COVID 19 Data Exploration and Visualization

Dinesh Adhithya

October 11, 2021

1 Introduction

Through this work I try use COVID 19 data made available by google [1] to create visualization to understand the pandemic better. These visualizations and plots were produced using matplotlib and seaborn modules implemented in python language.The data set contains time series data pertaining to the COVID pandemic such as cases , deaths and tests information for each region in the world. It also contains data about each country's economy, demography, weather and public health.We wish to create intelligent plots and try to explain using logical and scientific arguments as the famous quote goes "correlation doesn't mean causation".

2 Data Cleaning

An important process during every machine learning project is to clean the data set and pandas module has been a great assert in regards to this.The data points pertaining to countries which had undefined values were removed and data points were grouped by country rather than its local region. Relevant features were normalized using population value of the country.

3 Data Visualization

3.1 Plots showing COVID positive cases across the globe

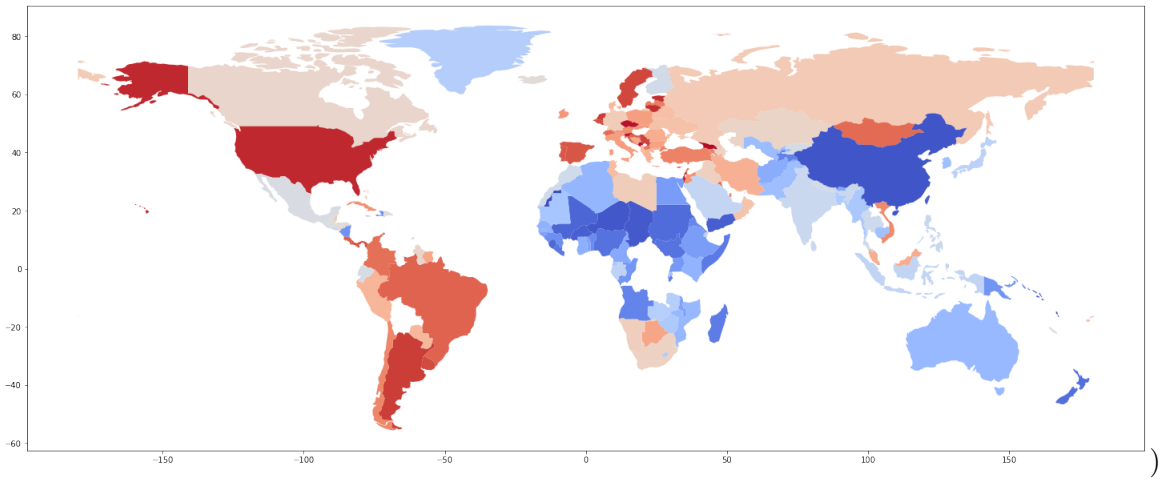


Figure 1: Figure showing cases per million people in different countries with red indicating large case load and blue the smaller case load.

The white spots refer to country whose data wasn't available in the data set.

