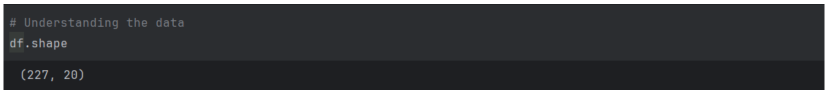
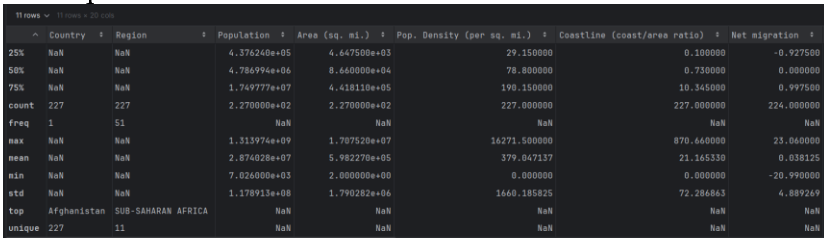
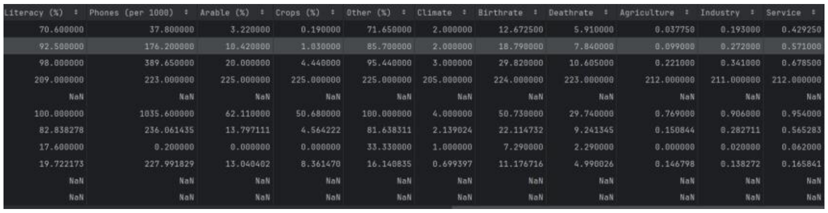
Data Collection and Preprocessing Phase

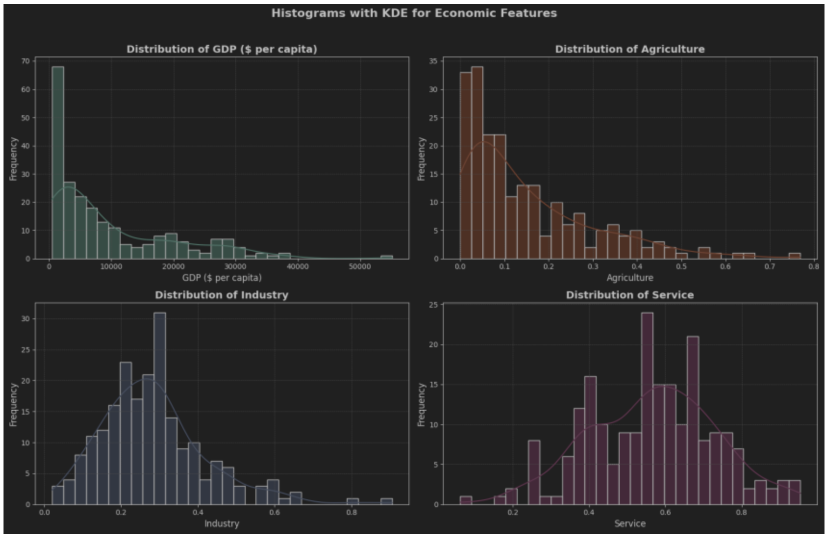
Date	16 June 2025
Team ID	SWTID1749653449
Project Title	Economic Growth: A Machine Learning Approach to GDP per Capita Prediction
Maximum Marks	6 Marks

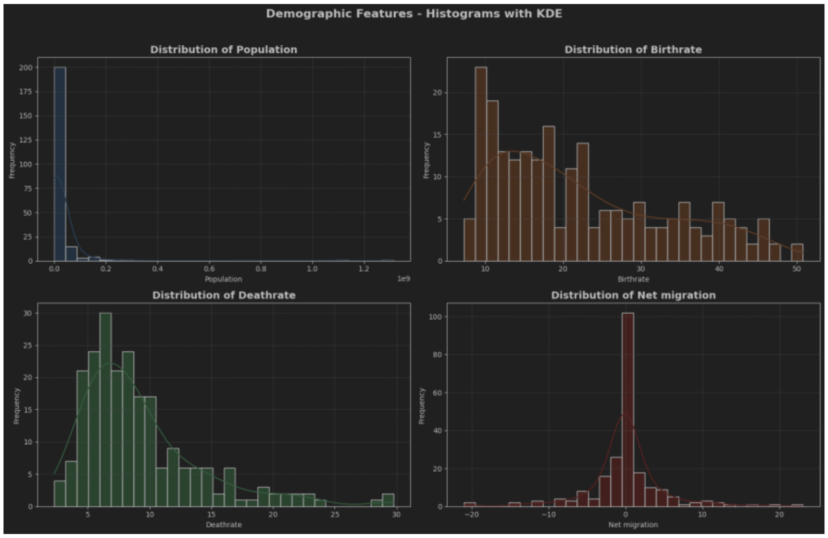
Data Exploration and Preprocessing Template

