**Analysis of Covid-19 Dataset**

Group 5 Report

horizontal line

# Introduction

In this report we are analyzing Covid-19 dataset of the following countries: *'Panama', 'Papua New Guinea', 'Paraguay', 'Peru', 'Philippines','Poland', 'Portugal', 'Qatar', 'Romania', 'Russia', 'Rwanda', 'Saint Kitts and Nevis', 'Saint Lucia', 'Saint Vincent and the Grenadines', 'San Marino','Sao Tome and Principe', 'Saudi Arabia', 'Senegal', 'Serbia', 'Seychelles', 'Sierra Leone', 'Singapore', 'Slovakia', 'Slovenia', 'Somalia', 'South Africa', 'South Korea', 'South Sudan'*. Our dataset consists of 5,452 records (rows) with 10 variables (columns) from 28 Countries. Our report consists of the following sections: **Data Cleaning, Data Analysis & Results and Conclusion.**

# 

# Data Cleaning

### Missing Values

In this section we looked at missing values from our ten (10) variables: **Date, Country/Region, Confirmed Deaths, Recovered, Active, New cases, New deaths, New recovered, WHO Region.** We found that there are no missing variables in our dataset. However if they were present, depending on the no of missing values present, we can do one of the following:

* Remove rows with missing values (advisable if there are not many)
* Replace the missing value with the mean/median/mode of the column if is a numeric value
* Replace the value with the mode of the column in categorical variables(e.g. if the missing value is female and there are many females than males, there are high chances the missing value is female)

### Anomalies

* No duplicates records (rows) in the data.
* According to descriptive statistics, we do have anomalies on the data as we have negative values for min of new death and new recovered fields (refer to anomalies section on the notebook).

# 

# Data Analysis & Results

### Trends (Refer to figure 1 below)

* All countries recorded zero in February 2020 which makes sense as we started seeing many countries going on lockdown in the middle of March 2020.
* Spain was the first country to record cases (Middle of March 2020 going to April 2020) and the number of cases remained steady throughout.
* Russia started recording cases in the middle of April and the cases spiked within a short period of time recording extremely high numbers from the rest of the countries.

**Figure 1: Confirmed cases over Time by Country**

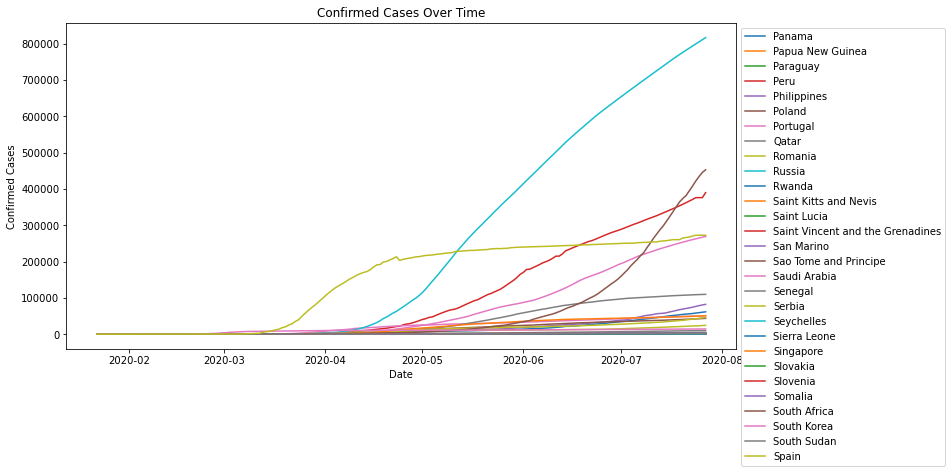


Table 1 below summarizes the results of the most and least confirmed cases, no of deaths and recovered cases.

|  |  |
| --- | --- |
|  | **Countries** |
| Most no of confirmed cases recorded | Russia (**45,408,411** cases) |
| Least no of confirmed cases recorded | Papua New Guinea (**1,185** cases) |
| Most no of deaths recorded | Spain with (**3,033,030** cases) |
| Least no of deaths recorded | Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Seychelles  ( **0** cases) |
| Most no of recovered cases recorded | Russia **(2,512,448** cases) |
| Least no of recovered cases recorded | Papua New Guinea ( **695** cases) |