3.2 Plots showing COVID deaths across the globe

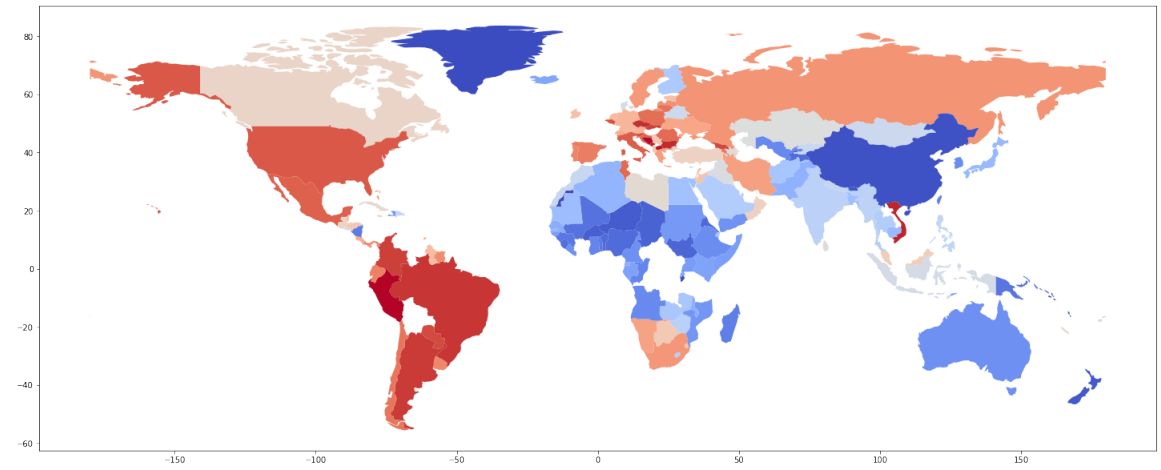


Figure 2: Figure showing deaths per million people in different countries with red indicating large case load and blue the smaller case load.

3.3 Does more testing means more cases?

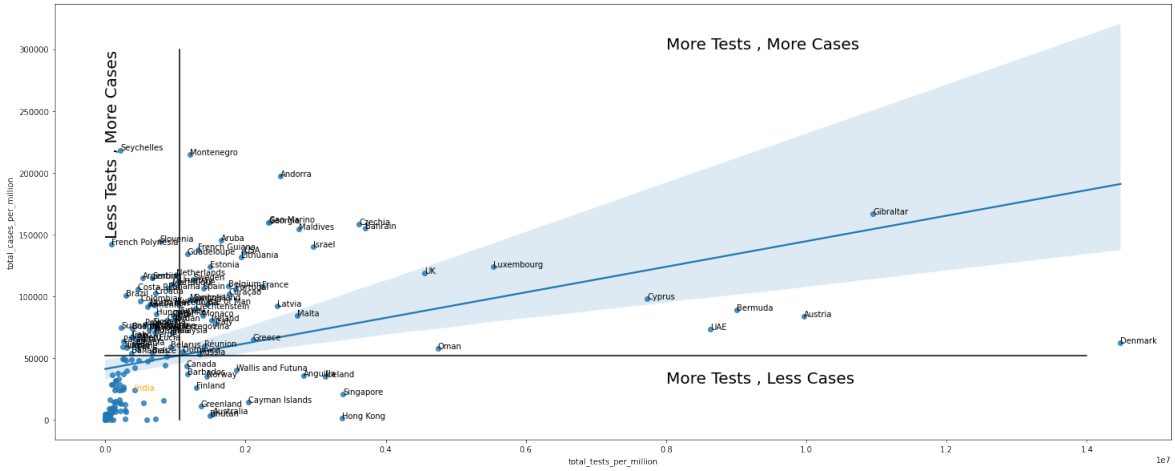


Figure 3: Scatter plot showing tests per million vs cases per million

In the above plot we fit a regression function fitting test per million to cases per million and the line of the regression plot acts as a baseline model which acts as the average no. of cases we would expect given the no. of test per million. The countries that fall in the more tests more cases region are mostly from the continent Europe , but these country's data points lie above the regression line and have had more cases than expected given their testing numbers. The countries that fall in the less tests per million and more cases per million mostly belong to the continent South America , in a sense have under tested which has led to loss of people's trust in the covid management in these countries. Then comes the region with more tests and less cases which have tested more than required. Countries that fall in this countries happen to be wealthy (high GDP per capita) islands. The countries falling in the less tests and less cases have generally tested less than global average.

India has generally done well with its COVID management except for the few gruesome months during the second wave. Her expected no. of cases per million given for the given testing numbers are large , but the actual values are much lower. Which could attribute to its large youth population , large rural population and style of living.

