Covid Cases Worldwide
Exploratory Analysis

Liezel Dela Cruz, ldelacruz@bellarmine.edu
Jaylin Roberts, jroberts12@bellarmine.edu

I. INTRODUCTION

The COVID-19 disease is an infectious disease caused by the SARS-CoV-2 virus. A large proportion of people were infected worldwide, and the number of COVID-19 cases continue to fluctuate daily. This dataset, extracted from Kaggle, contains information about the number of cases, deaths, recoveries, and active cases with respect to each country. We selected this dataset to analyze the correlation between these variables, in attempt to draw conclusions about the nature of COVID-19 worldwide.

II. DATA SET DESCRIPTION

This dataset contains 231 samples with 8 columns of various data types and some missing values. A complete listing is shown in **Table 1.**

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Serial Number	Ordinal/Int64	0 %
Country	Nominal/Object	0 %
Total Cases	Ratio/Float64	0 %
Total Deaths	Ratio/Float64	2.597403 %
Total Recovered	Ratio/Float64	9.090909 %
Active Cases	Ratio/Float64	8.225108 %
Total Test	Ratio/Float64	7.792208 %
Population	Ratio/Float64	1.298701 %

III. Data Set Summary Statistics

The quantitative data from this dataset can be further represented through various statistical measures. A complete listing is shown in Table 2.

Table 2: Summary Statistics for Covid Cases Worldwide

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Serial Number	195.0	1.071897e+02	6.430396e+01	1.0	51.5	104.0	162.0	2.250000e+02
Total Cases	195.0	3.329258e+06	1.019503e+07	1403.0	37865.5	297757.0	1723625.0	1.041969e+08
Total Deaths	195.0	3.378704e+04	1.125272e+05	1.0	313.0	3155.0	16877.0	1.132935e+06
Total Recovered	195.0	3.197261e+06	9.846399e+06	438.0	34699.5	288991.0	1708095.0	1.013228e+08
Active Cases	195.0	9.821017e+04	7.988122e+05	0.0	78.0	1253.0	10848.5	1.095262e+07
Total Test	195.0	3.375474e+07	1.220386e+08	7850.0	401044.0	2610114.0	14772746.5	1.159833e+09
Population	195.0	3.207370e+07	1.099820e+08	4965.0	1100457.0	6844597.0	27826923.0	1.406632e+09

Furthermore, we can analyze the frequency and proportion of each categorical data (in this dataset, "Country" is the only one that is categorical). A complete listing is shown in **Table 3**. Notice that the frequency is 1 and proportion is 0.5128% for each country.

Table 3: Frequency and Proportions for Country

Country	*Frequency*	*Proportion (%)*
USA	*1*	*0.5128 %*
Bahamas	*1*	*0.5128 %*
Malawi	*1*	*0.5128 %*
Ivory Coast	*1*	*0.5128 %*
New Caledonia	*1*	*0.5128 %*
...	*...*	*...*
Panama	*1*	*0.5128 %*
Mongolia	*1*	*0.5128 %*
Nepal	*1*	*0.5128 %*
Belarus	*1*	*0.5128 %*
Montserrat	*1*	*0.5128 %*

Next, the correlation coefficient of each continuous variable can be calculated to find the relationship between each variable. **Table 4** and **Figure 1** displays this information. Here we can identify the variables with a high and low correlation.

Table 4: Correlation Table

	Serial Number	*Total Cases*	*Total Deaths*	*Total Recovered*	*Active Cases*	*Total Test*	*Population*
Serial Number	*1.000000*	*-0.463011*	*-0.412296*	*-0.460312*	*-0.177270*	*-0.383558*	*-0.288476*
Total Cases	*-0.463011*	*1.000000*	*0.878457*	*0.997221*	*0.346938*	*0.842994*	*0.522161*
Total Deaths	*-0.412296*	*0.878457*	*1.000000*	*0.885198*	*0.159427*	*0.795416*	*0.572756*
Total Recovered	*-0.460312*	*0.997221*	*0.885198*	*1.000000*	*0.276273*	*0.851625*	*0.525893*
Active Cases	*-0.177270*	*0.346938*	*0.159427*	*0.276273*	*1.000000*	*0.149469*	*0.101207*
Total Test	*-0.383558*	*0.842994*	*0.795416*	*0.851625*	*0.149469*	*1.000000*	*0.665094*
Population	*-0.288476*	*0.522161*	*0.572756*	*0.525893*	*0.101207*	*0.665094*	*1.000000*

