Data Collection and Preprocessing Phase

Date	28 June 2024
Team ID	740079
Project Title	A Comprehensive Measure Of Well-Being:A Human Development Using Machine Learning
Maximum Marks	6 Marks

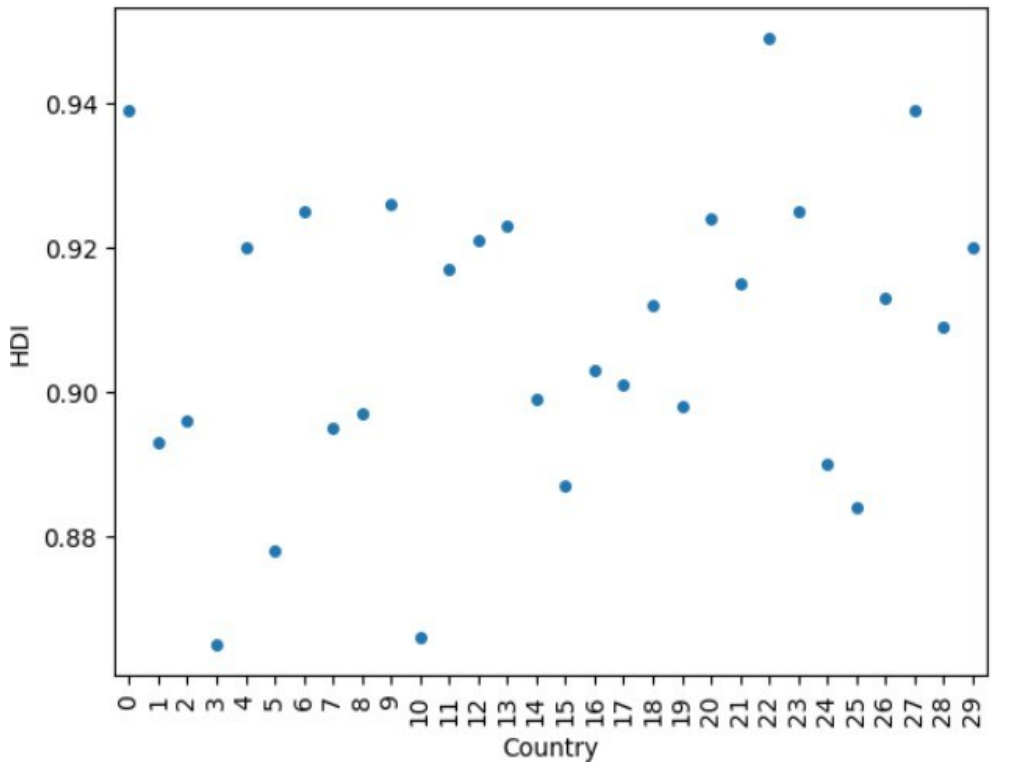
Data Exploration and Preprocessing Report

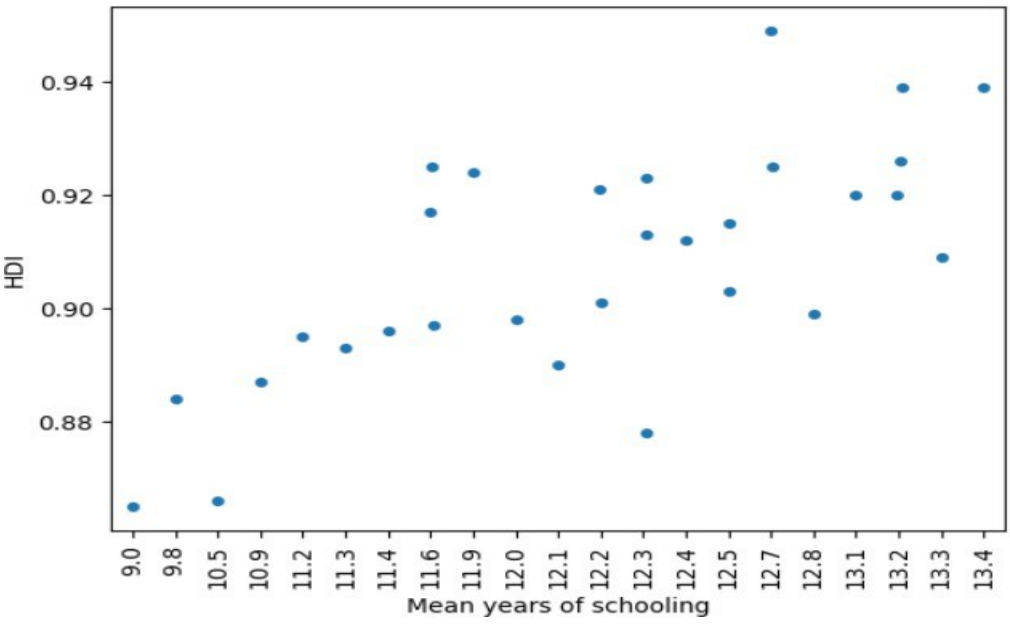
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong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