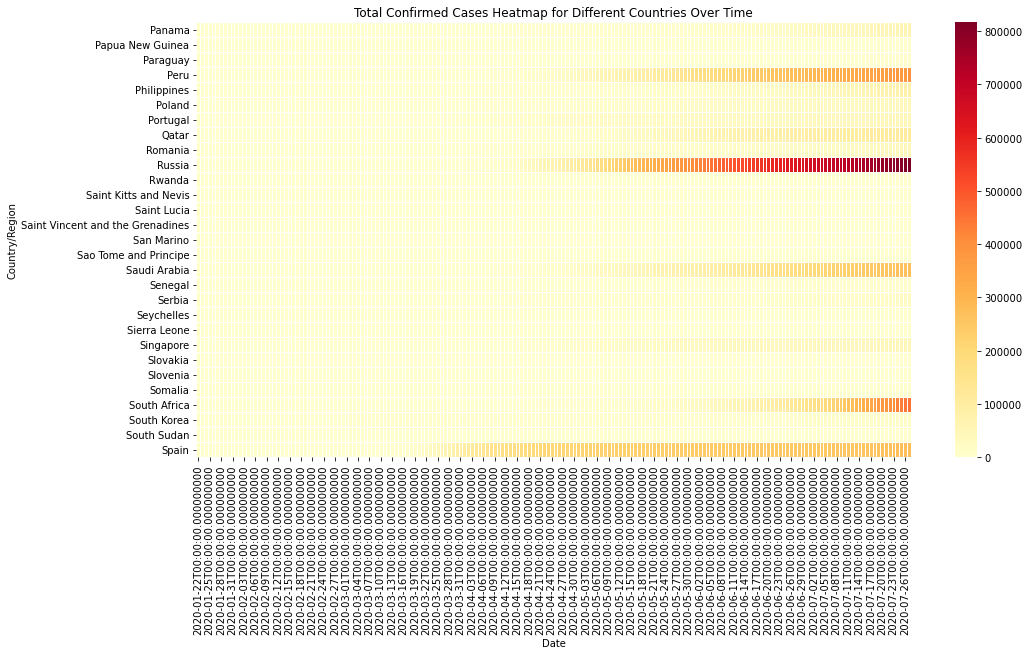
**Table1: Summary of Most and Least Cases**

Although Russia was the country with highest no of recoveries, it is not the country with the highest rate of recovery. Russia’s recovery rate is **55.32%** and the country with the highest recovery rate is **Saint Lucia** with **83.18 %**. The country with the lowest recovery rate is **South Sudan** with **19.46 %.** In terms of mortality rate, Spain is the country with the highest mortality rate of **11.07 %** while all countries who did not record death as per the table above will be the lowest in terms of mortality which is **0.0%**.

It is noted that most countries that recorded 0 deaths are countries with less than 5,000 confirmed cases except **Papua New Guinea**. Papua New Guinea although recorded 1,185 confirmed cases, they also recorded 2 deaths. This proves that small cases were easy to manage hence low mortality for such countries.

Records of Covid-19 cases started from 22 March 2020 which is when the first country started recording cases soon after other countries followed which corresponds to when most countries started to implement lockdown (refer to **figure 2** below). Refer to the **Bonus** section of Group5 Notebook to see the trends and numbers of the top 10 and bottom 10 countries for both recorded cases, death and recoveries.

**Figure 2: Confirmed Cases Compare between Countries over Time**



**Conclusion**

In conclusion Spain is the country that was affected the most by Covid-19 due to their highest mortality rate. Despite Russia having the highest confirmed cases which is understandable as a highly populated country, they have managed to keep their mortality rate low and have a good recovery rate. Covid-19 impact is seen mostly in 20 countries out of the 28 (see distribution plot of Group5 notebook in grouping section).