

# COVID-19 Exploratory Data Analysis

SMU Data Analytics Bootcamp

Project 1 | Group 1

Alyssa DiFurio, April Gao, Garrett Kidd

November 14, 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 affect on a country's total infection and death count during the COVID-19 pandemic. Data totals from 23JAN2020 to 6NOV2022, JHU.



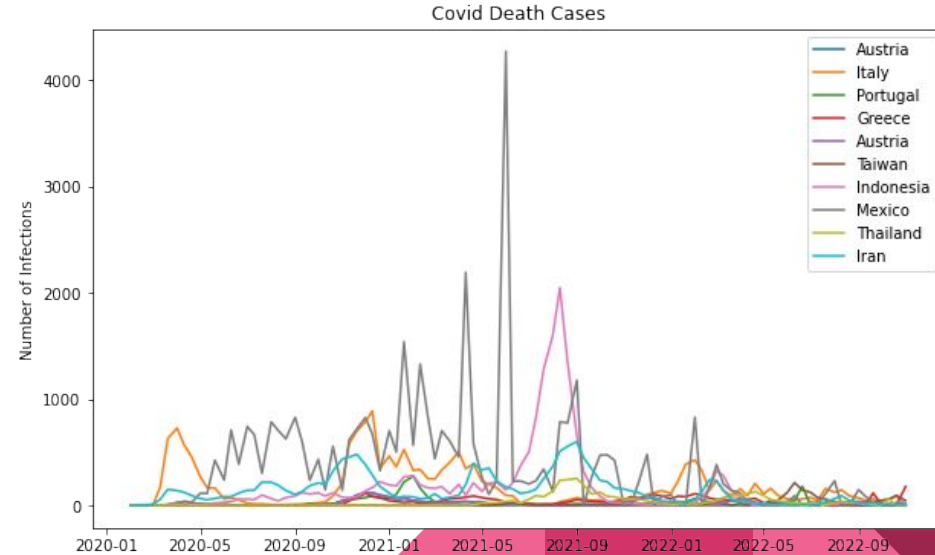
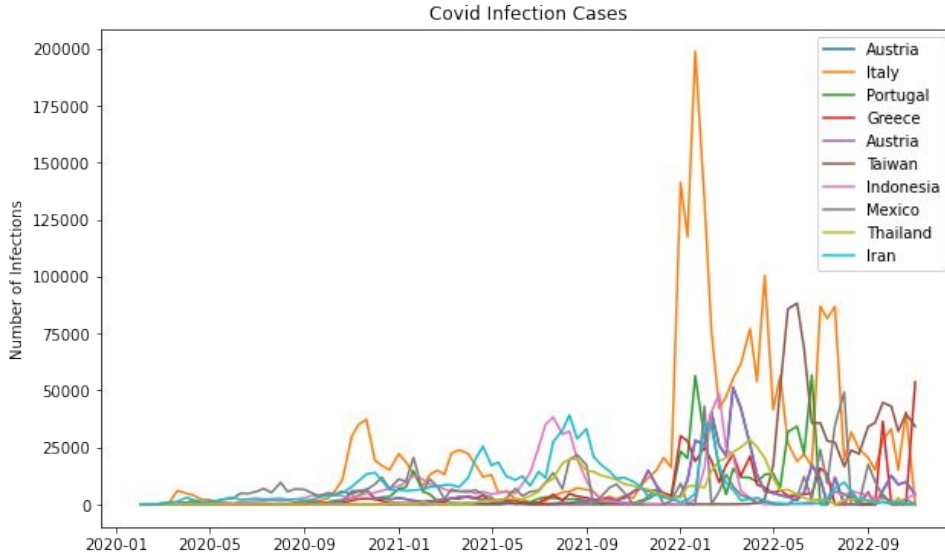
# Google Maps API: Map of Countries for COVID EDA

