EAST WEST INSTITUTE OF TECHNOLOGY

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

INTERNSHIP PRESENTATION ON

GDP ANALYSIS USING DATA SCIENCE

Presented By

EXTERNAL GUIDE
S.Romesh
Trainer, Technofly

Praveen R
(1EW20IS056)

INTERNAL GUIDE
Mrs. Veena N Iter
Asst. Prof
Dept. of ISE

TABLE OF CONTENTS

- ABSTRACT
- INTRODUCTION
- COMPANY PROFILE
- PROPOSED SYSTEM
- RESULTS
- CONCLUSION
- APPLICATIONS
- REFERENCES

ABSTRACT

Gross Domestic Product (GDP) analysis using data science is a critical and multifaceted approach to understanding and evaluating the economic performance of a country or region. GDP serves as a fundamental metric for assessing the overall health and growth of an economy, and data science techniques have become indispensable in extracting valuable insights from the vast amount of data generated in modern economies. The abstract provides an overview of how data science is applied to GDP analysis, highlighting its significance, methodologies, and key takeaways. GDP is a comprehensive measure of a nation's economic activity, encompassing the total value of goods and services produced within a specific time frame. It serves as a crucial indicator for policymakers, businesses, and investors to make informed decisions. conclusion, GDP analysis with data science represents a powerful approach to understanding and managing economic performance. By harnessing the capabilities of data science, stakeholders can gain deeper insights, enhance decision-making, and adapt more effectively to the dynamic nature of modern economies.

INTRODUCTION

- ❖ Gross Domestic Product (GDP) measures the economic performance of a country.
- ❖ Sum of all goods and services produced within a country in a specific time frame.
- ❖ It is one of the most widely used measures of a nation's economic performance and is essential for assessing and comparing the economic health and growth of different countries.
- ❖ Gross Domestic Product (GDP) is a key tool that guides investors, policymakers, and businesses in strategic decision-making.
- ❖ Data Science is a field of study that entails applying several scientific methods, algorithms, and processes to extract insights from massive amounts of data.
- ❖ In other words it is broadly defined as the capability of a machine to imitate intelligent human behaviour.

❖ COMPANY PROFILE

- Technofly was formed by professionals with formal qualifications and industrial experience in the fields of embedded systems, real-time software, process control and industrial electronics.

- The company is professionally managed and supported by qualified experienced specialists and consultants with experience in embedded systems - including hardware and software.

- Technofly Developed system software tools; these include C Compilers for micro-controllers and other supporting tools such as assembler, linker, simulator and Integrated Development Environment. Later Single Board Computers (SBCs) - were Developed and are still manufactured. Such hardware boards support a broad range of processors - including 8 bits, 16- and 32-bit processor. Since 2015, company also started offering design and development services

❖ PROPOSED SYSTEM

1. Data Collection
2. Data Cleaning
3. Exploratory Data Analysis (EDA)
4. Modeling and Forecasting
5. Interpretation and Reporting

➢ DATA COLLECTION

- Collecting data from a CSV file.
- Government economic reports.
- Population , Literacy rate , GDP ($ per capita)
- Birth and death rate

➢ DATA CLEANING

- Data cleaning is also referred to as data preparation, is a vital step that comprises reformatting the data, making data corrections, and merging data sets to enhance the data
- Normalizing data in data science refers to the process of rescaling or transforming your data so that it falls within a specific range or follows a particular distribution

SNAPSHOTS

➢ EXPLORATORY DATA ANALYSIS (EDA)

- Bar plot is a graphical representation of data using rectangular bars or columns. Bar plots are used to display categorical data and compare the values of different categories or groups.
- Heatmap is a graphical representation of data where individual values are represented as colors. Heatmaps are particularly useful for visualizing and exploring the patterns and relationships within large and complex datasets.
- Correlation analysis is a statistical method used to evaluate the strength and direction of the relationship between two or more variables. It measures the extent to which changes in one variable are associated with changes in another variable.

