**Project Report**

|  |  |
| --- | --- |
| **Course Name (WSQ)** | **Advanced Certificate in Data Science** |
| Product Name (Marketing &  Sales) | **Advanced Certificate in Data Science** |
| **Module Name (WSQ)** | **WSQ- Data Analysis and Visualization (SF)** |
| Product Name (Marketing &  Sales) | **WSQ- Data Analysis and Visualization (SF)** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Student name** |  |  | **Assessor name** |
| Toh Yi Ting |  |  | Muhammad Nasrullah Khairi |
| **Date issued** | **Completion date** | | **Submitted on** |
| 23 May 2024 | 5 June 2024 | | 5 June 2024 |
|  |  |  | |
| **Project title** | **Worldwide COVID-19 Data Analysis** | | |

|  |
| --- |
| **Learner declaration** |
|  |
| I certify that the work submitted for this assignment is my own and research sources are fully acknowledged.        Student signature: Date: 5 June 2024 |
|  |

**Index**

1. Project Overview--------------------------------------------------------------------------------Page 2

2. Project Technical Environment--------------------------------------------------------------Page 3

3. Activity 1------------------------------------------------------------------------------------------Pages 4-7

4. Activity 2------------------------------------------------------------------------------------------Pages 8-9

5. Activity 3------------------------------------------------------------------------------------------Pages 10-13

6. Activity 4------------------------------------------------------------------------------------------Pages 14-16

7. Activity 5------------------------------------------------------------------------------------------Pages 17-21

8. Activity 6------------------------------------------------------------------------------------------Pages 22-25

9. Activity 7------------------------------------------------------------------------------------------Pages 26-29

10. Conclusion and Future Improvements---------------------------------------------------Pages 30

1. **Project Overview:** Describe the Project along with Project Outcomes (Explain the Project in your own words in 15 – 20 lines)

This project is undertaken with the premise of ABC Health Analytics Company recognising the critical need for accurate and timely health data analysis following the global COVID-19 pandemic. There is an urgent need to use data to understand and respond to the evolving challenges caused by the pandemic. Hence, this project will be initiated for comprehensive data analysis and visualization of COVID-19 datasets using Python.

The goal of project is to perform a comprehensive analysis of worldwide COVID19 data. The dataset used for this data analysis is sourced from Our World in Data and contains information on various COVID-19 metrics across different countries. I will use data visualization techniques from Python libraries such as pandas, matplotlib, and seaborn to understand, analyse and visualise the data.

I will perform the tasks listed in the Project Brief under Project Task List and implement the tasks in Python. At the end of the project, the Python .ipynb notebook will be submitted as part of the Project Implementation Tasks. I will also prepare this Project Report with screenshots of the graphs and adding on my interpretations of the data. At the end of the report, I will give some recommendations for improvements should similar projects be conducted again.

1. **Project Technical Environment:**

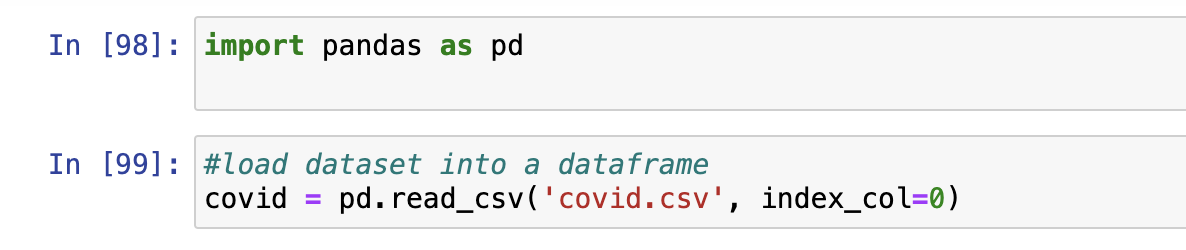
I will be completing the tasks in a Jupyter Notebook environment using the Python script. I will also be using the following Python libraries:

* Pandas: For data manipulation, for extracting, transforming and loading the dataset
* Matplotlib: For plotting graphs for visual analysis
* Seaborn: For plotting graphs for visual analysis

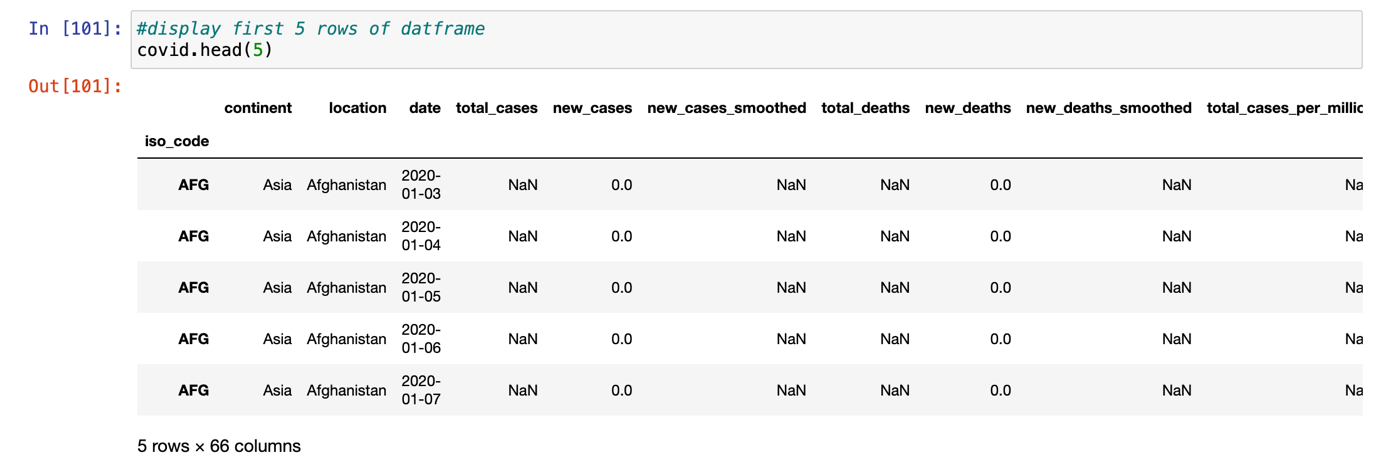
1. **Activity 1:** List the problem and screenshot for the same.

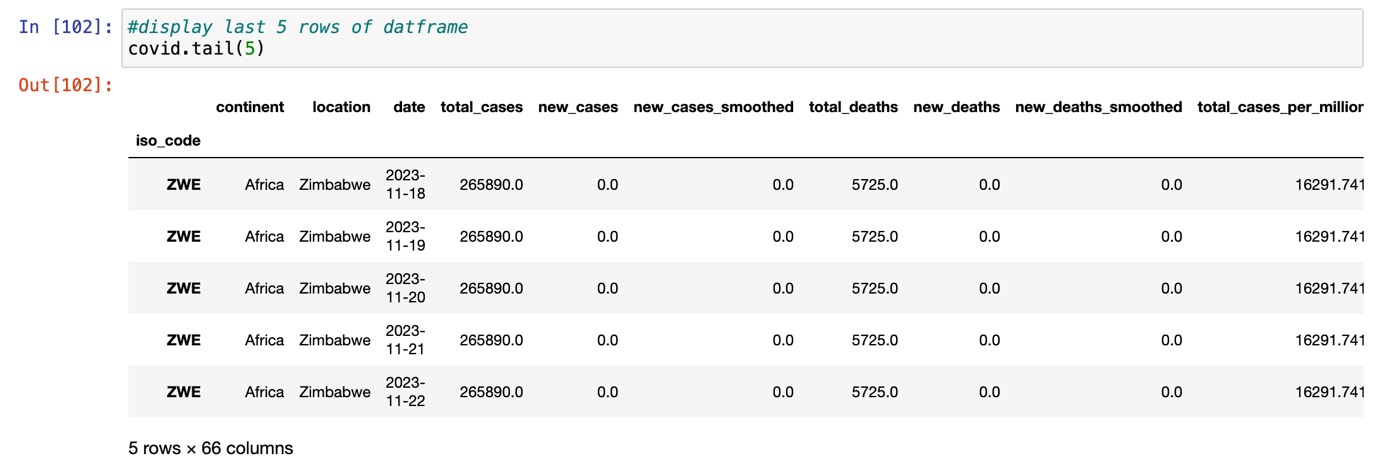
**Problem(Data Loading and Exploration): To load and explore the Covid-19 dataset, and to ensure data integrity and format consistency for analysis later.**

**1. Load the COVID-19 dataset using pandas from the provided dataset.**



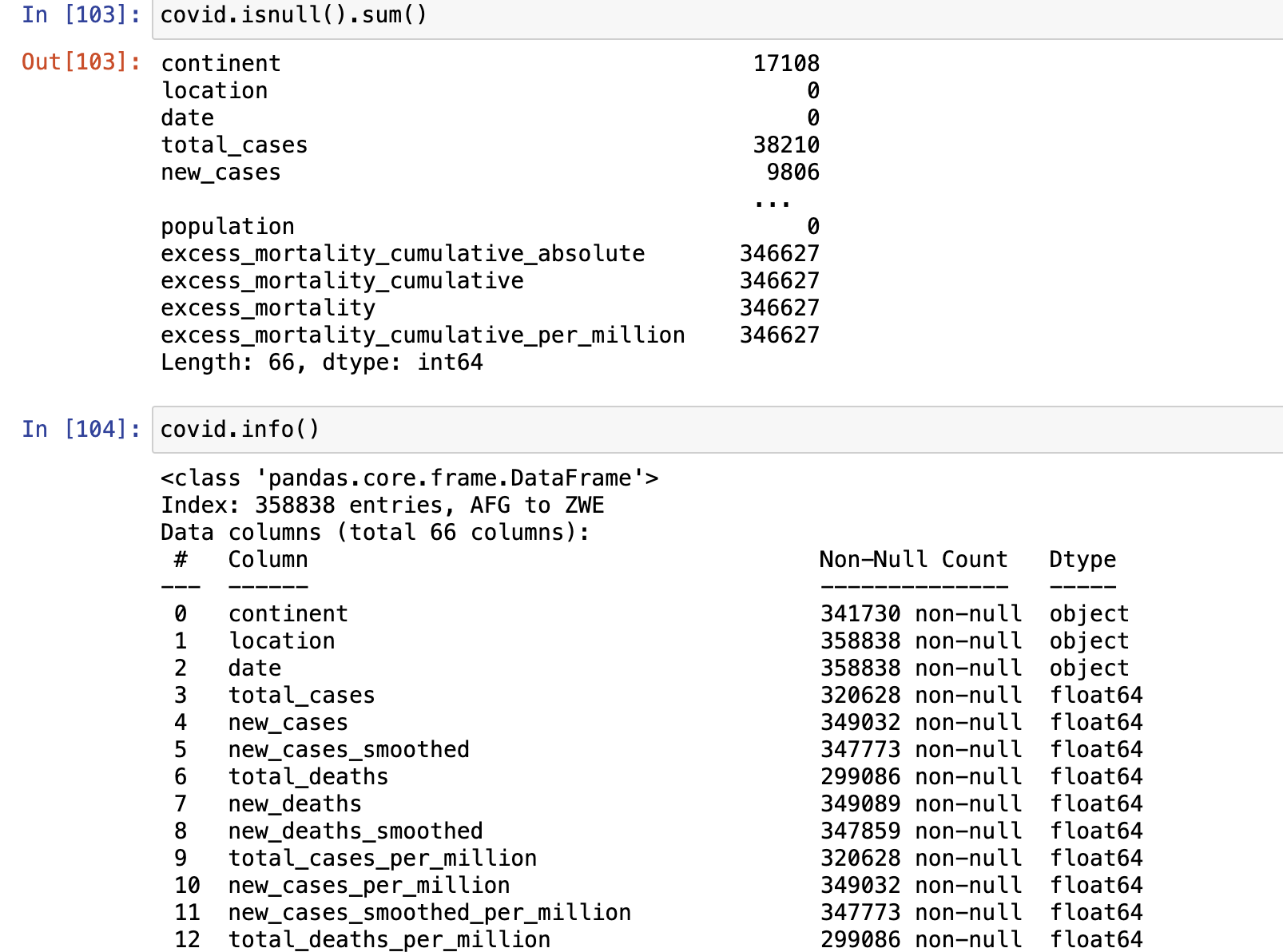
**2. Display the first 5 and last 5 rows of the DataFrame.**