3.4 Does more cases mean more deaths?

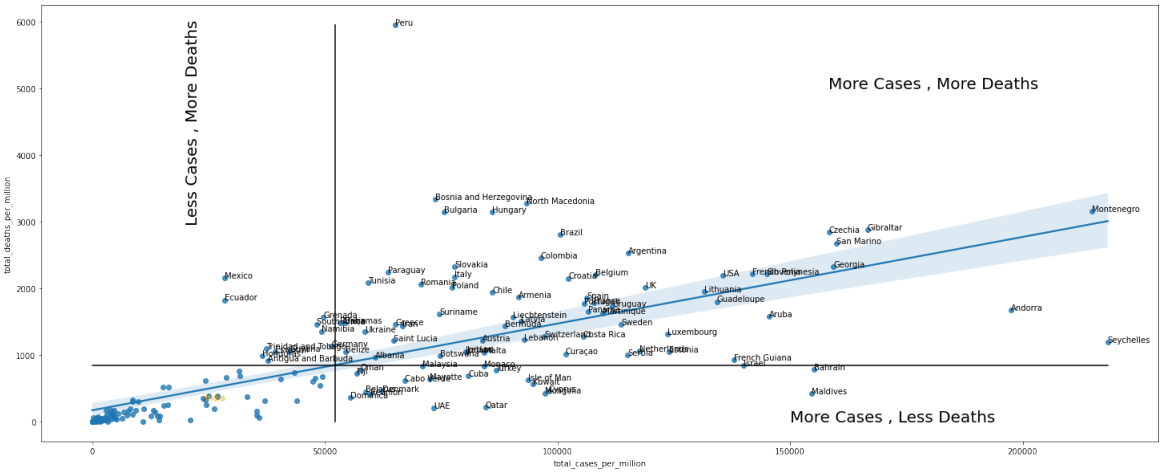


Figure 4: Scatter plot showing cases per million vs deaths per million

We try to a perform similar exercise to the one done before except using the death per million to cases per million . We observe that most Europe have had more deaths per million than the expected value. There are also countries which have had lesser deaths than expected such as India.

We see each country perform in varied ways when it comes to deaths per million and cases per million. We wish to try and understand this using various economic , health , demographic and climate related features describing a particular country.

3.5 How does population demographics affect cases per million

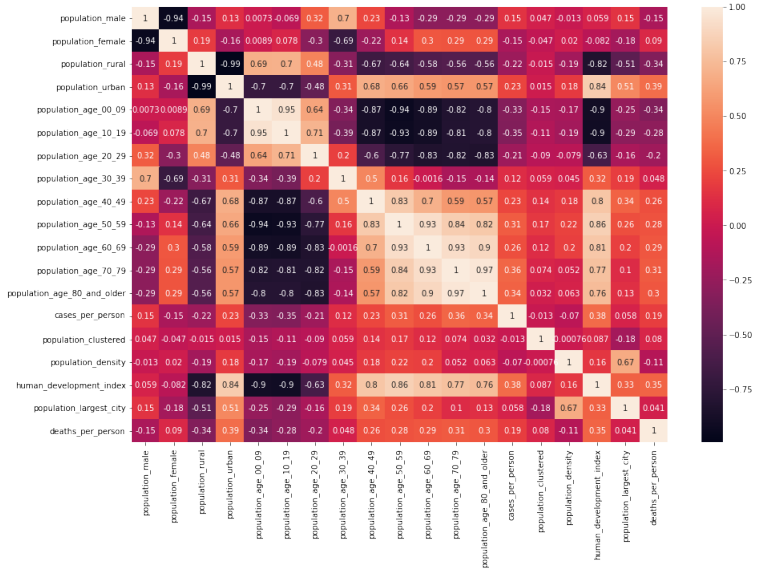


Figure 5: Correlation matrix showing cases per million and various population demographics features.

3.5.1 Deaths per million

shows negative correlation with population in range 0-29 and almost zero correlation in 30-39 age group. shows positive correlation with population in age groups above 40 years.Deaths per million shows large negative correlation with rural population and positive correlation with urban population. The interesting fact that this plot shows is that deaths per million shows positive correlation with Human development index.

3.5.2 Cases per million

shows negative correlation with population in range 0-29 and shows positive correlation with population in age groups above 30 years.Deaths per million shows large negative correlation with rural population and positive correlation with urban population. The interesting fact that this plot shows is that cases per million shows positive correlation with Human development index.Given the assumption that the case data is true , then the correlation with HDI seems counter intuitive , but countries with large HDI happen to have large proportion of old people (above 50 years).Which could be a possible explanation for this behaviour.

