CMPT 353 project - Data Analysis Report: COVID-19 Global Impact Assessment

Group member: Troy Zhu, 301249148

Executive Summary

This report outlines the analytical approach and findings from a comprehensive study on the global impact of COVID-19, focusing on the relationship between a country's Gross National Income (GNI) per capita and its COVID-19 death rate. Using time series data on the pandemic alongside economic indicators, we investigated whether higher national income correlates with the severity of the pandemic's impact, measured in terms of deaths per 100,000 people.

Introduction

The COVID-19 pandemic has presented unprecedented challenges globally, with varying impacts across different regions. In this project we discover how the virus is spreading over the world from early 2020 to early 2023. A pertinent question that arises is whether a country's economic standing has influenced its ability to cope with the pandemic. To explore this, we conducted a data-driven analysis to understand the relationship between a country's wealth and its COVID-19 death rates.

Data Acquisition and Preprocessing

The primary datasets for confirmed cases, deaths, and recoveries were sourced from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series).

Additional economic data, specifically GNI per capita, was obtained from the World Bank using their API.

The data was rigorously cleaned to ensure accuracy and relevance. This involved:

- Dropping irrelevant geographical coordinates.
- Aggregating data at the country level to account for total confirmed cases and deaths.
- Merging COVID-19 data with population statistics to calculate per capita rates.
- Sourcing and integrating World Bank data on GNI per capita for the year 2021.

Analytical Techniques

Our analysis entailed:

- Descriptive statistics to summarize the current state of the pandemic.
- Correlation analysis to assess the strength and direction of the relationship between economic indicators and death rates.
- Regression modeling to quantify the relationship and predict death rates based on GNI per capita.
- Log transformation to linearize the data and mitigate the effects of skewness.

Findings

Let's work through our findings along with visualizations.

We first obtained the datasets from Johns Hopkins University. Our datasets is consisted of 4 csv files:

- time_series_covid19_confirmed_global: daily reported cases at country/province level
- time_series_covid19_deaths_global: daily reported deaths at country/province level

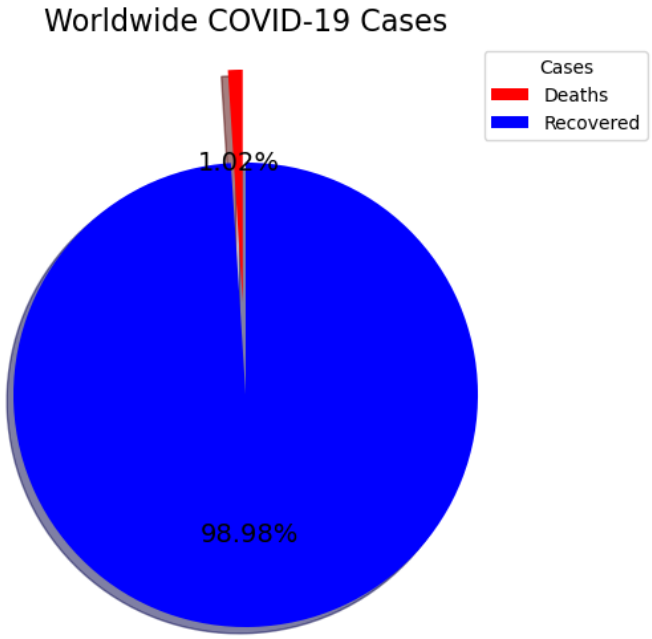
After gather the data, we dropped the `Lat` and `Long` columns as these are not needed. Since each dataframe has more than 1000 columns (one column per day from 01/2020 to 03/2023), We transformed them into two new dataframes with total number instead. We also calculate the recovered number of cases and mortality rate, then we have a table looks like following:

	Country	Total Confirmed	Total Deaths	Total Recovered	Mortality Rate (%)
0	Afghanistan	209451	7896	201555	3.77
1	Albania	334457	3598	330859	1.08
2	Algeria	271496	6881	264615	2.53
3	Andorra	47890	165	47725	0.34
4	Angola	105288	1933	103355	1.84