## Countries of Interest:

1. Austria
2. Italy
3. Portugal
4. Greece
5. Taiwan
6. Indonesia
7. India
8. Mexico
9. Thailand
10. Iran

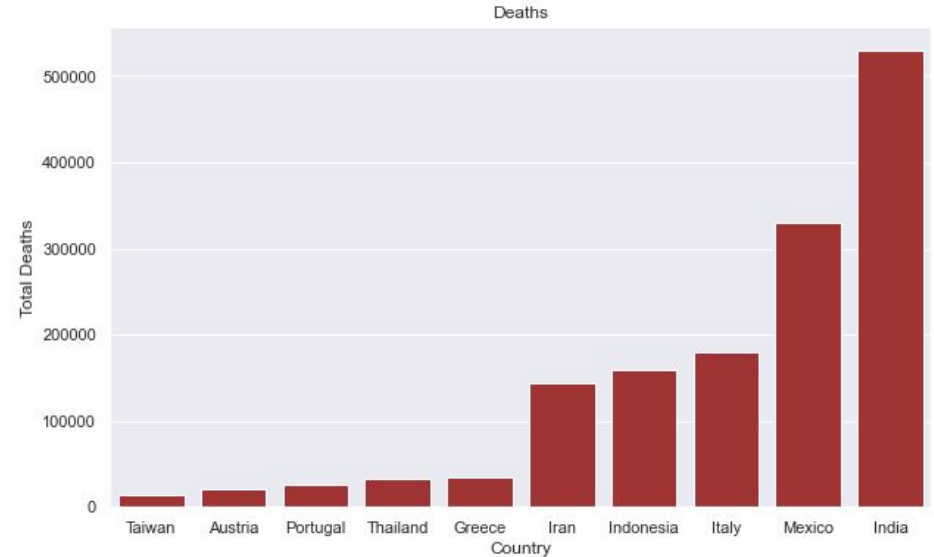
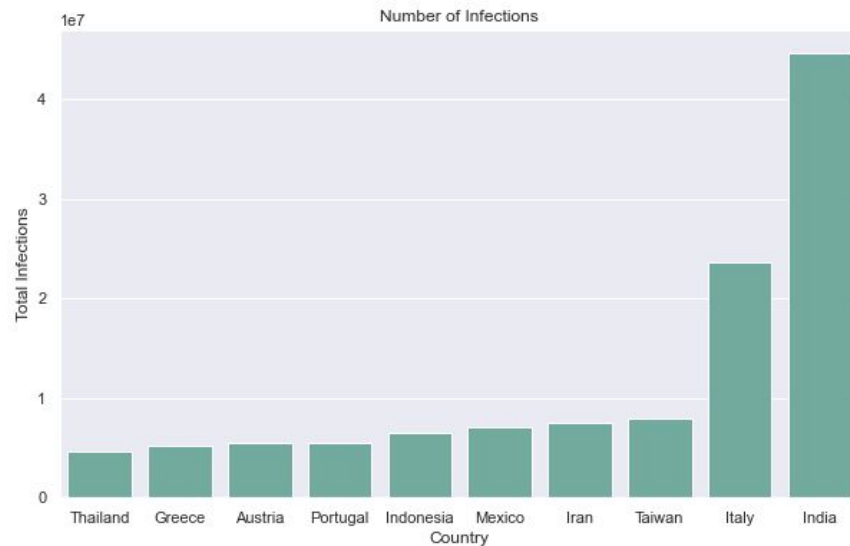


# COVID Death and Infection Line Graphs

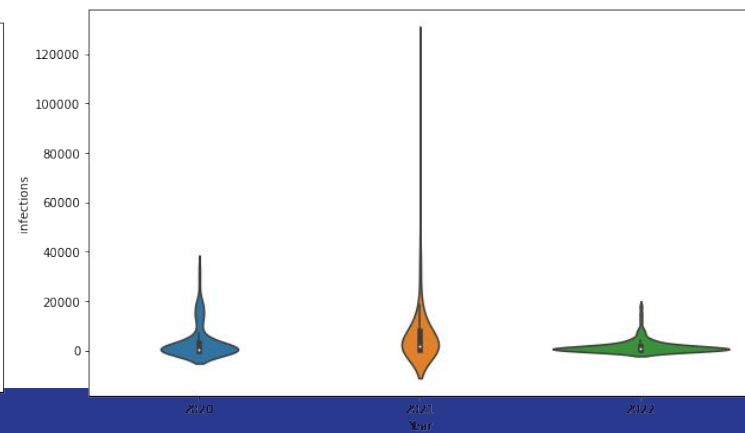
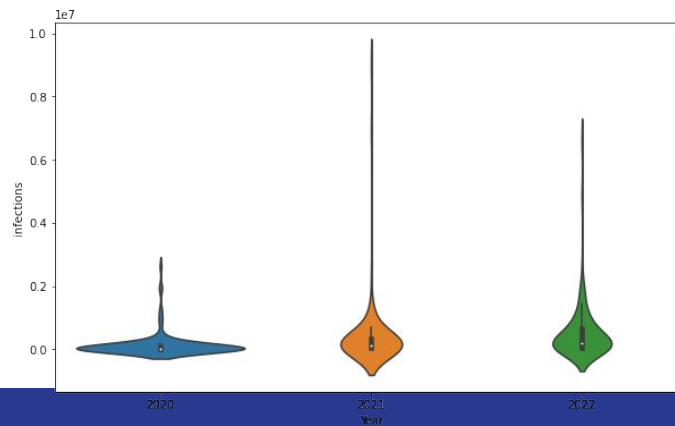
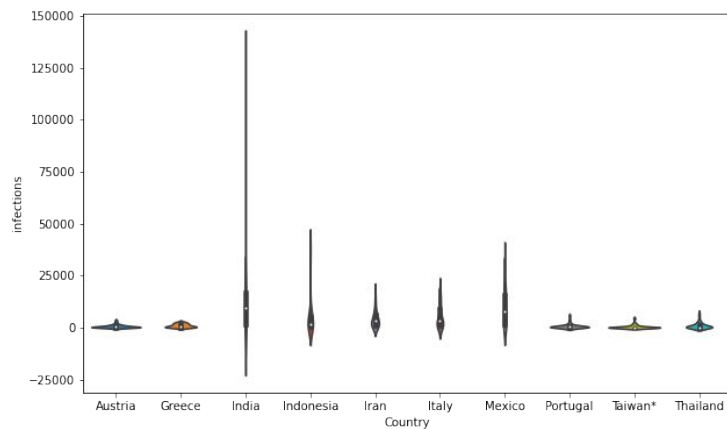
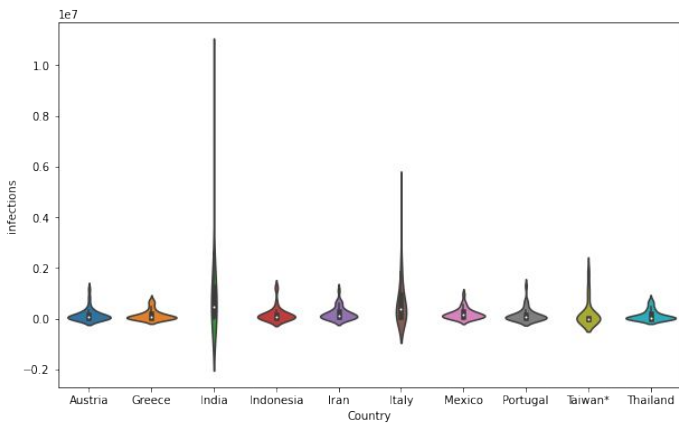