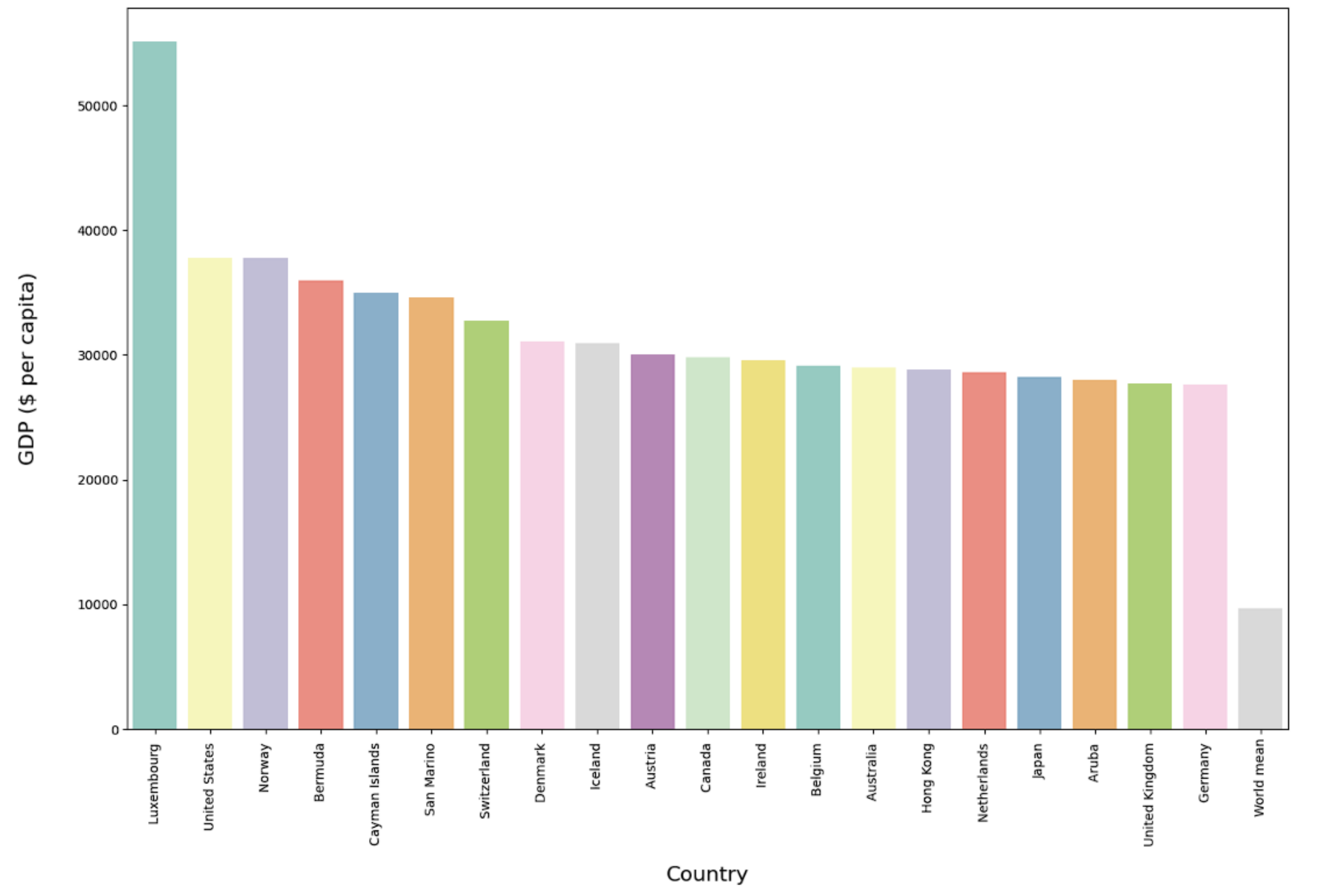
➢ PREDICTIVE MODELS

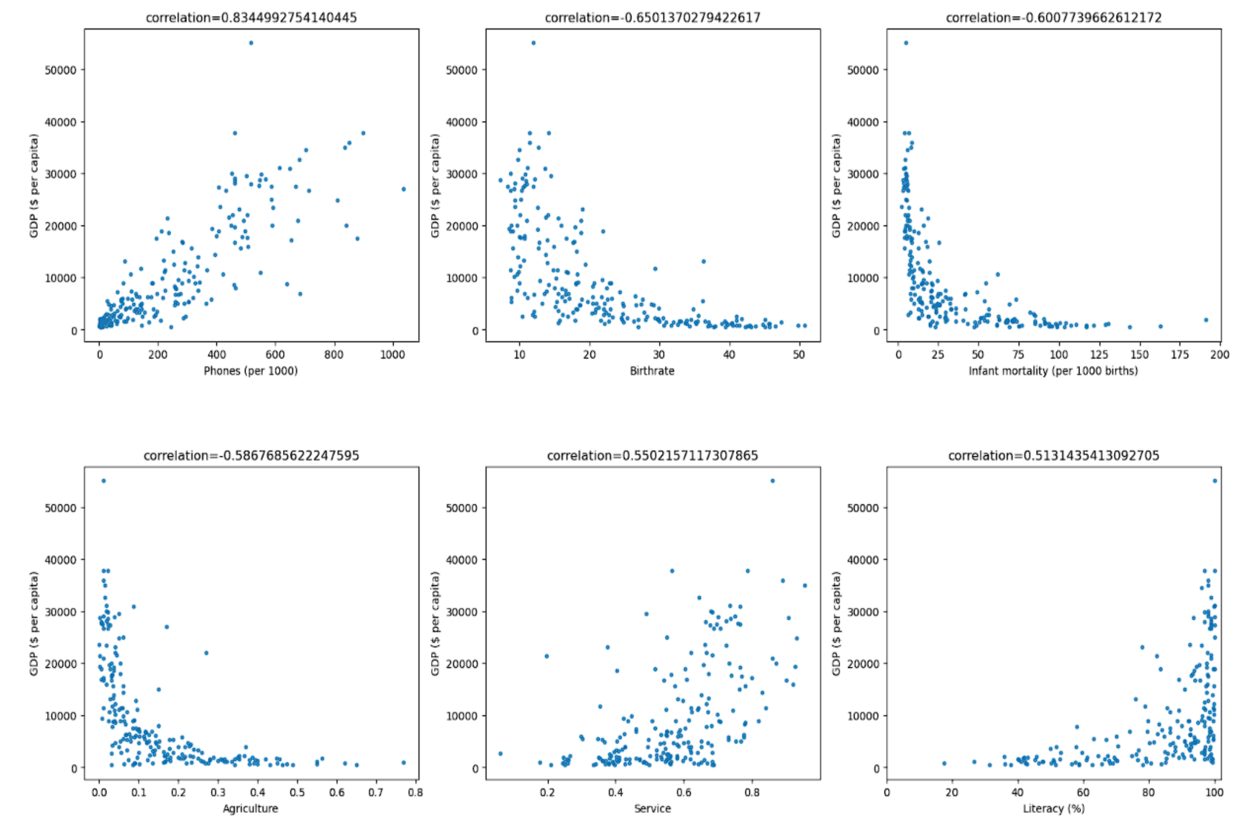
- Regression models are a class of statistical models used in data analysis and machine learning to explore and quantify the relationship between one or more independent variables (predictors or features) and a dependent variable (the target or outcome). The primary goal of regression analysis is to understand how changes in the independent variables are associated with changes in the dependent variable.
- Machine learning algorithms are computational methods or procedures that enable computers and machines to learn from data and make predictions or decisions without being explicitly programmed to do so. These algorithms are a fundamental component of machine learning, a subfield of artificial intelligence (AI).

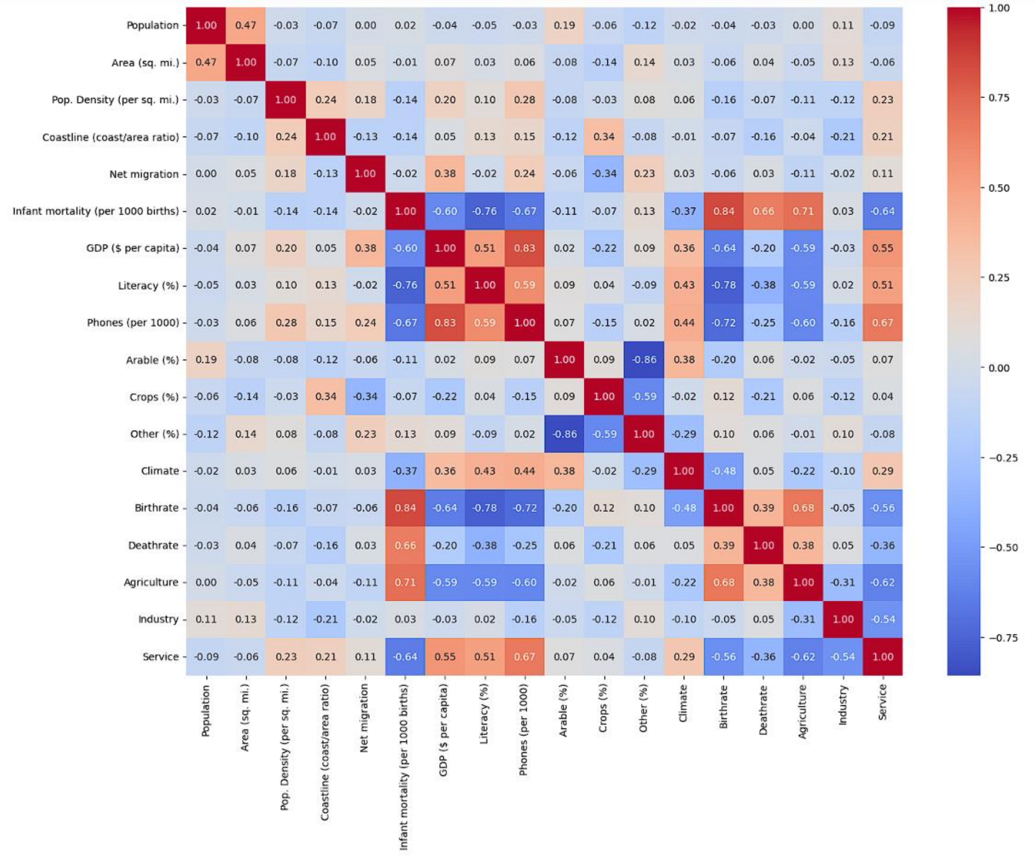
➢ INTERPRETATION AND REPORTING

- Analysing model results refers to the process of evaluating and interpreting the outcomes produced by a machine learning or statistical model. This step is crucial in understanding the model's performance, assessing its effectiveness, and gaining insights into the underlying patterns in the data.
- Making forecasts refers to the process of predicting future events, trends, or outcomes based on historical data, patterns, and models. Forecasting is a valuable tool used in various fields, including finance, economics, weather science, supply chain management, and many others, to make informed decisions and plans for the future
- Preparing reports involves the process of compiling, organizing, and presenting information in a structured and coherent manner to convey findings, analysis, or results to a specific audience

