Data Science 2nd Assignment

Dataset: OWID Covid Data

Mahan Madani – 99222092

1. Overview

This dataset is comprised of over 350’000 records of the Coronavirus pandemic, containing information in 67 different columns. The data was gathered by the Our World in Data (OWID) organization over the duration of the pandemic.

This dataset contains useful information on the number of confirmed cases and deaths during the pandemic, statistics about hospitalizations and intensive care units (ICUs), Testing for COVID-19, and vaccinations against COVID-19.

2. Dataset Exploration & Analysis

The dataset stores 67 different attributes each relating to a record of COVID-19 on a specific date and in a specific location.

Smoothed Data: Certain columns in the Covid dataset store a smoothed version of another column. Data Smoothing is a statistical method of removing outliers from datasets in order to make patterns more visible. It is accomplished by employing Algorithms to remove Statistical Noise from Datasets.

In this particular dataset, the data smoothing was performed on a 7-day period. This specific length allows us to understand patterns on a weekly basis.

Below you can find a list of some of the key attributes:

- **continent:** Continent of the geographical location
- **location:** Geographical location
- **date:** Date of observation
- **population:** Population of the location (latest available values)
- **population_density:** Number of people divided by land area, measured in square kilometers (most recent year available)

- **median_age:** Median age of the population, UN projection for 2020
- **gdp_per_capita:** Gross domestic product at purchasing power parity (constant 2011 international dollars)
- **extreme_poverty:** Share of the population living in extreme poverty (most recent year available since 2010)
- **female_smokers:** Share of women who smoke, most recent year available
- **male_smokers:** Share of men who smoke (most recent year available)
- **handwashing_facilities:** Share of the population with basic handwashing facilities on premises (most recent year available)
- **total_cases:** Total confirmed cases of COVID-19
- **new_cases:** New confirmed cases of COVID-19
- **total_deaths:** Total deaths attributed to COVID-19
- **new_deaths:** New deaths attributed to COVID-19
- **hosp_patients:** Number of COVID-19 patients in hospital on a given day
- **weekly_hosp_admissions:** Number of COVID-19 patients newly admitted to hospitals in a given week
- **stringency_index:** Government Response Stringency Index. A composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100
- **reproduction_rate:** Estimate of the effective reproduction rate of COVID-19
- **total_tests:** Total tests for COVID-19
- **new_tests:** New tests for COVID-19 (only calculated for consecutive days)
- **positive_rate:** The share of COVID-19 tests that are positive, given as a rolling 7-day average
- **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 initial vaccination protocol
- **total_boosters:** Total number of COVID-19 vaccination booster doses administered

3. Data Preprocessing

3.1 Drop Unnecessary Features:

Certain features in the dataset can be dropped due to various reasons:

- All **Weekly** attributes (e.g., “weekly_icu_admissions”). These features can be calculated using the rest of the stored features.

- All **Per Hundred/Thousand** attributes (e.g., “total_vaccinations_per_hundred”). These features can be calculated using the population if they are needed.

- All **Excess Mortality** attributes (e.g., “excess_mortality”). Excess mortality is a term that refers to the number of deaths from all causes during a crisis. This includes deaths caused by COVID-19 in addition to other source. These attributes are not necessary as this research is solely focused on COVID-19.

- The following attributes simply provide little value for this research, as such they can be dropped:

“life_expectancy” – “human_development_index” – “tests_units”

3.2 Check for Duplicate Records:

To ensure the dataset contains no duplicate records, a combination of the “iso_code” and “date” features can be used. If two ore more records share the same “iso_code” and “date”, that means they are potentially duplicates.

Even if the records aren’t completely identical, they may still contradict each other and need to be handled. **After testing the dataset, it seems that it contains no duplicate records** and no further action is required.

3.5 Convert Data Types:

- “date” is stored as a string in the dataset. By converting it to a datetime object, it can be used to generate useful features and also allow us to sort the data based on the date.