# COVID Death and Infection Totals Bar Plots

India had the most infections and deaths, with a total of 44,660,579 and 530,500 respectively. Thailand had the least infections with 4,695,203, and Taiwan had the least deaths with 13,198. (JHU COVID-19 Data Totals 23JAN2020 to 6NOV2022.)



# Violin Plots



# 1st Research Question

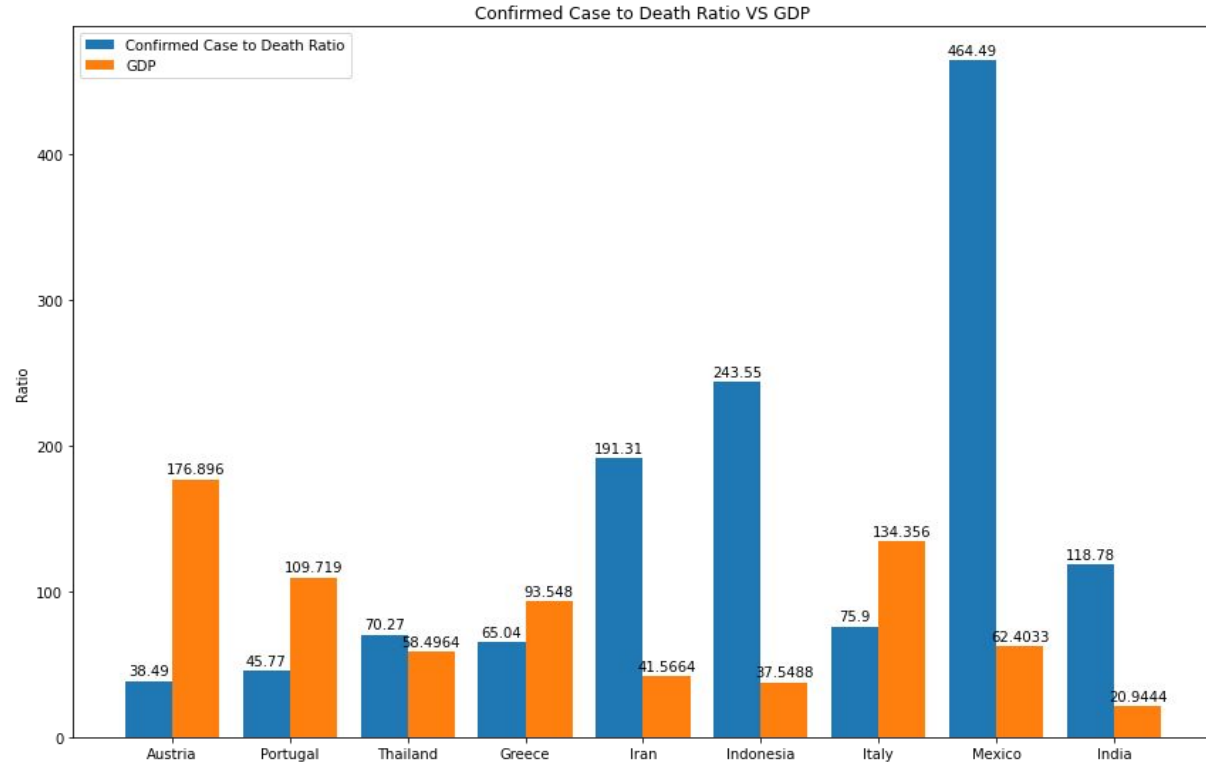
**Have COVID-19 infections and deaths been influenced by a country's GDP?**

