COVID-19 Exploratory Data Analysis: Analytics Write Up

SMU Data Analytics Bootcamp | Project 1

Group 1: Alyssa DiFurio, April Gao, Garrett Kidd

November 21, 2022

Introduction

The COVID-19 global pandemic affected virtually everyone on the planet in the past few years. Due to the disease's potential for rapid spread and harm, it has been studied extensively to understand how to mitigate its impact on populations around the world. Furthermore, countries around the world have had drastically different responses to the disease, with varying outcomes. Every country is unique in its GDP, Population Density, and quality of Health Care System. Our group's intent is to understand if these factors had an influential effect on a country's total infection and death count over the course of the COVID-19 pandemic.

We decided to select ten countries from around the world, to explore the COVID dataset from Johns Hopkins University. Our countries included: Austria, Italy, Portugal, Greece, Taiwan, Indonesia, India, Mexico, Thailand, and Iran. We chose these countries because five of them were in the top 10% for countries with case and death totals, and five of them were in the bottom 10% for case and death totals. By choosing countries from both ends of the spread of data, we felt that we could accurately model how countries, with different GDP, Population Density, and quality of Health Care Systems, would fare against dealing with the COVID-19 virus.

Figure 1 Google Maps API map of ten countries included in the EDA.

Data Acquisition and Cleaning

The data for our exploratory analysis was found on Kaggle.com. We elected to used the "COVID-19 data from Johns Hopkins University", as it has been updated daily since the beginning of the pandemic, was already in a useable CSV style format, included all countries in the world with a records for deaths and infections per day, and this dataset was also the same that has been referenced by multiple reputable news outlets and governmental and non-governmental organizations as an authoritative source for COVID-19 reporting. Within this dataset, we used two CSV tables of data, one for worldwide infections, and another for worldwide deaths.

The second dataset we referenced for our analysis also came from Kaggle, which was the "The World Factbook by CIA". The World Factbook is produced by the Central Intelligence Agency, as an "at-a-glance" reference for US policymakers, detailing facts and figures about countries around the world, ranging from population to economic to political to health care figures. The World Factbook provided data we used for establishing the numeric value of our

countries of interest's GDP, Population Density, and quality of Health Care Systems. This dataset was able to be downloaded as a massive CSV file, which required some cleaning and reformatting to make it useable for our purposes. Some of the data within the World Factbook was collected prior to the COVID-19 pandemic, however we made the determination that it was still relevant to proceed with, given our time constraints and veracity of this project. Fields from the World Factbook that we used included: Current Health Expenditure 2019, Physicians Density per 1000 2020 to 2018, Hospital Bed Density per 1000 2018 to 2017, Geography Area Total Sq Km, Real GDP per Capita 3 Year Average 2020 to 2018, GDP Official Exchange Rate in Billions, Population, and Real GDP Purchasing Power Parity 3 Year Average.

