Exploratory Analysis of Covid-19

General introduction, language and libraries used for the analysis

The objective that I had in mind when I was given this assessment was to try and predict certain patterns, trends and correlations.

Therefore, I will try to extract conclusions that will help me to comprehend behavior tendencies in order to solve potential problems and identify probable patterns that enhance the statistics of this outbreak.

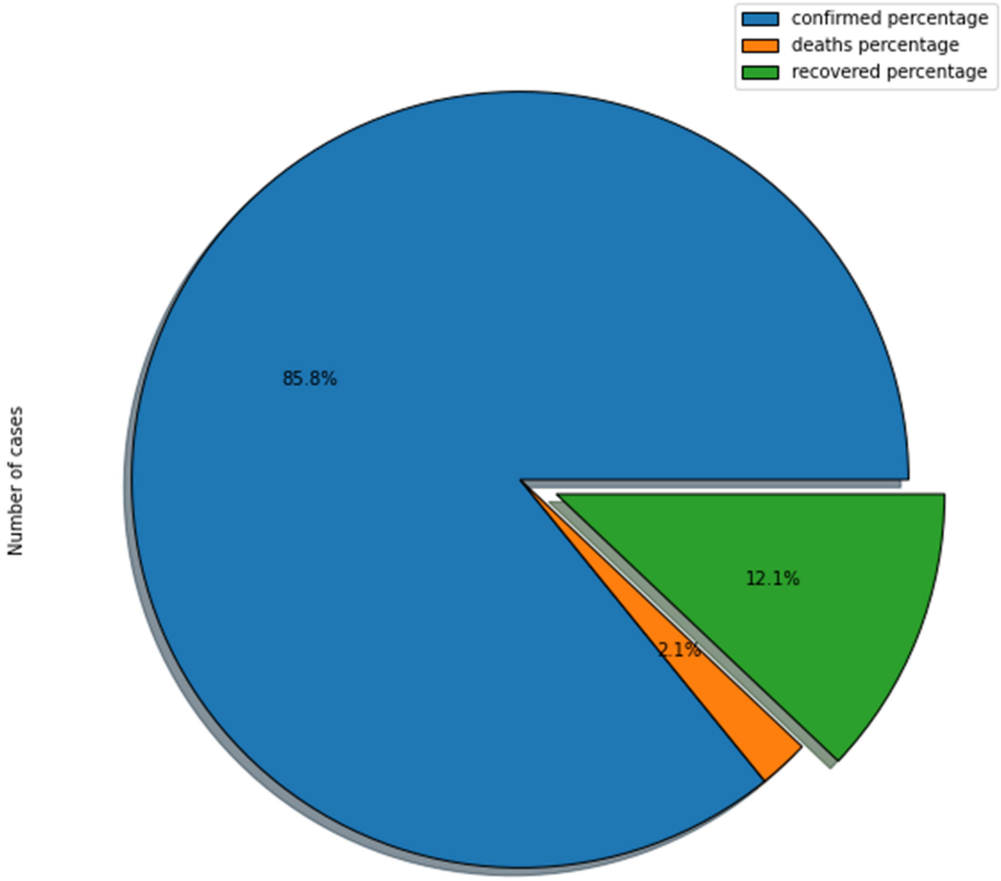
I've done this analysis with Python on Jupyter Notebook using libraries such as Pandas, Matplotlib, Requests and Pycountry.

I've also used the API:
https://covid-tracker-us.herokuapp.com/#/v2/get_latest_v2_latest_get
to get the data that I need for this analysis.

Data of each Continent's latest confirmed, deaths and recovered cases

Introduction:
I've requested the API on 'latest' and 'locations', which gave me the info from all the countries and some other places where the outbreak occurred (such as ships), and for each of these the latest deaths, recovered and confirmed cases of Covid-19. The first thing I did was to obtain all the continents from where each country belongs to, afterwards I've plot a bar graph to see the total cases of each of these columns that show the cases.

