DSC680 12.1 Project 3: Presentation/Milestone 3

Report/White Paper

Joseph Madden

1. **Background:**

The CIA World Factbook provides a comprehensive overview of countries worldwide, including economic indicators such as GDP, population demographics, health, education, and environmental factors. Analyzing this data can provide insights into global economic disparities in global economic development.

2. **Business Problem:**

This project aims to analyze and visualize economic indicators from "The world Factbook" dataset to identify patterns and disparities in global economic development. Key research questions include understanding the factors influencing economic growth, uncovering potential correlations, and assessing the impact of socioeconomic variables.

- The tasks consist of Data Loading: Importing "The World Factbook" dataset using pandas' library. Exploratory Data Analysis (EDA): Statistical summary and visualization of key economic indicators. Data Preprocessing: Cleaning data, handling missing values, and encoding categorical variables. Correlation Analysis: Identifying relationships between economic indicators. Cluster Analysis: Grouping countries based on economic indicators.
- The dataset consists of information on various countries across different economic indicators, such as GDP, population demographics, health, education, and environmental factors. Each country is represented by a set of attributes reflected its economic and social characteristics.

- The performance metrics for the analysis include correlation coefficients: Measure the strength and direction of relationships between economic indicators. Cluster quality metrics: Evaluate the quality of clustering results. Visualizing effectiveness: Assess the clarity and insightfulness of data visualizations in conveying patterns and disparities in global economic development.

1. **Data Explanation:**

- Data Loading: The analysis begins with importing data from 'The World Factbook" dataset using pandas' library. This dataset contains information on various countries across different economic indicators.
- Data Dictionary: The dataset includes attributes such as GDP, population demographics, health, education, and environmental factors for each country. The detailed data dictionary is given below.
    1. "**Country Name**" Name of the country
    2. "**GDP (Gross Domestic Product)**" The total monetary value of all goods and services produced within a country's borders in a specific time period, usually annually.
    3. "**Population**" The total number of people living in a country.
    4. "**Population Growth Rate**" The annual percentage change in population, indicating the rate at which the population is growing.
    5. "**Life Expectancy at Birth**" The average number of years a newborn is expected to live, based on currency mortality rates.

6. “**Infant Mortality Rate**” The number of deaths of infants under one year old per 1,000 live births each year.
7. “**Gini Index**” A measure of income inequality within a population, with higher values indicating greater inequality.
8. “**Unemployment Rate**” The percentage of the labor force that is unemployed and actively seeking employment.
9. “**Labor Force**” The total number of people who are employed or actively seeking employment.
10. “**Exports**” The total value of goods and services that a country sells to other countries.
11. “**Imports**” The total value of goods and services that a country buys from other countries.
12. “**Budget**” The government’s total revenues (taxes and other income) minus its total expenditures.
13. “**Debt**” The total amount of money that a country owes to its creditors.
14. “**Electricity Production**” The total amount of electricity generated within a country.
15. “**Internet Users**” The total number of people who use the internet in a country.
16. “**Telephones-Mobile Cellular**” The total number of mobile cellular telephone subscriptions in a country.
17. “**Railways**” The total length of railway lines in a country.
18. “**Roadways**” The total length of roadways in a country.

19. “**Airports**” The total number of airports in a country.
20. “**Military Expenditure**” The total amount of money spent by a country on its military.
21. “**Environmental Issues**” Various environmental issues faced by a country, such as air pollution, water pollution, and deforestation.
22. “**Target Variables**” In the context of exploring global economic disparities using “The world Factbook dataset, the target variable is not explicitly defined in the dataset itself. The target variable would depend on the specific analysis question being addressed.

