

Data Collection and Preprocessing Phase

Date	06 July 2024
Team ID	739899
Project Title	SmartLender – Envisioning Success: Predicting University Scores With Machine Learning
Maximum Marks	6 Marks

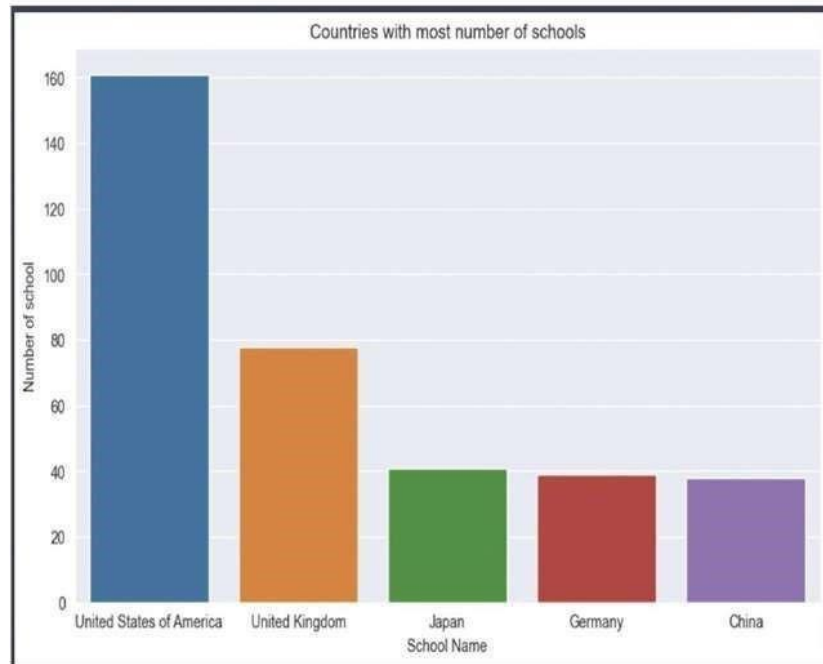
Data Exploration and Preprocessing Report

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.