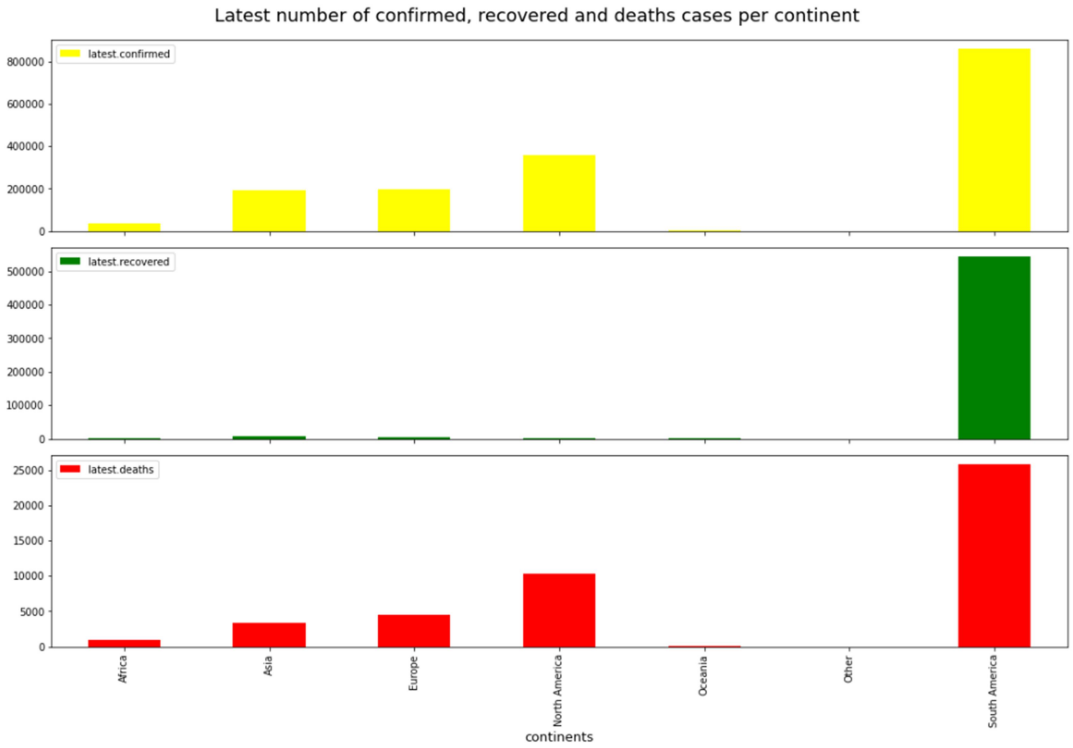
Latest number of cases based on confirmed, deaths and recovered cases



I've calculated the percentage of the recovered, deaths and confirmed cases (*) out of the total cases of Covid-19. From this analysis I can conclude that the ratio of recovered patients is 600% higher than the deaths percentage, the death outcome is 1 in 7 –but please check the note below.

(*) NOTE: The API provided for this exercise does NOT match the info from Worldometer's (https://www.worldometers.info/coronavirus/). Recovered cases here accounts for 7.5+ M, while the other accounts for 38+ M, therefore this analysis is only valid as an exercise.

