Data Science

Assignment 2 – COVID19

Farimah Rashidi – 99222040

1. Import library and dataset

At the first we should import dataset and libraries. I did this part in Kaggle.

2. Data Exploration

The complete COVID-19 dataset is a collection of the COVID-19 data maintained and provided by Our World in Data.

This dataset has these 67 columns:

continent: Continent of the geographical location

location: Geographical

location date : Date of observation

total_cases : Total confirmed cases of COVID-19

new_cases : New confirmed cases of COVID-19

new_cases_smoothed : New confirmed cases of COVID-19 (7-day smoothed)

total_deaths : Total deaths attributed to COVID-19

new_deaths : New deaths attributed to COVID-19

new_deaths_smoothed : New deaths attributed to COVID-19 (7-day smoothed) total_cases_per_million : Total confirmed cases of COVID-19 per 1,000,000 people new_cases_per_million : New confirmed cases of COVID-19 per 1,000,000 people new_cases_smoothed_per_million : New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people

total_deaths_per_million : Total deaths attributed to COVID-19 per 1,000,000 people new_deaths_per_million : New deaths attributed to COVID-19 per 1,000,000 people new_deaths_smoothed_per_million : New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people

reproduction_rate : Real-time estimate of the effective reproduction rate (R) of COVID-19. See https://github.com/crondonm/TrackingR/tree/main/Estimates-Database

icu_patients : Number of COVID-19 patients in intensive care units (ICUs) on a given day icu_patients_per_million : Number of COVID-19 patients in intensive care units (ICUs) on a given day per 1,000,000 people

hosp_patients : Number of COVID-19 patients in hospital on a given day

hosp_patients_per_million : Number of COVID-19 patients in hospital on a given day per 1,000,000 people weekly_icu_admissions : Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week

weekly_icu_admissions_per_million : Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week per 1,000,000 people

weekly_hosp_admissions : Number of COVID-19 patients newly admitted to hospitals in a given week

weekly_hosp_admissions_per_million : Number of COVID-19 patients newly admitted to hospitals in a given week per 1,000,000 people

total_tests : New tests for COVID-19 (only calculated for consecutive days)

new_tests : Total tests for COVID-19

total_tests_per_thousand : Total tests for COVID-19 per 1,000 people

new_tests_per_thousand : New tests for COVID-19 per 1,000 people

new_tests_smoothed : New tests for COVID-19 (7-day smoothed). For countries that don't report testing data on a daily basis, we assume that testing changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window

new_tests_smoothed_per_thousand : New tests for COVID-19 (7-day smoothed) per 1,000 people
positive_rate : The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of

tests_per_case) tests_per_case : Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of

positive_rate) tests_units : Units used by the location to report its testing data

total_vaccinations : Total number of COVID-19 vaccination doses administered

people_vaccinated : Total number of people who received at least one vaccine dose people_fully_vaccinated : Total number of people who received all doses prescribed by the vaccination protocol

total_boosters : Total number of COVID-19 vaccination booster doses administered (doses administered beyond the number prescribed by the vaccination protocol)

new_vaccinations new_vaccinations_smoothed : New COVID-19 vaccination doses administered (only calculated for consecutive days)

total_vaccinations_per_hundred : Total number of COVID-19 vaccination doses administered per 100 people in the total population

people_vaccinated_per_hundred : Total number of people who received at least one vaccine dose per 100 people in the total population

people_fully_vaccinated_per_hundred : Total number of people who received all doses prescribed by the vaccination protocol per 100 people in the total population

total_boosters_per_hundred : Total number of COVID-19 vaccination booster doses administered per 100 people in the total population

new_vaccinations_smoothed_per_million : New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population new_people_vaccinated_smoothed : Daily number of people receiving their first vaccine dose (7-day smoothed)

new_people_vaccinated_smoothed_per_hundred : Daily number of people receiving their first vaccine dose (7-day smoothed) per 100 people in the total population

stringency_index : Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)

population_density : Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population_latest.csv for full list of sources

median_age : Number of people divided by land area, measured in square kilometers, most recent year available aged_65_older : Median age of the population, UN projection for 2020 aged_70_older : Share of the population that is 65 years and older, most recent year available

gdp_per_capita : Share of the population that is 70 years and older in 2015

extreme_poverty : Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available

cardiovasc_death_rate : Share of the population living in extreme poverty, most recent year available since 2010

diabetes_prevalence : Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)

female_smokers : Diabetes prevalence (% of population aged 20 to 79) in 2017

male_smokers : Share of women who smoke, most recent year available

handwashing_facilities : Share of men who smoke, most recent year available

hospital_beds_per_thousand : Share of the population with basic handwashing facilities on premises, most recent year available

life_expectancy : Hospital beds per 1,000 people, most recent year available since 2010 human_development_index : Life expectancy at birth in 2019 population : A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506

excess_mortality_cumulative_absolute : Percentage difference between the reported number of weekly or monthly deaths in 2020–2021 and the projected number of deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality

excess_mortality_cumulative : Percentage difference between the cumulative number of deaths since 1 January 2020 and the cumulative projected deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality

excess_mortality : Cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality excess_mortality_cumulative_per_million : Cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years, per million people. For more information, see https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality

3. EDA

Initial Data Exploration:

I started the analysis by examining the first and last 10 rows of the dataset using `df.head(10)` and `df.tail(10)`. This allowed me to get a quick overview of the structure of the data, the available columns, and some sample records.

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...	male_smokers	handwashing_facilities	hospital_beds_per_thousand	life_expect
0	AFG	Asia	Afghanistan	2020-01-03	NaN	0.0	NaN	NaN	0.0	NaN	...	NaN	37.746	0.5	6
1	AFG	Asia	Afghanistan	2020-01-04	NaN	0.0	NaN	NaN	0.0	NaN	...	NaN	37.746	0.5	6
2	AFG	Asia	Afghanistan	2020-01-05	NaN	0.0	NaN	NaN	0.0	NaN	...	NaN	37.746	0.5	6
3	AFG	Asia	Afghanistan	2020-01-06	NaN	0.0	NaN	NaN	0.0	NaN	...	NaN	37.746	0.5	6
4	AFG	Asia	Afghanistan	2020-01-07	NaN	0.0	NaN	NaN	0.0	NaN	...	NaN	37.746	0.5	6
5	AFG	Asia	Afghanistan	2020-01-08	NaN	0.0	0.0	NaN	0.0	0.0	...	NaN	37.746	0.5	6
6	AFG	Asia	Afghanistan	2020-01-09	NaN	0.0	0.0	NaN	0.0	0.0	...	NaN	37.746	0.5	6
7	AFG	Asia	Afghanistan	2020-01-10	NaN	0.0	0.0	NaN	0.0	0.0	...	NaN	37.746	0.5	6
8	AFG	Asia	Afghanistan	2020-01-11	NaN	0.0	0.0	NaN	0.0	0.0	...	NaN	37.746	0.5	6
9	AFG	Asia	Afghanistan	2020-01-12	NaN	0.0	0.0	NaN	0.0	0.0	...	NaN	37.746	0.5	6

10 rows × 67 columns

	iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	...	male_smokers	handwashing_facilities	hospital_beds_per_thousand	life_e
350075	ZWE	Africa	Zimbabwe	2023-10-09	265771.0	0.0	0.857	5718.0	0.0	0.0	...	30.7	36.791	1.7	
350076	ZWE	Africa	Zimbabwe	2023-10-10	265808.0	37.0	5.286	5718.0	0.0	0.0	...	30.7	36.791	1.7	
350077	ZWE	Africa	Zimbabwe	2023-10-11	265808.0	0.0	5.286	5718.0	0.0	0.0	...	30.7	36.791	1.7	
350078	ZWE	Africa	Zimbabwe	2023-10-12	265808.0	0.0	5.286	5718.0	0.0	0.0	...	30.7	36.791	1.7	
350079	ZWE	Africa	Zimbabwe	2023-10-13	265808.0	0.0	5.286	5718.0	0.0	0.0	...	30.7	36.791	1.7	
350080	ZWE	Africa	Zimbabwe	2023-10-14	265808.0	0.0	5.286	5718.0	0.0	0.0	...	30.7	36.791	1.7	
350081	ZWE	Africa	Zimbabwe	2023-10-15	265808.0	0.0	5.286	5718.0	0.0	0.0	...	30.7	36.791	1.7	
350082	ZWE	Africa	Zimbabwe	2023-10-16	265808.0	0.0	5.286	5718.0	0.0	0.0	...	30.7	36.791	1.7	
350083	ZWE	Africa	Zimbabwe	2023-10-17	265808.0	0.0	0.000	5718.0	0.0	0.0	...	30.7	36.791	1.7	
350084	ZWE	Africa	Zimbabwe	2023-10-18	265808.0	0.0	0.000	5718.0	0.0	0.0	...	30.7	36.791	1.7	

10 rows × 67 columns

Dataset Information:

I utilized the `df.info()` function to obtain detailed information about the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 350085 entries, 0 to 350084
Data columns (total 67 columns):
#   Column                                      Non-Null Count   Dtype
---  ------                                      --------------   -----
0   iso_code                                    350085 non-null  object
1   continent                                   333420 non-null  object
2   location                                    350085 non-null  object
3   date                                        350085 non-null  object
4   total_cases                                 312088 non-null  float64
5   new_cases                                   340457 non-null  float64
6   new_cases_smoothed                          339198 non-null  float64
7   total_deaths                                290501 non-null  float64
8   new_deaths                                  340511 non-null  float64
9   new_deaths_smoothed                         339281 non-null  float64
10  total_cases_per_million                     312088 non-null  float64
11  new_cases_per_million                       340457 non-null  float64
12  new_cases_smoothed_per_million              339198 non-null  float64
13  total_deaths_per_million                    290501 non-null  float64
```

```
14  new_deaths_per_million                     340511 non-null  float64
15  new_deaths_smoothed_per_million            339281 non-null  float64
16  reproduction_rate                          184817 non-null  float64
17  icu_patients                               37615 non-null   float64
18  icu_patients_per_million                   37615 non-null   float64
19  hosp_patients                              38902 non-null   float64
20  hosp_patients_per_million                  38902 non-null   float64
21  weekly_icu_admissions                      10205 non-null   float64
22  weekly_icu_admissions_per_million          10205 non-null   float64
23  weekly_hosp_admissions                     23253 non-null   float64
24  weekly_hosp_admissions_per_million         23253 non-null   float64
25  total_tests                                79387 non-null   float64
26  new_tests                                  75403 non-null   float64
27  total_tests_per_thousand                   79387 non-null   float64
28  new_tests_per_thousand                     75403 non-null   float64
29  new_tests_smoothed                         103965 non-null  float64
30  new_tests_smoothed_per_thousand            103965 non-null  float64
31  positive_rate                              95927 non-null   float64
32  tests_per_case                             94348 non-null   float64
33  tests_units                                106788 non-null  object
34  total_vaccinations                         79308 non-null   float64
35  people_vaccinated                          75911 non-null   float64
36  people_fully_vaccinated                    72575 non-null   float64
37  total_boosters                             47562 non-null   float64
38  new_vaccinations                           65346 non-null   float64
39  new_vaccinations_smoothed                  180718 non-null  float64
40  total_vaccinations_per_hundred             79308 non-null   float64
41  people_vaccinated_per_hundred              75911 non-null   float64
42  people_fully_vaccinated_per_hundred        72575 non-null   float64
43  total_boosters_per_hundred                 47562 non-null   float64
44  new_vaccinations_smoothed_per_million      180718 non-null  float64
45  new_people_vaccinated_smoothed             180489 non-null  float64
46  new_people_vaccinated_smoothed_per_hundred 180489 non-null  float64
47  stringency_index                           197651 non-null  float64
48  population_density                         297178 non-null  float64
49  median_age                                 276367 non-null  float64
50  aged_65_older                              266708 non-null  float64
51  aged_70_older                              273597 non-null  float64
52  gdp_per_capita                             270863 non-null  float64
53  extreme_poverty                            174561 non-null  float64
54  cardiovasc_death_rate                      271487 non-null  float64
55  diabetes_prevalence                        285303 non-null  float64
56  female_smokers                             203659 non-null  float64
57  male_smokers                               200889 non-null  float64
58  handwashing_facilities                     132973 non-null  float64
59  hospital_beds_per_thousand                 239669 non-null  float64
60  life_expectancy                            322072 non-null  float64
61  human_development_index                    263138 non-null  float64
62  population                                 350085 non-null  float64
63  excess_mortality_cumulative_absolute       12184 non-null   float64
64  excess_mortality_cumulative                12184 non-null   float64
65  excess_mortality                           12184 non-null   float64
66  excess_mortality_cumulative_per_million    12184 non-null   float64
dtypes: float64(62), object(5)
memory usage: 179.0+ MB
```

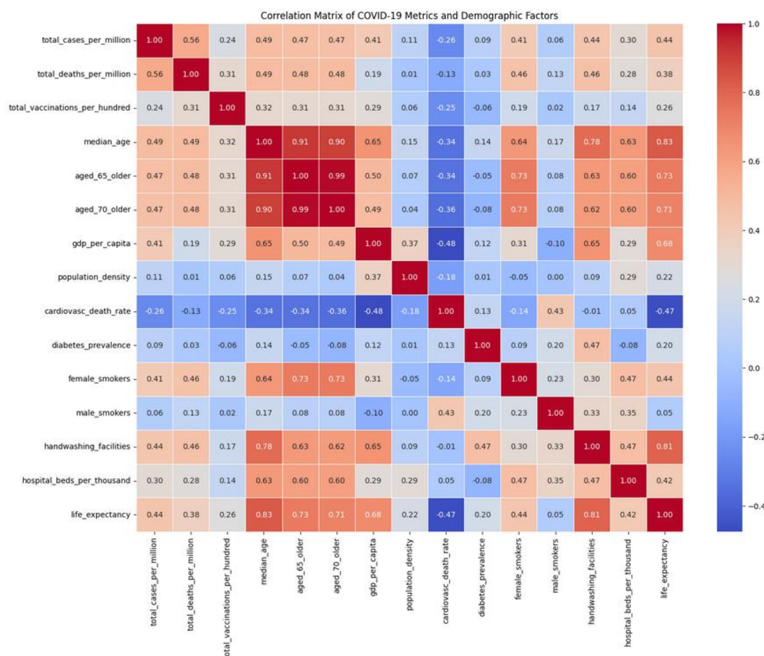
Check for missing values:

To address missing values in the dataset, I calculate the sum of missing values for each column in the dataset, providing valuable insights into areas that require attention during data preprocessing.

Correlation analysis:

For understanding the relationships between various COVID-19 metrics and demographic factors, I conducted a correlation analysis.

The heatmap provides an intuitive visualization, with warmer colors indicating stronger correlations.

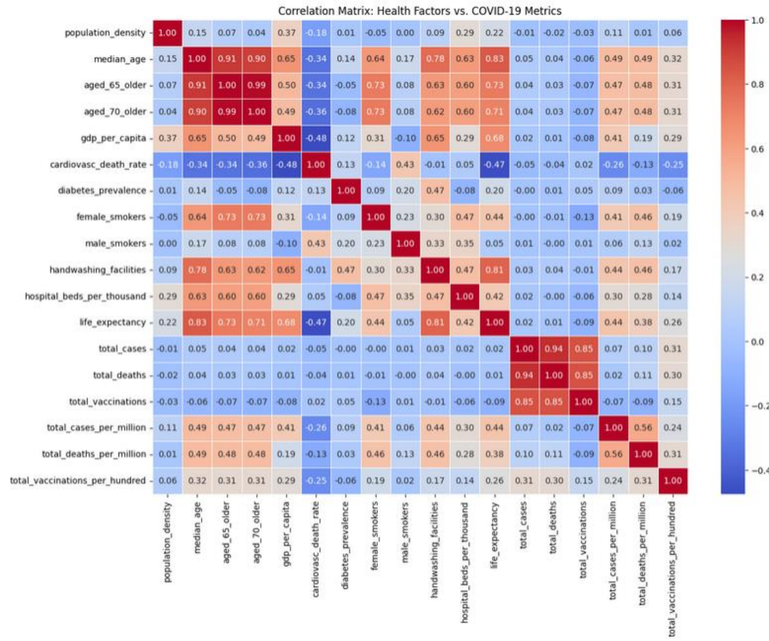


As you can see in heatmap, there is a strong correlation between life_expectancy and median_age. But there is a weak correlation between aged_70_older and cardiovasc_death_rate.

These two are just some examples of heatmap analyzing. A lot of other things can be understood from this correlation matrix.

Health Factors vs. COVID-19 Metrics:

The correlation matrix heatmap above explores the relationships between various health factors and key COVID-19 metrics.



For example, based on this heatmap, there is a high correlation between total cases and total deaths.

Smoking Rates: Positive correlations with male smokers and female smokers suggest potential connections between smoking rates and COVID-19 metrics.

Also, it seems that there is not a strong correlation between life expectancy and total deaths. After reviewing the heatmap completely, we can identify which health factors may have a significant impact on COVID-19 metrics.

Removal of Duplicate Rows:

In this part I remove all duplicate rows from dataset.

Identify numerical columns:

I identify numerical columns with selecting them.

```
Index(['total_cases', 'new_cases', 'new_cases_smoothed', 'total_deaths',
       'new_deaths', 'new_deaths_smoothed', 'total_cases_per_million',
       'new_cases_per_million', 'new_cases_smoothed_per_million',
       'total_deaths_per_million', 'new_deaths_per_million',
       'new_deaths_smoothed_per_million', 'reproduction_rate', 'icu_patients',
       'icu_patients_per_million', 'hosp_patients',
       'hosp_patients_per_million', 'weekly_icu_admissions',
       'weekly_icu_admissions_per_million', 'weekly_hosp_admissions',
       'weekly_hosp_admissions_per_million', 'total_tests', 'new_tests',
       'total_tests_per_thousand', 'new_tests_per_thousand',
       'new_tests_smoothed', 'new_tests_smoothed_per_thousand',
       'positive_rate', 'tests_per_case', 'total_vaccinations',
       'people_vaccinated', 'people_fully_vaccinated', 'total_boosters',
       'new_vaccinations', 'new_vaccinations_smoothed',
       'total_vaccinations_per_hundred', 'people_vaccinated_per_hundred',
       'people_fully_vaccinated_per_hundred', 'total_boosters_per_hundred',
       'new_vaccinations_smoothed_per_million',
       'new_people_vaccinated_smoothed',
       'new_people_vaccinated_smoothed_per_hundred', 'stringency_index',
       'population_density', 'median_age', 'aged_65_older', 'aged_70_older',
       'gdp_per_capita', 'extreme_poverty', 'cardiovasc_death_rate',
       'diabetes_prevalence', 'female_smokers', 'male_smokers',
       'handwashing_facilities', 'hospital_beds_per_thousand',
       'life_expectancy', 'human_development_index', 'population',
       'excess_mortality_cumulative_absolute', 'excess_mortality_cumulative',
       'excess_mortality', 'excess_mortality_cumulative_per_million'],
      dtype='object')
```

Conversion of 'date' Column to Datetime Format

Outlier Removal Using Z-score:

At the first I decided to remove outliers using z-score, but it seems that this decision doesn't have good effect on visualizations. So, I decided to keep outliers.

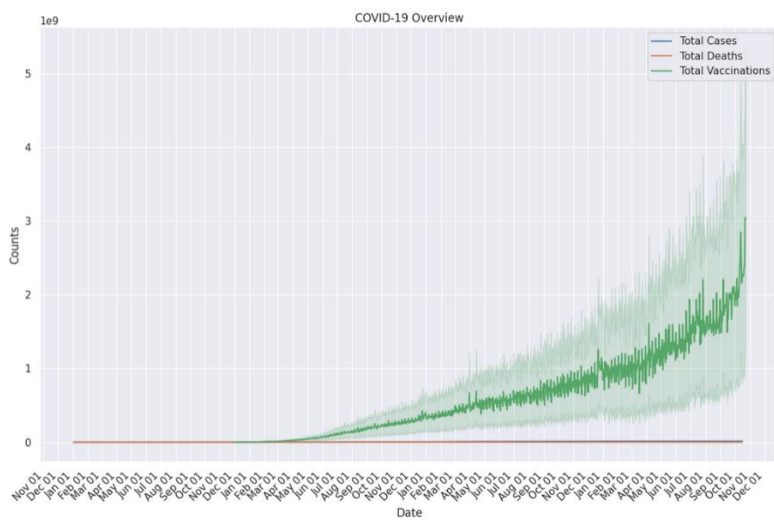
I will decide about null values after visualizations.

4. Visualization

COVID-19 Overview – 2020-2023:

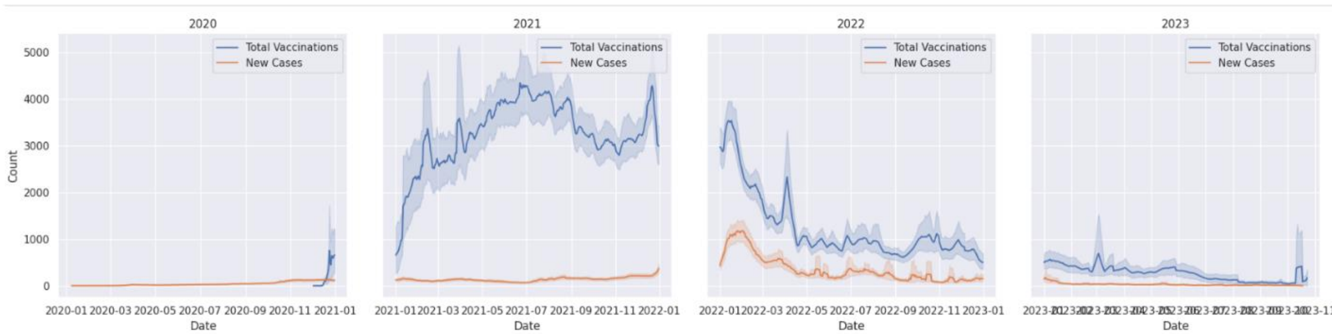
Understanding the correlation between increasing vaccination rates and potential reductions in infections and fatalities is crucial. However, does this pattern consistently hold true? To explore this, let's visualize the data and draw our conclusions.

The enhanced time series plot provides a clearer representation of the trends in COVID-19 metrics over time.



According to our data, the immunization effort started in the middle of 2021 and picked up steam in the years that followed. However, the huge discrepancy between immunization rates and reported instances makes things complicated. It is difficult to reach firm findings because of this scale disparity. Interestingly, there appears to be a notable upsurge in late 2023—especially significant considering that this is the period with the greatest vaccination rates. This unexpected finding, which contradicts common wisdom regarding the association between immunization campaigns and the occurrence of COVID-19 patients, calls for additional research.

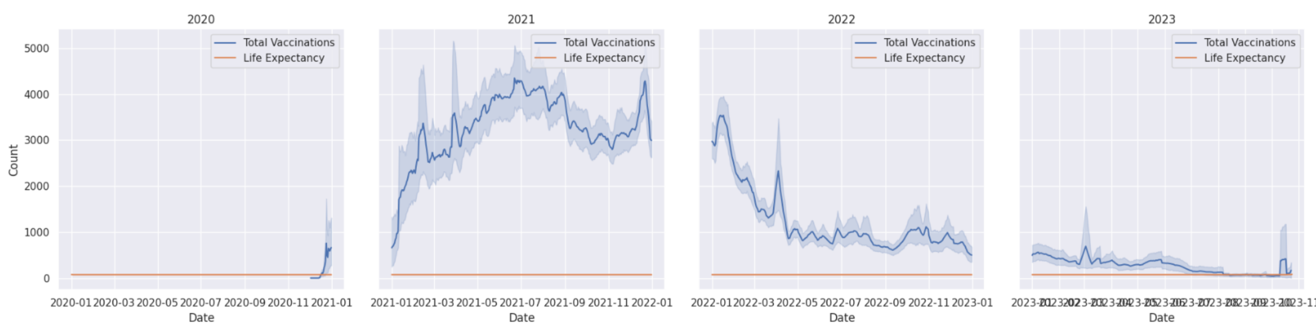
Comparison of Total Vaccinations and New Cases over Time:



The correlation between vaccines and new cases is shown in the above visual aid. Interestingly, there is a noticeable increase in instances in early 2022 that coincides with a decrease in vaccines. Furthermore, a notable pattern that shows a steady decrease in the number of new cases is noted as the years go by. Increased vaccination rates or the emergence of public immunity are two possible causes of this reduction.

Comparison of Total Vaccinations and Life Expectancy over Time:

The series of plots above provides a comparative analysis of total vaccinations and life expectancy over the years from 2020 to 2023. Each subplot represents a different year, allowing for a nuanced examination of trends.

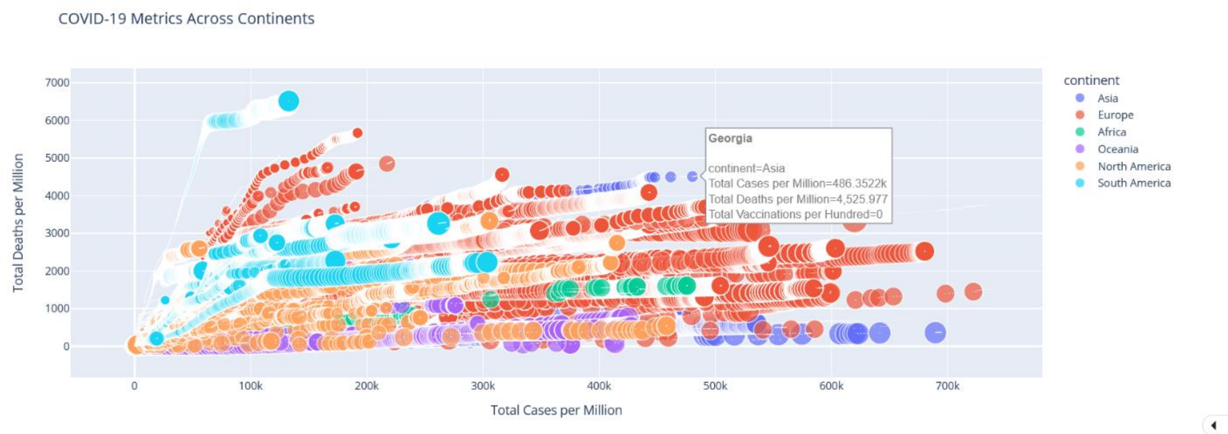


In this plot, the leftmost subplot (2020) demonstrates the dynamics between new vaccinations per million people and life expectancy throughout the year.

Surprisingly, it seems that there is not any correlation between total vaccinations and life expectancy.

COVID-19 Metrics Across Continents:

The Plotly Express visualization provides an insightful representation of COVID-19 metrics across continents, focusing on total cases and deaths per million people, as well as the vaccination rate per hundred people.

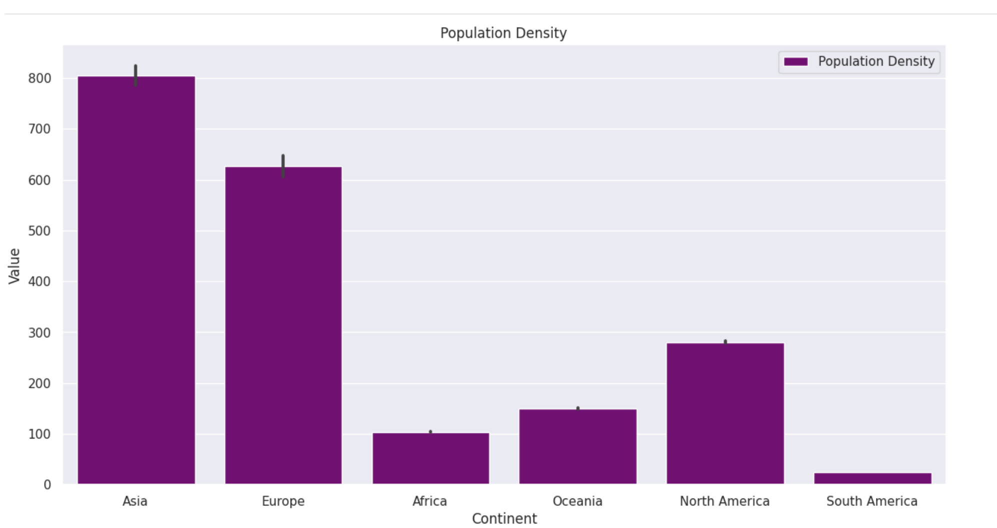


It is a dynamic plot. As you can see in photo, if you put your pointer on a special part of plot, it shows you the location, continent, total cases per million, total deaths per million and total vaccinations per hundred.

Also, we can see that total deaths in south America were quickly increasing with more speed than other continent.

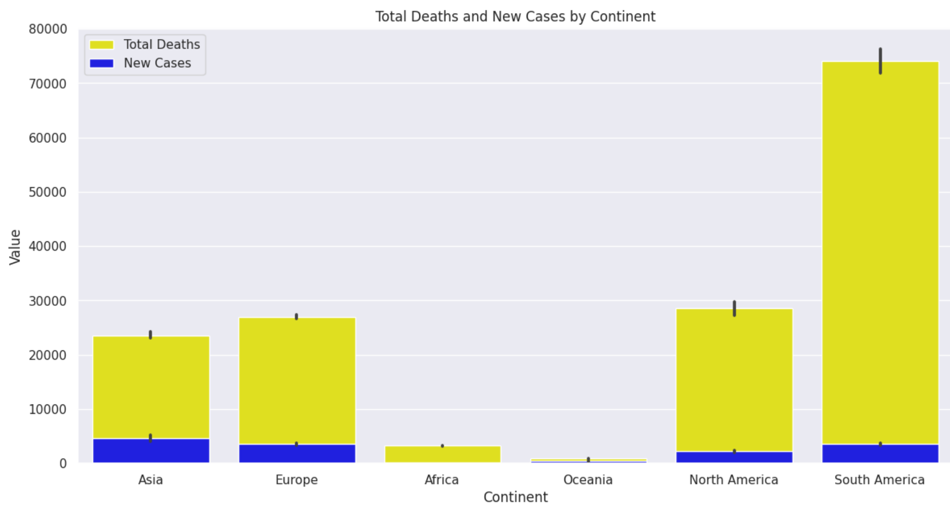
Now we want to analyze total death and cases in each continent according to population density:

Population Density:



This diagram shows the population density in each continent.

Total death and new cases:

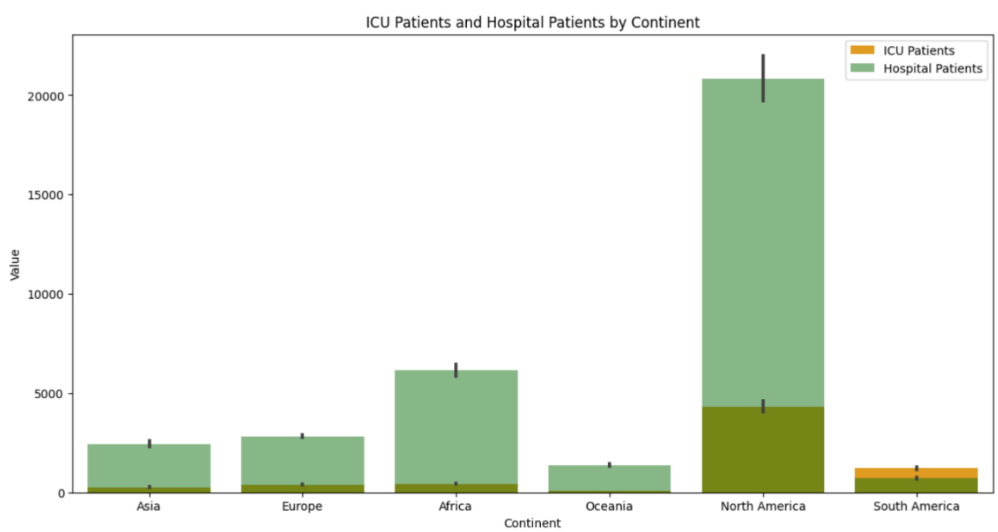


This diagram shows the number of deaths and new cases in each continent.

As you can see in these two plots, Asia is the densest continent, but its total death and total cases are fewer than south America.

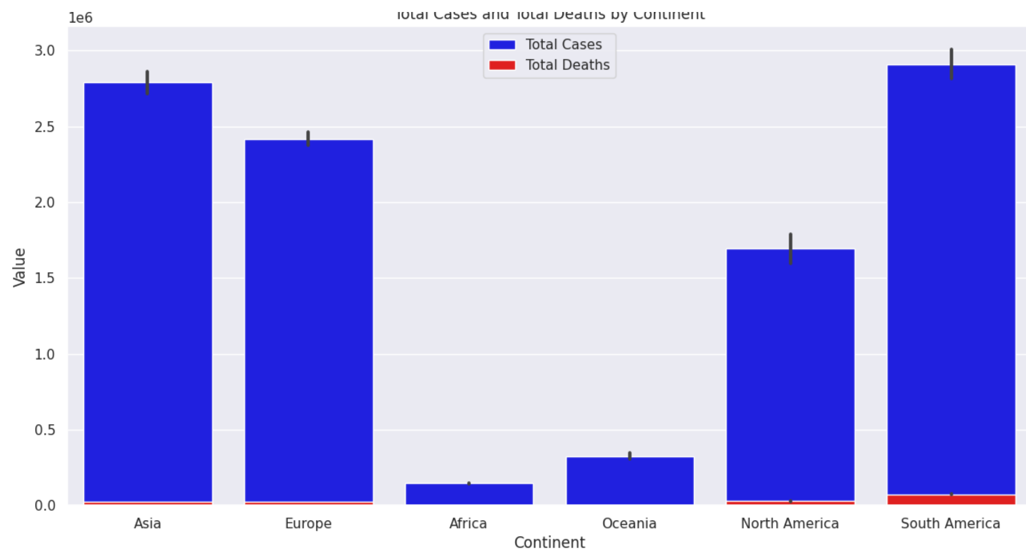
We can see that South America has a lot of deaths compared to its population density. Oceania handled this pandemic better than others according to population density.

ICU Patients and Hospital Patients by Continent:



It seems that North America has the most hospital patients. And, we can understand that even though Africa has a lot of hospital patients, but it doesn't have much ICU patients.

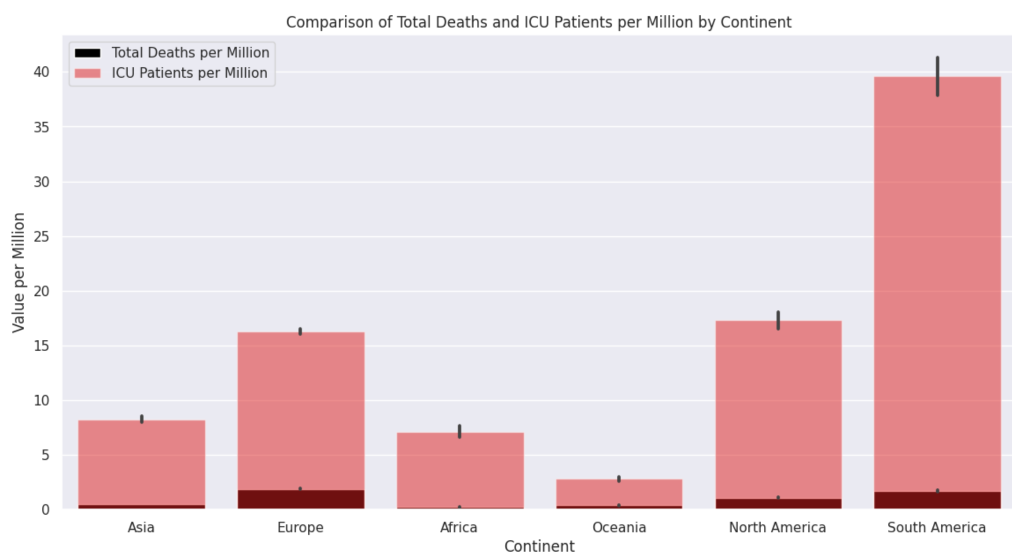
Total cases and total deaths by continent:



If we want to compare Asia and Europe together, total death of total cases in Asia is smaller than Europe.

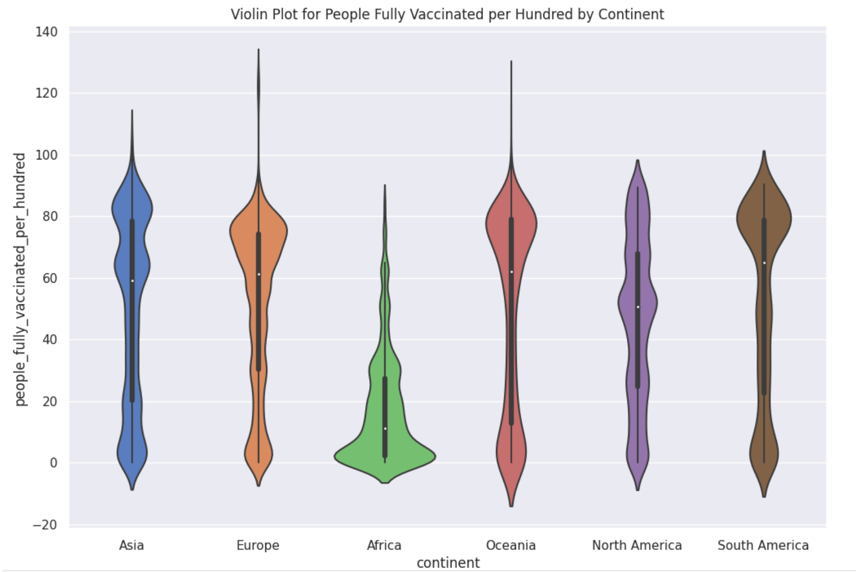
Also, we can see that Oceania was very good at handling pandemic and total death in this continent is lower compared to other continents.

Comparison of Total Deaths and ICU Patients per Million by Continent:



This diagram compares icu patients and their total deaths. North America is better at handling ICU patients and saving Covid cases in their hospital.

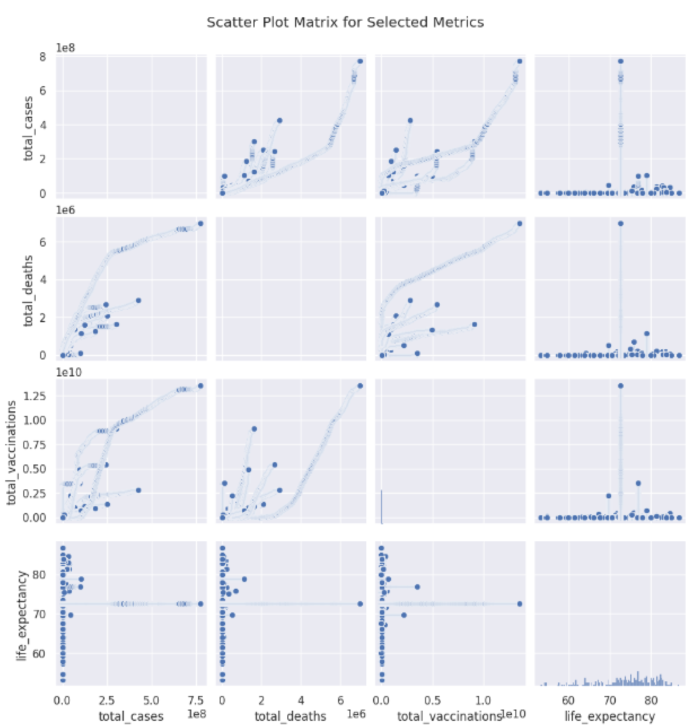
Violin Plot for People Fully Vaccinated per Hundred by Continent:



In Africa, the number of people vaccinated in each day is by far lower than other continents and the highest number of vaccinations is around 90 in 1000.

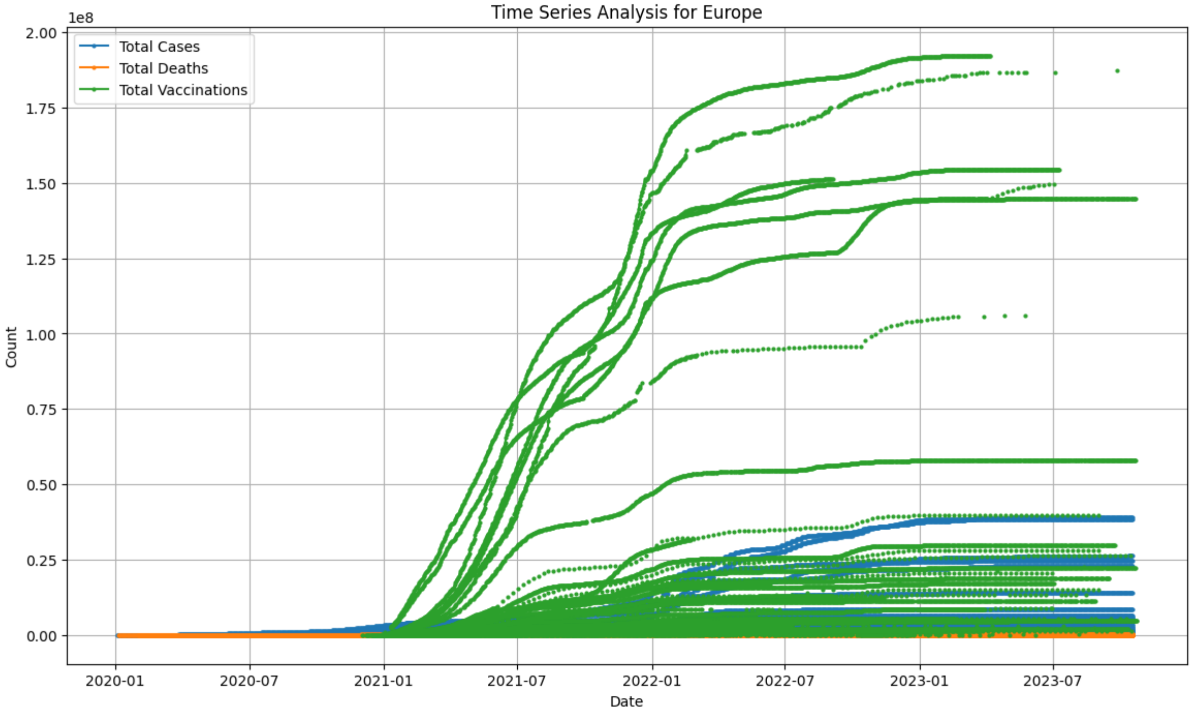
North America has an almost normal distribution in vaccination rates per day.

Scatter plot matrix for 'total_cases', 'total_deaths', 'total_vaccinations', 'life_expectancy':



Here we have a visual overview of the relationships and distributions between the selected metrics ('total_cases', 'total_deaths', 'total_vaccinations', 'life_expectancy').

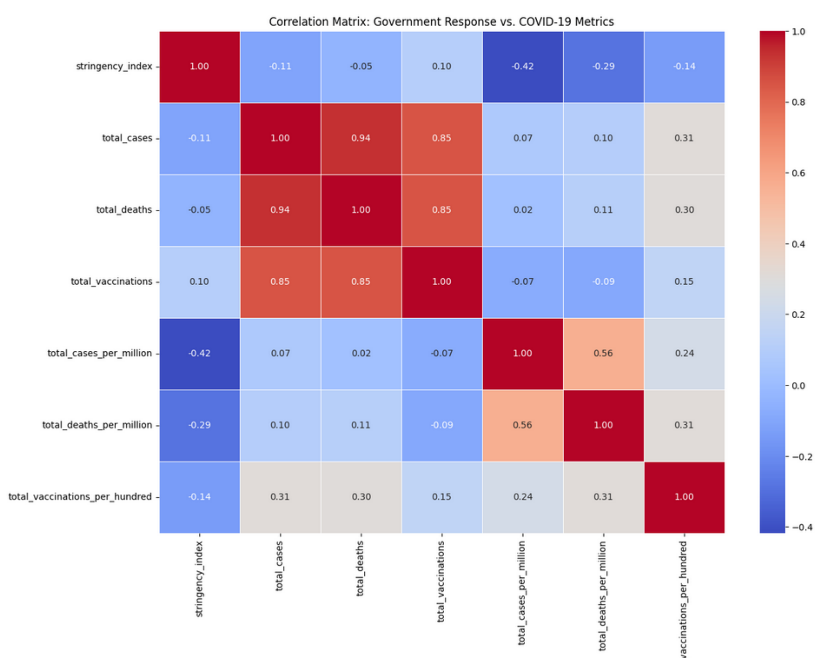
Time series analysis for Europe:



In this time series analysis for the Europe continent, we are examining the progression of COVID-19 metrics over time.

We can easily see that the number of vaccinations is increasing. Also, Total cases is quickly increasing in 2022.

Government Response vs. COVID-19 Metrics:

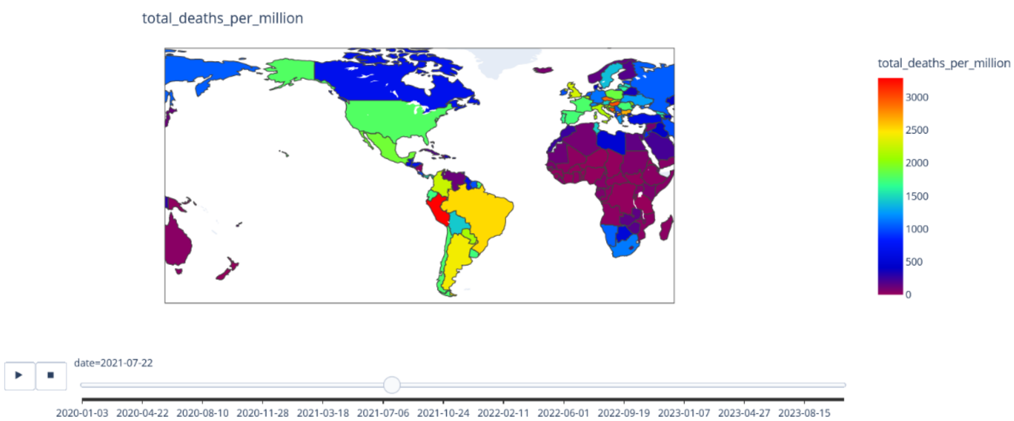


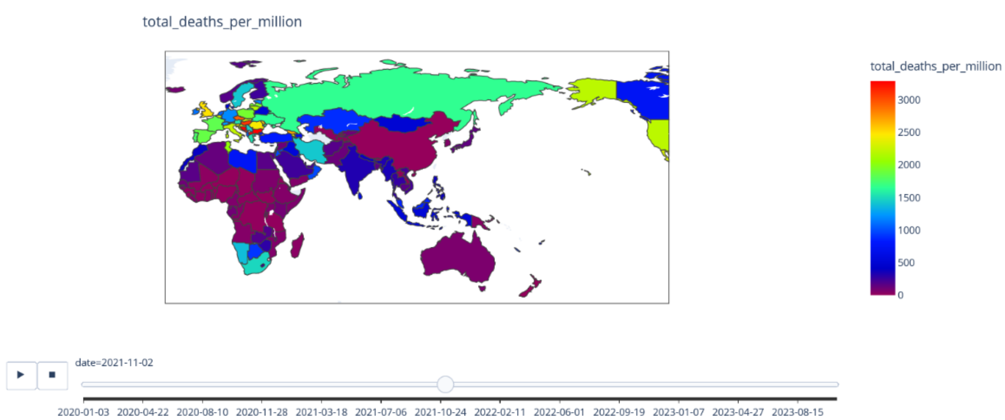
In this analysis, we are examining the correlation between government response measures and COVID-19 metrics.

You can see a strong correlation between total vaccinations and total cases. And there is a strong correlation between total vaccinations and total deaths. The correlation between other government response measures and COVID-19 metrics can be seen in heatmap.

Total Death per Million" Dynamic Visualization:

In this visualization we create an animated choropleth map using Plotly Express to visualize the 'total_deaths_per_million' values for different countries over time. It is a dynamic representation of how 'total_deaths_per_million' values evolve across countries over the specified period, offering insights into geographical patterns and trends in mortality rates. Here is just a photo of dynamic visualization for date 2023-02-15. You can check the complete format in my notebook.

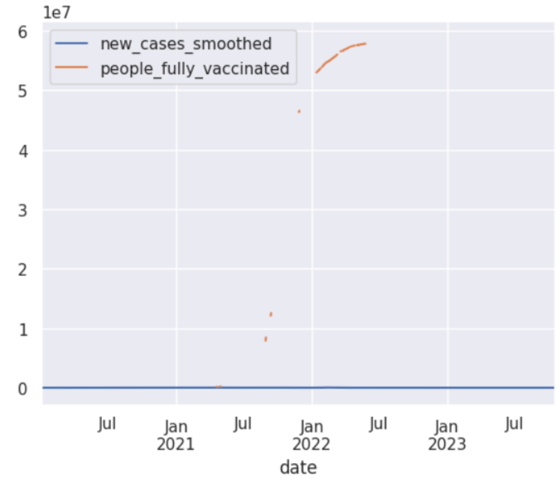


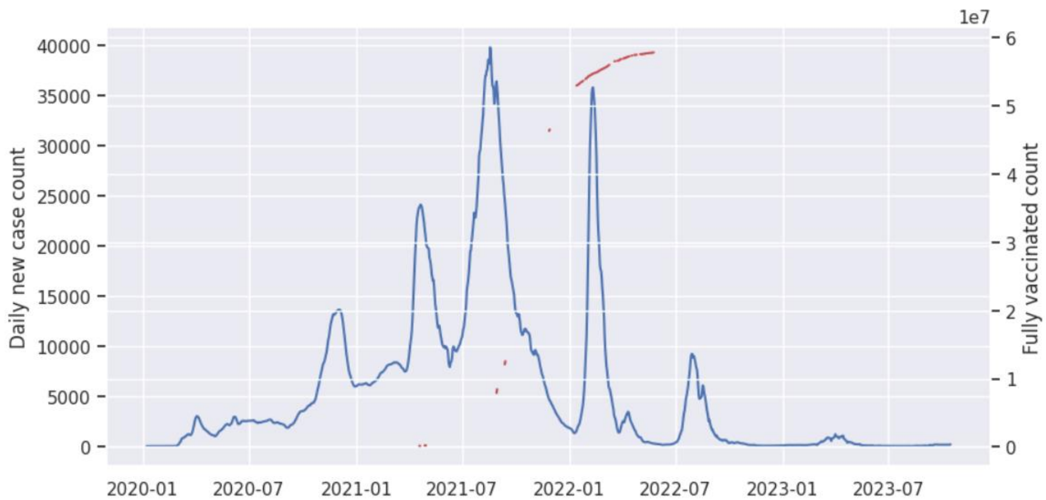


This map shows that the number of deaths began to increase especially in North America, South America , and Europe around January 2021, and grew at the same rate in these three regions. Fatalities in the U.S. , Brazil, Peru, the U.K., Italy, Hungary, and surrounding countries is more than 3 billion each in May 2023. Other than the large number of deaths, it is noteworthy that it in Asia and Africa is small. Although there are deaths in both regions, the fluctuations are smaller than those in other regions, and the total number of deaths is also smaller.

COVID-19 data for Iran:

In this analysis, we are focusing on the COVID-19 data for the country of Iran.

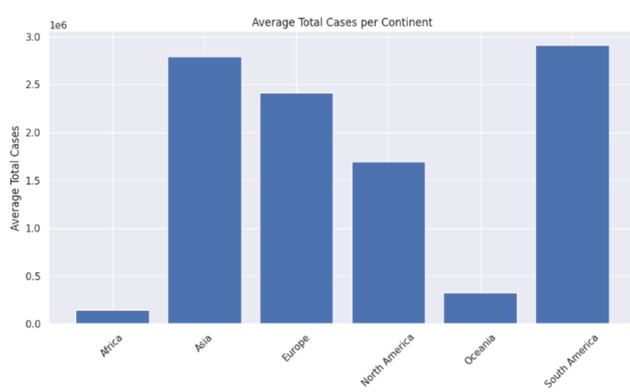


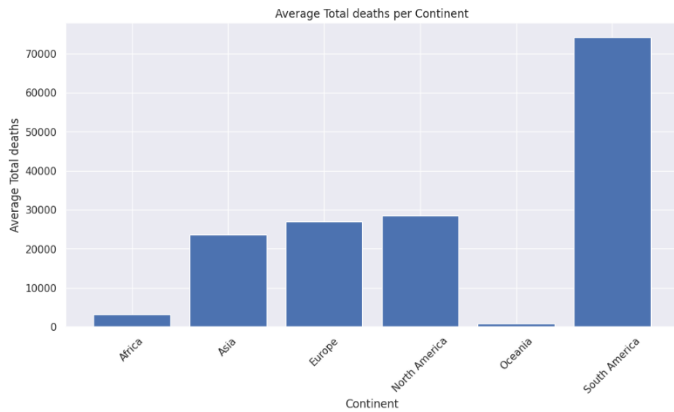


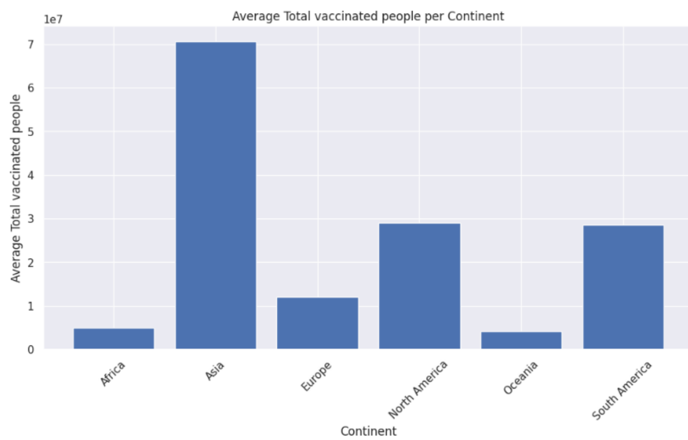
It seems that total new cases decreased when vaccinations started seriously.

New cases quickly increased in **second** half of 2021(autumn and winter of 2021).

Average total cases, total deaths and total vaccinated people per Continent:







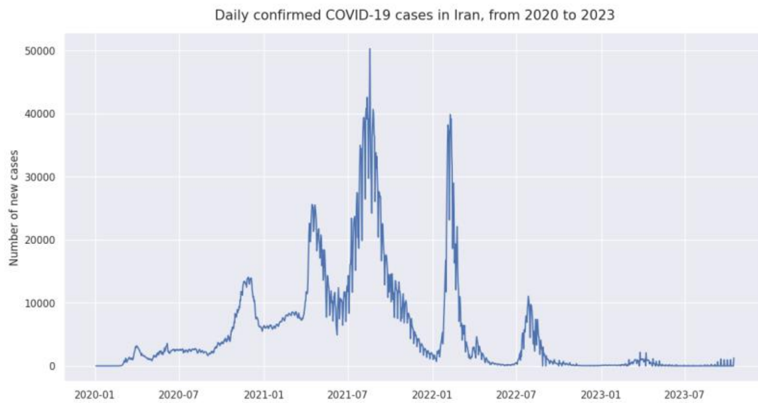
Africa and Oceania have low numbers of infected. Africa does not have accurate results since it mainly consists of poor countries. So, we can say one of the reasons for low COVID cases in Africa is poor technology. Europe, Asia and South America have the greatest number of cases by far.

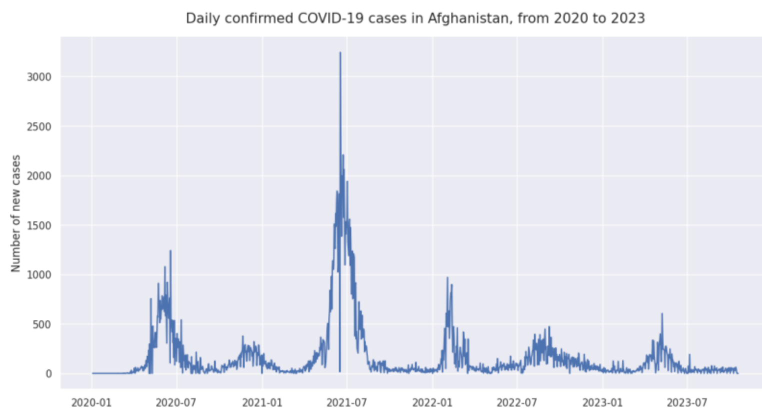
Oceania has more cases than Africa, it has a lower number of deaths.

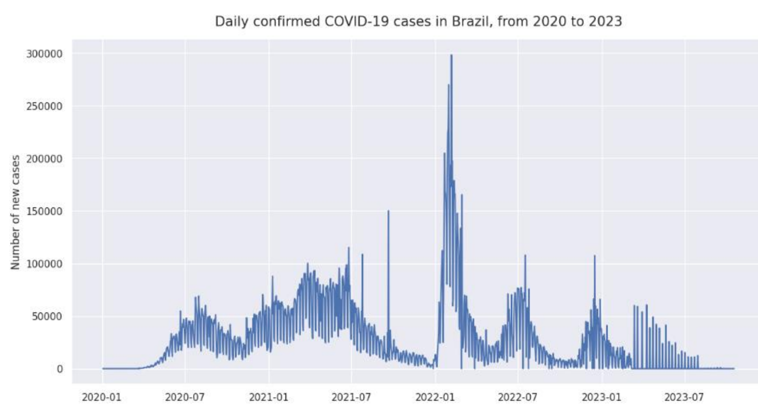
Asia has the most vaccinated people, and it has the greatest population among other continents. So, it makes sense.

Daily confirmed cases in Iran, Afghanistan and Brazil:

let’s have a look at the overall daily cases in some of the countries.





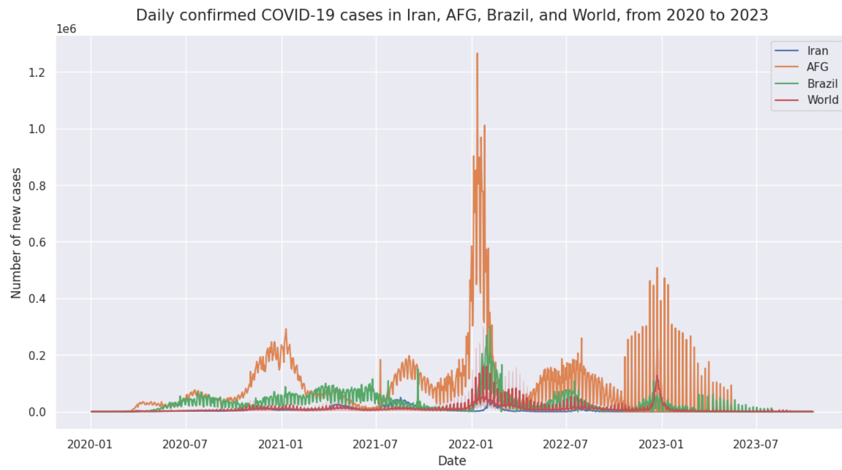


Iran has 3 major outbreaks.

And in the early 2022, all countries had outbreak.

Iran has the most total cases compared to Afghanistan and Brazil.

Number of new cases in Afghanistan suddenly increased in the middle of 2021.



5.Handling null values

Now after analyzing a lot of different complicated visualizations, it's time to decide about null values of c olumns. There are a lot of null values in numerical and categorical features. We have three choices for nu ll values:

1.Fill them according to other non-null values in that feature

2.Delete the feature if it just has a little non-null value and it is useless

3.Do not fill null values

Null values of continent column:

1) Country-to-Continent mapping:

In the first step, I crafted a mapping, connecting each country to its corresponding continent. This mapping was established by going through the rows of the DataFrame and associating countries with their known continents.

2) Filling null values in the Continent column:

Then, I defined a function, fill_continent, to address null values in the 'continent' column. This function utilizes the mapping created earlier. Through the application of this function to the DataFrame, I ensured that missing continent information was appropriately filled.

3) Manually assigning Continents for specific countries:

Recognizing the significance of certain countries, I manually assigned continents to them. This step involved specifying continent information for countries like 'Africa,' 'Asia,' 'Europe,' 'North America,' 'Oceania,' and 'South America' to ensure accurate geographical representation.

4) Excluding rows with specific country categories:

To refine the dataset and focus on individual countries, I excluded rows associated with specific country categories such as 'High income,' 'Low income,' 'Lower middle income,' 'Upper middle income,' and 'World.' This selective exclusion aimed to streamline the analysis and concentrate on individual country-level data.

Handling null values of date column:

To handle missing values in the 'total_cases' column, I grouped the data by 'country' and 'date' and filled the null values with the mean of each group. This approach ensures that missing case counts are replaced with values that reflect the typical trend for a specific country on a given date, maintaining accuracy in the analysis of total cases over time.

Handling null values of toal_deaths column:

To address missing values in the 'total_deaths' column, I employed **linear interpolation**. This method estimates the unknown values based on the known surrounding data points, providing a continuous and plausible representation of the total deaths over time. This helps maintain the integrity of the dataset and facilitates a more accurate analysis of mortality rates.

Handling null values of weekly_icu_admissions column:

This column is 97% null and the only 3% does not give us any special information. so, I decide to delete the column.

Handling null values of median_age column:

For the 'median_age' column, I first checked for missing values and found the count. Then, I addressed these missing values by imputing the median age. The median age serves as a representative central value, ensuring a more robust dataset for the analysis. This approach helps maintain the dataset's statistical characteristics while handling missing information in a pragmatic way.