WDI 2022 Report

Jinhao Tian

2025-10-04

Table of contents

Data Loading

In this section, I load the World Development Indicators dataset for 2022 (Bank 2022).

```
import pandas as pd

# Load the dataset
df = pd.read_csv("wdi.csv")

# Show a small preview (as a figure-like output)
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

Figure 1: Preview of the WDI sample (first 5 rows). Source: World Bank WDI.

Exploratory Data Analysis (EDA)

This section explores **GDP per capita**, **Life expectancy**, and **Inflation (CPI, annual %)** for 2022.

1) GDP per capita (levels and distribution)

I examine summary statistics, missingness, and the distribution in both raw and log scales.

```
# quick summary in text output
import numpy as np
print("Missing values (gdp_per_capita):", df["gdp_per_capita"].isna().sum())
print(df["gdp_per_capita"].describe())
```

```
Missing values (gdp_per_capita): 14
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
```

```
import matplotlib.pyplot as plt
x = df["gdp_per_capita"].dropna()
plt.figure()
plt.hist(x, bins=30)
plt.title("GDP per Capita (USD, 2022)")
```

```
plt.xlabel("USD")
plt.ylabel("Count")
plt.show()
```

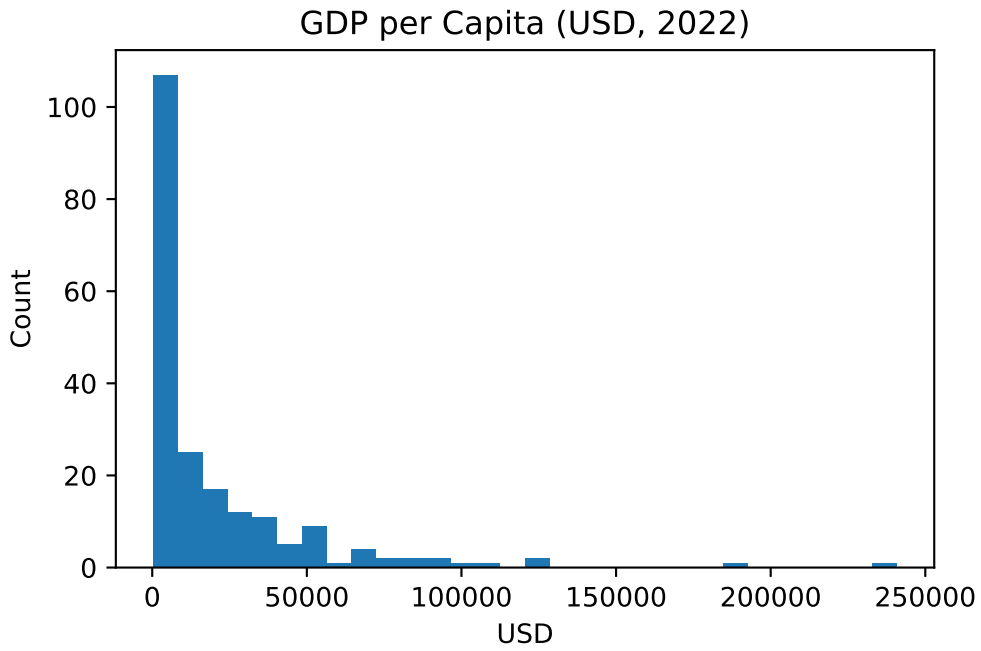


Figure 2: Distribution of GDP per capita (USD, 2022). Source: World Bank WDI.