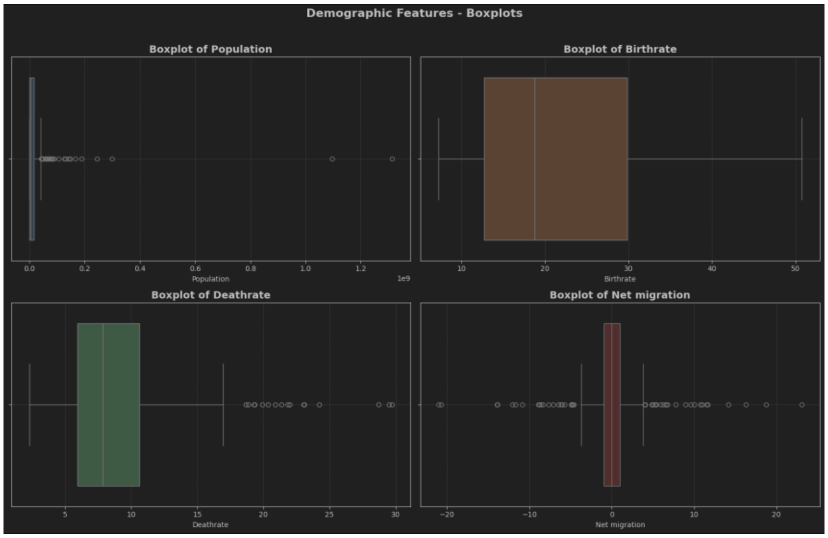
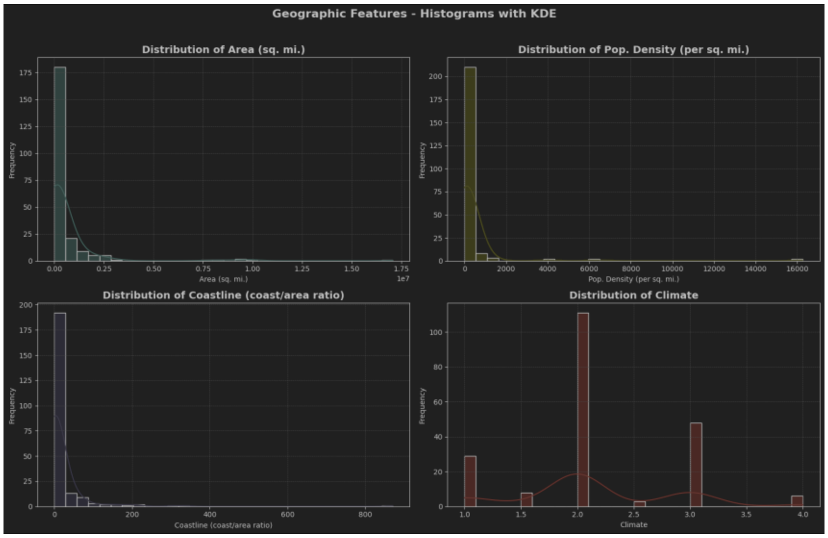
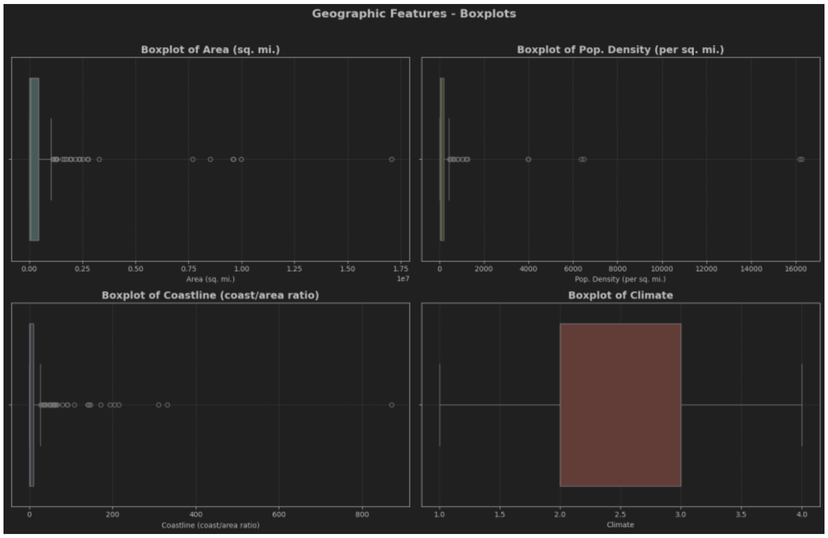
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