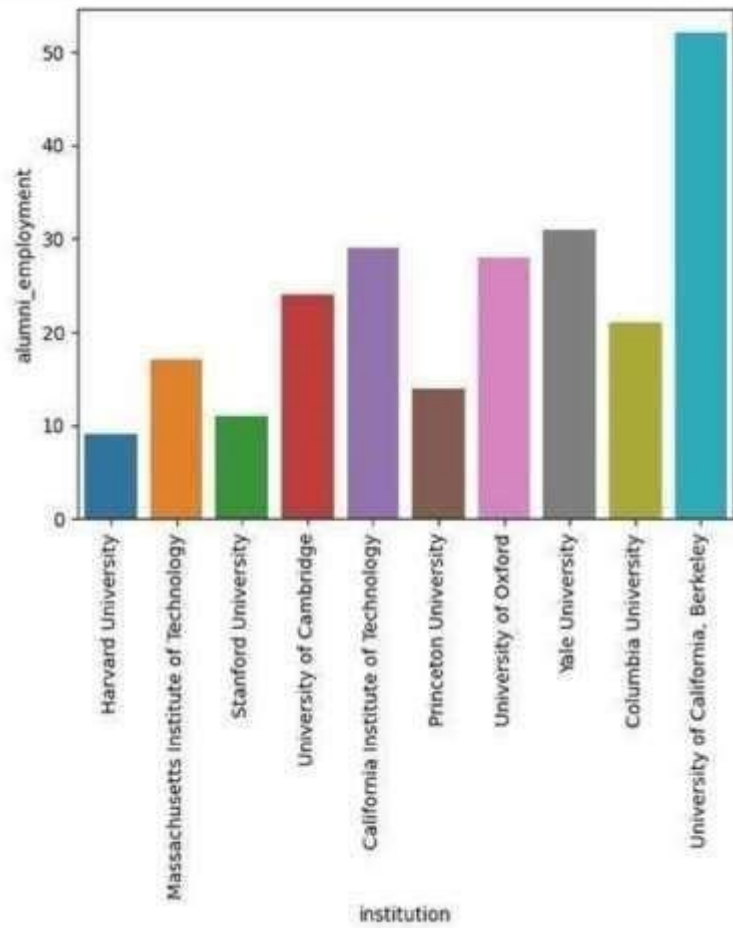
Section	Description																																																																																										
Data Overview	<div><pre>car.describe(include = "all")</pre></div> <table><thead><tr><th></th><th>world_rank</th><th>institution</th><th>costly</th><th>national_rank</th><th>quality_of_education</th><th>alumni_employment</th><th>quality_of_faculty</th><th>publications</th><th>20</th></tr></thead><tbody><tr><td>count</td><td>2200.000000</td><td>2200.000000</td><td>2200.000000</td><td>2200.000000</td><td>2200.000000</td><td>2200.000000</td><td>2200.000000</td><td>2200.000000</td><td>2200</td></tr><tr><td>mean</td><td>459.399909</td><td>519.390909</td><td>34.110435</td><td>34.181818</td><td>275.100455</td><td>857.116818</td><td>168.660000</td><td>459.828888</td><td>478</td></tr><tr><td>std</td><td>304.330383</td><td>294.608607</td><td>19.311020</td><td>25.642332</td><td>121.928100</td><td>166.779252</td><td>41.673073</td><td>303.780281</td><td>369</td></tr><tr><td>min</td><td>1.000000</td><td>0.000000</td><td>0.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>112.317000</td><td>1.000000</td><td>1</td></tr><tr><td>25%</td><td>175.750000</td><td>241.750000</td><td>11.000000</td><td>6.000000</td><td>175.750000</td><td>175.750000</td><td>175.750000</td><td>175.750000</td><td>175</td></tr><tr><td>50%</td><td>459.399909</td><td>519.390909</td><td>31.000000</td><td>21.000000</td><td>255.000000</td><td>856.900000</td><td>218.000000</td><td>459.828888</td><td>478</td></tr><tr><td>75%</td><td>725.250000</td><td>775.250000</td><td>54.000000</td><td>49.000000</td><td>367.000000</td><td>478.000000</td><td>218.000000</td><td>725.000000</td><td>725</td></tr><tr><td>max</td><td>1000.000000</td><td>1033.000000</td><td>50.000000</td><td>113.000000</td><td>367.000000</td><td>567.000000</td><td>218.000000</td><td>1000.000000</td><td>881</td></tr></tbody></table>		world_rank	institution	costly	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	20	count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200	mean	459.399909	519.390909	34.110435	34.181818	275.100455	857.116818	168.660000	459.828888	478	std	304.330383	294.608607	19.311020	25.642332	121.928100	166.779252	41.673073	303.780281	369	min	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	112.317000	1.000000	1	25%	175.750000	241.750000	11.000000	6.000000	175.750000	175.750000	175.750000	175.750000	175	50%	459.399909	519.390909	31.000000	21.000000	255.000000	856.900000	218.000000	459.828888	478	75%	725.250000	775.250000	54.000000	49.000000	367.000000	478.000000	218.000000	725.000000	725	max	1000.000000	1033.000000	50.000000	113.000000	367.000000	567.000000	218.000000	1000.000000	881
		world_rank	institution	costly	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	20																																																																																	
	count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200																																																																																	
	mean	459.399909	519.390909	34.110435	34.181818	275.100455	857.116818	168.660000	459.828888	478																																																																																	
	std	304.330383	294.608607	19.311020	25.642332	121.928100	166.779252	41.673073	303.780281	369																																																																																	
	min	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	112.317000	1.000000	1																																																																																	
	25%	175.750000	241.750000	11.000000	6.000000	175.750000	175.750000	175.750000	175.750000	175																																																																																	
	50%	459.399909	519.390909	31.000000	21.000000	255.000000	856.900000	218.000000	459.828888	478																																																																																	
	75%	725.250000	775.250000	54.000000	49.000000	367.000000	478.000000	218.000000	725.000000	725																																																																																	
	max	1000.000000	1033.000000	50.000000	113.000000	367.000000	567.000000	218.000000	1000.000000	881																																																																																	

Univariate Analysis

a. Univariate Analysis



Bivariate Analysis

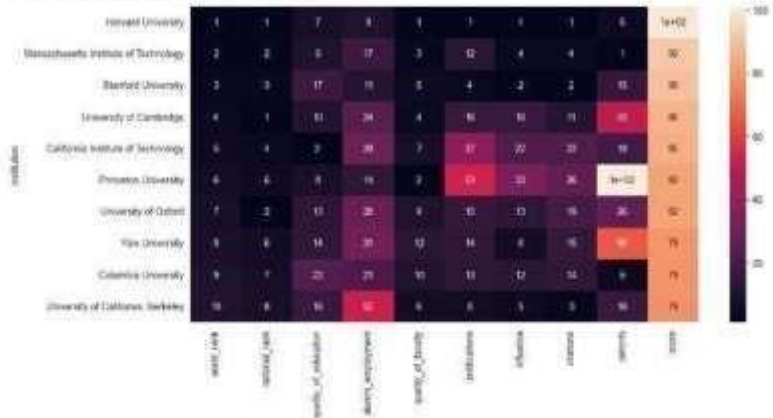


Multivariate Analysis

c. Multivariate Analysis

```
1 topic_csr.head(10)
2
3 etest.info()
4
5 topic_F = topic_model['world_rank', 'national_rank', 'quality_of_education', 'alumni_employment',
6                       'quality_of_family', 'publications', 'influence', 'citations', 'patents', 'score']
7
8 plt.figure(figsize=(12,8))
9
10 sns.heatmap(data=topic_F, annot=True)
11
12 plt.show
```

function sns.heatmap, pyplot.show, imshow, imshowc



Data Preprocessing Code Screenshots

Loading Data

```
csr = pd.read_csv('content/lowData.csv')
```

```
csr.head()
```

	world_rank	institution	country	national_rank	quality_of_education	alumni_employment	quality_of_family	publications	influence
0	1	Harvard University	USA	1	7	3	1	1	1
1	2	Massachusetts Institute of Technology	USA	2	5	17	3	12	4
2	3	Stanford University	USA	3	17	11	5	4	2
3	4	University of Cambridge	United Kingdom	1	11	24	4	16	16
4	5	California Institute of Technology	USA	4	2	29	7	27	20