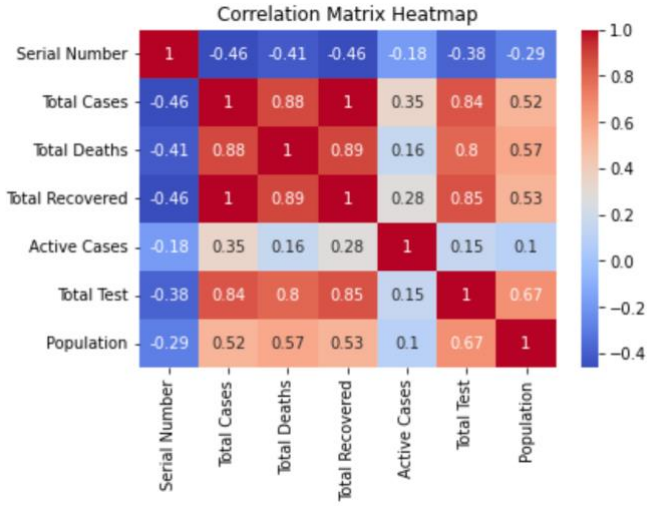

Figure 1: Heatmap of Correlation Matrix

IV. DATA SET GRAPHICAL EXPLORATION

Graphical representations of datasets can help make sense of the relationship between variables. Using Python, NumPy, Pandas, Matplotlib.pyplot, Seaborn, and Tableau, we will analyze the distribution of these graphs.

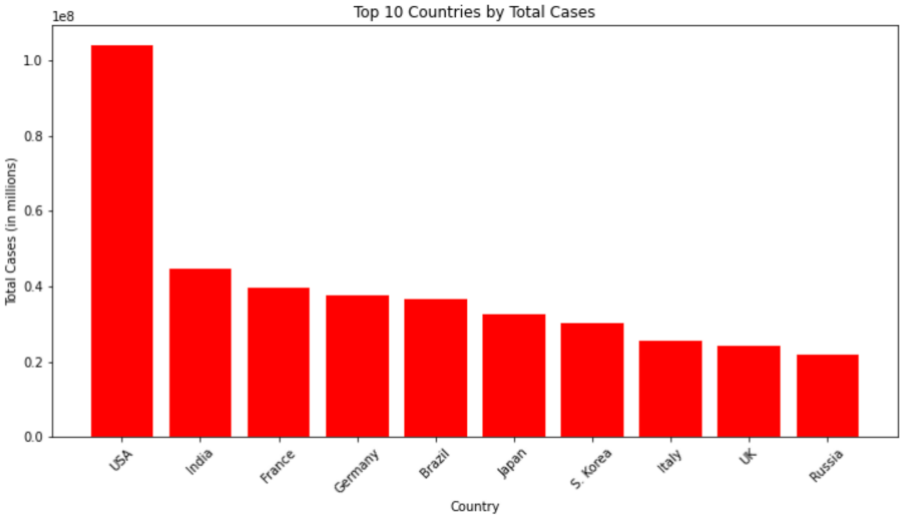

Figure 2: Bar Chart Comparison of Top 10 Countries against Total Number of Cases

Figure 2 shows the total number of COVID-19 cases in the top 10 countries: USA, India, France, Germany, Brazil, Japan, South Korea, Italy, UK, and Russia. The USA has the highest number of cases (104,196,861) whilst Russia has the lowest out of the 10 countries with 21,958,696 cases. This means the USA has approximately 5 times more than the number of cases in Russia, which is a significant difference.

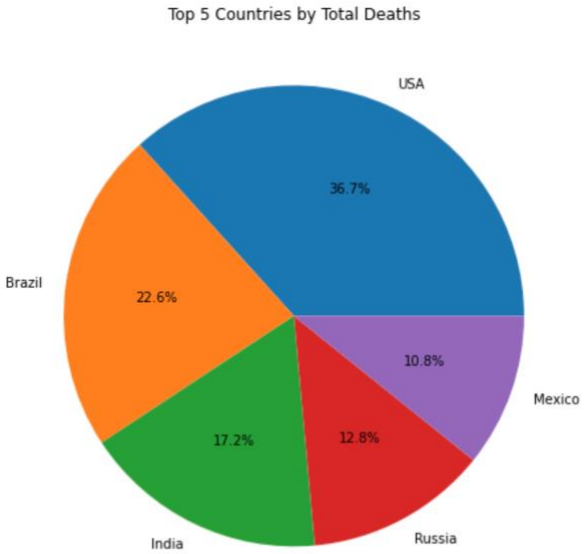

Figure 3: Pie Chart showing the Total Deaths in the Top 5 Countries

Figure 3 displays the proportion of deaths among the top 5 countries: USA, Brazil, India, Russia, and Mexico. Here, we can see the highest proportion of total deaths due to COVID-19 is in the USA with 36.7%. Whereas, Mexico has the lowest proportion of total deaths due to COVID-19 with 10.8%.