3.4 Handle Null Values:

Out of the 47 remaining features, several contain null values that must be dealt with. The exact number of null values for each attribute is displayed in the following table:

	Null Count
iso_code	0
continent	16872
location	0
date	0
total_cases	37962
new_cases	9558
new_cases_smoothed	10817
total_deaths	59675
new_deaths	9510
new_deaths_smoothed	10740
total_cases_per_million	37962
new_cases_per_million	9558
new_cases_smoothed_per_million	10817
total_deaths_per_million	59675
new_deaths_per_million	9510
new_deaths_smoothed_per_million	10740
reproduction_rate	170357
icu_patients	317451
icu_patients_per_million	317451
hosp_patients	316119
hosp_patients_per_million	316119
total_tests	275787
new_tests	279771

new_tests_smoothed	251209
positive_rate	259247
tests_per_case	260826
total_vaccinations	275506
people_vaccinated	278922
people_fully_vaccinated	282244
total_boosters	307274
new_vaccinations	289508
new_vaccinations_smoothed	172616
new_vaccinations_smoothed_per_million	172616
new_people_vaccinated_smoothed	172840
stringency_index	157523
population_density	53588
median_age	74735
aged_65_older	84520
aged_70_older	77547
gdp_per_capita	80302
extreme_poverty	178014
cardiovasc_death_rate	79652
diabetes_prevalence	65631
female_smokers	148488
male_smokers	151300
handwashing_facilities	220196
population	0

This dataset was created and expanded by importing data from several different sources. As such, many of the records have missing values. There are numerous methods for handling null values:

- For **Categorical** attributes, I replaced all null values with “Unknown”. From the list of remaining features, the only categorical feature with null values is “continent”.
- For **Numerical** attributes, different approaches should be taken based on the number of null values:

    - **Columns with a low quantity of null values:**

      Features that are found on the majority of the dataset (more than 90%) fall into this category. These are all important features and we would rather not reduce their accuracy by imputing data. Therefore, the records that are missing these values should be dropped.

      About 11’000 records are dropped after performing the operation.

    - **Columns with an average quantity of null values:**

      Due to the fact that we’re dealing with time series data, we can utilize **interpolation** and fill in the missing values with acceptable accuracy.

      Additionally, a backwards-fill can be used to fill the missing values that cannot be interpolated.

    - **Columns with a high quantity of null values:**

      Features that are only included in a small portion of the dataset (less than 20%) fall into this category. The amount of data available in these columns is limited and cannot be used to perform an accurate interpolation.

      One method is to create a number of **sub datasets** that only includes the records that have genuine values for these columns. This is the method I chose for this report, and three sub datasets were created for Hospitals, COVID-19 testing, and Vaccinations.

      *Another method is to use **regression** to predict these values for the records that are missing these features. This method is more time consuming as first we’ll need to find a number of appropriate features that correlate to the missing values and then perform the regression.

3.5 Detect Outlier Values:

Using the z-score method, outlier data can be identified. Over 20’000 records were detected as outliers and should be dropped from the dataset. Note that the standard deviation of the majority of the columns is rather high, and with a high standard deviation, it is more challenging to set a threshold for what is considered an outlier.

3.6 Feature Generation:

Based on the available data, the following features can be added to further improve the dataset:

- **year:** The year of the record. Extracted from “date (integer)
- **month:** The month of the record. Extracted from “date”. (string)
- **total_smokers:** The share of smokers, regardless of gender. Extracted from “male_smokers” and “female_smokers” by adding both features. (Integer)

4. Data Visualization:

Below you can find a number of plots visualizing the data so that it can be more easily understood.

Various different plots and charts can be used to visualize the data. I chose these 6 figures as a sample of what can be deduced from this dataset.