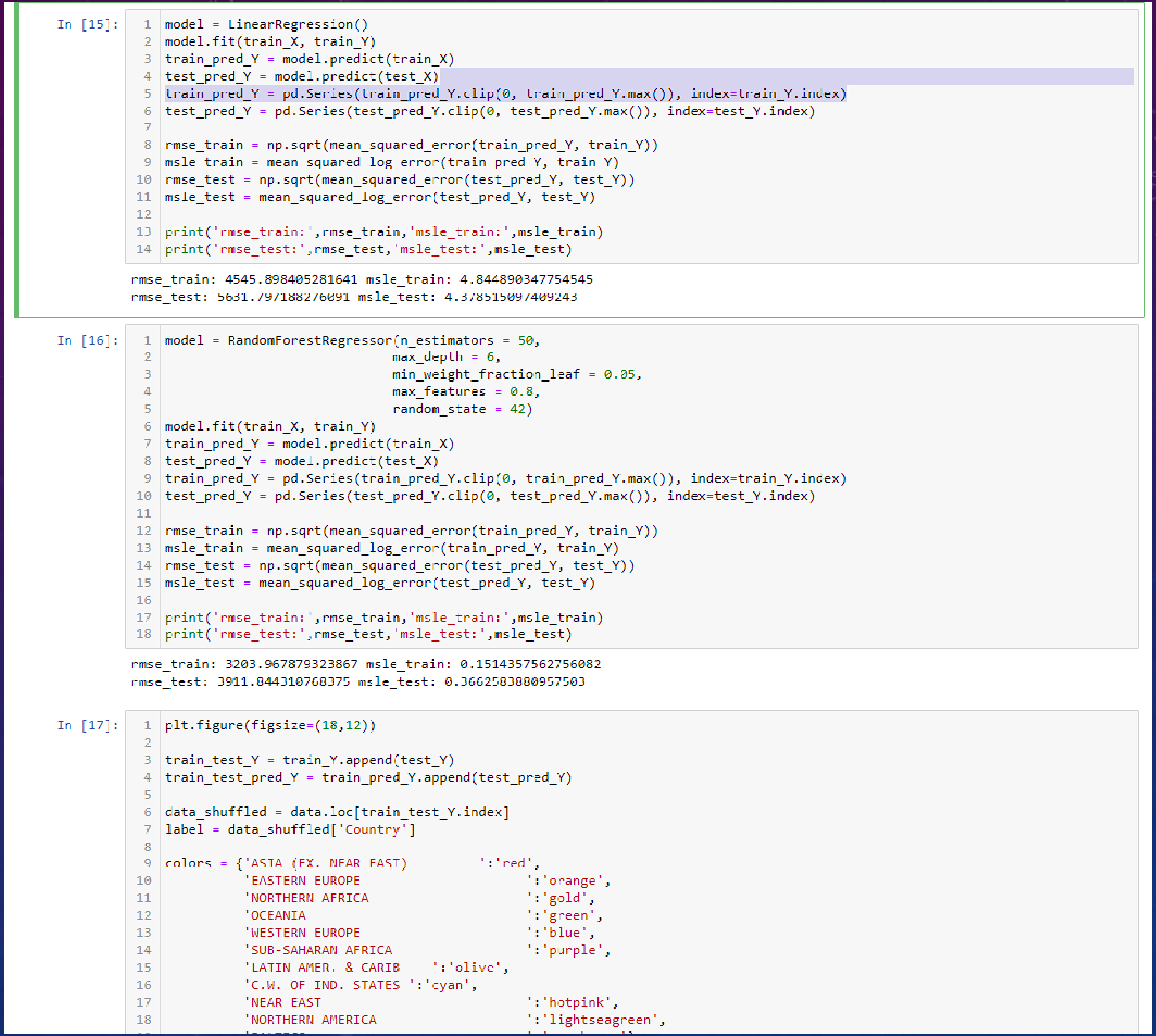
RESULTS


GDP ($ per capita)
50000
40000
30000
20000
10000
0
Luxembourg
United States
Norway
Bermuda
Cayman Islands
San Marino
Switzerland
Denmark
Iceland
Austria
Canada
Ireland
Belgium
Australia
Hong Kong
Netherlands
Japan
Aruba
United Kingdom
Germany
World mean
Country