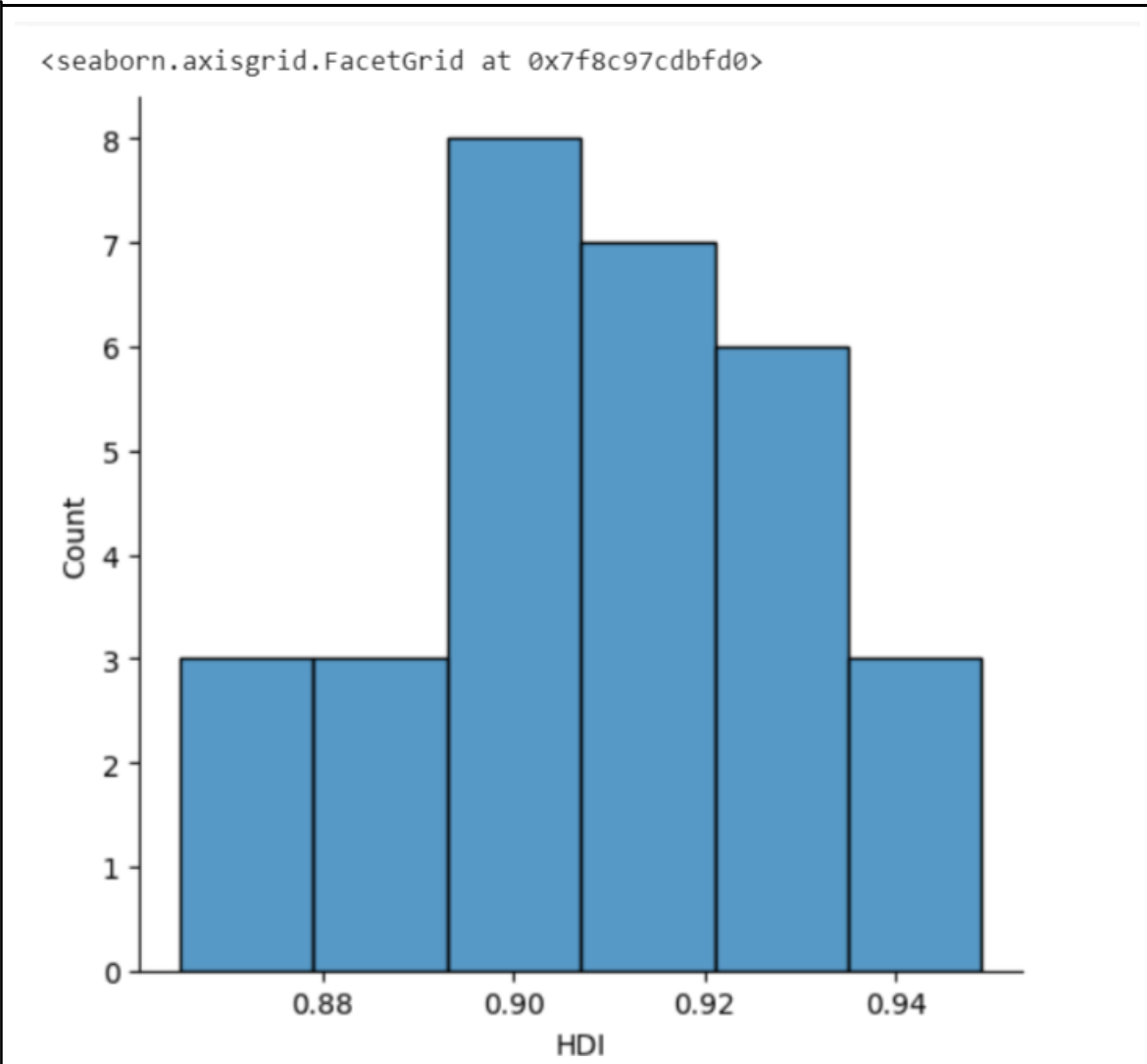
Section	Description
Data Overview	Descriptive Analysis:-
Univariate Analysis	