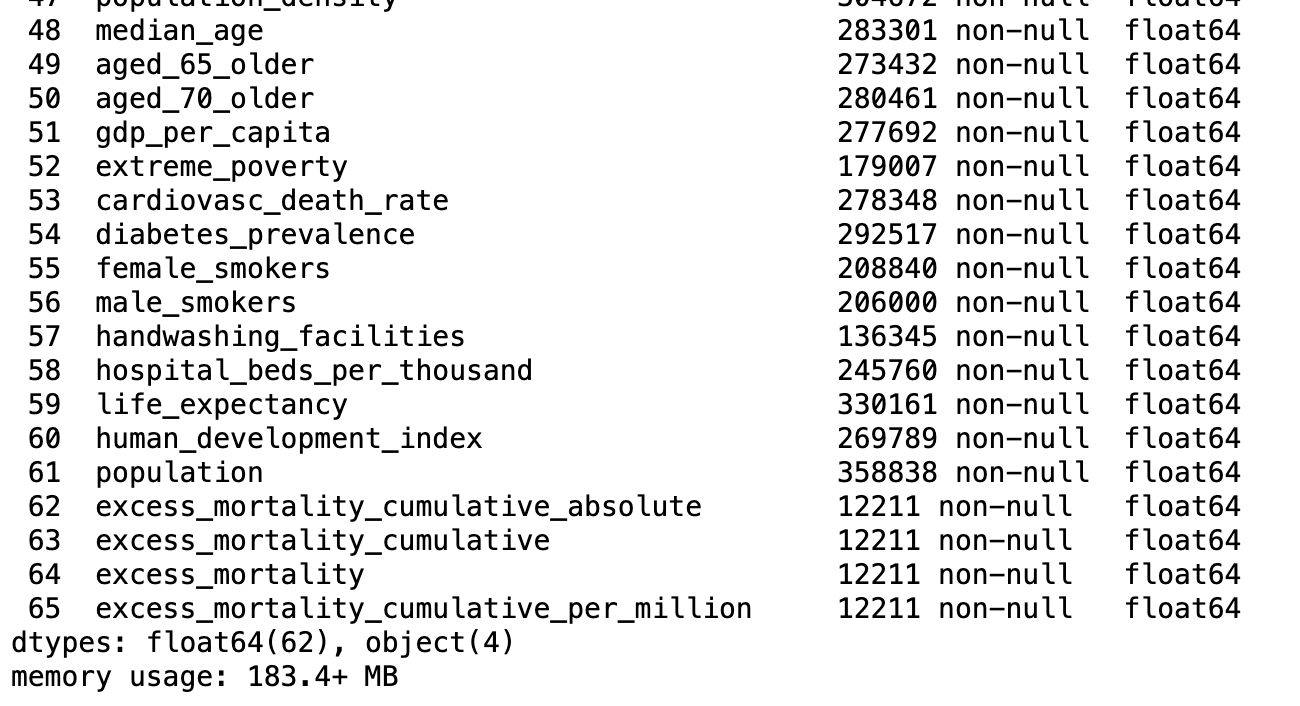
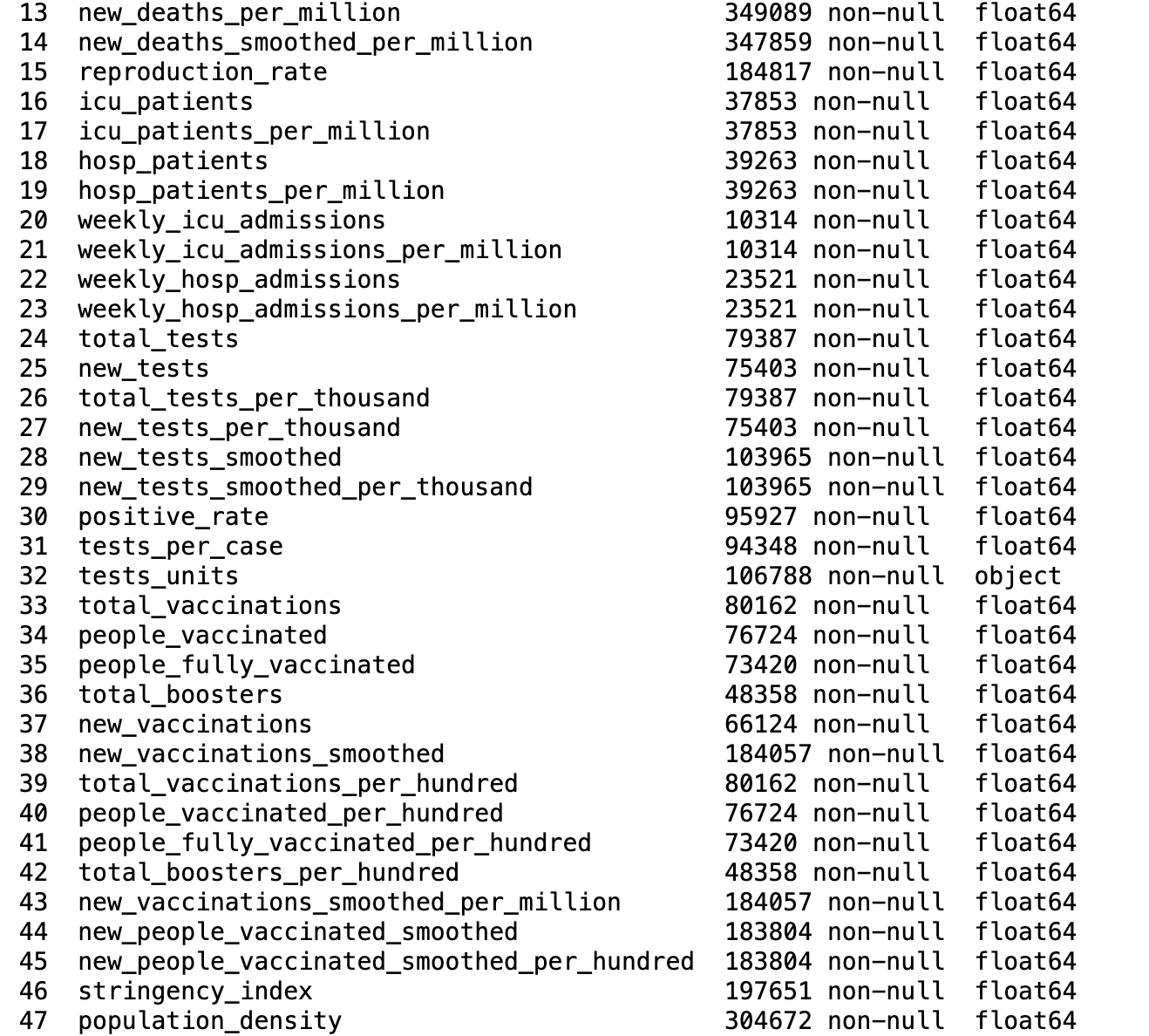




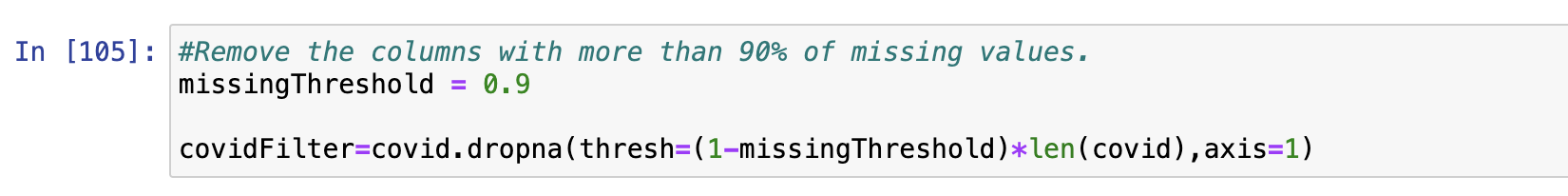
1. **Check for missing values in the dataset and decide on a strategy to handle them.**

There are too many columns/features in the dataframe, however, not all are necessary for the analysis. I will select the relevant columns/features to create another dataframe with only features needed for the analysis later. I will also fill in median values for numerical columns and fill in null values as ‘unknown’ for categorial columns.

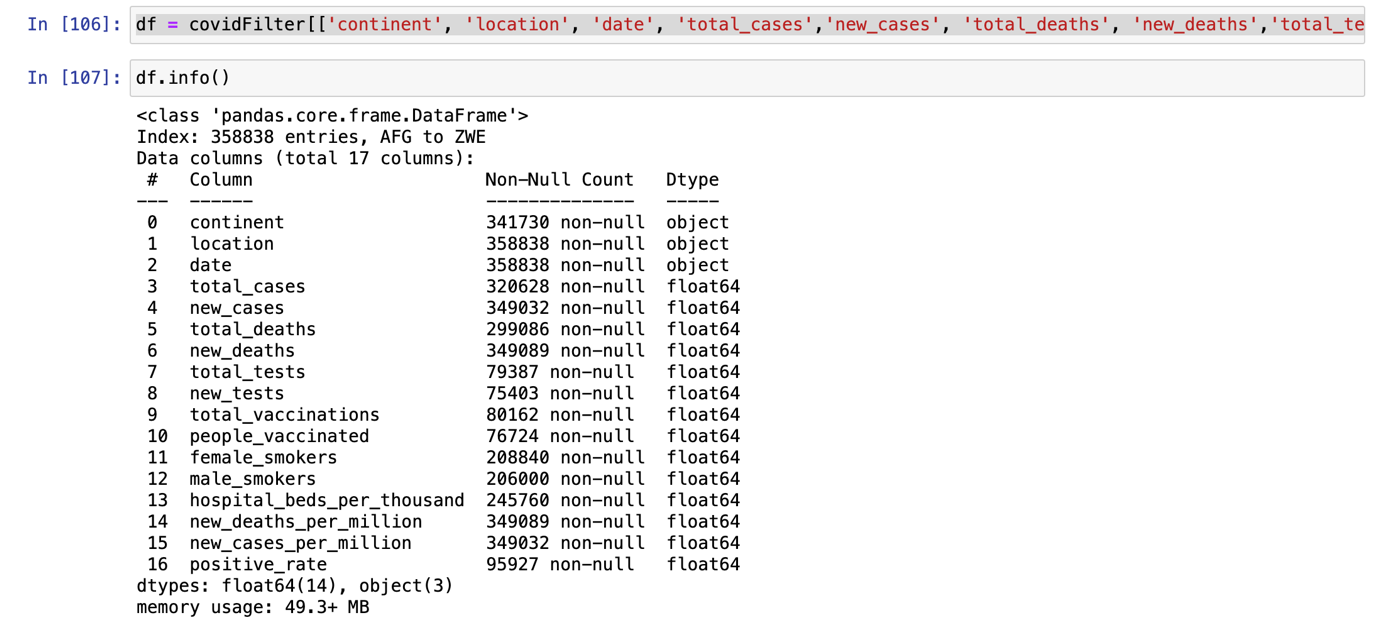




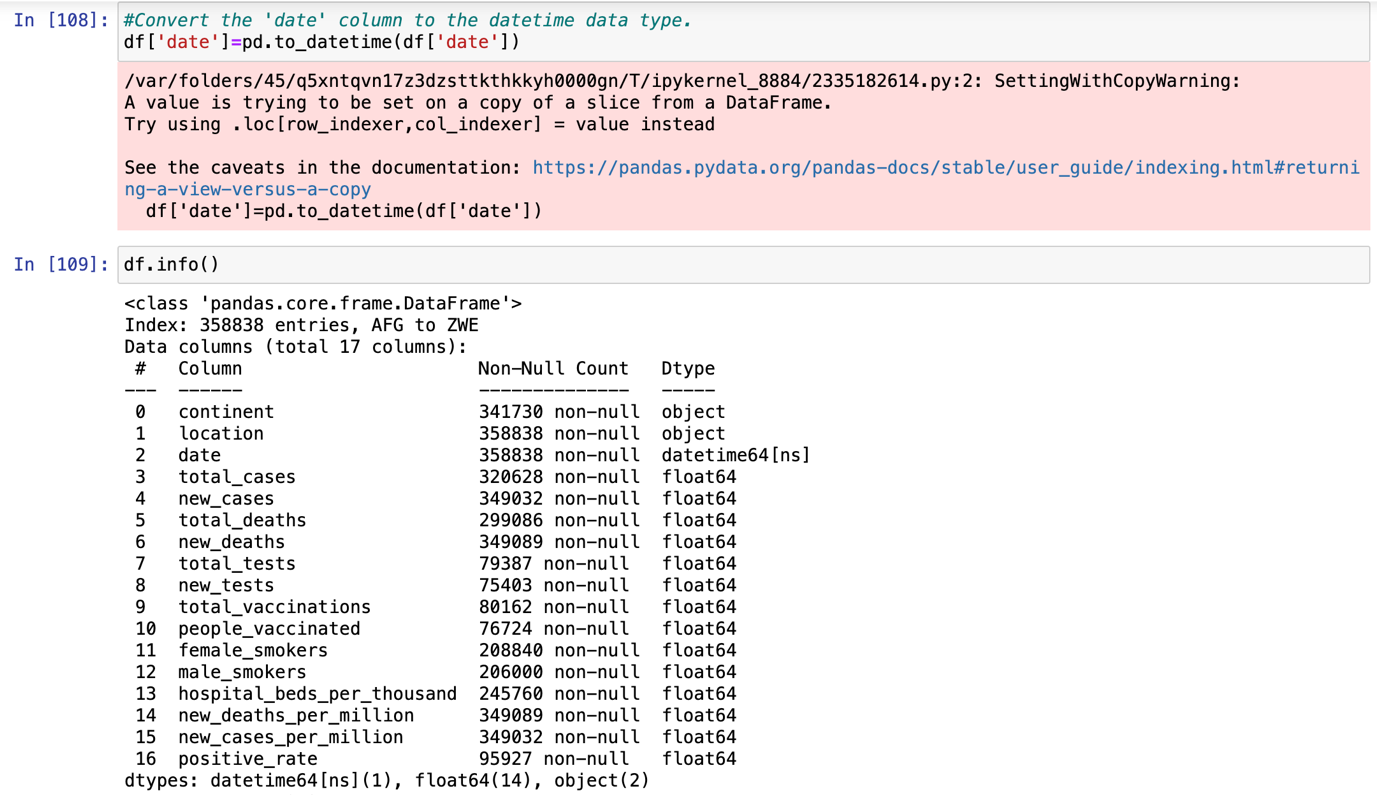
**4. Remove the columns with more than 90% of missing values.**



