Exploratory Data Analysis Report on WorldBank Data

Mingke Tian

2024-02-24

Introduction

This report aims to present an exploratory data analysis towards the WordBank data for the "World" region in 2022 (World Bank 2022). We will focus on three key indicators - "GDP per Capita, Inflation Rate and Unemployment Rate" - and includes visualisations and statistical summaries. Some of the methodologies used can be found in (McKinney 2018).

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Data Loading

```
df = pd.read_csv("/Users/mingketian/Desktop/MATH300/wdi.csv")
df.head()
```

	country	inflation_rate	exports_gdp_share	gdp_growth_rate	gdp_per_capita	adult_lite
0	Afghanistan	NaN	18.380042	-6.240172	352.603733	NaN
1	Albania	6.725203	37.395422	4.856402	6810.114041	98.5
2	Algeria	9.265516	31.446856	3.600000	5023.252932	NaN
3	American Samoa	NaN	46.957520	1.735016	19673.390102	NaN
4	Andorra	NaN	NaN	9.563798	42350.697069	NaN

Exploratory Data Analysis for Three Factors

Data Overview

```
df.info()
df.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 217 entries, 0 to 216
Data columns (total 14 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   country                         217 non-null    object
 1   inflation_rate                  169 non-null    float64
 2   exports_gdp_share               169 non-null    float64
 3   gdp_growth_rate                 202 non-null    float64
 4   gdp_per_capita                  203 non-null    float64
 5   adult_literacy_rate             49 non-null     float64
 6   primary_school_enrolment_rate   114 non-null    float64
 7   education_expenditure_gdp_share 105 non-null    float64
 8   measles_immunisation_rate       193 non-null    float64
 9   health_expenditure_gdp_share    20 non-null     float64
 10  income_inequality               28 non-null     float64
 11  unemployment_rate               186 non-null    float64
 12  life_expectancy                 209 non-null    float64
 13  total_population                217 non-null    float64
dtypes: float64(13), object(1)
memory usage: 23.9+ KB
```

	inflation_rate	exports_gdp_share	gdp_growth_rate	gdp_per_capita	adult_literacy_rate	pr
count	169.000000	169.000000	202.000000	203.000000	49.000000	11
mean	12.493936	46.170395	4.368901	20345.707649	79.574801	10
std	19.682433	34.001404	6.626811	31308.942225	19.375539	12
min	-6.687321	1.571162	-28.758591	259.025031	27.280001	64
25%	5.518129	24.526642	2.438593	2570.563284	72.400002	94
50%	7.967574	40.221277	4.204431	7587.588173	83.779999	10
75%	11.665567	55.460067	6.200000	25982.630050	95.500000	10
max	171.205491	211.278206	63.439864	240862.182448	99.999977	13

Checking for Missing Values

```
df.isnull().sum()
```

```
country                              0
inflation_rate                      48
exports_gdp_share                   48
gdp_growth_rate                     15
gdp_per_capita                      14
adult_literacy_rate                168
primary_school_enrolment_rate      103
education_expenditure_gdp_share    112
measles_immunisation_rate           24
health_expenditure_gdp_share       197
income_inequality                  189
unemployment_rate                   31
life_expectancy                      8
total_population                     0
dtype: int64
```

Factor 1: GDP per Capita

```
df['gdp_per_capita'].describe()
```

```
count       203.000000
mean      20345.707649
std       31308.942225
min         259.025031
25%        2570.563284
50%        7587.588173
75%       25982.630050
max      240862.182448
Name: gdp_per_capita, dtype: float64
```

Factor 2: Inflation Rate

```
df['inflation_rate'].describe()
```

```
count    169.000000
mean      12.493936
std       19.682433
min       -6.687321
25%        5.518129
50%        7.967574
75%       11.665567
max      171.205491
Name: inflation_rate, dtype: float64
```