2. Methods & Analysis:

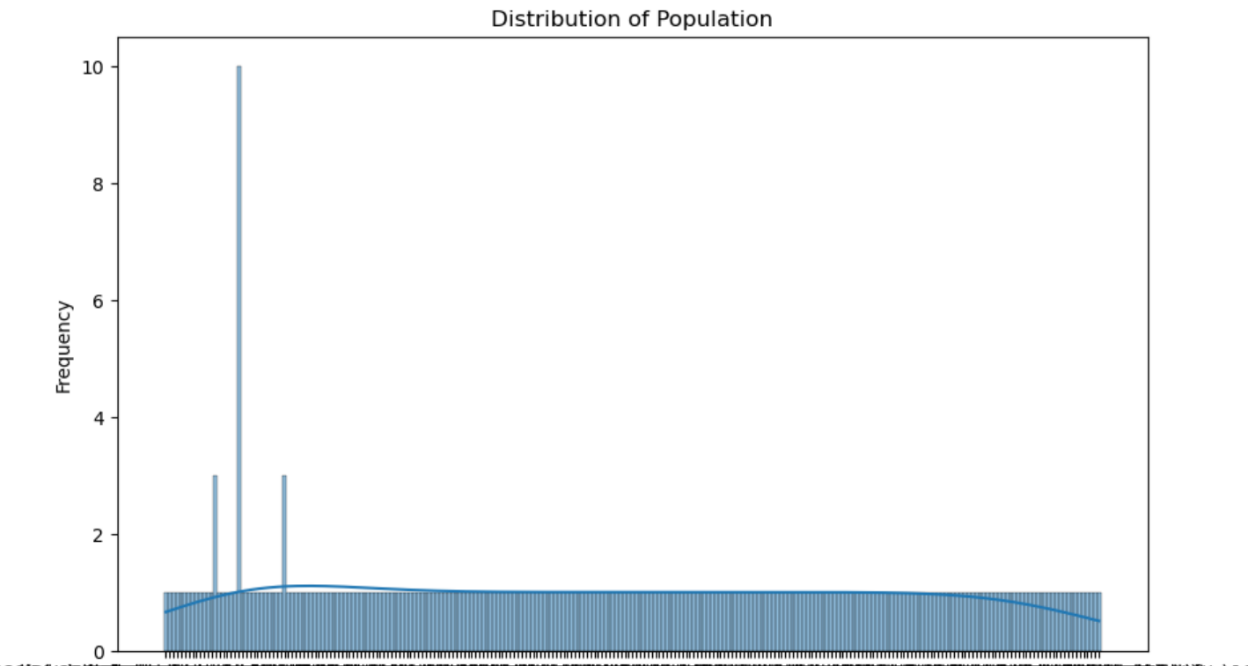
2.1 Exploratory Data Analysis (EDA)

Let’s see the statistical summary of the data.

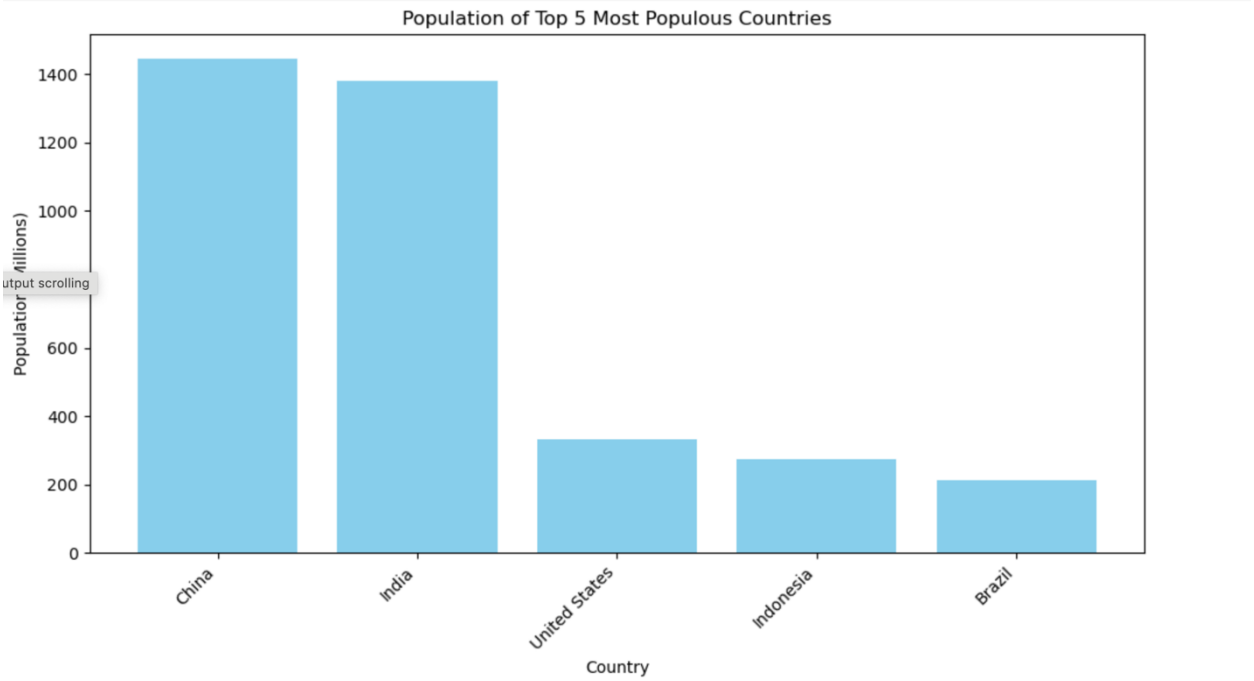
- Statistical summary of the data.
- Visualization of key economic indicators.

