CHRIST(Deemed to be University)

BDS471L – Machine Learning

Date: 11/01/2024

LAB-1

Siddharth R Bhardwaj
22112028

1) Dataset: "Top American Colleges 2022"
2) Source: https://www.kaggle.com/datasets/kabhishm/top-american-colleges-2022

```
In [3]: df.head(10)
```

Out[3]:

	description	rank	organizationName	state	studentPopulation	campusSetting	medianBaseSalary	longitude	latitude	website	...
0	A leading global research university, MIT attr...	1	Massachusetts Institute of Technology	MA	12195	Urban	173700.0	-71.093539	42.359006	http://web.mit.edu	...
1	Stanford University sits just outside of Palo ...	2	Stanford University	CA	20961	Suburban	173500.0	-122.168924	37.431370	http://www.stanford.edu	...
2	One of the top public universities in the coun...	2	University of California, Berkeley	CA	45878	Urban	154500.0	-122.258393	37.869236	http://www.berkeley.edu	...
3	Princeton is a leading private research univer...	4	Princeton University	NJ	8532	Urban	167600.0	-74.659119	40.349855	http://www.princeton.edu	...
4	Located in upper Manhattan, Columbia Universit...	5	Columbia University	NY	33882	Urban	148800.0	-73.961288	40.806515	http://www.columbia.edu	...

yearFounded	stateCode	collegeType	carnegieClassification	studentFacultyRatio	totalStudentPop	undergradPop	totalGrantAid	percentOfStudentsFinAid	percen
1861.0	MA	Private not-for-profit	Doctoral Universities: Very High Research Acti...	3	12195	4582	35299332.0	75.0	
1891.0	CA	Private not-for-profit	Doctoral Universities: Very High Research Acti...	4	20961	8464	51328461.0	70.0	
1868.0	CA	Public	Doctoral Universities: Very High Research Acti...	19	45878	33208	64495611.0	63.0	
1746.0	NJ	Private not-for-profit	Doctoral Universities: Very High Research Acti...	4	8532	5516	44871096.0	62.0	
1754.0	NY	Private not-for-profit	Doctoral Universities: Very High Research Acti...	6	33882	8689	44615007.0	58.0	
1919.0	CA	Public	Doctoral Universities: Very High Research Acti...	18	46947	33641	61100980.0	73.0	

The dataset talks about the rankings and the factors on which the universities depends on. The dataset has 496 rows and 25 columns.

Nominal data:

1) Description
2) organizationName,
3) State
4) campusSetting
5) City
6) Country
7) State
8) region
9) stateCode
10) college-type
11) carnegieClassification

Ordinal data:

1) rank

Continous data:

1) medianBaseSalary
2) longitude
3) latitude
4) totalStudentPop
5) undergradPop
6) totalGrantAid
7) percentOfStudentsFinAid
8) percentOfStudentsGrant

Discrete data:

1) yearFounded
2) studentFacultyRatio

2) What is the purpose of your analysis?

- The purpose of analysing the USA universities is to compare some key attributes of the universities like ranking, student population, financial aid and geographical location and how they affect the rankings of these universities in the USA.

3) What business problem or question are you trying to address?

- Using the information provided, the educational institutions can know which factors are more important or relevant to have a higher ranking.

4) What are the goals or objectives of the analysis?
   - We can examine the distribution of universities across the nation and how important the location is or where most universities are situated?
   - We can examine the ranking distributions of the universities
   - We can observe the need for financial aid to be provided to students by the university.
   - We can examine the student and faculty ratio in the university.
   - We can figure out the factors that make the top-ranked universities stand out from the rest of the universities in America.

5) Are there specific variables or columns that are crucial to your analysis?
   - rank :
   - organizationName, yearFounded, collegeType
   - state and region
   - studentPopulation
   - campusSetting
   - medianBaseSalary
   - totalGrantAid, percentOfStudentsFinAid , percentOfStudentsGrant
   - studentFacultyRatio