Creating another dataframe with the required features for analysis later:



**5. Convert the 'date' column to the datetime data type.**



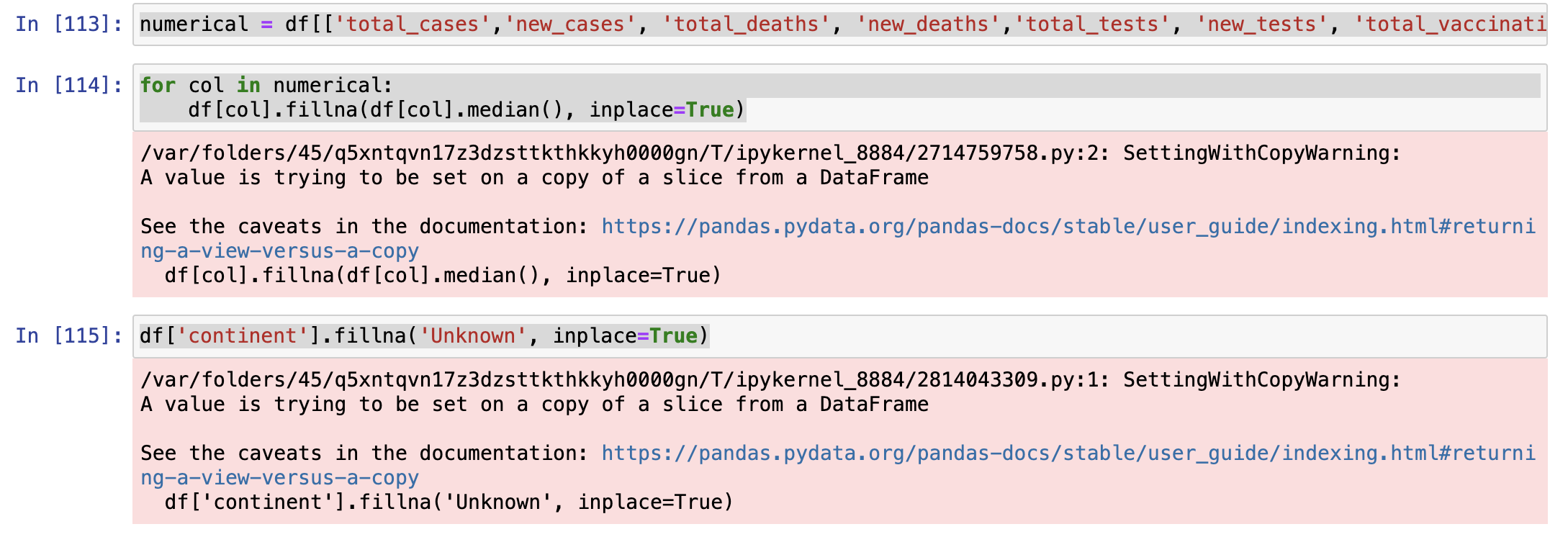
This new dataframe ‘df’ will be used as the main dataframe for further data and graph analysis later.

1. **Activity 2:** List the problem and screenshot for the same.

**Problem (Data Cleaning and Feature Engineering): To further enhance data relevance and structure by cleaning the dataset and creating new features, thereby enabling focused analysis on key COVID-19 indicators and demographic factors later.**

**1. Impute missing values in the dataset columns**

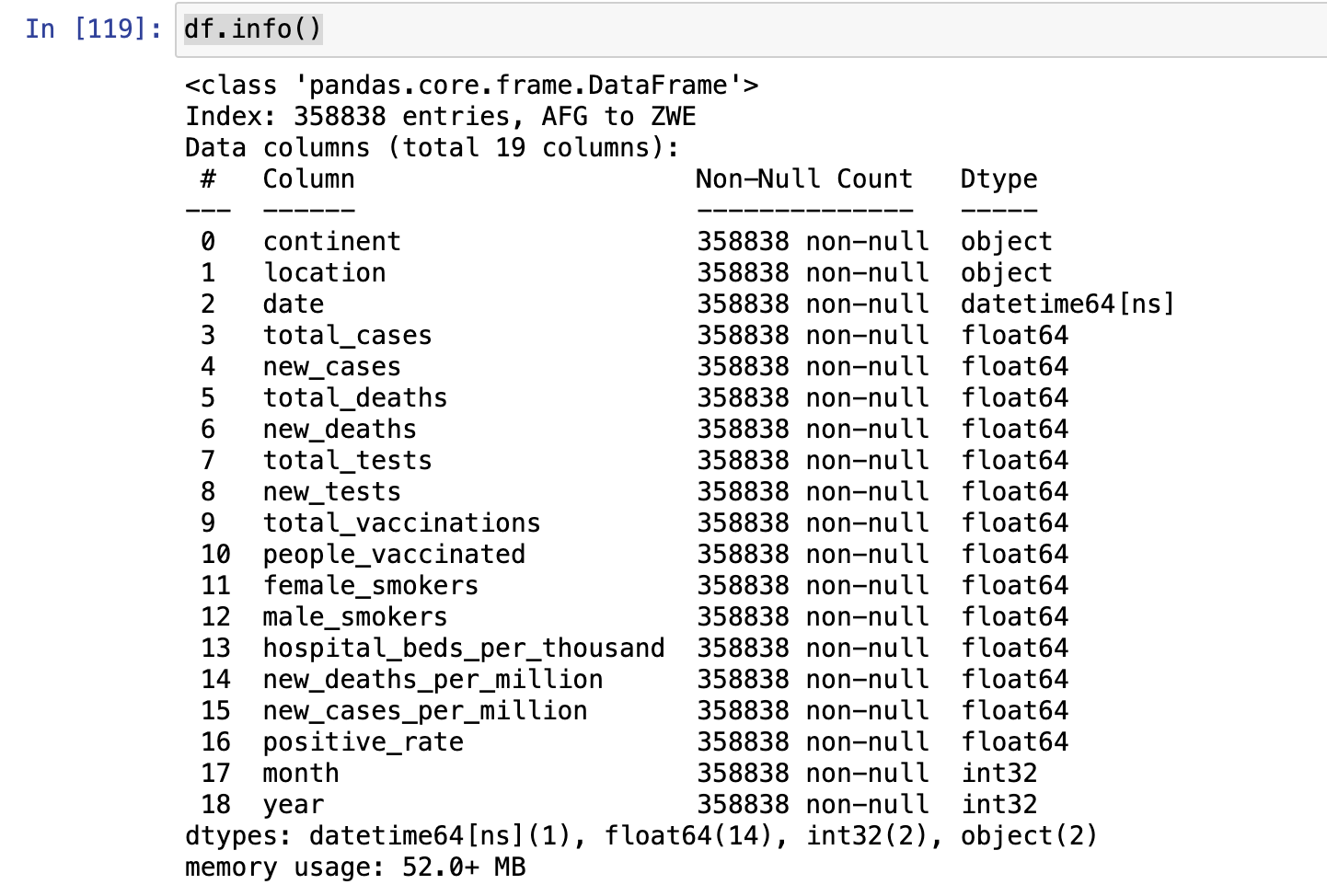
To fill in null values in numerical columns with median values and to replace null values in categorial column as ‘unknown’.