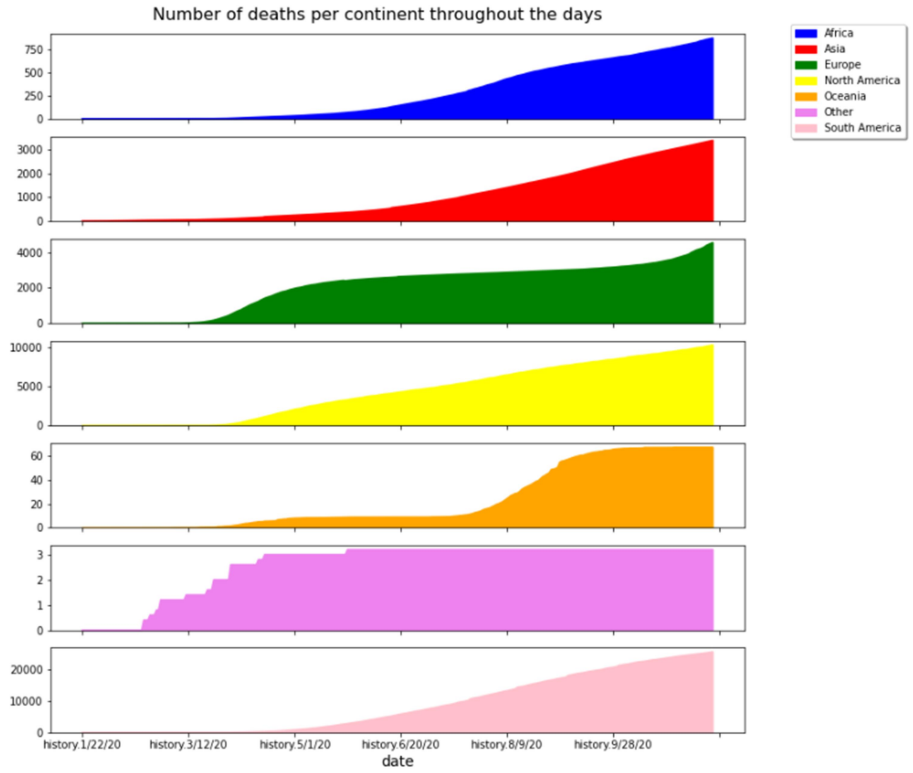
Latest number of cases based on confirmed, deaths and recovered cases per continent