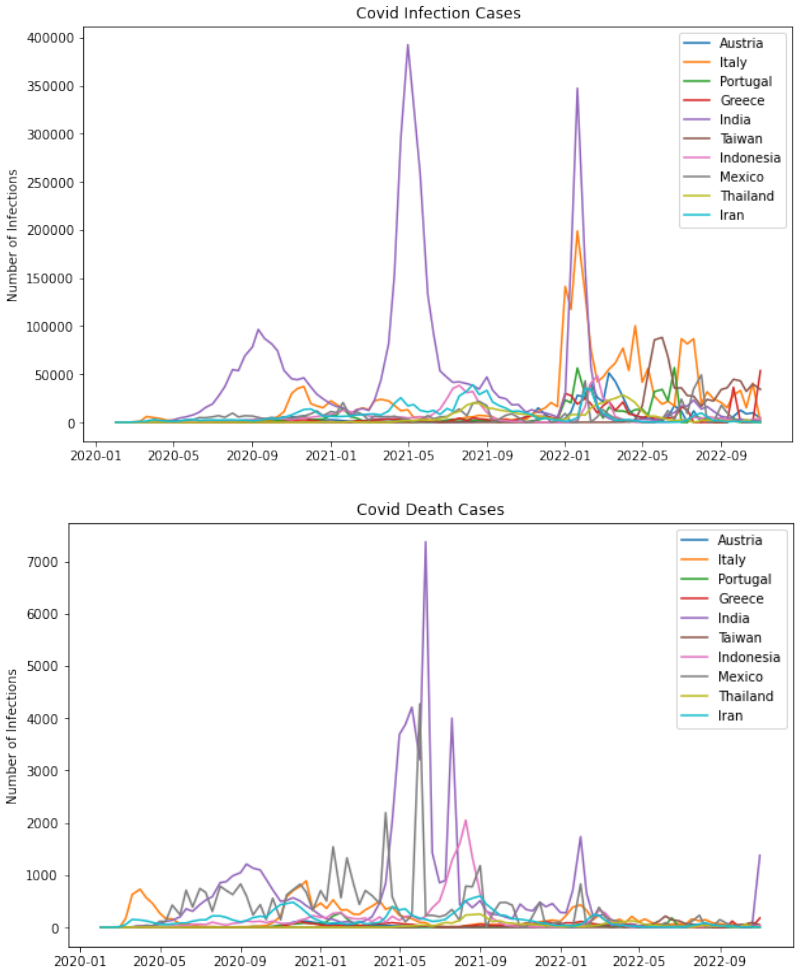

Figure 2 COVID-19 Infection and Death counts for ten countries of interest, entire study period (23JAN2020 to 6NOV2022).

Additional steps in data cleaning we undertook included creating several CSV files with reorganized columns and rows. In order to have CSV style data that would work well for transforming into Pandas data frames, we needed to move the names of countries to row values, as opposed to header/title values, with one column for "Countries". We accomplished this in a combination of Pandas functions and manual manipulation of the data within Microsoft Excel.

Research Question #1: Have COVID-19 infections and deaths been influenced by a country's GDP?

The first part of the research is finding out if a country's economic position measures a country's ability to fight the virus. Wealthier countries are preforming better at controlling COVID death. We decided to compare countries' average GDP per Capita from 2018 to 2020 to their COVID death rate to see if there is in fact a relation between the two. We summed each country's average GDP per Capita together to get the total. Then each country GDP per Capita ratio is calculated using their individual GDP per Capita divided by the total GDP per Capita. We then time the ratio by 10,000 to minimize the decimals for a cleaner presentation. The death rate is simply dividing each country's COVID death count by their COVID infection count. We then time the ratio by 1000 to minimize the decimals.

First, we used a Bar chart with the 10 sample countries, comparing their death rate and their average GDP per capita ratio as shown below. GDP per capita is orange and Death Rate is blue. Just from looking at the chart below, we can say when GDP per capita is higher the death rate is lower. From right to left, the death rate goes higher as the GDP per capita drops down.

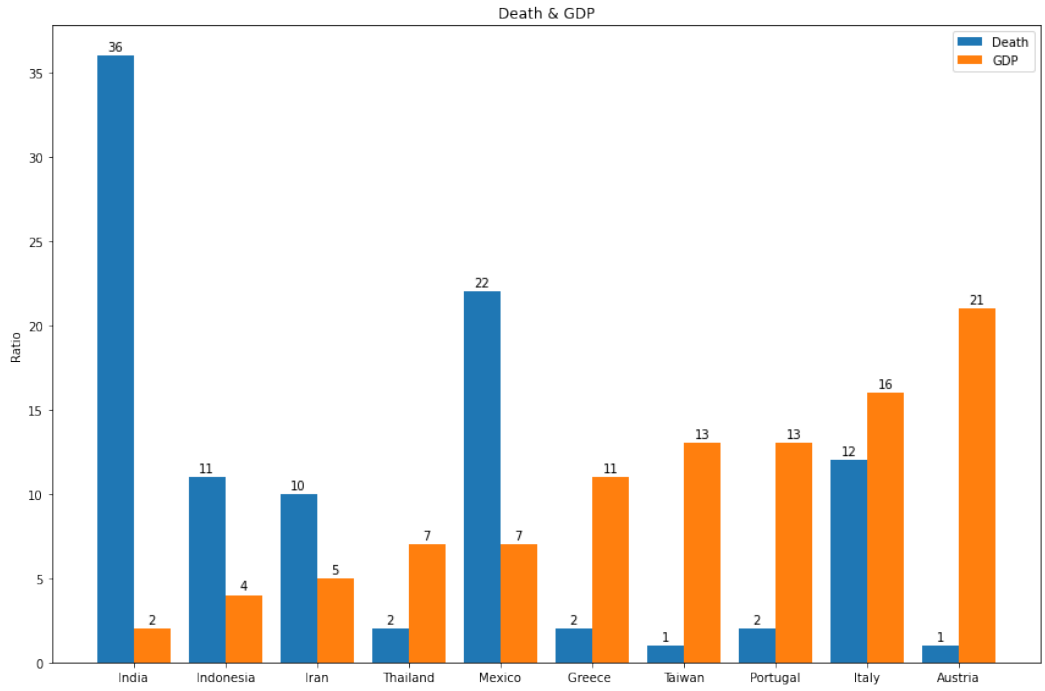

Figure 3 Bar chart reflecting the ratio of COVID-19 death to country GDP.