Factor 3: Unemployment Rate

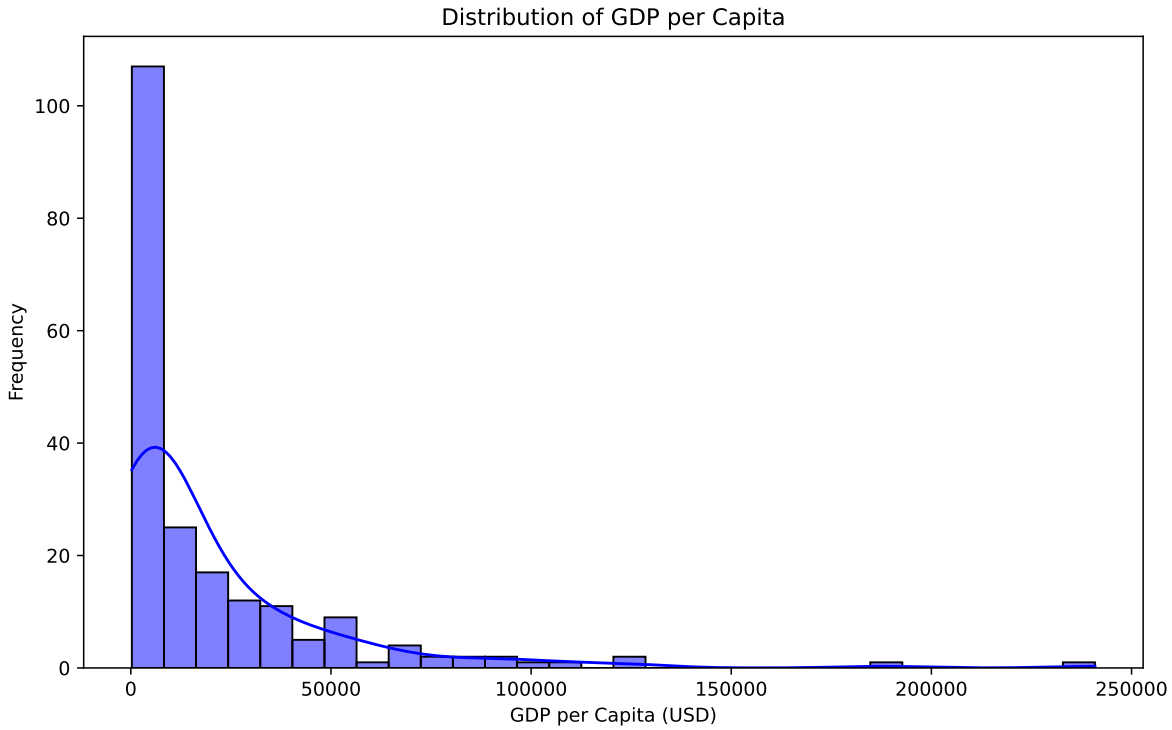
```
df['unemployment_rate'].describe()
```

```
count    186.000000
mean       7.268661
std        5.827726
min        0.130000
25%        3.500750
50%        5.537500
75%        9.455250
max       37.852000
Name: unemployment_rate, dtype: float64
```

Visualisations

Distribution of GDP per Capita

```
plt.figure(figsize=(10,6))
sns.histplot(df['gdp_per_capita'], bins=30, kde=True, color='blue')
plt.title("Distribution of GDP per Capita")
plt.xlabel("GDP per Capita (USD)")
plt.ylabel("Frequency")
plt.show()
```