6) What is the structure of the dataset?

```
: df.shape
: (498, 25)
```

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 498 entries, 0 to 497
Data columns (total 25 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   description              498 non-null    object
 1   rank                     498 non-null    int64
 2   organizationName         498 non-null    object
 3   state                    498 non-null    object
 4   studentPopulation        498 non-null    int64
 5   campusSetting            498 non-null    object
 6   medianBaseSalary         491 non-null    float64
 7   longitude                458 non-null    float64
 8   latitude                 458 non-null    float64
 9   website                  477 non-null    object
 10  phoneNumber              428 non-null    object
 11  city                     498 non-null    object
 12  country                  498 non-null    object
 13  state.1                  498 non-null    object
 14  region                   489 non-null    object
 15  yearFounded              451 non-null    float64
 16  stateCode                489 non-null    object
 17  collegeType              498 non-null    object
 18  carnegieClassification   498 non-null    object
 19  studentFacultyRatio      498 non-null    int64
 20  totalStudentPop          498 non-null    int64
 21  undergradPop             498 non-null    int64
 22  totalGrantAid            495 non-null    float64
 23  percentOfStudentsFinAid  495 non-null    float64
 24  percentOfStudentsGrant   495 non-null    float64
dtypes: float64(7), int64(5), object(13)
memory usage: 97.4+ KB
```

7) What do you want to learn or discover from the data?
   - To learn the impact of the attributes such as, state, longitude, latitude, totalpopulation, Grant, Financial Aid on the ranks of the Universities.

8) Break down your main question into smaller 5 specific questions.
   a) To sort the colleges based on the rankings.
   b) To sort ranking of colleges based on the median salary.
   c) To sort top organisations by student population and medianBaseSalary.
   d) To sort the ranking of colleges based on Financial Aid.
   e) To find Grant ratio per college.
   f) What is the correlation between student population and total grant aid?
   g) What is the median_Salary based on the ranking and organisation's name?
   h) What is the count of the organisation in each state?
   i) What is the count of organisations in each city?
   j) What is the distribution of universities across different campus settings?
   k) How does the median base salary vary across different states?
   l) How does the student-faculty ratio vary among different university types?

9) Depending on your goals, use exploratory data analysis (EDA) for all the questions and display the findings.

- These are the steps of data preprocessing being showcased below:

```
df.nunique()
```

```
description                498
rank                       491
organizationName           498
state                       51
studentPopulation          492
campusSetting                3
medianBaseSalary           324
longitude                  455
latitude                   455
website                    477
phoneNumber                428
city                       374
country                      1
state.1                     51
region                       4
yearFounded                175
stateCode                   51
collegeType                  2
carnegieClassification      12
studentFacultyRatio         29
totalStudentPop            492
undergradPop               489
totalGrantAid              495
percentOfStudentsFinAid     56
percentOfStudentsGrant      67
dtype: int64
```

```
df.isna().sum()
```

```
description                 0
rank                        0
organizationName            0
state                       0
studentPopulation           0
campusSetting               0
medianBaseSalary            7
longitude                  40
latitude                   40
website                    21
phoneNumber                70
city                        0
country                     0
state.1                     0
region                      9
yearFounded                47
stateCode                   9
collegeType                 0
carnegieClassification      0
studentFacultyRatio         0
totalStudentPop             0
undergradPop                0
totalGrantAid               3
percentOfStudentsFinAid     3
percentOfStudentsGrant      3
dtype: int64
```

```
df.duplicated().any()
```

```
False
```

```
df=df.dropna()
df=df.drop(["description","website","phoneNumber","stateCode",],axis=1)
print(df.shape)
```

```
(422, 21)
```

```
df.describe()
```

	rank	studentPopulation	medianBaseSalary	longitude	latitude	yearFounded	studentFacultyRatio	totalStudentPop	undergradPop	totalGra
count	498.000000	498.000000	491.000000	458.000000	458.000000	451.000000	498.000000	498.000000	498.000000	4.950000
mean	249.485944	16073.983936	116382.077393	-88.871596	39.082492	1879.944568	14.120482	16073.983936	12075.550201	2.491186
std	143.899350	16284.865007	17161.769465	18.490733	4.481500	50.805940	5.203074	16284.865007	12628.161452	1.936898
min	1.000000	421.000000	77300.000000	-157.820047	21.299373	1636.000000	3.000000	421.000000	421.000000	2.307030
25%	125.250000	3112.500000	104300.000000	-96.924510	36.105500	1851.000000	10.000000	3112.500000	2613.750000	1.219794
50%	249.500000	9850.000000	112800.000000	-84.251869	40.195954	1878.000000	13.500000	9850.000000	6844.500000	1.904380
75%	373.750000	24363.000000	125000.000000	-75.471093	42.210113	1908.500000	17.000000	24363.000000	18655.250000	3.218453
max	498.000000	102826.000000	173700.000000	85.501600	48.752350	2013.000000	49.000000	102826.000000	84202.000000	1.575583

- df2 is created by fixing some of crucial attributes of the data and by using df2, insights will be as the size of the dataset has been decreased so we know which columns are beneficial for drawing inferences.