To further approve our conclusion, we pulled all the countries into a scatter chart below. We grouped the countries by continents. They are North & Central America in blue, Australia in orange, Asia in green, Africa in red, Europe in purple, South America in brown. The scatter chart clearly shows when the GDP is higher death rate is lower, when the death rate goes higher GDP falls to the lower part of the chart except Africa. Africa's countries are scattered into a red flat line flows on the bottom of the chart. GDP does not react to death rate. However, most, if not all the countries in Africa were under many years of colonialism, civil war, and government corruption which led to deep poverty and inequality. Most definitely there are many challenges in economic growth and maintaining a healthy, wealthy, and productive society. Underdeveloped healthcare system on top of poverty, is the reason why the death rate is not reacting to the GDP ratio.

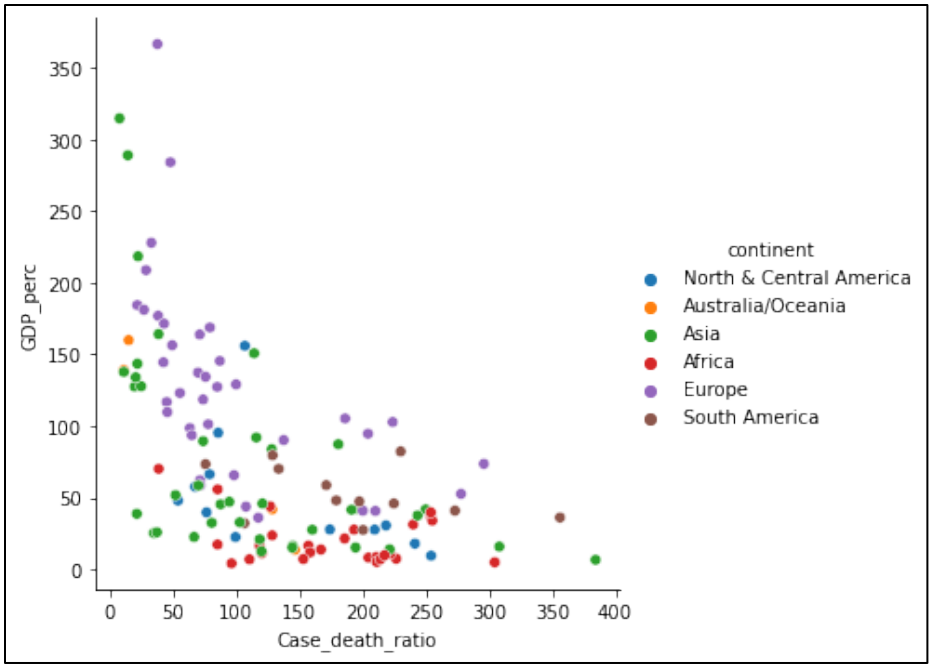

Figure 4 Scatter plot showing the ratio of COVID deaths and country GDP.

At last, we feed the GDP ratio and death rate into a regression model. Our independent variable X is death ratio, and the dependent variable y is GDP per capita ratio. Initially the R – squared is at 0.14 nearly to 0 as shown in the chart below. There are outliners in the chart is driving the fit not working out. From looking at the data, the largest outliner is Yemen. It has a death rate of 187% meaning it has almost twice of death than infection. Since is impossible for a country has more death than infection. A person must get infected to COVID before she or he included in the death count. We decided to exclude 7 countries with more than 40% death rate. After eliminating the 7 outliners, the rest of the 130 countries arrived at the R- squared of 0.34. even though the R- squared is not high, but with the 2 charts above we can conclude there is a negative relationship between GDP and death rate.

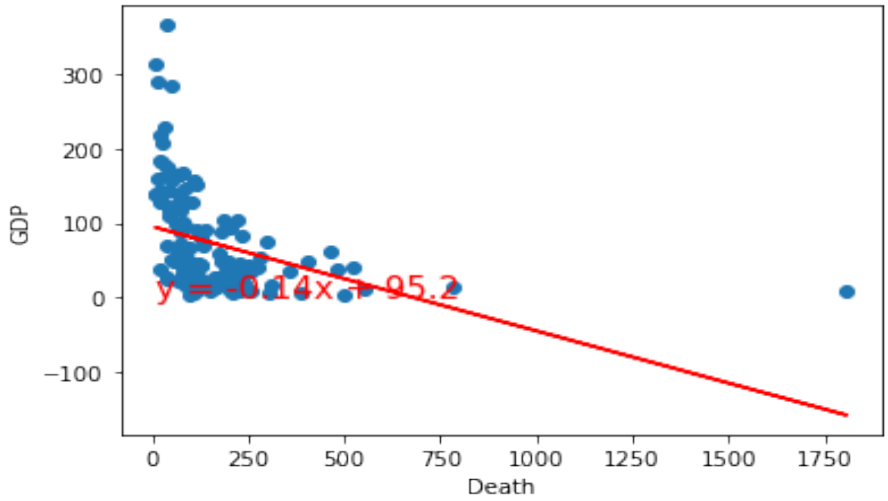

Figure 5 The R-Squared is: 0.1466576.