3.6 How does climatic conditions affect cases per million in a country

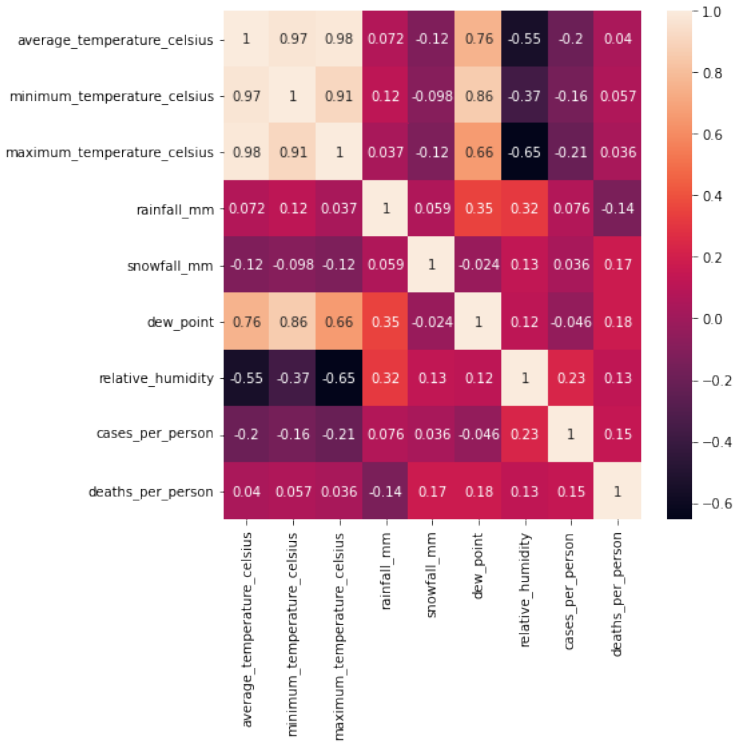


Figure 6: Correlation matrix showing cases per million and various climatic conditions features.

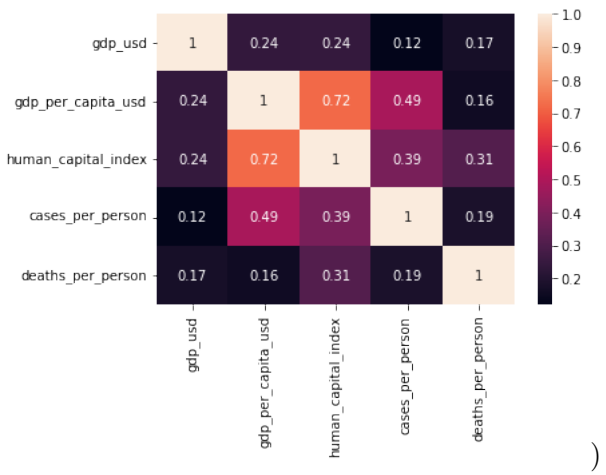
3.6.1 Deaths per million

Deaths per million shows zero correlation with a region's temperature. Whereas shows positive correlation with humidity and snowfall although it is quite low. Deaths per million shows negative correlation with average rainfall received which is quite counter intuitive but most high rainfall receiving countries lie in south east Asia which has relatively done well with its COVID case load.

3.6.2 Cases per million

Shows negative correlation with average temperature which makes sense and positive correlation with relative humidity and isn't affected much by other parameters.

3.7 How does economy affect cases per million in a country



)

Figure 7: Correlation matrix showing cases per million and economy describing features.

3.7.1 Deaths per million

Deaths per million shows a large correlation with human capital index and GDP per capita , which is mostly due to wealthy countries having small proportion of people in 0-30 age group.

3.7.2 Cases per million

Cases per million also shows a similar behaviour , given the fact that rich countries can afford lock downs should have a negative correlation with GDP per capita.But we see a positive correlation which is quite counter intuitive.Which shows how their economies are still dependent on travel and vulnerable to a pandemic.