```
df2=df.loc[:,['rank', 'organizationName','latitude','longitude','medianBaseSalary','studentPopulation','totalGrantAi
```

```
df2
```

	rank	organizationName	latitude	longitude	medianBaseSalary	studentPopulation	totalGrantAid
0	1	Massachusetts Institute of Technology	42.359006	-71.093539	173700.0	12195	35299332.0
1	2	Stanford University	37.431370	-122.168924	173500.0	20961	51328461.0
2	2	University of California, Berkeley	37.869236	-122.258393	154500.0	45878	64495611.0
3	4	Princeton University	40.349855	-74.659119	167600.0	8532	44871096.0
4	5	Columbia University	40.806515	-73.961288	148800.0	33882	44615007.0
...	...	...	...	...	...	...	...
490	491	Loyola University New Orleans	29.953690	-90.077714	102300.0	4972	26114959.0
491	492	Xavier University	39.149037	-84.476379	104900.0	8079	28294277.0
493	494	St. Joseph's College (NY)	40.690548	-73.968304	100900.0	5901	11919881.0
494	495	Moravian University	40.630303	-75.381596	109800.0	2961	12685943.0
497	498	University of Memphis	35.118453	-89.939618	90700.0	25128	27575189.0

422 rows × 7 columns

```
## Sorting the colleges based on the rankings
a=df2[['organizationName','rank']]
a=a.sort_values('rank',ascending=True)
print("The top 10 colleges in USA are: ")
print(a)
```

```
The top 10 colleges in USA are:
                          organizationName  rank
0    Massachusetts Institute of Technology     1
1                      Stanford University     2
2       University of California, Berkeley     2
3                     Princeton University     4
4                      Columbia University     5
..                                     ...   ...
490          Loyola University New Orleans   491
491                      Xavier University   492
493              St. Joseph's College (NY)   494
494                    Moravian University   495
497                   University of Memphis   498

[422 rows x 2 columns]
```

- To sort the colleges based on the rankings we get MIT to be at 1st rank and University of Memphis to be at 497th.

```
## The ranking of colleges based on the median salar
b=df2[['organizationName','medianBaseSalary']]
b=b.sort_values('medianBaseSalary',ascending=False)
b[0:10]
```

	organizationName	medianBaseSalary
0	Massachusetts Institute of Technology	173700.0
1	Stanford University	173500.0
14	Harvard University	169000.0
113	Harvey Mudd College	167800.0
3	Princeton University	167600.0
44	California Institute of Technology	164600.0
259	SUNY Maritime College	164100.0
9	University of Pennsylvania	164000.0
7	Yale University	163700.0
46	Claremont McKenna College	161700.0

- MIT has the highest Median Base Salary among all the American Colleges whereas Claremont McKenna College lies on the 10th position.

```
## Top organisations by student population and medianBaseSalary
d = df2[['organizationName', 'studentPopulation', 'medianBaseSalary']]
d = d.sort_values("medianBaseSalary", ascending = False)
d
```

	organizationName	studentPopulation	medianBaseSalary
0	Massachusetts Institute of Technology	12195	173700.0
1	Stanford University	20961	173500.0
14	Harvard University	41024	169000.0
113	Harvey Mudd College	1132	167800.0
3	Princeton University	8532	167600.0
...	...	...	...
448	John Brown University	2749	87800.0
423	University of Texas, Rio Grande Valley	41681	87300.0
485	Belmont University	9023	86300.0
394	Texas Woman's University	19733	86000.0
440	Berea College	1707	77300.0

422 rows × 3 columns

- MIT has the highest median salary based on the population of student and Berea College lies on the last in the terms of student population and the median Salary.