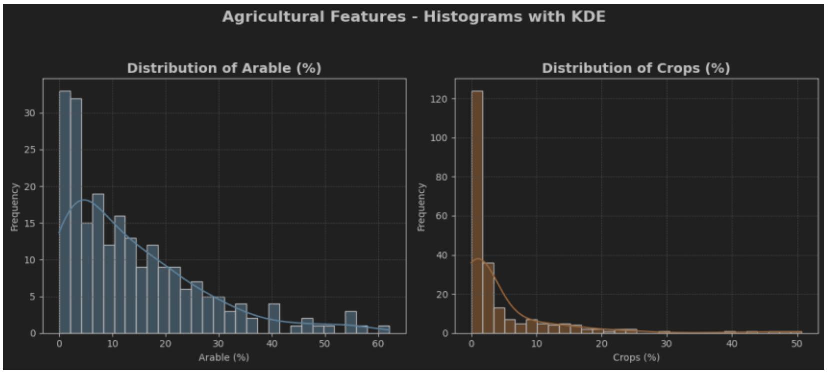
<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>

<pre>
Understanding the data
df.shape

 (227, 20)
</pre>
Dimensions: 227 rows x 20 Columns
Descriptive Stats:

11 rows
11 rows × 20 cols
<table>
<tr><th></th><th>Country</th><th>Region</th><th>Population</th><th>Area (sq. mi.)</th><th>Pop. Density (per sq. mi.)</th><th>Coastline (coast/area ratio)</th><th>Net migration</th></tr>
<tr><td>25%</td><td>NaN</td><td>NaN</td><td>4.376240e+05</td><td>4.647500e+03</td><td>29.150000</td><td>0.100000</td><td>-0.927500</td></tr>
<tr><td>50%</td><td>NaN</td><td>NaN</td><td>4.786994e+06</td><td>8.660000e+04</td><td>78.800000</td><td>0.730000</td><td>0.000000</td></tr>
<tr><td>75%</td><td>NaN</td><td>NaN</td><td>1.749777e+07</td><td>4.418110e+05</td><td>190.150000</td><td>10.345000</td><td>0.997500</td></tr>
<tr><td>count</td><td>227</td><td>227</td><td>2.270000e+02</td><td>2.270000e+02</td><td>227.000000</td><td>227.000000</td><td>224.000000</td></tr>
<tr><td>freq</td><td>1</td><td>51</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr>
<tr><td>max</td><td>NaN</td><td>NaN</td><td>1.313974e+09</td><td>1.707520e+07</td><td>16271.500000</td><td>870.660000</td><td>23.060000</td></tr>
<tr><td>mean</td><td>NaN</td><td>NaN</td><td>2.874028e+07</td><td>5.982270e+05</td><td>379.047137</td><td>21.165330</td><td>0.038125</td></tr>
<tr><td>min</td><td>NaN</td><td>NaN</td><td>7.026000e+03</td><td>2.000000e+00</td><td>0.000000</td><td>0.000000</td><td>-20.990000</td></tr>
<tr><td>std</td><td>NaN</td><td>NaN</td><td>1.178913e+08</td><td>1.790282e+06</td><td>1660.185825</td><td>72.286863</td><td>4.889269</td></tr>
<tr><td>top</td><td>Afghanistan</td><td>SUB-SAHARAN AFRICA</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr>
<tr><td>unique</td><td>227</td><td>11</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr>
</table>