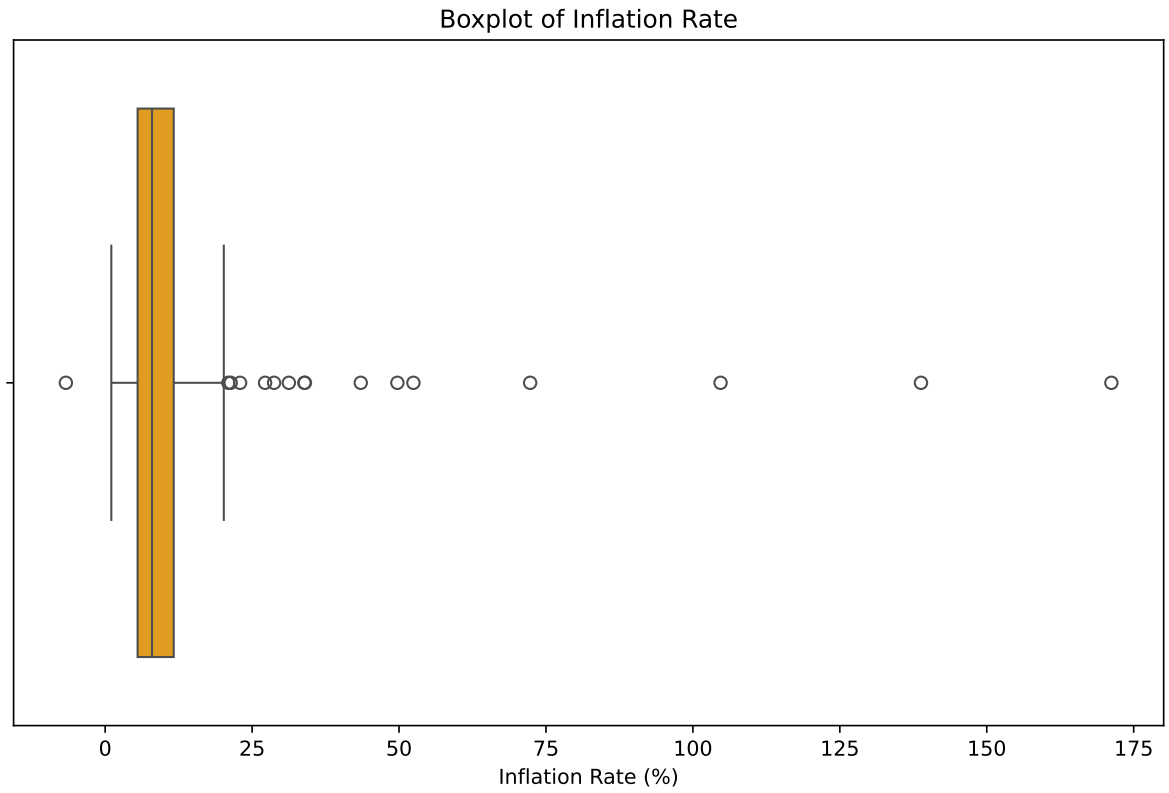


Findings 1:

As shown in figure (**ref?**)(fig:gdp-distribution): The distribution is right-skewed, with most countries having low GDP per capita. A few wealthy nations push the upper limit, highlighting global economic disparity. The majority cluster at lower values, suggesting a large income gap between countries.

Inflation Rate Analysis

```
plt.figure(figsize=(10,6))
sns.boxplot(x=df['inflation_rate'], color='orange')
plt.title("Boxplot of Inflation Rate")
plt.xlabel("Inflation Rate (%)")
plt.show()
```



Findings 2:

As shown in figure (**ref?**)(fig:inflation-analysis): Inflation varies widely, with many outliers indicating extreme cases. Most countries have inflation rates within a moderate range, but a few experience hyperinflation. The long upper whisker suggests economic instability in some regions.

