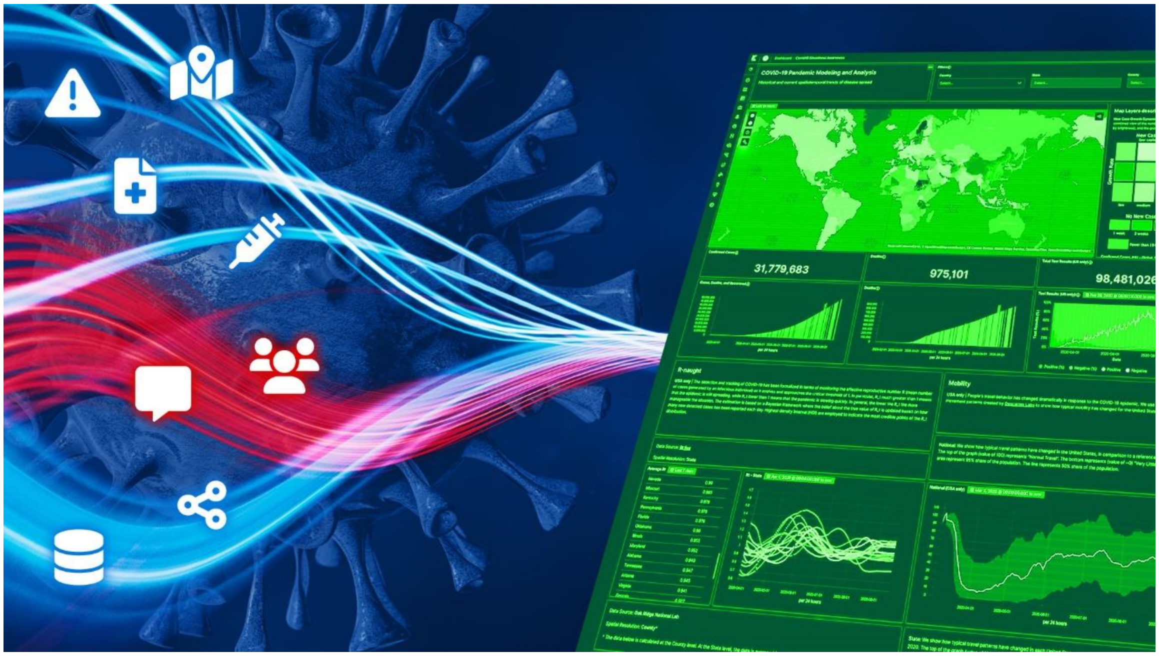


Covid-19-data

Farzam Manafzadeh | Data Science

The data and data sources

- **Confirmed cases and deaths:** this data is collected from the World Health Organization Coronavirus Dashboard. The cases & deaths dataset is updated daily.
    - ➢ Note 1: Time/date stamps reflect when the data was last updated by WHO. Due to the time required to process and validate the incoming data, there is a delay between reporting to WHO and the update of the dashboard.
    - ➢ Note 2: Counts and corrections made after these times will be carried forward to the next reporting cycle for that specific region. Delayed reporting for any specific country, territory or area may result in pooled counts for multiple days being presented, with a retrospective update to counts on previous days to accurately reflect trends. Significant data errors detected or reported to WHO may be corrected at more frequent intervals.
- **Hospitalizations and intensive care unit (ICU) admissions:** our data is collected from official sources and collated by Our World in Data. The complete list of country-by-country sources is available here.
- **Testing for COVID-19:** this data is collected by the *Our World in Data* team from official reports; you can find further details in our post on COVID-19 testing, including our checklist of questions to understand testing data, information on geographical and temporal coverage, and detailed country-by-country source information. **On 23 June 2022, we stopped adding new datapoints to our COVID-19 testing dataset.** You can read more here.
- **Vaccinations against COVID-19:** this data is collected by the *Our World in Data* team from official reports.
- **Other variables:** this data is collected from a variety of sources (United Nations, World Bank, Global Burden of Disease, Blavatnik School of Government, etc.). More information is available in our codebook.

THE DATASET HAS THE FOLLOWING ATTRIBUTES:

1. **iso_code:**
    - **Description:** ISO 3166-1 alpha-3 – three-letter country codes. Note that OWID-defined regions (e.g. continents like 'Europe') contain prefix 'OWID_'.
2. **continent:**
    - **Description:** Continent of the geographical location.
3. **location:**
    - **Description:** Geographical location.
4. **date:**
    - **Description:** Date of observation.
5. **total_cases:**
    - **Description:** Total confirmed cases of COVID-19. Counts can include probable cases, where reported.
6. **new_cases:**
    - **Description:** New confirmed cases of COVID-19. Counts can include probable cases, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA.
7. **new_cases_smoothed:**
    - **Description:** New confirmed cases of COVID-19 (7-day smoothed). Counts can include probable cases, where reported.
8. **total_deaths:**
    - **Description:** Total deaths attributed to COVID-19. Counts can include probable deaths, where reported.
9. **new_deaths:**
    - **Description:** New deaths attributed to COVID-19. Counts can include probable deaths, where reported. In rare cases where our source reports a negative daily change due to a data correction, we set this metric to NA.
10. **new_deaths_smoothed:**
    - **Description:** New deaths attributed to COVID-19 (7-day smoothed). Counts can include probable deaths, where reported.

11. **total_cases_per_million:**

    - **Description:** Total confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported.

12. **new_cases_per_million:**

    - **Description:** New confirmed cases of COVID-19 per 1,000,000 people. Counts can include probable cases, where reported.

13. **new_cases_smoothed_per_million:**

    - **Description:** New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable cases, where reported.

14. **total_deaths_per_million:**

    - **Description:** Total deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported.

15. **new_deaths_per_million:**

    - **Description:** New deaths attributed to COVID-19 per 1,000,000 people. Counts can include probable deaths, where reported.

16. **new_deaths_smoothed_per_million:**

    - **Description:** New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people. Counts can include probable deaths, where reported.

17. **reproduction_rate:**

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

18. **icu_patients:**

    - **Description:** Number of COVID-19 patients in intensive care units (ICUs) on a given day.

19. **icu_patients_per_million:**

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

20. **hosp_patients:**

    - **Description:** Number of COVID-19 patients in hospital on a given day.

21. **hosp_patients_per_million:**

    - **Description:** Number of COVID-19 patients in hospital on a given day per 1,000,000 people.

22. **weekly_icu_admissions:**

    - **Description:** Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week (reporting date and the preceding 6 days).

23. **weekly_icu_admissions_per_million:**

    - **Description:** Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week per 1,000,000 people (reporting date and the preceding 6 days).

24. **weekly_hosp_admissions:**

    - **Description:** Number of COVID-19 patients newly admitted to hospitals in a given week (reporting date and the preceding 6 days).

25. **weekly_hosp_admissions_per_million:**

    - **Description:** Number of COVID-19 patients newly admitted to hospitals in a given week per 1,000,000 people (reporting date and the preceding 6 days).

26. **total_tests:**

    - **Description:** Total tests for COVID-19.

27. **new_tests:**

    - **Description:** New tests for COVID-19 (only calculated for consecutive days).

28. **total_tests_per_thousand:**

    - **Description:** Total tests for COVID-19 per 1,000 people.

29. **new_tests_per_thousand:**

    - **Description:** New tests for COVID-19 per 1,000 people.

30. **new_tests_smoothed:**

    - **Description:** 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.

31. **new_tests_smoothed_per_thousand:**

    - **Description:** New tests for COVID-19 (7-day smoothed) per 1,000 people.

32. **positive_rate:**

    - **Description:** The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests_per_case).

33. **tests_per_case:**

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

34. **tests_units:**

    - **Description:** Units used by the location to report its testing data. A country file can't contain mixed units. All metrics concerning testing data use the specified test unit. Valid units are 'people tested' (number of people tested), 'tests performed' (number of tests performed. a single person can be tested more than once in a given day) and 'samples tested' (number of samples tested. In some cases, more than one sample may be required to perform a given test.)

35. **total_vaccinations:**

    - **Description:** Total number of COVID-19 vaccination doses administered.

36. **people_vaccinated:**

    - **Description:** Total number of people who received at least one vaccine dose.

37. **people_fully_vaccinated:**

    - **Description:** Total number of people who received all doses prescribed by the initial vaccination protocol.

38. **total_boosters:**

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

39. **new_vaccinations:**

    - **Description:** New COVID-19 vaccination doses administered (only calculated for consecutive days).

40. **new_vaccinations_smoothed:**

    - **Description:** New COVID-19 vaccination doses administered (7-day smoothed). For countries that don't report vaccination data on a daily basis, we assume that vaccination 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.

41. **total_vaccinations_per_hundred:**

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

42. **people_vaccinated_per_hundred:**

    - **Description:** Total number of people who received at least one vaccine dose per 100 people in the total population.

43. **people_fully_vaccinated_per_hundred:**

    - **Description:** Total number of people who received all doses prescribed by the initial vaccination protocol per 100 people in the total population.

44. **total_boosters_per_hundred:**

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

45. **new_vaccinations_smoothed_per_million:**

    - **Description:** New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population.

46. **new_people_vaccinated_smoothed:**

    - **Description:** Daily number of people receiving their first vaccine dose (7-day smoothed).

47. **new_people_vaccinated_smoothed_per_hundred:**

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

48. **stringency_index:**

    - **Description:** 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).

49. **population:**

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

50. **population_density:**

    - **Description:** Number of people divided by land area, measured in square kilometers, most recent year available.

51. **median_age:**

    - **Description:** Median age of the population, UN projection for 2020.

52. **aged_65_older:**

    - **Description:** Share of the population that is 65 years and older, most recent year available.

53. **aged_70_older:**

    - **Description:** Share of the population that is 70 years and older in 2015.

54. **gdp_per_capita:**

    - **Description:** Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available.

55. **extreme_poverty:**

    - **Description:** Share of the population living in extreme poverty, most recent year available since 2010.

56. **cardiovasc_death_rate:**

    - **Description:** Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people).

57. **diabetes_prevalence:**

    - **Description:** Diabetes prevalence (% of population aged 20 to 79) in 2017.

58. **female_smokers:**

    - **Description:** Share of women who smoke, most recent year available.

59. **male_smokers:**

    - **Description:** Share of men who smoke, most recent year available.

60. **handwashing_facilities:**

    - **Description:** Share of the population with basic handwashing facilities on premises, most recent year available.

61. **hospital_beds_per_thousand:**

    - **Description:** Hospital beds per 1,000 people, most recent year available since 2010.

62. **life_expectancy:**

    - **Description:** Life expectancy at birth in 2019.

63. **human_development_index:**

    - **Description:** 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.

64. **excess_mortality:**

- **Description:** 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.

65. **excess_mortality_cumulative:**

- **Description:** 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.

66. **excess_mortality_cumulative_absolute:**

- **Description:** 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.

67. **excess_mortality_cumulative_per_million:**

- **Description:** 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.

EDA

Data Exploration and Cleaning Report

Initial Data Inspection:

- The dataset contains 355,434 entries and 66 columns.
- The 'iso_code' column was dropped as it was deemed unnecessary.
- the latest date in the dataset: 2023-11-09
- North Korea's total cases on the last day: nan

It seems that the value for North Korea's total cases on the last day is NaN (Not a Number), which typically indicates missing or unavailable data. In many cases, countries may not provide accurate or up-to-date information.

investigating why the data is missing. It could be due to North Korea not reporting COVID-19 cases, lack of data availability, or reporting delays those countries with closed or restricted information policies, may be limited.

- Filter the data for countries with NaN total cases on the last date

```
   ['North Korea' 'Turkmenistan']
```

- droped rows for 'North Korea' and 'Turkmenistan'

Missing Values Analysis:

- Checked for missing values in each column
- Several columns have missing values, ranging from 4.75% to 96.56%. Columns like 'reproduction_rate,' 'icu_patients,' and 'stringency_index' have high percentages of missing values.

```
continent                                     16884
location                                          0
date                                              0
total_cases                                   35164
new_cases                                      9579
new_cases_smoothed                            10828
total_deaths                                  56883
new_deaths                                     9527
new_deaths_smoothed                           10747
total_cases_per_million                       35164
new_cases_per_million                          9579
new_cases_smoothed_per_million                10828
total_deaths_per_million                      56883
```

```
new_deaths_per_million                          9527
new_deaths_smoothed_per_million                10747
reproduction_rate                             167803
icu_patients                                  314889
icu_patients_per_million                      314889
hosp_patients                                 313542
hosp_patients_per_million                     313542
weekly_icu_admissions                         342363
weekly_icu_admissions_per_million             342363
weekly_hosp_admissions                        329238
weekly_hosp_admissions_per_million            329238
total_tests                                   273297
new_tests                                     277219
total_tests_per_thousand                      273297
new_tests_per_thousand                        277219
new_tests_smoothed                            249377
new_tests_smoothed_per_thousand               249377
positive_rate                                 256693
tests_per_case                                258272
tests_units                                   246561
total_vaccinations                            272939
people_vaccinated                             276356
people_fully_vaccinated                       279677
total_boosters                                304707
new_vaccinations                              286936
new_vaccinations_smoothed                     170926
total_vaccinations_per_hundred                272939
people_vaccinated_per_hundred                 276356
people_fully_vaccinated_per_hundred           279677
total_boosters_per_hundred                    304707
new_vaccinations_smoothed_per_million         170926
new_people_vaccinated_smoothed                171158
new_people_vaccinated_smoothed_per_hundred    171158
stringency_index                              156063
population_density                             53635
median_age                                     74798
aged_65_older                                  84589
aged_70_older                                  77612
gdp_per_capita                                 78961
extreme_poverty                               175334
cardiovasc_death_rate                          79717
diabetes_prevalence                            65686
female_smokers                                145787
male_smokers                                  148601
handwashing_facilities                        218953
hospital_beds_per_thousand                    112019
life_expectancy                                28367
human_development_index                        86791
population                                         0
excess_mortality_cumulative_absolute          340409
excess_mortality_cumulative                   340409
excess_mortality                              340409
excess_mortality_cumulative_per_million       340409
```

Handling Missing Data:

Continent

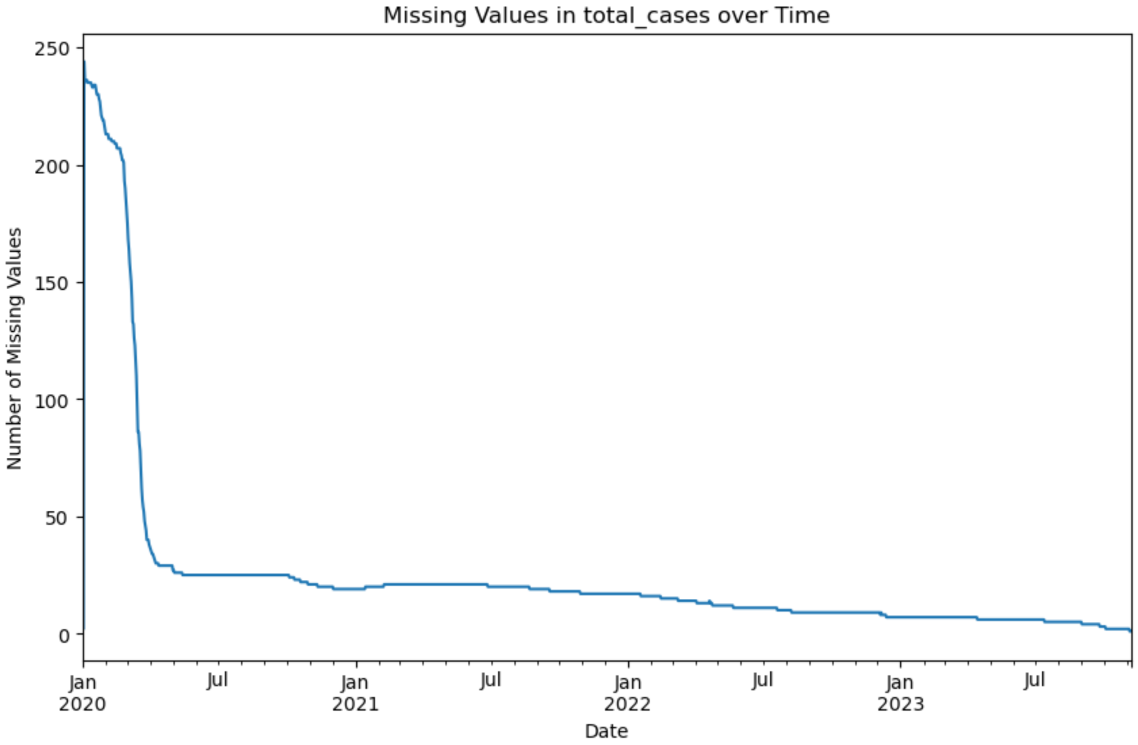
- Find unique names of countries with missing values in 'continent'
- ['Africa' 'Asia' 'Europe' 'European Union' 'High income' 'Low income'
- 'Lower middle income' 'North America' 'Oceania' 'South America'
- 'Upper middle income' 'World']

using the loc accessor to locate the rows where the 'location' column matches the specified countries and fills the 'continent' column with the value "iso_continent" for those rows.

After running the code, checking the specific rows for the countries filled to ensure that the 'continent' column has been updated as expected.

```
continent                                               0.000000

location                                                0.000000
date                                                    0.000000
```

total_case



Created a pivot table to count the occurrences of missing 'total_cases' for each combination of 'location' and 'date'

```
0                           Afghanistan 2020-01-03             1
1                           Afghanistan 2020-01-04             1
2                           Afghanistan 2020-01-05             1
3                           Afghanistan 2020-01-06             1
4                           Afghanistan 2020-01-07             1
5                           Afghanistan 2020-01-08             1
6                           Afghanistan 2020-01-09             1
7                           Afghanistan 2020-01-10             1
8                           Afghanistan 2020-01-11             1
9                           Afghanistan 2020-01-12             1
```

Now Replaced missing values in 'total_cases' with zeros.

```
Number of missing values in 'total_cases' after replacement: 0
```

more locations to drop

Create a pivot table to count the occurrences of missing 'total_cases' for each combination of 'location' and 'date'

it seems like wales have missing values.

```
Total cases in Wales on the last reported day: 0.0
```

now I found some more countries that have bad report and i want to find those counries that have 0 total_case on their last day of data of that specific location has been report

```
Countries with 0 total cases on their last reported day: ['England'
'Hong Kong' 'Macao' 'Northern Cyprus' 'Northern Ireland'
 'Scotland' 'Taiwan' 'Wales' 'Western Sahara']

Number of missing values in 'new_cases' after replacement: 0
Number of missing values in 'new_cases_smoothed'after replacement: 0
```

new_deaths

Created a pivot table to count the occurrences of missing 'total_cases' for each combination of 'location' and 'date'

it seems like USA have that condition.

```
New deaths in the United States on 2022-03-10: 0.0
```

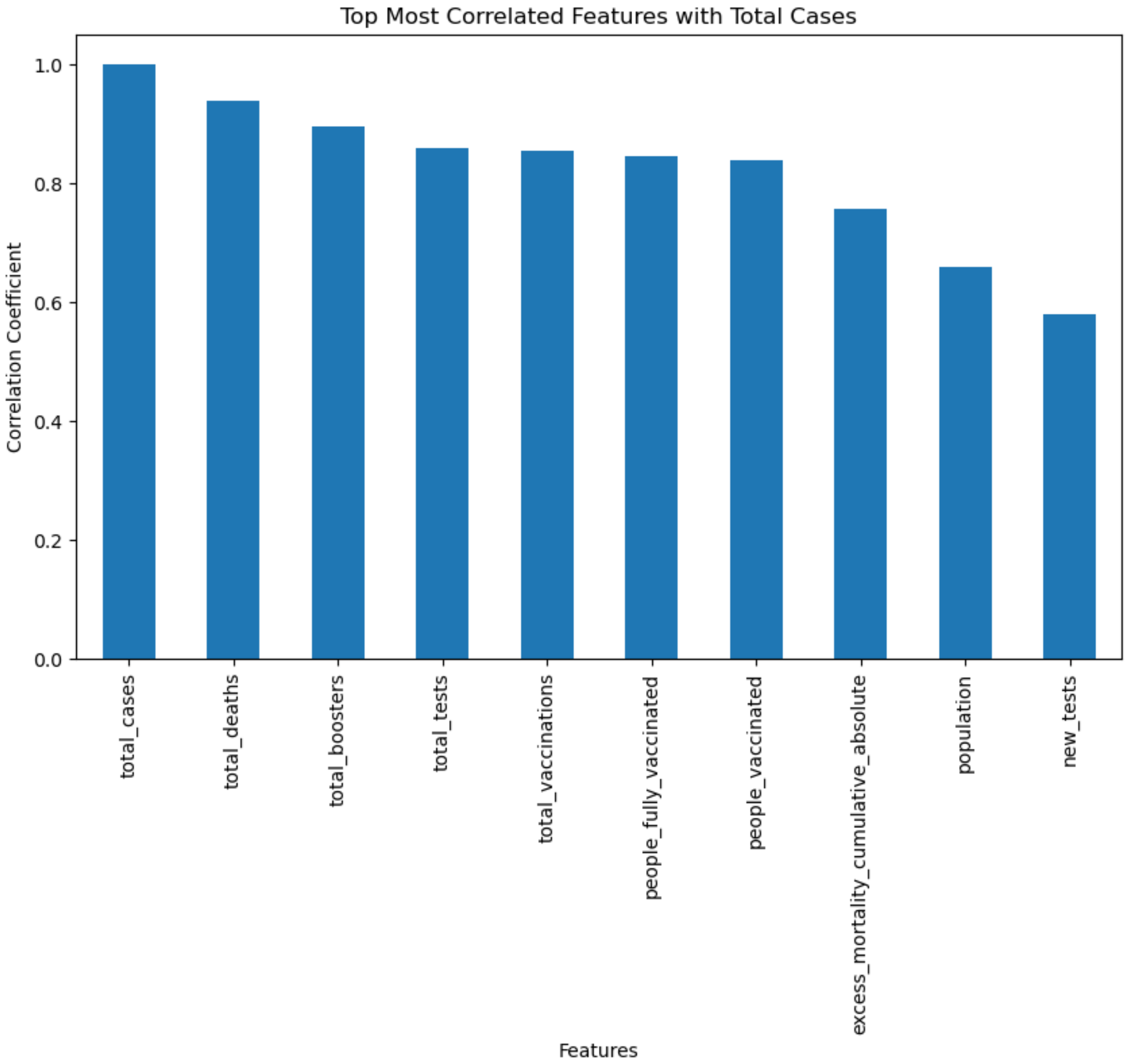
Filled missing values of total_deaths and new_deaths with 0

Now I have these features with zero missing values:

```
continent                                    0.000000
location                                     0.000000
date                                         0.000000
total_cases                                  0.000000
new_cases                                    0.000000
new_cases_smoothed                           0.000000
total_deaths                                 0.000000
new_deaths                                   0.000000
new_deaths_smoothed                          0.000000
```

- Imputed missing values or handled them appropriately based on the nature of the data.

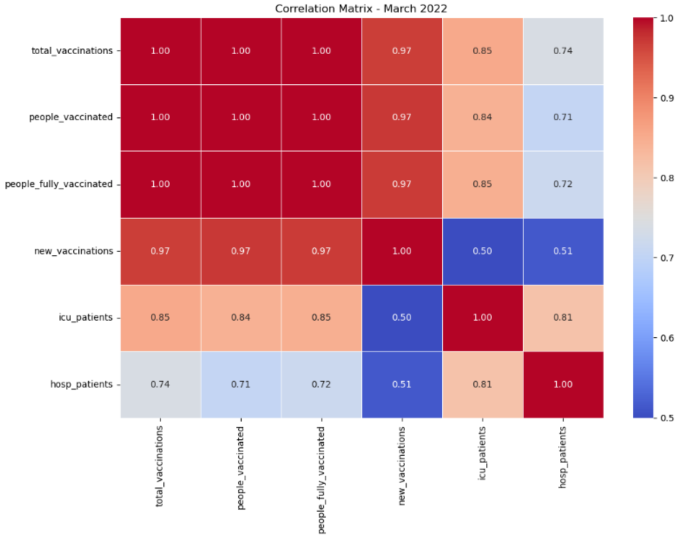
Correlation

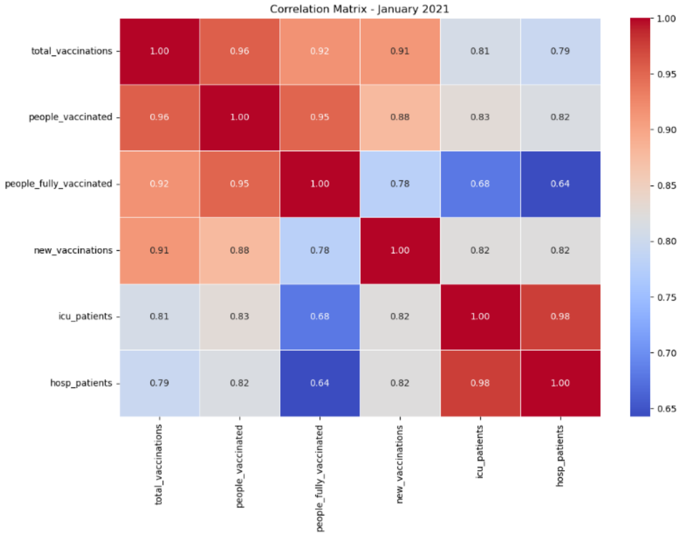


Selecting the correlation values for 'total_cases' and sorting them

```
total_cases                             1.000000
total_deaths                            0.938820
total_boosters                          0.895218
total_tests                             0.859420
total_vaccinations                      0.854184
people_fully_vaccinated                 0.844591
people_vaccinated                       0.838974
excess_mortality_cumulative_absolute    0.755098
population                              0.658435
new_tests                               0.578069
```

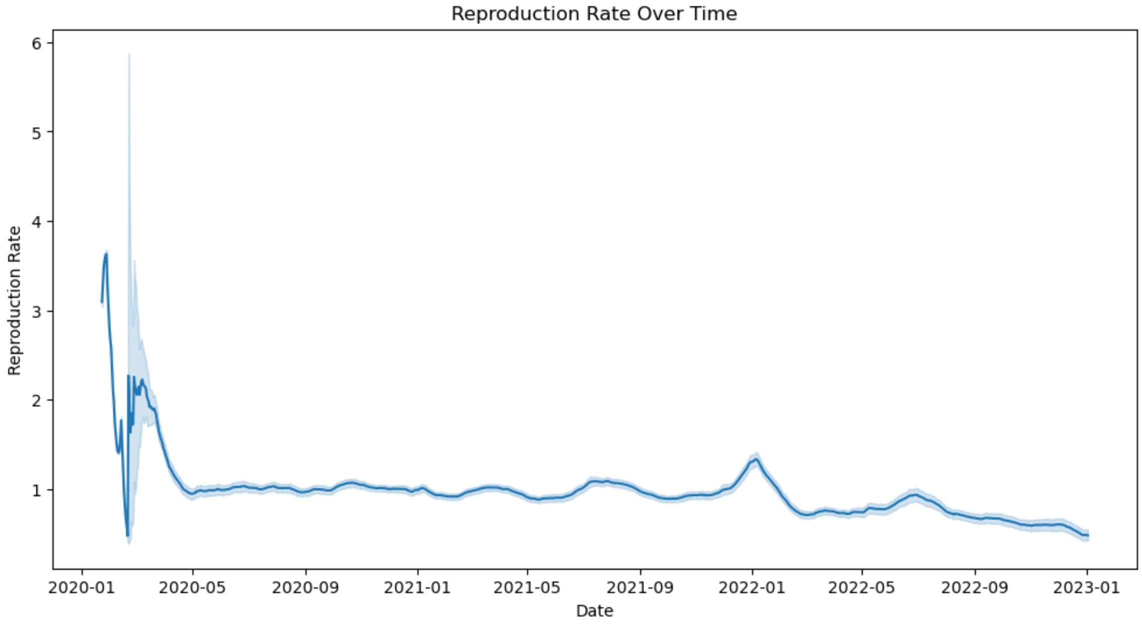
These values represent the strength and direction of the linear relationship between 'total_cases' and each of the listed variables.



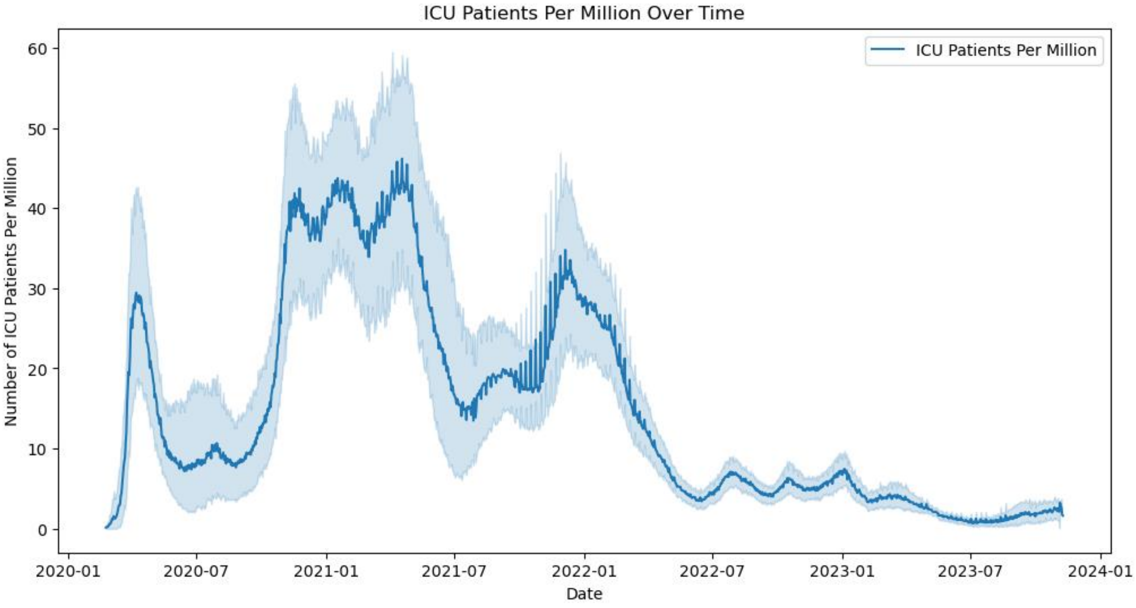


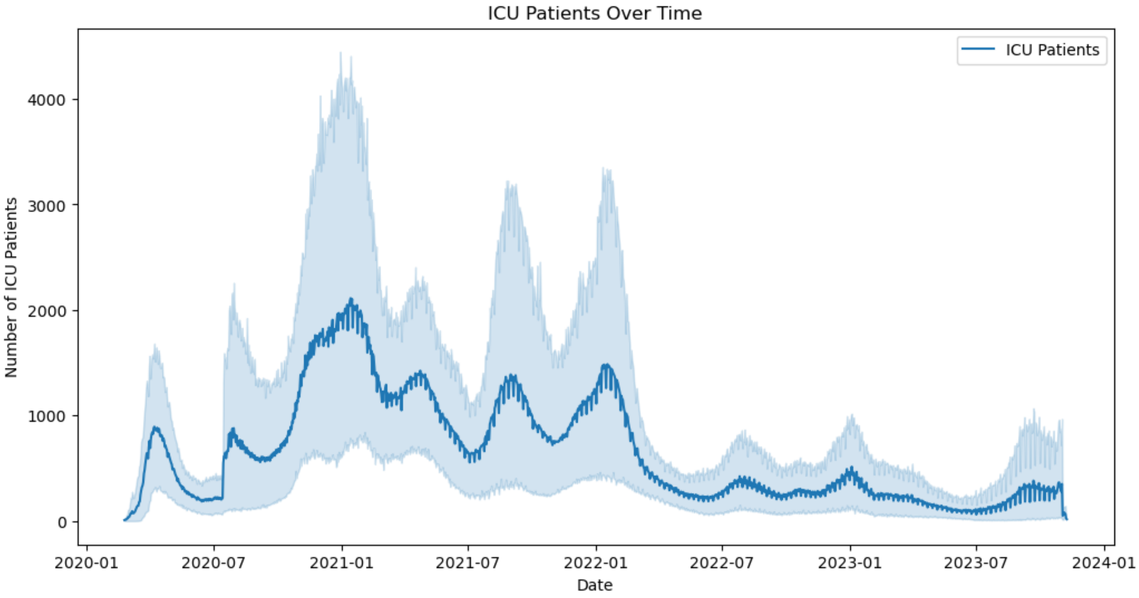
Reproduction Rate:

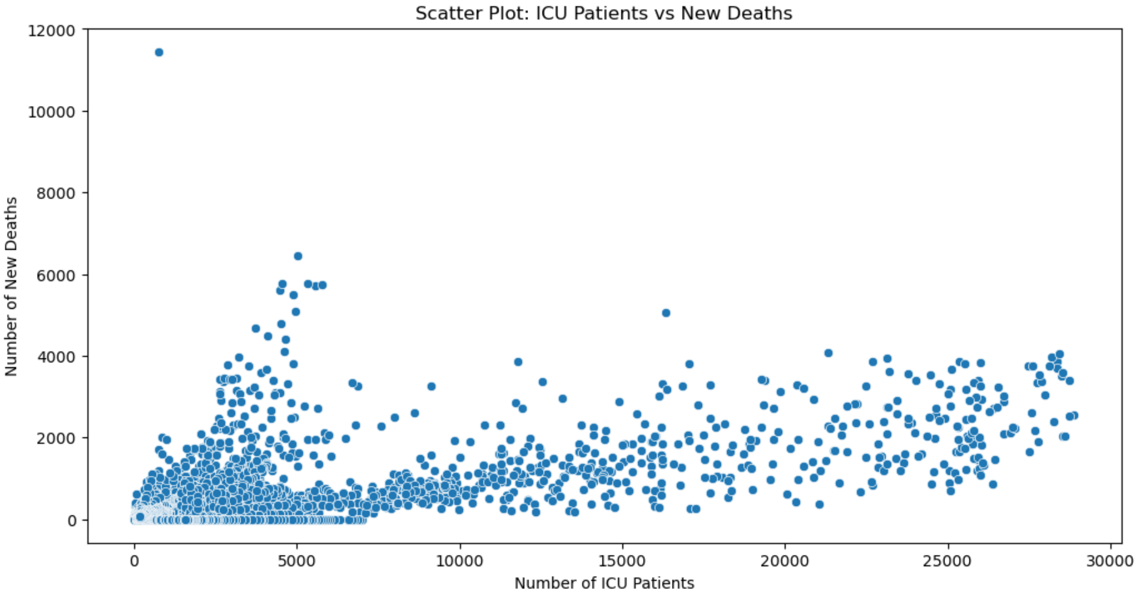
Line chart over time: Show how the reproduction rate changes over time. Insights: Identify periods of high and low reproduction rates.

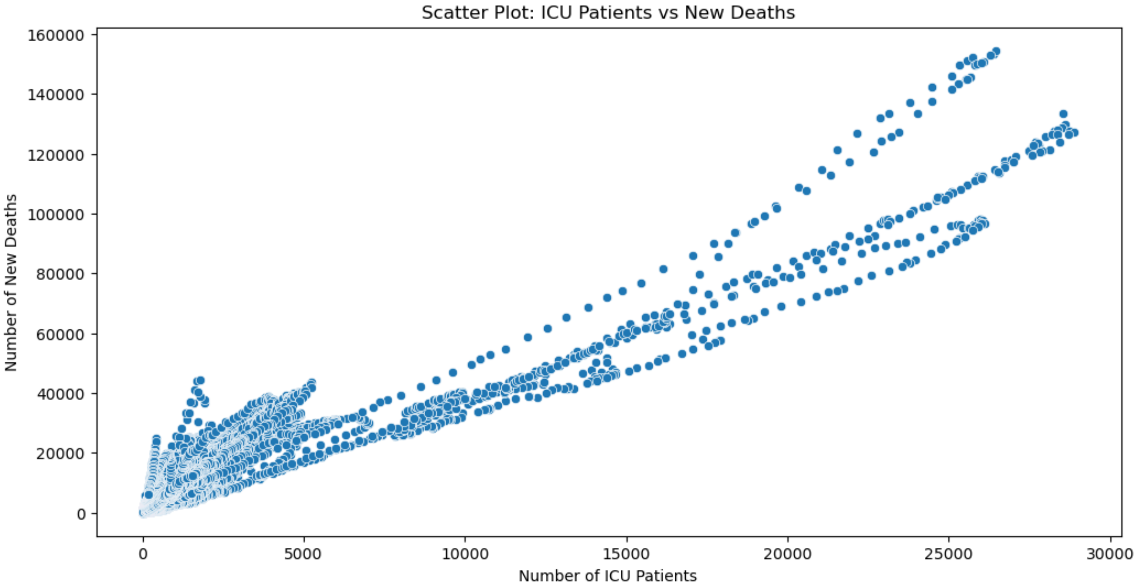


ICU Patients and ICU Patients Per Million:


ICU Patients Per Million Over Time


ICU Patients Over Time


Scatter Plot: ICU Patients vs New Deaths


Scatter Plot: ICU Patients vs New Deaths

Extracted correlation values related to 'icu_patients'

```
icu_patients                                 1.000000
hosp_patients                                0.941822
weekly_icu_admissions                        0.941201
weekly_hosp_admissions                       0.932974
new_deaths_smoothed                          0.903094
new_tests_smoothed                           0.735466
```

Summary statistics for ICU patients and total cases

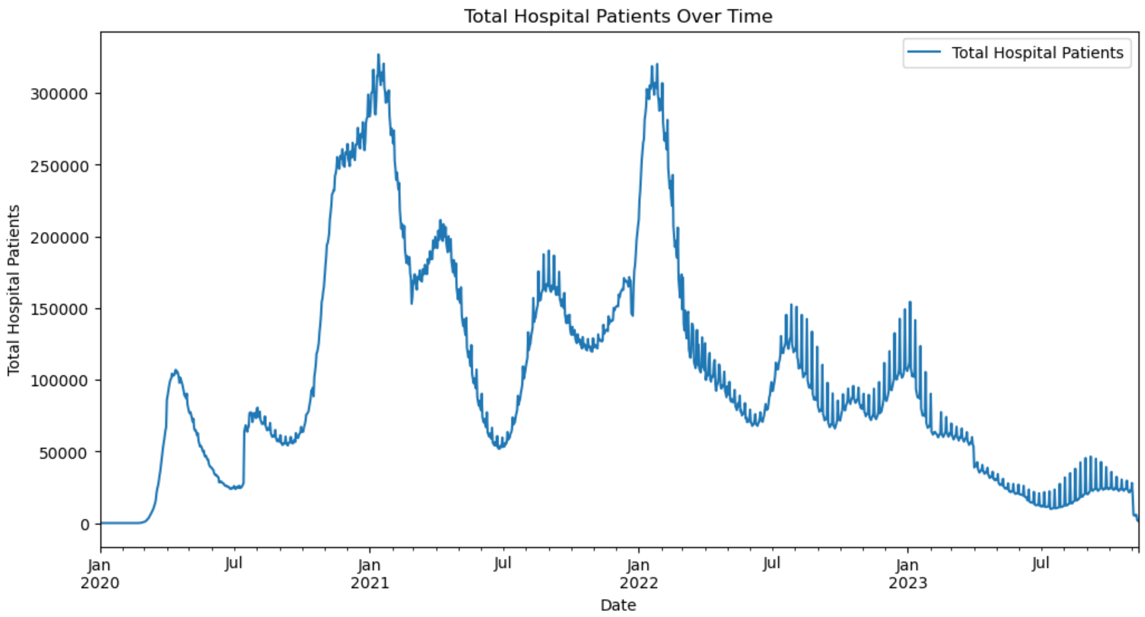
```
ICU Patients Summary:
count    33440.000000
mean       739.720156
std       2294.376172
min          0.000000
25%         25.000000
50%        112.000000
75%        492.000000
max      28891.000000
Name: icu_patients, dtype: float64

Total Cases Summary:
count    3.433120e+05
mean     6.285071e+06
std      3.969616e+07
min      0.000000e+00
25%      4.312000e+03
50%      4.996300e+04
75%      6.385288e+05
max      7.718202e+08
Name: total_cases, dtype: float64
```

- The average number of ICU patients is approximately 740. The standard deviation is relatively high (2294), indicating significant variability in ICU occupancy. The minimum value is 0, suggesting instances where there were no ICU patients reported. The maximum value is 28,891, indicating a wide range in the number of ICU patients across the dataset. Total Cases:

- The average total number of cases is around 6.29 million. The standard deviation is substantial (about 39.7 million), indicating a large variation in total cases globally. The minimum value is 0, which may represent missing or erroneous data for certain locations. The maximum value is approximately 771.82 million, indicating a broad range in the total number of cases across different regions.

Hospital Patients and Hospital Patients Per Million:

Group the data by date and calculate the total hospital patients for each date. Plot a line chart with the date on the x-axis and the total hospital patients on the y-axis.



- during March 2022 and January 2021 could be indicative of significant events or factors influencing healthcare systems.

- Investigate if there were specific events or factors during March 2022 and January 2021 that could explain the spikes. This could include the emergence of new variants, changes in public health policies, or peaks in infection rates.

- Examine the vaccination rates during these periods. Higher vaccination rates may influence the severity of cases and contribute to a potential decrease in hospitalizations. Public Health Interventions:

- Review any implemented public health interventions or restrictions during these times. Changes in mitigation measures can impact the spread of the virus and subsequent hospitalization rates

Dropping these features

After this analysis I decided to drop these features:

FEATURES_TO_DROP:

- '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'

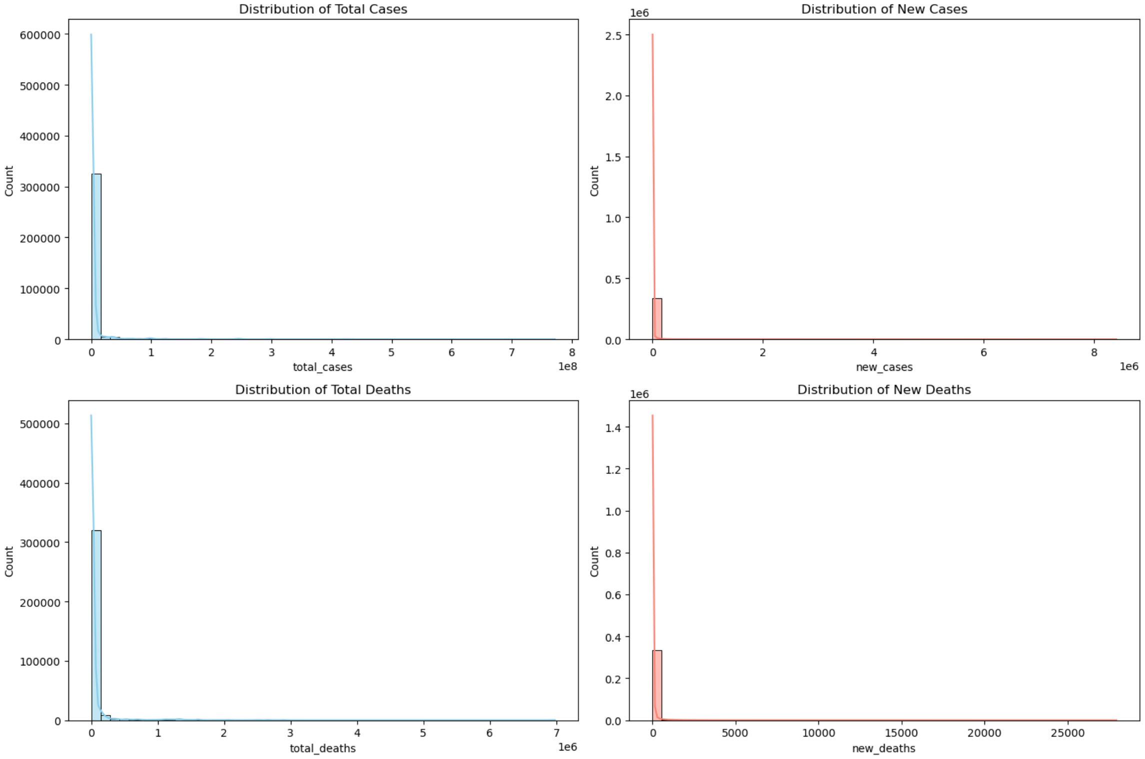
some of features that have many nan values, in the farther analysis has been dropped and do our work on data has remained.

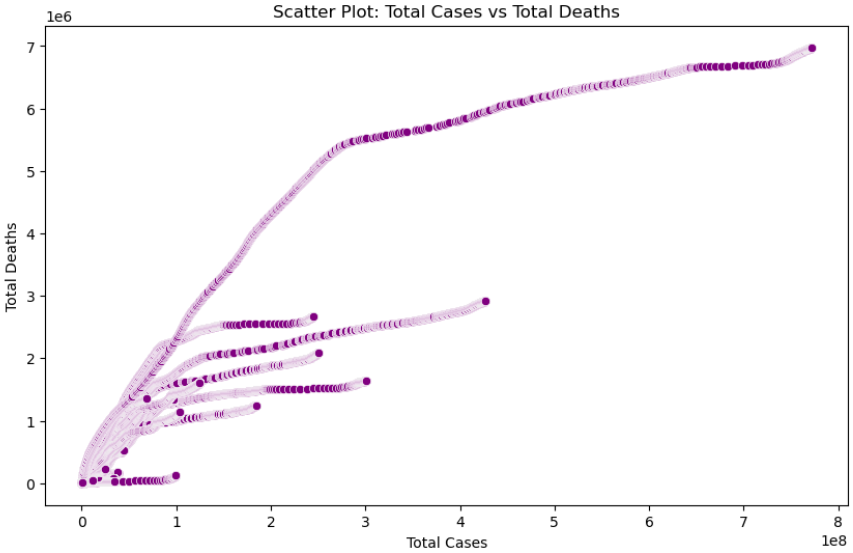
Since after the analysis I realized that null values are meaningful, the graphs drawn will be based on this criterion:

- **INCLUDE NULLS**
- **NO NULLS**

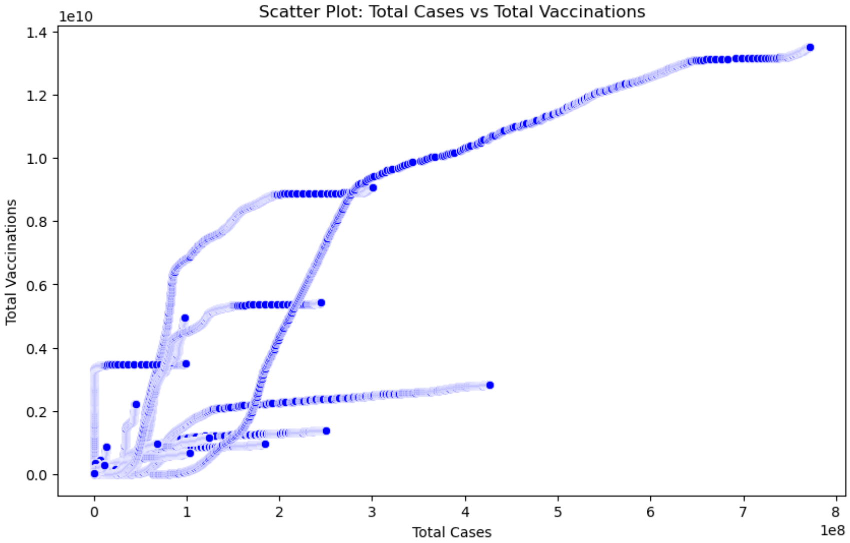
2. Data visualization

visualizing the distribution of 'total_cases', 'new_cases', 'total_deaths', and 'new_deaths'. This will give us an idea of the spread of these variables



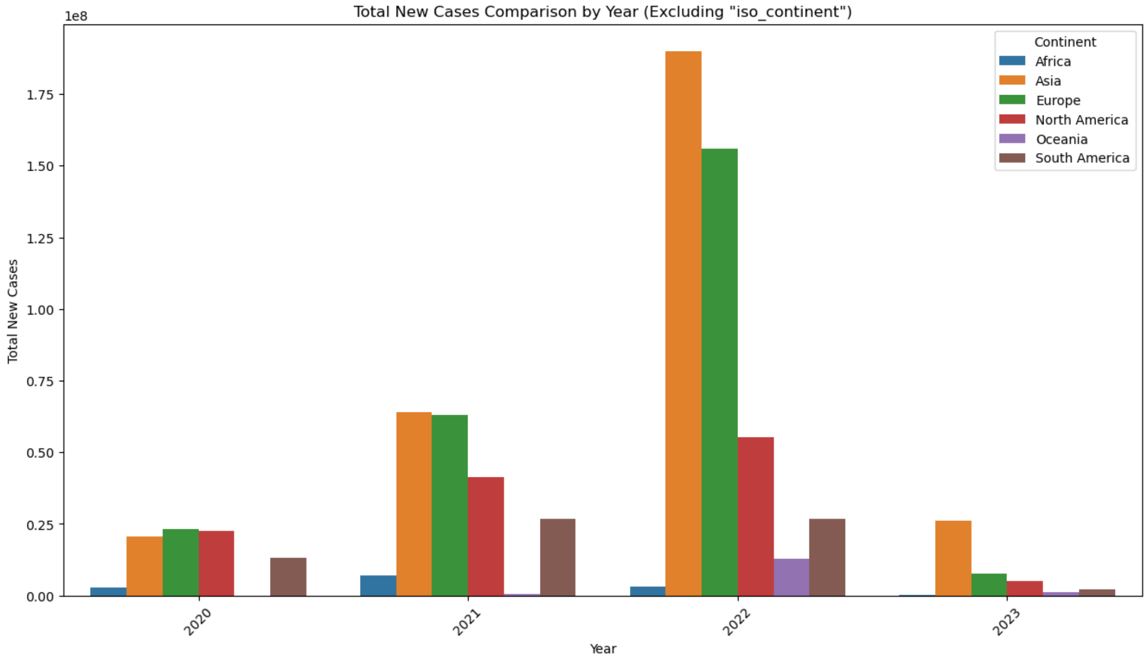


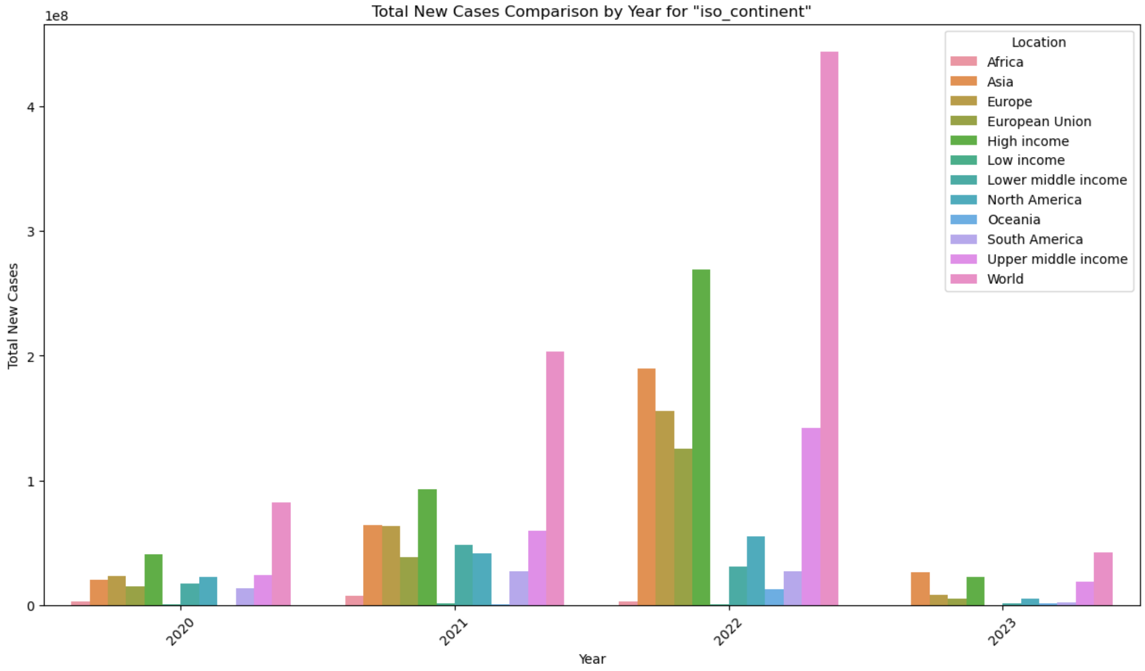
Indeed, there seems to be a positive correlation between 'total_cases' and 'total_deaths,' which is expected—higher total cases often correlate with higher total deaths

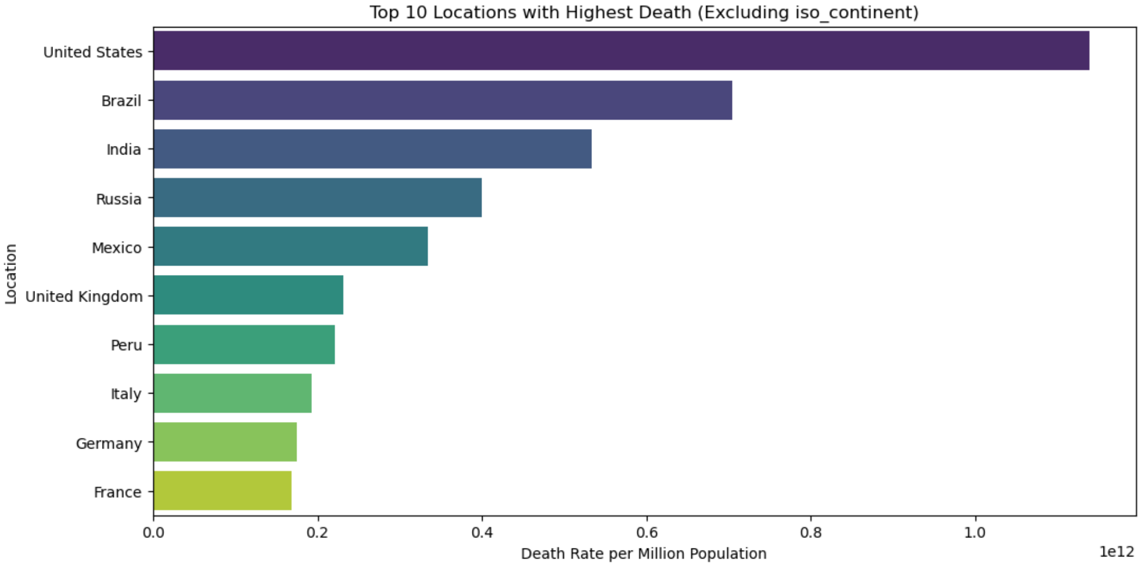


Indeed, a linear relationship suggests that as the total number of cases increases, the total number of vaccinations also tends to increase. This could be an indication of efforts to vaccinate the population in response to the rising number of cases.

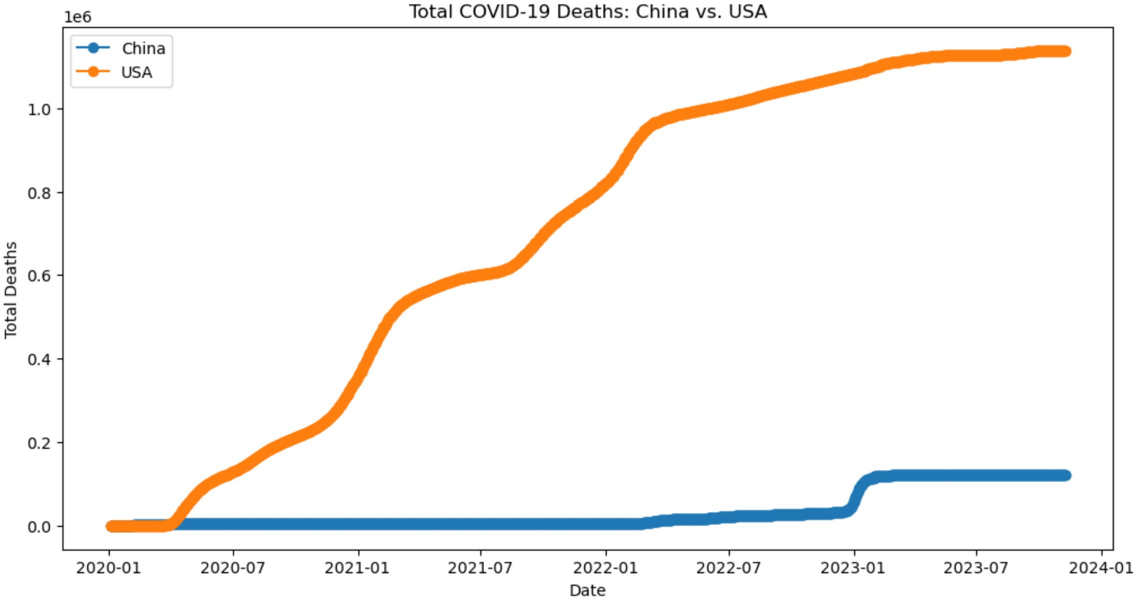
Total New Cases



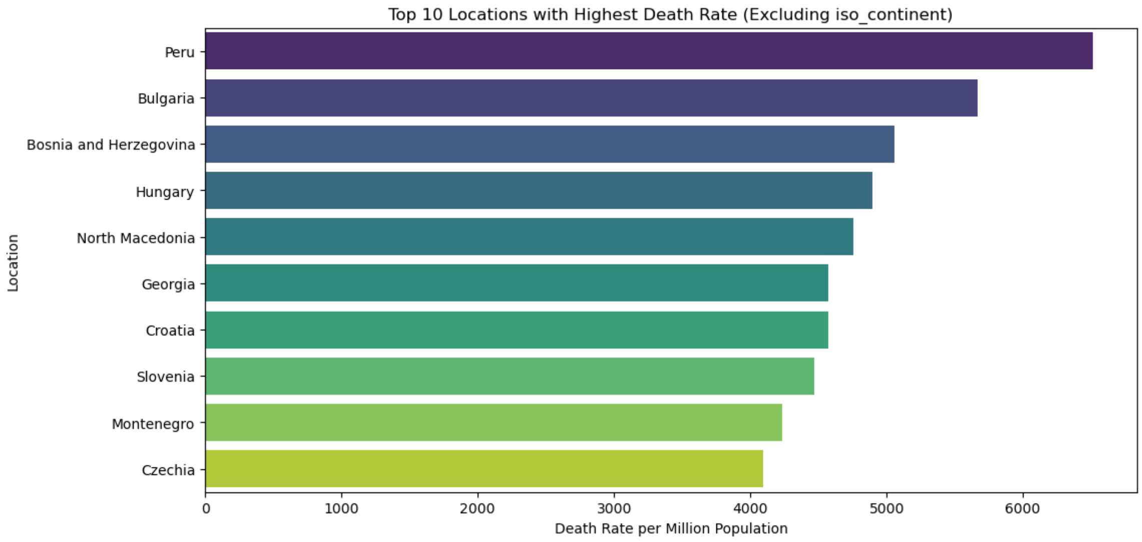




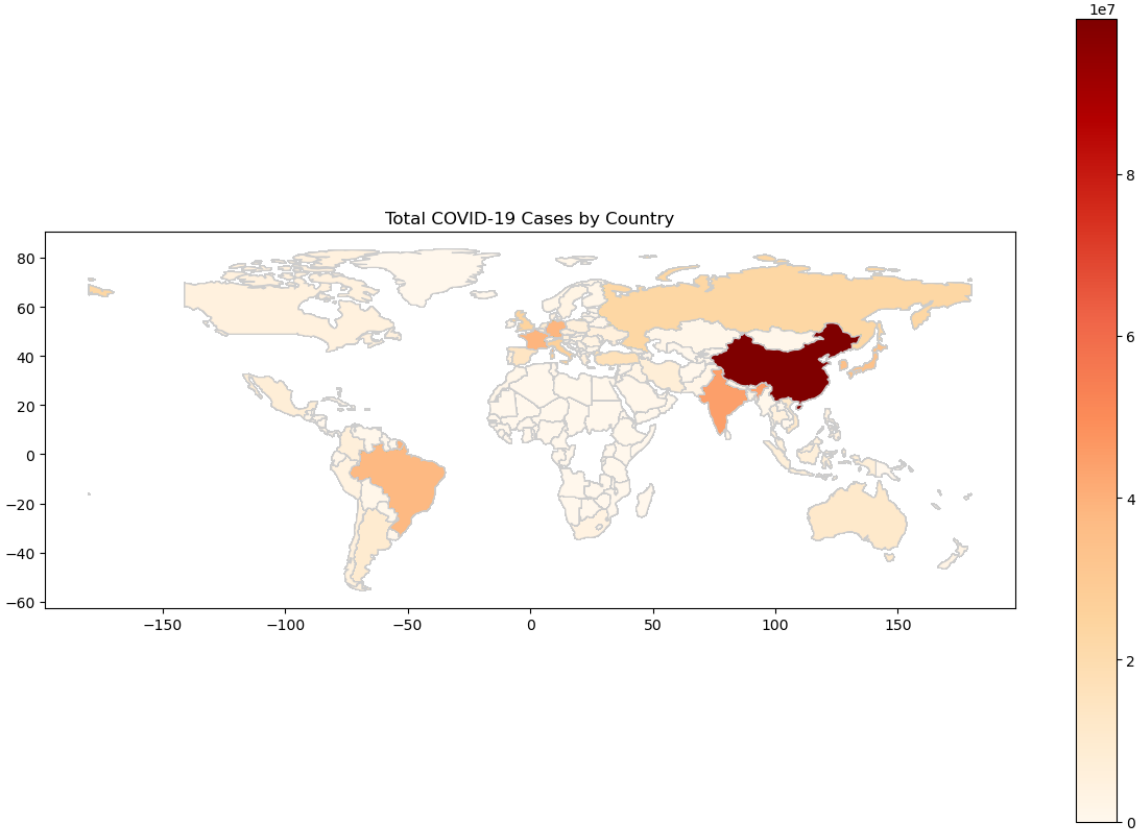
- After this chart, I doubted China's statistics and found this information

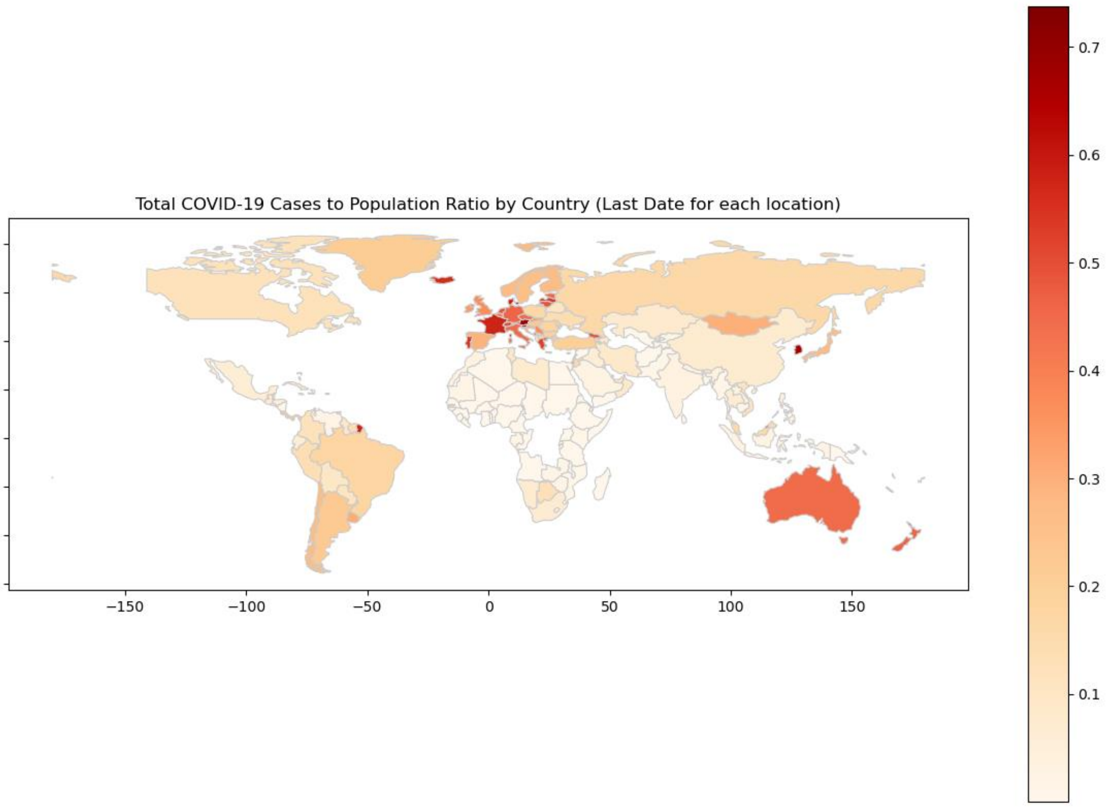


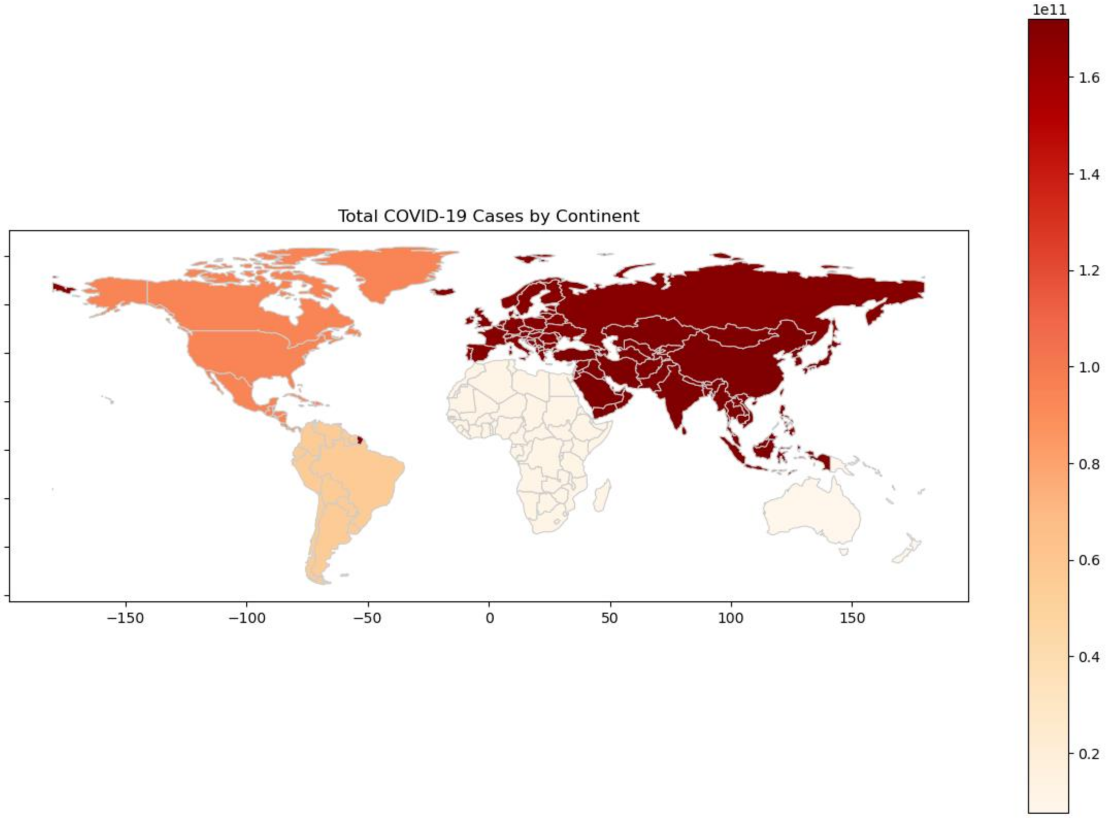
```
Total deaths in China: 121790.0
```



In addition to these approaches, also used map to explore geographical analysis on Covid-19.

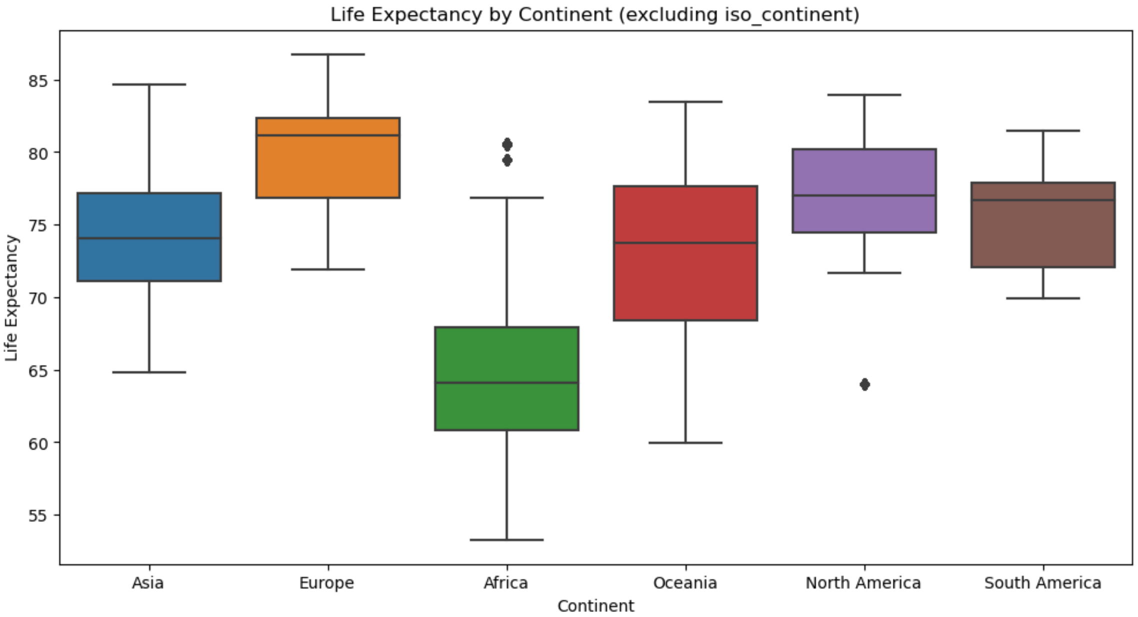



Total COVID-19 Cases to Population Ratio by Country (Last Date for each location)
80
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150
0.7
0.6
0.5
0.4
0.3
0.2
0.1

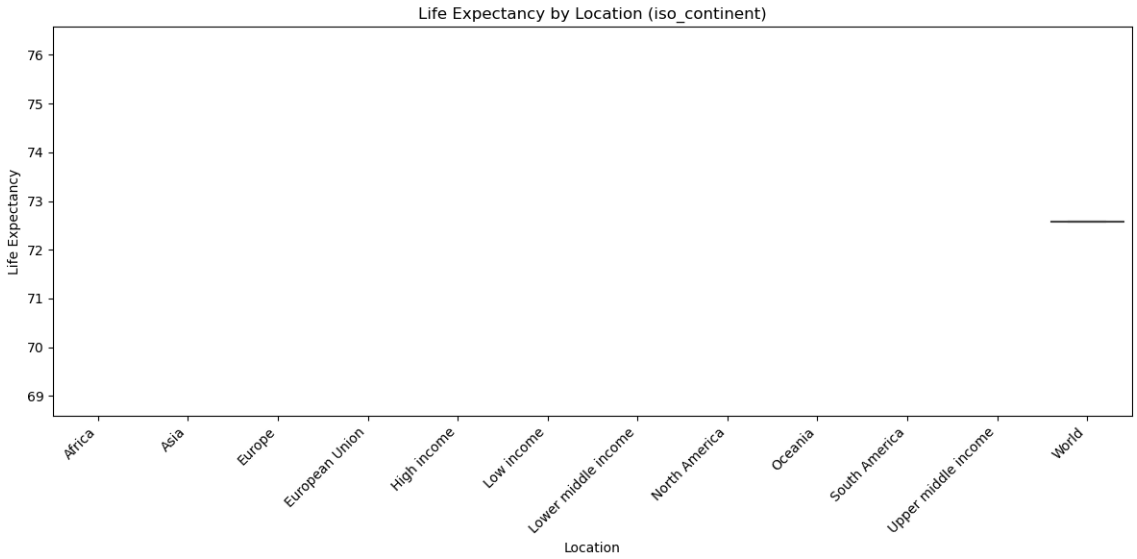

Total COVID-19 Cases by Continent
80
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150
1e11
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2

life_expectancy

```
location
Monaco                              86.75
San Marino                          84.97
Japan                               84.63
Cayman Islands                      83.92
Switzerland                         83.78
Andorra                             83.73
Singapore                           83.62
Spain                               83.56
Italy                               83.51

Central African Republic            53.28
Chad                                54.24
Lesotho                             54.33
Nigeria                             54.69
Sierra Leone                        54.70
Somalia                             57.40
Cote d'Ivoire                       57.78
South Sudan                         57.85
```



```
       continent  life_expectancy
0         Africa            64.10
1           Asia            74.05
2         Europe            81.15
3  North America            77.02
4        Oceania            73.70
5  South America            76.67
```



It seems data is not collected correctly for “iso_continent”

```
Iran's life expectancy: 76.68
```

Age over 65:

```
location
Japan                               18.493
Italy                               16.240
Germany                             15.957
Portugal                            14.924
Greece                              14.524
Latvia                              14.136
Spain                               13.799
```

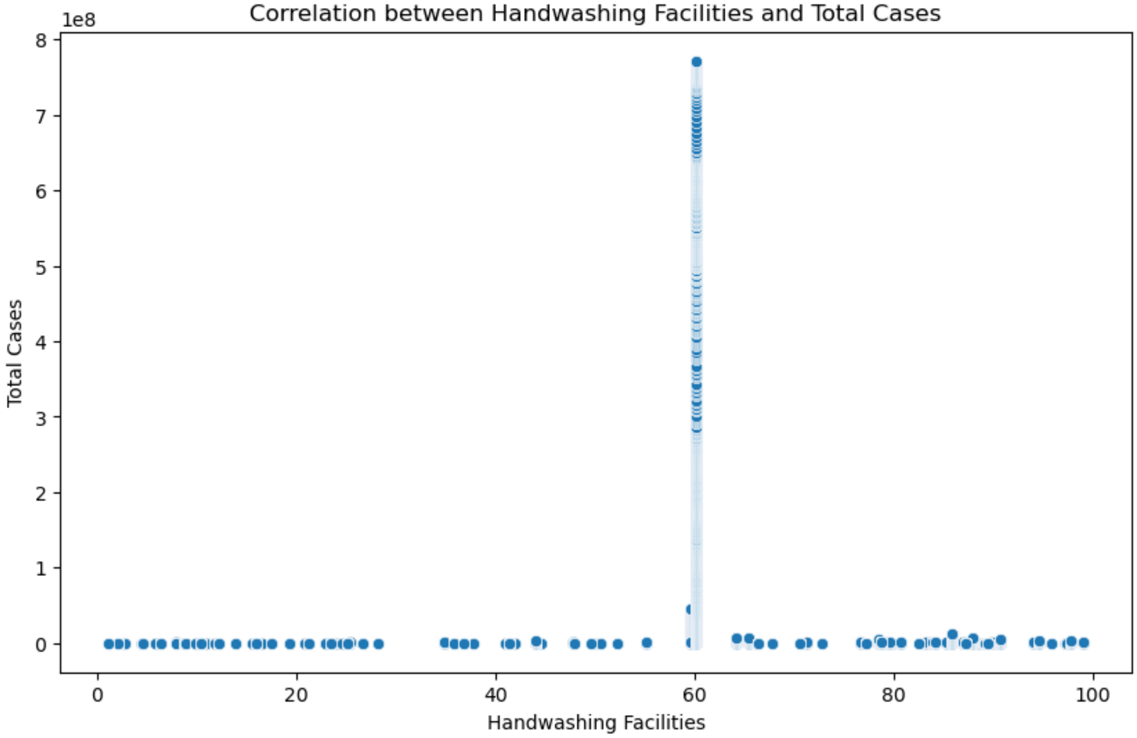
Mean Age:

```
location
Japan                               48.2
Italy                               47.9
Germany                             46.6
Portugal                            46.2
Martinique                          45.7
Spain                               45.5
Greece                              45.3
```

- As is characteristic of the countries that have the highest number of 65-year-olds, they also have a higher average age

HANDWASHING

The Pearson correlation coefficient between "handwashing_facilities" and "total_cases" is approximately 0.0331. The p-value is very close to zero (9.81e-34), indicating that the correlation is statistically significant. However, the correlation coefficient is relatively low, suggesting a weak positive correlation between the availability of handwashing facilities and the total number of COVID-19 cases. This means that as handwashing facilities increase, there is a slight tendency for total cases to increase, but the relationship is not very strong.



Explored relationships between variables, e.g., 'Life Expectancy' and 'GDP per Capita.' the correlation coefficient between 'reproduction_rate' and 'new_cases' is approximately 0.0197. This indicates a very weak positive correlation between these two variables. In other words, there is a slight tendency for an increase in the reproduction rate to be associated with a slight increase in the number of new cases, but the relationship is not strong.

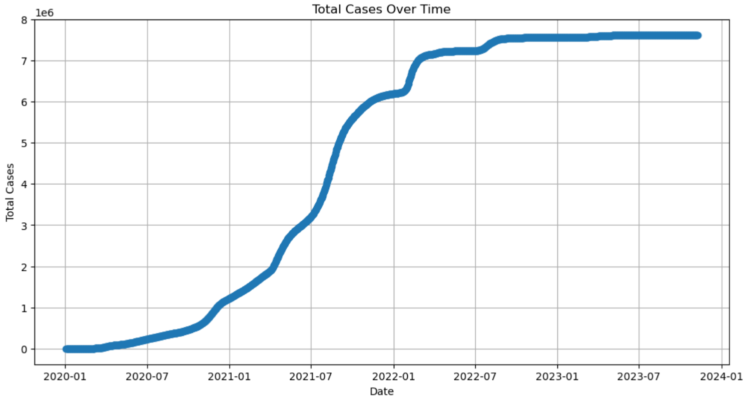
- Analyzed the correlation between 'Age' and 'Total Cases,' including statistical analysis.

```
Correlation Coefficient: 0.047103495045963714

P-Value: 1.4928627922317368e-134

The correlation is statistically significant.

Mean Age: 30.344619107843815

Mean Total Cases: 3709936.6972507825

Standard Deviation Age: 9.070943013411519

Standard Deviation Total Cases: 34038178.97808736
```

The correlation coefficient between 'median_age' and 'total_cases' is approximately 0.0471, and the p-value is very close to zero (1.49e-134). With such a low p-value, we can conclude that the correlation is statistically significant.
Correlation Coefficient (r): The positive correlation coefficient indicates a weak positive relationship between the median age of a population and the total number of COVID-19 cases. However, the strength of the correlation is relatively low.Statistical Significance: The very low p-value (close to zero) suggests that the observed correlation is unlikely to be due to random chance. Therefore, we have evidence to reject the null hypothesis and accept the alternative hypothesis that there is a significant correlation.

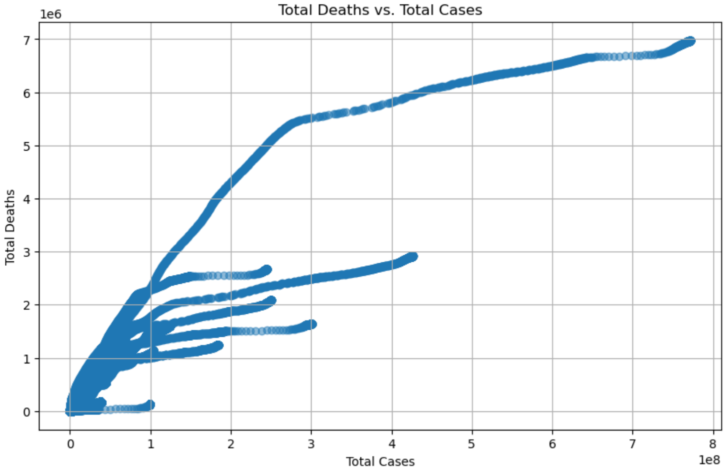
Total Cases Over Time:

Line plot showing the total cases over time.



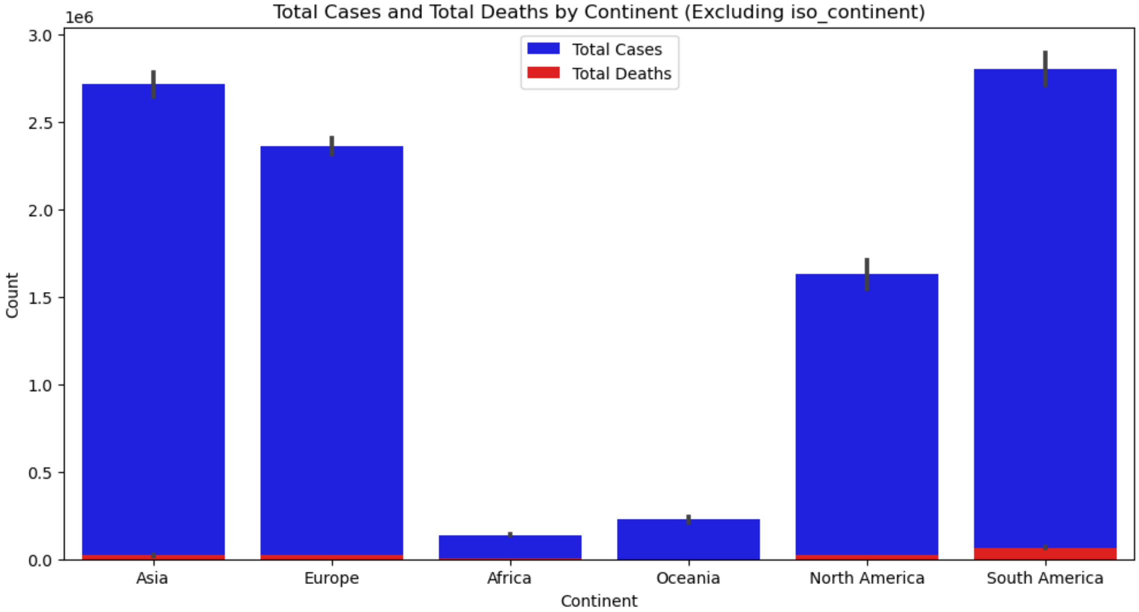
Total Deaths vs. Total Cases:

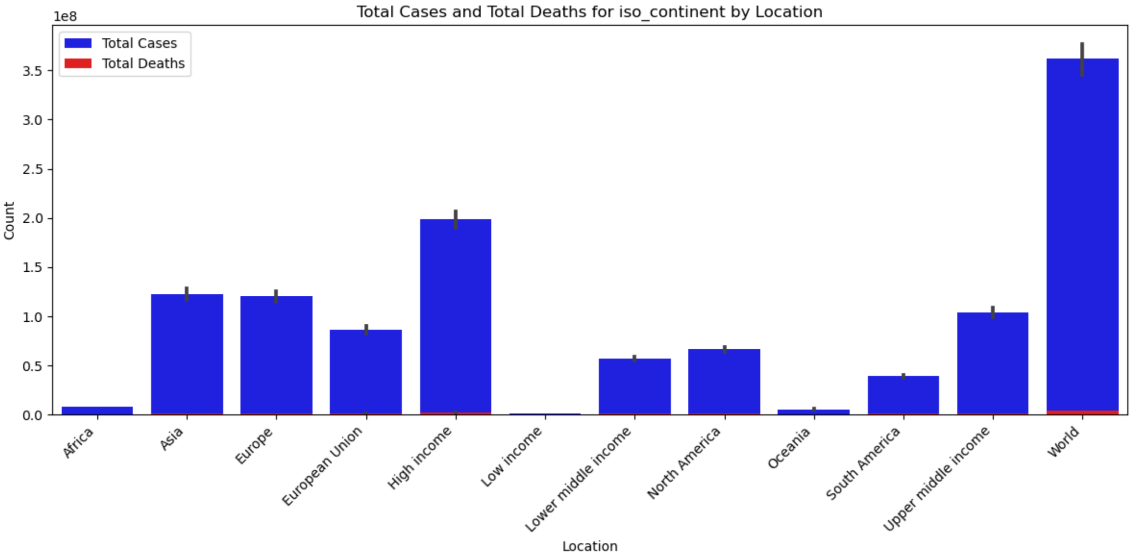
Scatter plot showing the relationship between total deaths and total cases.



Total Cases and Total Deaths by Continent:

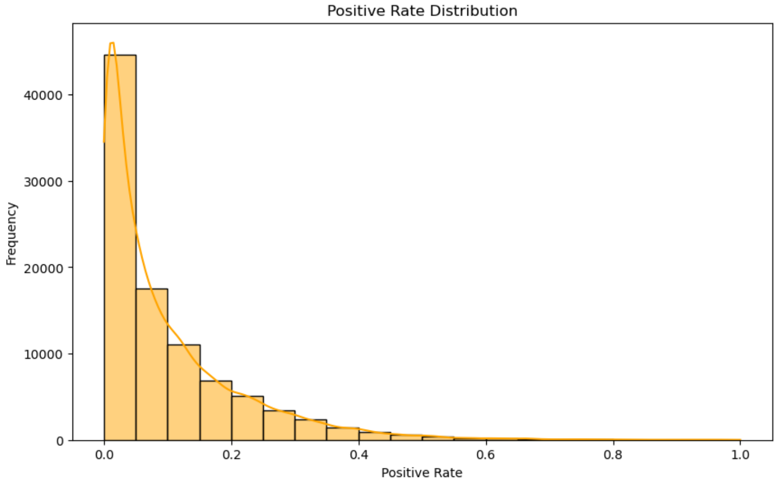
Bar plot showing the total cases and total deaths for each continent.





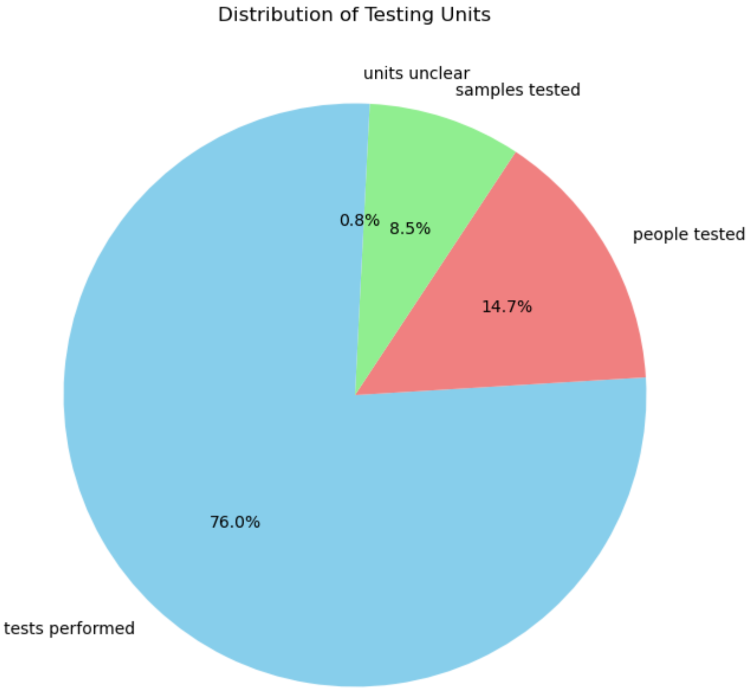
Positive Rate Distribution:

Histogram showing the distribution of positive rates.



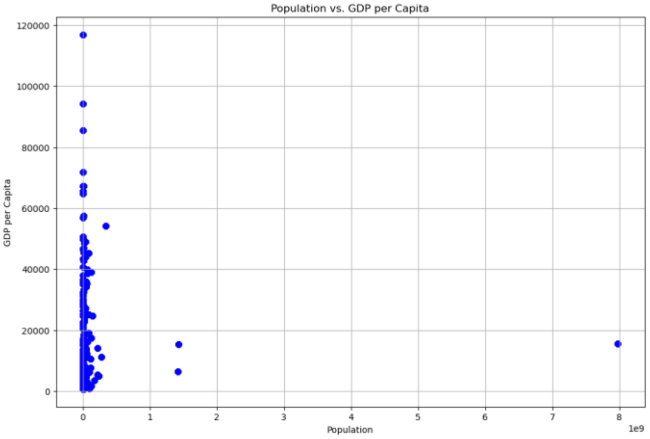
Testing Units Distribution:

Pie chart showing the distribution of testing units.

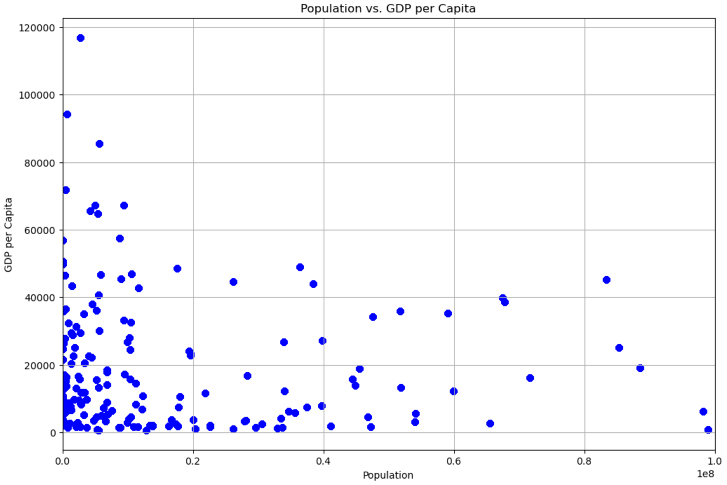


Population vs. GDP per Capita:

Scatter plot showing the relationship between population and GDP per capita.



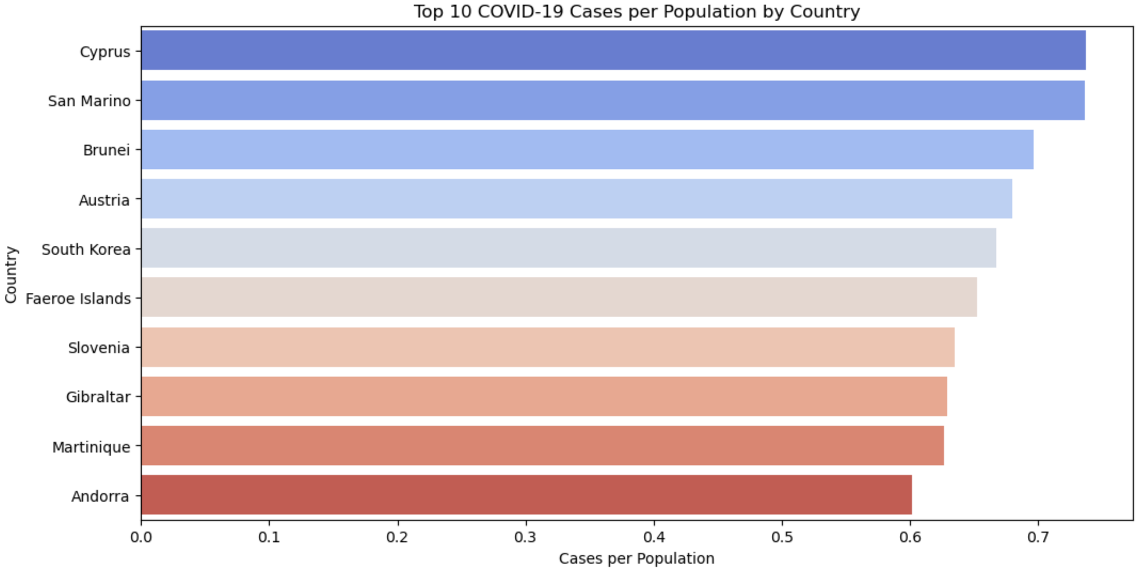
After plotting, I realized that the criteria was not correct and drew a new chart again with the new criteria.



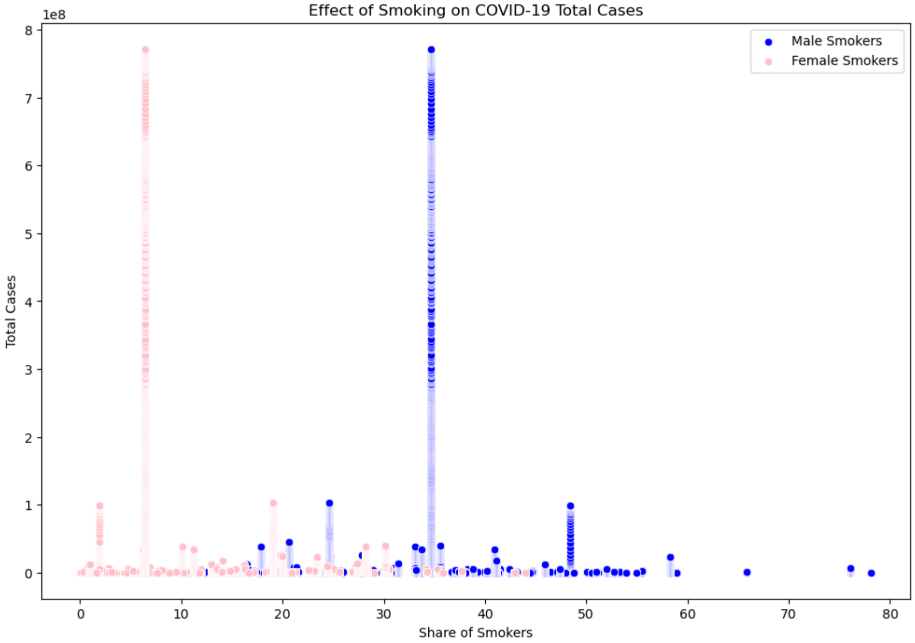
Well, as the correlation function showed, there is no linear relationship

Top COVID-19 cases:

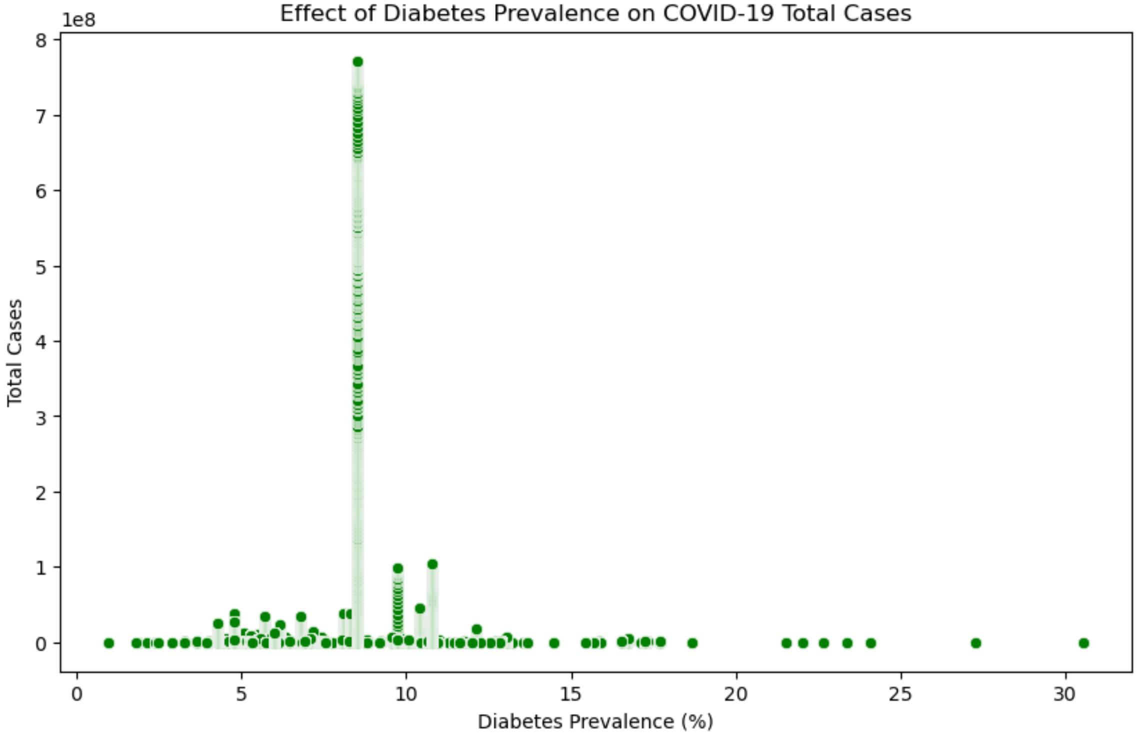
creating a bar chart to visualize the top 10 countries with the highest COVID-19 cases per population ratio



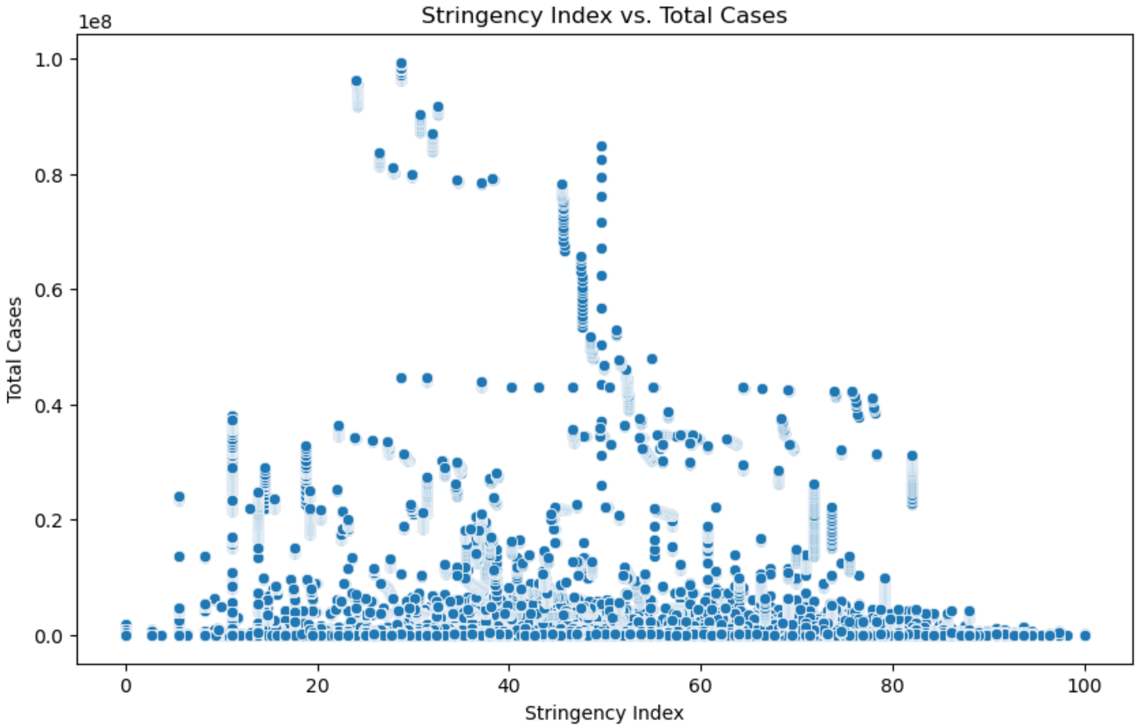
SMOKING



DIABETES

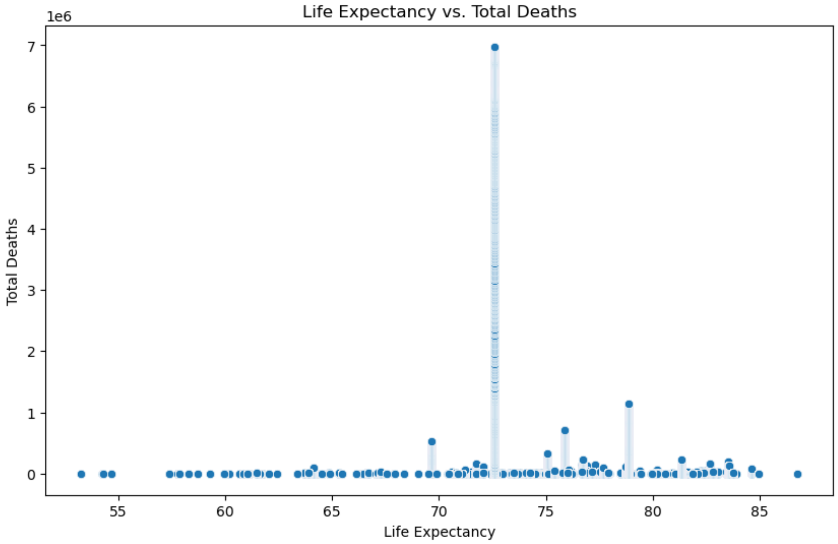


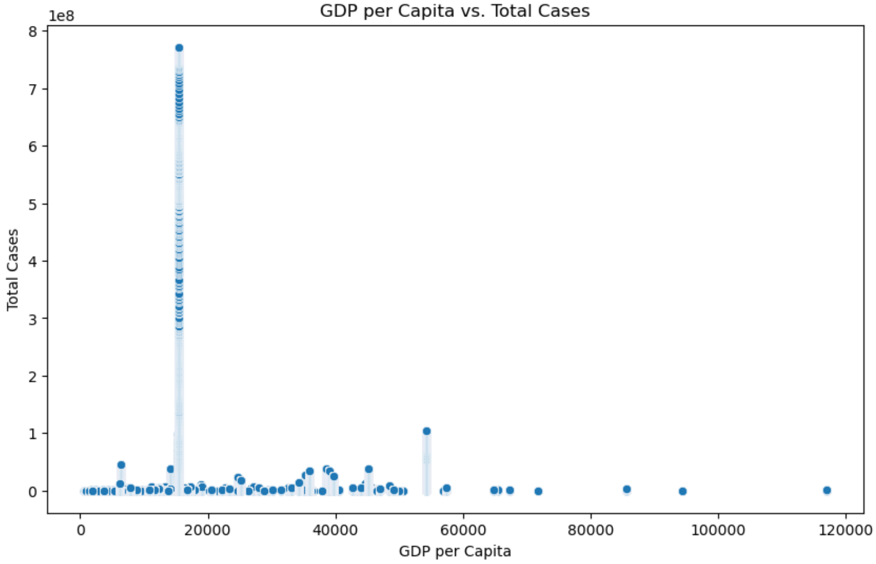
```
Correlation between Diabetes Prevalence and Total COVID-19 Cases: -0.0057131158
68240373

Correlation between Male Smokers Percentage and Total COVID-19 Cases: 0.0036120
134780312355
cardiovasc_death_rate                        0.425771
hospital_beds_per_thousand                   0.347238
handwashing_facilities                       0.327041
excess_mortality_cumulative_per_million      0.248458
diabetes_prevalence                          0.201121
extreme_poverty                              0.192306
```



According to the plot, it can be understood that there is no relationship:

- The correlation between 'stringency_index' and 'total_cases' is -0.0631996766158127.





- The mean of 'life_expectancy' is approximately 73.72, and the mean of 'gdp_per_capita' is approximately 17418.15.

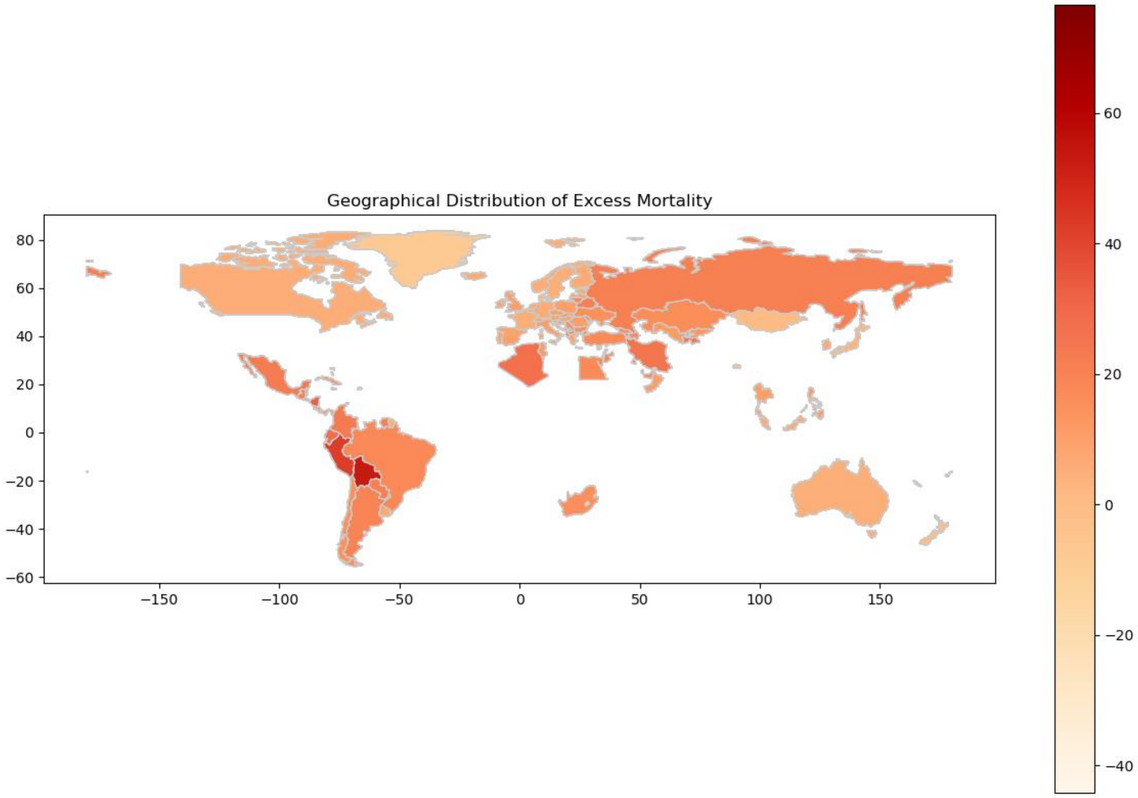
- It seems that there is no relationship and the plot is shown only on very high values which is the Mean.

Excess mortality

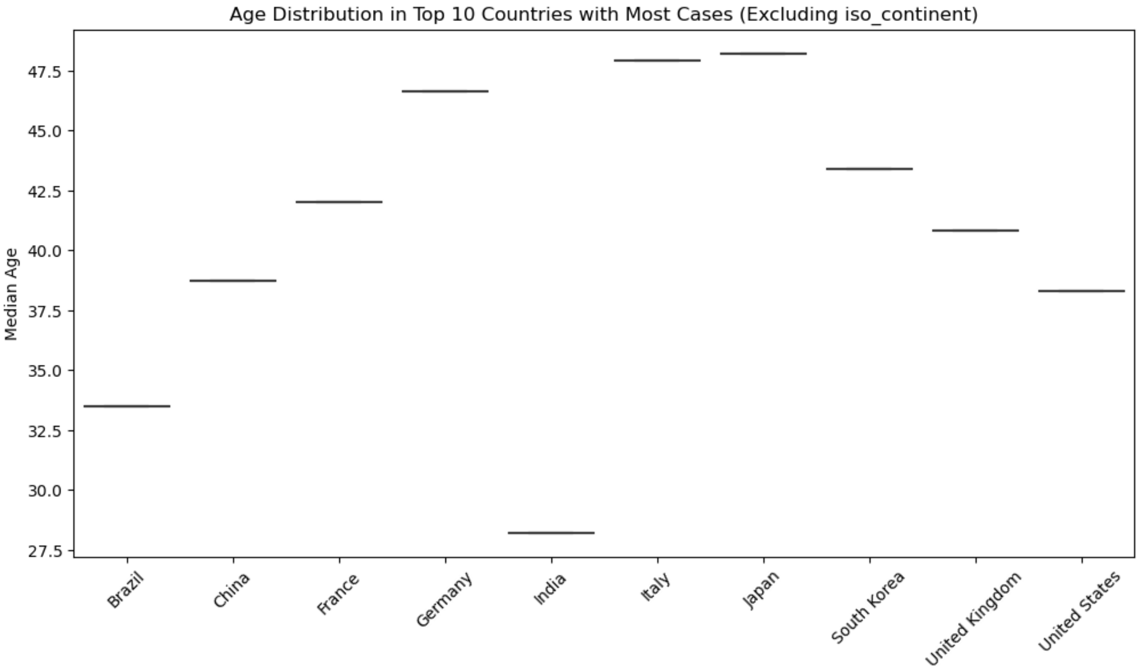
- the excess mortality can be calculated by taking the observed number of deaths during a specific period and subtracting the expected number of deaths based on historical data for the same period. This measure provides a more comprehensive understanding of the impact of a health crisis on mortality, especially when there may be underreporting or misclassification of causes of death.

- In the context of COVID-19, excess mortality has been used to assess the true impact of the virus, considering both confirmed COVID-19 deaths and any additional deaths that may be related to the pandemic's indirect effects.

- In summary, a higher excess mortality suggests that the observed number of deaths exceeds what would be expected under normal circumstances, indicating the broader impact of a health crisis on mortality.

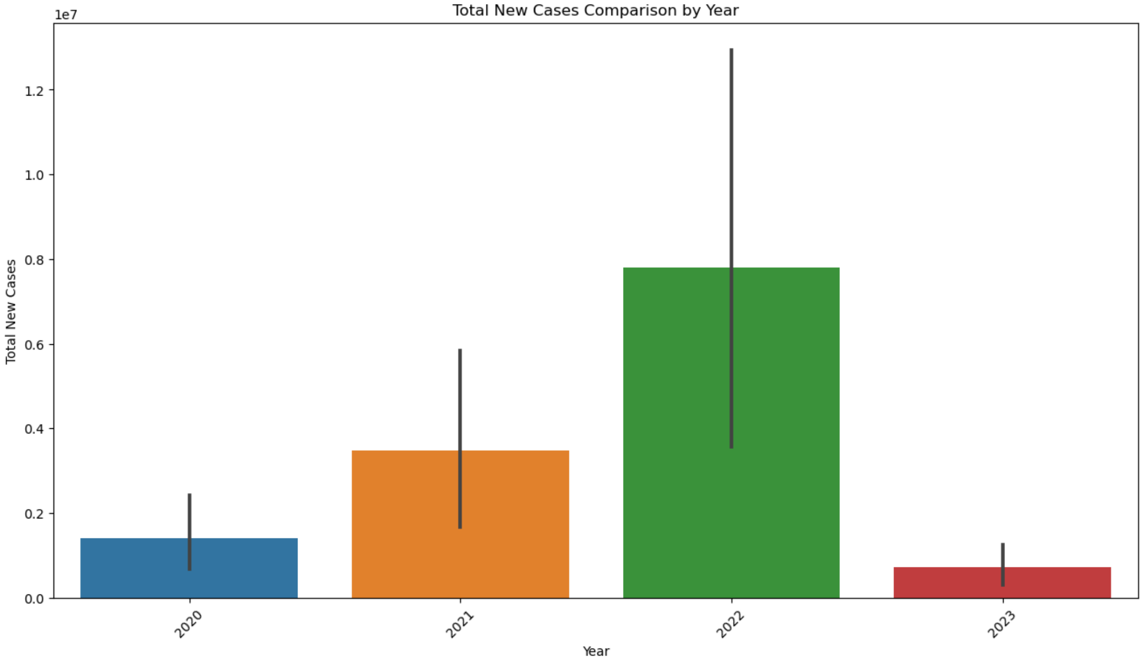


Age



Age Distribution in Top 10 Countries with Most Cases (Excluding iso_continent)
Median Age
47.5
45.0
42.5
40.0
37.5
35.0
32.5
30.0
27.5
Brazil
China
France
Germany
India
Italy
Japan
South Korea
United Kingdom
United States
Country

BY Year



Total New Cases Comparison by Year
1e7
Total New Cases
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2020
2021
2022
2023
Year

3. Summary

This report delves into a detailed analysis of a COVID-19 dataset, encompassing 355,434 entries and 66 columns. The exploration begins with an initial inspection, revealing missing values and anomalies, leading to meticulous data cleaning strategies.

Initial Data Inspection:

- The dataset spans until 2023-11-09, with the last day data reported for North Korea showing NaN, indicating potential data unavailability.
- Investigation identifies 'North Korea' and 'Turkmenistan' as countries with NaN total cases on the last date; these entries are subsequently dropped.

Missing Values Analysis:

- Various columns exhibit missing values, with 'reproduction_rate,' 'icu_patients,' and 'stringency_index' having notably high percentages.
- Imputation and handling strategies are applied based on the nature of the data, considering the meaningfulness of null values.

Handling Missing Data:

- The 'continent' column is updated using the 'iso_continent' prefix for specific countries.
- Missing values in 'total_cases' are replaced with zeros, and countries with '0' total cases on their last reported day are identified and handled accordingly.

Correlation Analysis:

- Correlation coefficients are calculated, revealing strong positive correlations between 'total_cases' and 'total_deaths,' 'total_cases' and 'total_vaccinations,' among others.
- Features with high correlation to 'total_cases' are identified, including 'total_deaths,' 'total_boosters,' and 'total_tests.'

Exploration of Specific Variables:

- In-depth analyses are conducted on variables such as 'reproduction_rate,' 'icu_patients,' 'hosp_patients,' and 'weekly_icu_admissions.'
- Hospitalization spikes during March 2022 and January 2021 prompt further investigation into potential contributing factors.

Data Visualization:

- A wide array of visualizations, including line charts, bar plots, scatter plots, and histograms, offer insights into the distribution and relationships of key variables.
- Geographical analysis using maps provides a global perspective on COVID-19 statistics.

Feature Removal:

- Features with high percentages of missing values or limited relevance are dropped post-analysis.

Health and Demographic Analysis:

- Life expectancy, age distribution, and other demographic factors are explored, revealing intriguing insights.

Excess Mortality:

- Excess mortality is discussed as a crucial metric for assessing the broader impact of the health crisis.

Conclusion:

- The report concludes with a summary of the dataset's characteristics and a thorough exploration of various dimensions, providing a comprehensive understanding of the COVID-19 data.

This report not only offers a meticulous exploration of the dataset but also demonstrates a thoughtful approach to handling missing data and deriving meaningful insights from diverse variables. The combination of statistical analyses and visualizations enhances the clarity of findings, making this report a valuable resource for understanding the dynamics of the COVID-19 dataset.