3. Conclusion:

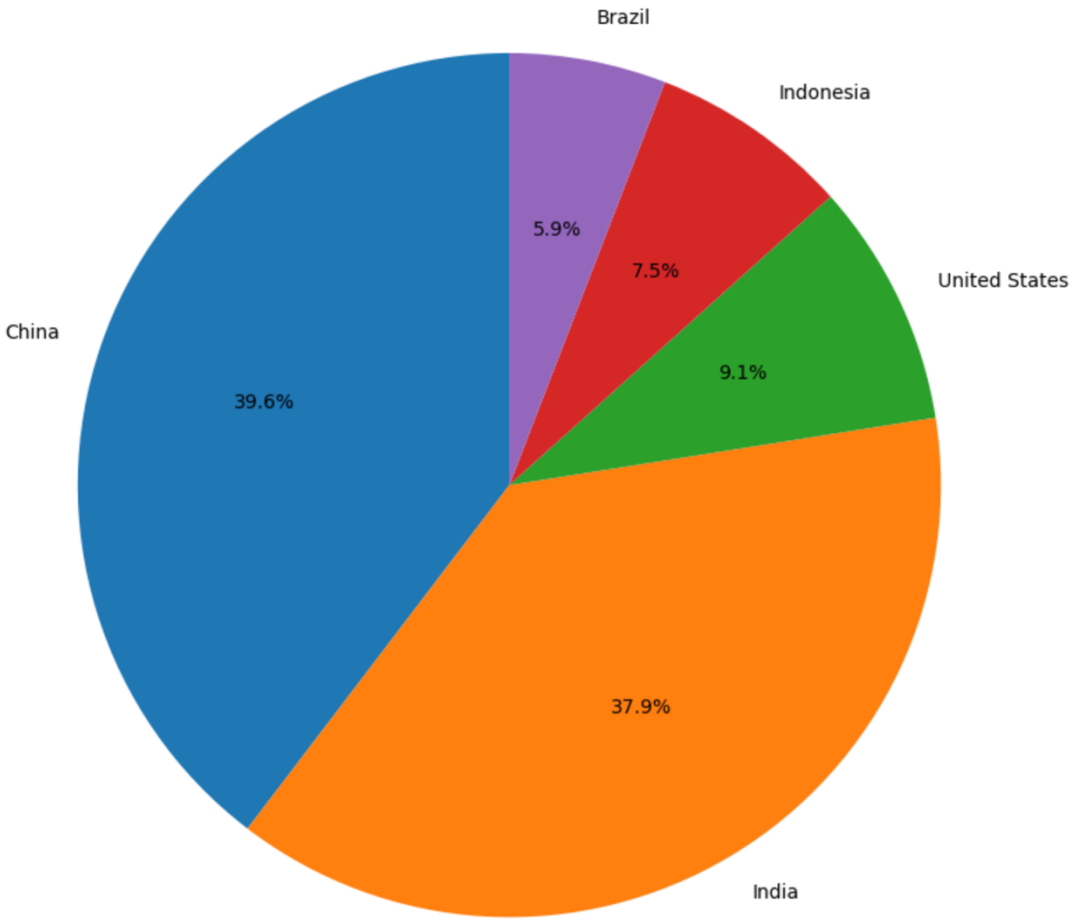
In conclusion, this study has provided valuable insights into the global economic landscape, highlighting significant disparities among nations. Through the analysis of CIA World Factbook data.



This is a univariate analysis on some key variables from The World Factbook dataset to gain insights into their distributions and characteristics.



This is a bar chart showing the top 5 most populous countries according to the dataset The World Factbook.



This is the pie chart showing the top 5 most populous countries according to The World Factbook.

3.1 Data Preprocessing:

- Data cleaning and handling missing values.
- Encoding categorical variables.