**2. Remove duplicate rows from the DataFrame.**

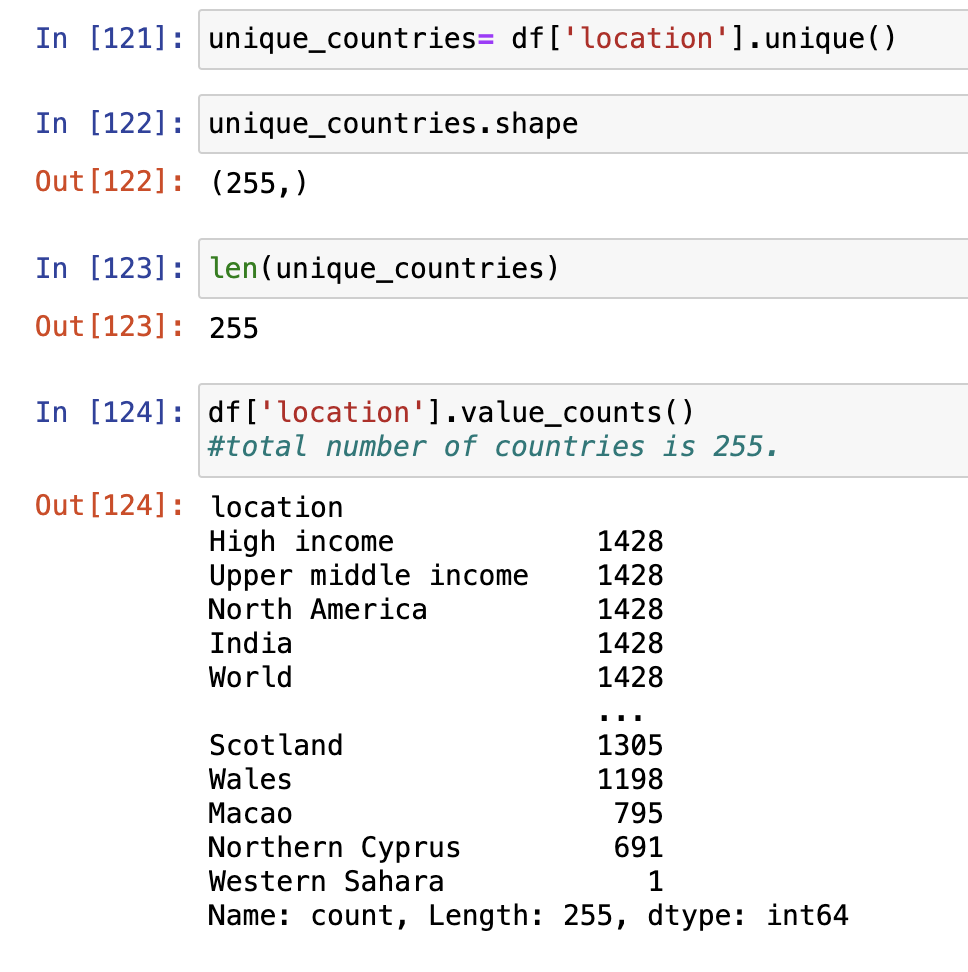


**3. Create new features if needed (e.g., extract year and month from the 'date' column).**



**4. Explore unique countries in the dataset and count the total number of countries.**

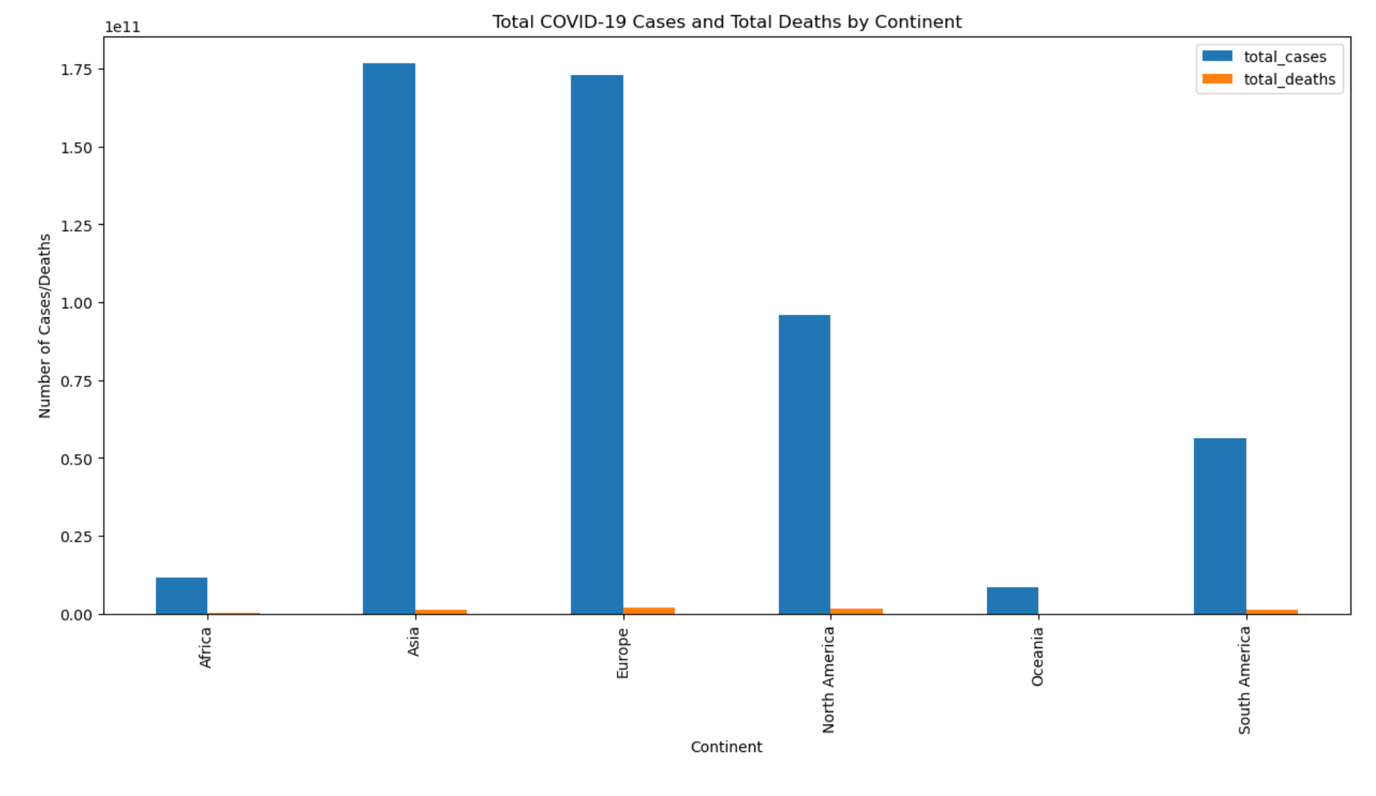
The total number of unique countries in the dataset is 255 countries.



1. **Activity 3:** List the problem and screenshot for the same.

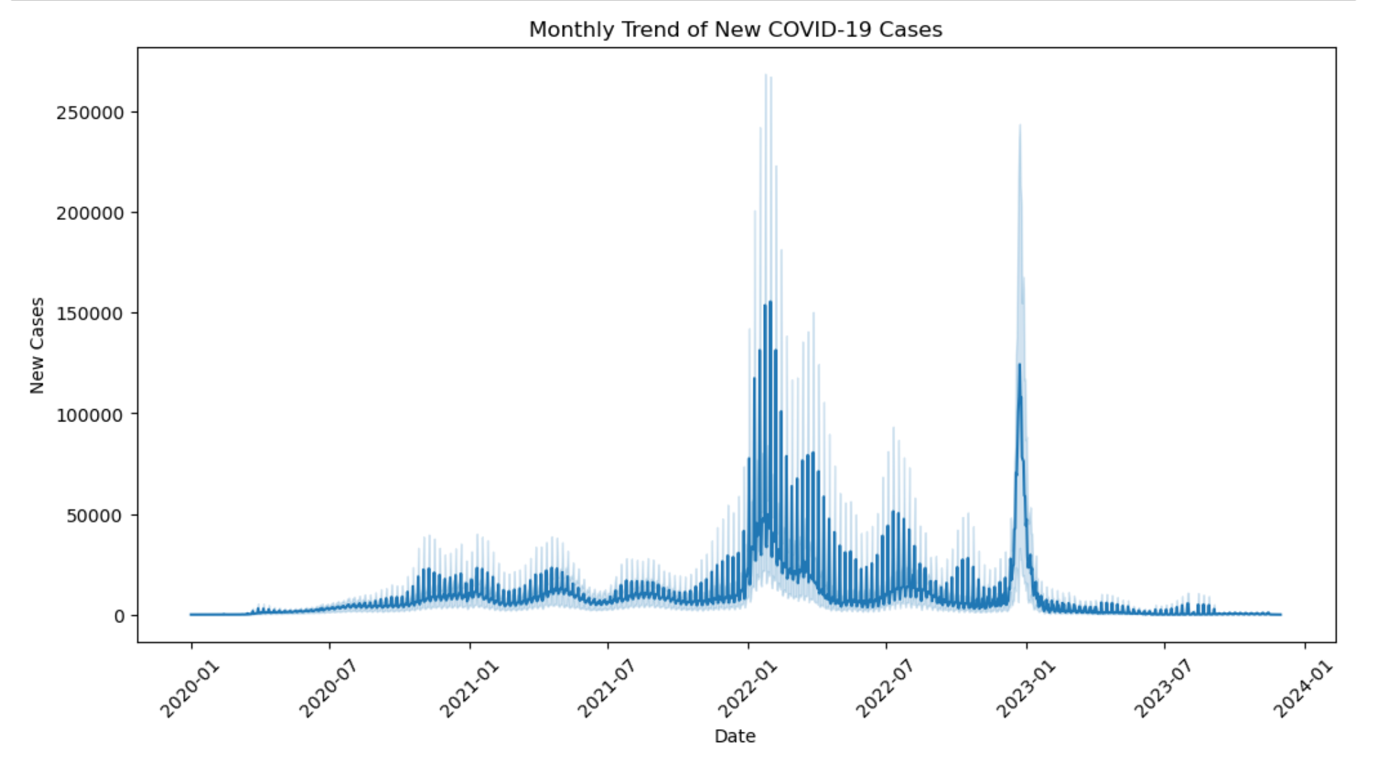
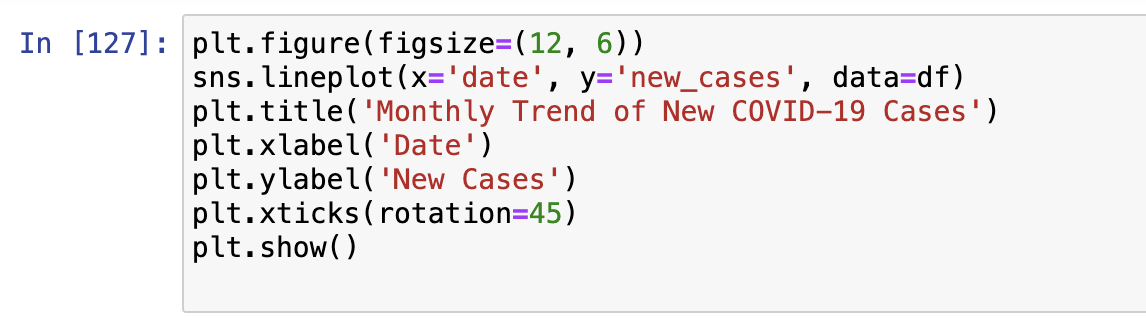
**Problem (Worldwide COVID-19 Overview): There is a need to identify patterns and correlations in the COVID-19 pandemic across a global and temporal scale. This is helpful for strategic decision-making. Analysis can be visualised using graphs.**

1. **Visualize the total COVID-19 cases and total deaths by continent using bar plots.**



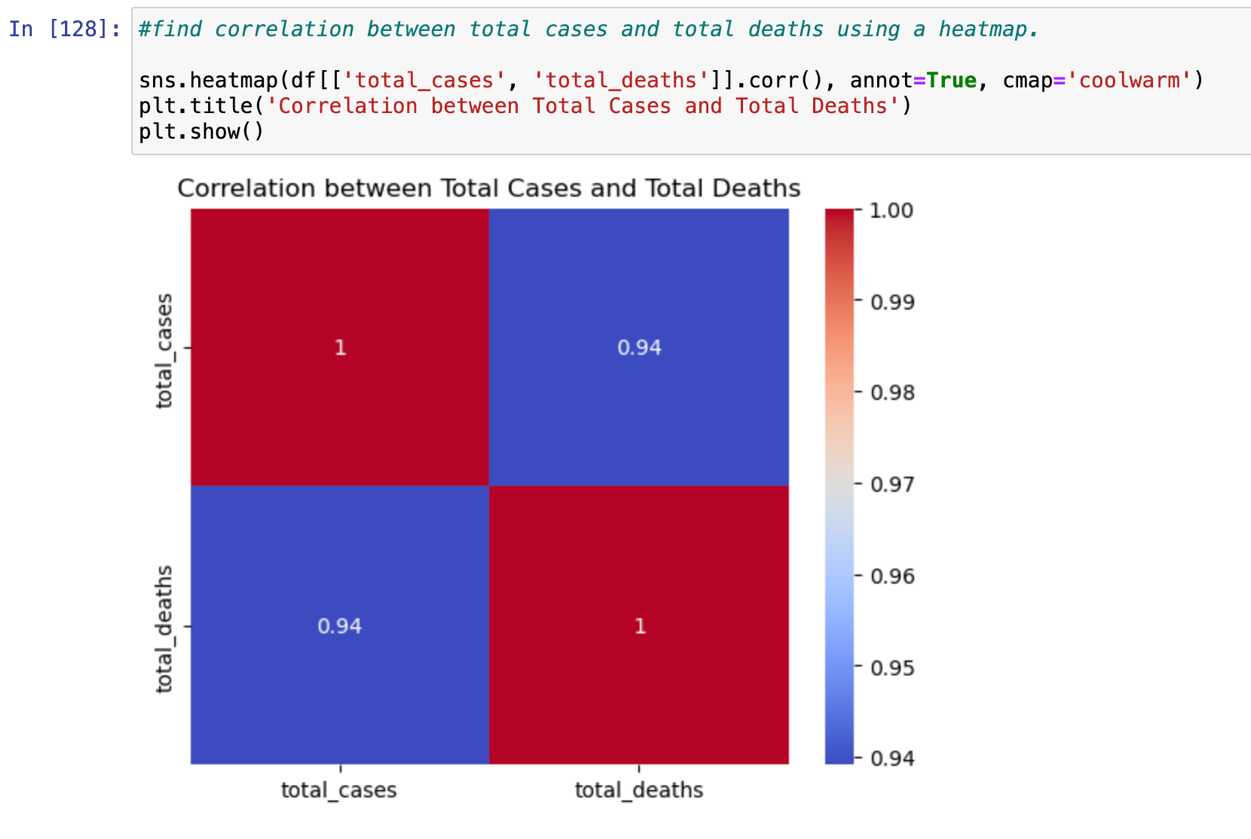
Asia has the highest total COVID-19 cases while Europe has the highest total deaths. The pandemic seems particularly serious in the Asian and Europe continents.