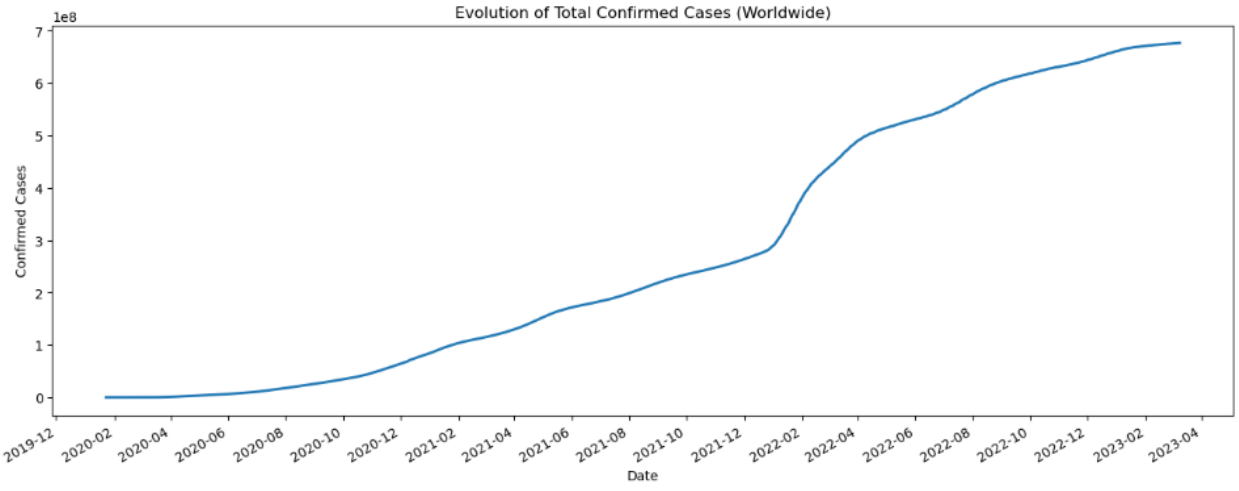
We then summed up the number for all countries with a final worldwide totals:

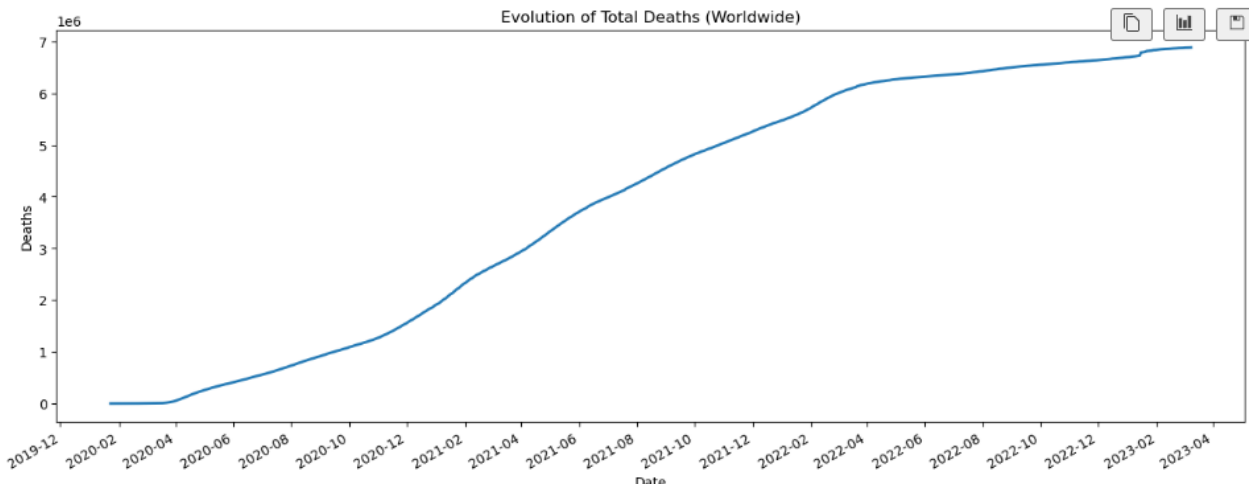
	Number of Cases
Total Confirmed	676570149
Total Deaths	6881802
Total Recovered	669688347

Putting the number into a pie chart and we get:

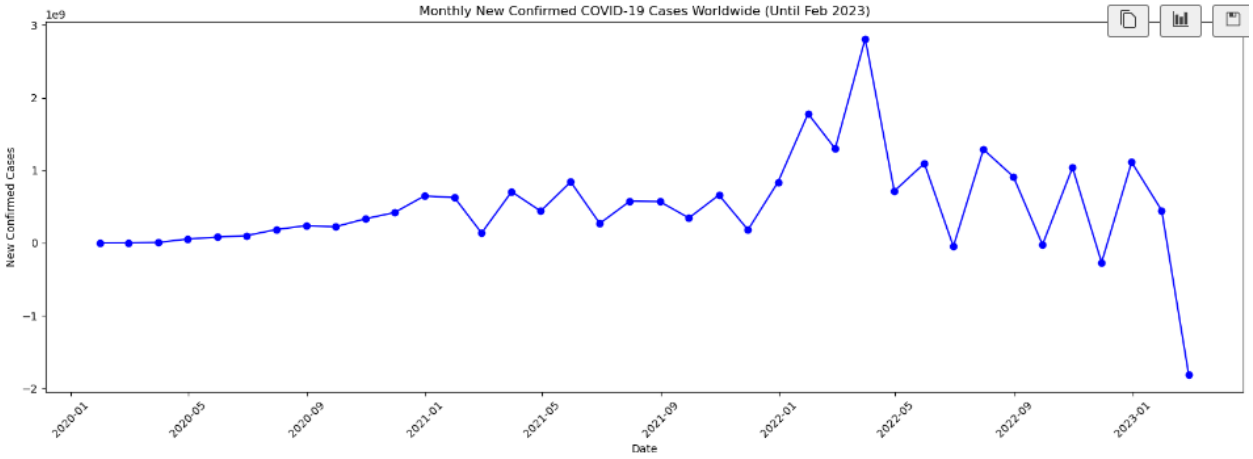


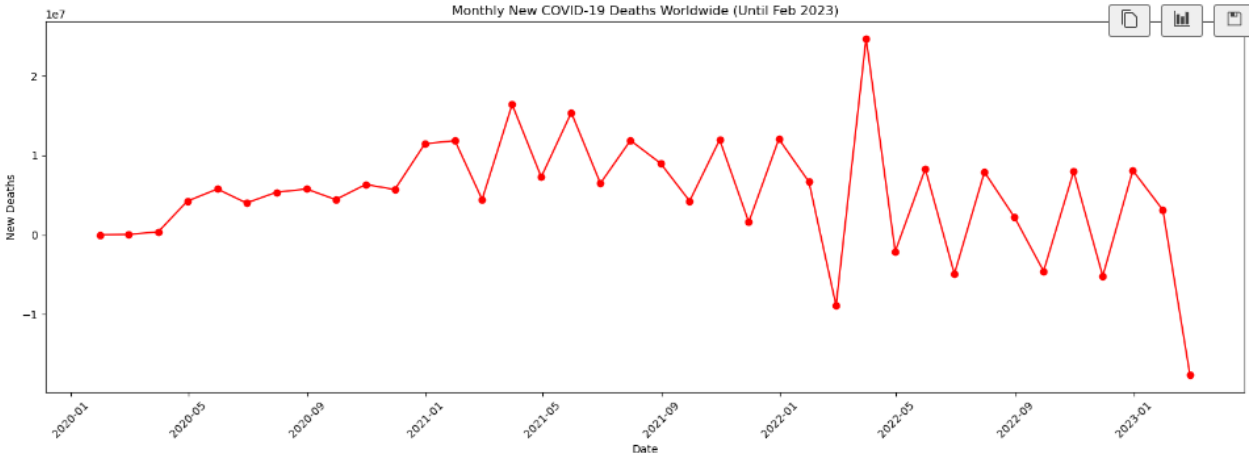
We then want to see how the COVID was spreading over the world. For each confirmed cases and deaths cases, we constructed a line chart for each dataframe:





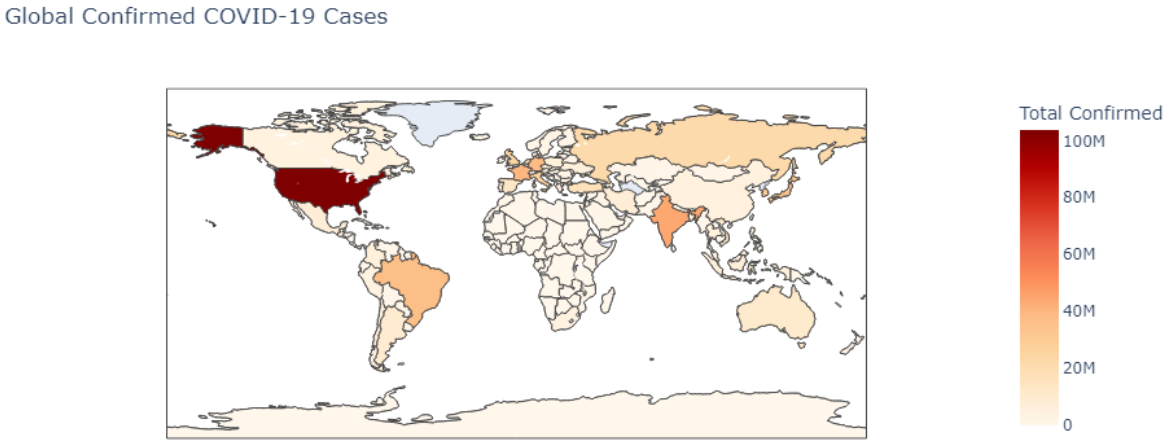
From the charts above, it looks like the confirmed cases jumped up quite a lot in early 2022. To further investigate this, let's plot the change of new cases each month instead:

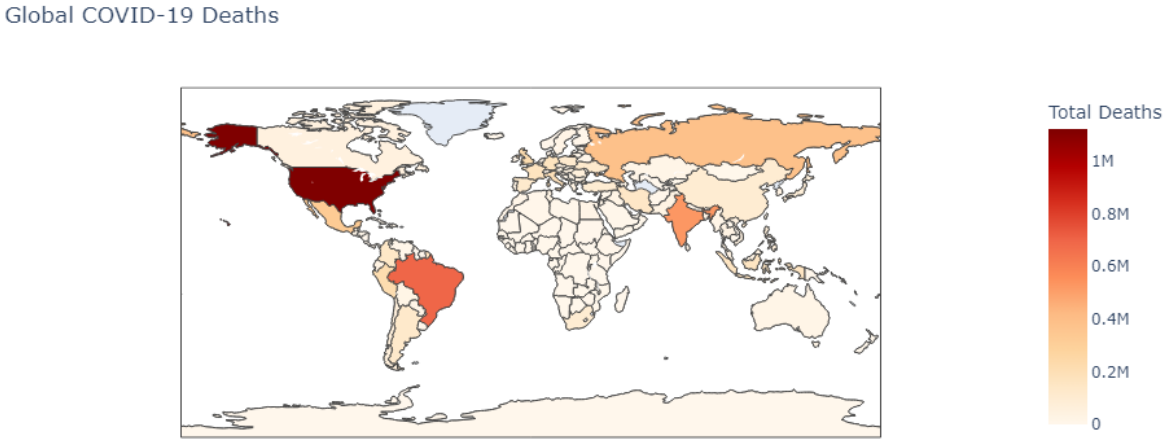




It looks like indeed the number of confirmed and deaths cases had a jump between 2022-01 and 2022-05. Combining with the fact that the Omicoron variant was discovered in late 2021, we have a reason to believe that Omicoron was possibly the main cause of the sudden jump.

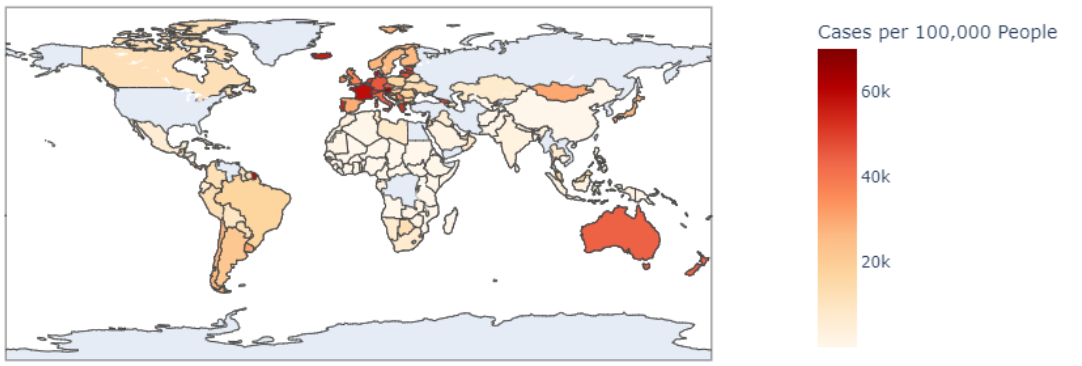
So we know how fast the COVID was spreading, we then want to find out which countries had most cases. In this scenario, we used Python `plotly` library to create a map distribution:



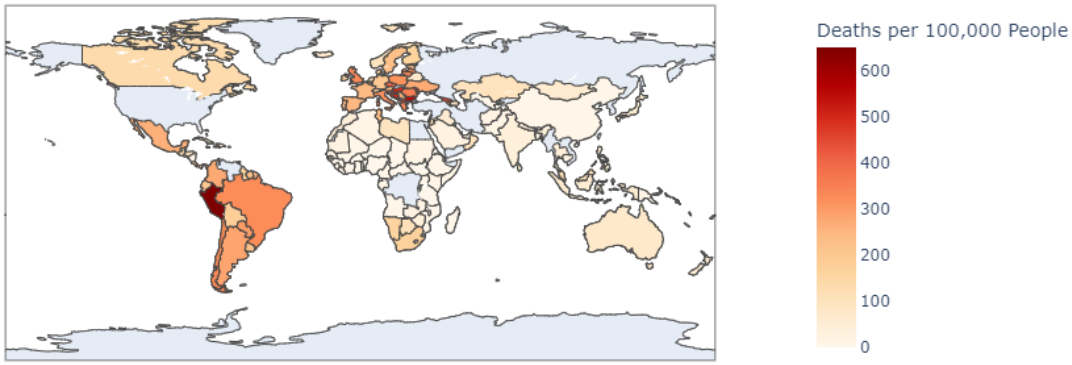


From the map distributions, it looks like US had most confirmed and deaths cases followed by Brazil. However, the population of a country could impact the results above as many countries have less population but with higher proportion of people got infected. To resolve this issue, we then used cases per 100,000 people instead of total population:

Global Confirmed COVID-19 Cases

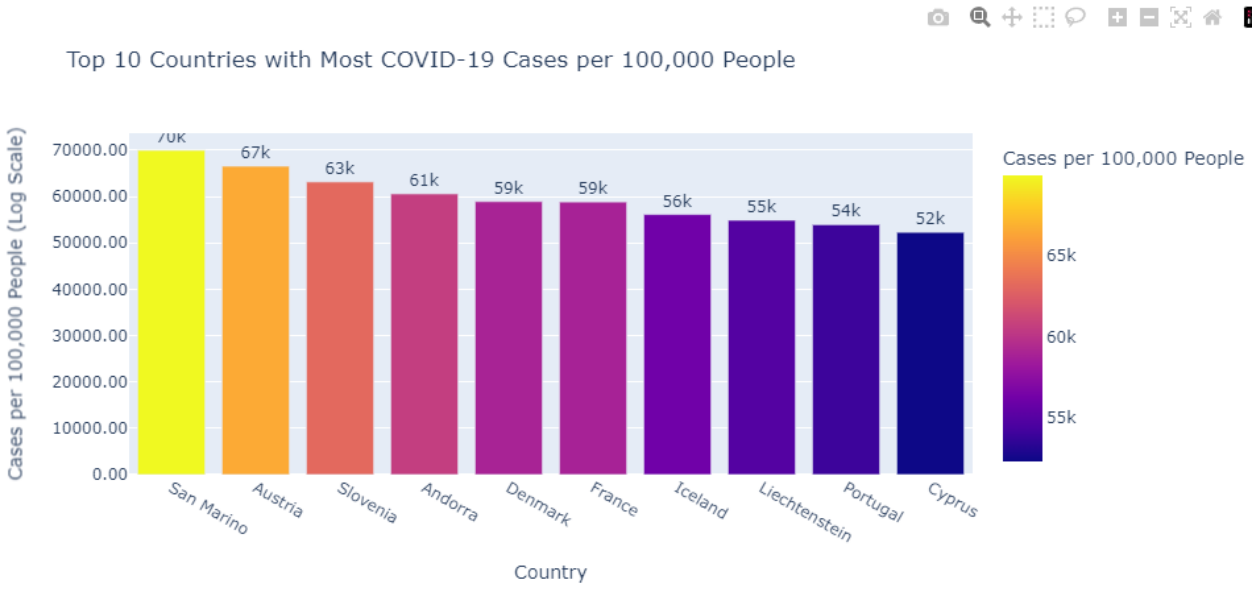


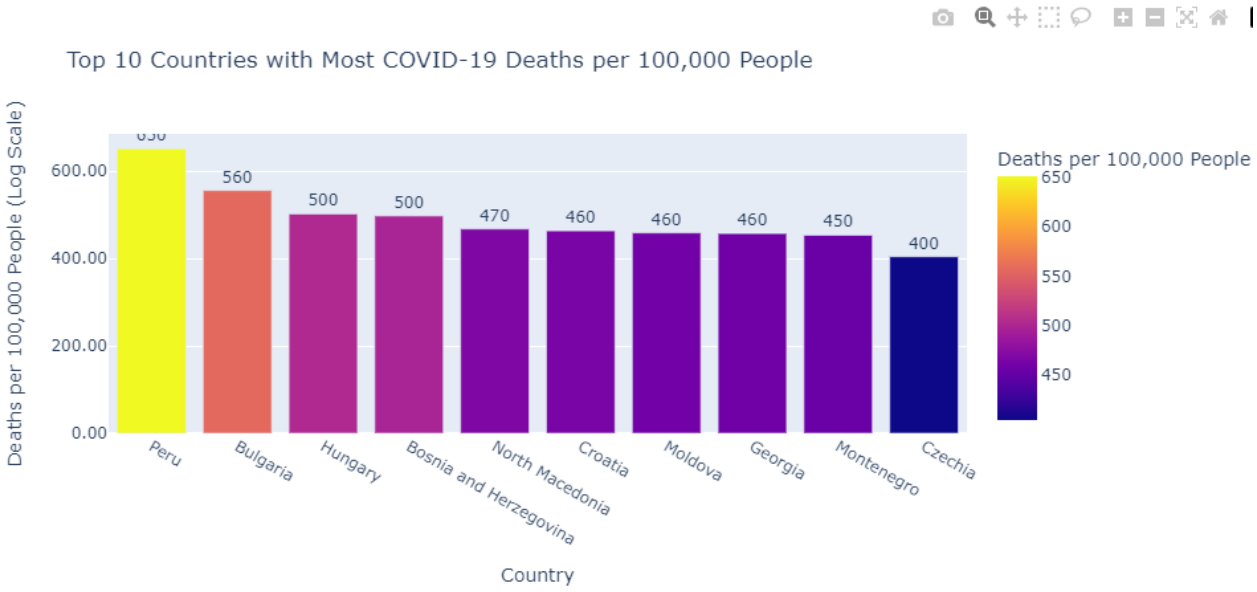
Global COVID-19 Deaths



After using cases per 100,000 people, we can tell that US is no longer the highest, and it looks like Latina America and Europe were worse than before.