```
## The ranking of colleges based on Financial Aid
c = df2[['organizationName',"totalGrantAid"]]
c = c.sort_values("totalGrantAid", ascending = False)
c
```

	organizationName	totalGrantAid
124	Arizona State University, Tempe	157558319.0
59	New York University	98732499.0
167	Drexel University	96187904.0
249	St. John's University (NY)	89415786.0
78	Northeastern University	84882594.0
...	...	...
428	Principia College	1853786.0
259	SUNY Maritime College	1662985.0
404	New College of Florida	1583167.0
354	Montana Tech of the University of Montana	1522295.0
266	California State University Maritime Academy	680549.0

422 rows × 2 columns

- Arizona State University has the highest Grant_Aid in American University, as i was going through the GrantAid and the total student population, I came up with a new column named “Aid_Ratio”.
- It is the ratio of = totalGrantAid / studentPopulation
- It gives us the ratio of the grant provided by the oragnisation by the population, which is the ratio of grant each student can avail in that organisation.

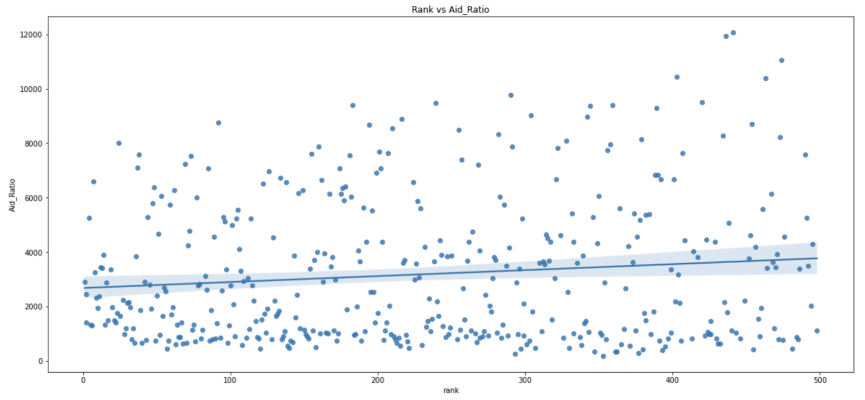
```
## Grant ratio per college
df["Aid_Ratio"] = df['totalGrantAid']/ df['studentPopulation']
selected_columns = ['organizationName', 'Aid_Ratio']
selected_df = df[selected_columns]
g = selected_df.sort_values("Aid_Ratio", ascending = False)
g[0:10]
```

	organizationName	Aid_Ratio
440	Berea College	12079.929701
435	Albion College	11938.506667
473	Ursinus College	11051.087802
402	Allegheny College	10440.402062
462	Wheaton College (MA)	10376.988379
289	College of Wooster	9776.718515
419	Susquehanna University	9515.663969
238	St. Lawrence University	9481.839859
358	Austin College	9410.593514
182	Kalamazoo College	9400.826490

- Berea College which has the least median based salary provides with the highest ratio of grant per student.

```
## Line Graph
plt.figure(figsize=(20, 9))
sns.regplot(x= 'rank', y='Aid_Ratio', data=df)
plt.title('Rank vs Aid_Ratio')
```

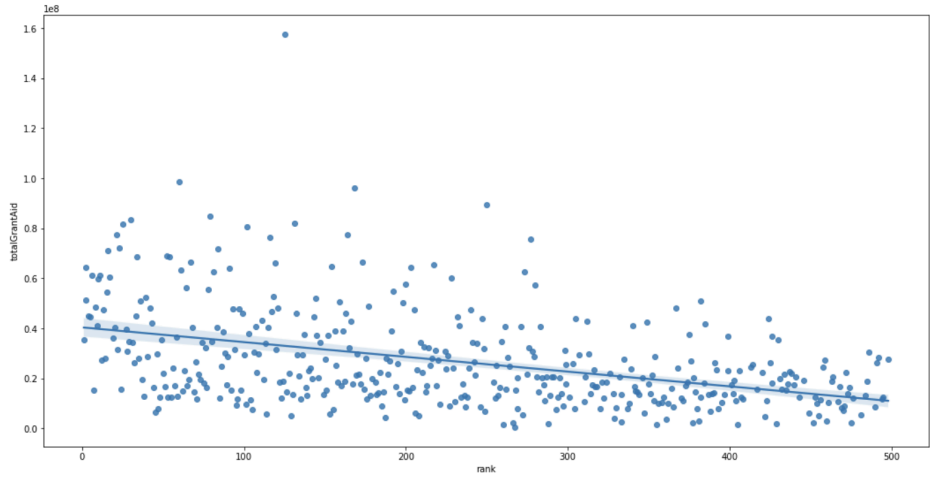
```
Text(0.5, 1.0, 'Rank vs Aid_Ratio')
```



- This shows that as the rank increases the Aid_Ratio also increases.