Let's preprocess the dataset, encode categorical features, and split the data into training and test.

```
# Preprocess target variable (Geography: Elevation - mean elevation)
# Handle missing values (if any) appropriately, e.g., impute with mean/median
target_var = "Geography: Elevation - mean elevation"  # Assuming this is the target

# Select categorical features
categorical_features = [
    # Add the categorical features you want to encode here
    # e.g., "Geography: Location", "People and Society: Languages"
]

# Encode categorical features using one-hot encoding
encoder = OneHotEncoder(sparse=False)
encoded_features = encoder.fit_transform(df[categorical_features])

# Concatenate numerical features and encoded categorical features
features = pd.concat(
    [df.drop(categorical_features + [target_var], axis=1), pd.DataFrame(encoded_features)], axis=1
)

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(features, df[target_var], test_size=0.2, random_state=42)

print("X_train shape:", X_train.shape)
print("X_test shape:", X_test.shape)
print("y_train shape:", y_train.shape)
print("y_test shape:", y_test.shape)
```

```
X_train shape: (206, 1070)
X_test shape: (52, 1070)
y_train shape: (206,)
y_test shape: (52,)
```

3.2 Modeling

- Correlation analysis to identify relationships between variables.
- Cluster analysis to group countries based on economic indicators.

```
# Check for missing values
print(df.isnull().sum())

# Impute missing values using appropriate methods (e.g., mean/median imputation)
# ...
```

```
Country                                                          0
url                                                              0
Introduction: Background                                         0
Geography: Location                                              2
Geography: Geographic coordinates                               11
                                                               ...
Transnational Issues: Trafficking in persons - Tier 2 Watch List   257
Transnational Issues: Trafficking in persons - Tier 3              257
Transnational Issues: Illicit drugs - cocaine                      257
Transnational Issues: Illicit drugs - opiates                      257
Energy: Electricity access - population without electricity        257
Length: 1071, dtype: int64
```

```python
# ... (pre-processing steps mentioned in previous response)

# Print data types and check for non-numeric values
print(X_train.dtypes)

# Handle missing values or non-numeric data as needed
# ...

# Ensure consistent data types in y_train (adjust for classification/regression)
# ...

# Define the ANN model based on data types and task (classification/regression)
# ...

model.compile(loss="...", optimizer="adam", metrics=["accuracy"])  # Adjust loss and activation for task

# Train the model
model.fit(X_train, y_train, epochs=10, batch_size=32)  # Adjust hyperparameters

# Evaluate the model
loss, accuracy = model.evaluate(X_test, y_test)
print("Test accuracy:", accuracy)
```

```
Country                                                          object
url                                                              object
Introduction: Background                                         object
Geography: Location                                              object
Geography: Geographic coordinates                                object
                                                                  ...
Transnational Issues: Trafficking in persons - Tier 2 Watch List   object
Transnational Issues: Trafficking in persons - Tier 3              object
Transnational Issues: Illicit drugs - cocaine                      object
Transnational Issues: Illicit drugs - opiates                      object
Energy: Electricity access - population without electricity        object
Length: 1071, dtype: object
```

```python
import pandas as pd

df = pd.read_csv("countries.csv")

# Print basic information about the data
print(df.head())
print(df.info())
```

```
                    Geography: Area - comparative  ...  \
0  almost six times the size of Virginia; slightl...  ...
1         about 0.7 times the size of Washington, DC  ...
2                   slightly smaller than Maryland  ...
3     slightly less than 3.5 times the size of Texas  ...
4                slightly larger than Washington, DC  ...

  Transportation: Waterways - note 2  \
0                                NaN
1                                NaN
2                                NaN
3                                NaN
4                                NaN

  Transportation: Waterways - top ten largest natural lakes (by surface area)  \
0                                                NaN
1                                                NaN
2                                                NaN
3                                                NaN
4                                                NaN

  Transportation: Waterways - note 3  \
0                                NaN
1                                NaN
2                                NaN
3                                NaN
4                                NaN

  Transportation: Ports and terminals - top twenty container ports as measured by Twenty-Foot Equivalent Units (TEUs) throughput  \
0                                                NaN
1                                                NaN
2                                                NaN
3                                                NaN
4                                                NaN
```

```
  Transnational Issues: Refugees and internally displaced persons  \
0                                                  NaN
1                                                  NaN
2                                                  NaN
3                                                  NaN
4                                                  NaN

  Transnational Issues: Trafficking in persons - Tier 2 Watch List  \
0                                                  NaN
1                                                  NaN
2                                                  NaN
3                                                  NaN
4                                                  NaN

  Transnational Issues: Trafficking in persons - Tier 3  \
0                                                  NaN
1                                                  NaN
2                                                  NaN
3                                                  NaN
4                                                  NaN

  Transnational Issues: Illicit drugs - cocaine  \
0                                          NaN
1                                          NaN
2                                          NaN
3                                          NaN
4                                          NaN

  Transnational Issues: Illicit drugs - opiates  \
0                                          NaN
1                                          NaN
2                                          NaN
3                                          NaN
4                                          NaN
```

```
  Energy: Electricity access – population without electricity
0                                                  NaN
1                                                  NaN
2                                                  NaN
3                                                  NaN
4                                                  NaN

[5 rows x 1071 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 258 entries, 0 to 257
Columns: 1071 entries, Country to Energy: Electricity access – population without electricity
dtypes: float64(5), object(1066)
memory usage: 2.1+ MB
None
```