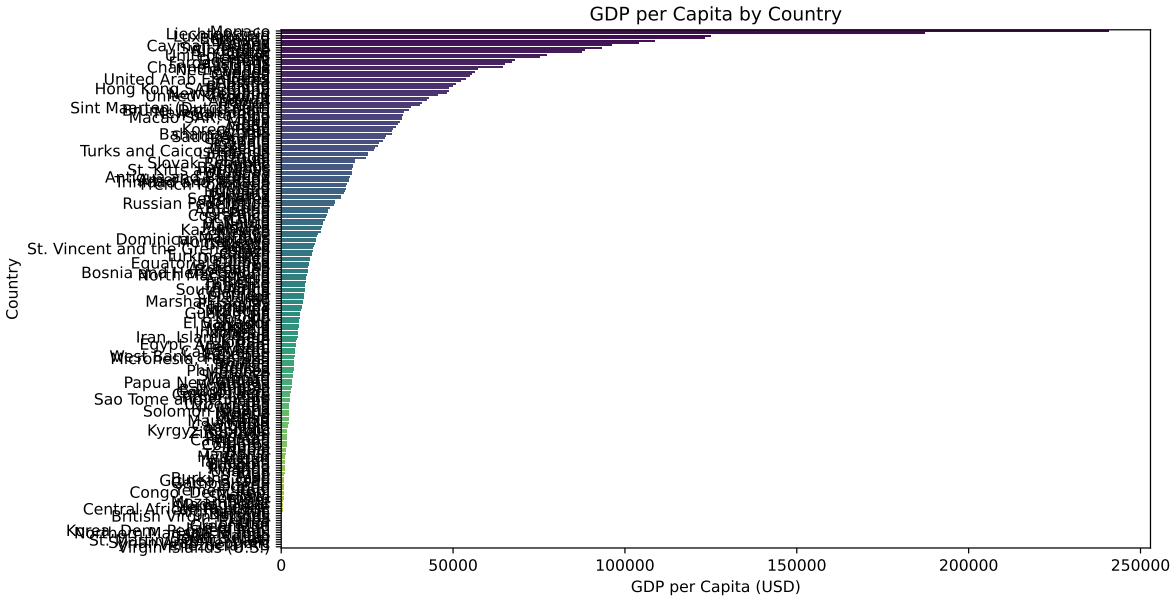
Bar Chart of GDP per Capita

```
plt.figure(figsize=(10,6))
df_sorted = df.sort_values('gdp_per_capita', ascending=False)
sns.barplot(data=df_sorted, x='gdp_per_capita', y='country', palette='viridis')
plt.title("GDP per Capita by Country")
plt.xlabel("GDP per Capita (USD)")
plt.ylabel("Country")
plt.yticks(rotation=0, ha='right')
plt.show()
```

```
/var/folders/20/wpnwbjwj2y75lf83mlmkjjcw0000gn/T/ipykernel_93701/3729854541.py:3: FutureWarni

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assig

  sns.barplot(data=df_sorted, x='gdp_per_capita', y='country', palette='viridis')
```

