Data Collection and Preprocessing Phase

Date	20 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	Dimension:
	2200 rows×14 columns
	Descriptive statistics:
	`snc_df.describe(include='all')`
	`cwur.describe(include='all')`

snc_df.describe(include='all')

	school_name	country
count	818	818
unique	818	70
top	Harvard University	United States of America
freq	1	161

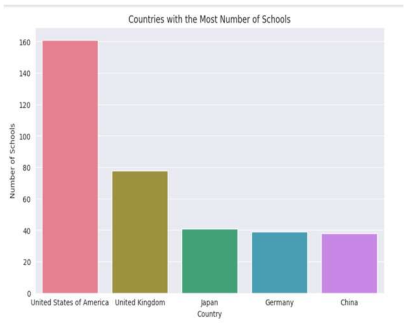
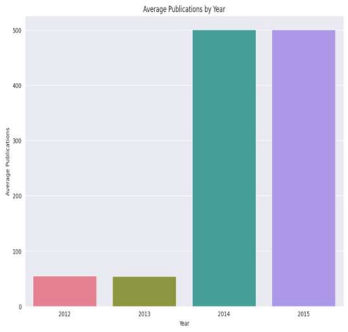
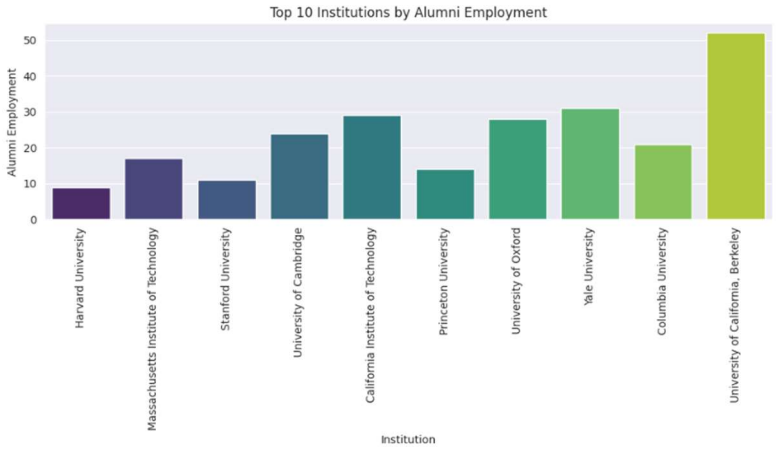
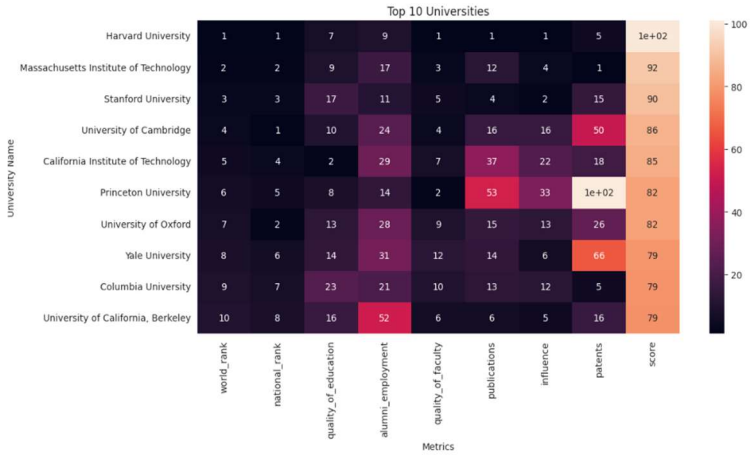
cwur.describe(include='all')

	world_rank	institution	country	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	infl
count	2200.000000	2200	2200	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.0
unique	NaN	1024	59	NaN	NaN	NaN	NaN	NaN	
top	NaN	Harvard University	USA	NaN	NaN	NaN	NaN	NaN	
freq	NaN	4	573	NaN	NaN	NaN	NaN	NaN	
mean	459.590909	NaN	NaN	40.278182	275.100455	357.116818	178.888182	459.908636	459.7
std	304.320363	NaN	NaN	51.740870	121.935100	186.779252	64.050885	303.760352	303.3
min	1.000000	NaN	NaN	1.000000	1.000000	1.000000	1.000000	1.000000	1.0
25%	175.750000	NaN	NaN	6.000000	175.750000	175.750000	175.750000	175.750000	175.7
50%	450.500000	NaN	NaN	21.000000	355.000000	450.500000	210.000000	450.500000	450.5
75%	725.250000	NaN	NaN	49.000000	367.000000	478.000000	218.000000	725.000000	725.2
max	1000.000000	NaN	NaN	229.000000	367.000000	567.000000	218.000000	1000.000000	991.0

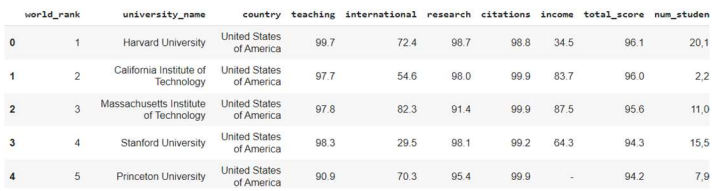
```
times.describe(include='all')
```

	world_rank	university_name	country	teaching	international	research	citations	income	total_s
count	1112.000000	2603	2603	2603.000000	2594.000000	2603.000000	2603.000000	2385.000000	1201.00
unique	NaN	818	72	NaN	NaN	NaN	NaN	NaN	
top	NaN	Harvard University	United States of America	NaN	NaN	NaN	NaN	NaN	
freq	NaN	6	659	NaN	NaN	NaN	NaN	NaN	
mean	98.201439	NaN	NaN	37.801498	52.007440	35.910257	60.921629	48.979874	59.84
std	58.097290	NaN	NaN	17.604218	22.103825	21.254805	23.073219	21.179938	12.80
min	1.000000	NaN	NaN	9.900000	7.100000	2.900000	1.200000	24.200000	41.40
25%	48.000000	NaN	NaN	24.700000	33.425000	19.600000	45.500000	33.000000	50.30
50%	97.000000	NaN	NaN	33.900000	50.300000	30.500000	62.500000	41.000000	56.00
75%	148.000000	NaN	NaN	46.400000	69.000000	47.250000	79.050000	59.000000	66.20
max	200.000000	NaN	NaN	99.700000	100.000000	99.400000	100.000000	100.000000	96.10

Univariate Analysis

	 
Bivariate Analysis	
Multivariate Analysis	

Outliers and Anomalies	-
Data Preprocessing Code Screenshots	

Loading Data	`times=pd.read_csv("timesData.csv")`
	`times.head()`
	 
	`cwur=pd.read_csv("cwurData.csv")`
	`cwur.head()`
	 

Handling Missing Data	`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)`
	
	`cwur['broad_impact'].fillna(cwur['broad_impact'].mean(), inplace=True)`

Data Transformation	`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')`

Feature Engineering	Attached the codes in final submission.
Save Processed Data	-