**Is richer countries really have lower death rate?**



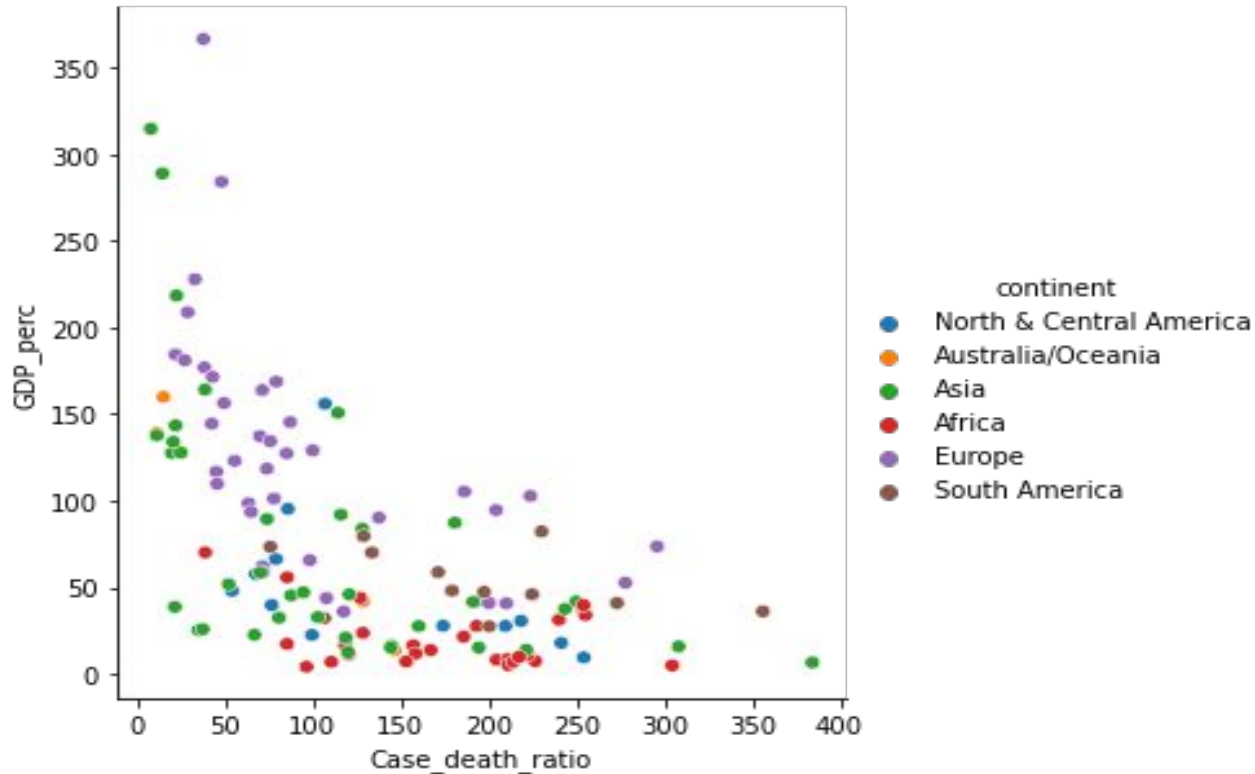
# GDP & Death Ratio

Richer countries have lower death rate VS poor countries.