1. **Explore the monthly trend of new COVID-19 cases using a line plot.**

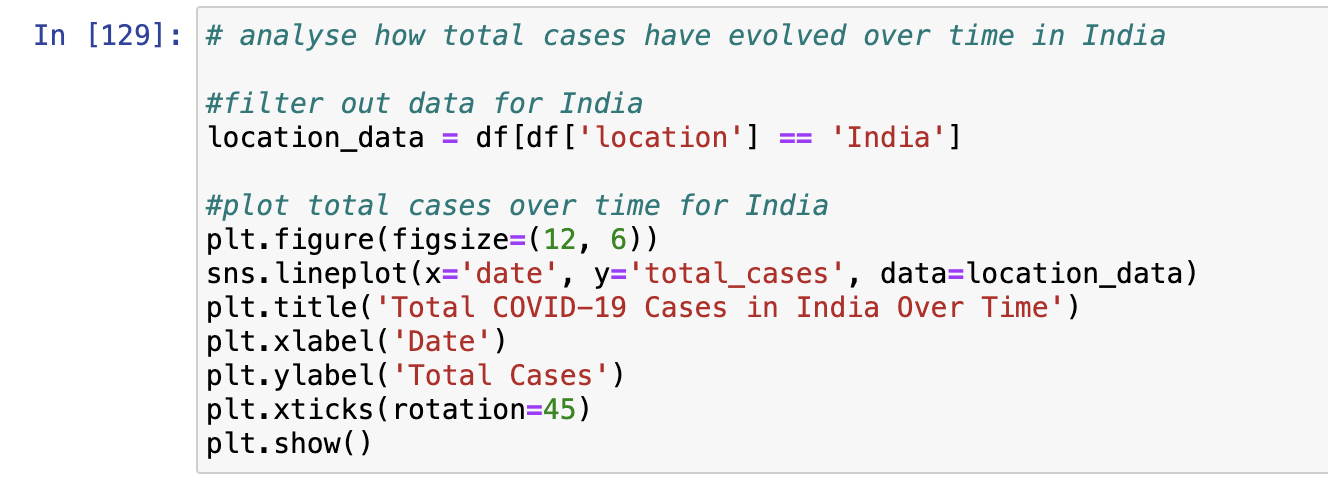


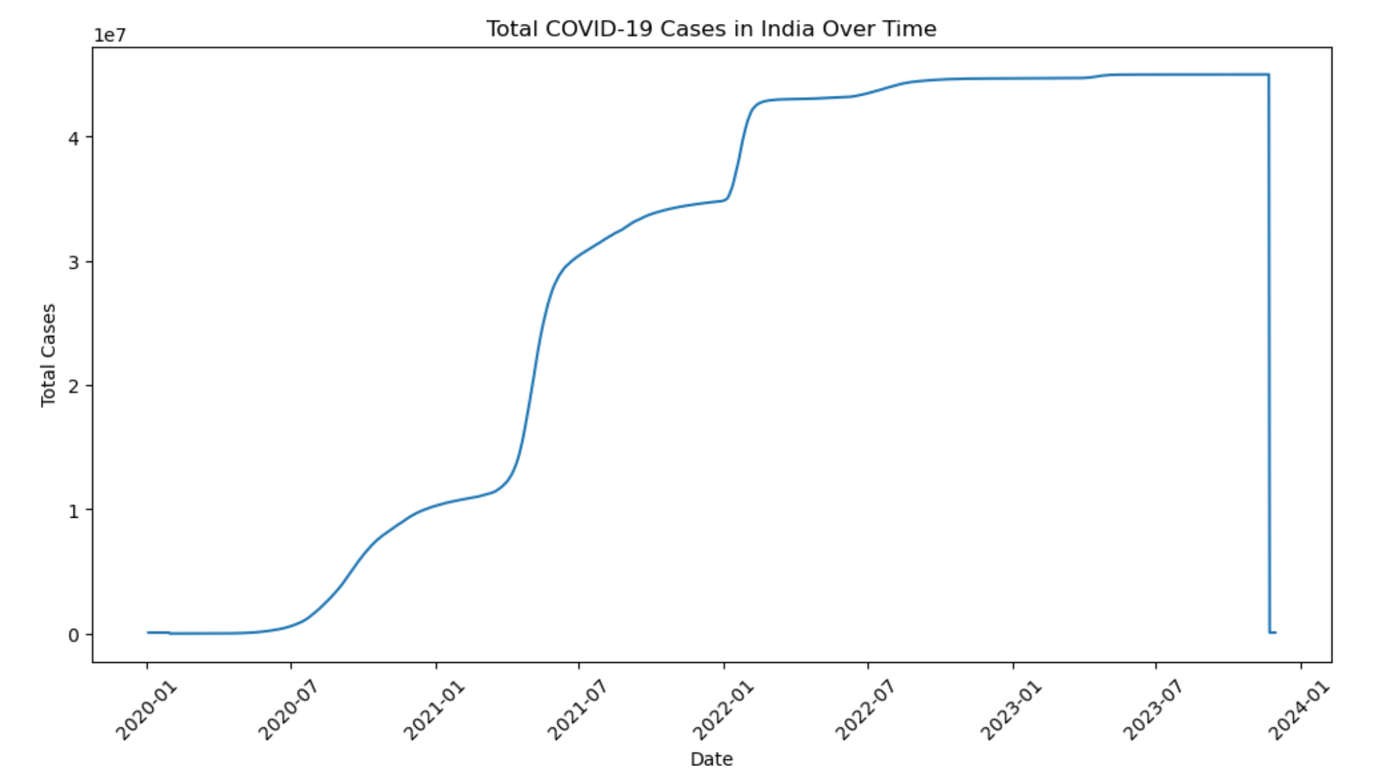
New cases seemed to spike at the start of the year in 2022 and later on again in 2023. There may be a seasonal aspect to the spread of COVID-19 or it may be due to mutations/variations of the COVID-19 viruses that can cause people to be infected/to be diagnosed with COVID-19 again at a second/third time. Air travel has opened up in the years 2022 and 2023 and that may also be a reason for the spread/increase in cases of COVID-19.

1. **Investigate the correlation between total cases and total deaths using a heatmap.**



There is a high and positive correlation between total cases and total deaths.

1. **Analyze how total cases have evolved over time for a specific location (e.g., India).**

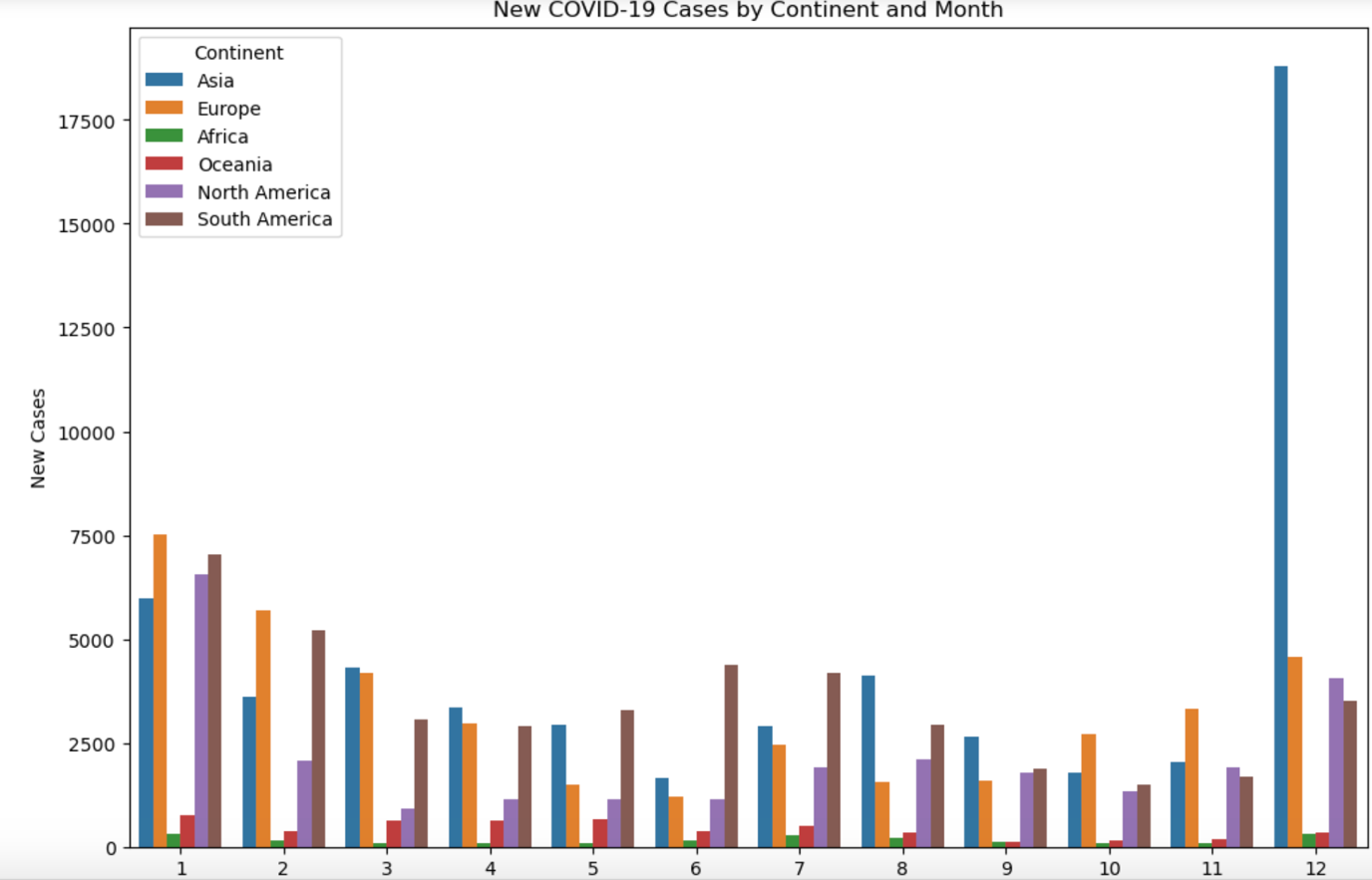
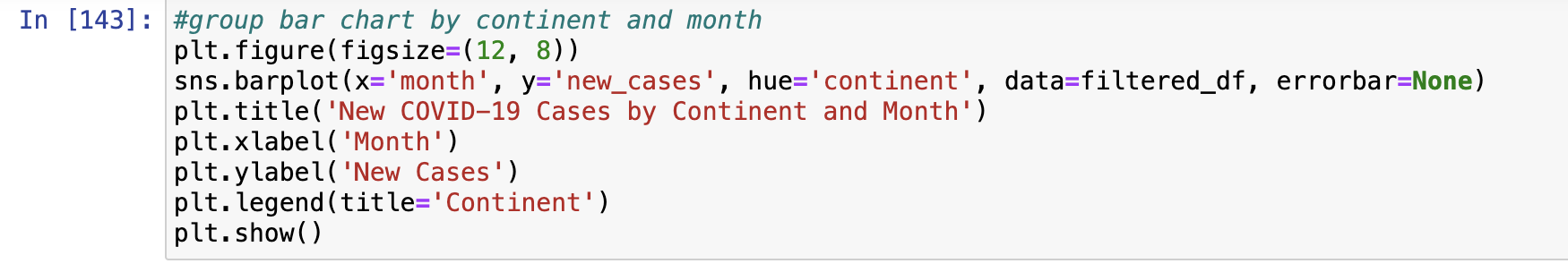


Total COVID-19 cases in India increased over time from 2020 and peaked from 2022 January to the end of 2023. There is also a sharp increase in cases in the early half of the year 2021 and another significant increase in cases at the start of 2022.

1. **Activity 4:** List the problem and screenshot for the same.

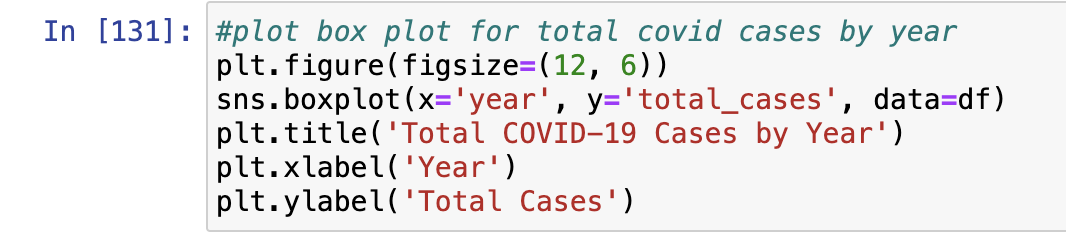
**Problem (Regional Analysis): To look at and identify patterns in Covid-19 cases and fatality across time and continents.**

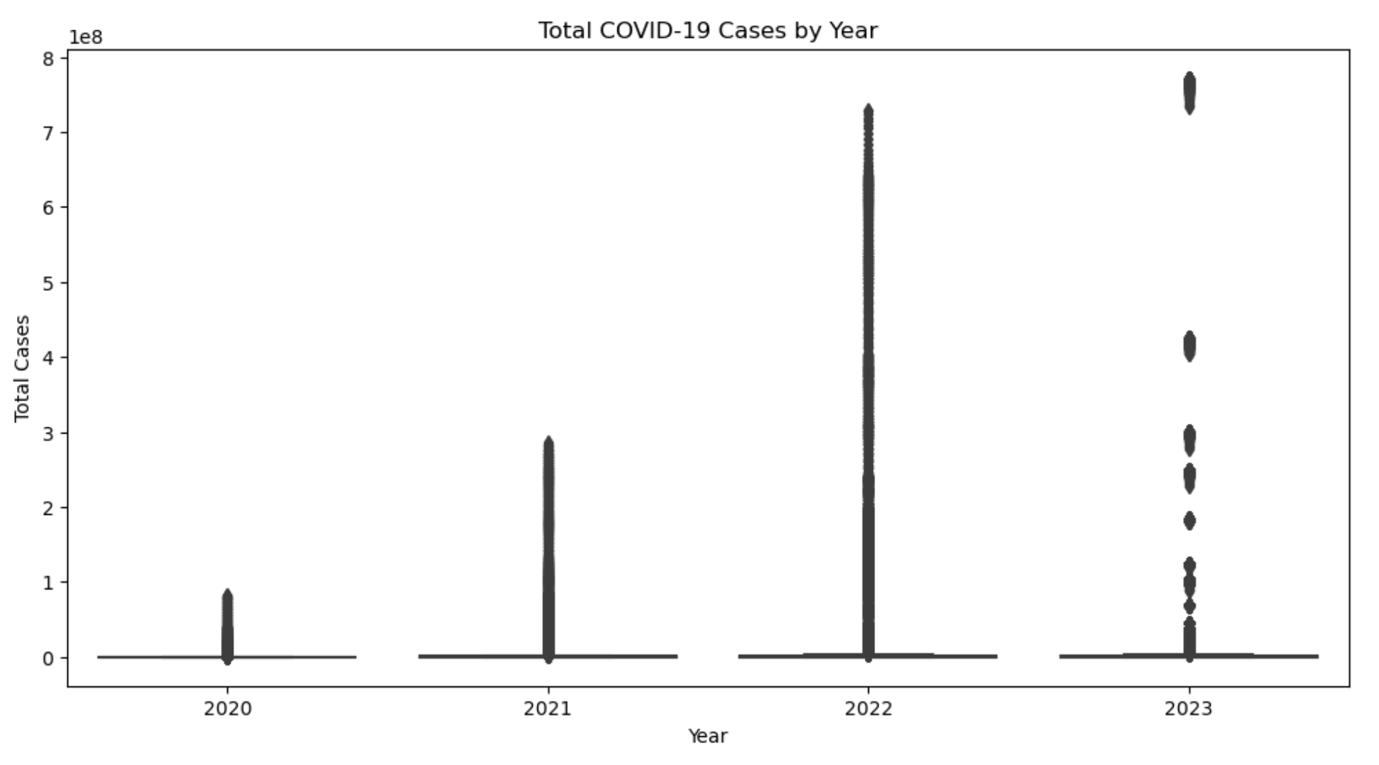
1. **Create a grouped bar chart to visualize new cases by continent and month.**



There is a sudden spike in new cases for the Asia continent in the month of December.