4. **Assumptions:**

It is assumed that the data from The World Factbook is accurate and representative of the economic conditions in each country. Additionally, the analysis assumes that economic indicators are sufficient to capture the complexity of global economic disparities.

5. **Limitations:**

Limitations include the availability and quality of data, which may vary across countries. Additionally, the analysis may not capture all factors influencing economic disparities, such as political stability or cultural factors.

6. **Challenges:**

Challenges include dealing with missing or incomplete data, ensuring data consistency across countries, and addressing cultural sensitives when interpreting social indicators. Additionally, there may be challenges in representing a diverse range of countries accurately.

7. **Future Uses/Additional Applications:**

Future uses of the analysis could include forecasting economic trends, evaluating the effectiveness of development policies, and comparing economic performance across regions.

8. **Recommendations:**

Recommendations based on the analysis could indicate targeted interventions to address economic disparities, policy reforms to promote inclusive growth, and initiatives to improve data collections and reporting.

9. **Implementation Plan:**

The implementation plan could involve sharing the findings with policymakers, development agencies, and researchers to inform decision-making and research agendas. Additionally, the analysis could be updated regularly to track progress and identify emerging trends.

10. **Ethical Assessment:**

Ethical considerations include ensuring privacy and confidentiality of data, avoiding biases in data interpretation, and transparently communicating findings to stakeholders.

Questions:

1. What are the key economic indicators used in the analysis?
2. How does the analysis identify patterns and disparities in global economic development?
3. What are the assumptions made regarding the accuracy and representativeness of the data?
4. How are missing or incomplete data handled in the analysis?
5. What are the limitations of using economic indicators to measure global economic disparities?
6. How are cultural factors considered in the analysis of social indicators?
7. What are some potential future uses of the analysis beyond identifying economic disparities?
8. How can the findings of the analysis inform policymaking and development efforts?
9. What recommendations are provided based on the analysis?
10. How can the analysis be updated and maintained to track progress and identify emerging trends?

Answers:

1. The key economic indicators used in the analysis include GDP (Gross Domestic Product), GDP per capita, economic growth rate, income distribution, poverty rates, and unemployment rates.

2. The analysis identifies patterns and disparities by comparing economic indicators across countries and regions. It looks for trends such as income inequality, differences in economic growth rates, and disparities in poverty and unemployment rates.

3. The analysis assumes that the data provided by the CIA World Factbook is accurate and representative of the economic conditions in each country. It also assumes that any missing or incomplete data can be estimated or imputed accurately.

4. Missing or incomplete data are handled by either imputing values based on available data or excluding the affected observations from the analysis, depending on the extent of missingness and the impact on the results.

5. Economic indicators may not capture the full complexity of economic disparities as they often focus on macroeconomic factors and may not account for regional or local variations. They also do not account for non-economic factors that influence economic development.

6. Cultural factors are considered in the analysis by examining how they may influence or interact with economic indicators, for example, cultural attitudes towards work and wealth distribution can impact income inequality and poverty rates.

7. Some potential future uses of the analysis include informing policy decisions, guiding international development efforts, and providing insights for academic research education.

8. The findings can inform policymaking and development efforts by highlighting areas where intervention is needed, identifying effective strategies for reducing economic disparities, and providing benchmarks for tracking progress.

9. Recommendations based on the analysis may include policy recommendations, such as implementing social welfare programs to reduce poverty, or recommendations for further research to explore specific aspects of economic disparities.

10. The analysis can be updated and maintained by regularly collecting and updating the data, reviewing the methodology to ensure its relevance and accuracy, and incorporating new indicators or variables to capture emerging trends.

References:

1. The World Factbook by CIA (https://www.cia.gov/the-world-factbook/)
2. Economic Development and Growth: A Comparative Analysis by Michael P. Todaro and Stephen C. Smith
3. “Data Science for Business” by Foster Provost and Tom Fawcett

Website link

https://madden0121.wixsite.com/joseph-madden-data-1

Github Link

https://github.com/josephmadden121/josephmadden121.github.io.git