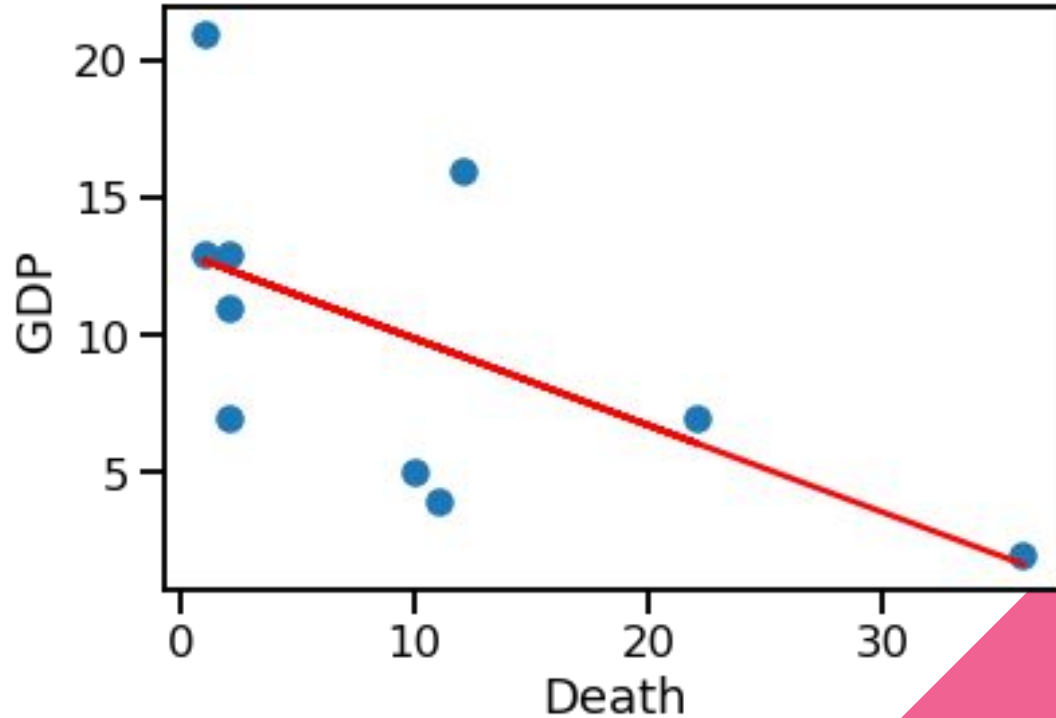




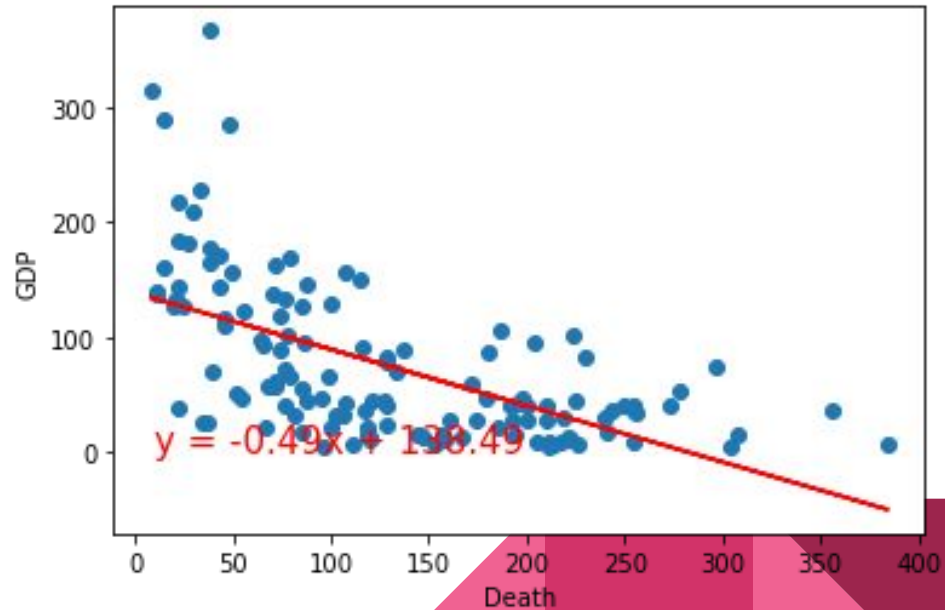
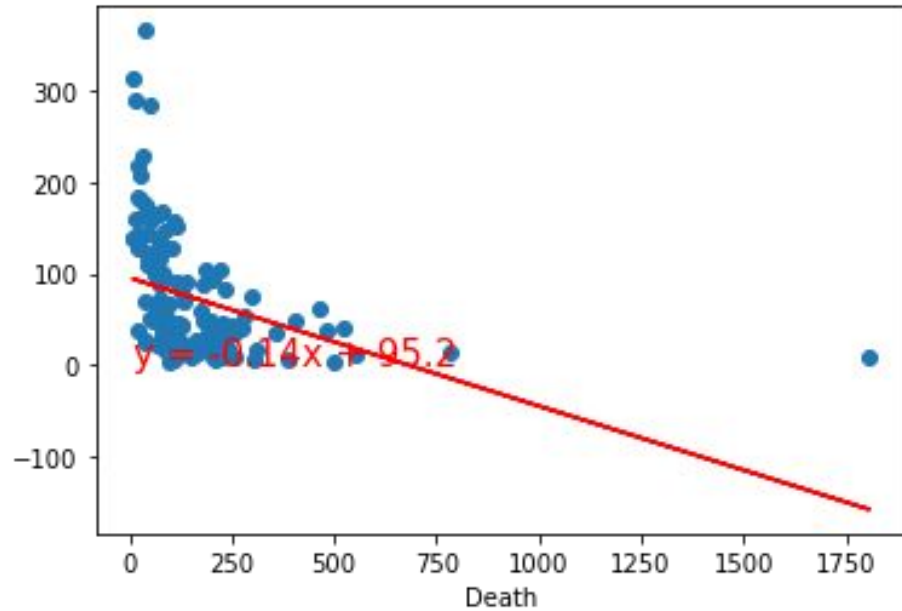
# GDP to Death Ratio Scatter Plot



# Regression Model: Sample Countries




# Regression Model: All Countries



## Second Research Question

Have COVID-19 infections and deaths been influenced by a country's population density?

Early on in the pandemic it was understood that the COVID-19 virus was easily transmissible through the air, and close proximity to 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, and it is suggested that we increase the sample size of county data in the future to better accurately determine any correlations.

