Data Science in Practice CA Report

Yuzhe Shi

This report analyzes the OWID[1] COVID-19[2] dataset, exploring several key questions. Through data analysis and visualization, this study aims to provide insights for understanding global pandemic patterns and improving control strategies.

March 22, 2025

Yuzhe Shi
20108862@mail.wit.ie°

Code is available on https://github.com/bkmashiro/DSP-report°

Contents

1. Problem Definition

This study aims to address three key research questions regarding the COVID-19 pandemic's impact on global health and economy:

1. What are the long-term health consequences of COVID-19 across **different regions** and **income levels**?
2. How has the pandemic affected different **economic sectors** and **income groups**? This includes analyzing the relationship between **GDP changes**, **stringency measures**, and their **effectiveness** across different economic levels.
3. What is the relationship between poverty levels and COVID-19 mortality rates? This question explores the complex interaction between socioeconomic factors and health outcomes during the pandemic.

These questions are particularly relevant as they address both the immediate and long-term implications of the pandemic, while considering the varying capacities and responses of different countries and regions.

2. General Preprocessing

2.1. Data cleaning

The dataset is from the OWID COVID-19 dataset, which is a comprehensive collection of COVID-19 data from around the world.

The dataset has a size of $(429435, 72)$.

This report focuses on the social and economic indicators and healthcare system capacity metrics:

- `gdp_per_capita`
- `extreme_poverty`
- `life_expectancy`
- `human_development_index`

which all contain missing values. The percentage of avalibility values are:

- gdp_per_capita: 76.45%
- extreme_poverty: 49.37%
- life_expectancy: 90.89%
- human_development_index: 74.31%

The distribution of the data is shown below:

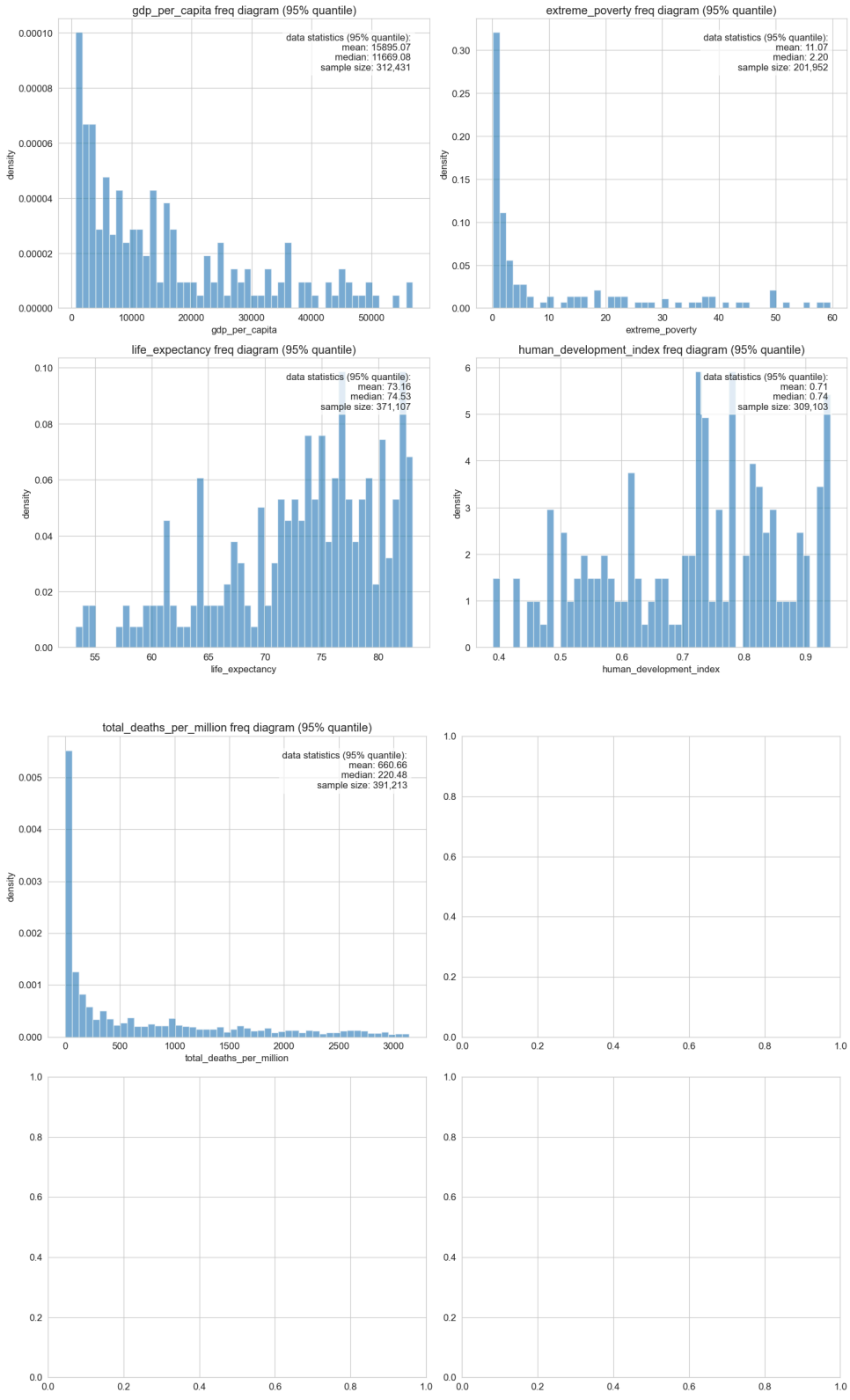

Figure 1: Distribution of the selected data

2.2. Further cleaning of the data

The data is grouped by `location`, and the data rows without GDP data or life expectancy data are removed.

There are 195 countries in the final dataset, with a size of $(326618, 72)$

The distribution of the data is shown below:

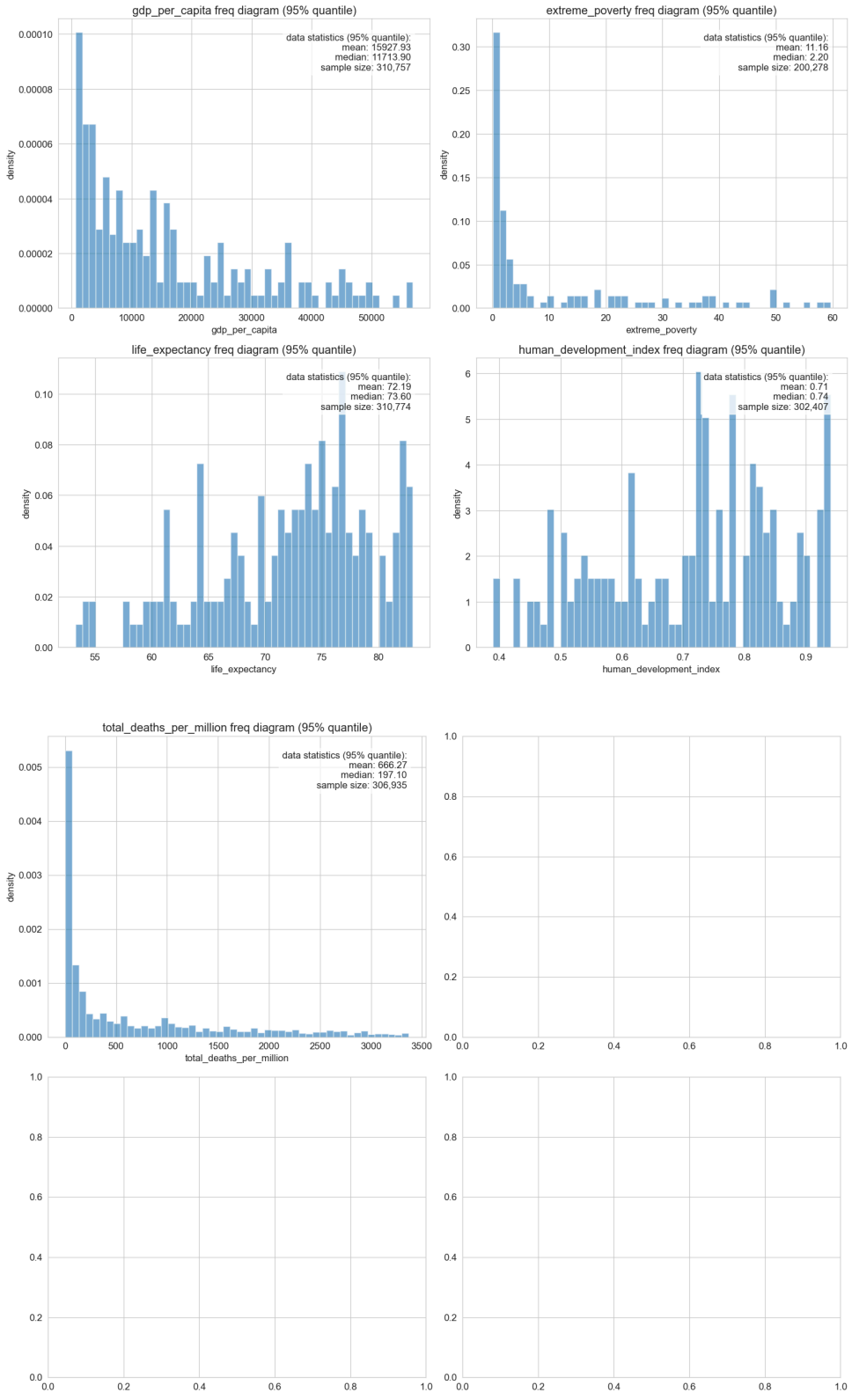

Figure 2: Distribution of the cleaned selected data

3. Health and Economy Impact

3.1. Research Question 1: Long-term Health Consequences of COVID-19

3.1.1. Data and Preprocessing

3.1.1.1. Data selection

The analysis of COVID-19's long-term health consequences utilized the following data:

- Social and economic indicators from OWID COVID-19 dataset
  - `gdp_per_capita`
  - `extreme_poverty`
  - `life_expectancy`
  - `human_development_index`
- Healthcare system capacity metrics from OWID COVID-19 dataset
  - `total_deaths_per_million`

3.1.1.2. Data preparation

- data is grouped by `location`
- data with missing mortality are removed

3.1.2. Results and Findings

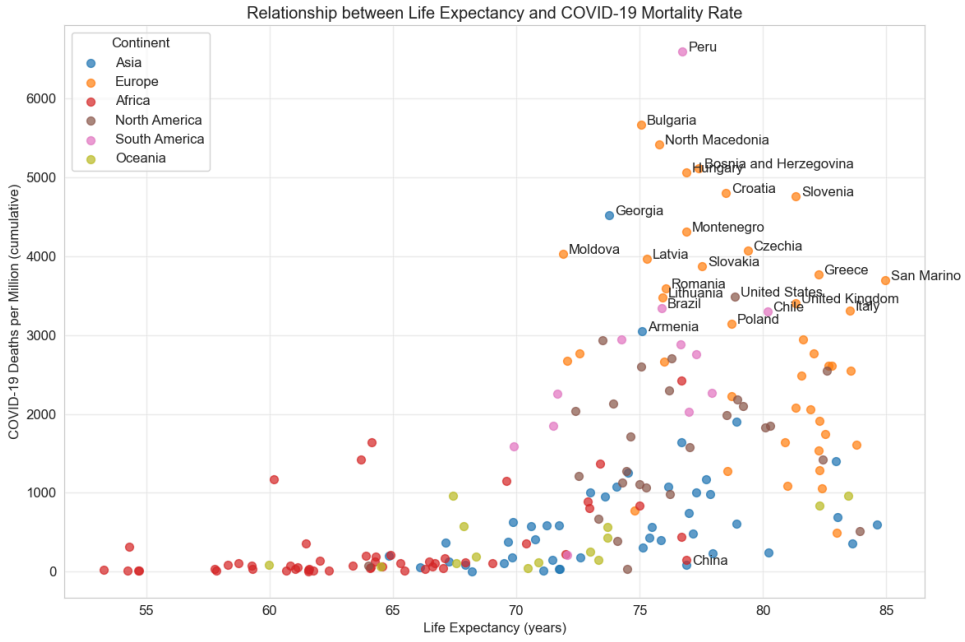

Figure 3: Scatter plot of Life Expectancy and COVID-19 Mortality Rate

R^2 :	0.286
#Observations:	193
Variable	**Coefficient**
const	−6052.11

life_expectancy	101.09
P-value	**< 0.001**

Table 1: OLS Analysis of Life Expectancy and COVID-19 Mortality Rate

From the scatter plot and OLS analysis, we can observe:

- **Correlation Analysis**:
  - Correlation coefficient between life expectancy and COVID-19 mortality rate: 0.535 ($p < 0.001$).
  - This positive correlation indicates that:
    - Countries with higher life expectancy.tend to have higher COVID-19 mortality rates
    - The relationship is **statistically significant** ($p < 0.001$).
    - The moderate strength of correlation (0.535) suggests a meaningful but not perfect relationship.
- **Life Expectancy Loss Analysis**:

Continent	Estimated Life Expectancy Loss (Years)	Relative Impact (Africa = 1.0)
Europe	0.294	9.8
South America	0.267	8.9
North America	0.162	5.4
Asia	0.070	2.3
Oceania	0.038	1.7
Africa	0.030	1.0