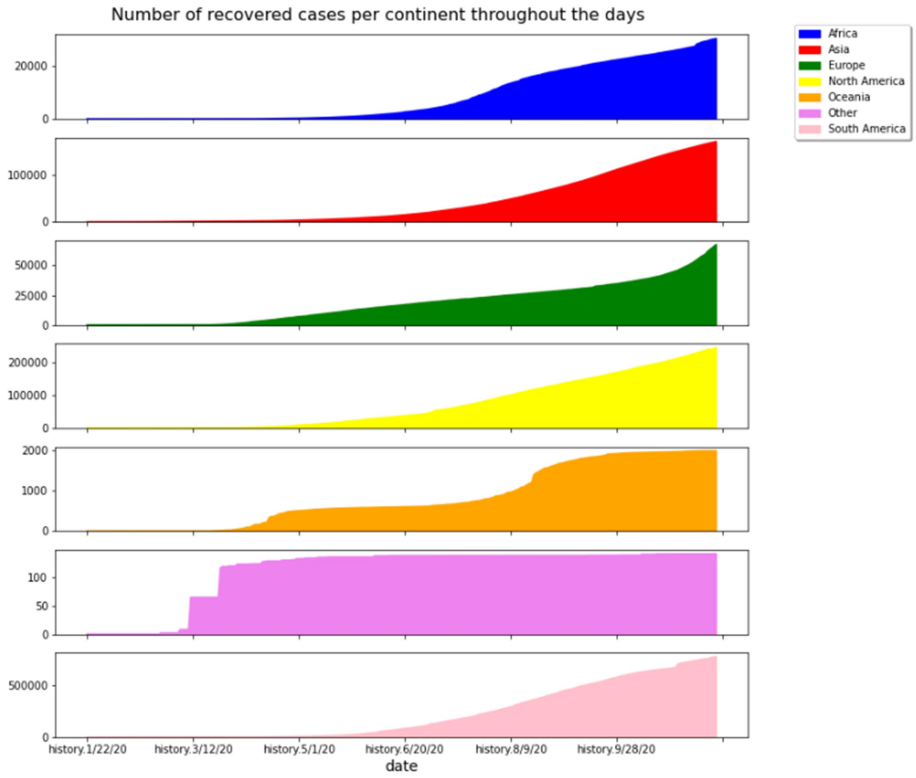


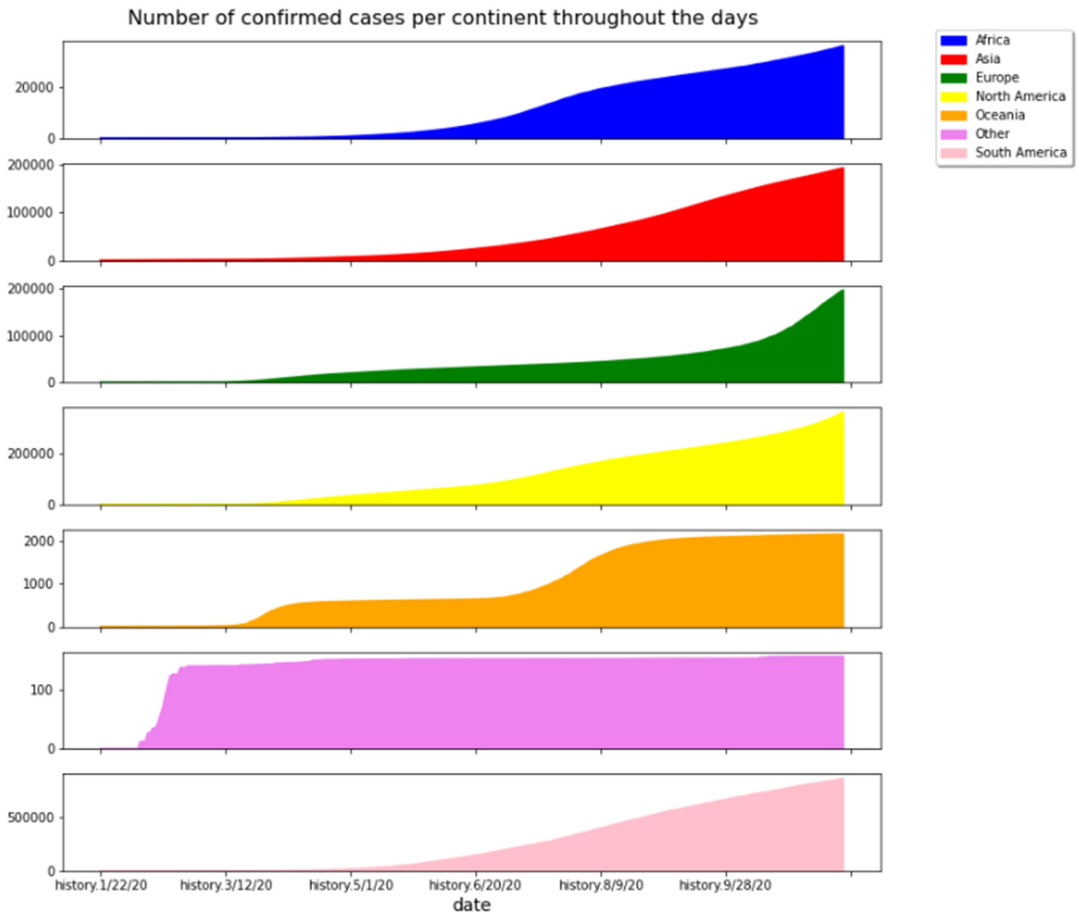
The data I've analyzed was in a good state (from a consistency perspective), I haven't found any Nulls that'd carry some kind of problem or error during the processing and analysis of these.

As we see here, the cases of Covid-19 seem to be concentrated on South America with approximately 800K confirmed, 500K recovered and 250K deaths from the latest data that we have, whilst on the second place is North America with about 400K confirmed cases, less than 100K recovered cases and less than 150K deaths.

Latest number of cases based on confirmed, deaths and recovered cases per Continent throughout the days