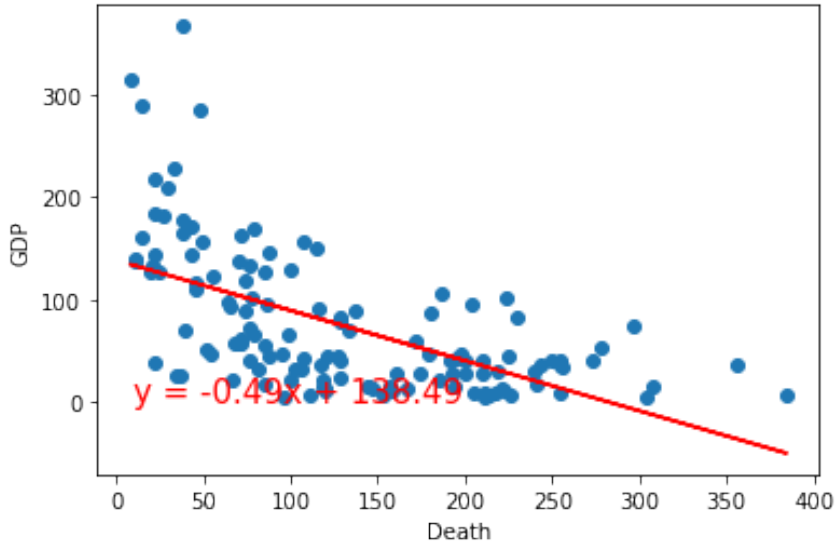

Figure 6 The R-Squared is: 0.3449913.

Research Question #2: Have COVID-19 infections and deaths been influenced by a country's population density?

Early in the pandemic it was understood that the COVID-19 virus was easily transmissible through the air and being near other people was discouraged. One of our hypotheses was that countries that had a higher population density (population total / country area sq km) would experience higher COVID infections and deaths. The results of our analysis

using the ten countries we selected had mixed results. The scatter plots produced for the analysis show a slight correlation in population density and COVID infections and deaths, however several of the countries skewed the expected result, in particular India and Taiwan, which both have relatively high population density, however drastically different infection and death totals.

