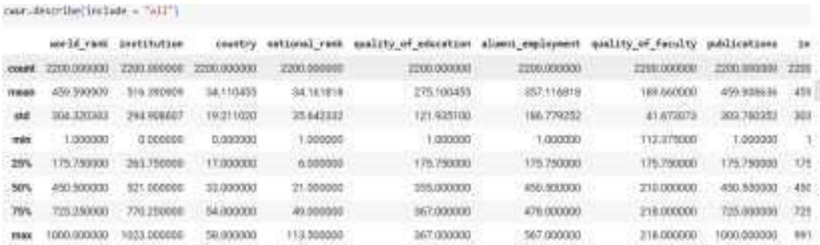
Data Collection and Preprocessing Phase

Date	06 July 2024
Team ID	739824
Project Title	SmartLender – Envisioning Success: Predicting University Scores With Machine Learning
Maximum Marks	6 Marks

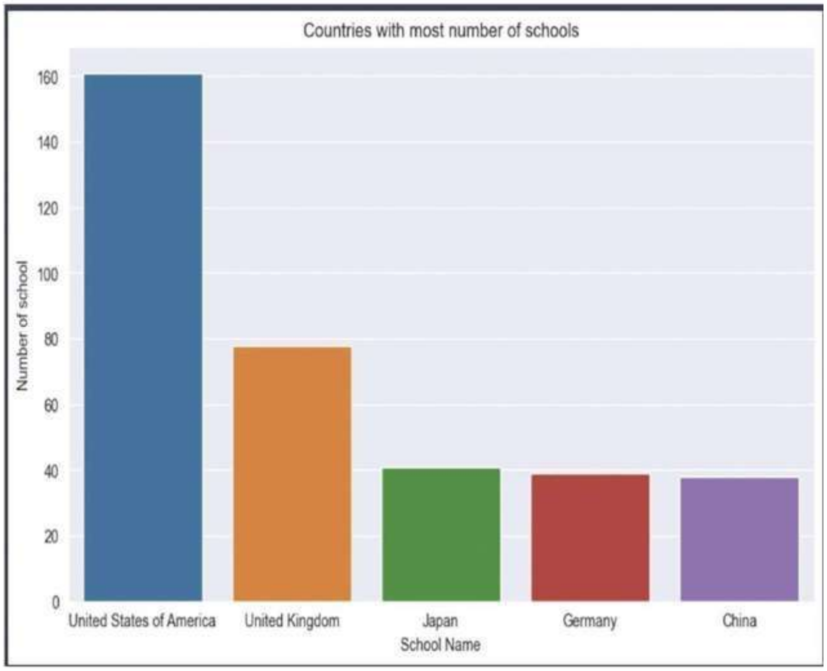
Data Exploration and Preprocessing Report

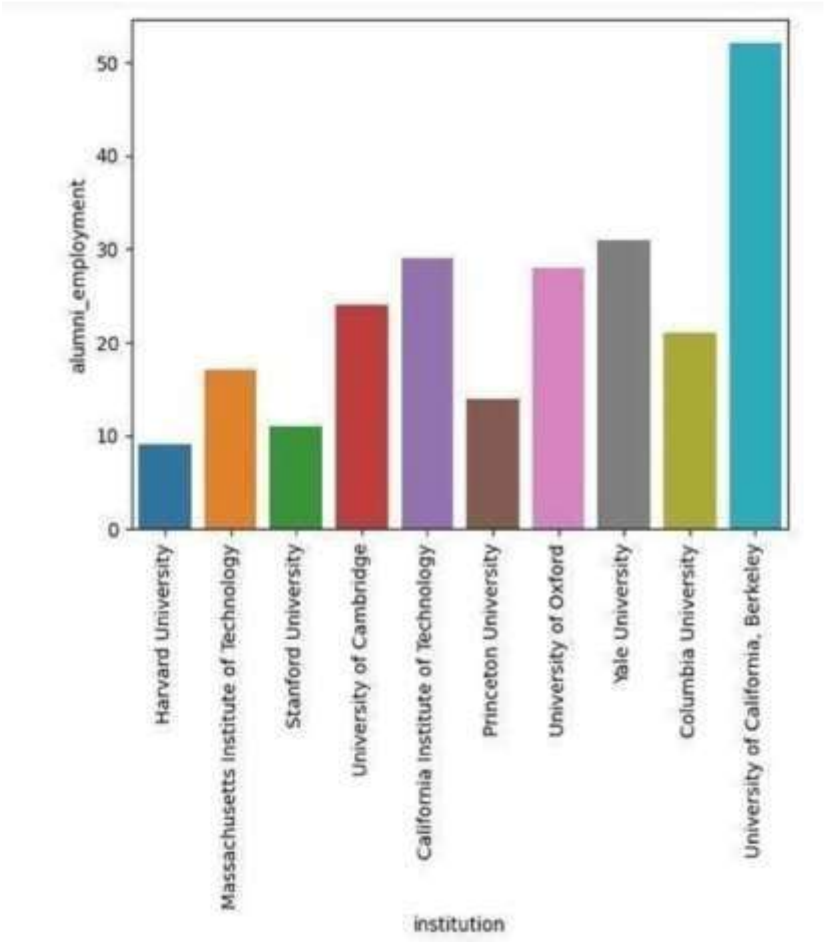
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	**Description**
Data Overview	