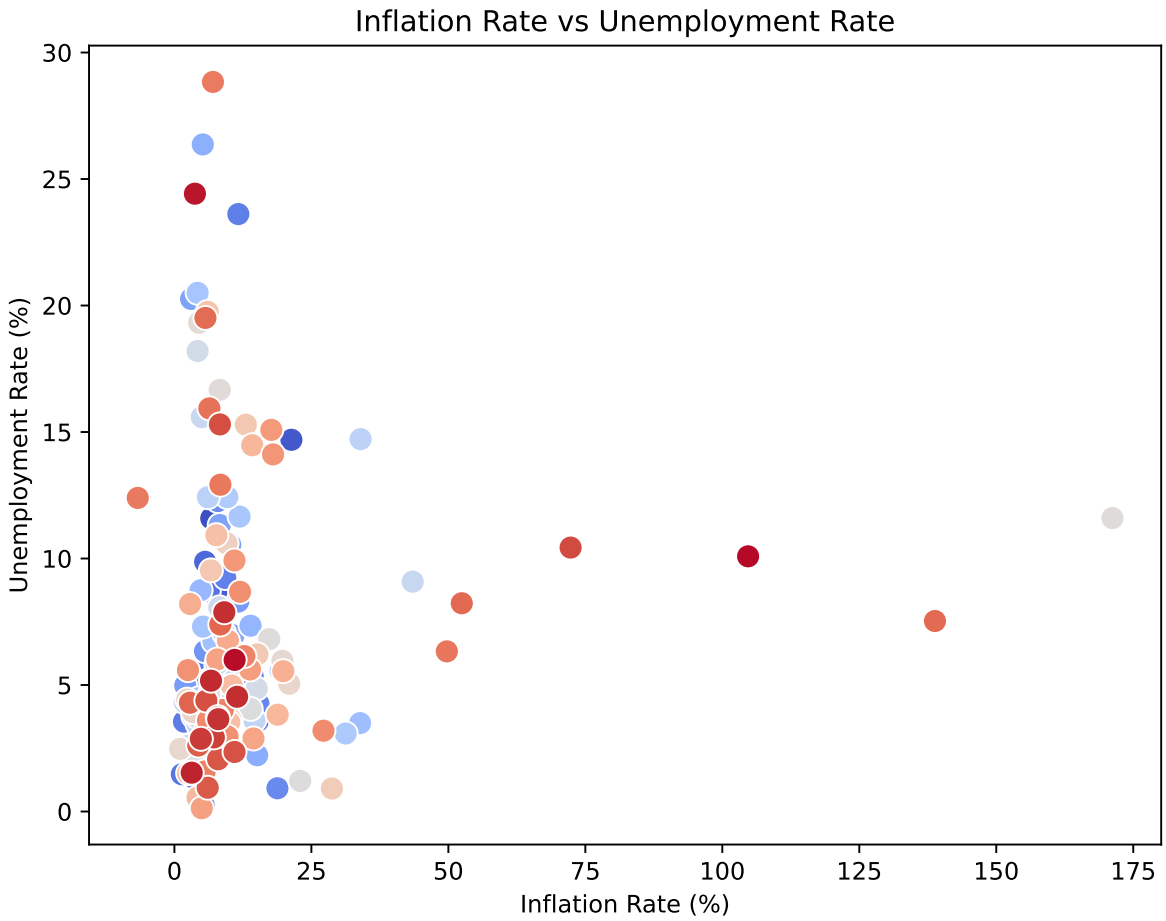


Findings 3:

As shown in figure (**ref?**)(fig:bar-chart): The bar chart shows a high concentration of countries with low GDP per capita, while a few nations have significantly higher values. The skewed distribution highlights global economic inequality, with wealth concentrated in a small number of countries.

Scatter Plot of Inflation Rate vs Unemployment Rate

```
plt.figure(figsize=(8,6))
sns.scatterplot(data=df, x='inflation_rate', y='unemployment_rate', hue='country', palette='
plt.title("Inflation Rate vs Unemployment Rate")
plt.xlabel("Inflation Rate (%)")
plt.ylabel("Unemployment Rate (%)")
plt.show()
```



As shown in figure (**ref?**)(fig:scatter-plot): The scatter plot reveals no clear correlation between inflation and unemployment, with most countries clustering at low inflation and unemployment rates. However, a few outliers show extreme values, indicating economic instability in certain nations.

Summary Table

```
summary_table = df[['gdp_per_capita', 'inflation_rate', 'unemployment_rate']].describe()
summary_table
```

	gdp_per_capita	inflation_rate	unemployment_rate
count	203.000000	169.000000	186.000000

	gdp_per_capita	inflation_rate	unemployment_rate
mean	20345.707649	12.493936	7.268661
std	31308.942225	19.682433	5.827726
min	259.025031	-6.687321	0.130000
25%	2570.563284	5.518129	3.500750
50%	7587.588173	7.967574	5.537500
75%	25982.630050	11.665567	9.455250
max	240862.182448	171.205491	37.852000

Table (**ref?**)(tbl:summary) presents key statistics for the three indicators.

References

McKinney, Wes. 2018. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 2nd ed. O'Reilly Media.

World Bank. 2022. "World Bank Data." https://data.worldbank.org/.