<table>
<tr><th>Literacy (%)</th><th>Phones (per 1000)</th><th>Arable (%)</th><th>Crops (%)</th><th>Other (%)</th><th>Climate</th><th>Birthrate</th><th>Deathrate</th><th>Agriculture</th><th>Industry</th><th>Service</th></tr>
<tr><td>70.600000</td><td>37.800000</td><td>3.220000</td><td>0.190000</td><td>71.650000</td><td>2.000000</td><td>12.672500</td><td>5.910000</td><td>0.037750</td><td>0.193000</td><td>0.429250</td></tr>
<tr><td>92.500000</td><td>176.200000</td><td>10.420000</td><td>1.030000</td><td>85.700000</td><td>2.000000</td><td>18.790000</td><td>7.840000</td><td>0.099000</td><td>0.272000</td><td>0.571000</td></tr>
<tr><td>98.000000</td><td>389.650000</td><td>20.000000</td><td>4.440000</td><td>95.440000</td><td>3.000000</td><td>29.820000</td><td>10.605000</td><td>0.221000</td><td>0.341000</td><td>0.678500</td></tr>
<tr><td>209.000000</td><td>223.000000</td><td>225.000000</td><td>225.000000</td><td>225.000000</td><td>205.000000</td><td>224.000000</td><td>223.000000</td><td>212.000000</td><td>211.000000</td><td>212.000000</td></tr>
<tr><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr>
<tr><td>100.000000</td><td>1035.600000</td><td>62.110000</td><td>50.680000</td><td>100.000000</td><td>4.000000</td><td>50.730000</td><td>29.740000</td><td>0.769000</td><td>0.906000</td><td>0.954000</td></tr>
<tr><td>82.838278</td><td>236.061435</td><td>13.797111</td><td>4.564222</td><td>81.638311</td><td>2.139024</td><td>22.114732</td><td>9.241345</td><td>0.150844</td><td>0.282711</td><td>0.565283</td></tr>
<tr><td>17.600000</td><td>0.200000</td><td>0.000000</td><td>0.000000</td><td>33.330000</td><td>1.000000</td><td>7.290000</td><td>2.290000</td><td>0.000000</td><td>0.020000</td><td>0.062000</td></tr>
<tr><td>19.722173</td><td>227.991829</td><td>13.040402</td><td>8.361470</td><td>16.140835</td><td>0.699397</td><td>11.176716</td><td>4.990026</td><td>0.146798</td><td>0.138272</td><td>0.165841</td></tr>
<tr><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr>
<tr><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr>
</table>
</td></tr>
<tr><td>Univariate Analysis</td><td></td></tr>
</table>


Histograms with KDE for Economic Features
Distribution of GDP ($ per capita)
Frequency
GDP ($ per capita)
0
10000
20000
30000
40000
50000
Distribution of Agriculture
Agriculture
Distribution of Industry
Industry
Distribution of Service
Service

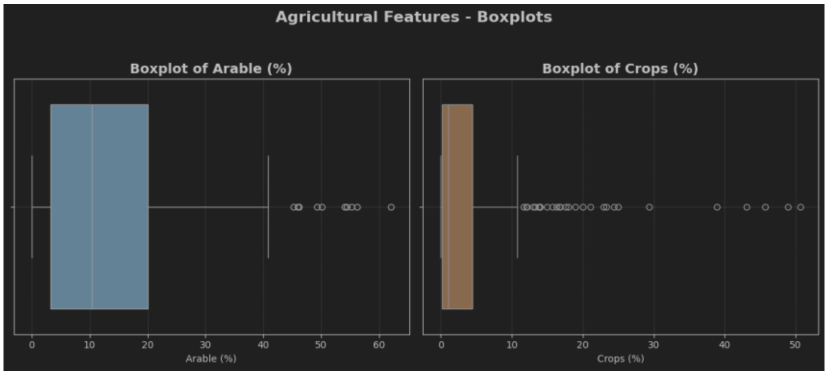

Demographic Features - Histograms with KDE
Distribution of Population
Frequency
Population
1e9
Distribution of Birthrate
Birthrate
Distribution of Deathrate
Deathrate
Distribution of Net migration
Net migration

	Demographic Features - Boxplots Boxplot of Population Population 1e9 Boxplot of Birthrate Birthrate Boxplot of Deathrate Deathrate Boxplot of Net migration Net migration Geographic Features - Histograms with KDE Distribution of Area (sq. mi.) Area (sq. mi.) 1e7 Frequency Distribution of Pop. Density (per sq. mi.) Pop. Density (per sq. mi.) Distribution of Coastline (coast/area ratio) Coastline (coast/area ratio) Distribution of Climate Climate Geographic Features - Boxplots Boxplot of Area (sq. mi.) Area (sq. mi.) 1e7 Boxplot of Pop. Density (per sq. mi.) Pop. Density (per sq. mi.) Boxplot of Coastline (coast/area ratio) Coastline (coast/area ratio) Boxplot of Climate Climate

