COVID-19 Dataset

Visualization and Exploratory Data Analysis

Abtin Mahyar

1. Data Summary

The COVID-19 dataset which is provided by OWID organization has valuable information about the cases, deaths, vaccinations, performed tests, and other related variables in different countries around the world which has been updated daily. This dataset has around 130,000 records with various types of fields which each record refers to the situation of a particular location in an exact date. The dataset has records from the begging of 2020 to 11/16/2021. Each record has 67 attributes which are described completely in a CSV file. As it can be suspected based on these attributes, some of the features are useless or illegible, which means they must be discarded during the data preprocessing. Also, dataset has some null values in some of its' features which should be managed during preprocessing. Moreover, there are null values in every single column of the dataset except in location and date which should be replaced by a meaningful value or drop in data preprocessing section.

2. Data Preprocessing

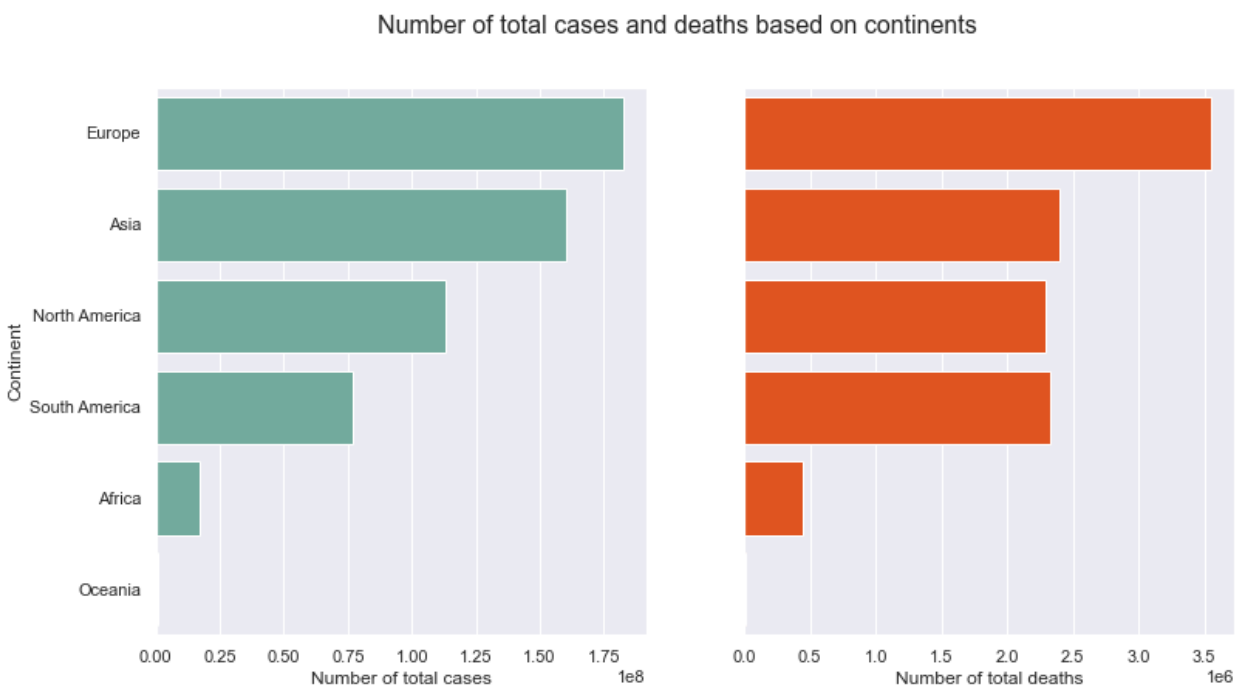
The preprocessing done in the notebook consist of the following steps:

1. Some of unrelated, unvaluable, and redundant columns were dropped. The resulted dataset has 30 columns and dropped columns list can be seen at the notebook.
2. Null values are eliminated or filled with a specific value in some columns or filled based on forward propagate method, in other columns null values remained unchanged due to further analysis and not losing information.
3. Some values in location column were meaningless or illegible; therefore, rows with those values were dropped in analyses related to location.
4. Data types are fixed and each attribute convert to its' own data type.
5. New features are extracted from the dataset such as 'new cases per new deaths ratio' and 'ICU per hospital patients'
6. There were some outliers in number of new deaths which were ignored in the analyses were related to this column.
7. Because there are still some null values in the dataset, different data frames are made up from the original dataset for different analysis in 'number of cases and deaths', 'vaccinations', 'patients', 'performed tests', and other variables, so that we do not lose meaningful data in each analysis.

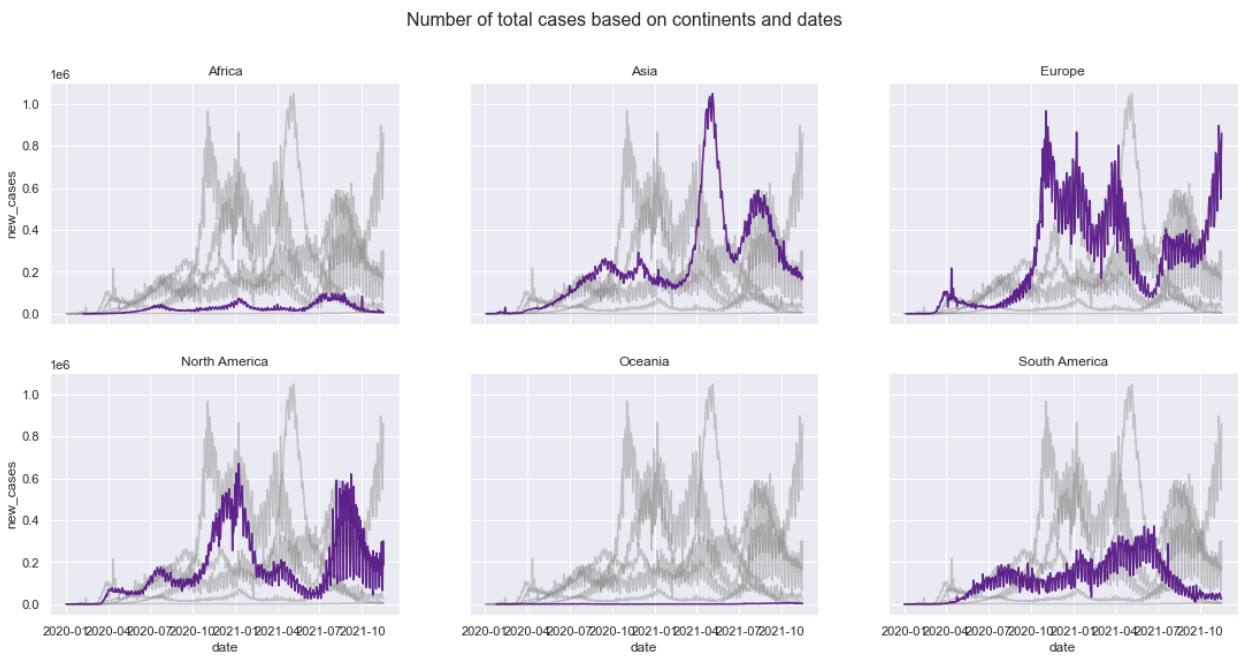
3. Exploratory Visualization

3.1. Cases and Deaths

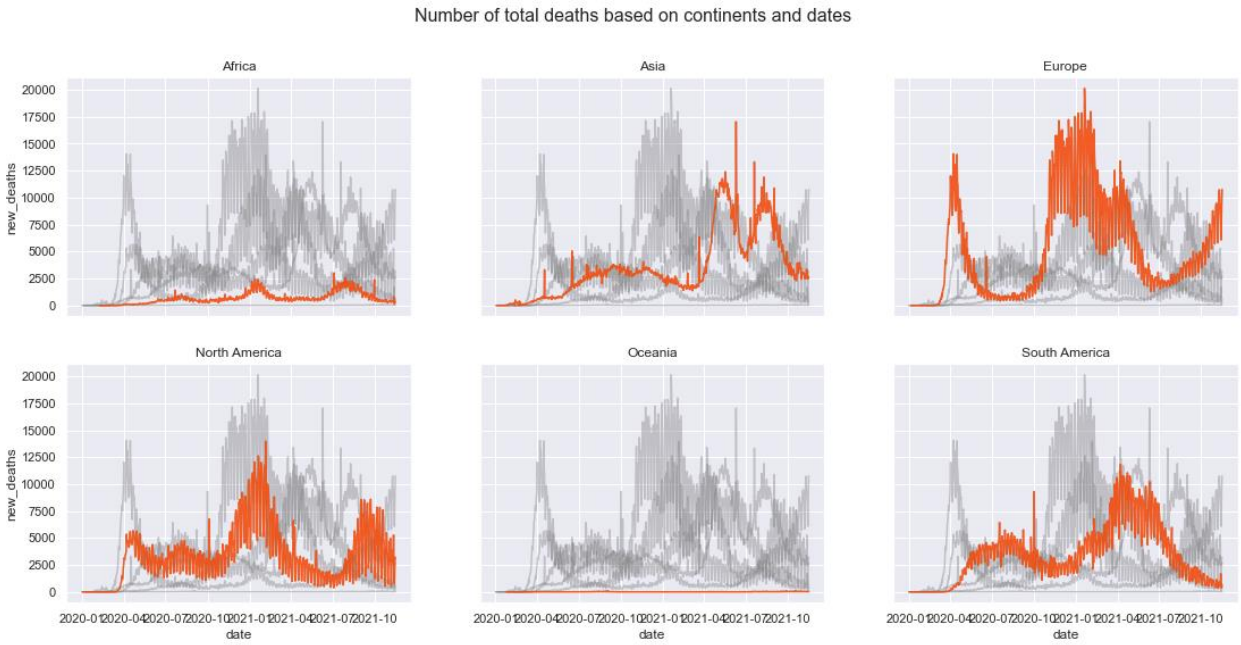
The following plots demonstrate number of total cases and deaths globally on different dates. The line graphs show the trend of number of cases and deaths in different dates. Each point in scatter plot refers to a certain date with the specific number of cases and deaths. Also, a regression line is drawn on this plot which shows that these two variables are positively correlated. There is an increasing trend in number of cases and deaths from 2020 to 2021. After that, these numbers fluctuated until the current day. However, as it can be seen from the line charts and the scatter plot, these two variables are highly correlated which means that with the increase of cases, number of deaths increases. Note that cases and deaths have different scales therefore we can just compare the trend between these two variables. Generally, number of cases in each date is much more than number of deaths.



A figure showing the number of total cases and deaths in each continent in the whole dataset. As it can be understood from the bar charts Europe has the greatest number of total cases and deaths compared to the other continents and there is a considerable difference in number of deaths in Europe and other continents. Moreover, Africa and Oceania have the least number of cases and deaths which is mainly because of lack of data from these continents compared to the others. Also, South America has fewer number of cases compared to North America but, these continents have the same number of total deaths which means that the probably of death of confirmed cases in South America are greater than this probability in other continents. Note that cases and deaths have different scales.

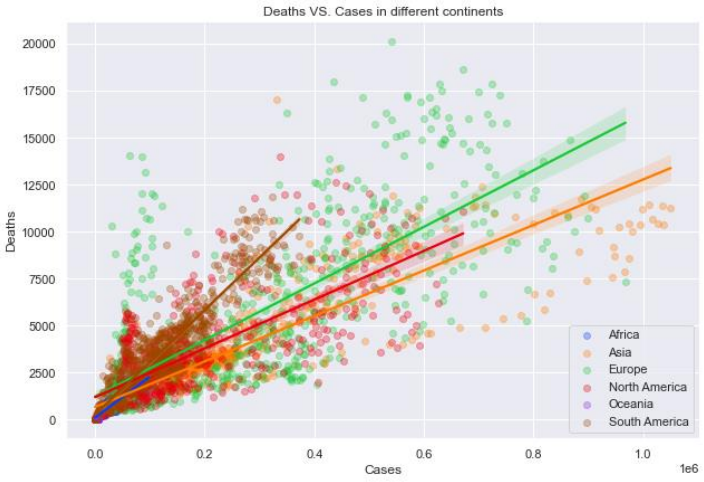


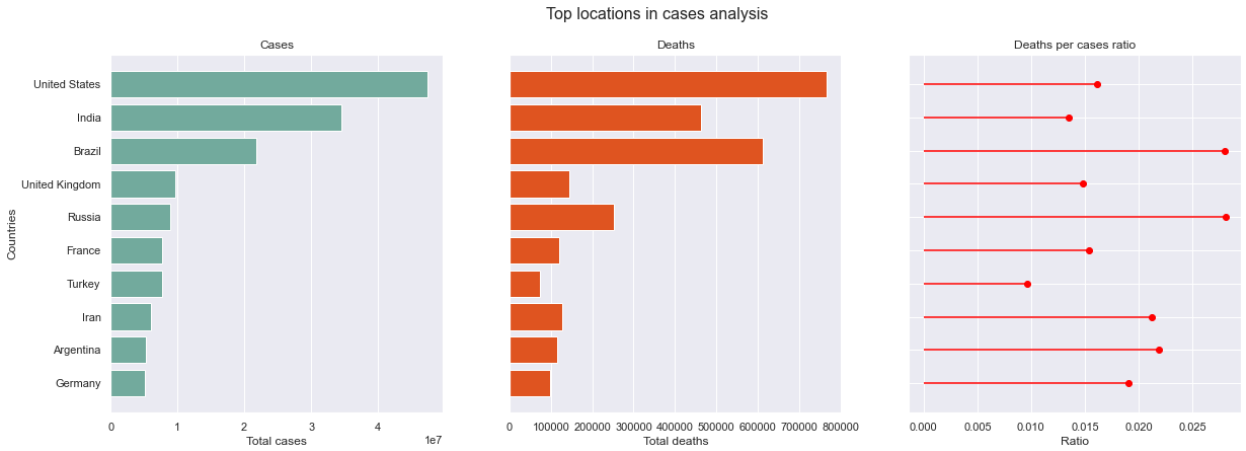
The following plots illustrates the number of total cases in each date in different continents. As it can be seen from the plots, trend of new cases in different continents are different from each other.

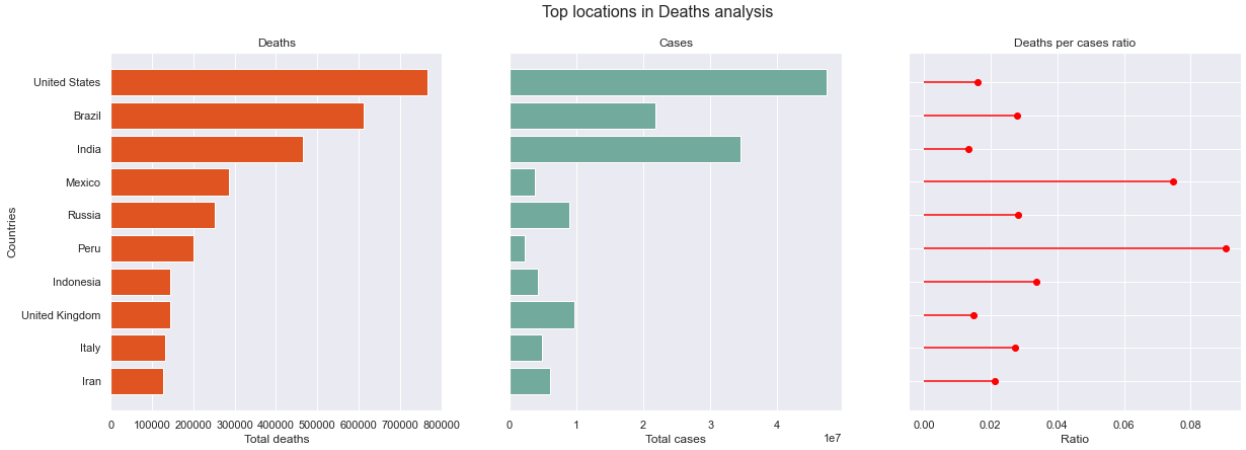


The following plots illustrates the number of total deaths in each date in different continents. As it can be seen from the plots, trend of new deaths in different continents are different from each other. As it expected, trend of new cases is same as new deaths for each continent.

A scatter plot which every single point refers to a single record with certain number of cases and deaths in a particular continent and date which specified with a color. As it can be seen from the plot, correlation between number of deaths and cases are different from each continent. For instance, South America has the highest correlation coefficients for These two variables and Asia has the least coefficients.



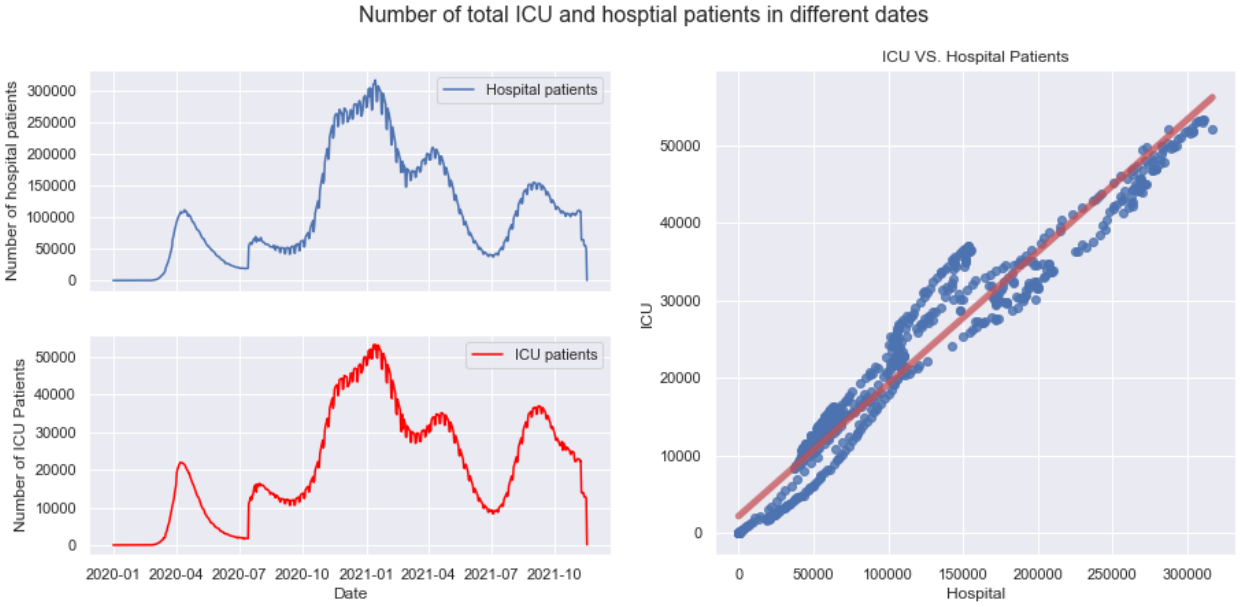




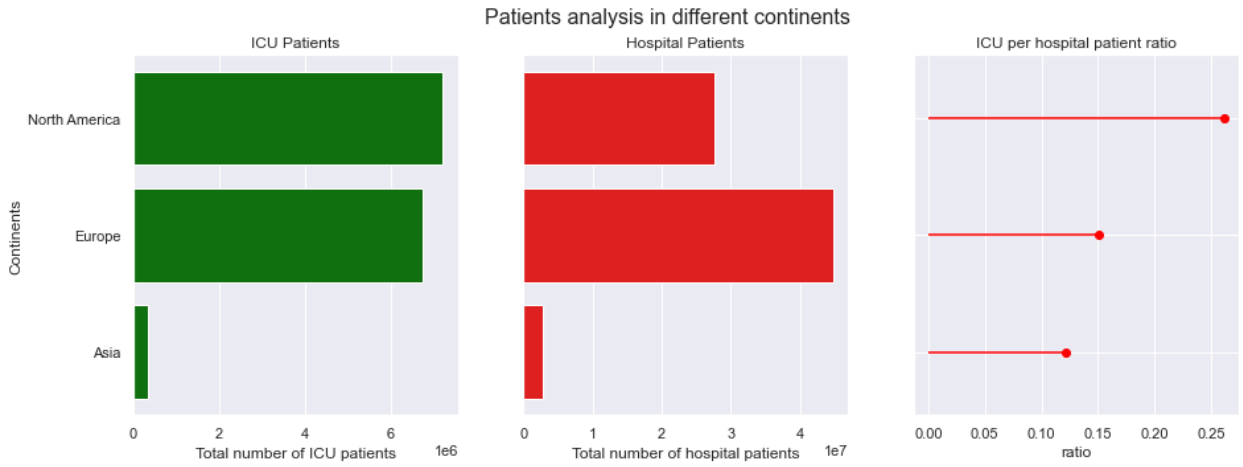
Following plots demonstrates total number of cases and deaths in locations with the greatest number of cases or deaths. Plots in the first row shows the countries with highest number of cases. As it can be seen, United States has the greatest number of cases and deaths among the others. Also, Brazil and Russia are the two countries with highest ratio of deaths per cases and Turkey has

the lowest ratio compared to the others. Plots in the second row shows the countries with highest number of deaths. Mexico and Peru have high value of death per case ratio too. USA, Brazil, UK, Russia, Iran are the countries which were existed in both analysis; therefore, in further analyses I use these countries as location. Note that cases and deaths have different scales.

3.2. ICU and Hospital Patients



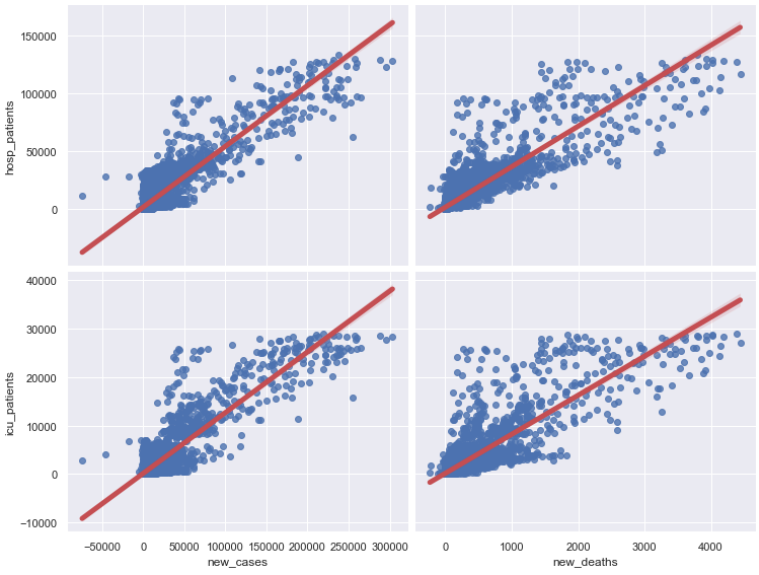
The above plots demonstrate number of total ICU and hospital patients globally on different dates. The line graphs show the trend of total number of hospital and ICU patients in different dates. Each point in scatter plot refers to a certain date with the specific number of ICU and hospital patients. Also, a regression line is drawn on this plot which shows that these two variables are positively correlated. There is an increasing trend in number of ICU and hospital patients from 2020 to 2021. After that, this trend fell dramatically until July of 2021, then trend rises again and reaches a peak in September. However, as it can be seen from the line charts and the scatter plot, these two variables are highly correlated which means that with the increase of hospital patients, number of ICU patients increases. Note that ICU and hospital patients have different scales therefore we can just compare the trend between these two variables. Generally, number of hospital patients in each date is much more than number of ICU patients.



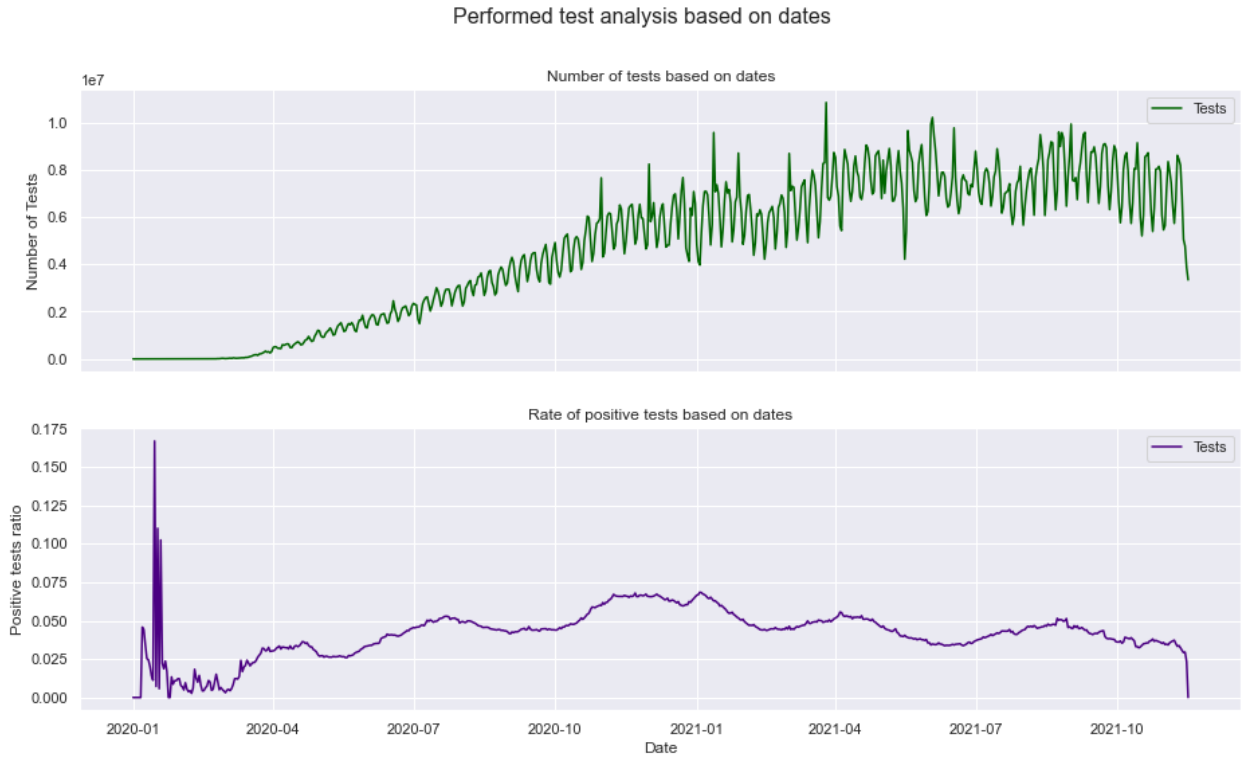
Note: due to lack of data in ICU and hospital patients only North America, Europe, and Asia are considered in this analysis.

A figure showing the number of total ICU and hospital patients in each continent in the whole dataset. As it can be understood from the bar charts North America has the greatest number of total ICU patients. Europe has the same but a little less than ICU patients compared to the North America, but it has the highest number of hospital patients between the others. Asia has the lowest number of ICU and hospital patients compared to the others. North America has the high rate of ICU per hospital patient which means that the probability of hospital patients become ICU patient is greater than other continents. Note that ICU and hospital patients have different scales.

The following plot demonstrates the correlation between number of new cases and deaths and ICU and hospital patients. Each point refers to a record with a specific date and a regression line is drawn and fit the data. As it can be seen there is a strong positive correlation between these variables which means that when number of cases increases, number of patients also increases.



3.3. Performed Tests

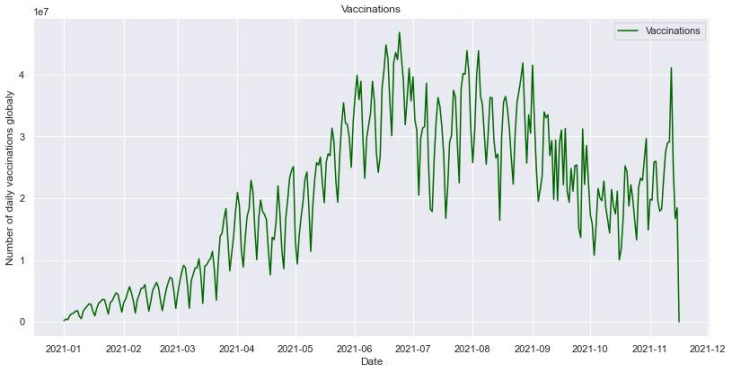


The plots above illustrate the total number of tests which performed globally and rate of positive tests in different dates. As it can be seen, total number of tests has increased during this period till reach its peak after that this number remained unchanged. Positive test ratio has the same rate during this period and there is no considerable change in this trend.

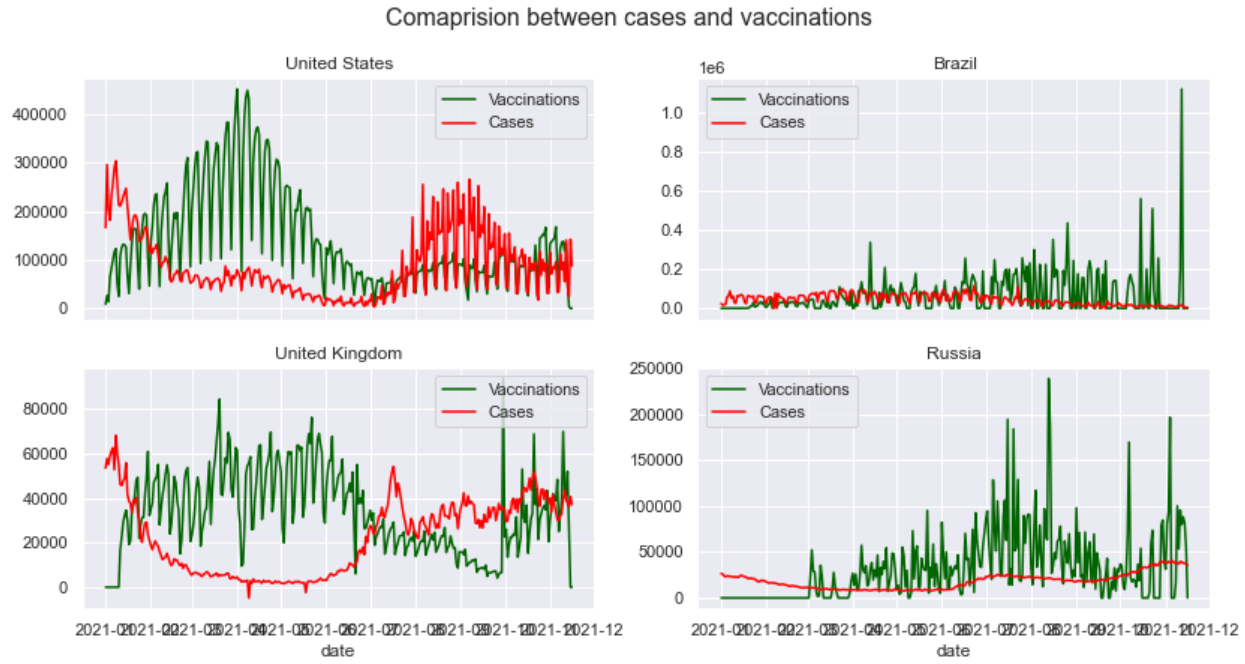
The following plots shows total number of performed test and ratio of positive tests in different continents. As it can be seen, total number of tests in Europe and Asia are much more than this number in other continents due to their population. South America has the highest rate of positive test which means that probability of getting a positive result in a COVID-19 test in this continent is more than other continents.

3.4. Vaccinations

Note: because vaccine for this virus is discovered in the begging of 2021, data from before this year is dropped for these analyses due to clarity of plots.

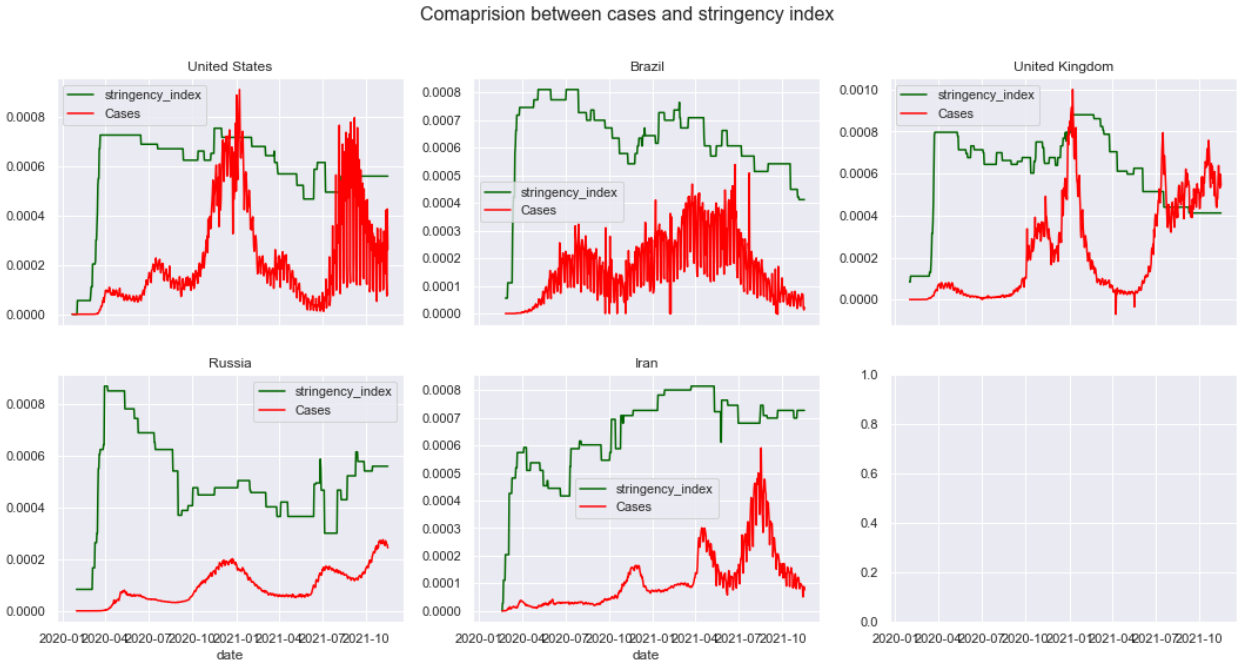


The following plot demonstrates total number of vaccinations in different date from 2021. As it can be seen, this trend increased until the middle of the 2021 and after that, decreased slightly.



The plot above demonstrates the comparison between vaccinations and new cases in worst locations (Iran excluded due to noisy and useless data). As it can be seen, these two variables are negatively correlated since with the increase of vaccinations, number of new cases in most countries decreases and vice versa.

3.4. Stringency Index



The above figure shows the correlation between stringency index and new cases in worst locations. As it can be seen, there is a negative correlation between these variables which means that with the increase of stringency index, number of new cases decreases and vice versa.

4. Results

The most important and valuables conclusions that gathered during the processes which were discussed above, are listed as follows:

- Confirmed cases and deaths are positively correlated. Europe has the highest total number of cases and deaths in this period. Number of cases and deaths depends on location for instance South American countries have the highest rate of death per case and highest rate of positive tests; therefore, South America has the worst situation in dealing with this virus among the other continents.
- ICU and hospital patients are positively correlated. Europe has the highest number of ICU and hospital patients. Number of patients is depended on location for example North American countries have the highest rate of ICU per hospital patient.
- Confirmed cases, deaths and patients are positively correlated to each other.
- Vaccination is negatively correlated to these variables which means that with increasing number of vaccinations in most countries, number of new cases, deaths, and patients decrease.
- Stringency index is negatively correlated to these variables which means that with increasing stringency in most countries, number of new cases, deaths, and patients decrease.

Note: there could be some more analyses in this report but due to lack of time some of them are mentioned in the following notebook.