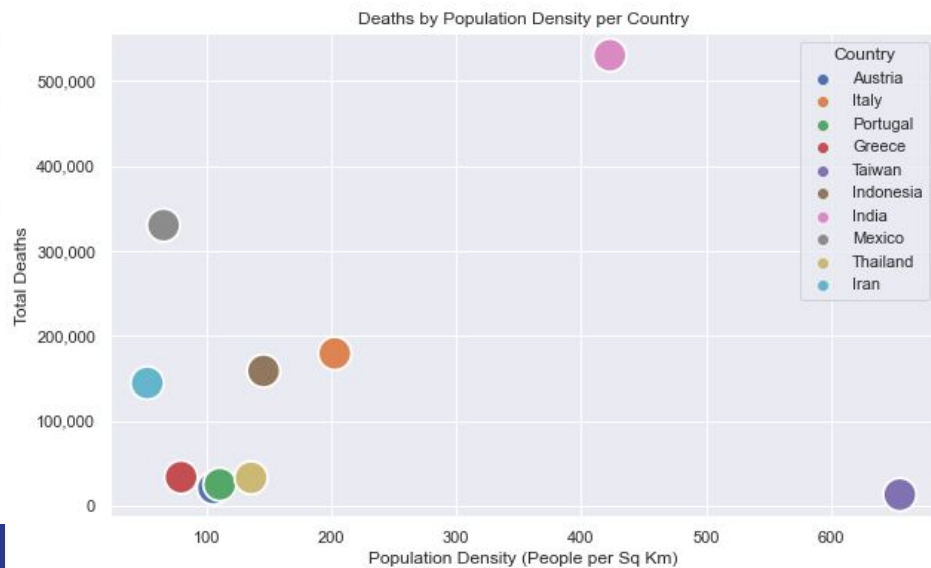
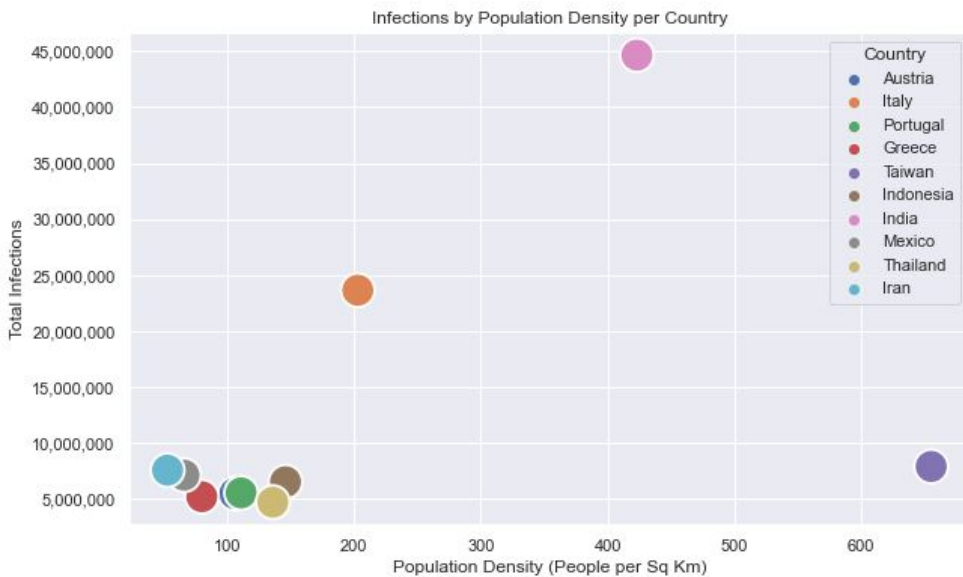


## Second Research Question

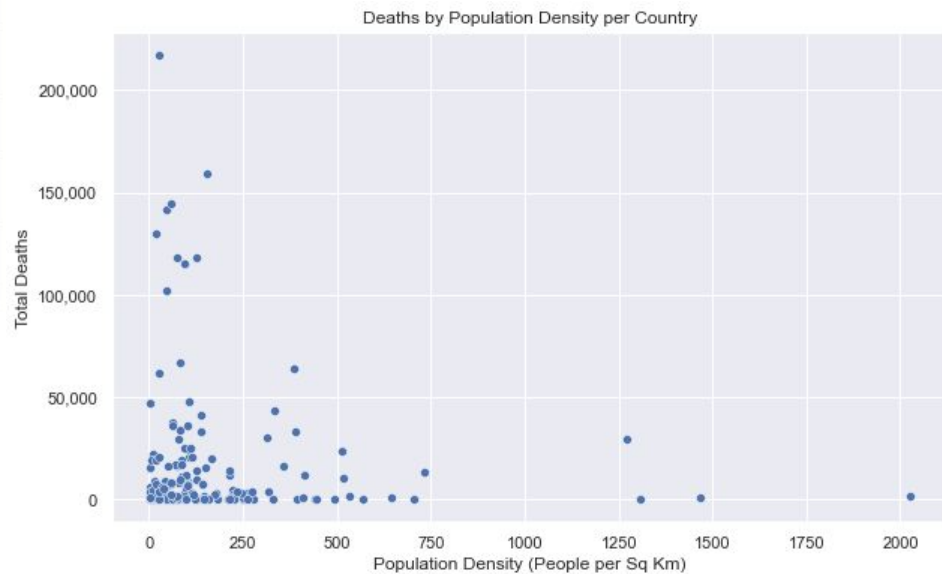
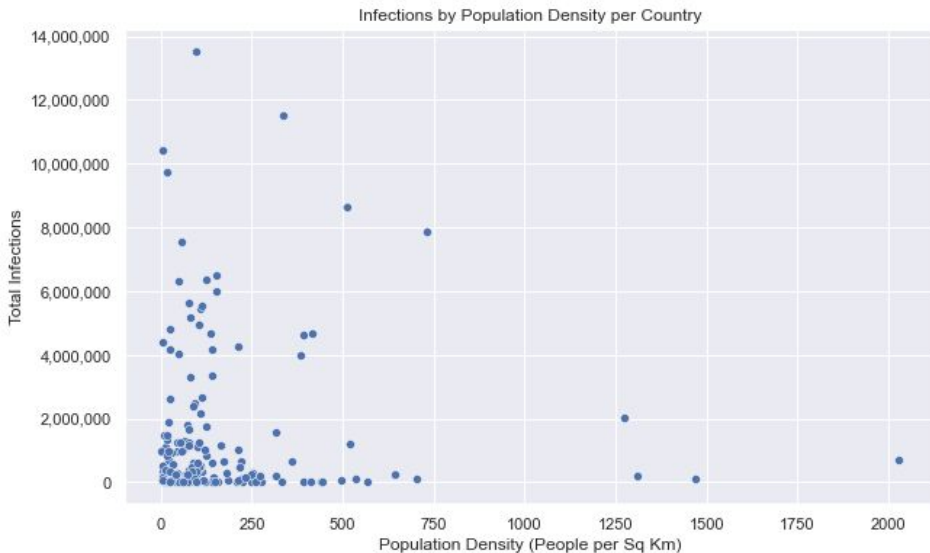
Dataframe for all countries of interest, and their corresponding data.