<table>
<tr>
<td></td>
<td>

Agricultural Features - Histograms with KDE

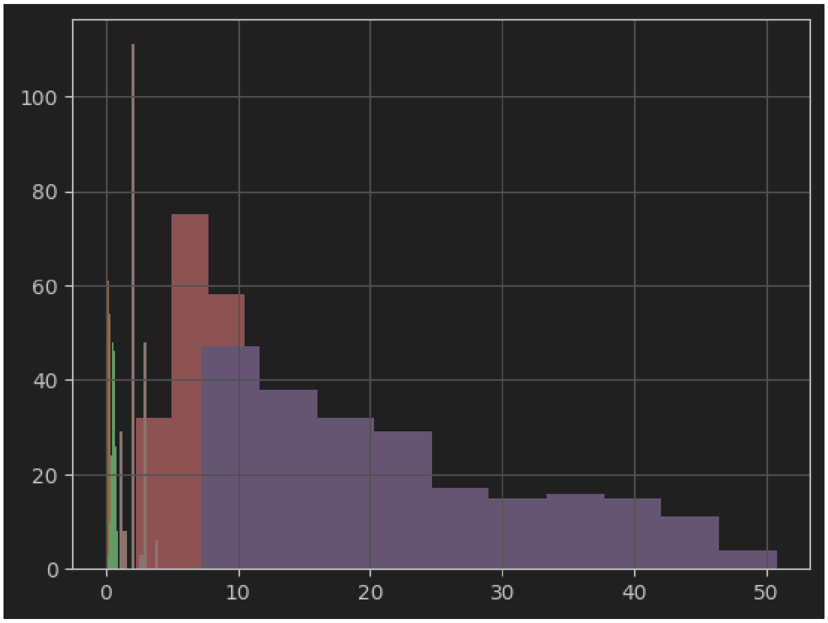
Distribution of Arable (%)

Distribution of Crops (%)

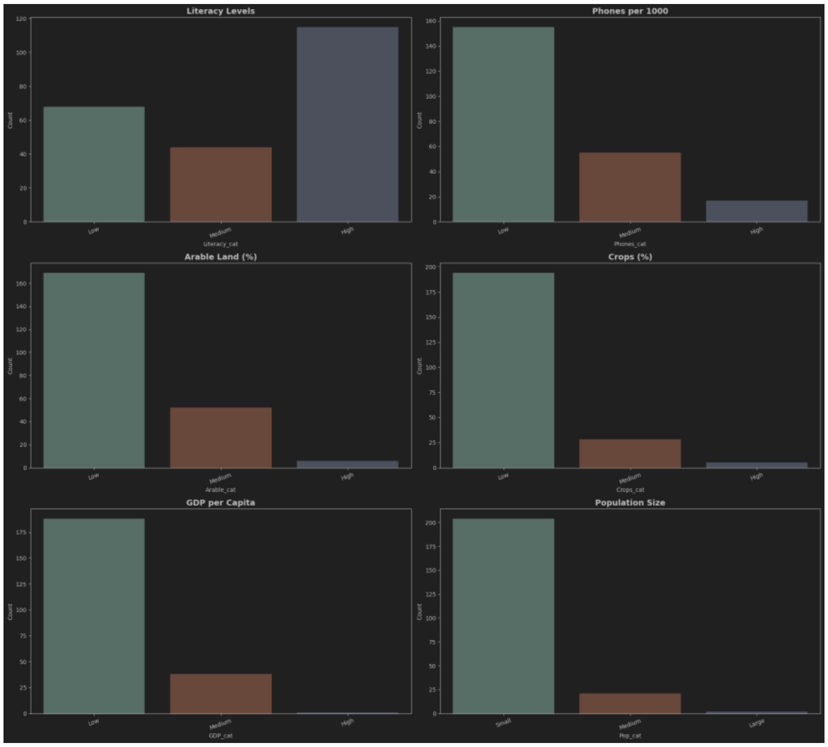

Agricultural Features - Boxplots

Boxplot of Arable (%)