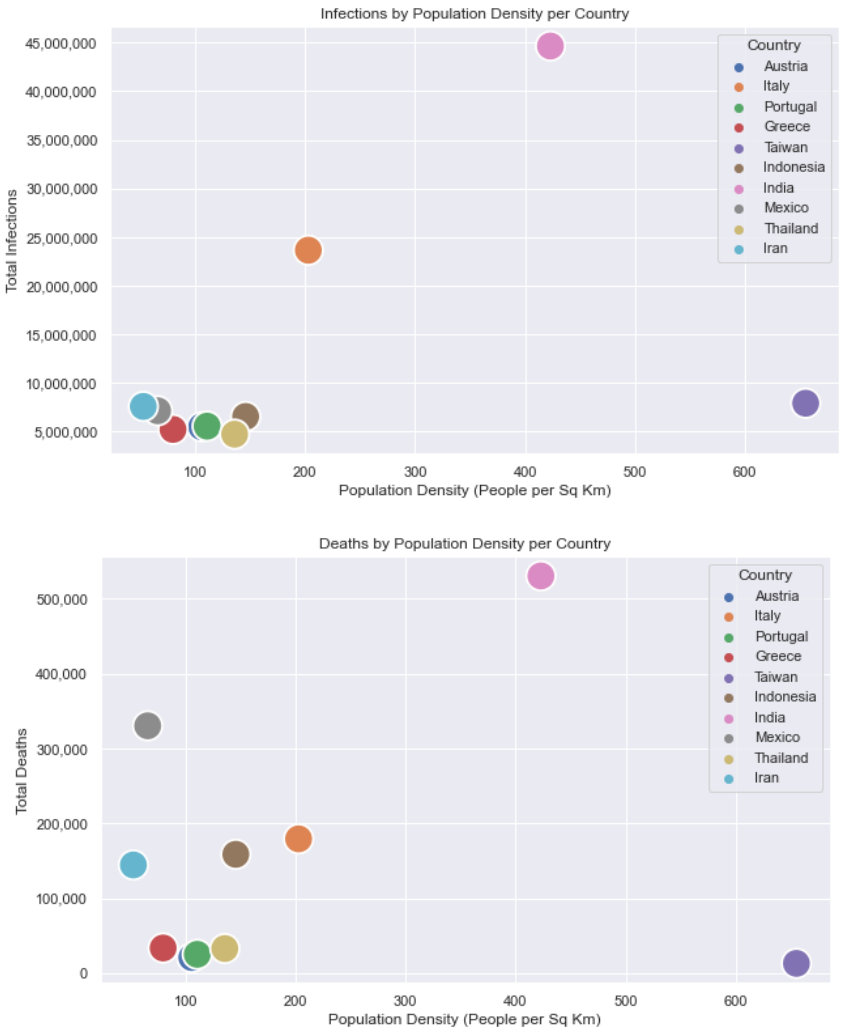

Figure 7 Scatter plots showing infection and death totals by country compared to the country's population density.

Because the scatter plots with our original ten countries of interest did not reveal any insights in to a possible correlation between population density and COVID infections and deaths, we decided to expand the analysis to include all of the countries for which we had data.

Some more data cleaning was required in order to produce workable data frames, however this was able to be accomplished with some time spent in Microsoft Excel. With data frames for all countries we had data for, we were able to visualize this data by way of scatter plots. Initially, the data contained some outlier data with very high case and death counts that skewed the plots and made them difficult to interpret. However, after removing these outliers, we were able to produce two scatter plots, one for infections and one for deaths, that were much more legible.

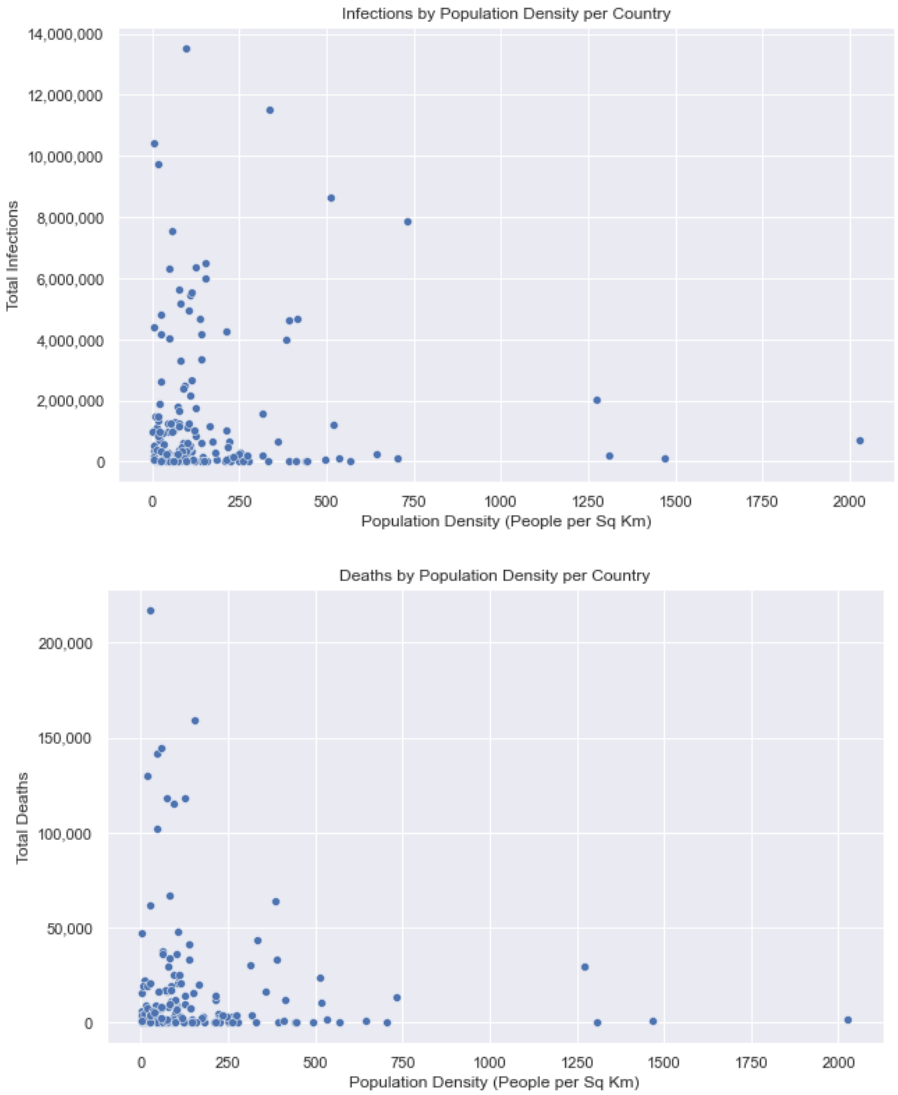

Figure 8 Scatter plots for COVID infections and deaths compared to country population density for over 170 countries.