	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (%)	Crops (%)	Other (%)	Climate	Birthrate	Deathrate	Agriculture	Industry	Service
Population	1.00	0.47	-0.03	-0.07	0.00	0.02	-0.04	-0.05	-0.03	0.19	-0.06	-0.12	-0.02	-0.04	-0.03	0.00	0.11	-0.09
Area (sq. mi.)	0.47	1.00	-0.07	-0.10	0.05	-0.01	0.07	0.03	0.06	-0.08	-0.14	0.14	0.03	-0.06	0.04	-0.05	0.13	-0.06
Pop. Density (per sq. mi.)	-0.03	-0.07	1.00	0.24	0.18	-0.14	0.20	0.10	0.28	-0.08	-0.03	0.08	0.06	-0.16	-0.07	-0.11	-0.12	0.23
Coastline (coast/area ratio)	-0.07	-0.10	0.24	1.00	-0.13	-0.14	0.05	0.13	0.15	-0.12	0.34	-0.08	-0.01	-0.07	-0.16	-0.04	-0.21	0.21
Net migration	0.00	0.05	0.18	-0.13	1.00	-0.02	0.38	-0.02	0.24	-0.06	-0.34	0.23	0.03	-0.06	0.03	-0.11	-0.02	0.11
Infant mortality (per 1000 births)	0.02	-0.01	-0.14	-0.14	-0.02	1.00	-0.60	-0.76	-0.67	-0.11	-0.07	0.13	-0.37	0.84	0.66	0.71	0.03	-0.64
GDP ($ per capita)	-0.04	0.07	0.20	0.05	0.38	-0.60	1.00	0.51	0.83	0.02	-0.22	0.09	0.36	-0.64	-0.20	-0.59	-0.03	0.55
Literacy (%)	-0.05	0.03	0.10	0.13	-0.02	-0.76	0.51	1.00	0.59	0.09	0.04	-0.09	0.43	-0.78	-0.38	-0.59	0.02	0.51
Phones (per 1000)	-0.03	0.06	0.28	0.15	0.24	-0.67	0.83	0.59	1.00	0.07	-0.15	0.02	0.44	-0.72	-0.25	-0.60	-0.16	0.67
Arable (%)	0.19	-0.08	-0.08	-0.12	-0.06	-0.11	0.02	0.09	0.07	1.00	0.09	-0.86	0.38	-0.20	0.06	-0.02	-0.05	0.07
Crops (%)	-0.06	-0.14	-0.03	0.34	-0.34	-0.07	-0.22	0.04	-0.15	0.09	1.00	-0.59	-0.02	0.12	-0.21	0.06	-0.12	0.04
Other (%)	-0.12	0.14	0.08	-0.08	0.23	0.13	0.09	-0.09	0.02	-0.86	-0.59	1.00	-0.29	0.10	0.06	-0.01	0.10	-0.08
Climate	-0.02	0.03	0.06	-0.01	0.03	-0.37	0.36	0.43	0.44	0.38	-0.02	-0.29	1.00	-0.48	0.05	-0.22	-0.10	0.29
Birthrate	-0.04	-0.06	-0.16	-0.07	-0.06	0.84	-0.64	-0.78	-0.72	-0.20	0.12	0.10	-0.48	1.00	0.39	0.68	-0.05	-0.56
Deathrate	-0.03	0.04	-0.07	-0.16	0.03	0.66	-0.20	-0.38	-0.25	0.06	-0.21	0.06	0.05	0.39	1.00	0.38	0.05	-0.36
Agriculture	0.00	-0.05	-0.11	-0.04	-0.11	0.71	-0.59	-0.59	-0.60	-0.02	0.06	-0.01	-0.22	0.68	0.38	1.00	-0.31	-0.62
Industry	0.11	0.13	-0.12	-0.21	-0.02	0.03	-0.03	0.02	-0.16	-0.05	-0.12	0.10	-0.10	-0.05	0.05	-0.31	1.00	-0.54
Service	-0.09	-0.06	0.23	0.21	0.11	-0.64	0.55	0.51	0.67	0.07	0.04	-0.08	0.29	-0.56	-0.36	-0.62	-0.54	1.00



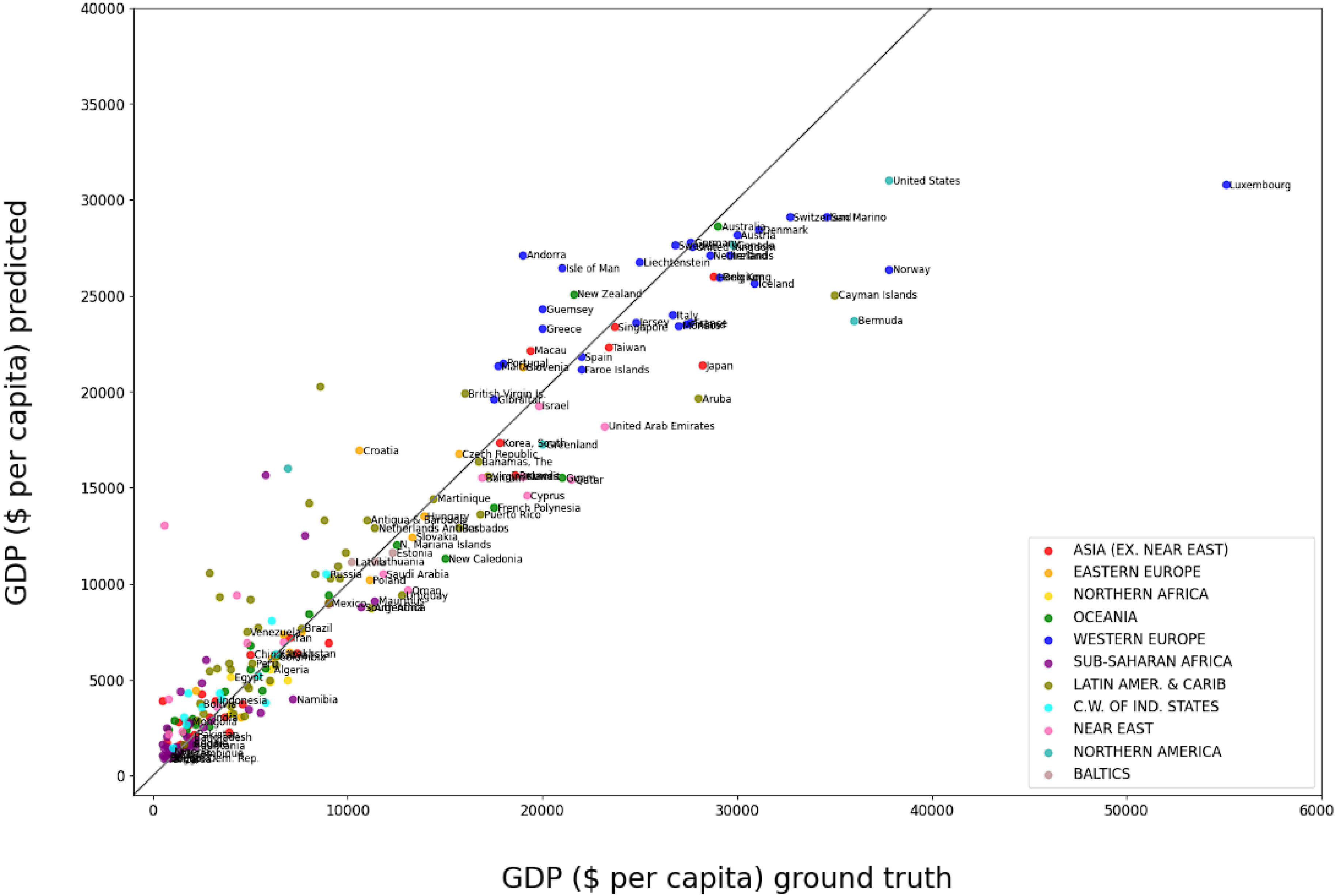
```python
model = LinearRegression()
model.fit(train_X, train_Y)
train_pred_Y = model.predict(train_X)
test_pred_Y = model.predict(test_X)
train_pred_Y = pd.Series(train_pred_Y.clip(0, train_pred_Y.max()), index=train_Y.index)
test_pred_Y = pd.Series(test_pred_Y.clip(0, test_pred_Y.max()), index=test_Y.index)

rmse_train = np.sqrt(mean_squared_error(train_pred_Y, train_Y))
msle_train = mean_squared_log_error(train_pred_Y, train_Y)
rmse_test = np.sqrt(mean_squared_error(test_pred_Y, test_Y))
msle_test = mean_squared_log_error(test_pred_Y, test_Y)

print('rmse_train:',rmse_train,'msle_train:',msle_train)
print('rmse_test:',rmse_test,'msle_test:',msle_test)
```