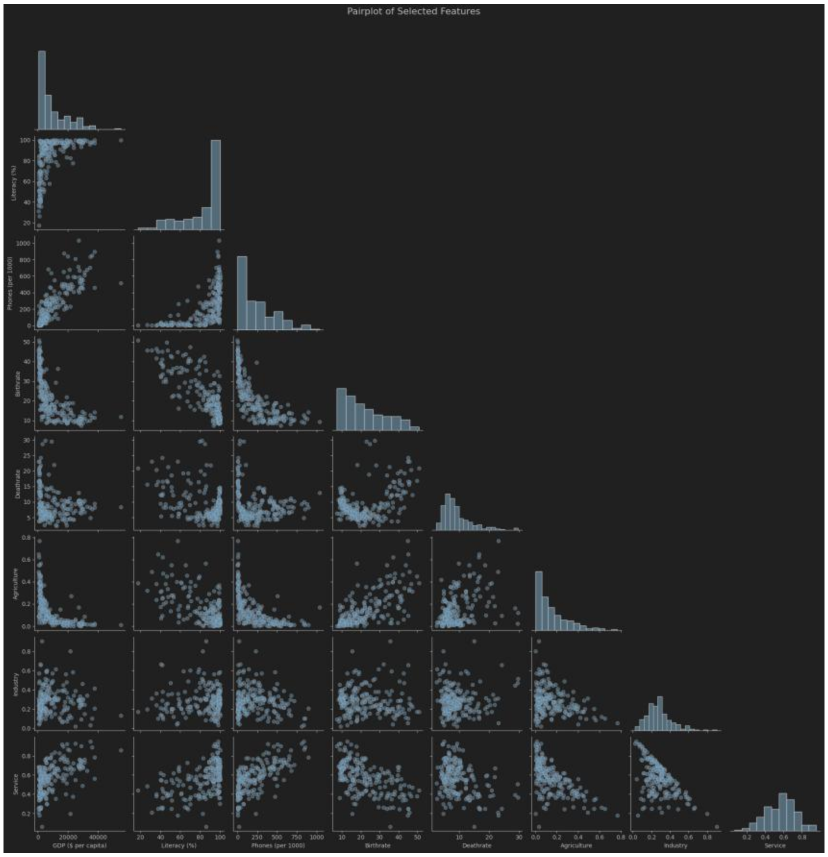
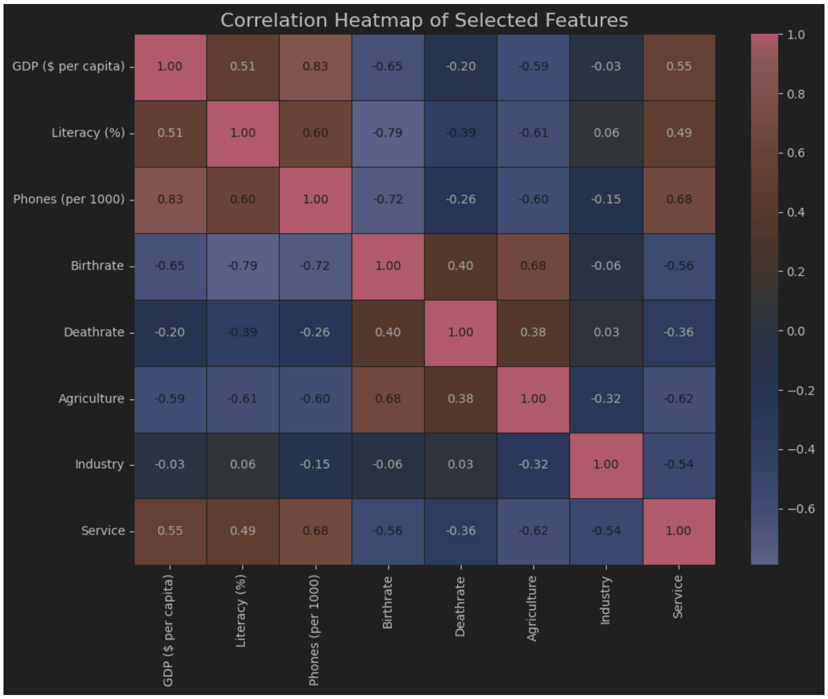
Boxplot of Crops (%)

