

Data Collection and Preprocessing Phase

Date	19 April 2024
Team ID	Team-738178
Project Title	Envisioning Success : Predicting University Scores With Machine Learning
Maximum Marks	6 Marks

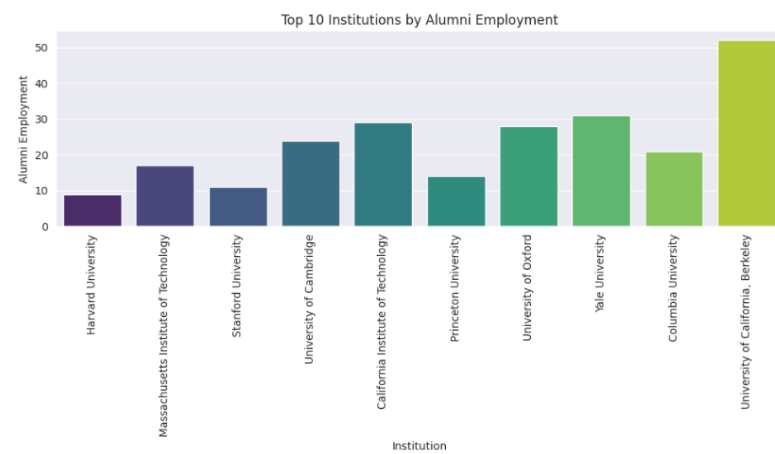
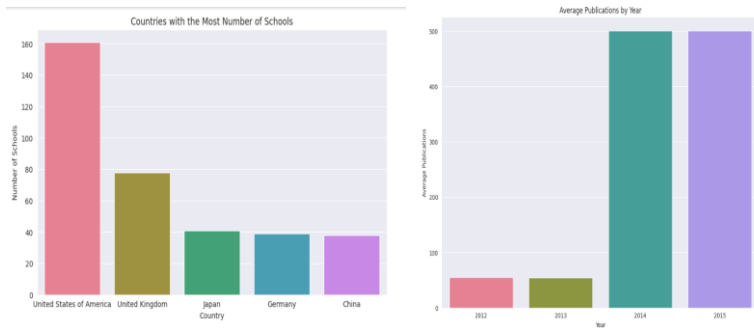
Data Exploration and Preprocessing Report

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.