While these new scatter plots comparing infection and death totals to population density included seventeen time more data, the were still unable to show any real correlative connection. With outlier removed, there appeared to be additional outliers, of a smaller magnitude. This analysis could probably be run again, with more outliers removed, however that would require additional qualitative evaluation to determine which data should be retained and considered.

Research Question #3: Have COVID-19 infections and deaths been influenced by a country's quality of health care system?

Utilizing a merged dataset of the World Factbook and John's Hopkins Covid Cases, we were able to utilize two variables as numeric values for our healthcare comparisons. We drew from the "hospital beds per 1000 people" and the "physicians per 1000 people" to gauge how the amount of healthcare availability would affect a country's overall infections and rates. This dataset is inconclusive for Taiwan due to our datasets missing information of these variables in the World Factbook dataset. This country was not considered in our final outcomes. Our first set of graphs allowed us to compare the impact of a country's physician density to their overall infection and death rates. Our data showed that in our higher infection rate countries, the physician density and potential access to healthcare was higher. Whereas in selected countries that had higher counts of death reported, the physician density and potential access to healthcare was significantly lower.

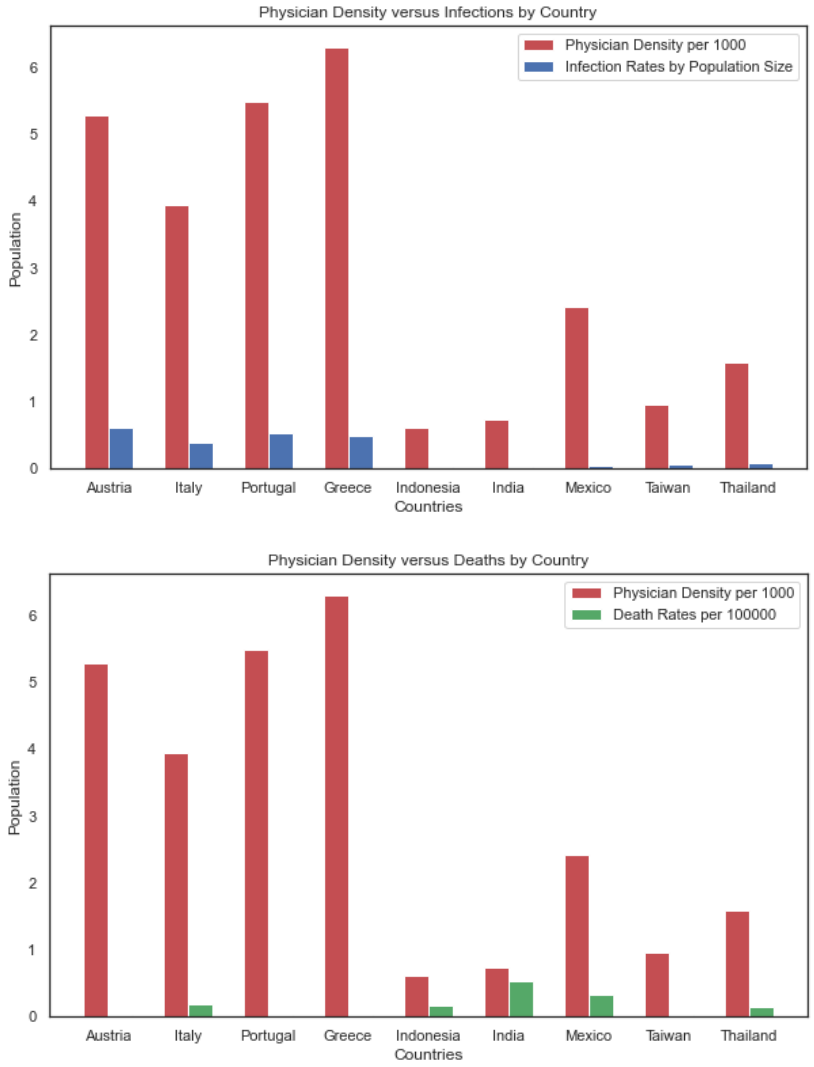

Figure 9 Bar charts showing physician density per country compared to COVID infections and deaths, per 1,000 people.

Our second set of graphs compared the impact of hospital bed density on infection and death rates in our selected countries. This graph proved to have similar results as our first; the countries with larger percentages of hospital beds had a larger rates of cases, and our countries that had less access to hospital beds and care had higher numbers of deaths.