```
import numpy as np
import matplotlib.pyplot as plt
x = df["gdp_per_capita"].dropna()
plt.figure()
plt.hist(np.log10(x[x>0]), bins=30)
plt.title("GDP per Capita (log10 scale, 2022)")
plt.xlabel("log10(USD)")
plt.ylabel("Count")
plt.show()
```

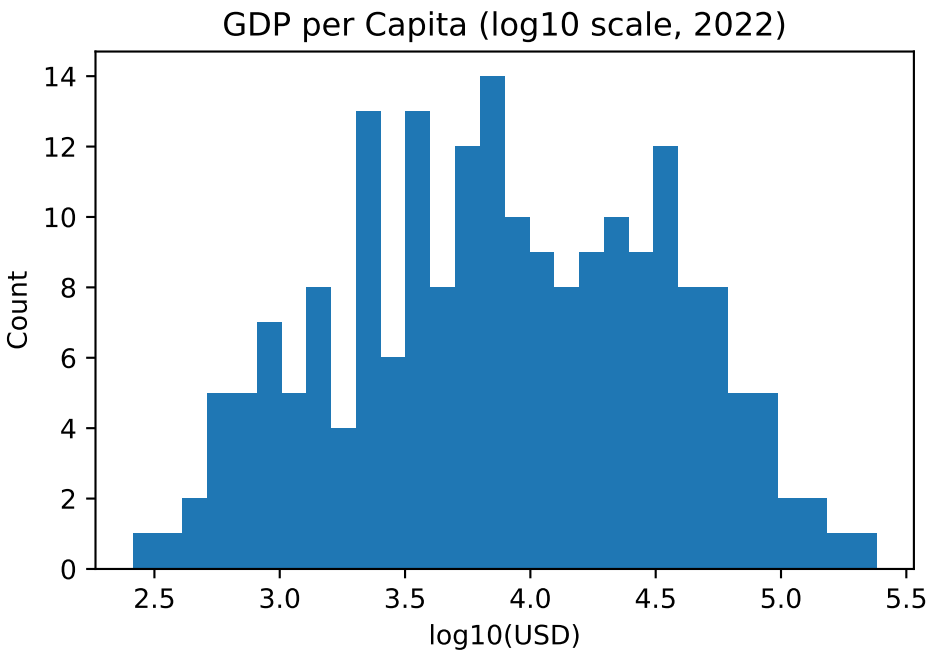


Figure 3: Distribution of GDP per capita on log scale (2022). Source: World Bank WDI.

Summary — GDP per capita (2022)
- The distribution is **highly right-skewed**: most countries cluster below ~$20k, with a **long tail** of high-income economies.
- On a **log scale**, the distribution looks much closer to symmetric (roughly log-normal), which is typical for income variables.
- Interpretation: comparing countries on the **log** of GDP per capita (rather than raw USD) is more informative and reduces the influence of outliers.

---

2) Life expectancy (levels and relation to income)

I inspect the distribution of life expectancy and its relationship with income (diminishing returns expected), a pattern first described by Preston (Preston 1975).

```
# brief summary
print("Missing values (life_expectancy):", df["life_expectancy"].isna().sum())
print(df["life_expectancy"].describe())
```

```
Missing values (life_expectancy): 8
count    209.000000
mean      72.416519
std        7.713322
min       52.997000
25%       66.782000
50%       73.514634
75%       78.475000
max       85.377000
Name: life_expectancy, dtype: float64
```

```python
import numpy as np
import matplotlib.pyplot as plt


gdp_le = df[["gdp_per_capita","life_expectancy"]].dropna().copy()
x = np.log10(gdp_le["gdp_per_capita"].values)
y = gdp_le["life_expectancy"].values

m, b = np.polyfit(x, y, 1)
plt.figure()
plt.scatter(x, y, alpha=0.7)
plt.plot(np.sort(x), m*np.sort(x)+b, linewidth=2)
plt.title("Life Expectancy vs log10(GDP per Capita), 2022")
plt.xlabel("log10(GDP per Capita, USD)")
plt.ylabel("Life Expectancy (years)")
plt.grid(True, linewidth=0.3)
plt.show()