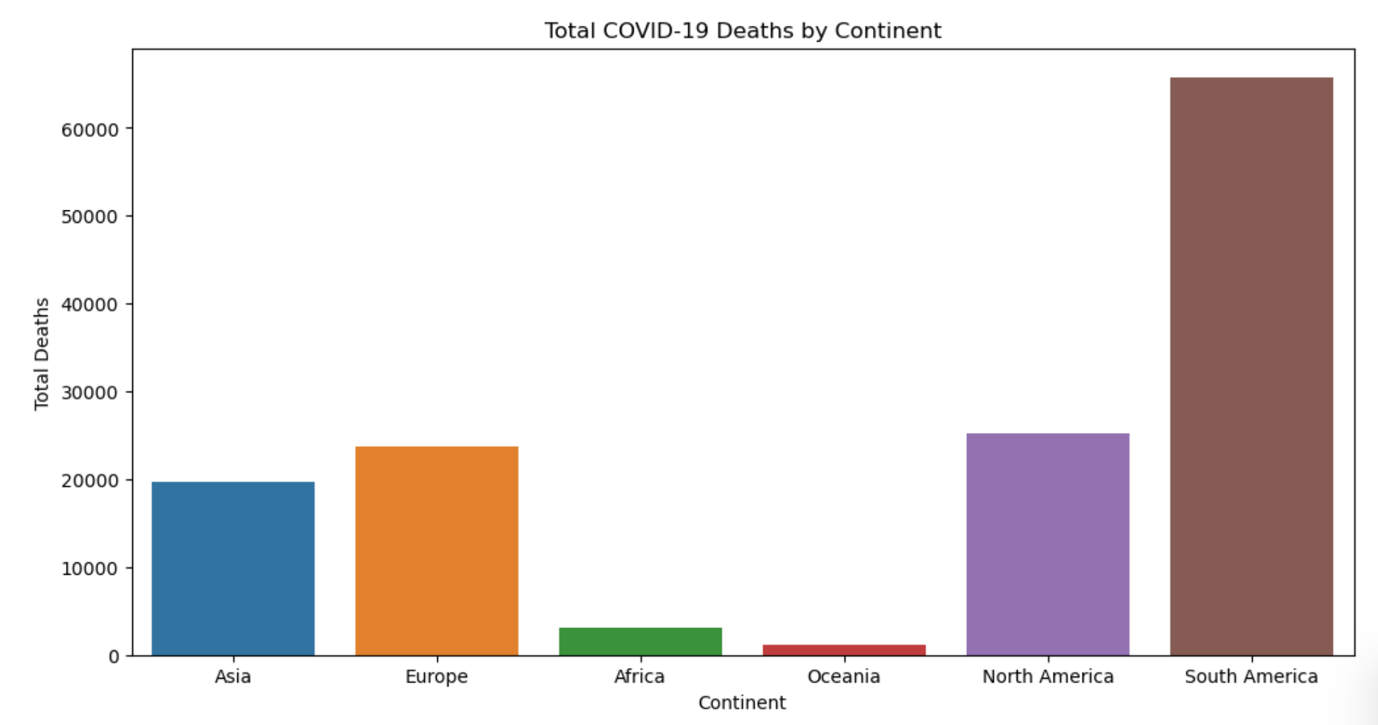
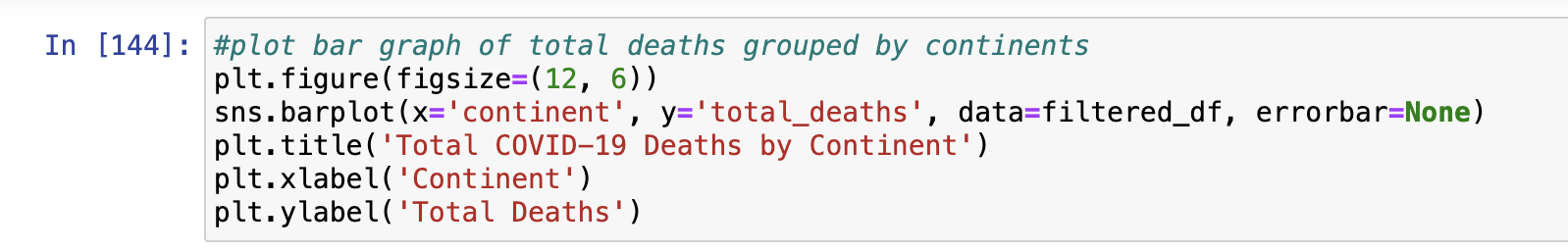
1. **Visualize the distribution of total COVID-19 cases by year using a box plot.**





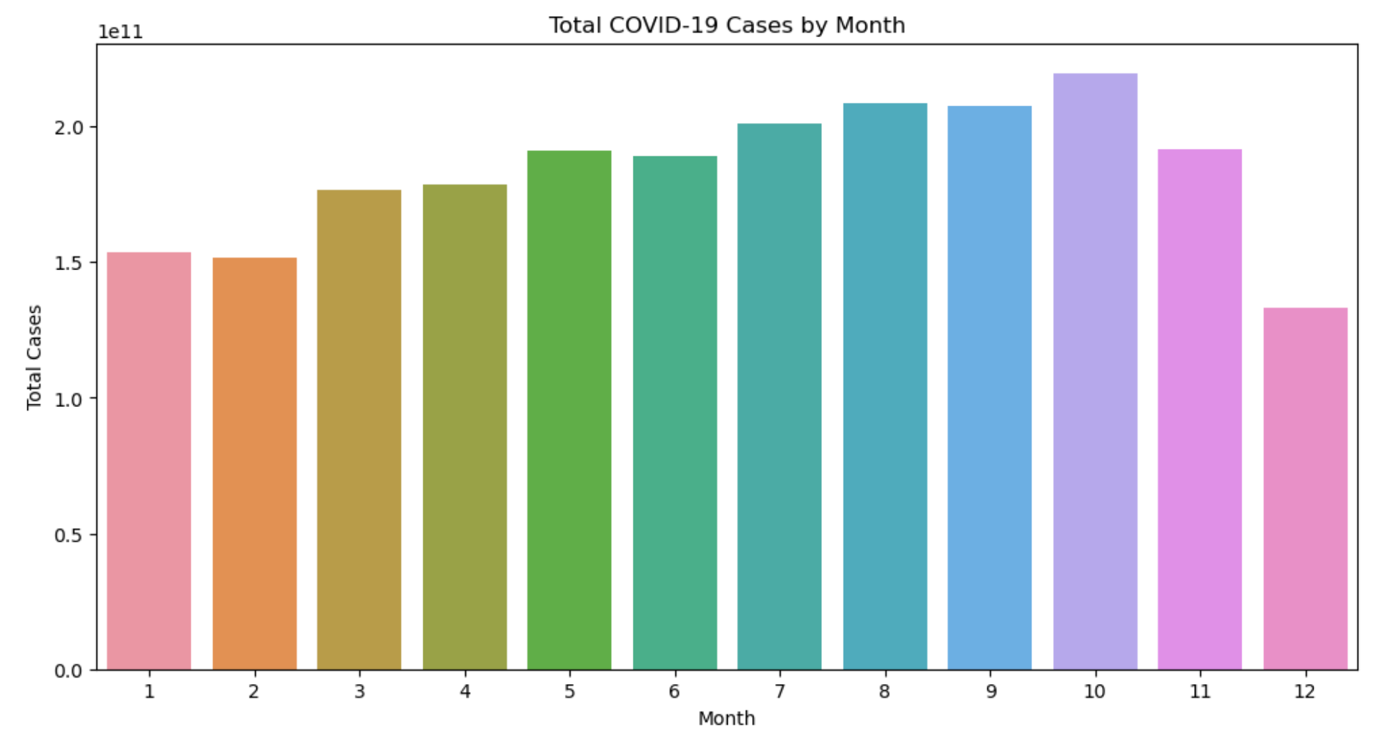
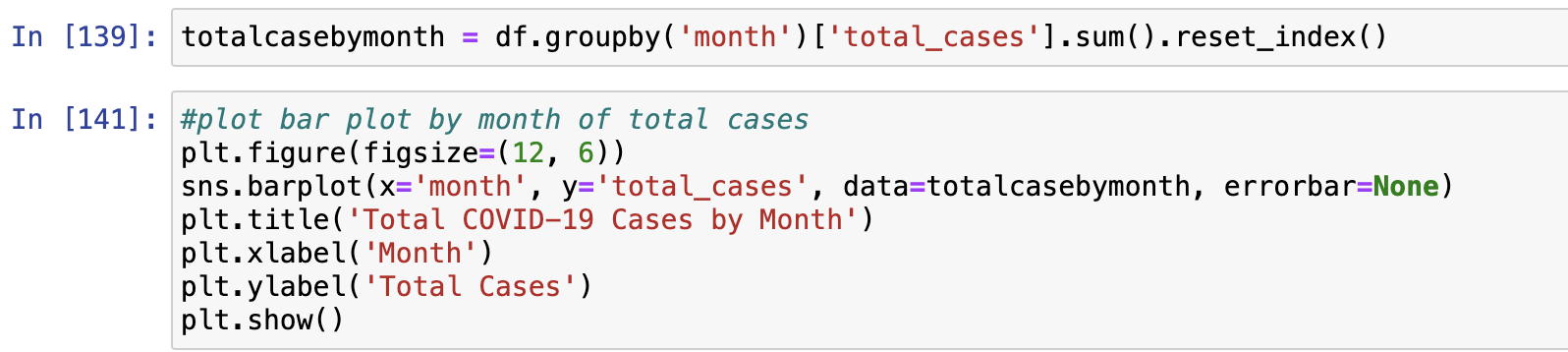
There is an increasing number of total cases from 2020 to 2023.

1. **Compare total deaths across different continents using a bar plot.**



South America has the highest total COVID-19 deaths while Oceania has the least total deaths.