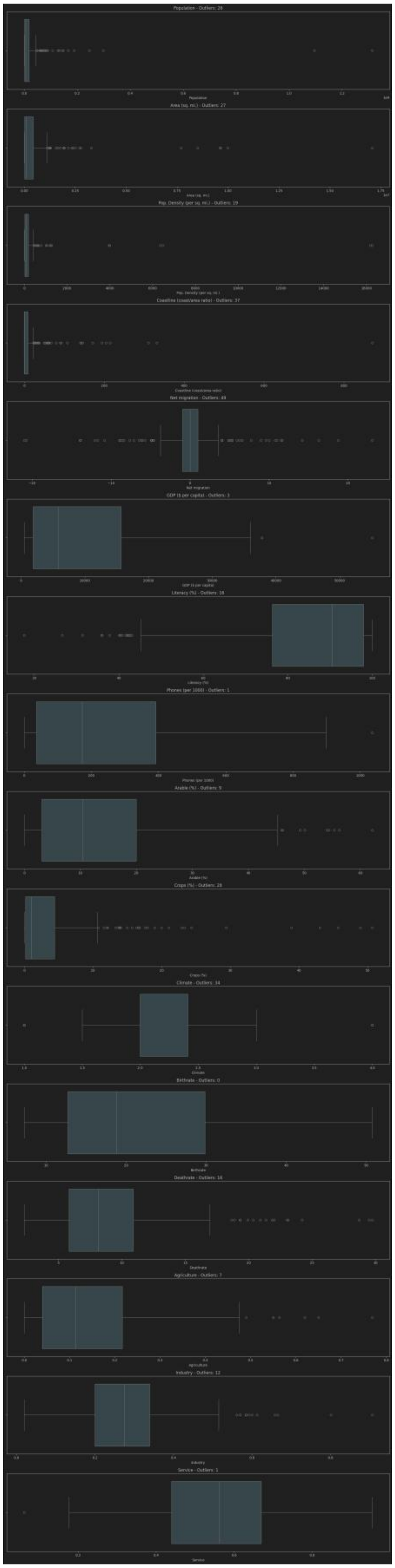


</td>
</tr>
<tr>
<td>Bivariate Analysis</td>
<td></td>
</tr>
</table>

<table>
<tr>
<td></td>
<td>

Literacy Levels
Count
0
20
40
60
80
100
120
Low
Medium
High
Literacy_cat
Phones per 1000
Count
0
20
40
60
80
100
120
140
160
Low
Medium
High
Phones_cat
Arable Land (%)
Count
0
20
40
60
80
100
120
140
160
Low
Medium
High
Arable_cat
Crops (%)
Count
0
25
50
75
100
125
150
175
200
Low
Medium
High
Crops_cat
GDP per Capita
Count
0
25
50
75
100
125
150
175
Low
Medium
High
GDP_cat
Population Size
Count
0
25
50
75
100
125
150
175
200
Small
Medium
Large
Pop_cat