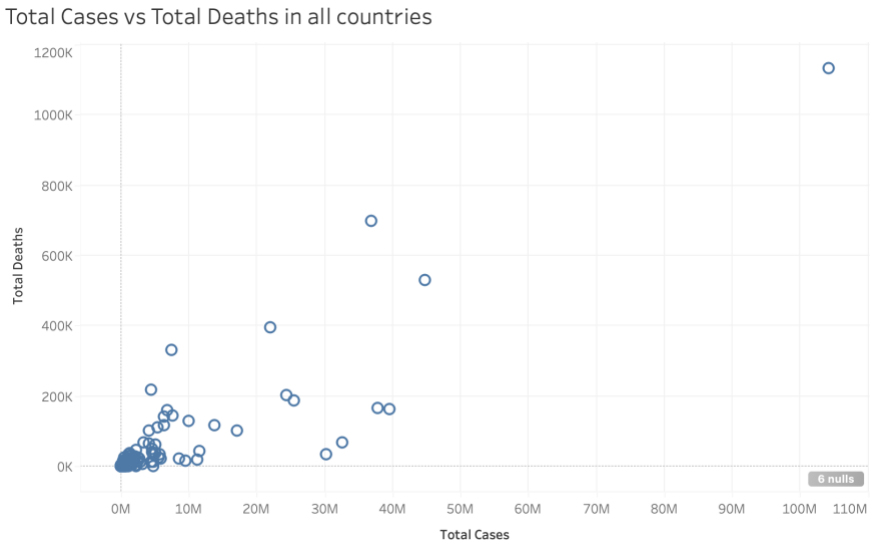

Figure 4: Scatter Graph of Total Cases against Total Tests in all countries

Figure 4 presents the total number of cases against the total number of tests in every country of the dataset. From the graph, we can see that most of the countries are highly concentrated around 10 million cases and 100,000 deaths. In general, there is a weak positive correlation between the two variables, which means that as the total number of cases increases, so does the total number of deaths. However, there is one extreme value (which may be considered as an anomaly) – USA. It has a total of 104,196,861 cases and 1,132,935 deaths which is significantly more than the other countries. We may consider this as an outlier.

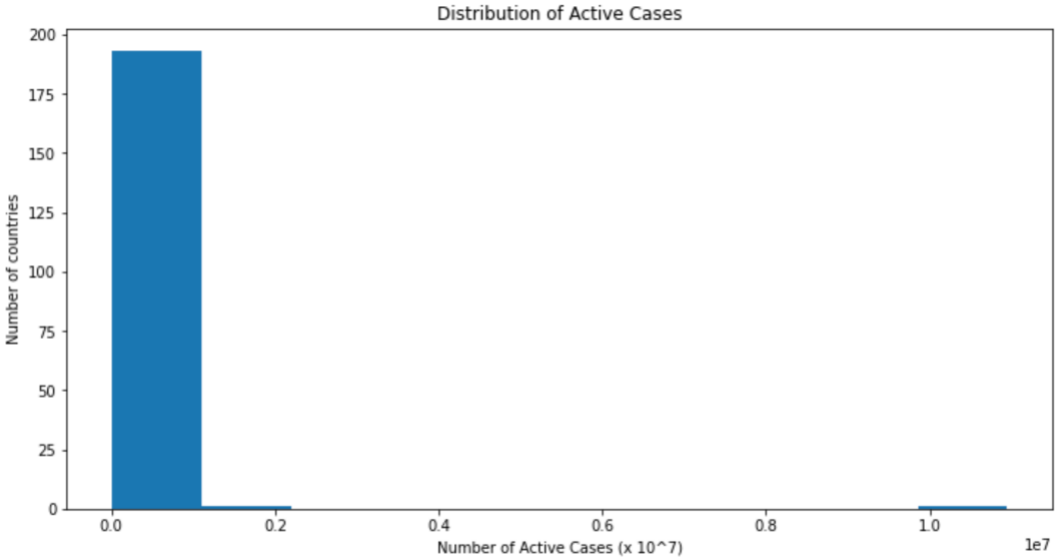

Figure 5: Distribution of Active Cases

Figure 5 illustrates the frequency for the number of active cases. Evidently, we can see that the majority of countries have active cases ranging from 0 to 1,000,000. However, there are 2 countries that have a total number of active cases above this range. USA with 1,741,147 and Japan with 10,952,618 active cases.

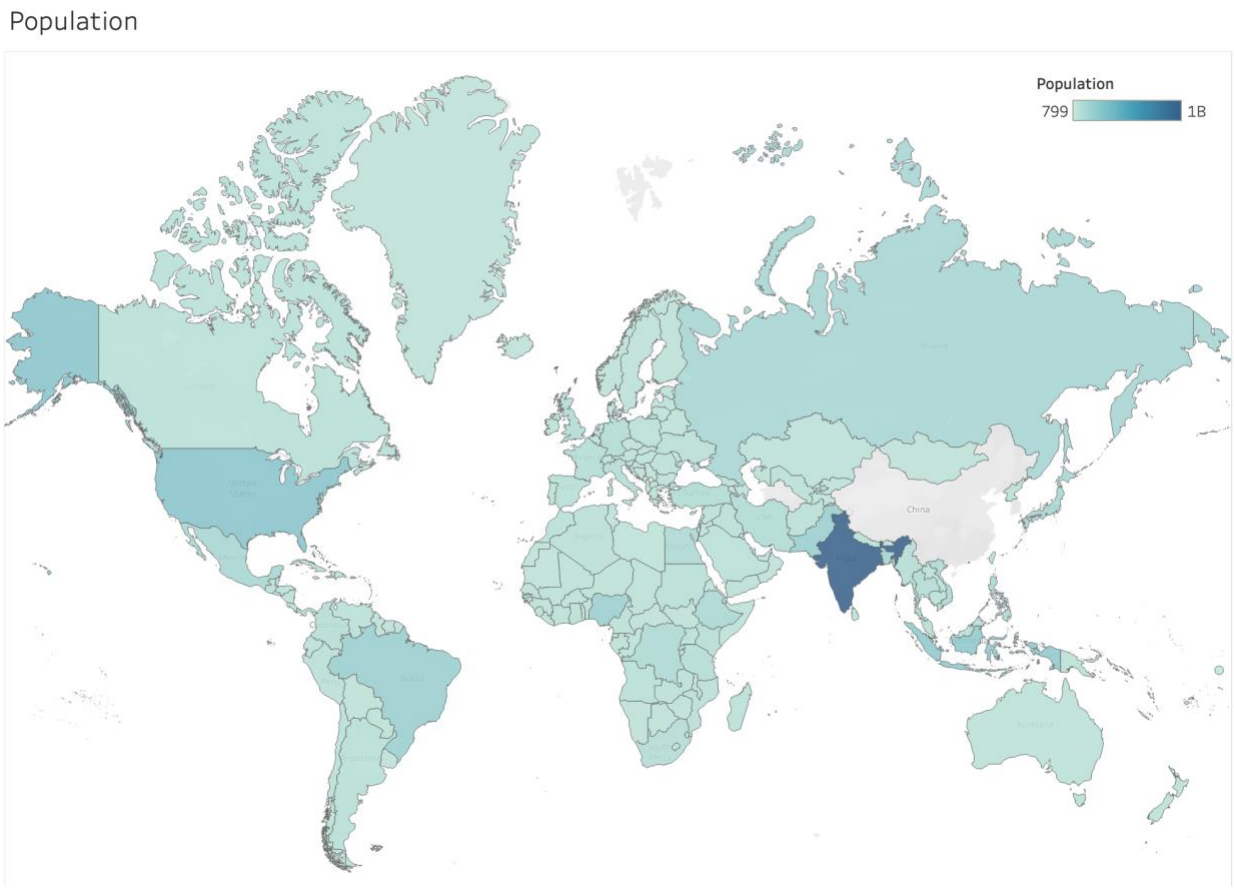

Figure 6: Population of Each Country

Figure 6 shows the population of each country in the world. The darker regions on the map signify a high population, whilst the lighter regions mean small population. In this dataset, we can see that India has the highest population with 1,406,631,776 people. However, we know that China has the highest population. But since the dataset has this entry as a null value, we are unable to see this demonstrated on the map. The next darkest region is the USA with 334,805,269 people.

V. SUMMARY OF FINDINGS

In conclusion, COVID-19 has spread across the planet, but more so in certain places of the world. Whilst this dataset displays India to be the most populated country, it does not have the highest number of cases. In fact, the USA significantly has the largest number of total cases, despite it only being the second largest populated country in this dataset. Moreover, the USA has the largest number of deaths with 1,132,935 and takes up 36.7% of total deaths among the top 5 countries. On the other hand, whilst the USA has the highest number of total cases and deaths, Japan has the largest number of active cases, by approximately 6 times more than the USA. There also seems to be a correlation between the total number of cases and number of deaths recorded per country. This suggests that lower number of cases can help result in fewer deaths. As such, Japan (the country with the highest number of active cases) must enforce stricter measures to help reduce the spread of the COVID-19 virus in order to prevent a high number of deaths, as a result of their high number of cases.