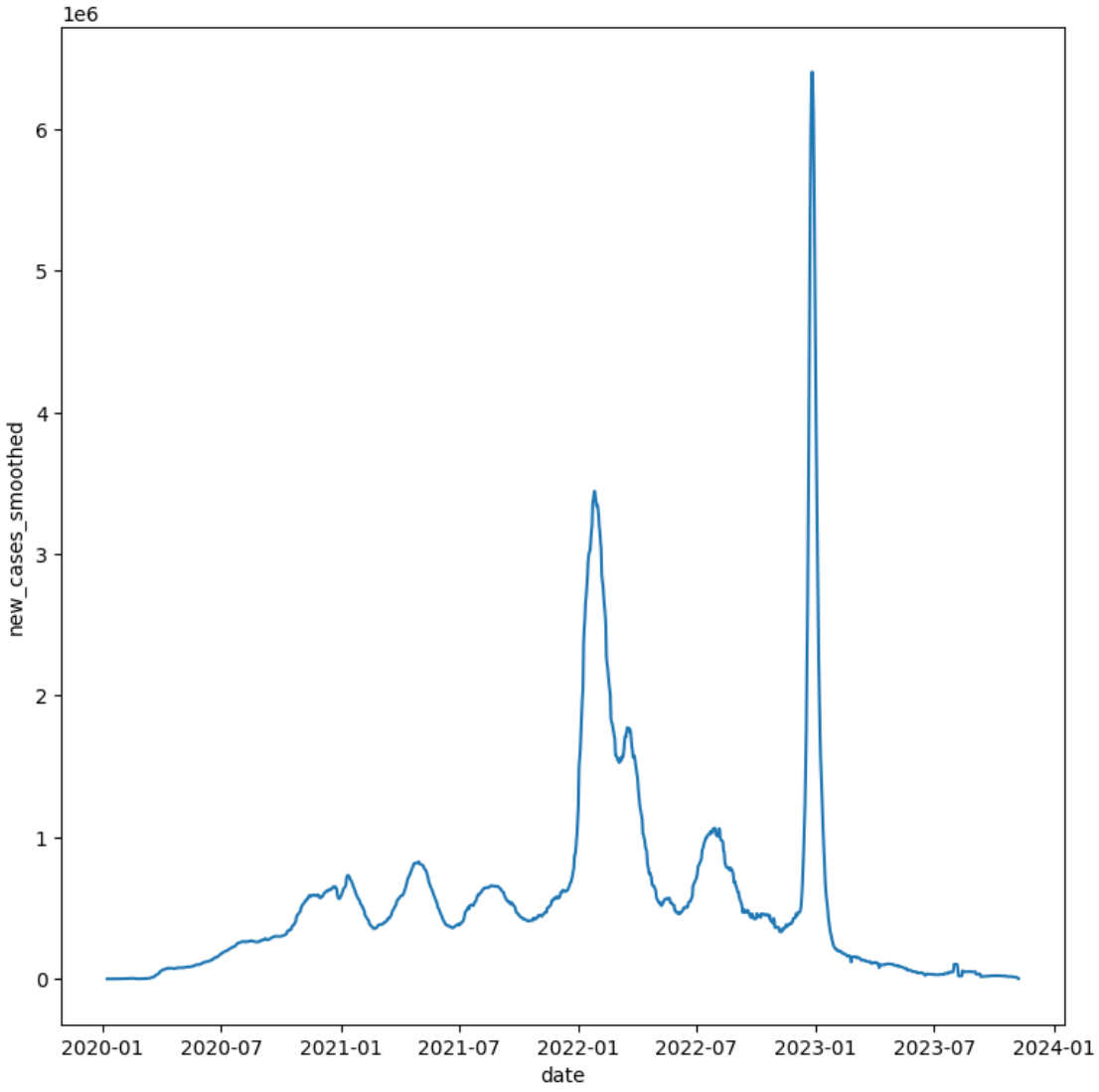


Figure 4.1

The plot depicts the numbers of new COVID-19 cases per record in the World. This is the total number of new cases (smoothed over a 7-day period) reported and compiled from all countries.

We can see the multiple peaks of new cases being reported, especially in early 2022 and early 2023. Additionally, the number of new cases has reached an all-time low in November 2023 since the start of the pandemic.