Table 2: Estimated Life Expectancy Loss by Continent

Note that:

1. Europe experienced the highest life expectancy loss (0.29 years)
2. Africa showed the lowest impact (0.03 years)

  - **Regional Patterns**:
    - European countries show higher mortality rates despite higher life expectancy
    - African nations generally show lower mortality rates but also lower life expectancy
    - This suggests that the relationship is influenced by multiple socioeconomic factors

3.1.3. Interpretation

- This finding might seem counterintuitive at first that **European countries show higher mortality rates than Africa despite higher life expectancy**.
- However, it can be explained by several factors:
  - Higher life expectancy countries often have older populations
  - More developed healthcare systems may have better reporting of COVID-19 deaths
  - Different testing policies and reporting standards

3.2. Research Question 2: Economic Sector Impact

3.2.1. Data and Preprocessing

3.2.1.1. Data selection

A new dataset is created by merging the OWID COVID-19 dataset with the **World Bank dataset**[3] on:

- GDP growth (annual %)
- GDP per capita growth (annual %)
- GDP (current US$)
- GDP per capita growth (annual %)

Economic impact analysis utilized:

- Economic indicators
  - GDP growth (annual %)
  - GDP per capita growth (annual %)
  - GDP (current US$)
  - GDP per capita growth (annual %)
- OWID COVID-19 dataset
  - `total_deaths_per_million`
  - `stringency_index`
  - `continent`

3.2.1.2. Data preparation

- Merge Data
  - Merge the OWID COVID-19 dataset with the World Bank dataset on `location` and `Country Name`
  - GDP change is added to the dataset
    - `gdp_change_2020` is the GDP change in 2020, in percentage
    - `gdp_change_2021` is the GDP change in 2021, in percentage
    - `gdp_change_2022` is the GDP change in 2022, in percentage
- Data Cleaning
  - These have missing values in `stringency_index`, `total_deaths_per_million` and `gdp_per_capita` are removed.
  - The final dataset has a size of $(152, 28)$
- Preprocessing
  - `gdp_group` is q-cut into 3 groups (Low, Medium, High) based on `gdp_per_capita`
  - `income_level` is q-cut into 4 groups (Low, Medium-Low, Medium-High, High) based on `gdp_per_capita`

3.2.2. Results and Findings

Breif view of the World Bank dataset:

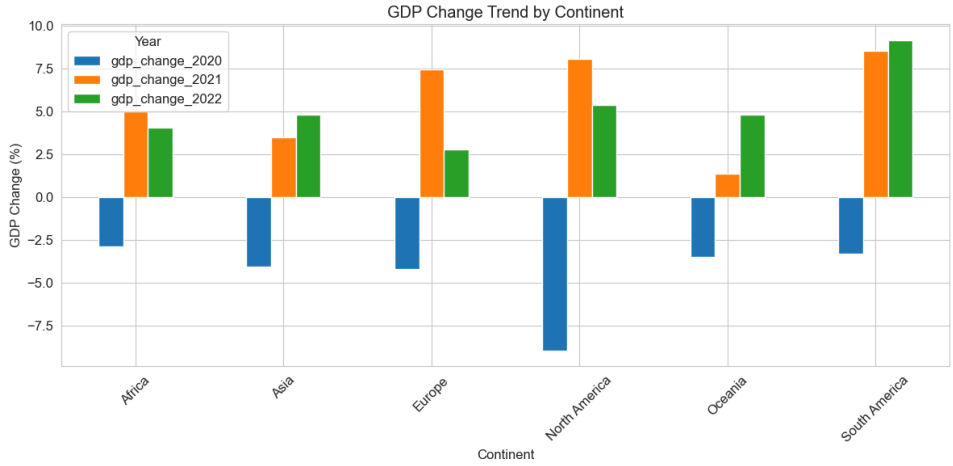

Figure 4: GDP change by continent

We can observe that:

- In the year of 2020, the GDP of all continents are decreased. While in the year of 2021 and 2022, the GDP of all continents are increased, which is a sign of economic recovery.