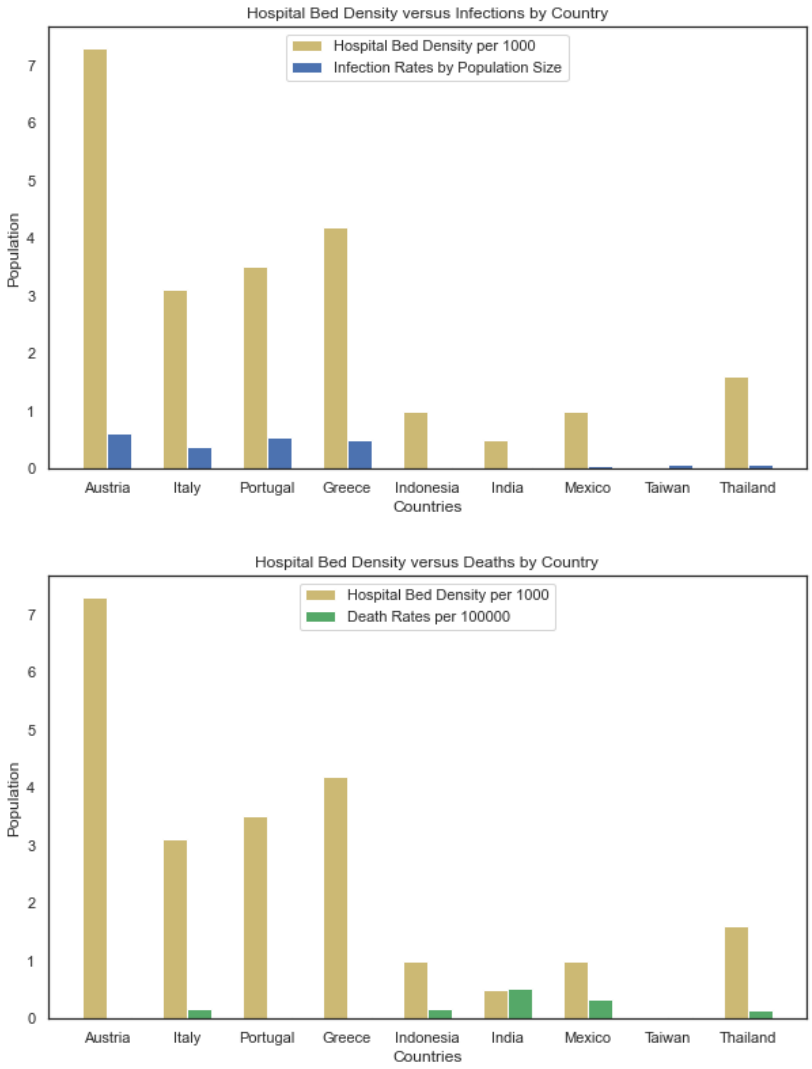

Figure 10 Bar charts showing Hospital bed density per country compared to COVID infections and deaths, per 1,000 people.

This data seems to reach a reasonable conclusion, countries that have more access to healthcare were more likely to have more citizens seeking help when sick and therefore more reported cases, or even just a larger population. Similarly, countries that didn't have as much access to healthcare or the capacity to keep citizens were more likely to have higher reports of death due to lack of ability to be seen or treated.

Regression Analysis

The regression that we focused on for our model was the infection cases over years. We used a linear regression model to predict the case outcome over time. Our initial graph showed to be less accurate, with a significant number of outliers. However, when we plugged our data into a Random Forest regression model, it narrowed our data into a cleaner set that lined up with our model more accurately.