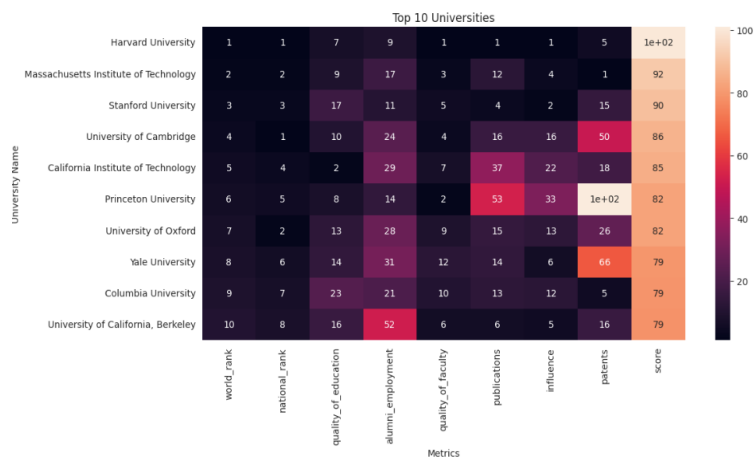
Section	Description																																																																																																																							
Data Overview	<u>Dimension:</u> 2200 rows×14 columns																																																																																																																							
	<u>Descriptive statistics:</u>																																																																																																																							
	<pre>snc_df.describe(include='all')</pre>																																																																																																																							
	<table><thead><tr><th></th><th>school_name</th><th>country</th></tr></thead><tbody><tr><td>count</td><td>818</td><td>818</td></tr><tr><td>unique</td><td>818</td><td>70</td></tr><tr><td>top</td><td>Harvard University</td><td>United States of America</td></tr><tr><td>freq</td><td>1</td><td>161</td></tr></tbody></table>		school_name	country	count	818	818	unique	818	70	top	Harvard University	United States of America	freq	1	161																																																																																																								
		school_name	country																																																																																																																					
	count	818	818																																																																																																																					
	unique	818	70																																																																																																																					
	top	Harvard University	United States of America																																																																																																																					
	freq	1	161																																																																																																																					
	<pre>cour.describe(include='all')</pre>																																																																																																																							
<table><thead><tr><th></th><th>world_rank</th><th>institution</th><th>country</th><th>national_rank</th><th>quality_of_education</th><th>alumi_employment</th><th>quality_of_faculty</th><th>publications</th><th>inf1</th></tr></thead><tbody><tr><td>count</td><td>2200.000000</td><td>2200</td><td>2200</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></tr><tr><td>unique</td><td>NaN</td><td>1024</td><td>59</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr><tr><td>top</td><td>NaN</td><td>Harvard University</td><td>USA</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr><tr><td>freq</td><td>NaN</td><td>4</td><td>573</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr><tr><td>mean</td><td>459.590909</td><td>NaN</td><td>NaN</td><td>40.278182</td><td>275.100455</td><td>357.116818</td><td>178.888182</td><td>459.908636</td><td>459.71</td></tr><tr><td>std</td><td>304.320363</td><td>NaN</td><td>NaN</td><td>51.740670</td><td>121.935100</td><td>186.779252</td><td>64.050885</td><td>303.760352</td><td>303.31</td></tr><tr><td>min</td><td>1.000000</td><td>NaN</td><td>NaN</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.000000</td><td>1.00</td></tr><tr><td>25%</td><td>175.750000</td><td>NaN</td><td>NaN</td><td>6.000000</td><td>175.750000</td><td>175.750000</td><td>175.750000</td><td>175.750000</td><td>175.71</td></tr><tr><td>50%</td><td>450.500000</td><td>NaN</td><td>NaN</td><td>21.000000</td><td>355.000000</td><td>450.500000</td><td>210.000000</td><td>450.500000</td><td>450.51</td></tr><tr><td>75%</td><td>725.250000</td><td>NaN</td><td>NaN</td><td>49.000000</td><td>367.000000</td><td>478.000000</td><td>218.000000</td><td>725.000000</td><td>725.21</td></tr><tr><td>max</td><td>1000.000000</td><td>NaN</td><td>NaN</td><td>229.000000</td><td>367.000000</td><td>567.000000</td><td>218.000000</td><td>1000.000000</td><td>991.01</td></tr></tbody></table>		