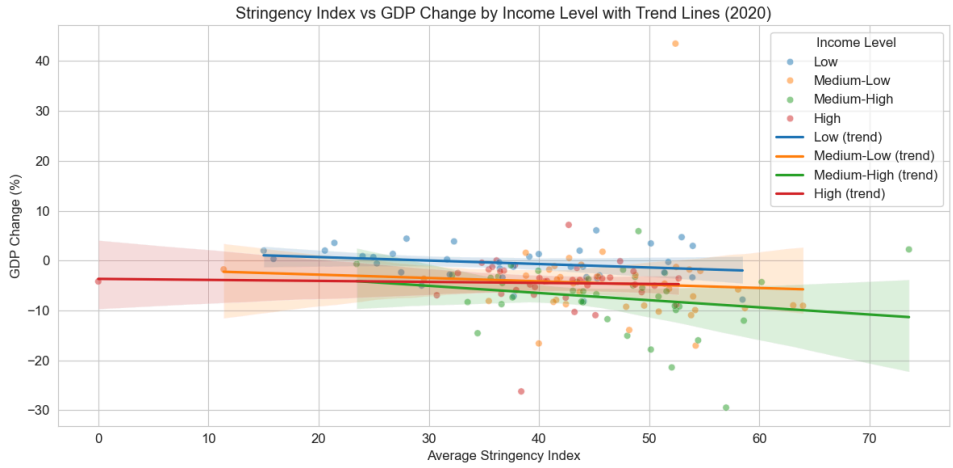

Figure 5: GDP change and stringency index by income level

We can observe that:

- The relationship between the stringency index and the GDP change is not very strong.

GDP Group	Stringency Index	Deaths/Million	GDP/Capita	Count
Low Income	38.66	201.65	2,898.27	51
Middle Income	47.26	1,702.12	12,757.97	50
High Income	41.54	2,173.71	42,895.97	51

Table 3: Health and Economic Impacts by Income Level

The table shows that:

- high income level countries have higher stringency index and deaths per million. This is possibly due to different testing policies and reporting standards.

Continent	Stringency Index	Deaths/Million	GDP/Capita	Count

Africa	39.00	322.52	5,360.89	44
Asia	48.50	664.93	23,883.40	33
Europe	39.61	2,809.86	35,321.85	37
North America	42.97	1,627.93	20,385.54	19
Oceania	39.86	428.00	13,224.41	8
South America	48.73	2,890.62	13,576.77	11

Table 4: Health and Economic Impacts by Continent

Efficiency of the policy is defined as below:

$$\text{Efficiency} = \frac{\text{Deaths per Million}}{\text{Stringency Index}} \tag{1}$$

GDP Group	Mean Efficiency	Median Efficiency	Std Dev	Count
Low Income	5.22	1.89	8.76	51
Middle Income	36.02	29.84	27.91	50
High Income	52.33	48.77	31.45	51

Table 5: Policy Efficiency by Income Level[1]

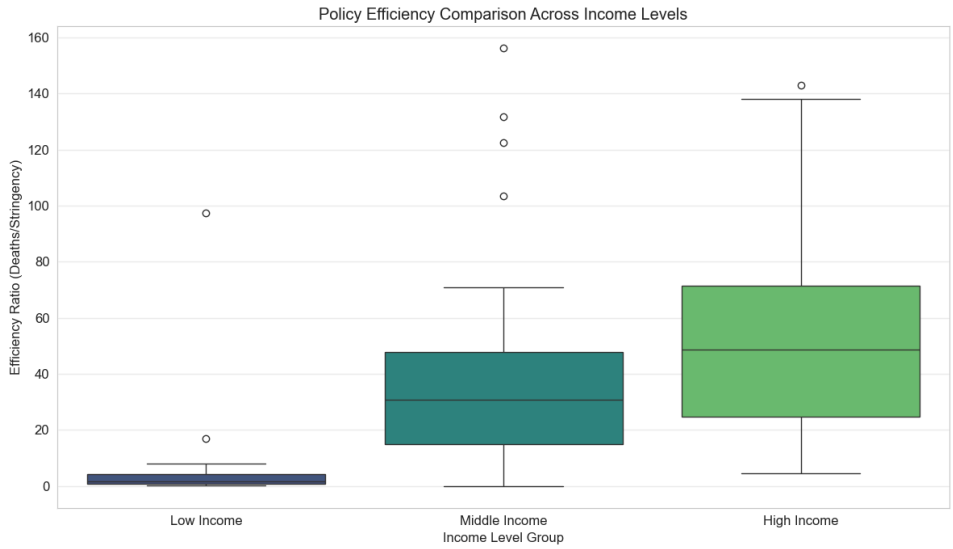

Figure 6: Policy Efficiency by GDP Group

We can observe that:

- The efficiency of the policy is **much higher in the high income level**.