Handling Null Values

Handling null Values

```
np.sum(cwur.isnull())
```

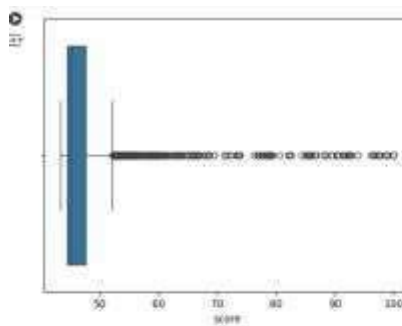
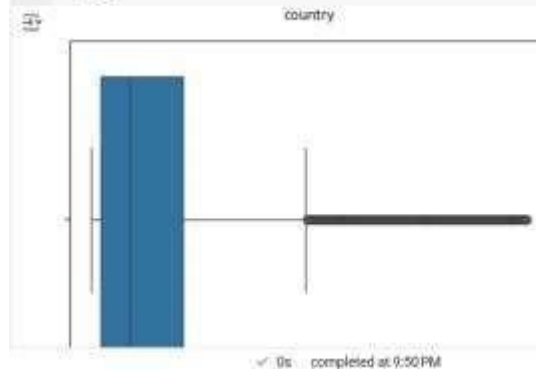
```
world_rank      0
institution      0
country         0
national_rank    0
quality_of_education  0
alumni_employment  0
quality_of_faculty  0
publications    0
influence        0
citations        0
broad_impact     0
patents          0
score            0
year            0
dtype: int64
```

Viewing outliers

Handling outliers

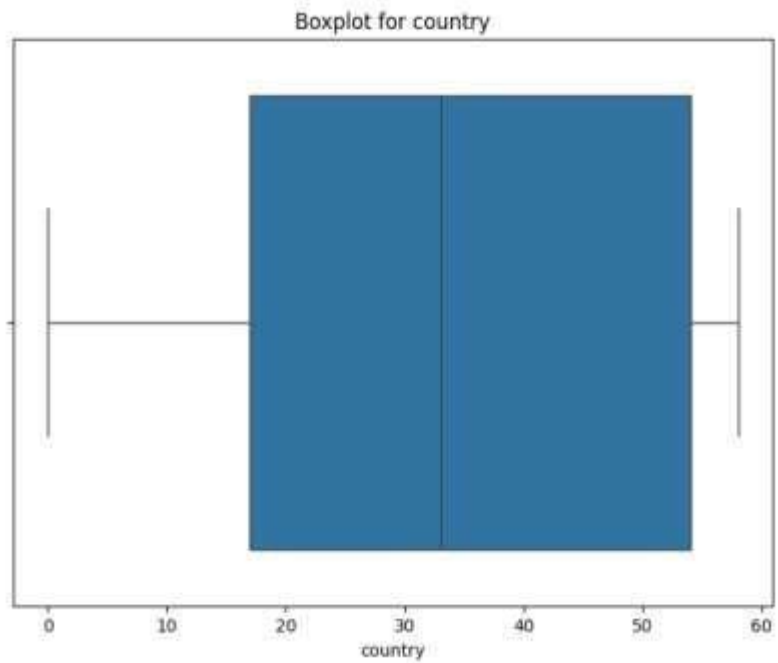
```
def fun(col):
    sns.boxplot(x=col,data=cwar)
    plt.show()
```

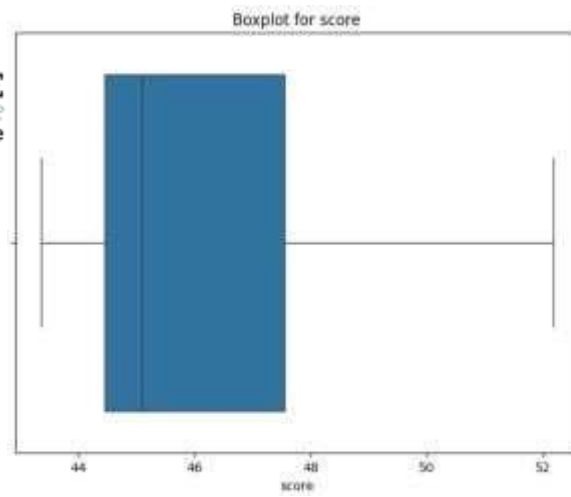
```
for i in cwar.columns:
    fun(i)
```



Handling outliers

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





Saved Processed Data

 `cwur.shape`

 `(2200, 14)`