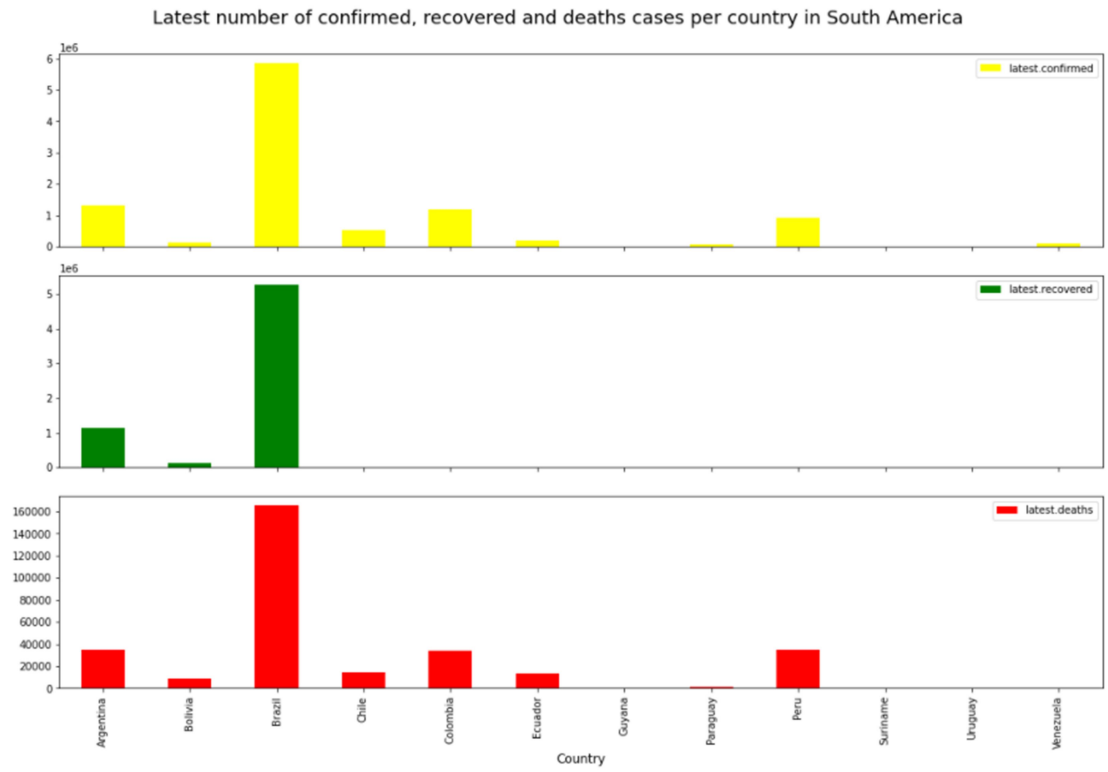






Here we can see that South America leads in deaths, recovered and confirmed cases, second to North America and followed by Asia, it's interesting to see here the exponential function of the virus, where people infect others and these infect others too, and while the recovery rate is slow compared to the rate of new confirmed cases, nonetheless the virus does seem to be slowing down and becoming a constant instead of growing, that seems to be the case in Oceania, while in Europe, it's happening the opposite, it's growing at an unexpected rate all of a sudden (approximately on October, which match the second wave of the pandemic. All of the continents above (excluding others) seem to have begun the outbreak between 3/12/20 and 5/1/20

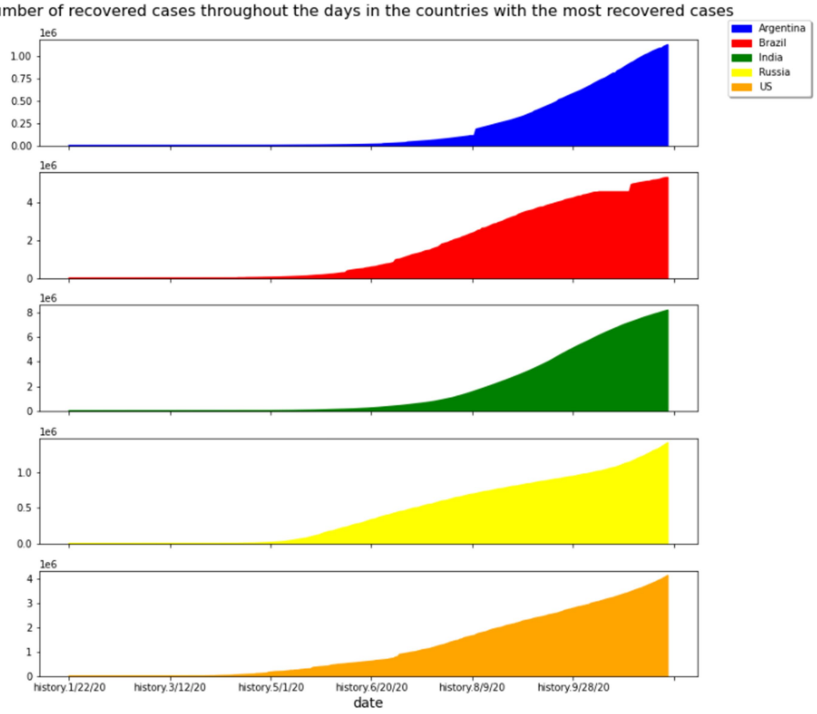
Based on this outcome, I've decided to focus the analysis on South America, since the quantity of confirmed cases may give more accurate info.



