

In [1]: `import pandas as pd`

In [2]: `import matplotlib`

In [3]: `import numpy`

In [4]: `import seaborn`

In [5]: `df = pd.read_csv('/home/ubuntu/Desktop/dataset/qs-world-university-rankings-2017-to-2022-V2.csv')`

In [6]: `df.head()`*#starting data*

Out[6]:

	university	year	rank_display	score	link	country	city	region	logc
0	Massachusetts Institute of Technology (MIT)	2017	1	100.0	https://www.topuniversities.com/universities/m...	United States	Cambridge	North America	https://www.topuniversities.com/sites/default/..
1	Stanford University	2017	2	98.7	https://www.topuniversities.com/universities/s...	United States	Stanford	North America	https://www.topuniversities.com/sites/default/..
2	Harvard University	2017	3	98.3	https://www.topuniversities.com/universities/h...	United States	Cambridge	North America	https://www.topuniversities.com/sites/default/..
3	University of Cambridge	2017	4	97.2	https://www.topuniversities.com/universities/u...	United Kingdom	Cambridge	Europe	https://www.topuniversities.com/sites/default/..
4	California Institute of Technology (Caltech)	2017	5	96.9	https://www.topuniversities.com/universities/c...	United States	Pasadena	North America	https://www.topuniversities.com/sites/default/..

In [7]: `df.info()`*#tells columns datatype*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6482 entries, 0 to 6481
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype
---  -
0   university             6482 non-null   object
1   year                   6482 non-null   int64
2   rank_display           6414 non-null   object
3   score                  2820 non-null   float64
4   link                   6482 non-null   object
5   country                6482 non-null   object
6   city                   6304 non-null   object
7   region                 6482 non-null   object
8   logo                   6482 non-null   object
9   type                   6470 non-null   object
10  research_output         6480 non-null   object
11  student_faculty_ratio   6407 non-null   float64
12  international_students  6318 non-null   object
13  size                   6480 non-null   object
14  faculty_count           6404 non-null   object
dtypes: float64(2), int64(1), object(12)
memory usage: 759.7+ KB
```

In [8]: `df.describe()`

Out[8]:

	year	score	student_faculty_ratio
count	6482.000000	2820.000000	6407.000000
mean	2019.693613	46.595532	13.264554
std	1.716683	18.813110	6.604294
min	2017.000000	23.500000	1.000000
25%	2018.000000	31.800000	9.000000
50%	2020.000000	40.600000	12.000000
75%	2021.000000	58.025000	17.000000
max	2022.000000	100.000000	67.000000

In [9]: `df.tail(10)`*#show last 10 values*

Out[9]:

	university	year	rank_display	score	link	country	city	region	log
6472	University of Sarajevo	2022	1201	NaN	https://www.topuniversities.com/universities/u...	Bosnia and Herzegovina	Sarajevo	Europe	https://www.topuniversities.com/sites/default/..
6473	University of Split	2022	1201	NaN	https://www.topuniversities.com/universities/u...	Croatia	Split	Europe	https://www.topuniversities.com/sites/default/..
6474	Università degli studi di Bergamo	2022	1201	NaN	https://www.topuniversities.com/universities/u...	Italy	Bergamo	Europe	https://www.topuniversities.com/sites/default/..
6475	Université Mohammed V de Rabat	2022	1201	NaN	https://www.topuniversities.com/universities/u...	Morocco	Rabat	Africa	https://www.topuniversities.com/sites/default/..
6476	Université de Caen Normandie	2022	1201	NaN	https://www.topuniversities.com/universities/u...	France	Caen	Europe	https://www.topuniversities.com/sites/default/..
6477	Université de Tunis	2022	1201	NaN	https://www.topuniversities.com/universities/u...	Tunisia	Tunis	Africa	https://www.topuniversities.com/sites/default/..
6478	Université de Tunis El Manar	2022	1201	NaN	https://www.topuniversities.com/universities/u...	Tunisia	Tunis	Africa	https://www.topuniversities.com/sites/default/..
6479	Yarmouk University	2022	1201	NaN	https://www.topuniversities.com/universities/y...	Jordan	Irbid	Asia	https://www.topuniversities.com/sites/default/..
6480	Yildiz Technical University	2022	1201	NaN	https://www.topuniversities.com/universities/y...	Turkey	Istanbul	Asia	https://www.topuniversities.com/sites/default/..
6481	Zagazig University	2022	1201	NaN	https://www.topuniversities.com/universities/z...	Egypt	Zagazig	Africa	https://www.topuniversities.com/sites/default/..

In [10]: `df = df.dropna()`*#to drop null*

In [11]: `df['year'] = df['year'].astype('string')`*#to change datatype of column*

In [12]: `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2716 entries, 0 to 5682
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype
---  -
0   university             2716 non-null   object
1   year                   2716 non-null   string
2   rank_display           2716 non-null   object
3   score                  2716 non-null   float64
4   link                   2716 non-null   object
5   country                2716 non-null   object
6   city                   2716 non-null   object
7   region                 2716 non-null   object
8   logo                   2716 non-null   object
9   type                   2716 non-null   object
10  research_output         2716 non-null   object
11  student_faculty_ratio   2716 non-null   float64
12  international_students  2716 non-null   object
13  size                   2716 non-null   object
14  faculty_count           2716 non-null   object
dtypes: float64(2), object(12), string(1)
memory usage: 339.5+ KB
```

In [13]: `pd.get_dummies(df['region'])`*#to change categorical data to quantitative data*

Out[13]:

	Africa	Asia	Europe	Latin America	North America	Oceania
0	0	0	0	0	1	0
1	0	0	0	0	1	0
2	0	0	0	0	1	0
3	0	0	1	0	0	0
4	0	0	0	0	1	0
...
5678	0	1	0	0	0	0
5679	0	0	0	0	1	0
5680	0	1	0	0	0	0
5681	0	0	1	0	0	0
5682	0	0	0	0	1	0

2716 rows × 6 columns

In []: