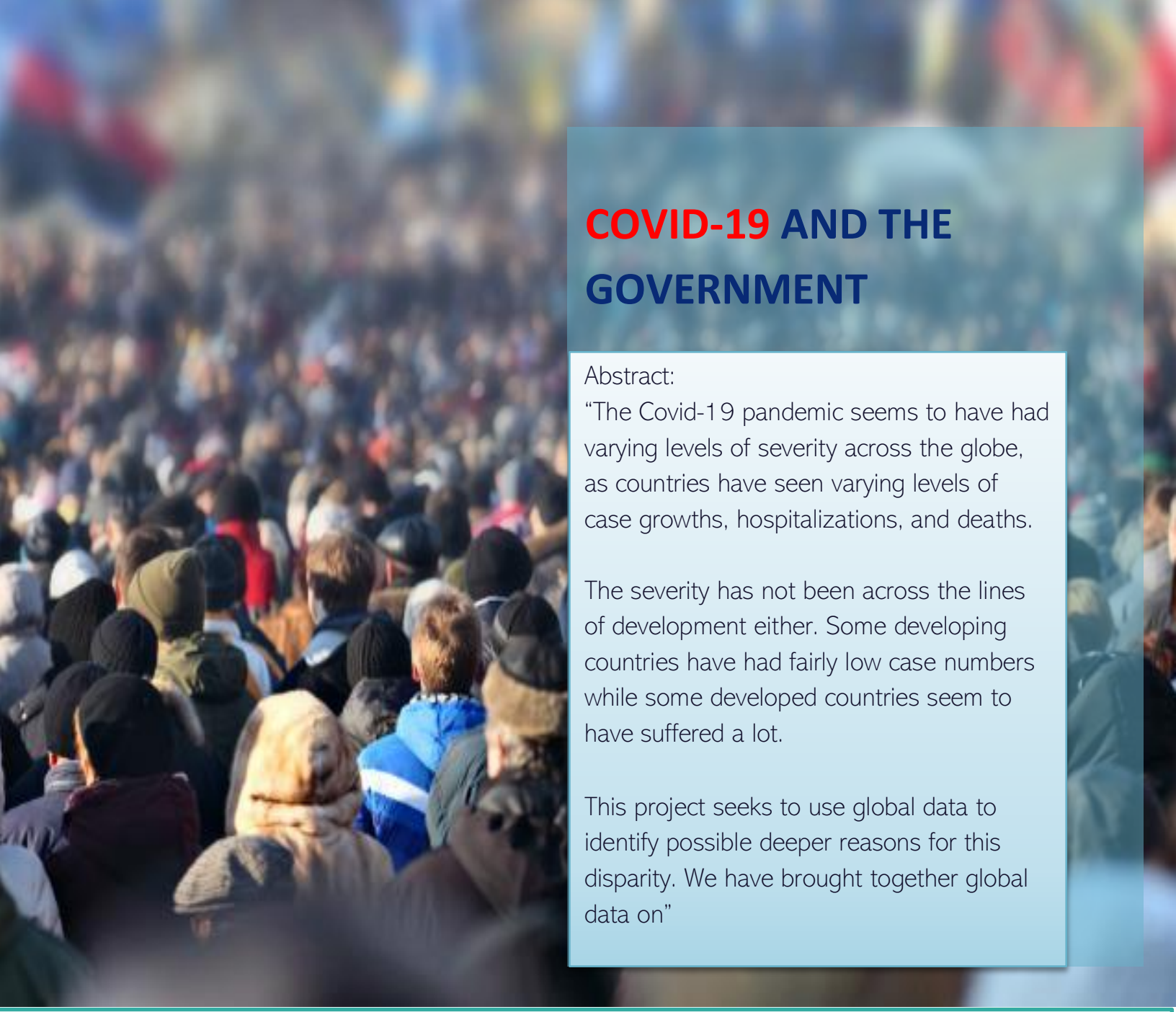


COVID-19 AND THE GOVERNMENT

Abstract:
"The Covid-19 pandemic seems to have had varying levels of severity across the globe, as countries have seen varying levels of case growths, hospitalizations, and deaths.

The severity has not been across the lines of development either. Some developing countries have had fairly low case numbers while some developed countries seem to have suffered a lot.

This project seeks to use global data to identify possible deeper reasons for this disparity. We have brought together global data on"

October 9, 2020

ISQS 6339 Business Intelligence
Angela Adjei-Mosi
Mohammad Ansari
Abdulmuizz Yusuf

Contents

Analysis of Data

The purpose of this project is to analyze and bring together global government efficiency data and medical infrastructure data and to see how that has impacted Covid-19 response in respective countries; with different shared datasets from public databases. We used three data sets, one being the master data set and the second and the third datasets were merged because they were potentially better off used when joined together than separately.

There are many questions that our datasets can help potentially answer. Our main project question is to understand the effect of government and infrastructure on Covid-19 response and local mortality rate. This project seeks to use global data to identify possible deeper reasons for this disparity. We can answer many questions like: how many cases and deaths have occurred because of COVID, different markers of each country's health infrastructure and workers, and markers like GDP, government effectiveness etc. There is a vast amount of information existing in the data sets. We determined which part of the data was deemed valuable towards our research to answer the questions at hand. The valuable data items consisted of test, cases, government efficiency index, medical markers, and GDP.

These are very trying times during the global crisis. The pandemic is affecting the economy thus affecting business. However with this data it could help to analyze forthcoming trends in the months and years to come. Applying the analysis of this data in business could help the economy and the government regroup and pivot in another direction with laws, government spending etc and globally help countries work together cohesively and efficiently.

The tools we used to analyze our data were Python and Tableau. We compiled the data input into python; clean, merged and used the merged data to answer the following questions in the rest of the report.

Data Cleaning

Because there were so many measures in the data set, the data set was not complete due to null values. Data from some countries were missing such as United States due to inaccessibility. The overall quality of the data was good because strong correlations were found in the dataset. To clean the merged data, simple drop duplicate and operations to drop null values were performed. Calculated columns were made in Tableau for data analysis. Max and average values over countries were calculated in Tableau to keep raw data file clean and reduce the time required to clean data.

A correlation matrix was made between independent variables. The dataset had values for New COVID cases as well as total number of cases for each country over time. Because the objective was to find correlations between locations as opposed to spread over time, the maximum value was taken from New Cases, Total Deaths, and Total Tests to represent correlations during the peak of Coronavirus in each country.

Data Merging

Following the cleaning of the two data sets: Data With COVID-19 stats broken down by date and country and Data with country's efficiency and infrastructural statistics, we were left with an idea of correlating the two data sets and thus creating a merged file. We decided to merge the datasets by comparing Countries within both data sets given that they were the common element amongst both data sets. Both sets had country as the primary key. To accomplish this task, we had to find websites and data sets that had date, country, country's efficiency and infrastructural statistics.

The variables were more valuable merged than separate because the variables that examine medical markers(infrastructure and personnel) in a country by themselves are interesting to look at, however when merged with COVID-19 datasets, we start to see a picture of how these medical markers might have an effect on each country's "performance" during the pandemic. Also, it shows in this way that merging these data sets would allow us to see how the markers would interact and generate new business insights. An example of a question that can be answered is if countries with higher numbers of physicians per 1000 people would have a lower COVID fatality rate. Answering these kinds of questions would be impossible without the datasets being merged.

Luckily, there were no issues with multi-level measurement. One thing we had to watch out for was that while some values changed daily(number of cases for example), others were static values(GDP per capita for example), and this meant we had to aggregate them different before visualizing.

Analysis of Visualizations

The COVID crisis has greatly had an effect on different aspects of the government and the world. Our team decided to look deeper into different entities affected by COVID. We chose to use different types of visualizations and graphs to depict the different variables and how they interact with each other. For each visualizations we will describe each of the variables and how they are intertwined and connected using Tableau and Python as seen below. We will analyze the trends, patterns, and correlations and also predictions based on our data. The data is a very good representation of natural processing and it follows the characteristics and principles of effective data visualization.

Graphs

We begin with discussing the first graph. This graph is looking at the relationship between stringency(quarantine) and the number of new cases. For this visualization we decided to focus on the US. We can see that there was a spike in cases(April 26, 2020), even though the stringency was high. However when the stringency slightly dipped(June 16, 2020), you can see the avg new cases slightly start to rise. We can conclude from the graph that initially when the stringency was rising, the cases were increasing but slightly dipped. But any decrease in stringency tends to have a greater effect on the total number of cases.

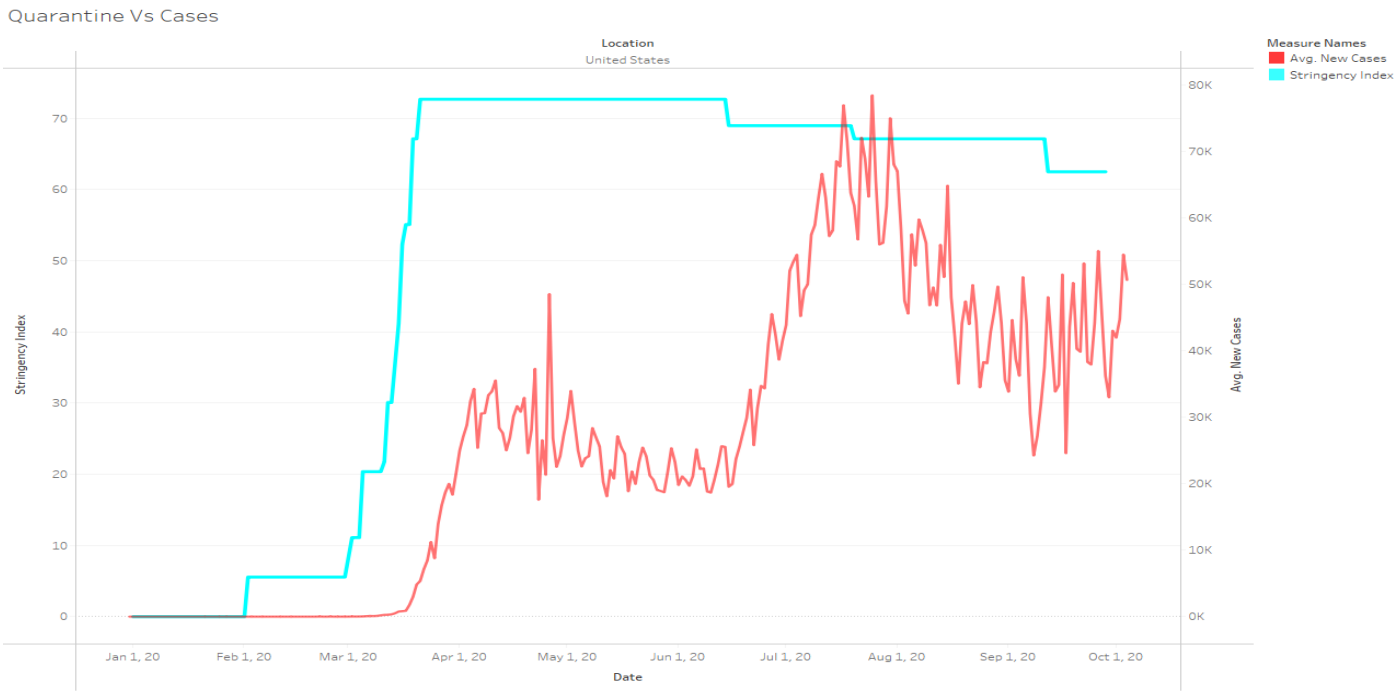


Figure 1

The next comparison we made was to try to determine if better governed countries managed to mitigate the spread of the virus better. We compared the 25 countries with the highest number of cases per million people, with the Governance index, a measure calculated by the World Bank that measures how well a government is perceived to be ran using a variety of variables.
What we saw was the correlation in this case was not the strongest. In the top 5 countries in cases per million we have countries like Israel, Andorra, and Aruba, countries which according to the measure, are very well run. We also have countries like Peru, Brazil, and the Maldives, which are not so well run according to the Governance Index.
While the correlation is not strong, we do see that most of the countries in the top 25 for cases per million are generally well run. A theory here is that well run countries would be doing more Covid testing, as well as being more transparent with case reporting, and this could lead their numbers to be higher than countries that are not as well run.

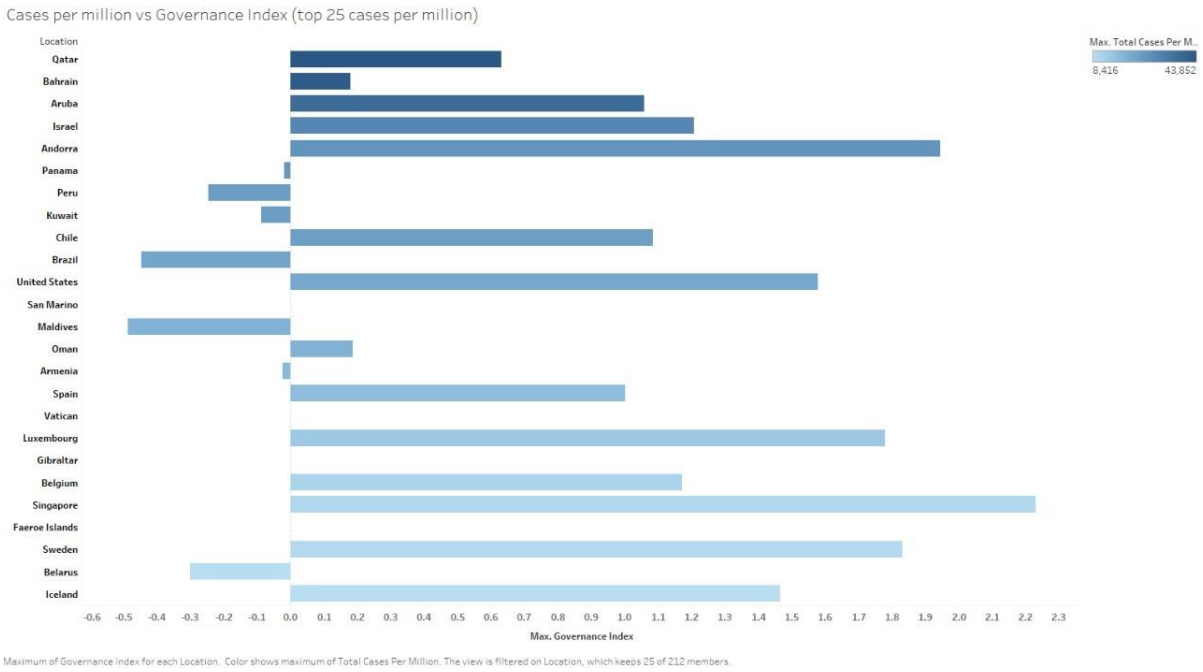


Figure 2

One area where we say quite a strong correlation was in the GDP per Capita and the Total Cases per Million (r=0.46). This suggests that more developed countries have seen more infections than less developed countries, which is unexpected. It is also in keeping with what we saw when we compared the government effectiveness and cases per million.
The most likely explanation for this is that countries with a higher GDP would be able to perform more COVID tests than less developed nations. Consequently, they would also report more cases than less developed countries.

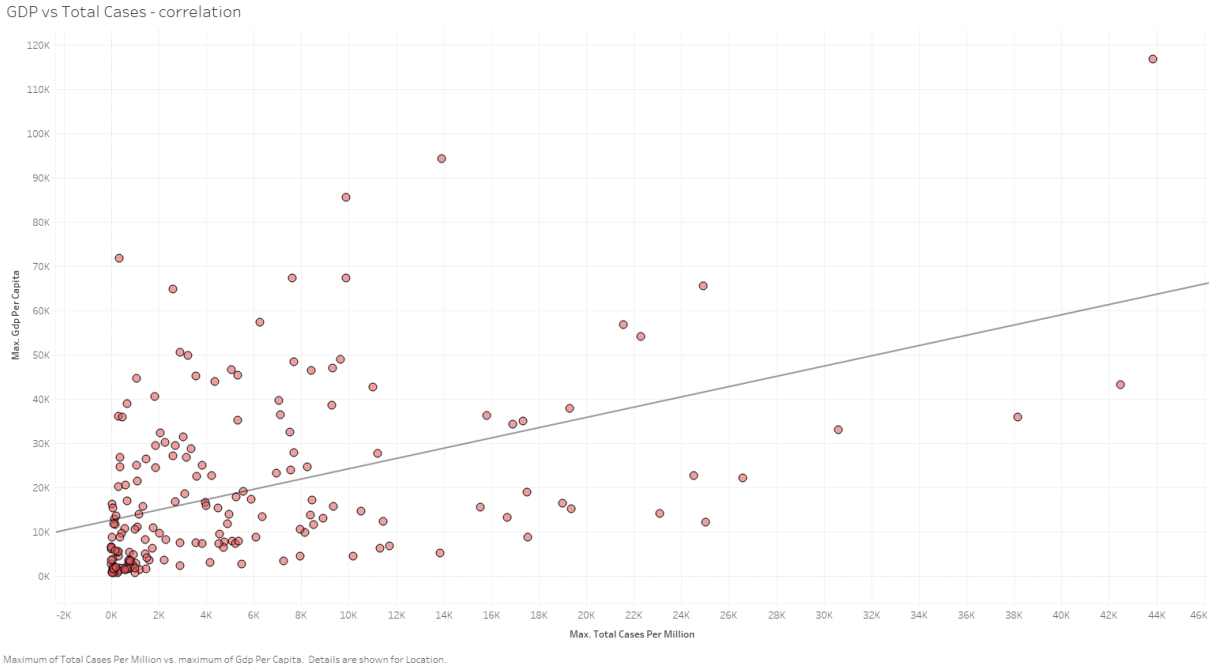


Figure 3

Lastly, a geovizualization map was created with Jupyter Notebooks and the python library Folium. A base map was created, then enhanced as a heat map. Circular markers were added to simply just show the total confirmed cases.

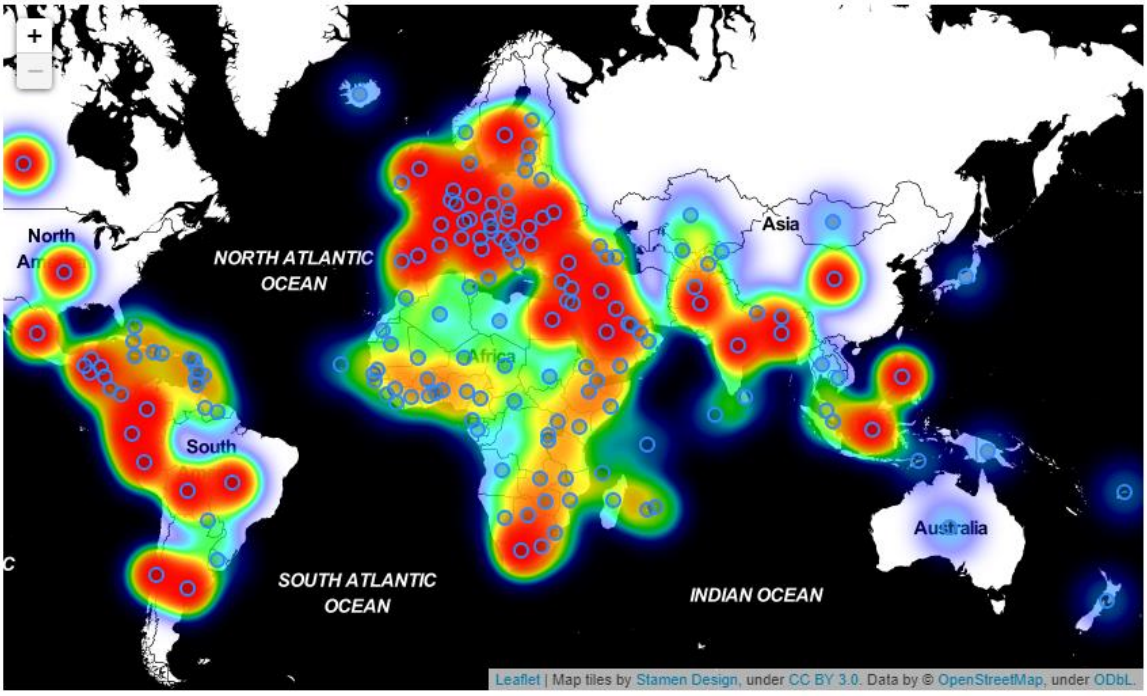


Figure 4

Correlation Matrix

A correlation matrix was made between independent variables. The dataset had values for New COVID cases as well as total number of cases for each country over time. Because the objective was to find correlations between locations as opposed to spread over time, the maximum value was taken from New Cases, Total Deaths, and Total Tests to represent correlations during the peak of Coronavirus in each country.

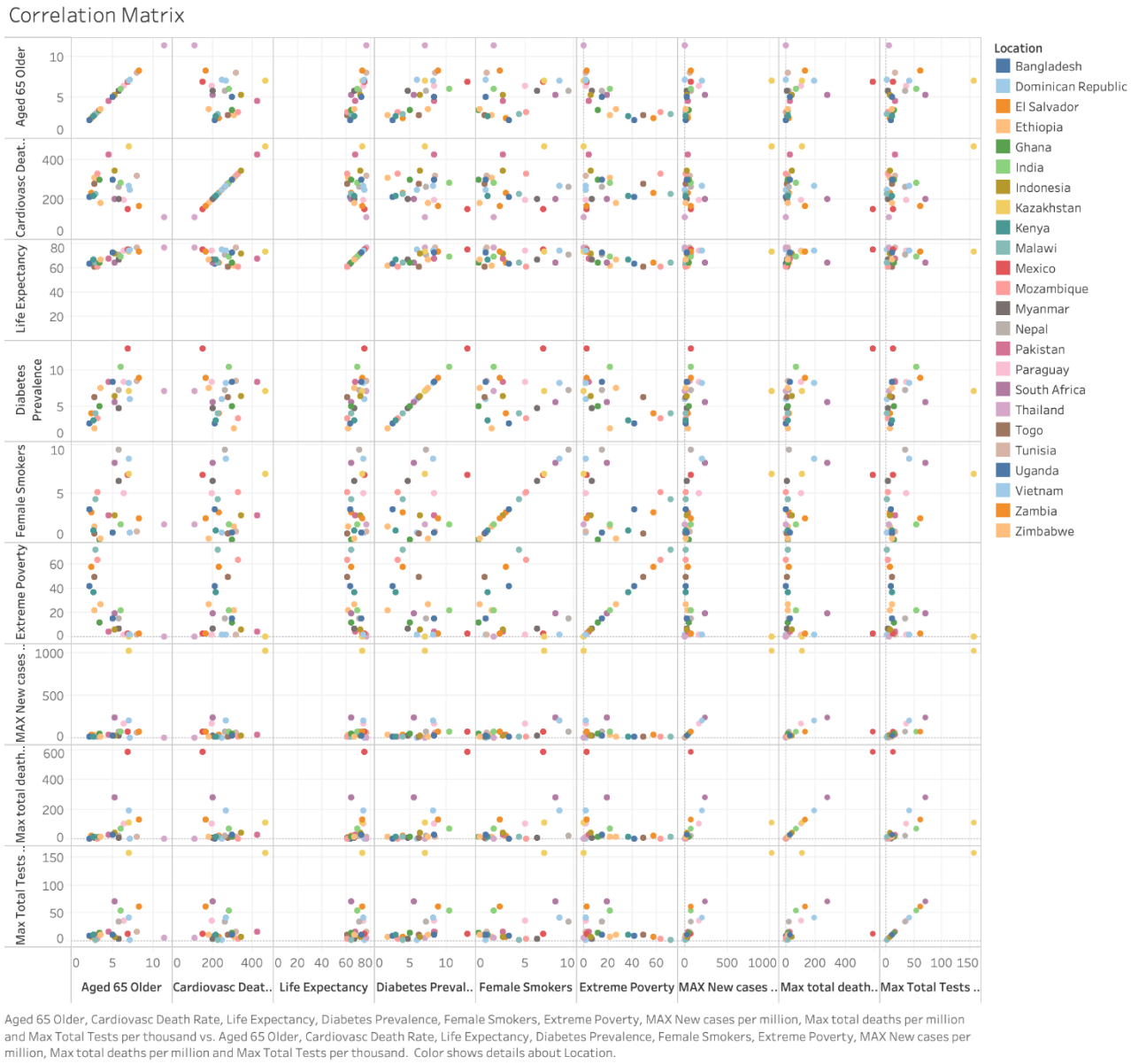
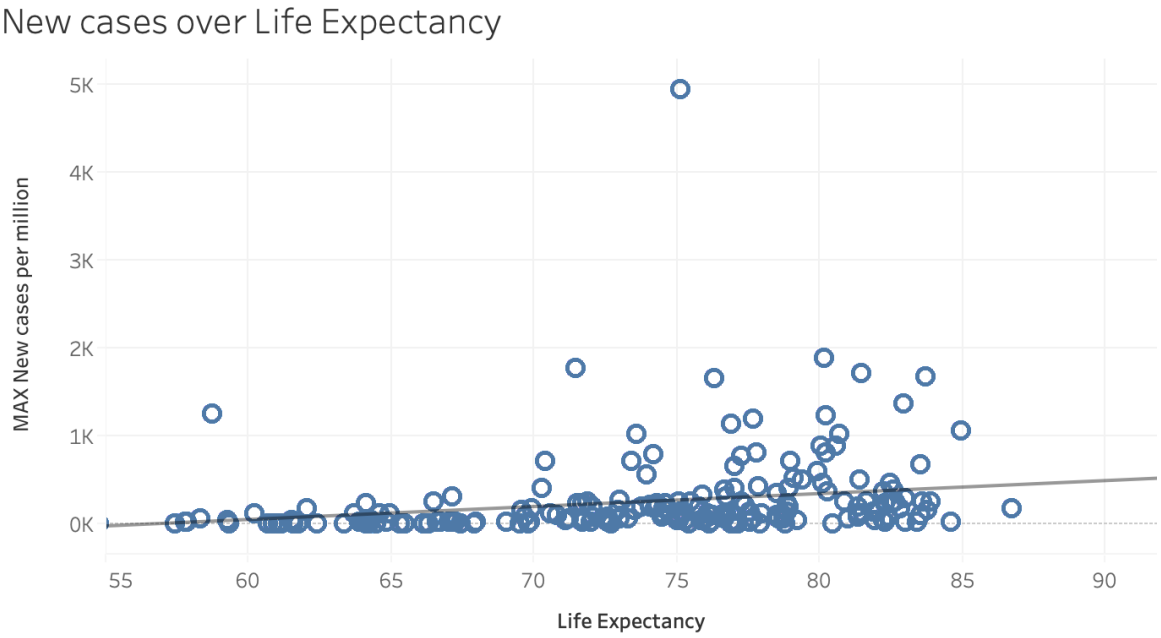


Figure 5

The correlation matrix allows for a quick visual of how different measures are related to one another. Once a correlation was identified, it was plotted to calculate P values for trend. The correlation between life expectancy and New cases was plotted. This shows that for countries with higher life expectancies had a higher peak in new cases.



Life Expectancy vs. MAX New cases per million.

Figure 6

P-value:	< 0.0001
Equation:	MAX New cases per million = 14.7371*Life Expectancy + -836.64

Coefficients				
Term	**Value**	**StdErr**	**t-value**	**p-value**
Life Expectancy	14.7371	0.285068	51.6969	< 0.0001
intercept	-836.64	21.2022	-39.4601	< 0.0001

Because each Country tests at different rates, it is difficult to gauge response to the pandemic based off total cases alone. Below is a chart showing the maximum value for total cases per million and total tests per thousands. Based off the visualization below, places which tested the largest percentage of their population at the time when COVID related deaths were highest with respect to population in that area. This shows that a higher rate of cases due to COVID is measured in countries where more people are tested.

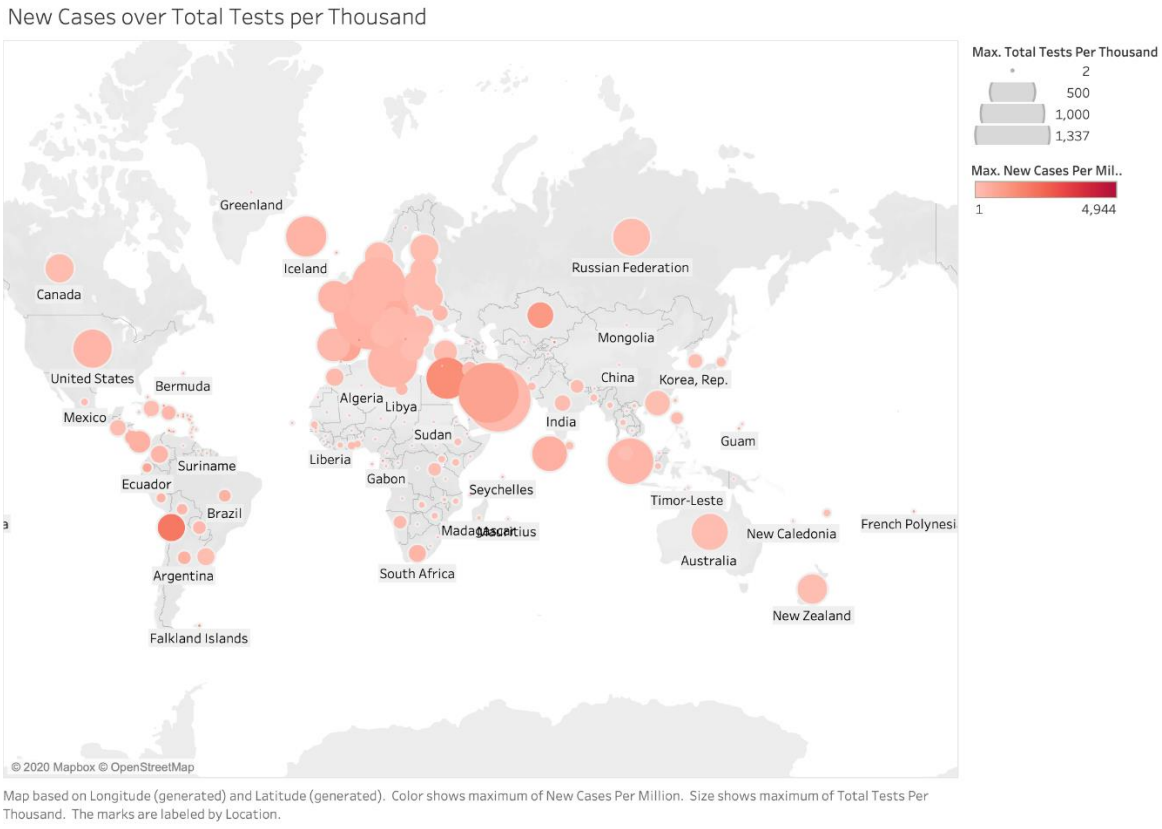


Figure 7

Lastly, total cases over time.

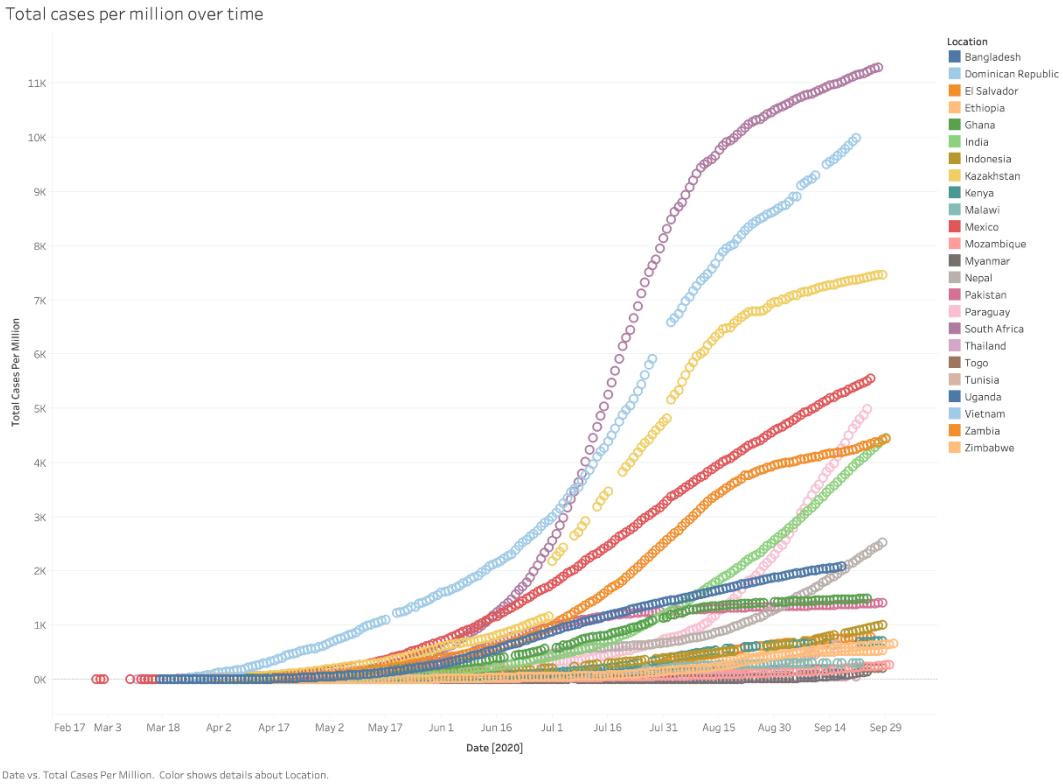


Figure 8

Flow Diagram

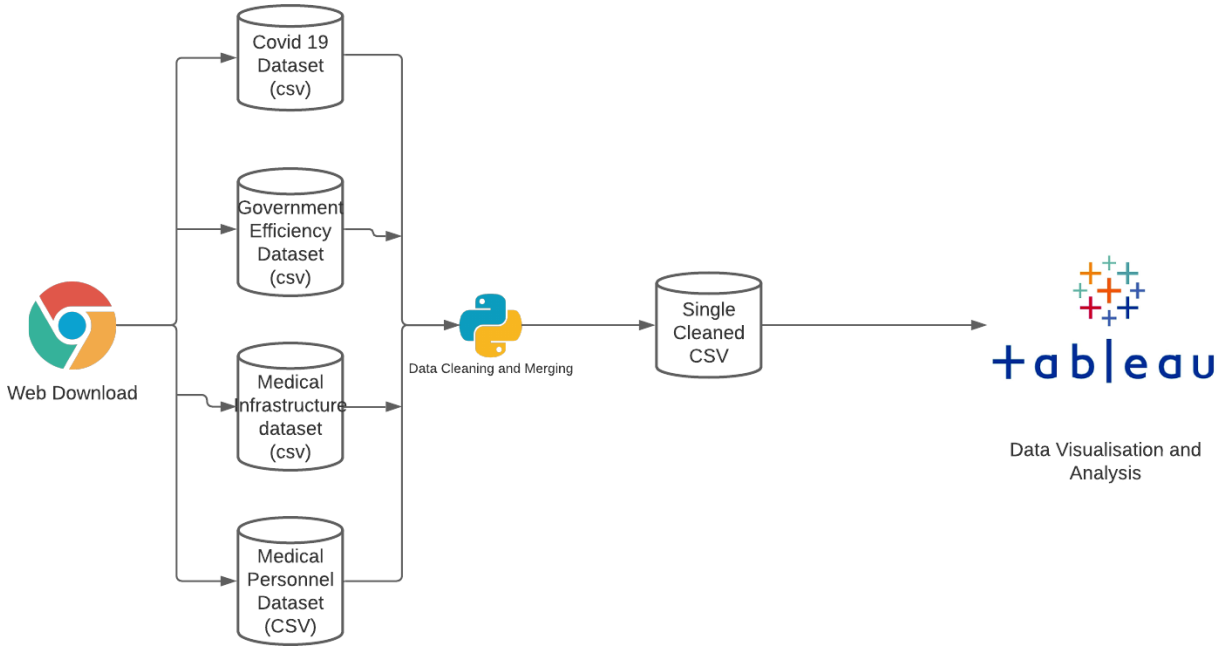


Figure 9

Instructions for Code

Cleaning data

```
#variables needed for ease of file access, files downloaded from web url are in path.

path = '/Users/mohammadansari/Downloads/'
file_1 = 'joined.csv'

file_out = 'cleaned_raw_data.csv'
file_out2 = 'healing_damage_position.csv'
file_out3 = 'total_data.csv'


#confirm the file downloaded has data needed
df = pd.read_csv(path + file_1)


#cleaning Data from duplicates and null values
df = df.drop_duplicates()
print(df)

df=df.dropna()

print(df)

#calculate column averages and replace null values with column averages

#calculate maximum values for COVID Death, tests, and cases for each country
#File write output
```

Data Merging

```
# import data files
df1 = pd.read_csv("covid_data.csv")

df2 = pd.read_csv("government_effectiveness.csv")

df3 = pd.read_csv("nurses_and_midwives.csv")

df4 = pd.read_csv("physicians.csv")

# replace location names in Covid dataset to match other datasets
df1['location'].replace(….)

# join 1st file to #2nd file on the country names
join_1 = df1.merge(df2, left_on='location', right_on='Country', how='left')

# Join first two datasets to 3rd dataset on the country names

# Join previous datasets to 4th dataset on the country names

# create subsets of columns for final export on the country names

final = join_1[['iso_code', 'continent', 'location', 'date', 'total_cases', 'new_cases',

      'new_cases_smoothed', '……]]

# rename some columns


# Export final dataset

final.to_csv("joined.csv")
```

References

https://ourworldindata.org/coronavirus-source-data
http://info.worldbank.org/governance/wgi/
https://data.worldbank.org/indicator/SH.MED.PHYS.ZS