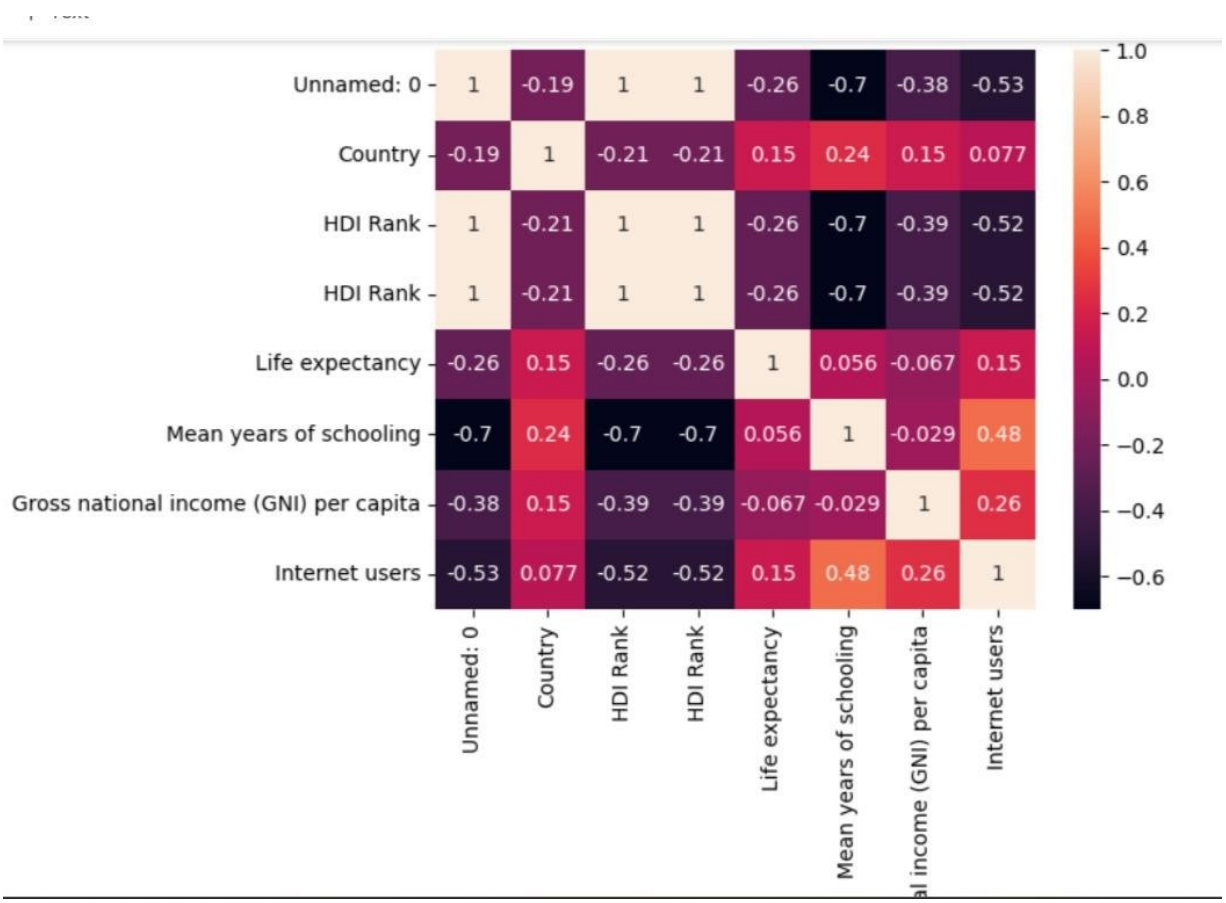
```
[ ] data.describe()
```

	Unnamed: 0	Id	HDI Rank	HDI	Life expectancy	Mean years of schooling	Gross national income (GNI) per capita	GNI per capita rank minus HDI rank	Change in HDI rank 2010-2015	Average annual HDI growth 1990-2000	...	Coefficient of human inequality	Inequality in life expectancy (%) 2010-2015	Inequality-adjusted life expectancy index	Inequality in education(%)	Inequality-adjusted education index	Inequality in income (%)
count	195.000000	195.000000	188.000000	188.000000	190.000000	188.000000	193.000000	188.000000	188.000000	144.000000	...	151.000000	185.000000	185.000000	160.000000	158.000000	154.000000
mean	97.000000	98.000000	94.319149	0.698835	71.267368	8.371277	17239.502591	2.239362	-0.021277	0.773611	...	20.609934	17.155135	0.666151	20.278125	0.527722	24.115584
std	56.435804	56.435804	54.435501	0.155153	8.328811	3.098028	19052.801649	17.309425	5.440940	0.762704	...	10.210353	12.197865	0.190741	14.631437	0.224236	10.511501
min	0.000000	1.000000	1.000000	0.352000	48.900000	1.400000	294.000000	-79.000000	-35.000000	-1.550000	...	5.300000	2.600000	0.263000	1.400000	0.134000	8.900000
25%	48.500000	49.500000	47.750000	0.577750	65.550000	6.075000	3527.000000	-7.000000	-2.000000	0.425000	...	10.950000	6.700000	0.524000	6.575000	0.324000	16.300000
50%	97.000000	98.000000	94.500000	0.728500	73.300000	8.650000	10383.000000	4.000000	0.000000	0.760000	...	20.200000	13.600000	0.707000	17.450000	0.527000	21.300000
75%	145.500000	146.500000	139.750000	0.811250	77.000000	11.125000	23394.000000	13.000000	2.000000	1.150000	...	29.500000	26.100000	0.813000	34.525000	0.732250	29.475000