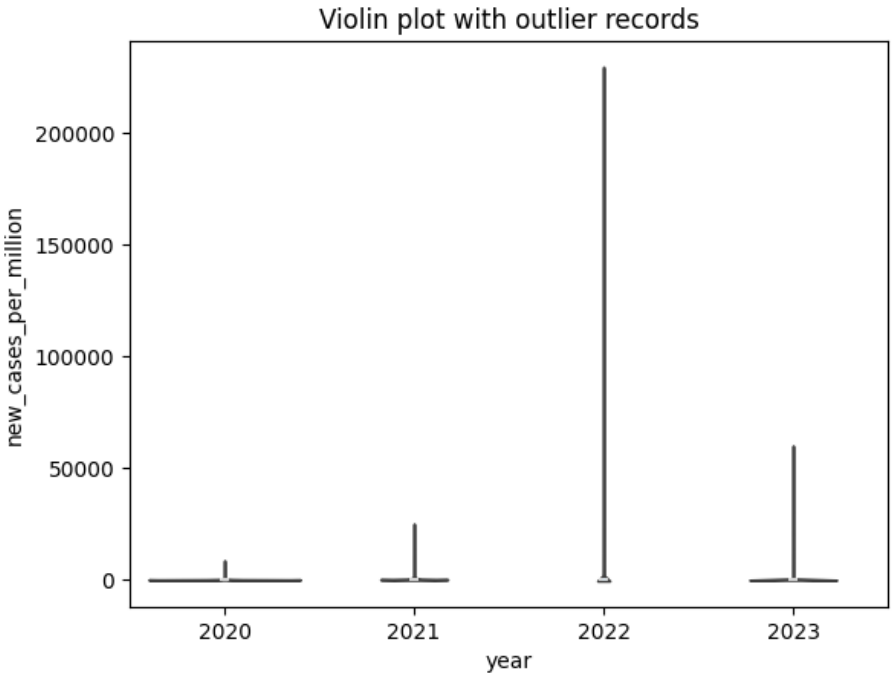
Figure 4.2

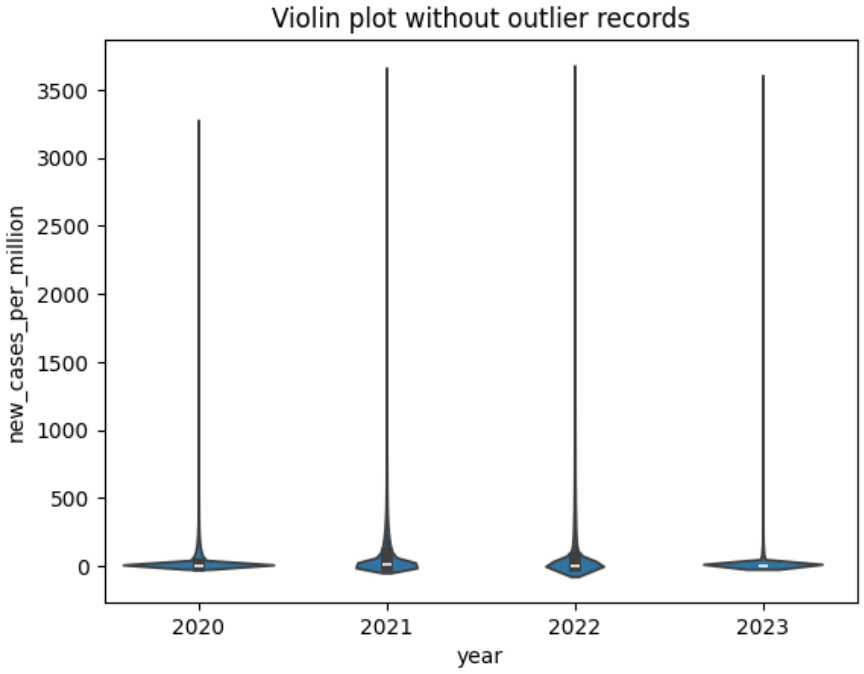
These two plots depict the numbers of new COVID-19 cases per million people.