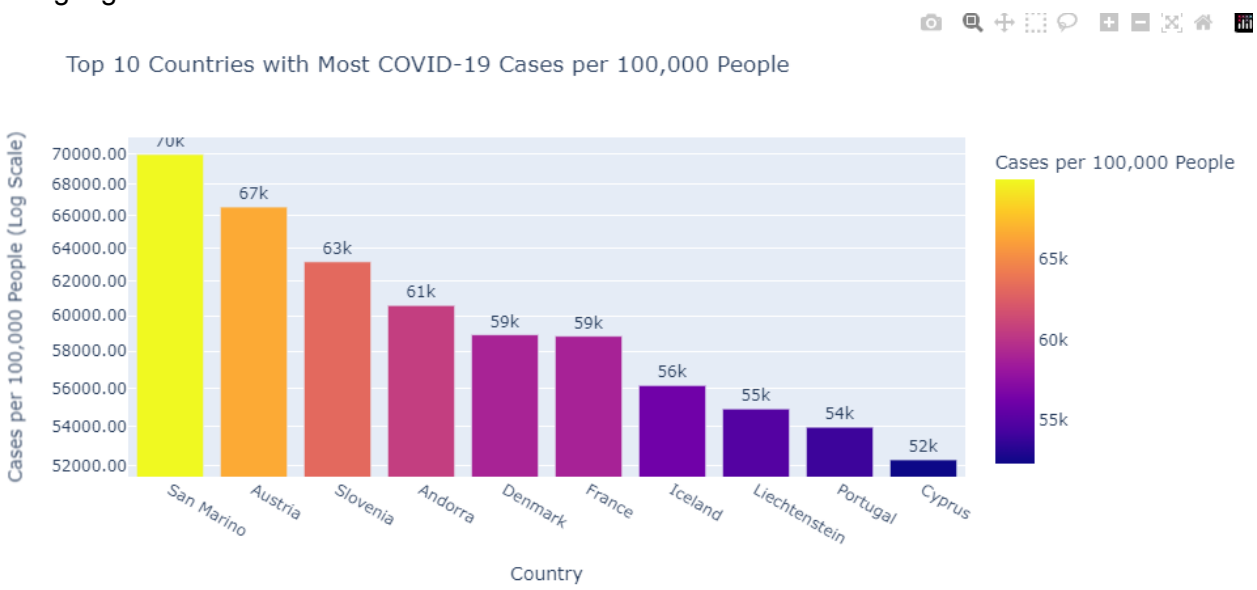
To further investigate this, we wanted to list out the top 10 countries with most confirmed and death cases:

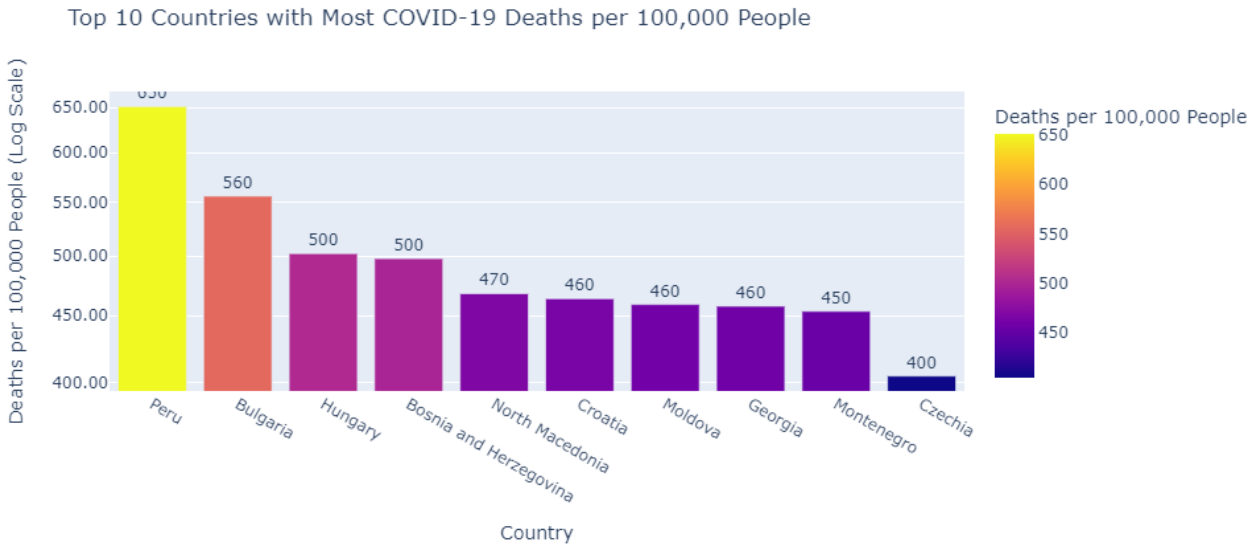




From the above plots, althoug we have found the top 10 countries with most cases, there are not obvious difference between them and it was hard to tell. Therefore, we transformed the plots

using logscale instead:





After changing the scale into logscale, the results are much more obvious now. Out of 10 countries with most confirmed case per capita, 9 of them are in Europe wich indicated that European people suffered most during COVID season.

