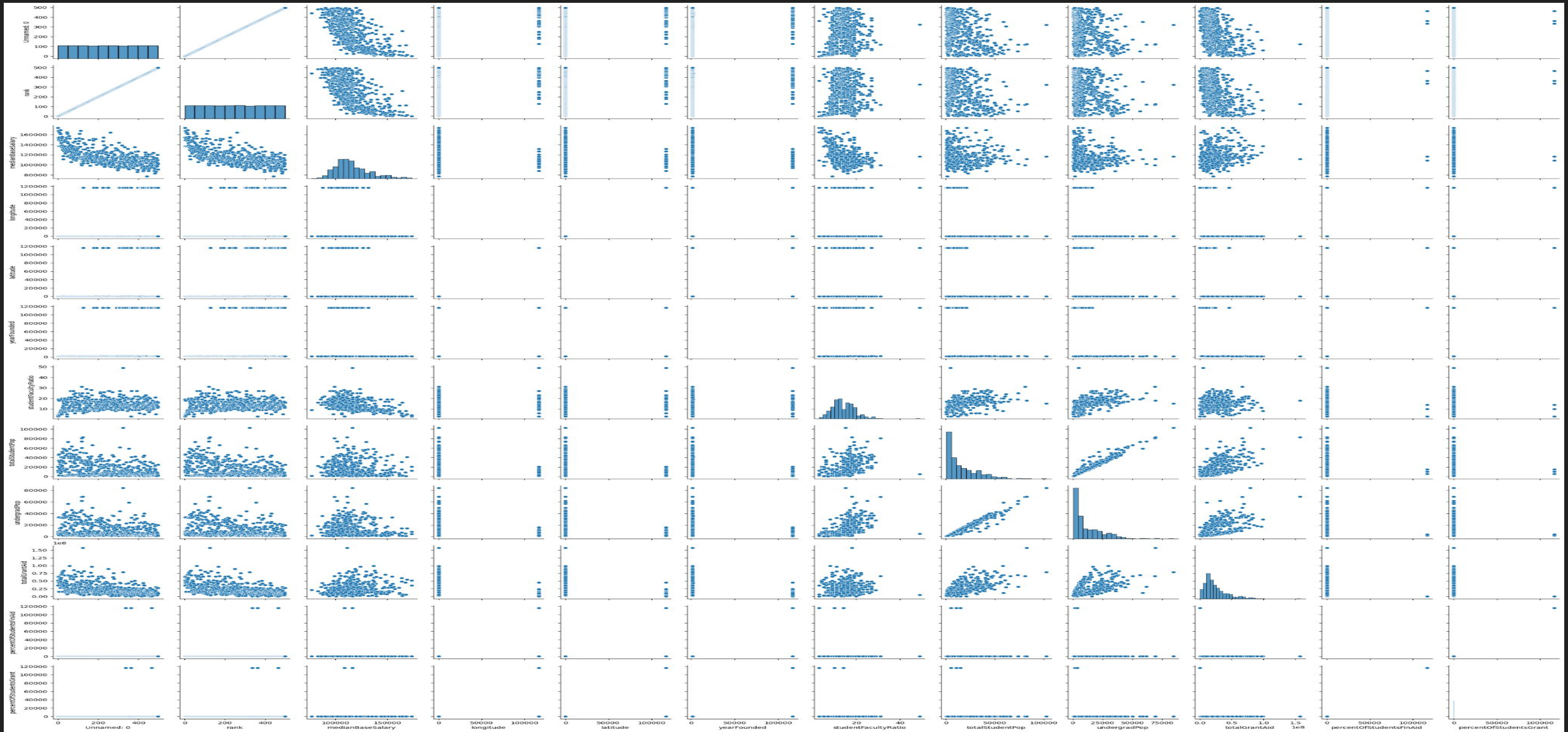
Here is the link to my Tableau story:

https://public.tableau.com/app/profile/shilpa.rao5350/viz/CollegeAnalytics-Story/Story1