The Major difference between the two plots is the fact that one is using a dataset with outlier data, and the other is using the same dataset but with some of the outlier data dropped.

Despite the fact that the distribution is still quite skewed, the effect of outlier data is obvious.

Regardless, both plots display that over this time period, the average number of new COVID-19 cases per record is smaller than 100.





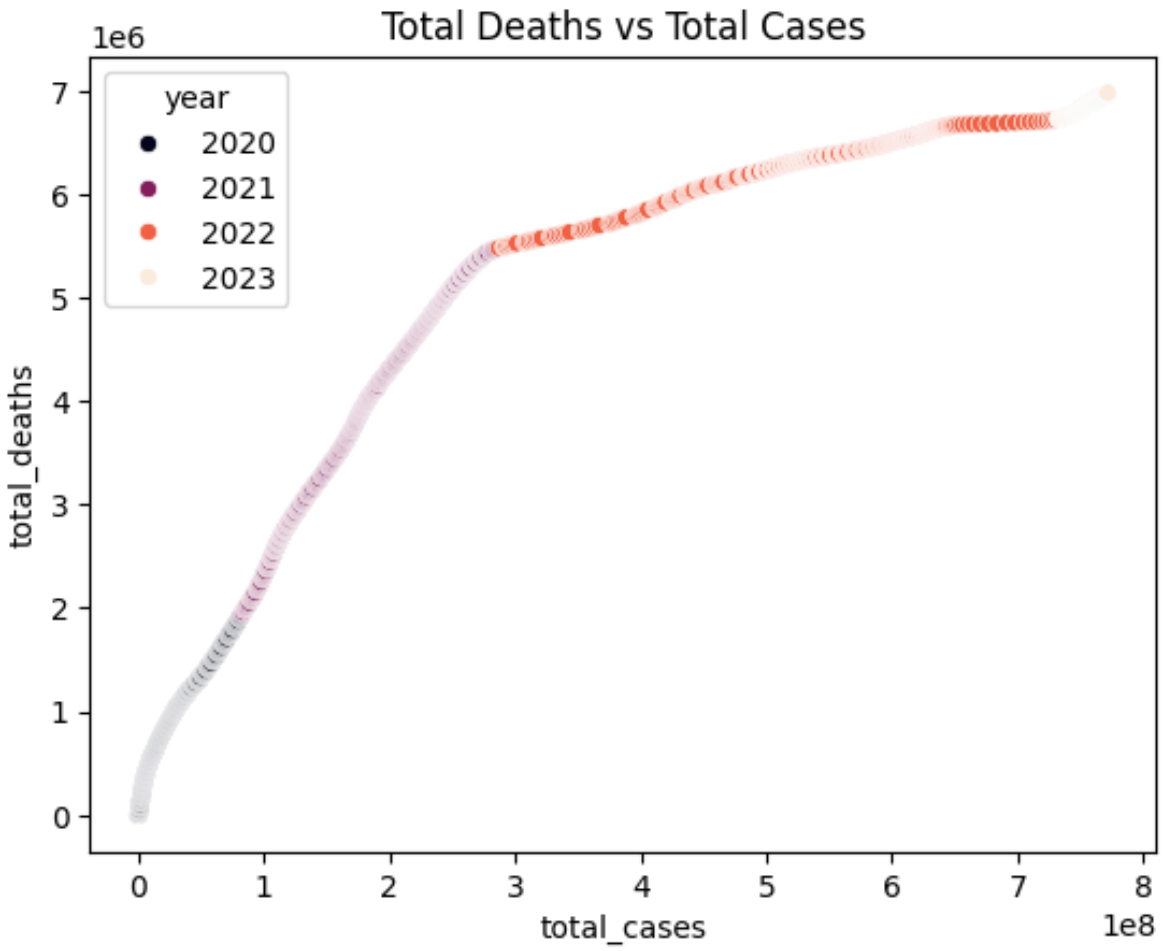


Figure 4.3

This scatter plot depicts the relationship between the total number of COVID-19 cases and the total number of COVID-19 cases for each record.