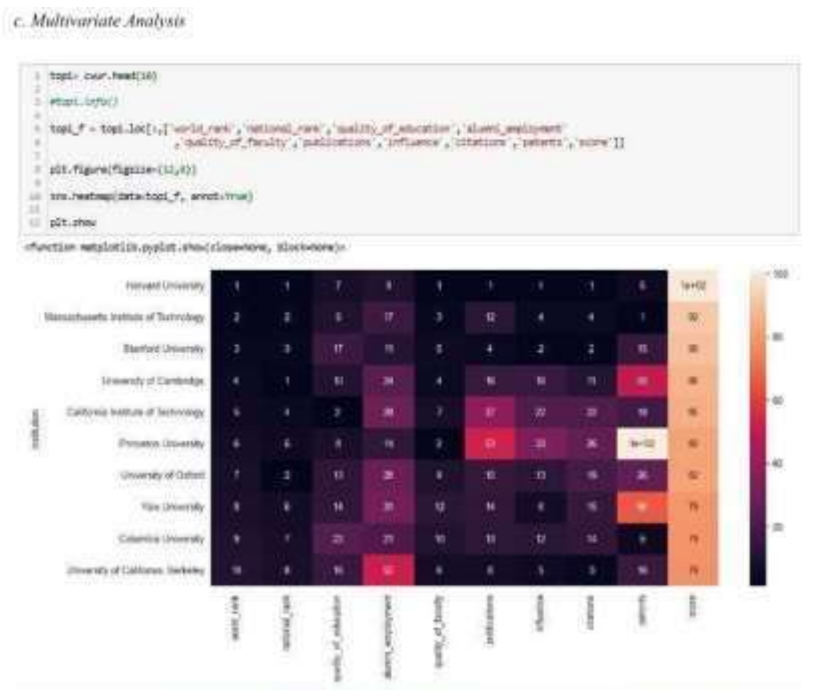
```
raw.describe(include = "all")
```

```
        world_rank  institution      country  national_rank  quality_of_education  alumni_employment  quality_of_faculty  publications   in
count  2200.000000  2200.000000  2200.000000    2200.000000           2200.000000        2200.000000         2200.000000   2200.000000  220
mean    459.590909   516.380909    34.110455      34.161818            275.100455         357.116818          189.660000    459.908636  459
std     304.320303   294.908607    19.211020      25.642332            121.935100         186.779252           41.673073    303.760353  303
min       1.000000     0.000000     0.000000       1.000000              1.000000           1.000000          112.375000      1.000000    1
25%     175.750000   263.750000    17.000000       6.000000            175.750000         175.750000          175.750000    175.750000  175
50%     450.500000   521.000000    33.000000      21.000000            355.000000         450.500000          210.000000    450.500000  450
75%     725.250000   770.250000    54.000000      49.000000            367.000000         478.000000          218.000000    725.000000  725
max    1000.000000  1023.000000    58.000000     113.500000            367.000000         567.000000          218.000000   1000.000000  991
```