Another question that we want to explore is that whether the economy of a country has an impact on the death rate of COVID.

To start, we first use Python `pandas_datareader` to fetch the Gross National Income (GNI) from World Bank, and then joined this with the original dataframe we created to make sure each country has the correct data. Then we found the correlation between the death rates and GNI which is 0.22, which refers a slightly positive correlation.

Continue on this topic, we then use the Python `statsmodel` to get some more information, and we obtained the following OLS Regression Results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Deaths per 100,000 People   R-squared:                       0.051
Model:                                 OLS   Adj. R-squared:                  0.045
Method:                      Least Squares   F-statistic:                     8.710
Date:                     Wed, 29 Nov 2023   Prob (F-statistic):            0.00363
Time:                             19:58:05   Log-Likelihood:                -1037.8
No. Observations:                      164   AIC:                             2080.
Df Residuals:                          162   BIC:                             2086.
Df Model:                                1
Covariance Type:                 nonrobust
==============================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          106.7579     13.232      8.068      0.000      80.628     132.888
GNI per Capita   0.0016      0.001      2.951      0.004       0.001       0.003
==============================================================================
Omnibus:                       45.167   Durbin-Watson:                   2.139
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               74.530
Skew:                           1.421   Prob(JB):                     6.55e-17
Kurtosis:                       4.682   Cond. No.                     3.06e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.06e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

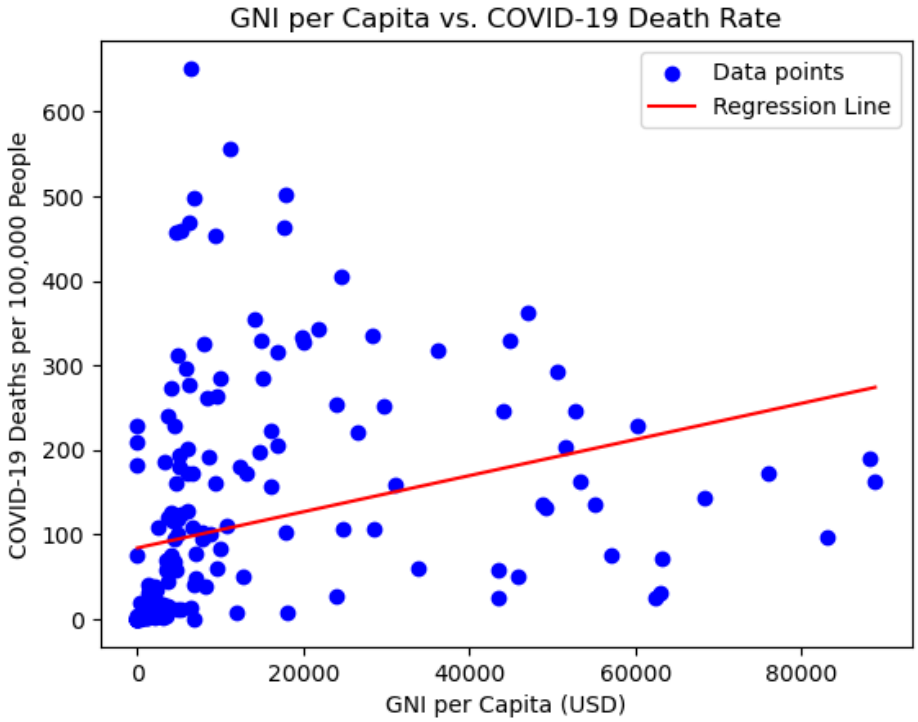
To interpret the result, we discovered:

R-squared: The R-squared value is 0.051, which means that approximately 5.1% of the variance in the COVID-19 death rate is explained by GNI per capita. This is a relatively low value, suggesting that GNI per capita alone does not strongly predict COVID-19 death rates.