3.2.3. Interpretation

- The high income level countries have higher stringency index and deaths per million. This is possibly due to different testing policies and reporting standards.[2]

[1]These has 0 stringency index values are removed

[2]This is a assumption.

- The low income level countries have lower efficiency in the policy. This is possibly due to the lack of resources and infrastructure.

3.3. Research Question 3: Health Measures and Economic Recovery Relationship

3.3.1. Data and Preprocessing

3.3.1.1. Data selection

The analysis used the following data:

- Poverty
  - `extreme_poverty`
- Economic data
  - `gdp_per_capita`
- Health data
  - `total_deaths_per_million`
  - `stringency_index`
- Miscellaneous
  - `continent`

3.3.1.2. Data preparation

If a country has missing values in the `extreme_poverty` and `total_deaths_per_million`, it will be removed.

210322(64.39%) records have poverty data, covered 125 countries.

3.3.2. Results and Findings

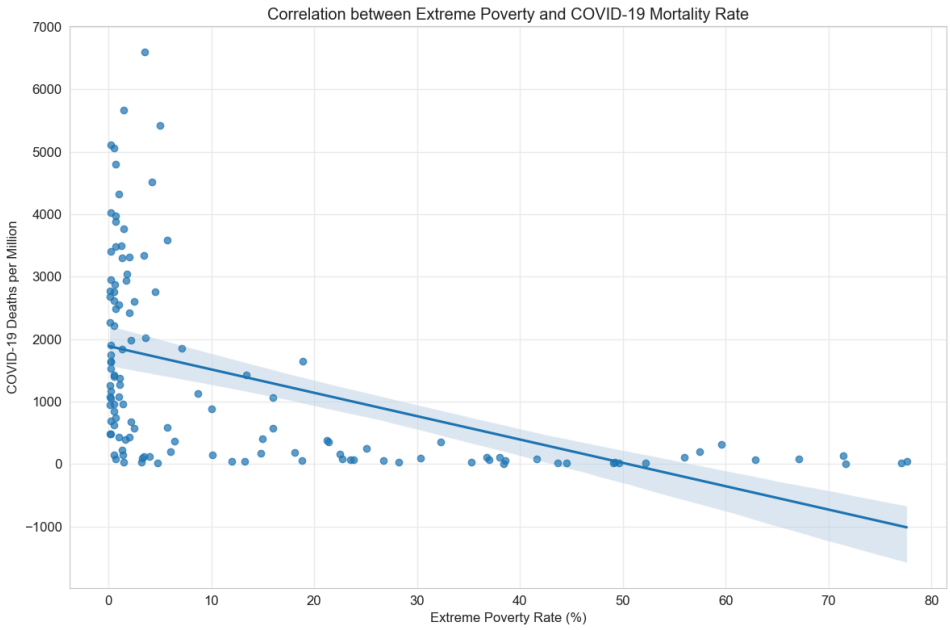

Figure 7: Correlation between Extreme Poverty and COVID-19 Mortality Rate

From the figure, we can observe that:

- The correlation between extreme poverty and COVID-19 mortality rate is negative.
- The relationship is not very strong.

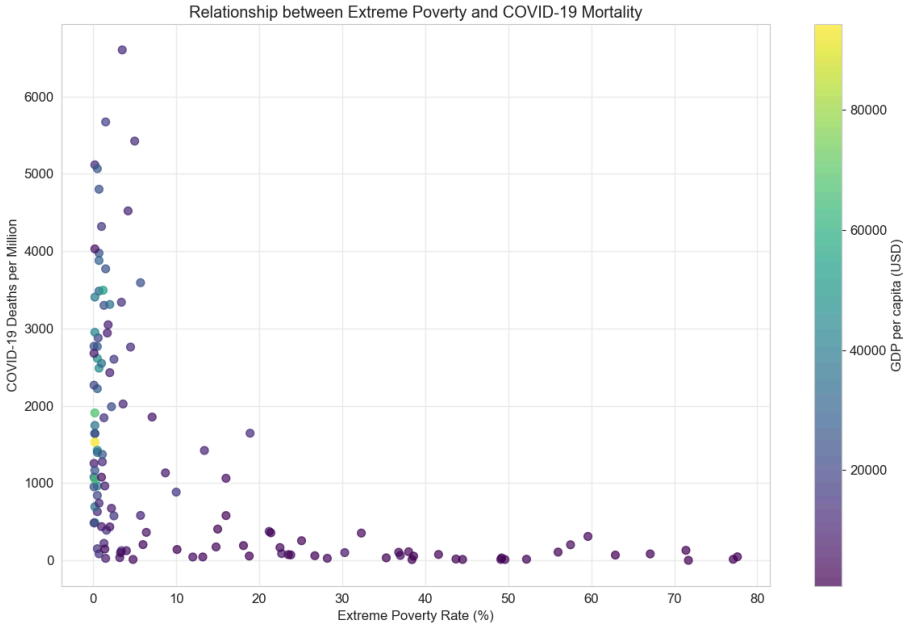

Figure 8: Scatter plot of Extreme Poverty and COVID-19 Mortality Rate

From the figure, we can observe that:

- These countries with higher GDP per capita have deaths per million rate.
- But some extreme poverty countries have lower deaths per million rate.

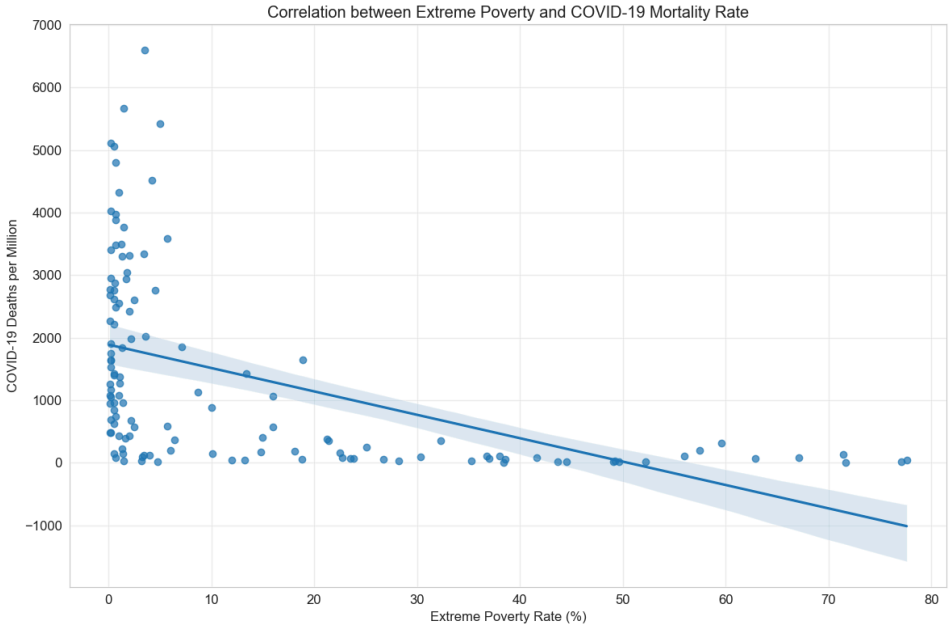

Figure 9: Correlation between Extreme Poverty and COVID-19 Mortality Rate

Variable	Coefficient	Std Error	t-value	P-value
Constant	1646.21	635.21	2.592	0.011
GDP per capita	0.0201	0.009	2.351	0.020
Extreme poverty	−28.99	7.30	−3.974	0.000
Stringency index	−5.08	12.56	−0.404	0.687

The regression analysis tested two main hypotheses:

- H0: There is no significant relationship between extreme poverty and COVID-19 mortality rate H1: There is a significant relationship between extreme poverty and COVID-19 mortality rate
- H0: There is no significant relationship between GDP per capita and COVID-19 mortality rate H1: There is a significant relationship between GDP per capita and COVID-19 mortality rate

Based on the p-values ($p < 0.001$ for extreme poverty and $p < 0.05$ for GDP per capita), we reject both null hypotheses, indicating statistically significant relationships exist.

From the regression results, we can observe that:
- The model explains 28.2% of the variance in COVID-19 mortality rate (R-squared = 0.282)
- GDP per capita has a positive relationship with mortality rate (coef = 0.0201, $p < 0.05$)
- Extreme poverty has a significant negative relationship with mortality rate (coef = −28.99, $p < 0.001$)
- Stringency index shows no significant relationship with mortality rate ($p = 0.687$)

Note that:
- The condition number is 1e+5, which is very high. This indicates that there is multicollinearity in the data.[3]