world_rank	institution	country	national_rank	quality_of_education	alumi_employment	quality_of_faculty	publications	inf1	count	2200.000000	2200	2200	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	unique	NaN	1024	59	NaN	NaN	NaN	NaN	NaN	NaN	top	NaN	Harvard University	USA	NaN	NaN	NaN	NaN	NaN	NaN	freq	NaN	4	573	NaN	NaN	NaN	NaN	NaN	NaN	mean	459.590909	NaN	NaN	40.278182	275.100455	357.116818	178.888182	459.908636	459.71	std	304.320363	NaN	NaN	51.740670	121.935100	186.779252	64.050885	303.760352	303.31	min	1.000000	NaN	NaN	1.000000	1.000000	1.000000	1.000000	1.000000	1.00	25%	175.750000	NaN	NaN	6.000000	175.750000	175.750000	175.750000	175.750000	175.71	50%	450.500000	NaN	NaN	21.000000	355.000000	450.500000	210.000000	450.500000	450.51	75%	725.250000	NaN	NaN	49.000000	367.000000	478.000000	218.000000	725.000000	725.21	max	1000.000000	NaN	NaN	229.000000	367.000000	567.000000	218.000000	1000.000000	991.01
	world_rank	institution	country	national_rank	quality_of_education	alumi_employment	quality_of_faculty	publications	inf1																																																																																																															
count	2200.000000	2200	2200	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000																																																																																																															
unique	NaN	1024	59	NaN	NaN	NaN	NaN	NaN	NaN																																																																																																															
top	NaN	Harvard University	USA	NaN	NaN	NaN	NaN	NaN	NaN																																																																																																															
freq	NaN	4	573	NaN	NaN	NaN	NaN	NaN	NaN																																																																																																															
mean	459.590909	NaN	NaN	40.278182	275.100455	357.116818	178.888182	459.908636	459.71																																																																																																															
std	304.320363	NaN	NaN	51.740670	121.935100	186.779252	64.050885	303.760352	303.31																																																																																																															
min	1.000000	NaN	NaN	1.000000	1.000000	1.000000	1.000000	1.000000	1.00																																																																																																															
25%	175.750000	NaN	NaN	6.000000	175.750000	175.750000	175.750000	175.750000	175.71																																																																																																															
50%	450.500000	NaN	NaN	21.000000	355.000000	450.500000	210.000000	450.500000	450.51																																																																																																															
75%	725.250000	NaN	NaN	49.000000	367.000000	478.000000	218.000000	725.000000	725.21																																																																																																															
max	1000.000000	NaN	NaN	229.000000	367.000000	567.000000	218.000000	1000.000000	991.01																																																																																																															