print(f"Correlation (life expectancy vs log10 GDP per capita): {np.corrcoef(x,y)[0,1]:.3f}")
```

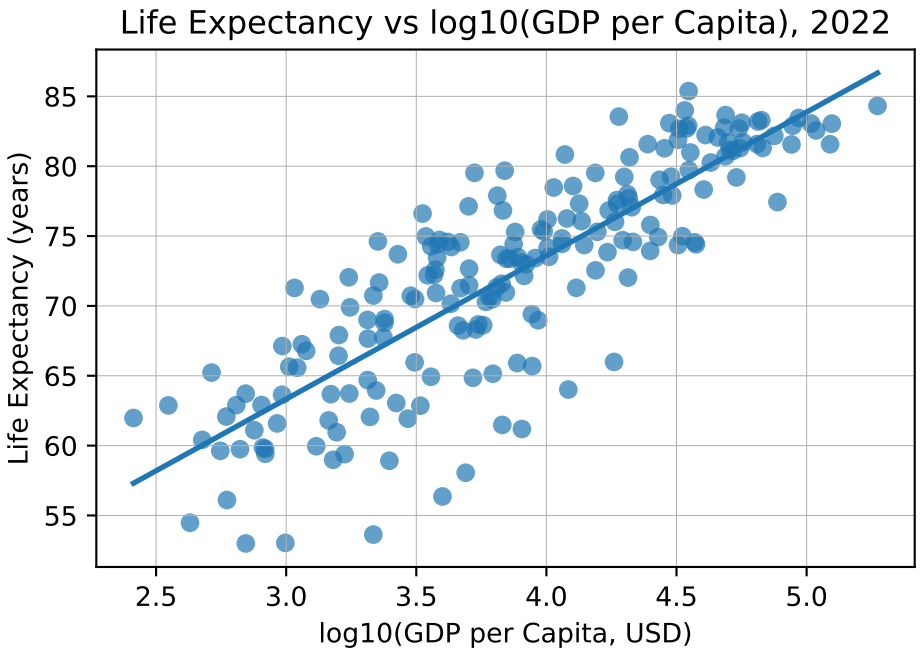


Figure 4: Life expectancy vs log (GDP per capita), 2022. Source: World Bank WDI.

```
Correlation (life expectancy vs log10 GDP per capita): 0.841
```

Summary — Life expectancy vs GDP per capita
- There is a **strong positive association**: the correlation with **log (GDP per capita)** is about **0.84** in the output.
- The **slope flattens at higher incomes**, consistent with **diminishing returns** (gains in life expectancy are large at low incomes and smaller at high incomes), this finding aligns with macroeconomic theories discussed by Mankiw (Mankiw 2019). .
- Takeaway: income is a powerful predictor of population health, especially among lower-income countries. —

3) Inflation (levels and relation to growth)

I examine inflation levels and its same-year association with GDP growth.

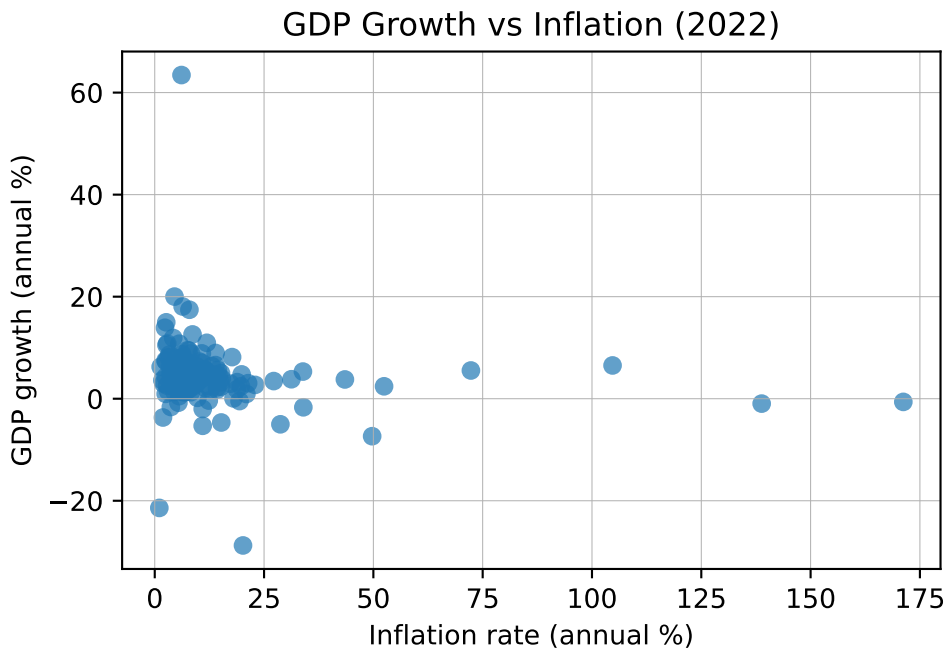
```
# brief summary
print("Missing values (inflation_rate):", df["inflation_rate"].isna().sum())
print(df["inflation_rate"].describe())
```