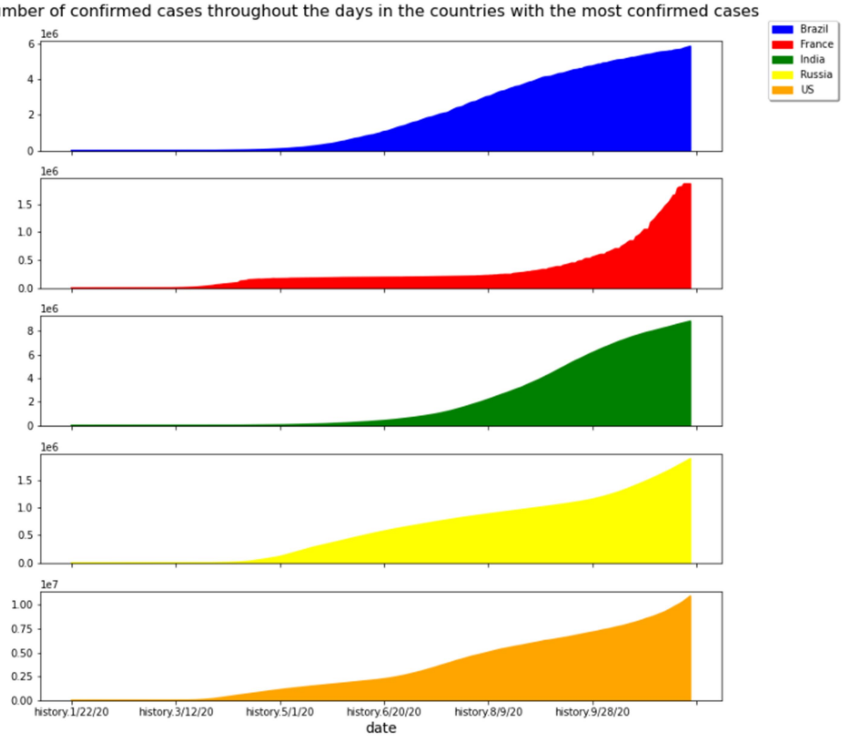
From this analysis, I can conclude that the focus of the outbreak –the place where there're more cases- is Brazil.