There is a positive correlation between the two attributes, one that can be displayed using a piece-wise function. Based on the plot, in early 2022 and once a large portion of the population had been vaccinated, the mortality rate of COVID-19 decreased.

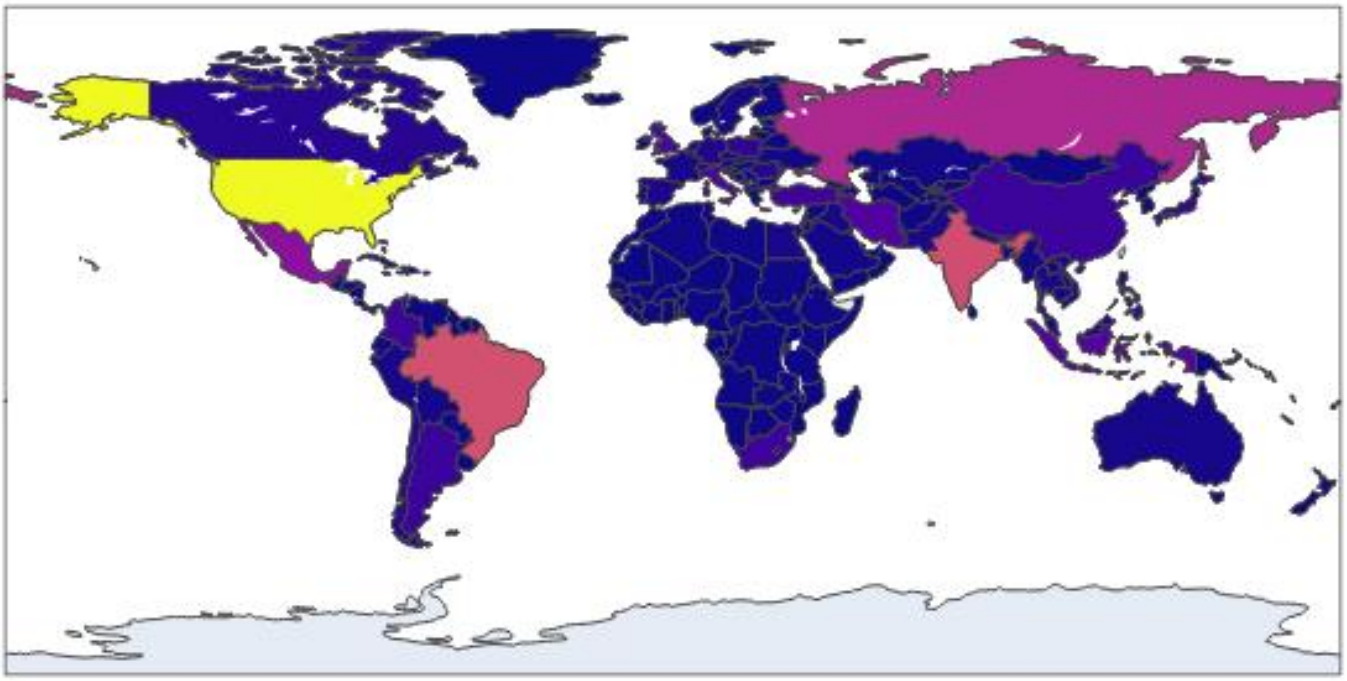


Figure 4.4: Total deaths based on country

This figure depicts the final number of deaths caused by COVID-19 as of November 2023. The Choropleth graph displays each country's total death count with a color.

This data can be useful to determine which countries had the highest share of the total deaths in the world, but cannot be used to compare countries to each other as it is not normalized.