```
#| label: fig-growth-vs-infl
#| fig-cap: "GDP growth vs inflation, 2022. Source: [World Bank WDI](https://databank.worldba
import numpy as np
import matplotlib.pyplot as plt

if "gdp_growth_rate" in df.columns:
    infl_growth = df[["inflation_rate","gdp_growth_rate"]].dropna()
    plt.figure()
    plt.scatter(infl_growth["inflation_rate"], infl_growth["gdp_growth_rate"], alpha=0.7)
    plt.title("GDP Growth vs Inflation (2022)")
    plt.xlabel("Inflation rate (annual %)")
    plt.ylabel("GDP growth (annual %)")
    plt.grid(True, linewidth=0.3)
    plt.show()

    r = np.corrcoef(infl_growth["inflation_rate"], infl_growth["gdp_growth_rate"])[0,1]
    print(f"Correlation (GDP growth vs inflation): {r:.3f}")
else:
    print("Column 'gdp_growth_rate' not found in the dataset.")
```

```
Missing values (inflation_rate): 48
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
```




```
Correlation (GDP growth vs inflation): -0.140
```

Summary — GDP growth vs inflation (2022)
- The cross-section shows a **weak negative correlation** (**−0.14**), and the scatter is **very noisy**.
- Several **extreme-inflation outliers** pull the pattern around; excluding them would likely make the relationship even weaker.
- Same-year cross-sectional comparisons don't identify causality—growth and inflation dynamics are **time-dependent**; panel/time-series analysis would be more appropriate.

Visualisations

Top GDP per Capita (Bar Chart)

As shown in (Figure 5), high-income economies dominate the upper tail of GDP per capita in 2022.

```
import matplotlib.pyplot as plt

gdp_top = (
    df[["country", "gdp_per_capita"]]
    .dropna()
    .sort_values("gdp_per_capita", ascending=False)
    .head(12)
    .iloc[::-1]
)