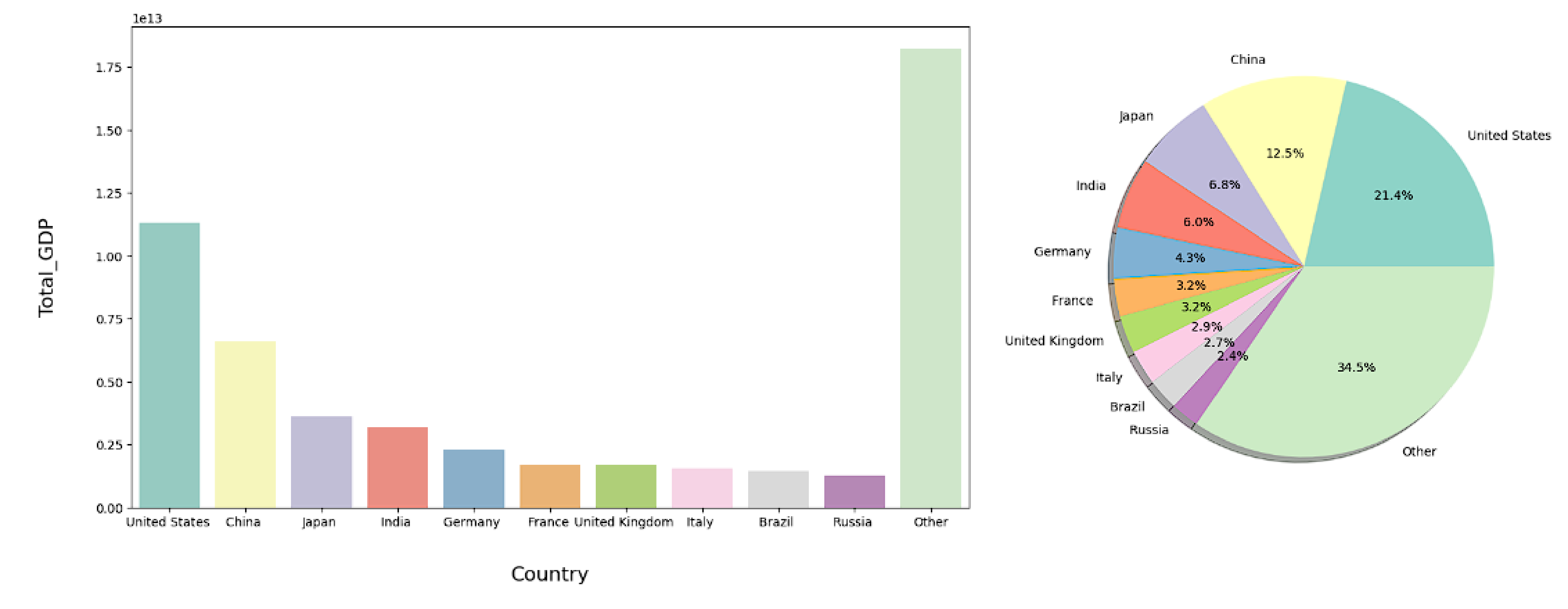
```
rmse_train: 4545.898405281641 msle_train: 4.844890347754545
rmse_test: 5631.797188276091 msle_test: 4.378515097409243
```

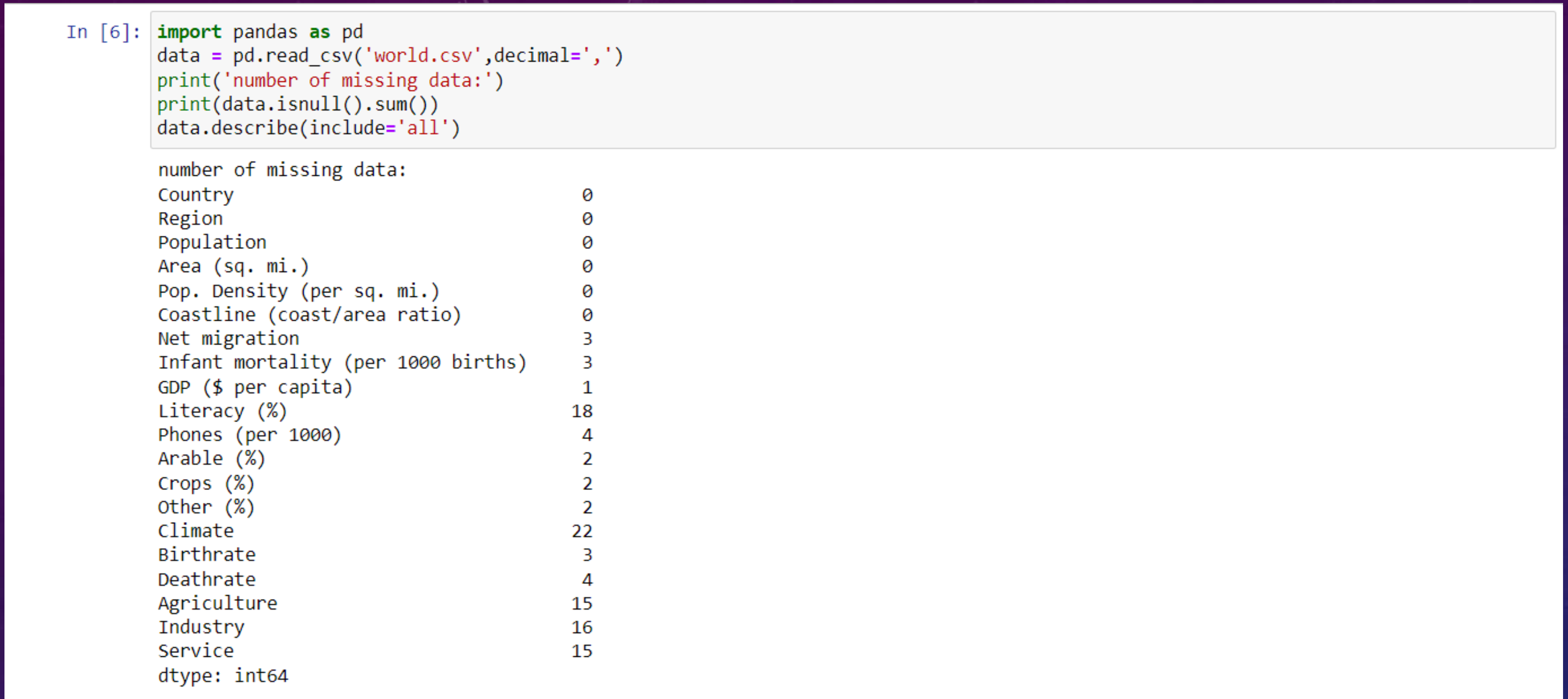
```python
model = RandomForestRegressor(n_estimators = 50,
                             max_depth = 6,
                             min_weight_fraction_leaf = 0.05,
                             max_features = 0.8,
                             random_state = 42)
model.fit(train_X, train_Y)
train_pred_Y = model.predict(train_X)
test_pred_Y = model.predict(test_X)
train_pred_Y = pd.Series(train_pred_Y.clip(0, train_pred_Y.max()), index=train_Y.index)
test_pred_Y = pd.Series(test_pred_Y.clip(0, test_pred_Y.max()), index=test_Y.index)

rmse_train = np.sqrt(mean_squared_error(train_pred_Y, train_Y))
msle_train = mean_squared_log_error(train_pred_Y, train_Y)
rmse_test = np.sqrt(mean_squared_error(test_pred_Y, test_Y))
msle_test = mean_squared_log_error(test_pred_Y, test_Y)

print('rmse_train:',rmse_train,'msle_train:',msle_train)
print('rmse_test:',rmse_test,'msle_test:',msle_test)
```

```
rmse_train: 3203.967879323867 msle_train: 0.1514357562756082
rmse_test: 3911.844310768375 msle_test: 0.3662583880957503
```

```python
plt.figure(figsize=(18,12))

train_test_Y = train_Y.append(test_Y)
train_test_pred_Y = train_pred_Y.append(test_pred_Y)

data_shuffled = data.loc[train_test_Y.index]
label = data_shuffled['Country']

colors = {'ASIA (EX. NEAR EAST)         ':'red',
          'EASTERN EUROPE                     ':'orange',
          'NORTHERN AFRICA                    ':'gold',
          'OCEANIA                            ':'green',
          'WESTERN EUROPE                     ':'blue',
          'SUB-SAHARAN AFRICA                 ':'purple',
          'LATIN AMER. & CARIB    ':'olive',
          'C.W. OF IND. STATES ':'cyan',
          'NEAR EAST                          ':'hotpink',
          'NORTHERN AMERICA                   ':'lightseagreen',
```