```
## Regresion Graph to see relationship between the rank and the totalGrantAid.
plt.figure(figsize=(18, 9))
sns.regplot(x= 'rank', y='totalGrantAid', data=df2)
```

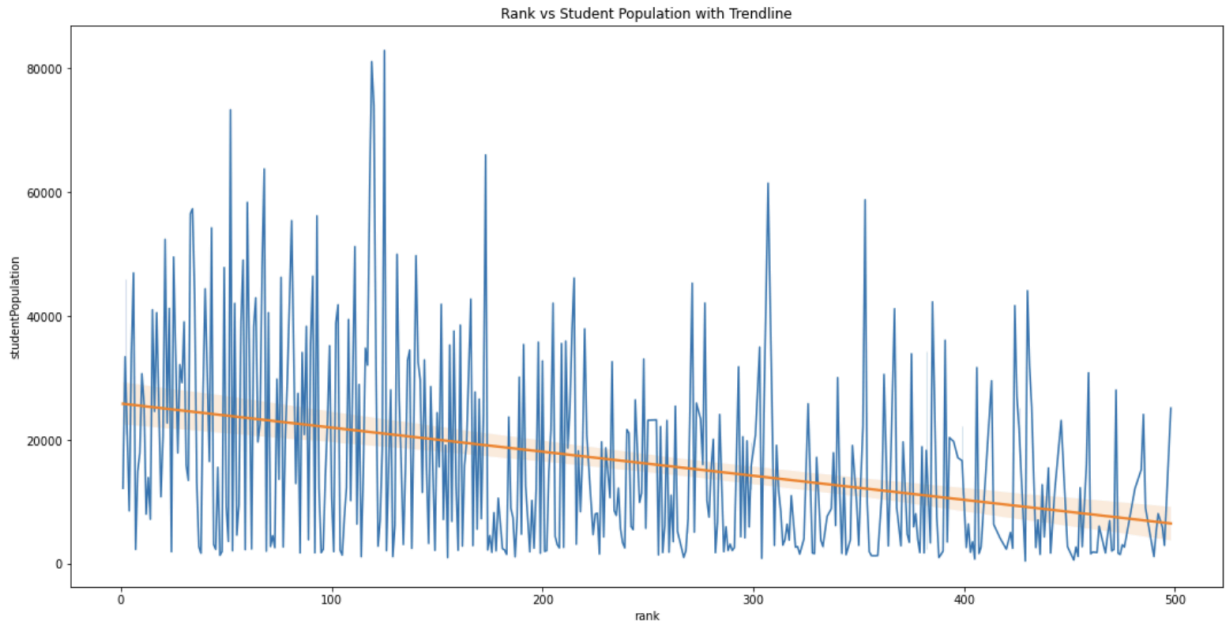
```
<AxesSubplot:xlabel='rank', ylabel='totalGrantAid'>
```



- This plot shows that as the Rank of the colleges increases the grantaid provided also decreases.

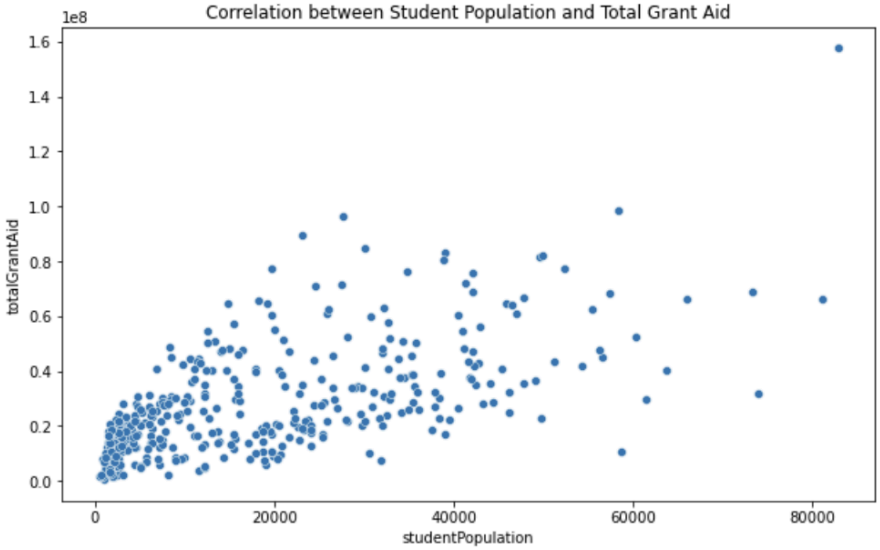
```
## TO SEE THE STUDENT POPULATION IN COLLEGES WITH RESPECT TO RANKS
plt.figure(figsize=(18, 9))
plt.title('Rank vs Student Population with Trendline')
sns.lineplot(x='rank', y='studentPopulation', data=df)
sns.regplot(x='rank', y='studentPopulation', data=df, scatter=False)
```