max	194.000000	195.000000	188.000000	0.949000	84.200000	13.400000	129916.000000	48.000000	15.000000	3.630000	...	44.800000	46.200000	0.959000	48.300000	0.921000	56.400000


HDI
0.94
0.92
0.90
0.88
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Country


HDI
0.94
0.92
0.90
0.88
9.0 9.8 10.5 10.9 11.2 11.3 11.4 11.6 11.9 12.0 12.1 12.2 12.3 12.4 12.5 12.7 12.8 13.1 13.2 13.3 13.4
Mean years of schooling

Bivariate Analysis	

Multivariate Analysis	

Outliers and Anomalies	-
Data Preprocessing Code Screenshots	

Loading Data

```
[4] data=pd.read_csv("/content/HDI.csv")
```

```
[5] data.head()
```

	Unnamed: 0	Id	Country	HDI Rank	HDI	Life expectancy	Mean years of schooling	Gross national income (GNI) per capita	GNI per capita rank minus HDI rank	Change in HDI rank 2010-2015	...	Coefficient of human inequality	Inequality in life expectancy (%) 2010-2015	Inequality-adjusted life expectancy index	Inequality in education(%)	Inequality-adjusted education index	Inequality in income (%)	Inequali adjus inc in
0	0	1	Norway	1.0	0.949	81.7	12.7	67614.0	5.0	0.0	...	5.4	3.3	0.918	2.4	0.894	10.4	0.
1	1	2	Australia	2.0	0.939	82.5	13.2	42822.0	19.0	1.0	...	8.0	4.3	0.921	1.9	0.921	17.7	0.
2	2	3	Switzerland	2.0	0.939	83.1	13.4	56364.0	7.0	0.0	...	8.4	3.8	0.934	5.7	0.840	15.7	0.
3	3	4	Germany	4.0	0.926	81.1	13.2	45000.0	13.0	0.0	...	7.0	3.7	0.905	2.6	0.891	14.8	0.
4	4	5	Denmark	5.0	0.925	80.4	12.7	44519.0	13.0	2.0	...	7.0	3.8	0.894	3.0	0.896	14.3	0.

5 rows × 82 columns

Handling Missing Data

```
[18] data_filled = data_encoded.fillna(data_encoded.mean())
```

```
data_filled.fillna(data_filled.mean(),inplace=True)
data_filled
```