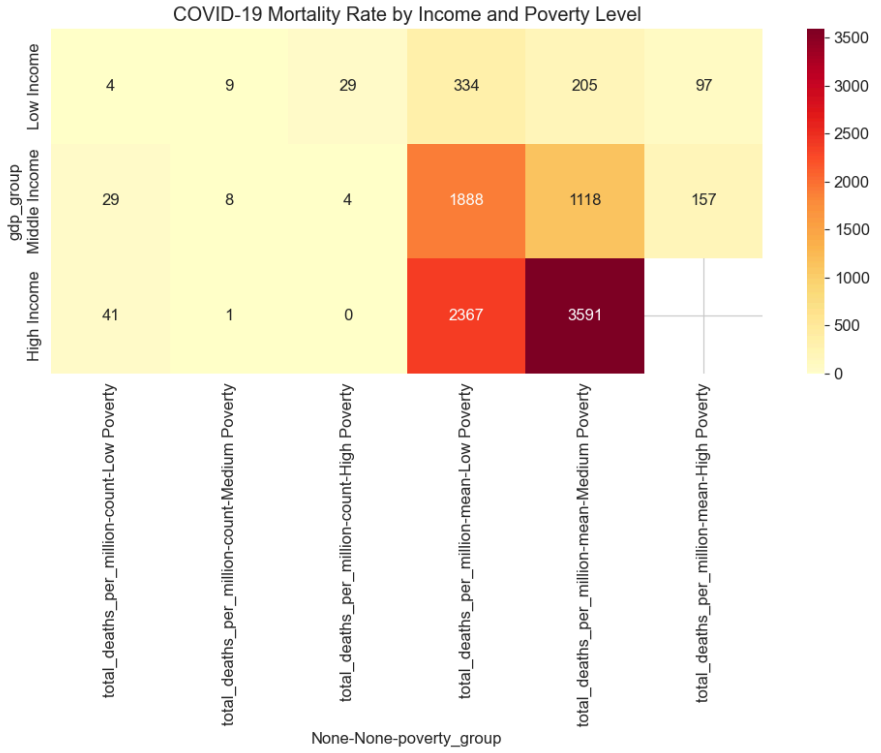

Figure 10: Heatmap of Extreme Poverty and COVID-19 Mortality Rate

From the heatmap, we can observe that:
- The extreme poverty and COVID-19 mortality rate are negatively correlated.

[3]I'm considering use other models like Generalized linear model(GLM)

3.3.3. Interpretation

- Same as the previous research question, it's anti-intuitive that the extreme poverty and COVID-19 mortality rate are negatively correlated.
- This might be due to:
  - Population structure
    - The population in extreme poverty are older, and thus more vulnerable to COVID-19.
  - Healthcare system
    - The healthcare systems in extreme poverty countries do not accurately report the deaths.

4. Conclusion

This study analyzed the OWID COVID-19 dataset to understand the pandemic's impact on global health and economy. Several counter-intuitive findings emerged[4]:

- Despite having better healthcare systems, developed countries showed higher COVID-19 mortality rates than developing nations **statistically**.
- Countries with higher life expectancy experienced higher mortality rates **statistically**.
- Extreme poverty showed a negative correlation with COVID-19 mortality rates **statistically**.

These seemingly paradoxical findings can be explained by several factors:

- **Reporting Standards**: Developed countries likely have more accurate death reporting systems, while underreporting may be more common in developing nations
- **Population Demographics**: Higher life expectancy countries tend to have older populations, which are more vulnerable to COVID-19
- **Healthcare Access**: While developed countries have better healthcare systems, they may also have higher rates of comorbidities and elderly populations
- **Testing Capacity**: Developed countries conducted more comprehensive testing, leading to higher case detection rates

The study also revealed that:

- Policy efficiency (measured as mortality rate per unit of stringency) was higher in high-income countries

These findings highlight the complex relationship between health and economic indicators during the COVID-19 pandemic, suggesting that traditional assumptions about healthcare system effectiveness may need to be reconsidered in the context of global health crises.

[4] I do not fully sure the code is correct. Because the result that 'Developed countries showed higher COVID-19 mortality rates than developing nations' is not very much aligned with my intuition.

Statement of Original Authorship

I, Yuzhe Shi, hereby declare that this report is my original work and has not been submitted for assessment in any other context. All sources of information have been duly acknowledged and referenced in accordance with the academic standards of the South East Technological University.

Signature(Seal): ..... **Date:** 22 March 2025

Bibliography

[1] "Our World in Data." [Online]. Available: https://ourworldindata.org/°

[2] "Our World in Data - COVID-19." [Online]. Available: https://github.com/owid/covid-19-data/°

[3] "World Bank." [Online]. Available: https://data.worldbank.org/°

Index of Figures

Index of Tables

[5]These has 0 stringency index values are removed