GDP ($ per capita) predicted
GDP ($ per capita) ground truth
40000
35000
30000
25000
20000
15000
10000
5000
0
0
10000
20000
30000
40000
50000
60000
● ASIA (EX. NEAR EAST)
● EASTERN EUROPE
● NORTHERN AFRICA
● OCEANIA
● WESTERN EUROPE
● SUB-SAHARAN AFRICA
● LATIN AMER. & CARIB
● C.W. OF IND. STATES
● NEAR EAST
● NORTHERN AMERICA
● BALTICS
United States
Luxembourg
Switzerland
San Marino
Australia
Denmark
Austria
Germany
United Kingdom
Netherlands
Andorra
Isle of Man
Liechtenstein
Norway
Hong Kong
Belgium
Iceland
New Zealand
Cayman Islands
Guernsey
Italy
Jersey
France
Monaco
Singapore
Bermuda
Greece
Macau
Taiwan
Spain
Portugal
Malta
Slovenia
Japan
Faroe Islands
British Virgin Is.
Gibraltar
Israel
Aruba
United Arab Emirates
Croatia
Korea, South
Greenland
Czech Republic
Bahamas, The
Bahrain
Guam
Qatar
Cyprus
Martinique
French Polynesia
Puerto Rico
Hungary
Antigua & Barbuda
Netherlands Antilles
Barbados
Slovakia
N. Mariana Islands
Estonia
New Caledonia
Latvia
Lithuania
Russia
Saudi Arabia
Poland
Oman
Uruguay
Mauritius
Mexico
South Africa
Argentina
Brazil
Venezuela
Iran
China
Kazakhstan
Colombia
Peru
Algeria
Egypt
Namibia
Indonesia
Bolivia
India
Mongolia
Pakistan
Bangladesh
Tanzania
Mozambique
Congo, Dem. Rep.

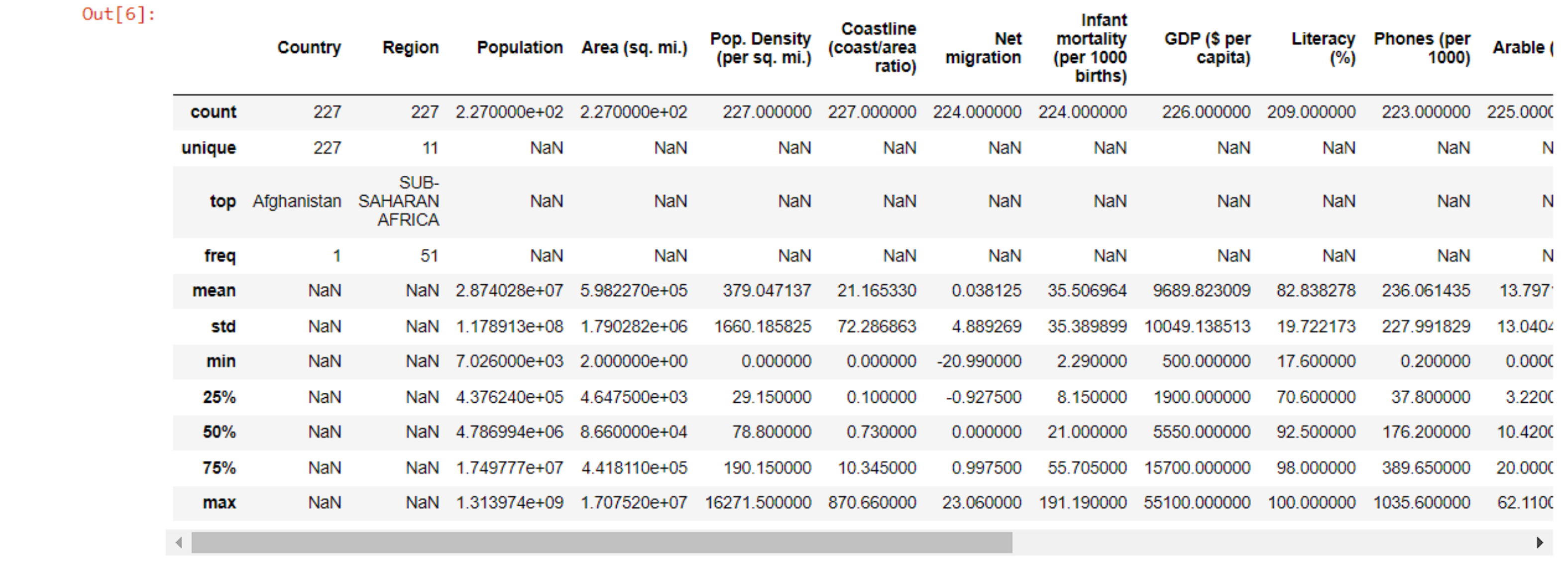

1e13
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Total_GDP
United States
China
Japan
India
Germany
France
United Kingdom
Italy
Brazil
Russia
Other
Country
China
Japan
United States
India
Germany
France
United Kingdom
Italy
Brazil
Russia
Other
12.5%
6.8%
21.4%
6.0%
4.3%
3.2%
3.2%
2.9%
2.7%
2.4%
34.5%



In [6]:

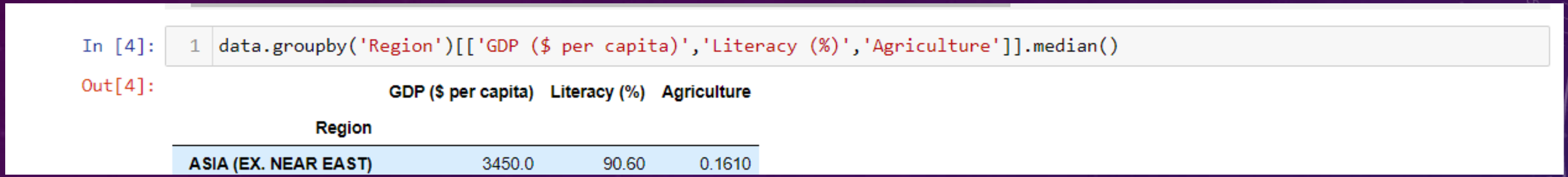
```python
import pandas as pd
data = pd.read_csv('world.csv',decimal=',')
print('number of missing data:')
print(data.isnull().sum())
data.describe(include='all')
```

```
number of missing data:
Country                               0
Region                                0
Population                            0
Area (sq. mi.)                        0
Pop. Density (per sq. mi.)            0
Coastline (coast/area ratio)          0
Net migration                         3
Infant mortality (per 1000 births)    3
GDP ($ per capita)                    1
Literacy (%)                         18
Phones (per 1000)                     4
Arable (%)                            2
Crops (%)                             2
Other (%)                             2
Climate                              22
Birthrate                             3
Deathrate                             4
Agriculture                          15
Industry                             16
Service                              15
dtype: int64
```

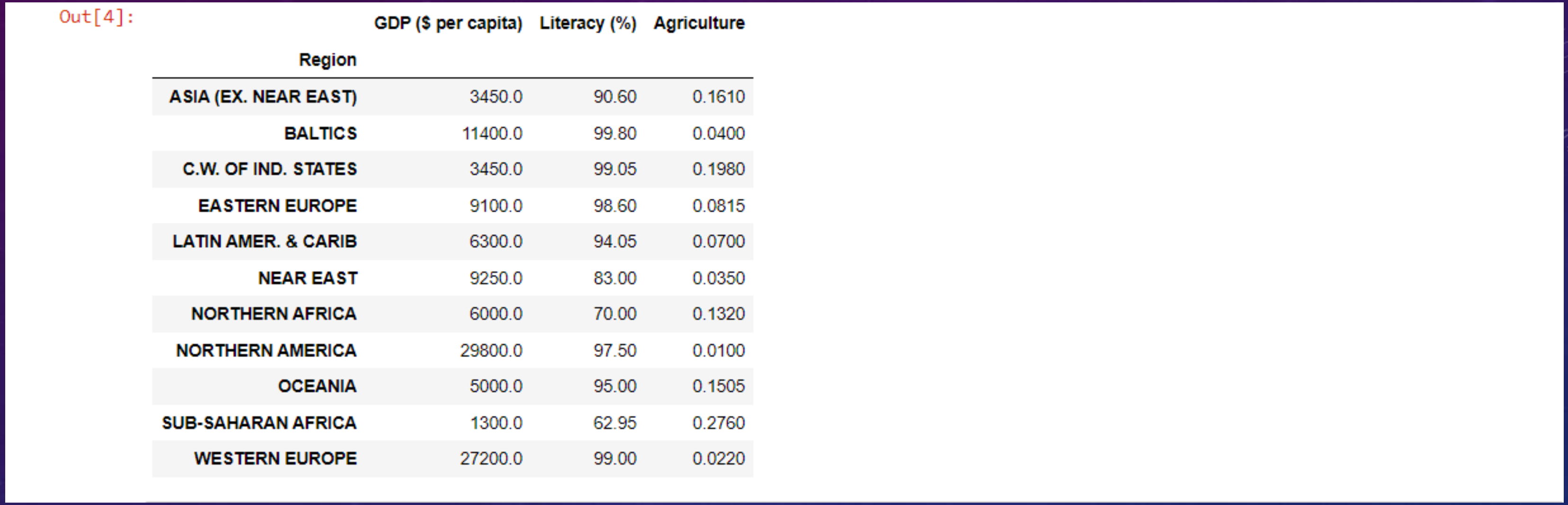


Out[6]:

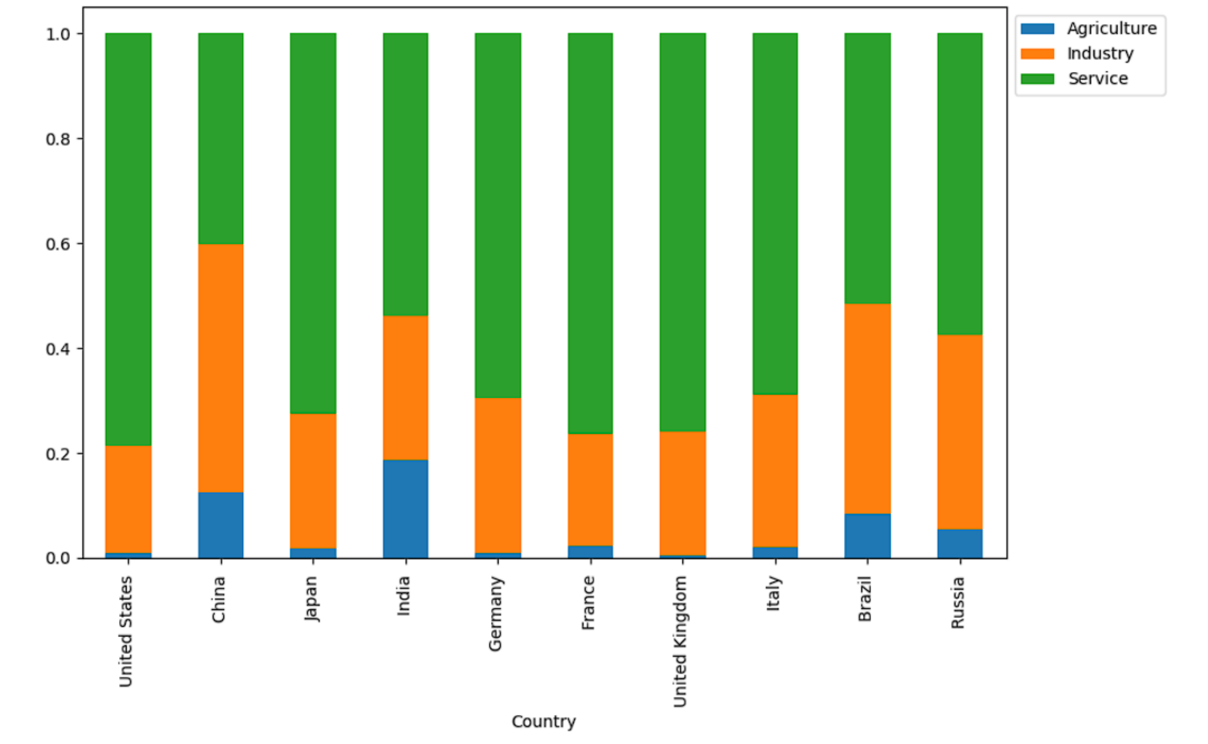
	Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP ($ per capita)	Literacy (%)	Phones (per 1000)	Arable (
count	227	227	2.270000e+02	2.270000e+02	227.000000	227.000000	224.000000	224.000000	226.000000	209.000000	223.000000	225.0000
unique	227	11	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
top	Afghanistan	SUB-SAHARAN AFRICA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
freq	1	51	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
mean	NaN	NaN	2.874028e+07	5.982270e+05	379.047137	21.165330	0.038125	35.506964	9689.823009	82.838278	236.061435	13.797
std	NaN	NaN	1.178913e+08	1.790282e+06	1660.185825	72.286863	4.889269	35.389899	10049.138513	19.722173	227.991829	13.040
min	NaN	NaN	7.026000e+03	2.000000e+00	0.000000	0.000000	-20.990000	2.290000	500.000000	17.600000	0.200000	0.0000
25%	NaN	NaN	4.376240e+05	4.647500e+03	29.150000	0.100000	-0.927500	8.150000	1900.000000	70.600000	37.800000	3.2200
50%	NaN	NaN	4.786994e+06	8.660000e+04	78.800000	0.730000	0.000000	21.000000	5550.000000	92.500000	176.200000	10.4200
75%	NaN	NaN	1.749777e+07	4.418110e+05	190.150000	10.345000	0.997500	55.705000	15700.000000	98.000000	389.650000	20.0000
max	NaN	NaN	1.313974e+09	1.707520e+07	16271.500000	870.660000	23.060000	191.190000	55100.000000	100.000000	1035.600000	62.1100