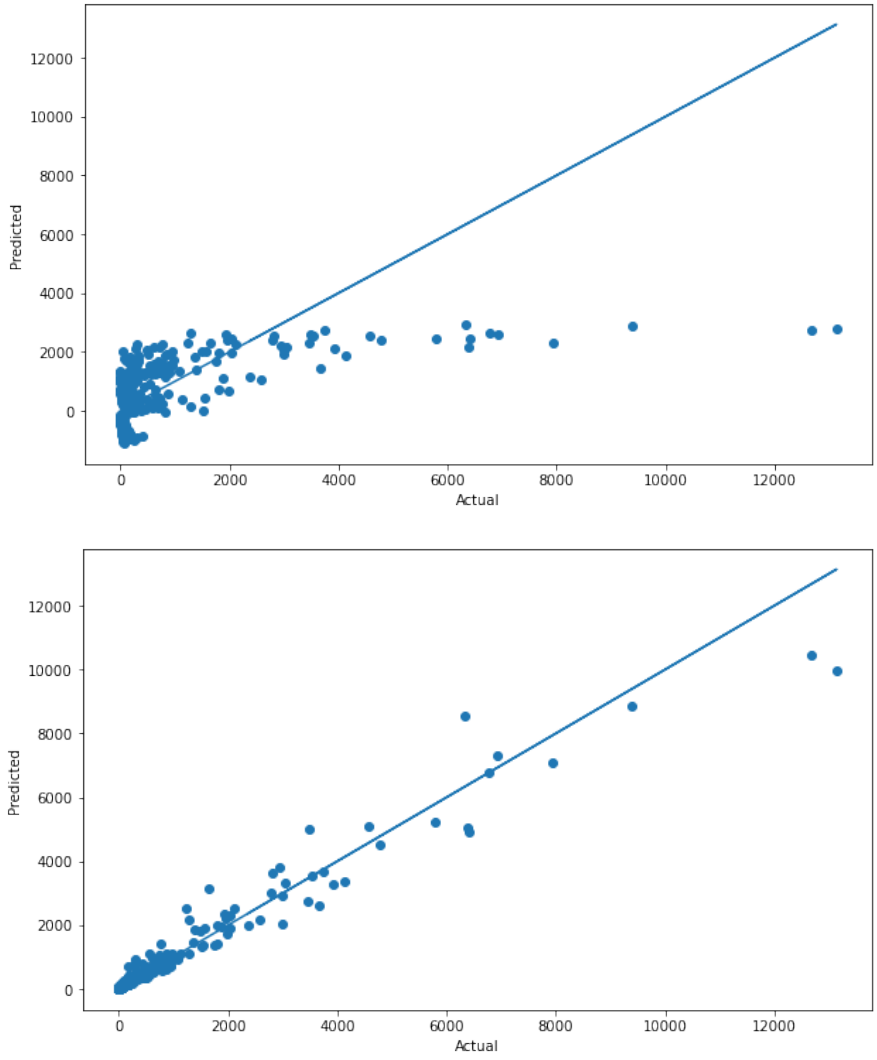

Figure 11 Regression models, with the Random Forest style regression at the bottom.

Neither model proved to be significantly accurate in predicting our datasets, and moreover, would have been more beneficial if the dataset was comparing our rates over months, which has a more trackable result than years. Overall, our regression model did not contribute much to our previous findings.

Conclusion

While the countries of the world continue to grapple with COVID-19, so does analysis of its data to better understand its impact and ways forward in the future. COVID-19 data will need to be continually reported and collected to maintain better accuracy in the future as it comes to analytic modeling. Through this group's analysis, we found that a country's factors such as economics and health care system quality tend to have more of an effect on death totals, but not on total infections. A country's GDP showed a negative relationship with death rates. Health care system quality within a country appeared to play a more significant role in the number of deaths compared to infection rates. Finally, population density had no discernable correlation with either infections or deaths and requires further study in the future.

Limitations & Bias

A significant drawback in our findings was the limitations caused by our CIA World Factbook dataset. This dataset contained values that were updated as information was found, meaning that not all data was current for the same year in every column. Some data was updated as far back as the beginning of the pandemic, with some being as current as the last month. This inconsistency put a limit on our data, however. With this dataset as well, was a lack of complete indexing of countries. This meant that we had to doctor which countries to use in our samples based on what could be found on this dataset as well as the covid set. Another issue in this

dataset was how messy and large it was. This made cleaning in python significantly more difficult.

We exhibited a bias towards countries that were more developed or had a larger economy in our samples which could have contributed to correlations we found with those factors in our datasets. To better explore these factors, we should have grabbed a larger sample size that had less bias and more reach across the world. A larger sample size would have helped our visuals more accurately identify trends.

Future Work

If our group was to do this project again, the primary change would be to our sample size. While our ten countries served a purpose for our intent, a larger sample would have provided an easier to identify visual trend for our research. This would allow us to find regressions and predictions more easily for our data as well. Limiting the data to year-based data sets would have also provided for a more distinct set of values that could be contextualized by vaccination timelines as well.

Going forward there needs to continue to be more studies into the effects countries' internal factors have on infectious diseases. Countries providing polls and surveys to citizens as they get vaccinated (or choose not to) helps contribute to these datasets and allow analysts more effective data to look at in the future.

Works Cited

COVID-19 Dataset: Worldwide Infections and Deaths, Johns Hopkins University

https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university?select=CONVENIENT_global_confirmed_cases.csv

https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university?select=CONVENIENT_global_deaths.csv

CIA World Factbook

https://www.kaggle.com/datasets/lucafrance/the-world-factbook-by-cia?select=countries.json