Univariate Analysis	*a.Univariate Analysis* 

Bivariate Analysis	

Multivariate Analysis



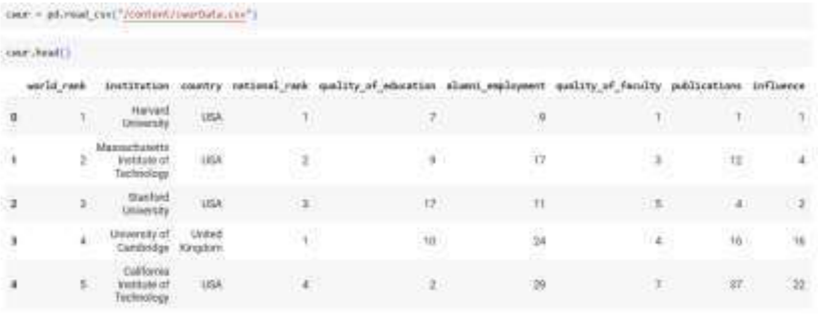
c. Multivariate Analysis

```
topi= cwur.head(10)

#topi.info()

topi_f = topi.loc[:,['world_rank','national_rank','quality_of_education','alumni_employment'
                     ,'quality_of_faculty','publications','influence','citations','patents','score']]

plt.figure(figsize=(12,8))

sns.heatmap(data=topi_f, annot=True)

plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

Data Preprocessing Code Screenshots

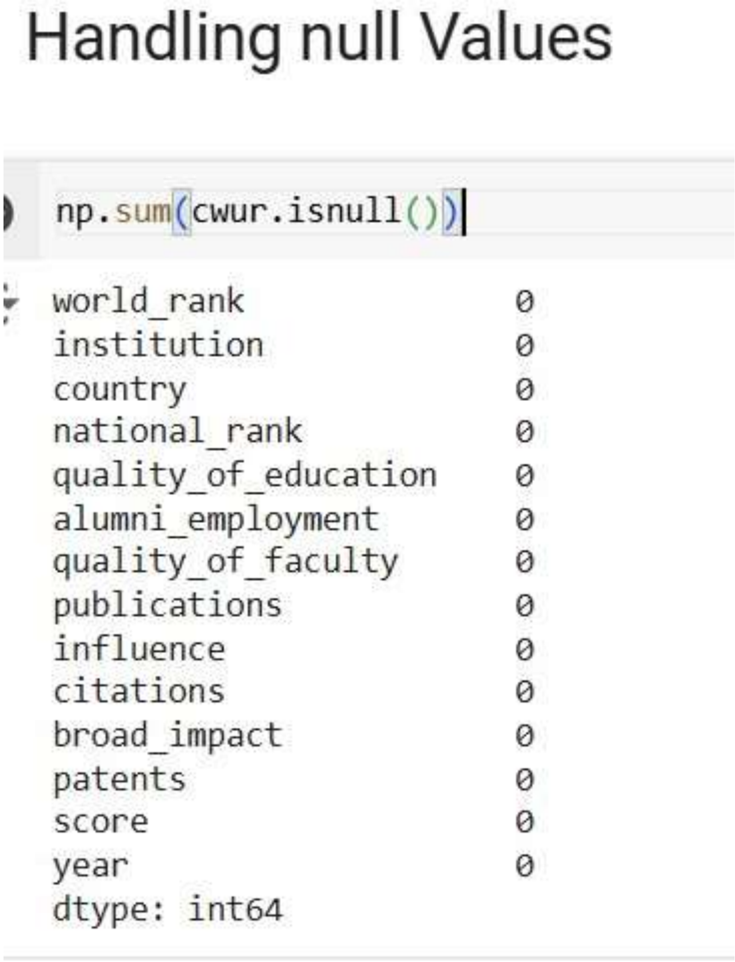
Loading Data

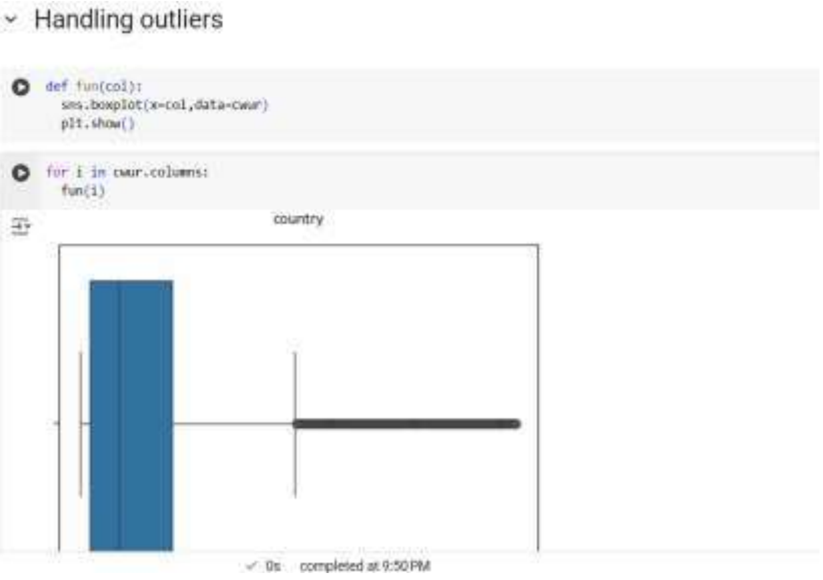