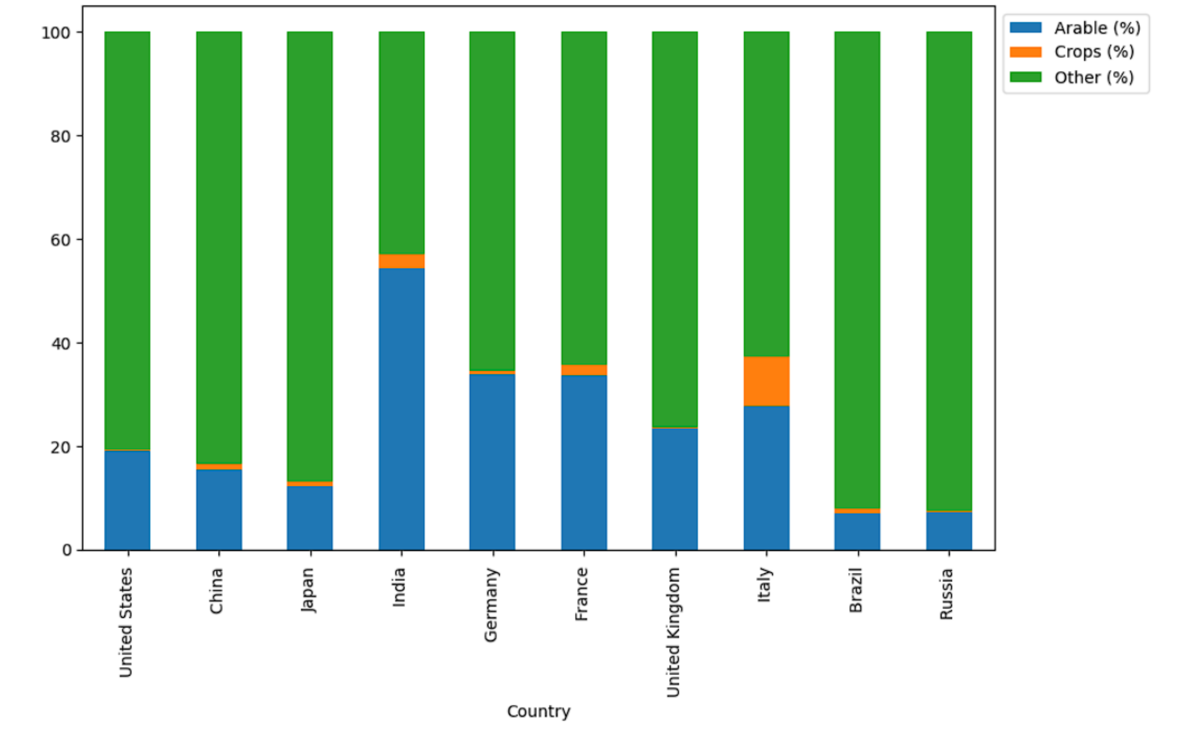


```
In [4]:  data.groupby('Region')[['GDP ($ per capita)','Literacy (%)','Agriculture']].median()
Out[4]:
                       GDP ($ per capita)  Literacy (%)  Agriculture
               Region
ASIA (EX. NEAR EAST)               3450.0         90.60       0.1610
```



```
Out[4]:
                       GDP ($ per capita)  Literacy (%)  Agriculture
               Region
ASIA (EX. NEAR EAST)               3450.0         90.60       0.1610
             BALTICS              11400.0         99.80       0.0400
C.W. OF IND. STATES                3450.0         99.05       0.1980
      EASTERN EUROPE               9100.0         98.60       0.0815
 LATIN AMER. & CARIB               6300.0         94.05       0.0700
           NEAR EAST               9250.0         83.00       0.0350
    NORTHERN AFRICA                6000.0         70.00       0.1320
   NORTHERN AMERICA               29800.0         97.50       0.0100
             OCEANIA               5000.0         95.00       0.1505
  SUB-SAHARAN AFRICA               1300.0         62.95       0.2760
      WESTERN EUROPE              27200.0         99.00       0.0220
```


1.0
0.8
0.6
0.4
0.2
0.0
Agriculture
Industry
Service
United States
China
Japan
India
Germany
France
United Kingdom
Italy
Brazil
Russia
Country


100
80
60
40
20
0
Arable (%)
Crops (%)
Other (%)
United States
China
Japan
India
Germany
France
United Kingdom
Italy
Brazil
Russia
Country

❖ CONCLUSION

In conclusion, our journey through GDP analysis using data science has provided us with valuable insights into the economic performance of our nation. We've seen how data science techniques can transform raw economic data into actionable information for policymakers, businesses, and individuals. The world is always changing, and we can keep using data science to understand it better. Our hope is that this helps everyone, from the government to businesses and individuals, make smarter decisions for a brighter economic future.

❖ APPLICATIONS

- **Economic Policy Formulation:** Governments use GDP analysis to design economic policies. By understanding the current economic situation and predicting future trends, policymakers can make informed decisions about taxation, spending, and interest rates to promote economic growth and stability.
- **Business Strategy:** Companies use GDP data to make strategic decisions. For instance, retailers may use GDP forecasts to plan inventory levels, and investment firms may adjust their portfolios based on economic outlooks
- **International Trade:** Exporters and importers use GDP data to identify markets with growth potential. A country with a rising GDP may represent an attractive export destination
- **Entrepreneurship Opportunities:** Entrepreneurs can spot business opportunities by analyzing GDP trends. For example, during economic expansion, there may be increased demand for new services or products.

❖ REFERENCES

- https://www.w3schools.com/datascience/
- https://www.investopedia.com/terms/p/per-capita-gdp.asp
- https://www.guru99.com/data-science-tutorial.html

THANK YOU