Univariate Analysis

Bivariate Analysis



Multivariate Analysis



Outliers and Anomalies	-																																																																																																																														
Data Preprocessing Code Screenshots																																																																																																																															
Loading Data	<div><pre>times=pd.read_csv("timesData.csv") times.head()</pre><table><thead><tr><th></th><th>world_rank</th><th>university_name</th><th>country</th><th>teaching</th><th>international</th><th>research</th><th>citations</th><th>income</th><th>total_score</th><th>num_studen</th></tr></thead><tbody><tr><td>0</td><td>1</td><td>Harvard University</td><td>United States of America</td><td>99.7</td><td>72.4</td><td>98.7</td><td>98.8</td><td>34.5</td><td>96.1</td><td>20.1</td></tr><tr><td>1</td><td>2</td><td>California Institute of Technology</td><td>United States of America</td><td>97.7</td><td>54.6</td><td>98.0</td><td>99.9</td><td>83.7</td><td>96.0</td><td>2.2</td></tr><tr><td>2</td><td>3</td><td>Massachusetts Institute of Technology</td><td>United States of America</td><td>97.8</td><td>82.3</td><td>91.4</td><td>99.9</td><td>87.5</td><td>95.6</td><td>11.0</td></tr><tr><td>3</td><td>4</td><td>Stanford University</td><td>United States of America</td><td>98.3</td><td>29.5</td><td>98.1</td><td>99.2</td><td>64.3</td><td>94.3</td><td>15.5</td></tr><tr><td>4</td><td>5</td><td>Princeton University</td><td>United States of America</td><td>90.9</td><td>70.3</td><td>95.4</td><td>99.9</td><td>-</td><td>94.2</td><td>7.9</td></tr></tbody></table></div> <div><pre>cwur=pd.read_csv("cwurData.csv") cwur.head()</pre><table><thead><tr><th></th><th>world_rank</th><th>institution</th><th>country</th><th>national_rank</th><th>quality_of_education</th><th>alumni_employment</th><th>quality_of_faculty</th><th>publications</th><th>infl</th></tr></thead><tbody><tr><td>0</td><td>1</td><td>Harvard University</td><td>USA</td><td>1</td><td></td><td>7</td><td>9</td><td>1</td><td>1</td></tr><tr><td>1</td><td>2</td><td>Massachusetts Institute of Technology</td><td>USA</td><td>2</td><td></td><td>9</td><td>17</td><td>3</td><td>12</td></tr><tr><td>2</td><td>3</td><td>Stanford University</td><td>USA</td><td>3</td><td></td><td>17</td><td>11</td><td>5</td><td>4</td></tr><tr><td>3</td><td>4</td><td>University of Cambridge</td><td>United Kingdom</td><td>1</td><td></td><td>10</td><td>24</td><td>4</td><td>16</td></tr><tr><td>4</td><td>5</td><td>California Institute of Technology</td><td>USA</td><td>4</td><td></td><td>2</td><td>29</td><td>7</td><td>37</td></tr></tbody></table></div>		world_rank	university_name	country	teaching	international	research	citations	income	total_score	num_studen	0	1	Harvard University	United States of America	99.7	72.4	98.7	98.8	34.5	96.1	20.1	1	2	California Institute of Technology	United States of America	97.7	54.6	98.0	99.9	83.7	96.0	2.2	2	3	Massachusetts Institute of Technology	United States of America	97.8	82.3	91.4	99.9	87.5	95.6	11.0	3	4	Stanford University	United States of America	98.3	29.5	98.1	99.2	64.3	94.3	15.5	4	5	Princeton University	United States of America	90.9	70.3	95.4	99.9	-	94.2	7.9		world_rank	institution	country	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	infl	0	1	Harvard University	USA	1		7	9	1	1	1	2	Massachusetts Institute of Technology	USA	2		9	17	3	12	2	3	Stanford University	USA	3		17	11	5	4	3	4	University of Cambridge	United Kingdom	1		10	24	4	16	4	5	California Institute of Technology	USA	4		2	29	7	37
	world_rank	university_name	country	teaching	international	research	citations	income	total_score	num_studen																																																																																																																					
0	1	Harvard University	United States of America	99.7	72.4	98.7	98.8	34.5	96.1	20.1																																																																																																																					
1	2	California Institute of Technology	United States of America	97.7	54.6	98.0	99.9	83.7	96.0	2.2																																																																																																																					
2	3	Massachusetts Institute of Technology	United States of America	97.8	82.3	91.4	99.9	87.5	95.6	11.0																																																																																																																					
3	4	Stanford University	United States of America	98.3	29.5	98.1	99.2	64.3	94.3	15.5																																																																																																																					
4	5	Princeton University	United States of America	90.9	70.3	95.4	99.9	-	94.2	7.9																																																																																																																					
	world_rank	institution	country	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	infl																																																																																																																						
0	1	Harvard University	USA	1		7	9	1	1																																																																																																																						
1	2	Massachusetts Institute of Technology	USA	2		9	17	3	12																																																																																																																						
2	3	Stanford University	USA	3		17	11	5	4																																																																																																																						
3	4	University of Cambridge	United Kingdom	1		10	24	4	16																																																																																																																						
4	5	California Institute of Technology	USA	4		2	29	7	37																																																																																																																						
Handling Missing Data	<div><pre>times['student_staff_ratio'].fillna(times['student_staff_ratio'].mean(), inplace=True) times['international_students'].fillna(times['international_students'].mode()[0], inplace=True) mode_value = times['female_male_ratio'].mode()[0] times['female_male_ratio'].fillna(mode_value, inplace=True) times['num_students'] = pd.to_numeric(times['num_students'], errors='coerce') times['num_students'].fillna(times['num_students'].mean(), inplace=True)</pre></div> <div><pre>cwur['broad_impact'].fillna(cwur['broad_impact'].mean(), inplace=True)</pre></div>																																																																																																																														
Data Transformation	<div><pre>times['world_rank'] = pd.to_numeric(times['world_rank'], errors='coerce') times['female_male_ratio'] = pd.to_numeric(times['female_male_ratio'], errors='coerce') times['income'] = pd.to_numeric(times['income'], errors='coerce') times['total_score'] = pd.to_numeric(times['total_score'], errors='coerce') times['international_students'] = pd.to_numeric(times['international_students'], errors='coerce') times['international'] = pd.to_numeric(times['international'], errors='coerce')</pre></div>																																																																																																																														
Feature Engineering	Attached the codes in final submission.																																																																																																																														
Save Processed Data	-																																																																																																																														