```
cwur = pd.read_csv("/content/cwurData.csv")
```

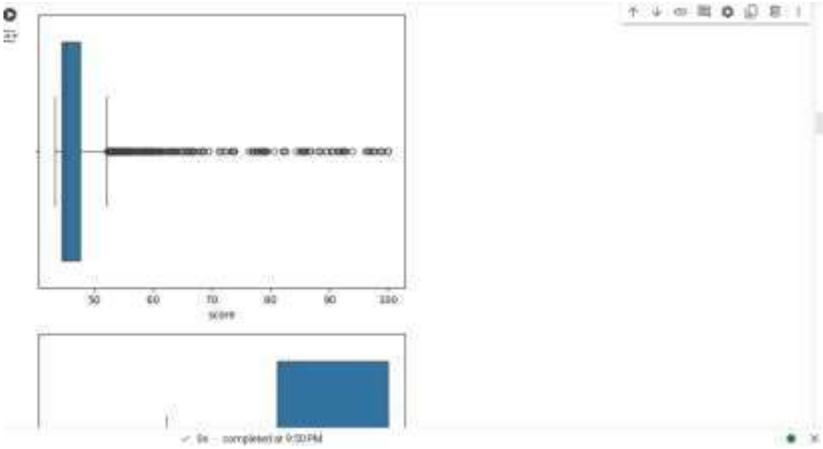
```
cwur.head()
```

	world_rank	institution	country	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	influence
0	1	Harvard University	USA	1	7	9	1	1	1
1	2	Massachusetts Institute of Technology	USA	2	9	17	3	12	4
2	3	Stanford University	USA	3	17	11	5	4	2
3	4	University of Cambridge	United Kingdom	1	10	24	4	16	16
4	5	California Institute of Technology	USA	4	2	29	7	37	22

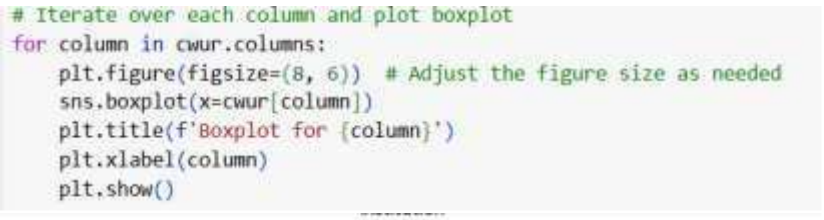
Handling Null Values	
	Handling null Values
	`np.sum(cwur.isnull())`
	`world_rank              0`
	`institution             0`
	`country                 0`
	`national_rank           0`
	`quality_of_education    0`
	`alumni_employment       0`
	`quality_of_faculty      0`
	`publications            0`
	`influence               0`
