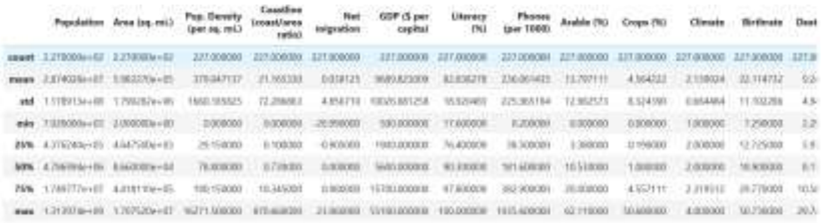
Data Collection and Preprocessing Phase

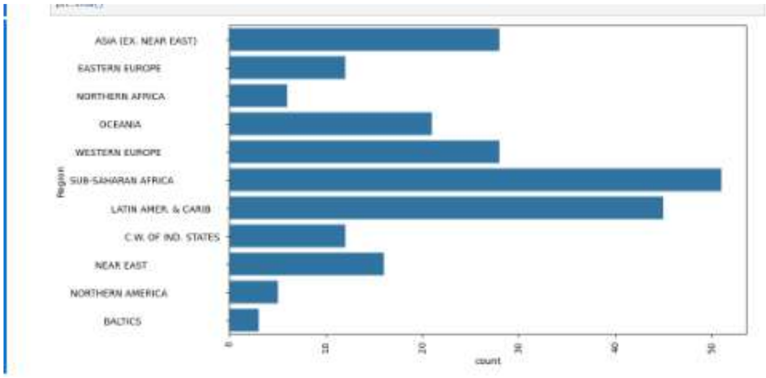
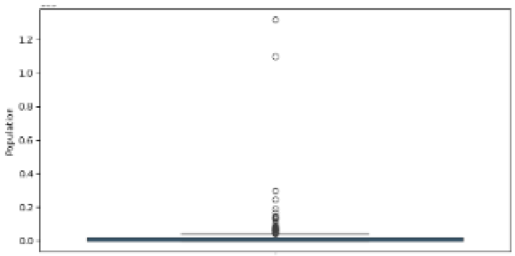
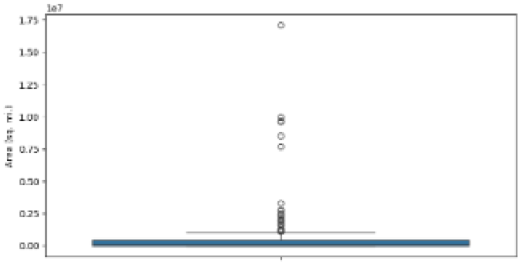
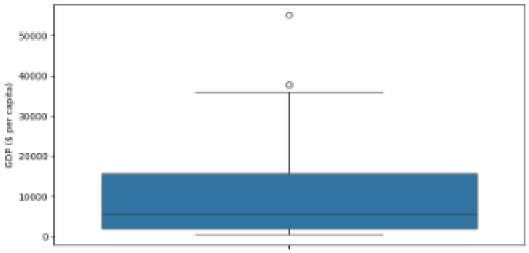
Date	2 July 2024
Team ID	SWTID1720086522
Project Title	Forecasting Economic Prosperity: Leveraging Machine Learning For GDP Per Capita Prediction
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

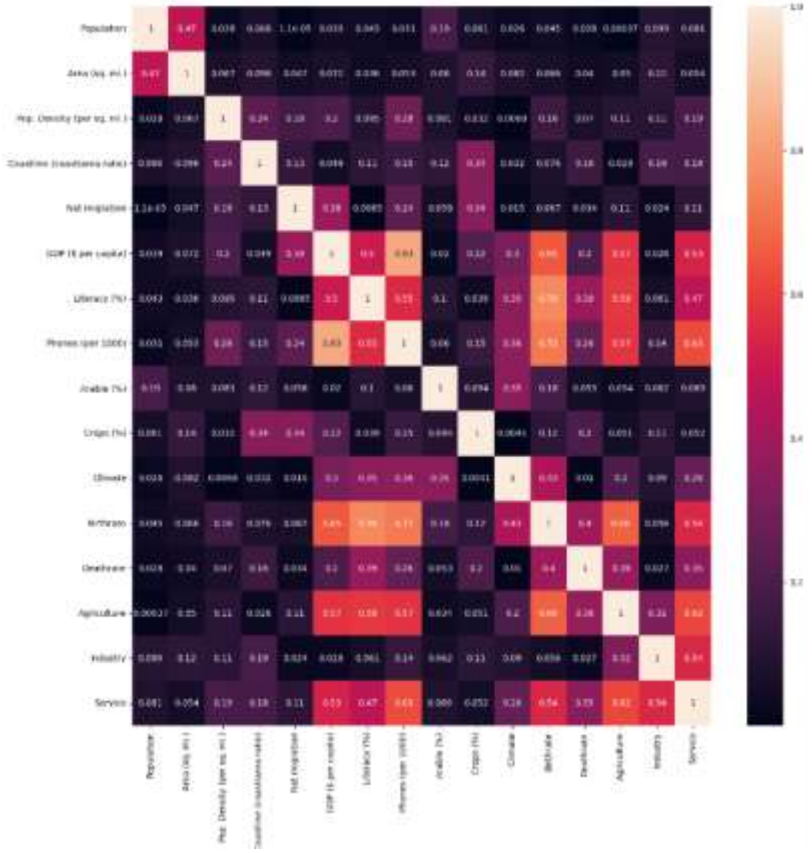
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	**Description**
Data Overview	**Dimension:** 227 rows X 20 columns **Descriptivestatistics:** 