```
<AxesSubplot:title={'center':'Rank vs Student Population with Trendline'}, xlabel='rank', ylabel='studentPopulatio
n'>
```



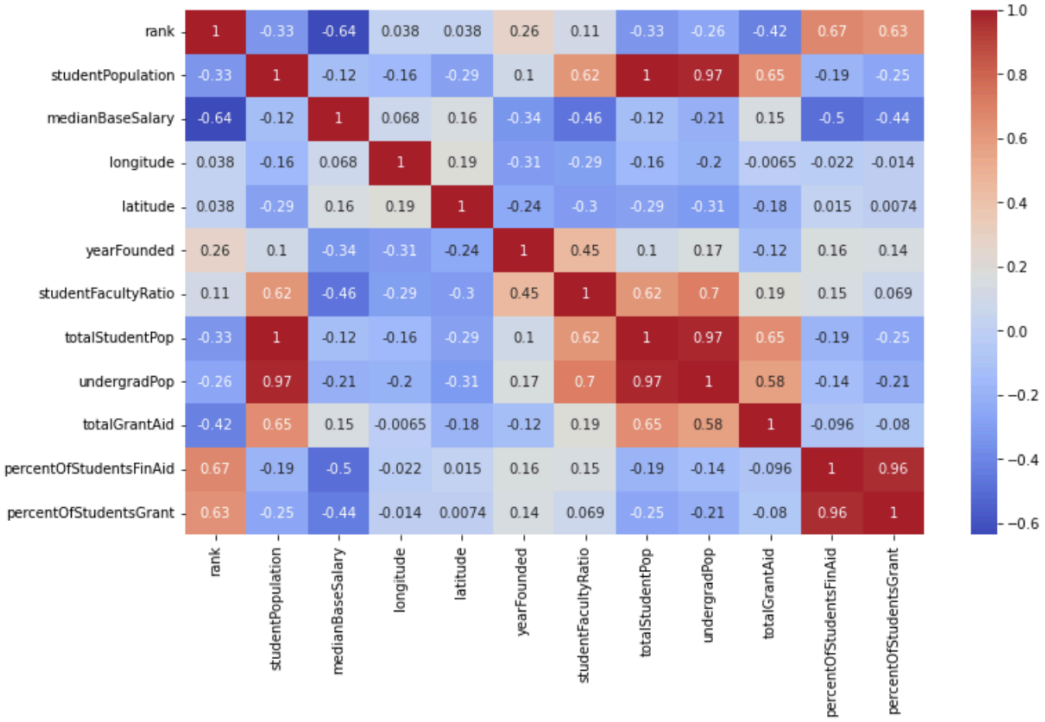
- The population of the students also tend to decrease as the ranking of the college decreases, the students tend to get enrolled in the college with better ranks.

```
## What is the correlation between student population and total grant aid?
plt.figure(figsize=(10, 6))
sns.scatterplot(x='studentPopulation', y='totalGrantAid', data=df)
plt.title('Correlation between Student Population and Total Grant Aid')
plt.show()
```



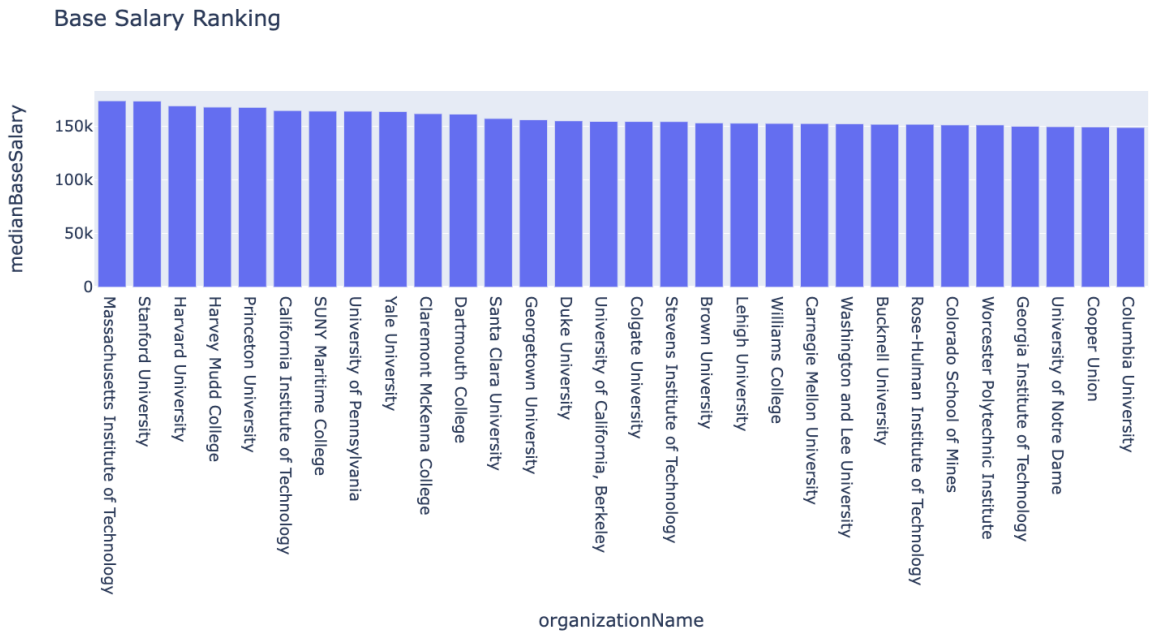
- The datapoints are dense in the region of 0-20000 students population, which mean as the population increases the totalGrantAid decreases for the colleges and maybe others are the outliers.