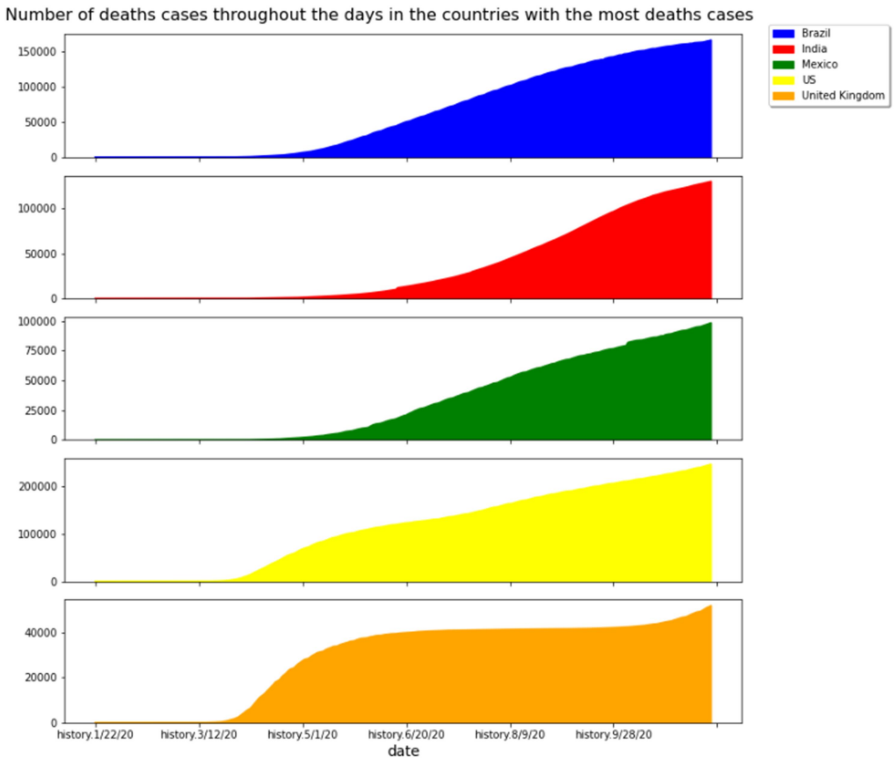
Let's analyze the development of the outbreak through time between the first five countries with the largest number of confirmed cases:

Number of recovered cases throughout the days in the countries with the most recovered cases



Number of confirmed cases throughout the days in the countries with the most confirmed cases





Here we have:

- Brazil, Argentina, India, Russia, US in the countries with the largest number of recovered cases.
- Brazil, India, Russia, US with the largest number of confirmed cases.
- Brazil, India, Mexico, US and United Kingdom with the largest number of death cases.

Leading with the largest number of recovered cases is **India** with **8M recovered** patients. In contrast, it has a little more than **8 M confirmed** cases and **130K deaths**, which mean that at the rate that it gets a new infected person it has a lot of chances to recover (**98.441%**).

In the second place is **Brazil**, with **5M recovered** patients, **6M confirmed** cases and **160K deaths**, this means a **97.4%** chance of recovery which is still pretty high.

In the third place is **US** with **5M recovered** patients, **11 M confirmed** cases and **250 K deaths**, this means a **95.23%** chance of recovery.

On the other countries, the ones that don't appear on all of these graphs, I've reached to some conclusions about it in comparison to the ones that do appear on all of the plots.

Argentina appears on the top 5 of recovered cases; that is a good sign, because it doesn't appear on the most confirmed cases and neither on the most deaths cases, which means that the percentage of recovery is high.

Russia has **1.5M recovered cases** and **2M confirmed cases** but it doesn't appear on the deaths top 5. This means that the survival rate is above average at least.

France doesn't appear on the most recovered cases, but it appears on the most confirmed cases. What's curious about France is that it's the only one that doesn't seem to have a linear growth, instead, it seems to have had a sudden explosion of confirmed cases in September, and therefore, it will probably take some time to calculate an accurate mortality rate, but with the latest data we can deduce since there's a lot of confirmed cases and not enough recovered cases, the recovery percentage has to be below average.

Mexico doesn't appear on the most recovered and neither does it appear on the most confirmed cases, which means that, in the total of cases, Mexico has the highest counts of death per confirmed cases out of these countries, which means that Mexico, as France, has a recovery percentage below average, which means the mortality rate is, in contrast, above average.

Final suggestions

I think I would have made a more accurate and better analysis if I had more datasets related to the quantity of the days of the quarantine (if available) for each country, the economic status throughout the months, maybe the compliance of the people with the quarantine too (for example, how much people actually comply to the quarantine throughout the days), perhaps the weather and conditions that would help the virus spread, and lastly, how advance each country is (in terms of medicine, economy and so on).

I could've drawn more conclusions if I were to have compared each and every country on the dataset on each day to see where exactly did the covid start in every country, how it compares to the other countries and how effective the measures taken by the government were.