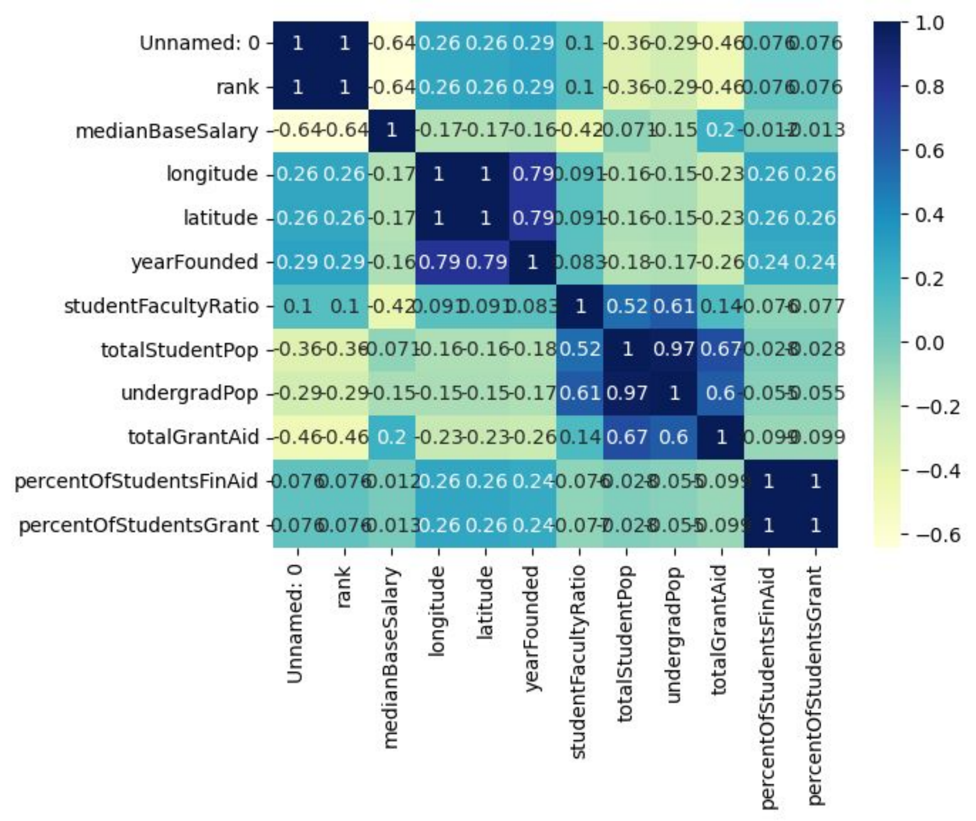
Here is a pdf version of the tableau story:

Checkpoint 5 : Modeling Data

correlation - pairplot for all columns

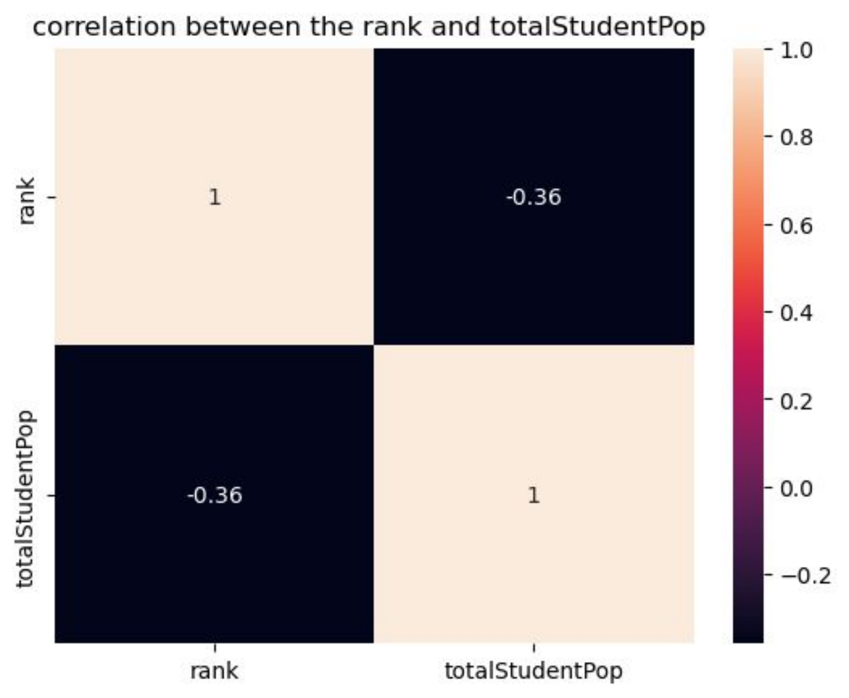


correlation - heatmap for all columns



	Unnamed: 0	rank	medianBaseSalary	longitude	latitude	yearFounded	studentFacultyRatio	totalStudentPop	undergradPop	totalGrantAid	percentOfStudentsFinAid	percentOfStudentsGrant
Unnamed: 0	1	1	-0.64	0.26	0.26	0.29	0.1	-0.36	-0.29	-0.46	0.076	0.076
rank	1	1	-0.64	0.26	0.26	0.29	0.1	-0.36	-0.29	-0.46	0.076	0.076
medianBaseSalary	-0.64	-0.64	1	-0.17	-0.17	-0.16	-0.42	0.071	-0.15	0.2	-0.012	-0.013
longitude	0.26	0.26	-0.17	1	1	0.79	0.091	-0.16	-0.15	-0.23	0.26	0.26
latitude	0.26	0.26	-0.17	1	1	0.79	0.091	-0.16	-0.15	-0.23	0.26	0.26
yearFounded	0.29	0.29	-0.16	0.79	0.79	1	0.083	-0.18	-0.17	-0.26	0.24	0.24
studentFacultyRatio	0.1	0.1	-0.42	0.091	0.091	0.083	1	0.52	0.61	0.14	-0.076	-0.077
totalStudentPop	-0.36	-0.36	0.071	-0.16	-0.16	-0.18	0.52	1	0.97	0.67	-0.028	-0.028
undergradPop	-0.29	-0.29	-0.15	-0.15	-0.15	-0.17	0.61	0.97	1	0.6	-0.055	-0.055
totalGrantAid	-0.46	-0.46	0.2	-0.23	-0.23	-0.26	0.14	0.67	0.6	1	-0.099	-0.099
percentOfStudentsFinAid	0.076	0.076	-0.012	0.26	0.26	0.24	-0.076	-0.028	-0.055	-0.099	1	1
percentOfStudentsGrant	0.076	0.076	-0.013	0.26	0.26	0.24	-0.077	-0.028	-0.055	-0.099	1	1

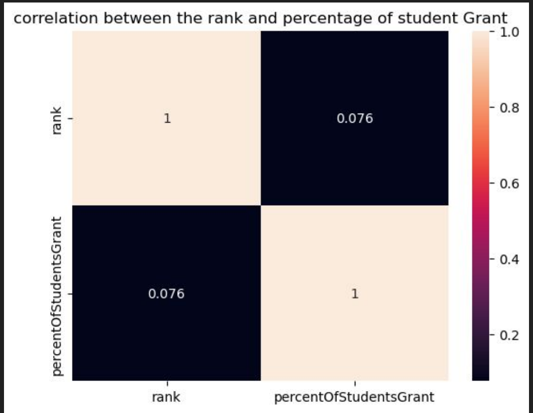
heat map for correlation between the rank of the colleges vs total student population



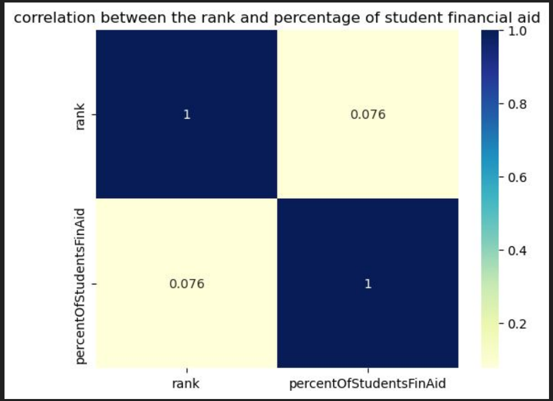
correlation between the rank and totalStudentPop

	rank	totalStudentPop
rank	1	-0.36
totalStudentPop	-0.36	1

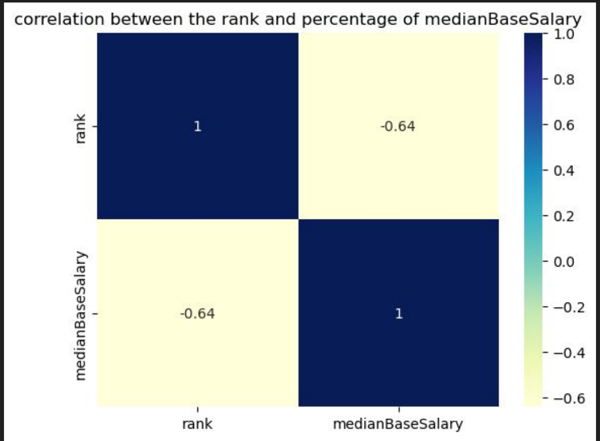
heat map for correlation between the rank of the colleges vs percent Of Students Grant



heat map for correlation between the rank of the colleges vs percentOfStudentsFinAid



correlation between the rank of the colleges vs medianBaseSalary



Conclusion

- colleges are concentrated in coastal locations, especially northeast and lesser in midwest
- looks like higher ranked and higher population schools are connected and are higher in number along the coast
- Higher the rank of the college, higher the median base salary. This means that after graduating from higher ranked colleges, possibility of getting higher salary is more likely
- lower the rank shows higher percentage of financial aids and higher rank shows lower percentage of financial aid
- lower the rank shows higher percentage of grants and higher rank shows lower percentage of grants