	country	total_deaths	total_infections	geography_area_total_sq_km	Population	pop_density
0	Austria	21,037.0	5,466,250.0	83,871.0	8,913,088.0	106.0
1	Italy	179,436.0	23,642,011.0	301,340.0	61,095,551.0	203.0
2	Portugal	25,290.0	5,525,459.0	92,090.0	10,242,081.0	111.0
3	Greece	33,750.0	5,188,890.0	131,957.0	10,533,871.0	80.0
4	Taiwan	13,198.0	7,887,537.0	35,980.0	23,580,712.0	655.0
5	Indonesia	158,829.0	6,521,292.0	1,904,569.0	277,329,163.0	146.0
6	India	530,500.0	44,660,579.0	3,287,263.0	1,389,637,446.0	423.0
7	Mexico	330,424.0	7,113,658.0	1,964,375.0	129,150,971.0	66.0
8	Thailand	32,995.0	4,695,203.0	513,120.0	69,648,117.0	136.0
9	Iran	144,596.0	7,558,142.0	1,648,195.0	86,758,304.0	53.0

# Population Density Scatter Plots, Select Countries




# Population Density Scatter Plots, All Countries



## Third Research Question

Have COVID-19 infections and deaths been influenced by a country's quality of its health care system?

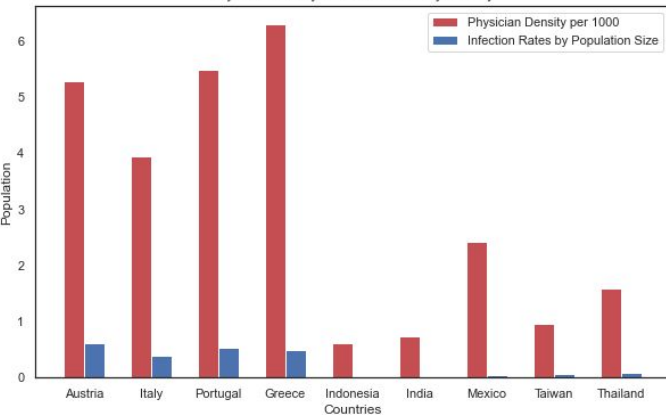
During COVID-19 a significant story in the news concerned the overflowing of hospitals and the lack of hospital beds/respiratory care. Our initial prediction was that countries that had lower spending on physicians and hospital care were going to have higher rates of infection and deaths. Our bar plots for data concerning different health care measurements did not follow our prediction entirely, while there was a correlation amongst the death rates, the infection rates did not show any significant values. Density of beds or physicians may have not been the ideal way to track the value of healthcare, however, more data points could have adjusted these correlations.



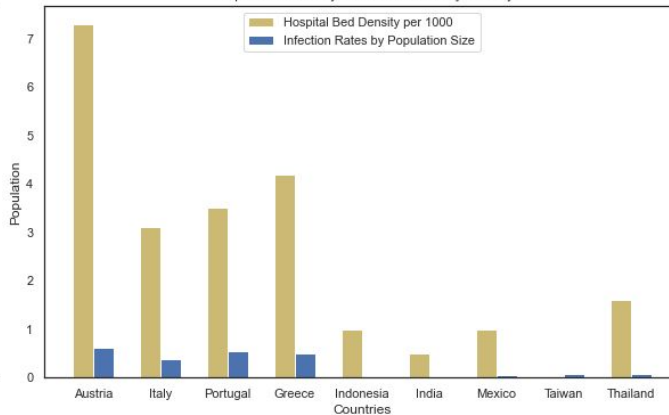


# Healthcare graphs

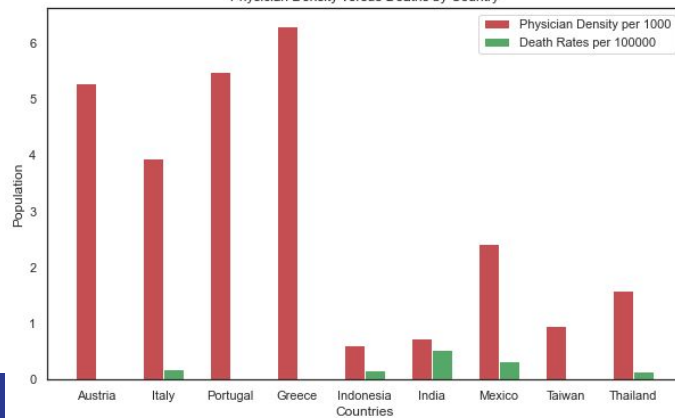
Physician Density versus Infections by Country



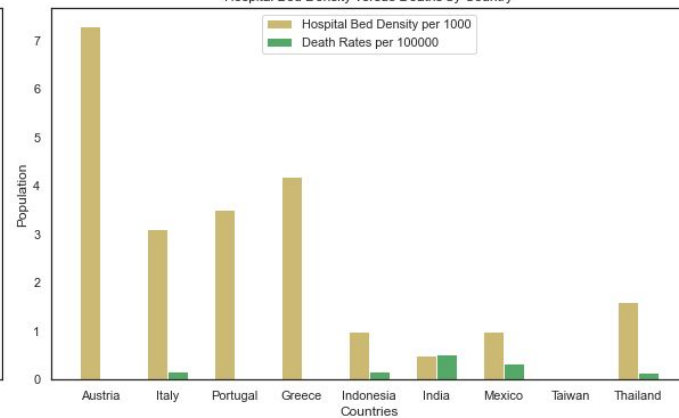
Hospital Bed Density versus Infections by Country