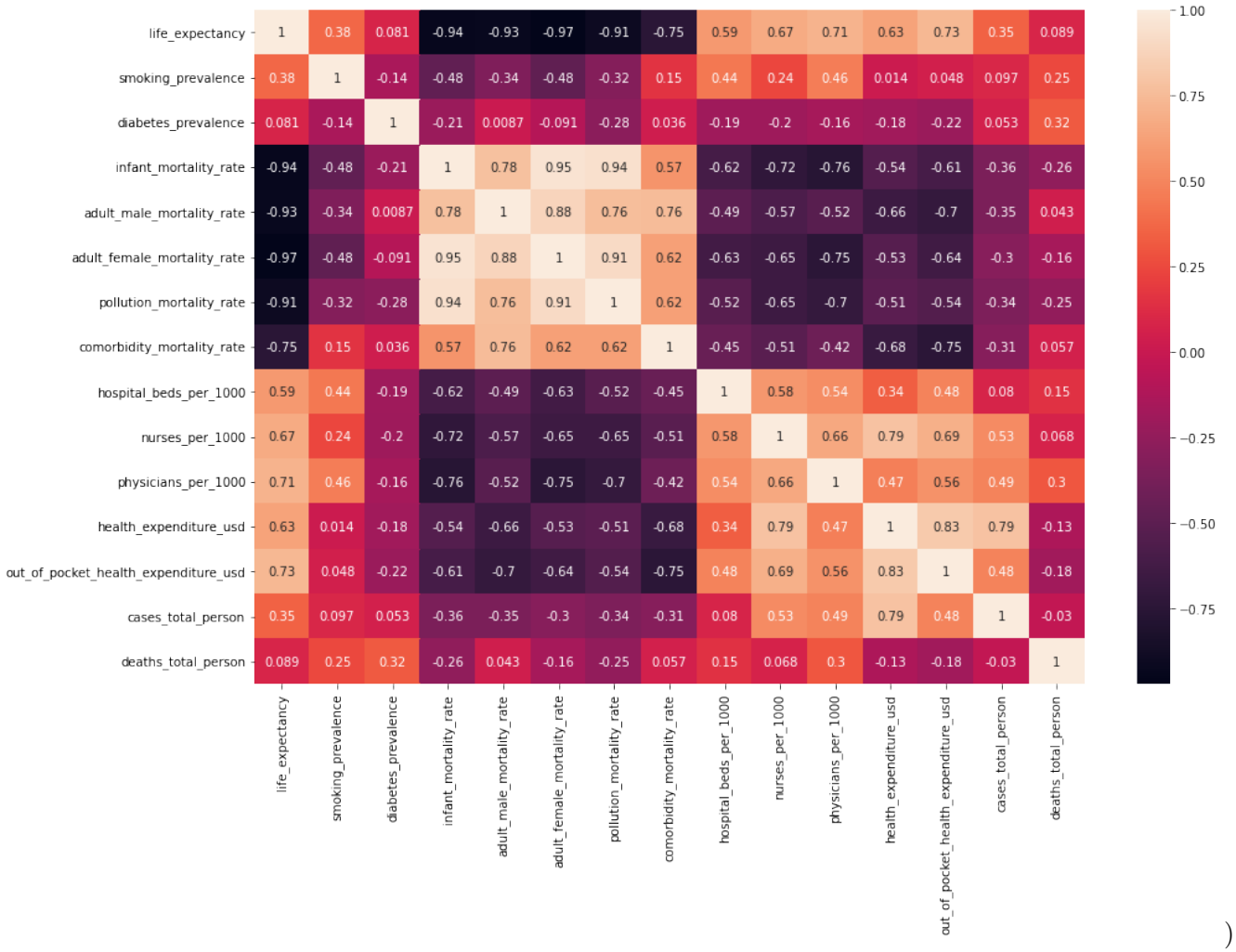
3.8 How does public health affect cases per million in a country

3.8.1 Deaths per million

Shows significant correlation with smoking prevalence and diabetes prevalence. Shows negative correlation with mortality rate , where most countries with high mortality rate have low HDI and large youth population.Deaths per million shows negative correlation with health expenditure which makes sense , otherwise Deaths per million isn't affected by any other features describing health.

3.8.2 Cases per million

Shows positive correlation with life expectancy as large proportion of old population means more cases. Which in turn shows positive correlation with larger health expenditure capacity as rich countries have large paying capacity but large proportion of old people.



)

Figure 8: Correlation matrix showing cases per million and public health describing features.

All codes were written in python and the codes can be found at this GitHub repository. Any changes to the codes can be suggested in GitHub itself.

References

[1] O. Wahltinez et al. Covid-19 open-data: curating a fine-grained, global-scale data repository for sars-cov-2. 2020. Work in progress.