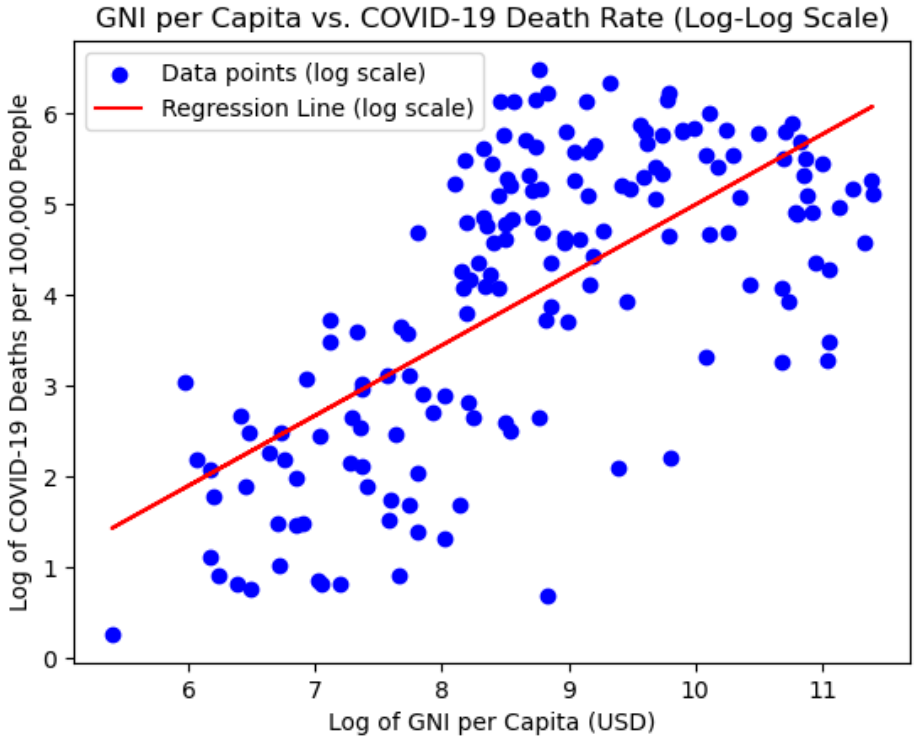
P-Value: The P-value for the GNI per capita coefficient is 0.004, which is less than the conventional alpha level of 0.05. This indicates that the relationship between GNI per capita and COVID-19 death rate is statistically significant.

F-statistic and its P-value: The F-statistic is 8.710 with a probability of 0.00363. This indicates the overall significance of the regression model. A low P-value here (below 0.05) suggests that the model is statistically significant.

To further exploring on this, we used Python `sklearn` and ran a linear regression model on this. Then we generated scatter plot and the linear regression:



As you can see, the scatter plot does not reveal much relationship for now, we then tried it with log-log scale:



When use log-log scale, graphically, there is a clear linear relationship between the GNI and death rate, which is actually unintuitive to think about. Ideally, we would expect wealthy countries have better medical treatment availble and hence the death rate should be lower.

The analysis revealed a statistically significant, yet weak, positive correlation between GNI per capita and COVID-19 death rates. The linear regression model, after log transformation to address non-linearity, indicated a higher explanatory power, suggesting that wealthier countries tend to have higher reported death rates. However, this relationship is complex and warrants further examination.

Given the nature of the data, this relationship could be due to high-income countries having more developed health systems that are better at recording actual COVID-19 mortality figures, whereas low-income countries might underreport due to limited testing and reporting infrastructure.

It's also possible that higher-income countries have older populations, which are more susceptible to severe outcomes from COVID-19, hence higher death rates.

Limitations and Challenges

The study faced several constraints:

- Variability in reporting standards and testing rates across countries.
- The exclusion of potential confounding variables such as healthcare system quality, government response, and population age structure.
- The assumption of a linear relationship in regression models, which may not capture the complexity of the pandemic's dynamics.

Given more time and resources, we would expand our model to include more variables and employ more sophisticated statistical techniques, such as multivariate regression or machine learning models, to improve predictive accuracy.

Conclusion

The preliminary findings suggest that a country's economic status, as measured by GNI per capita, has a measurable, albeit small, impact on COVID-19 death rates. However, due to the exploratory nature of this analysis and its limitations, we recommend a cautious interpretation of these results and suggest further investigation into the multifaceted effects of the pandemic.

Project Experience Summary

COVID-19 Global Impact Analysis Project

- Spearheaded a data analytics project to investigate the impact of GNI per capita on COVID-19 death rates using Python and statistical modeling, enhancing understanding of the pandemic's economic correlations.
- Managed data pipelines for preprocessing and analysis of extensive COVID-19 datasets from reputable sources, including Johns Hopkins University and the World Bank.
- Applied statistical techniques, including correlation analysis and regression modeling, to uncover subtle trends and provided a log-transformed data model to better interpret non-linear relationships.
- Synthesized findings into accessible visualizations and reports for stakeholders, contributing to data-driven decision-making processes.
- Navigated through data discrepancies and reporting inconsistencies, ensuring high data integrity and reliability in findings.
- Documented and communicated technical methodologies and insights effectively, balancing detail with high-level summarization to cater to diverse audiences.