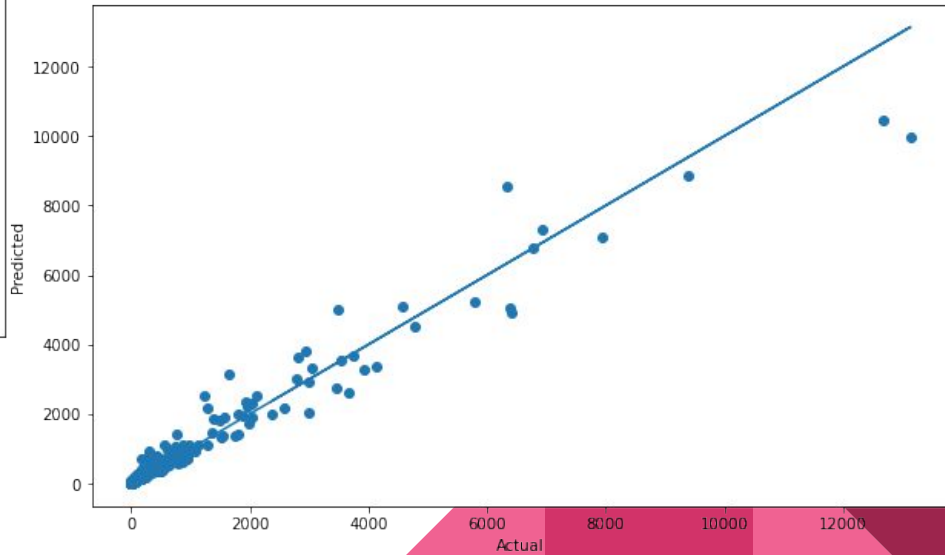
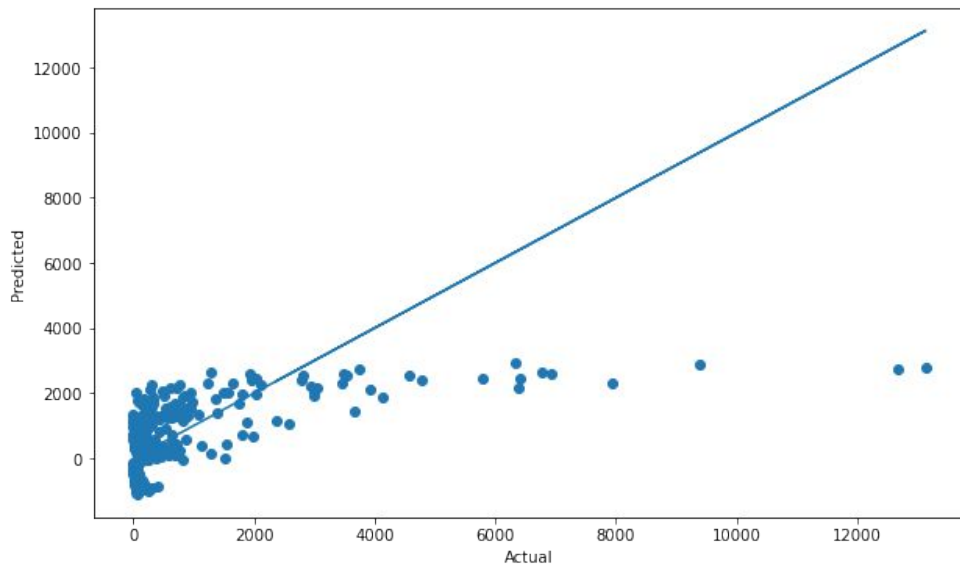
Physician Density versus Deaths by Country




Hospital Bed Density versus Deaths by Country



# Regressions



# Conclusion

- CTA - Covid Data needs to be more continually reported and collected to maintain better accuracy in the future
  - Our information can allow us to predict that internal factors such as economics and healthcare will have more effect on the rate of deaths in the country, not infections
  - GDP showed a negative relationship with death rates
  - Healthcare appeared to play a more significant role in the amount of deaths versus infection rates
  - Population Density had no apparent correlations
- 

# Limitations & Bias

- World Factbook contained data that was tabulated at different times for various countries; some data was before the start of the pandemic.
- CSV datasets required a lot of cleaning and reorganizing to become workable with the Python data analysis packages.
- Bias towards countries of higher development/larger economy being more apt to controlling the spread of infections and deaths from COVID-19.
- Sample size for study should be reconsidered, include more countries if we had more time.



# Future work

- Perform study again with larger sample size of countries.
- Break out Infections and Deaths datasets into yearly counts as opposed to the entire pandemic timeline.
- Use year-based datasets in consideration of vaccination availability timelines.
- Create a “fake country” with controlled GDP, Population Density, and Health Care System Quality to test against actual data to uncover additional insights and challenge expectations.



# 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_confirmed_cases.csv) |

[https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university?select=CONVENIENT\\_global\\_deaths.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>