	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)	Climate	Birthrate	Deat
count	2.270000e+02	2.270000e+02	227.000000	227.000000	227.000000	227.000000	227.000000	227.000000	227.000000	227.000000	227.000000	227.000000	227.0
mean	2.874028e+07	5.982270e+05	379.047137	21.165330	0.038125	9689.823009	82.838278	236.061433	13.797111	4.564222	2.139024	22.114732	9.2
std	1.178913e+08	1.790282e+06	1660.185825	72.286863	4.854718	10026.881258	19.552469	225.365184	12.982571	8.524590	0.654464	11.102286	4.9
min	7.026000e+03	2.000000e+00	0.000000	0.000000	-20.990000	500.000000	17.600000	0.200000	0.000000	0.000000	1.000000	7.290000	2.2
25%	4.376240e+05	4.647500e+03	29.150000	0.100000	-0.900000	1900.000000	76.400000	38.300000	3.380000	0.195000	2.000000	12.725000	5.9
50%	4.786994e+06	8.660000e+04	78.800000	0.730000	0.000000	5600.000000	91.100000	181.600000	10.510000	1.000000	2.000000	18.900000	8.1
75%	1.749777e+07	4.418110e+05	190.150000	10.345000	0.900000	15700.000000	97.800000	392.900000	20.000000	4.557111	2.319512	29.770000	10.5
max	1.313974e+09	1.707520e+07	16271.500000	870.660000	21.060000	55100.000000	100.000000	1035.600000	62.110000	50.680000	4.000000	50.730000	29.3

Univariate Analysis	   
Bivariate Analysis	-

Multivariate Analysis

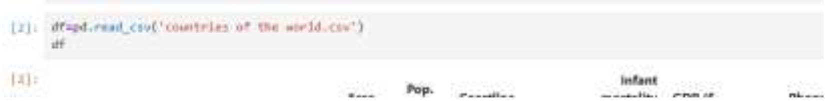


Outliers and Anomalies

-

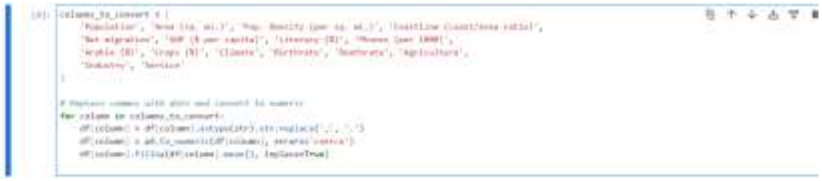
Data Preprocessing Code Screenshots

Loading Data



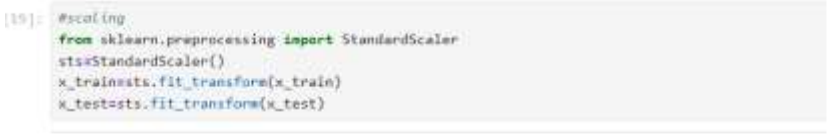
```
[2]: df=pd.read_csv('countries of the world.csv')
     df
[2]:
```

Handling Missing Data

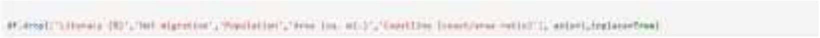


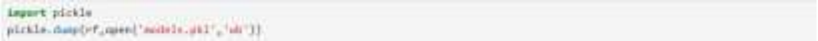
```
[3]: columns_to_convert = [
         'Population', 'Area (sq. mi.)', 'Pop. Density (per sq. mi.)', 'Coastline (coast/area ratio)',
         'Net migration', 'GDP ($ per capita)', 'Literacy (%)', 'Phones (per 1000)',
         'Arable (%)', 'Crops (%)', 'Climate', 'Birthrate', 'Deathrate', 'Agriculture',
         'Industry', 'Service'
     ]

     # Replace commas with dots and convert to numeric
     for column in columns_to_convert:
         df[column] = df[column].astype(str).str.replace(',', '.')
         df[column] = pd.to_numeric(df[column], errors='coerce')
         df[column].fillna(df[column].mean(), inplace=True)
```

Data Transformation



```
[15]: #scaling
      from sklearn.preprocessing import StandardScaler
      sts=StandardScaler()
      x_train=sts.fit_transform(x_train)
      x_test=sts.fit_transform(x_test)
```

Feature Engineering	 `df.drop(['Literacy (%)','Net migration','Population','Area (sq. mi.)','Coastline (coast/area ratio)'], axis=1,inplace=True)`

Save Processed Data	
	`import pickle`
	`pickle.dump(rf,open('models.pkl','wb'))`