	Unnamed: 0	Id	Country	HDI Rank	HDI	Life expectancy	Mean years of schooling	Gross national income (GNI) per capita	GNI per capita rank minus HDI rank	Change in HDI rank 2010-2015	...	Coefficient of human inequality	Inequality in life expectancy (%) 2010-2015	Inequality-adjusted life expectancy index	Inequality in education(%)	Inequality-adjusted education index	Inequality in income (%)	Inequality-adjusted income index	Income inequality (Quintile ratio) 2010-2015	Income inequality (Palma ratio) 2010-2015
0	0	1	22	1.0	0.949	81.7	12.7	67614.0	5.0	0.0	...	5.400	3.300000	0.918000	2.400	0.89400	10.400	0.88200	3.800000	0.900000
1	1	2	0	2.0	0.939	82.5	13.2	42822.0	19.0	1.0	...	8.000	4.300000	0.921000	1.900	0.92100	17.700	0.75300	6.000000	1.400000
2	2	3	27	2.0	0.939	83.1	13.4	56364.0	7.0	0.0	...	8.400	3.800000	0.934000	5.700	0.84000	15.700	0.80600	4.900000	1.200000
3	3	4	9	4.0	0.926	81.1	13.2	45000.0	13.0	0.0	...	7.000	3.700000	0.905000	2.600	0.89100	14.800	0.78700	4.600000	1.100000
4	4	5	6	5.0	0.925	80.4	12.7	44519.0	13.0	2.0	...	7.000	3.800000	0.894000	3.000	0.89600	14.300	0.78900	4.500000	1.000000
5	5	6	23	5.0	0.925	83.2	11.6	78162.0	-3.0	0.0	...	8.688	3.000000	0.943000	6.124	0.81848	16.184	0.76008	5.445833	1.245833
6	6	7	20	7.0	0.924	81.7	11.9	46326.0	8.0	-2.0	...	6.800	3.700000	0.914000	4.200	0.85900	12.400	0.81200	4.200000	1.000000
7	7	8	13	8.0	0.923	81.1	12.3	43798.0	11.0	1.0	...	7.700	3.700000	0.905000	3.000	0.88300	16.300	0.76900	5.300000	1.300000
8	8	9	12	9.0	0.921	82.7	12.2	37065.0	20.0	7.0	...	5.700	2.900000	0.937000	2.500	0.88400	11.700	0.78900	4.000000	1.000000
9	9	10	4	10.0	0.920	82.2	13.1	42582.0	12.0	1.0	...	8.700	4.700000	0.912000	3.900	0.85600	17.400	0.75500	5.800000	1.300000
10	10	11	29	10.0	0.920	79.2	13.2	53245.0	1.0	-3.0	...	12.900	6.100000	0.856000	5.600	0.85000	27.000	0.69200	9.100000	2.000000
11	11	12	11	12.0	0.917	84.2	11.6	54265.0	-2.0	3.0	...	8.688	2.800000	0.959000	6.124	0.81848	16.184	0.76008	5.445833	1.245833

Data Transformation

```
import pandas as pd
from sklearn.preprocessing import OneHotEncoder

# Identify categorical columns
categorical_columns = data1.select_dtypes(include=['object']).columns

# Apply One-Hot Encoding to categorical columns
data_encoded = pd.get_dummies(data1, columns=categorical_columns, drop_first=True)

# Display the first few rows of the encoded dataset
data_encoded.head()
```

	Unnamed: 0	Id	Country	HDI Rank	HDI	Life expectancy	Mean years of schooling	Gross national income (GNI) per capita	GNI per capita rank minus HDI rank	Change in HDI rank 2010-2015	...	Coefficient of human inequality	Inequality in life expectancy (%) 2010-2015	Inequality-adjusted life expectancy index	Inequality in education(%)	Inequality-adjusted education index	Inequality in income (%)	Inequality-adjusted income index	Income inequality (Quintile ratio) 2010-2015	Income inequality (Palma ratio) 2010-2015	co
0	0	1	22	1.0	0.949	81.7	12.7	67614.0	5.0	0.0	...	5.4	3.3	0.918	2.4	0.894	10.4	0.882	3.8	0.9	
1	1	2	0	2.0	0.939	82.5	13.2	42822.0	19.0	1.0	...	8.0	4.3	0.921	1.9	0.921	17.7	0.753	6.0	1.4	
2	2	3	27	2.0	0.939	83.1	13.4	56364.0	7.0	0.0	...	8.4	3.8	0.934	5.7	0.840	15.7	0.806	4.9	1.2	
3	3	4	9	4.0	0.926	81.1	13.2	45000.0	13.0	0.0	...	7.0	3.7	0.905	2.6	0.891	14.8	0.787	4.6	1.1	
4	4	5	6	5.0	0.925	80.4	12.7	44519.0	13.0	2.0	...	7.0	3.8	0.894	3.0	0.896	14.3	0.789	4.5	1.0	

5 rows × 82 columns

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