	`citations               0`
	`broad_impact            0`
	`patents                 0`
	`score                   0`
	`year                    0`
	`dtype: int64`

Viewing outliers	
	Handling outliers
	`def fun(col):`
	`  sns.boxplot(x=col,data=cwur)`
	`  plt.show()`
	`for i in cwur.columns:`
	`  fun(i)`
	
	✓ 0s completed at 9:50 PM

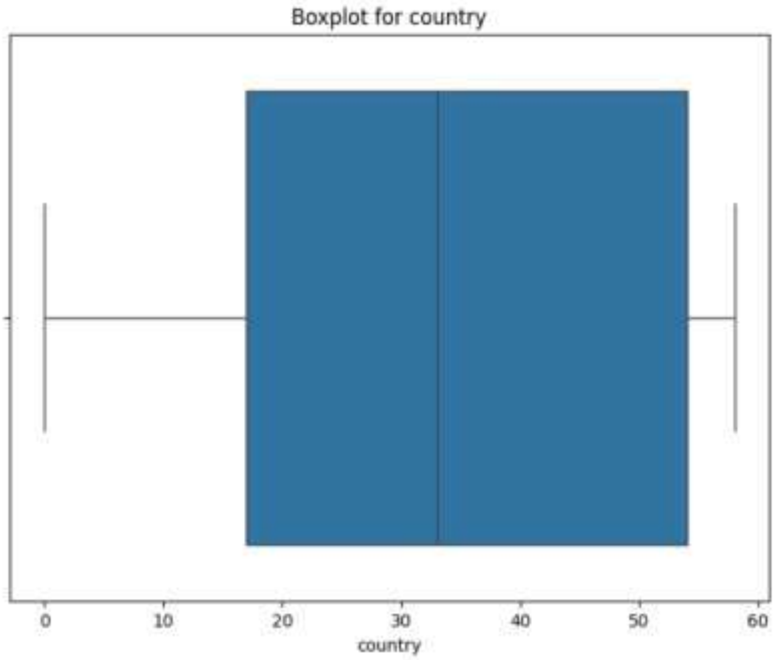
<table>
<tr>
<td></td>
<td>



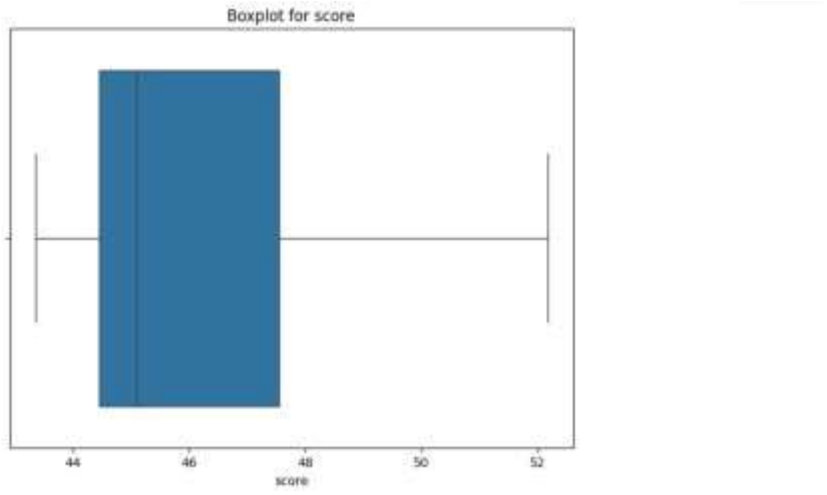
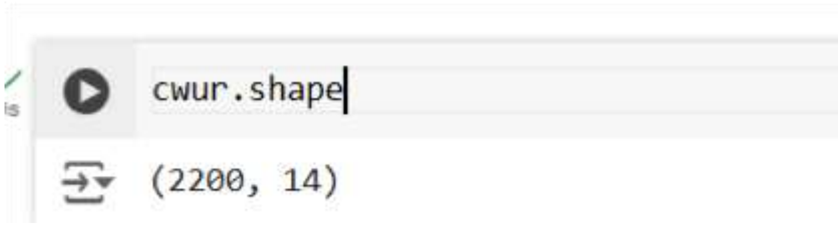
</td>
</tr>
<tr>
<td>Handling outliers</td>
<td>



```
# Iterate over each column and plot boxplot
for column in cwur.columns:
    plt.figure(figsize=(8, 6))  # Adjust the figure size as needed
    sns.boxplot(x=cwur[column])
    plt.title(f'Boxplot for {column}')
    plt.xlabel(column)
    plt.show()
```



</td>
</tr>
</table>

	
Saved Processed Data	 `cwur.shape` `(2200, 14)`