```
plt.figure(figsize=(12,7))
sns.heatmap(df.corr(), cmap='coolwarm', annot=True)
plt.show()
```



- This is a heatmap, it shows that there is strong positive correlation between:
- There is negative correlation between rank and medianBasesalary which tells that as the rank of a university gets better the university gives a higher salary.
- We see as the rank gets better the amount of financial aid decreases.
- From the correaltipn heat map we can also confirm that longitude and latitude play no role in effecting any other variable.

```
## The median_Salary based on the ranking and organisation's name
fig = px.bar(d[:30], x='organizationName', y='medianBaseSalary',title="Base Salary Ranking")
fig.show()
```



- The average salary decreases with the ranks of the college, being MIT at the top.

```
## Grouping the states to get the count of the organisation in each state.
state_counts = df['organizationName'].groupby(df['state']).count().sort_values(ascending=False)
top_10_states = state_counts.head(10)
```

```
top_10_states
```

```
state
NY    51
CA    45
PA    30
TX    23
MA    22
OH    15
IL    15
IN    12
VA    11
MN    11
Name: organizationName, dtype: int64
```

- This tells us the number of colleges/organisations in top 10 states of USA.

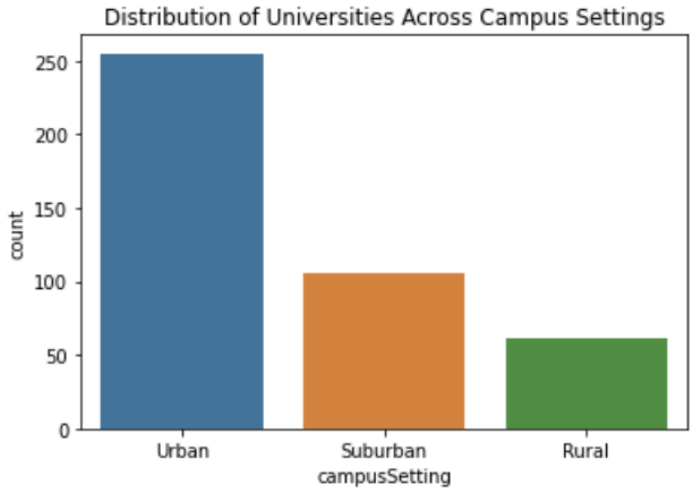
```
## Grouping based on the cities to get the total count of colleges in the cities
cities = df.groupby('city')['organizationName'].count().sort_values(ascending=False)
```