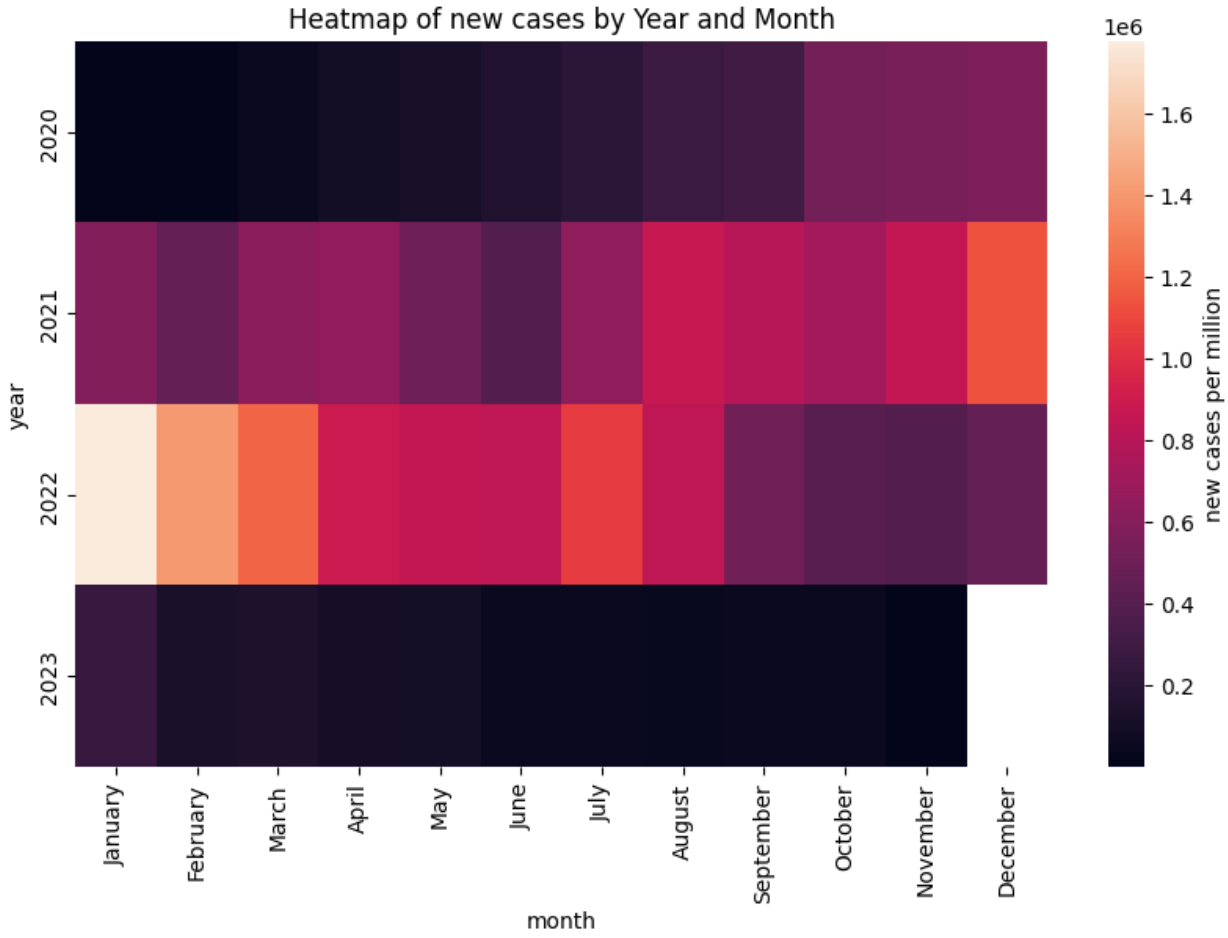


Figure 4.5

The heatmap plot depicts the severity of the pandemic over time in an easily understandable way. The number of new cases per million people in the world is displayed here, showing how it started off with a slow start, reached its peak in early 2022, and it is almost eradicated by the end of 2023.

Figure 4.6

The correlation heatmap displaying the correlation between all numerical values in the dataset.