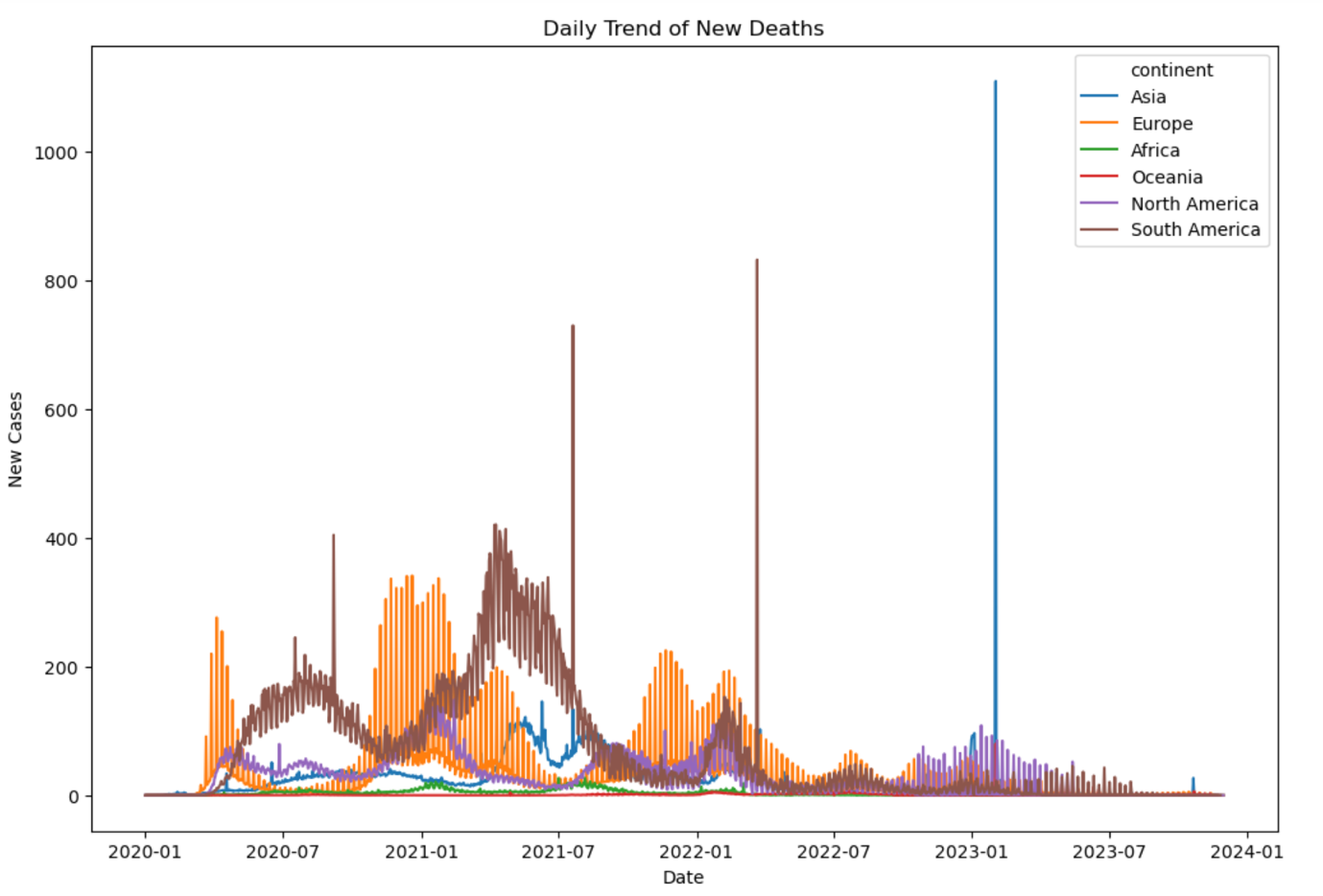
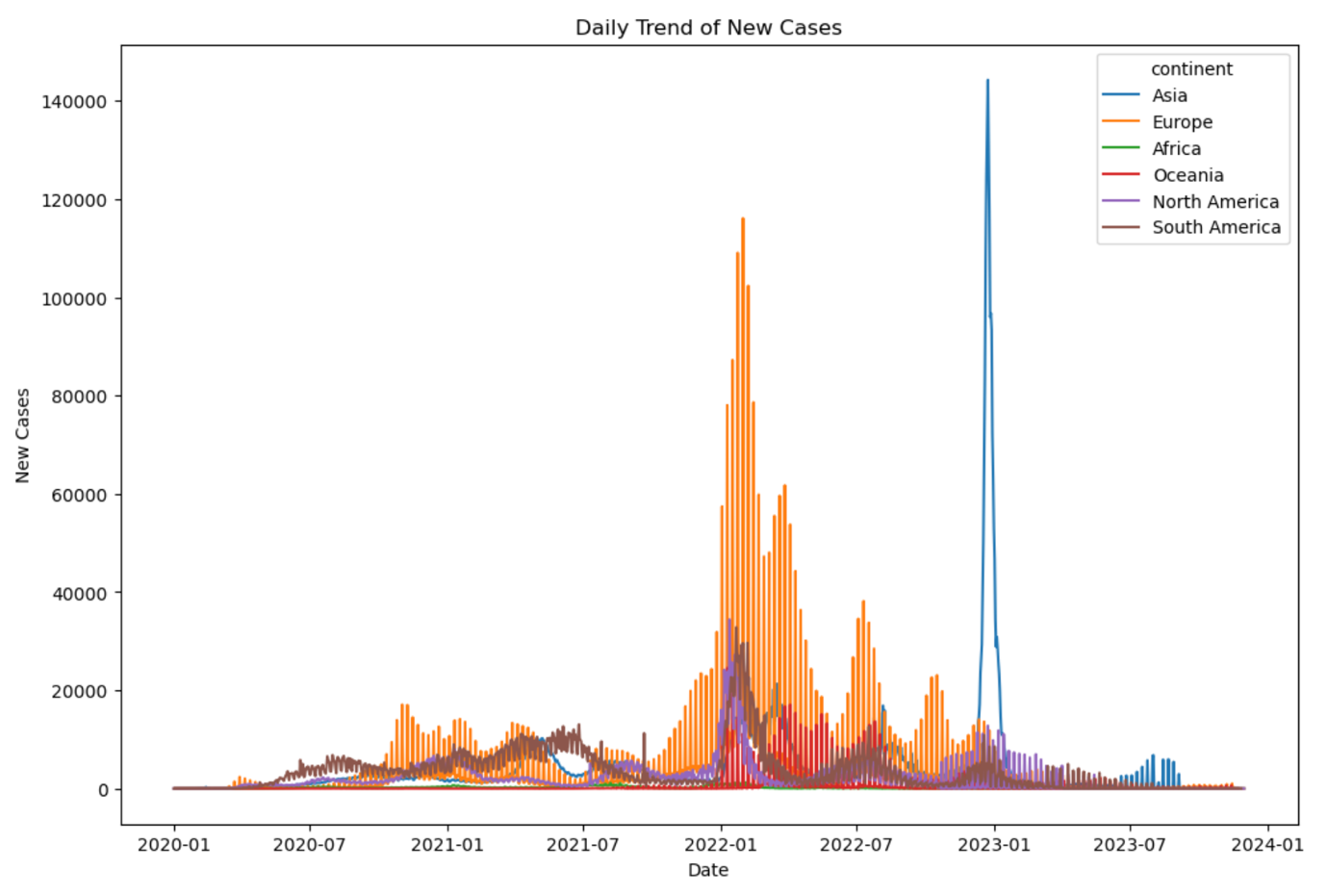
1. **Analyze the total cases on a month-by-month basis using a bar plot.**



COVID-19 cases showed an increasing trend from January to October before showing a decreasing trend in the months of November and December.

1. **Activity 5:** List the problem and screenshot for the same.

**Problem (Time Series Analysis): Using graphs to depict and capture the daily dynamics of COVID-19, assess the impact of vaccination, and monitor testing metrics for a comprehensive global overview across time.**

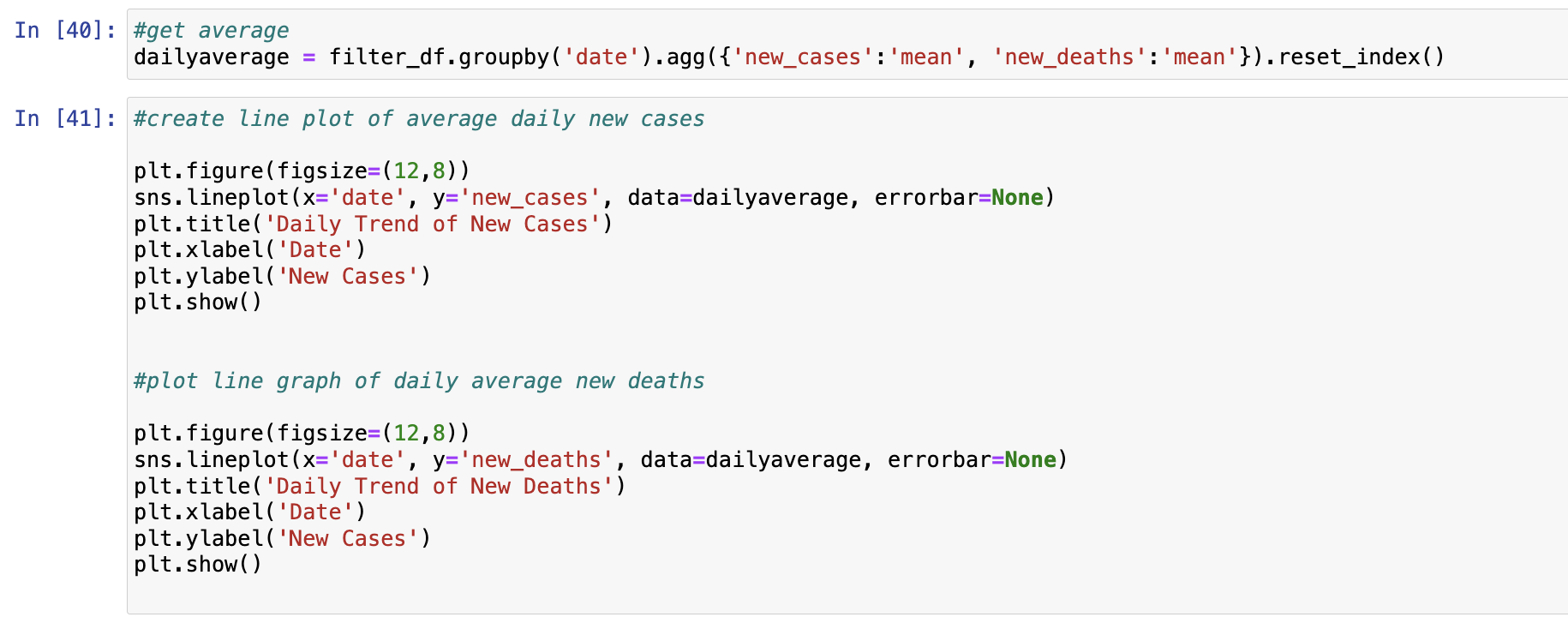
**1. Explore the daily trend of new cases and new deaths globally using line plots.**

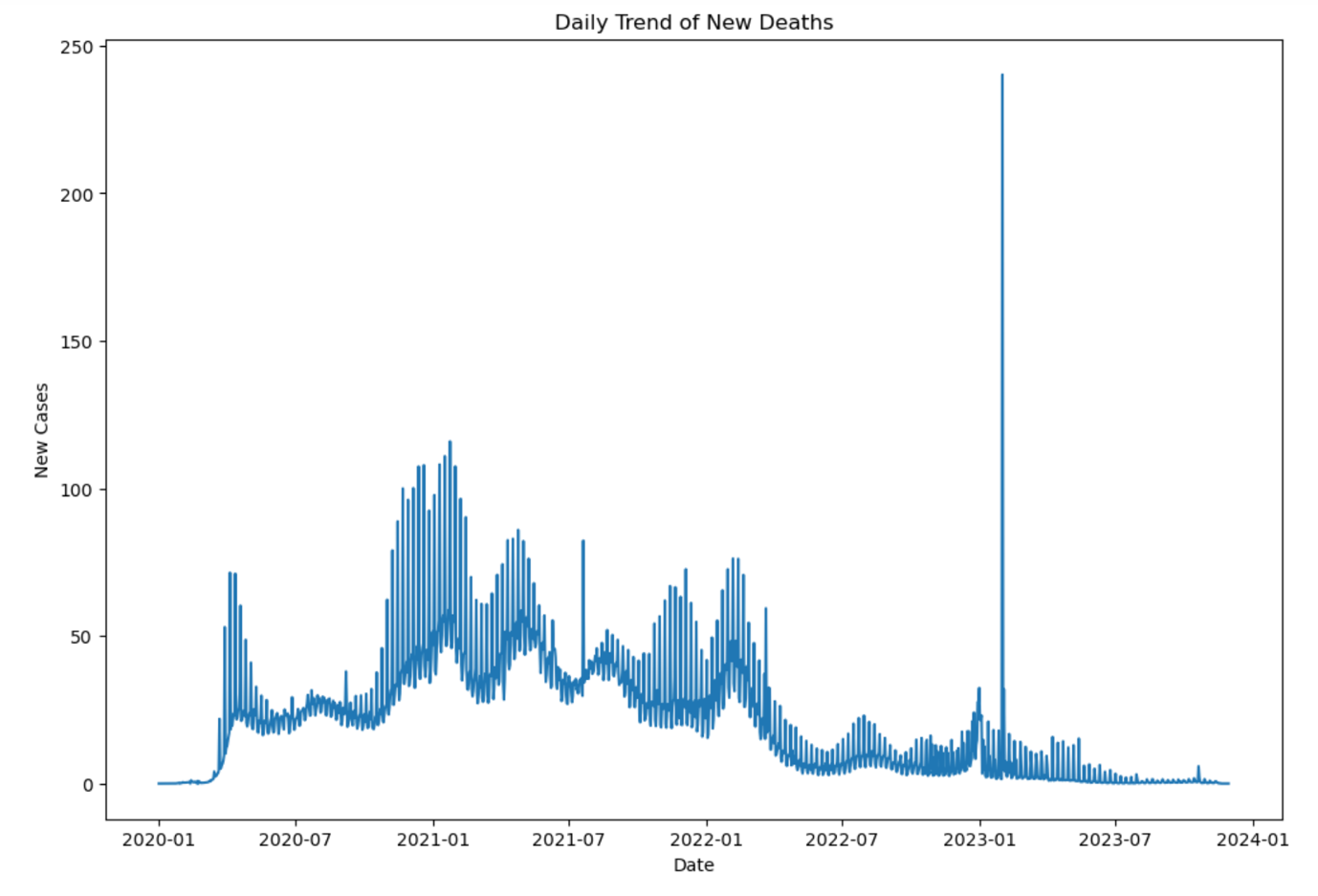
The graphs illustrated how the daily trend of new cases and new deaths differed from continent to continent over time. Both Oceania and Africa have low new cases and new deaths. North America and South America showed similar trends in new cases but South America had higher rates of new deaths than North America.