plt.figure(figsize=(9,6))
plt.barh(gdp_top["country"], gdp_top["gdp_per_capita"])
plt.title("Top 12 GDP per Capita (USD), 2022")
plt.xlabel("GDP per Capita (USD)")
plt.ylabel("Country")
plt.tight_layout()
plt.show()
```

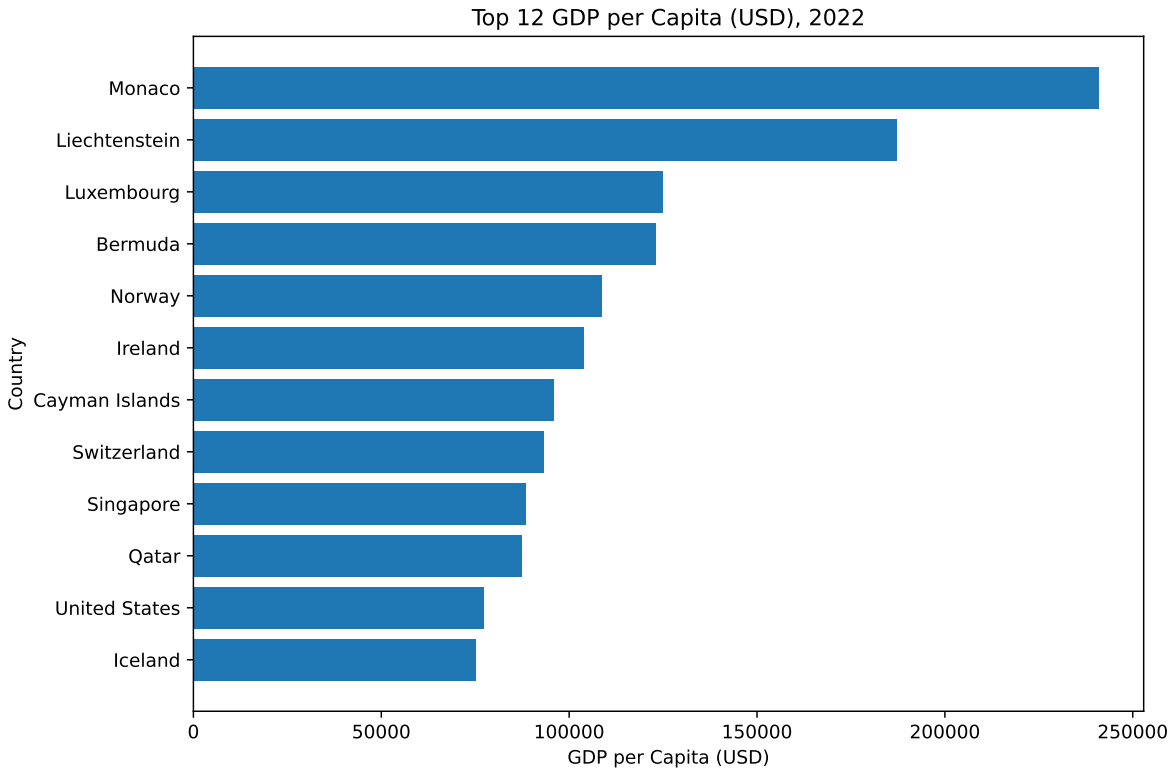


Figure 5: Top 12 countries by GDP per capita (2022). Source: World Bank WDI.

Life Expectancy vs Income (Scatter)

In (Figure 6), life expectancy rises with income, with diminishing returns at high income.

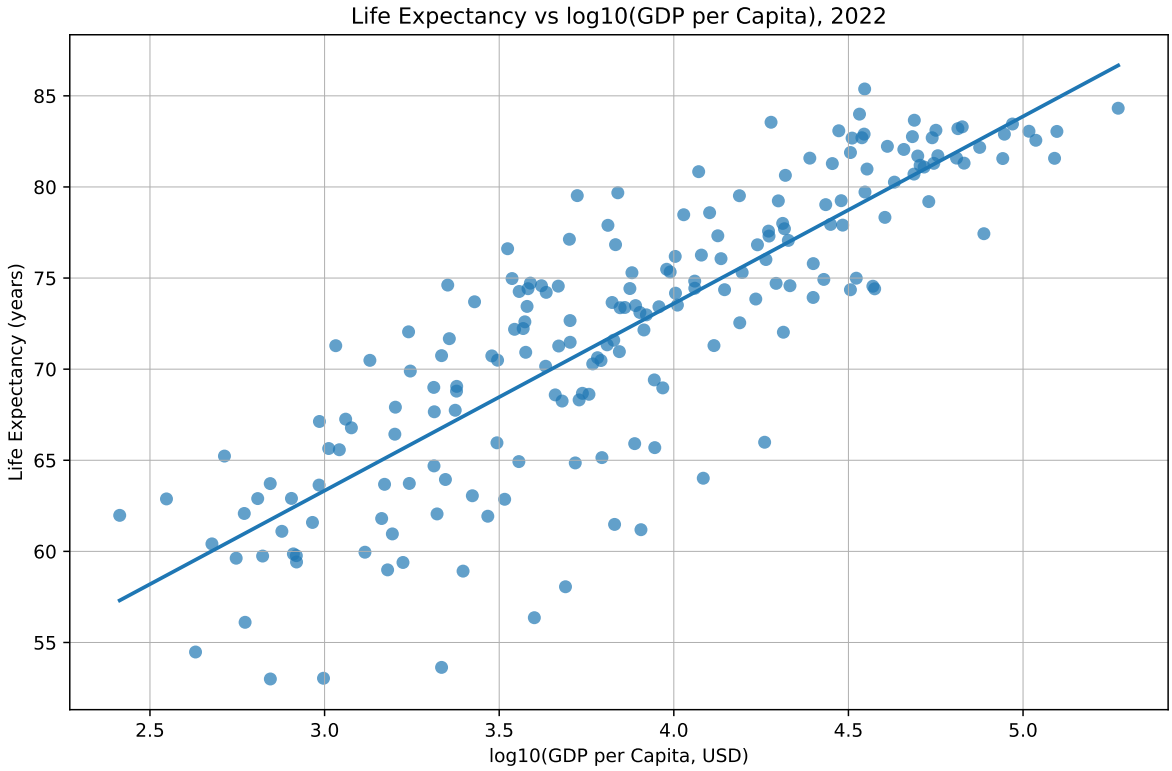


Figure 6: Life expectancy vs. log (GDP per capita), 2022. Source: World Development Indicators — World Bank.

Key Statistics Table

Table Table 2 summarizes the indicators used in the analysis and supports the distributional findings in Figure 2 and the relationship in Figure 4.

```
import pandas as pd
import numpy as np

indicators = {
    "gdp_per_capita": "GDP per capita (USD)",
    "life_expectancy": "Life expectancy (years)",
    "inflation_rate": "Inflation rate (%)",
```

```
}

use_cols = [c for c in indicators if c in df.columns]
desc = (
    df[use_cols]
    .describe(percentiles=[], include="all")
    .loc[["count","mean","50%","std","min","max"]]
    .rename(index={"50%":"median"})
    .T
)

def fmt_row(name, row):
    row = row.copy()
    if name == "gdp_per_capita":
        for c in ["mean","median","std","min","max"]:
            row[c] = f"{row[c]:,.0f}"
    elif name in ["life_expectancy"]:
        for c in ["mean","median","std","min","max"]:
            row[c] = f"{row[c]:.1f}"
    else:  # inflation_rate or other %
        for c in ["mean","median","std","min","max"]:
            row[c] = f"{row[c]:.2f}"
    row["count"] = int(row["count"])
    return row

display_df = pd.DataFrame(
    [fmt_row(name, desc.loc[name]) for name in desc.index],
    index=[indicators[name] for name in desc.index]
)[["count","mean","median","std","min","max"]]

display_df
```

Table 2: Key statistics for selected indicators (2022). Source: World Bank WDI.

	count	mean	median	std	min	max
GDP per capita (USD)	203	20,346	7,588	31,309	259	240,862
Life expectancy (years)	209	72.4	73.5	7.7	53.0	85.4
Inflation rate (%)	169	12.49	7.97	19.68	-6.69	171.21

Bank, World. 2022. “World Development Indicators 2022.” https://databank.worldbank.org/source/world-development-indicators.

Mankiw, N. Gregory. 2019. *Principles of Economics*. 9th ed. Cengage Learning.
Preston, Samuel H. 1975. "The Changing Relation Between Mortality and Level of Economic Development." *Population Studies* 29 (2): 231–48. https://doi.org/10.1080/00324728.1975.10410201.