```
cities[0:10]
```

```
city
New York        12
Washington       6
Philadelphia     5
Claremont        5
Los Angeles      5
Chicago          5
Portland         4
Boston           4
Worcester        3
Pittsburgh       3
Name: organizationName, dtype: int64
```

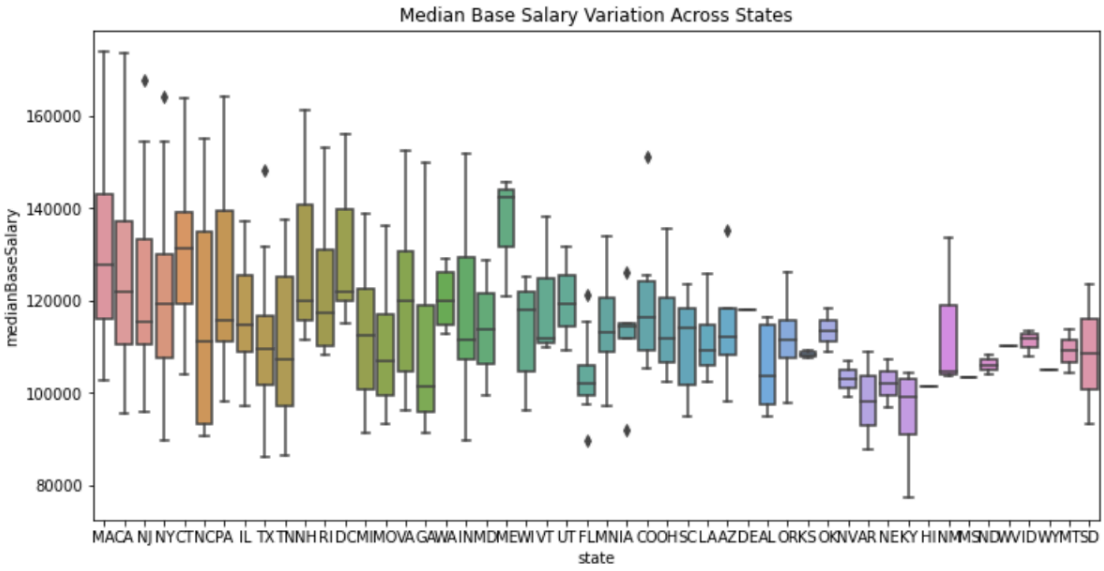
- This gives us the count of the colleges/organisations situated in top 10 cities of USA.

```
## What is the distribution of universities across different campus settings?
sns.countplot(x='campusSetting', data=df)
plt.title('Distribution of Universities Across Campus Settings')
plt.show()
```



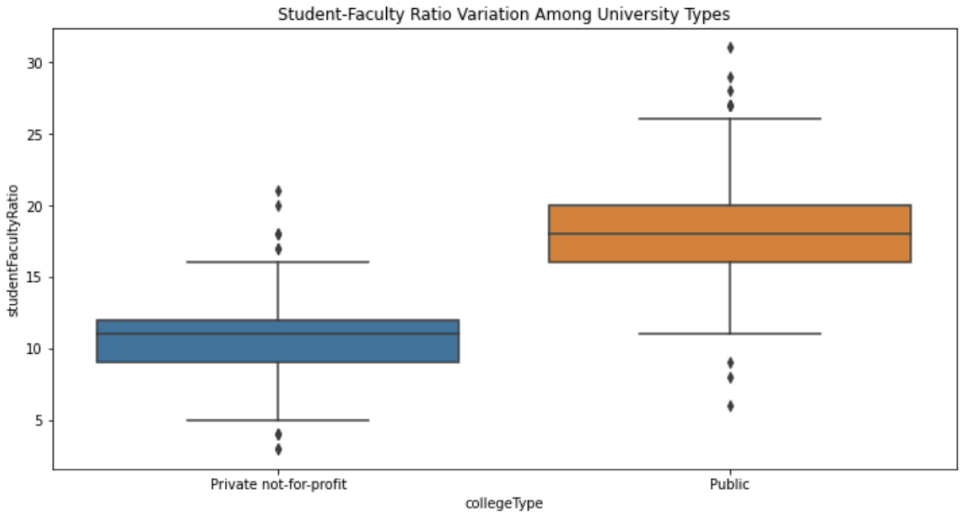
- There are three categories of campus settings available in the USA, Urban have most of the educational organisations while Rural having the least number of organisations.

```
##  How does the median base salary vary across different states?
plt.figure(figsize=(12, 6))
sns.boxplot(x='state', y='medianBaseSalary', data=df)
plt.title('Median Base Salary Variation Across States')
plt.show()
```



- This plot showcases the distribution of the median base salary across the states, there are some states which have more median base salary(outliers). MA state having the highest along with ME.

```
## How does the student-faculty ratio vary among different university types?
plt.figure(figsize=(12, 6))
sns.boxplot(x='collegeType', y='studentFacultyRatio', data=df)
plt.title('Student-Faculty Ratio Variation Among University Types')
plt.show()
```



-

- According to the plot we can say the student faculty ratio is better in Public universities rather than Private non-Profit univerities