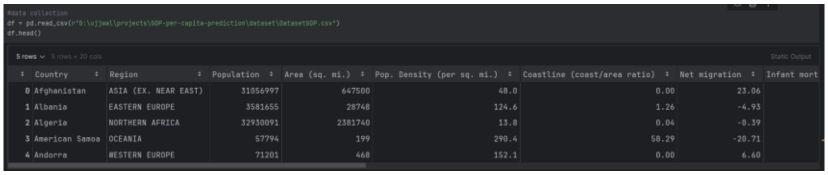
Pair Plot:
</td>
</tr>
</table>

	 .
Multivariate Analysis	

Outliers and Anomalies	

Data Preprocessing Code Screenshots

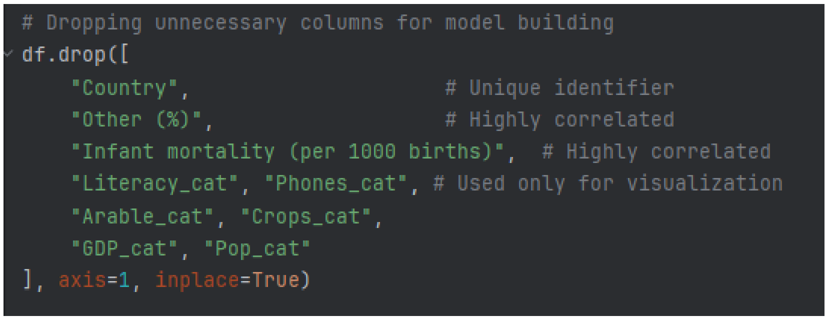
Loading Data



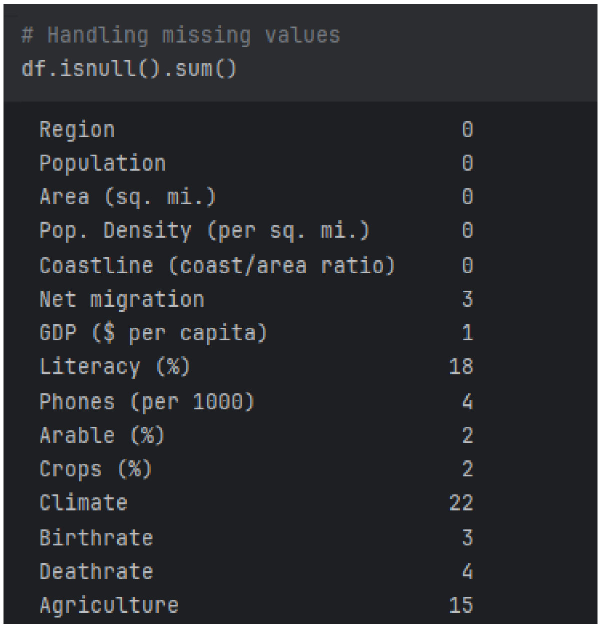
```
#data collection
df = pd.read_csv(r"D:\ujjwal\projects\GDP-per-capita-prediction\dataset\DatasetGDP.csv")
df.head()
```

```
5 rows ∨   5 rows × 20 cols                                                                                                   Static Output
  ⇕ Country         ⇕ Region               ⇕ Population ⇕ Area (sq. mi.) ⇕ Pop. Density (per sq. mi.) ⇕ Coastline (coast/area ratio) ⇕ Net migration ⇕ Infant mort
  0 Afghanistan       ASIA (EX. NEAR EAST)     31056997          647500                         48.0                            0.00            23.06
  1 Albania           EASTERN EUROPE            3581655           28748                        124.6                            1.26            -4.93
  2 Algeria           NORTHERN AFRICA          32930091         2381740                         13.8                            0.04            -0.39
  3 American Samoa    OCEANIA                     57794             199                        290.4                           58.29           -20.71
  4 Andorra           WESTERN EUROPE              71201             468                        152.1                            0.00             6.60
```

Handling Missing Data

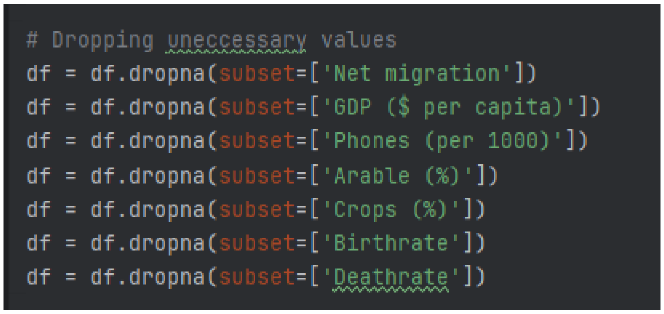


```
# Dropping unnecessary columns for model building
df.drop([
    "Country",                          # Unique identifier
    "Other (%)",                        # Highly correlated
    "Infant mortality (per 1000 births)",  # Highly correlated
    "Literacy_cat", "Phones_cat", # Used only for visualization
    "Arable_cat", "Crops_cat",
    "GDP_cat", "Pop_cat"
], axis=1, inplace=True)
```

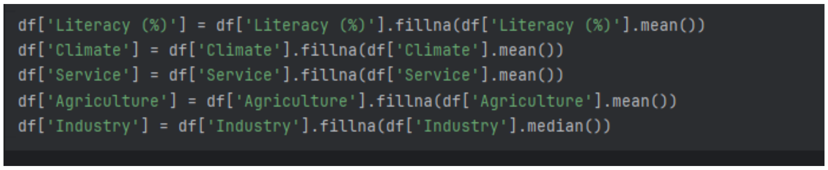


```
# Handling missing values
df.isnull().sum()
```

```
Region                              0
Population                          0
Area (sq. mi.)                      0
Pop. Density (per sq. mi.)          0
Coastline (coast/area ratio)        0
Net migration                       3
GDP ($ per capita)                  1
Literacy (%)                       18
Phones (per 1000)                   4
Arable (%)                          2
Crops (%)                           2
Climate                            22
Birthrate                           3
Deathrate                           4
Agriculture                        15
```



```
# Dropping uneccessary values
df = df.dropna(subset=['Net migration'])
df = df.dropna(subset=['GDP ($ per capita)'])
df = df.dropna(subset=['Phones (per 1000)'])
df = df.dropna(subset=['Arable (%)'])
df = df.dropna(subset=['Crops (%)'])
df = df.dropna(subset=['Birthrate'])
df = df.dropna(subset=['Deathrate'])
```

Data Transformation	
	`df['Literacy (%)'] = df['Literacy (%)'].fillna(df['Literacy (%)'].mean())`
	`df['Climate'] = df['Climate'].fillna(df['Climate'].mean())`
	`df['Service'] = df['Service'].fillna(df['Service'].mean())`
	`df['Agriculture'] = df['Agriculture'].fillna(df['Agriculture'].mean())`
	`df['Industry'] = df['Industry'].fillna(df['Industry'].median())`

Feature Engineering	Attached the codes in final submission.
Save Processed Data	__ Jupyter Notebook was Used __