For Europe, new deaths were high in early 2020 compared to other continents but the trend started decreasing around January 2022. While new deaths decreased, new cases increased sharply in the same time period and peaked at January 2022 before the new cases rate started decreasing. This showed that despite high rates of new cases, hospitals in Europe have since been more equipped to handle COVID-19 cases in 2022 as compared to when the pandemic first started in 2020.

For Asia, new cases and new deaths remained low compared to Europe and South America before suddenly spiking to have the highest new cases and new death rate of all continents in the year 2023. This showed that Asia were perhaps more affected by the later waves of the COVID-19 pandemic.

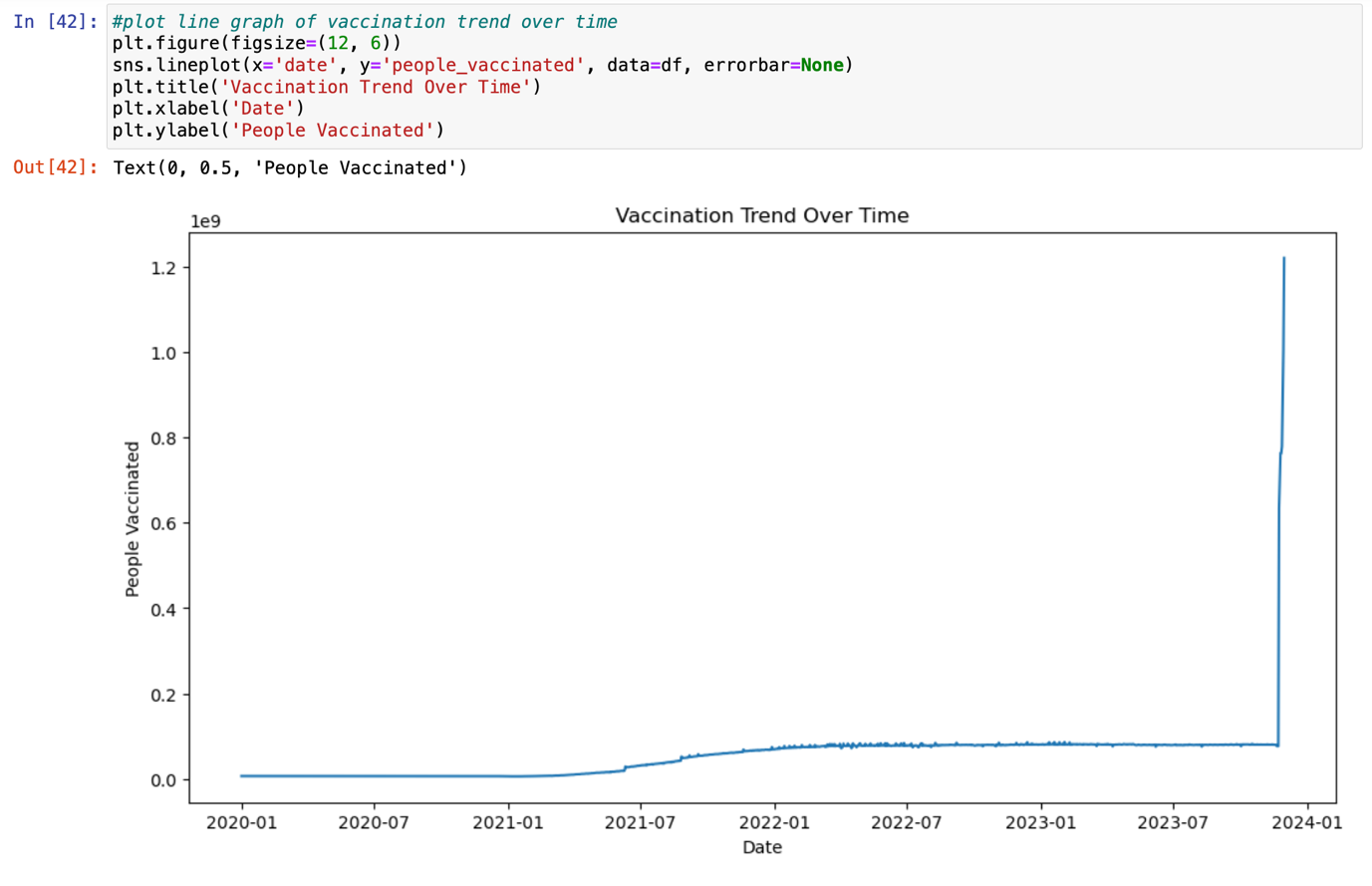
**2. Calculate and visualize the daily average new cases and deaths globally.**





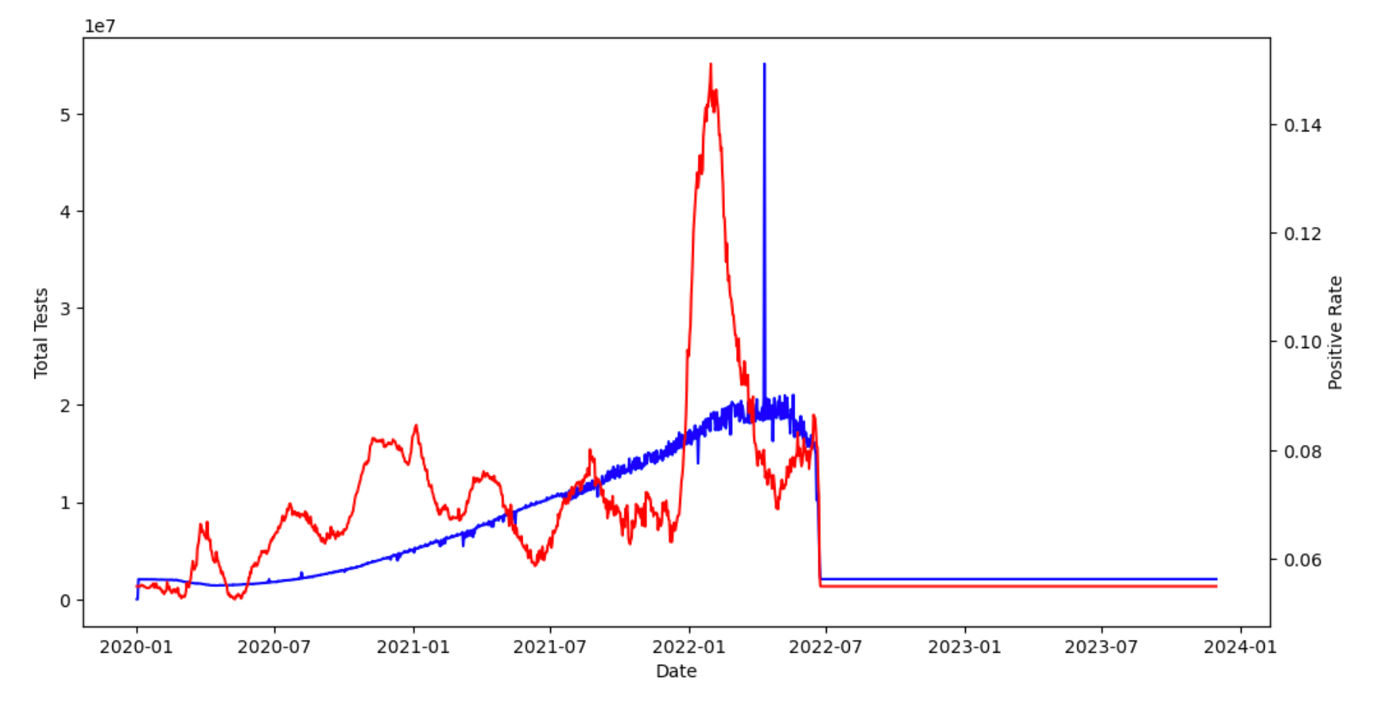
This is similar to the graphs from the previous task as it is derived from the average of new cases and new deaths from all the continents over time.

1. **Explore the trend of vaccination coverage over time globally.**



Vaccination rates increased over time. This showed that people had more access to vaccination over time and are also more willing to be vaccinated over time.

**4. Analyze the total tests and positive rate over time globally.**

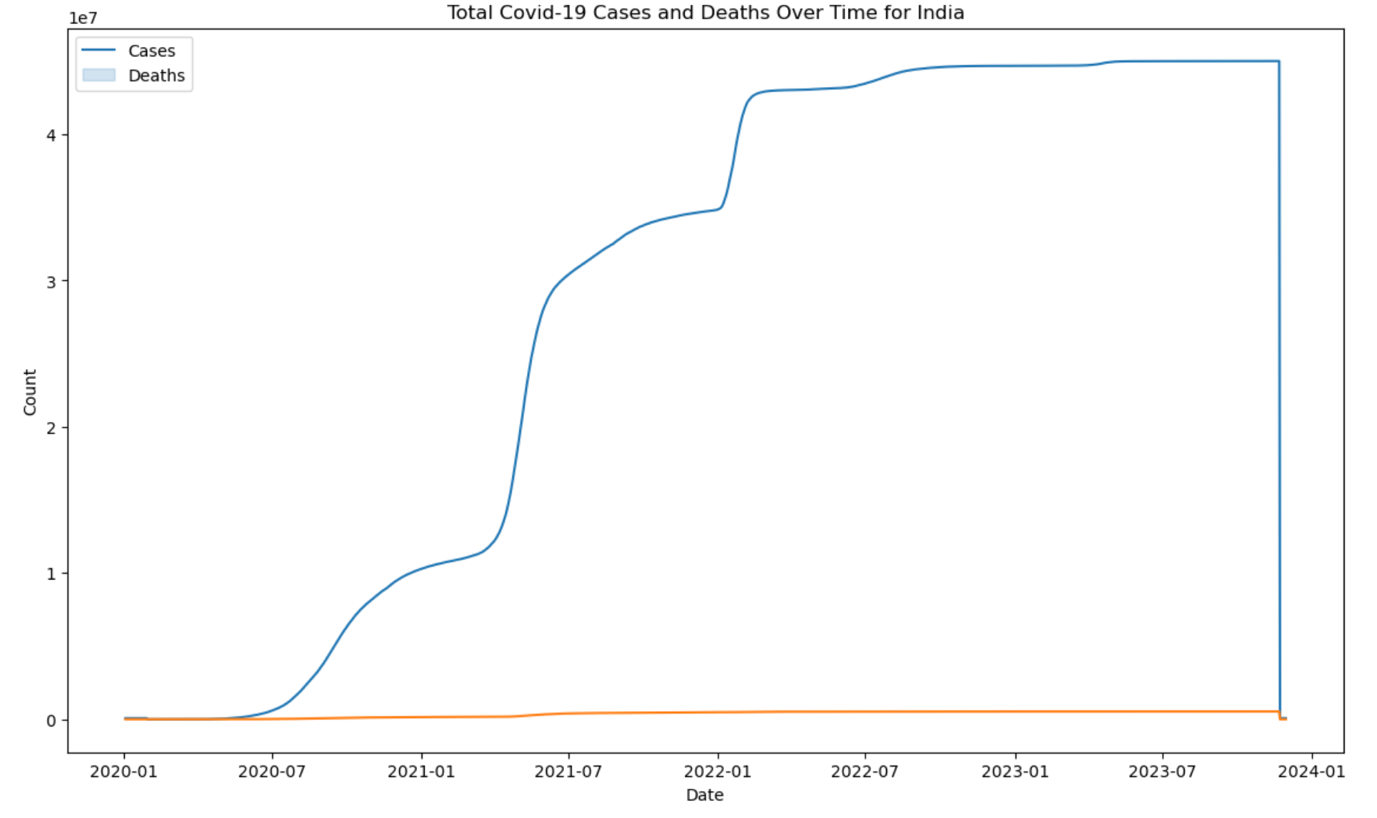
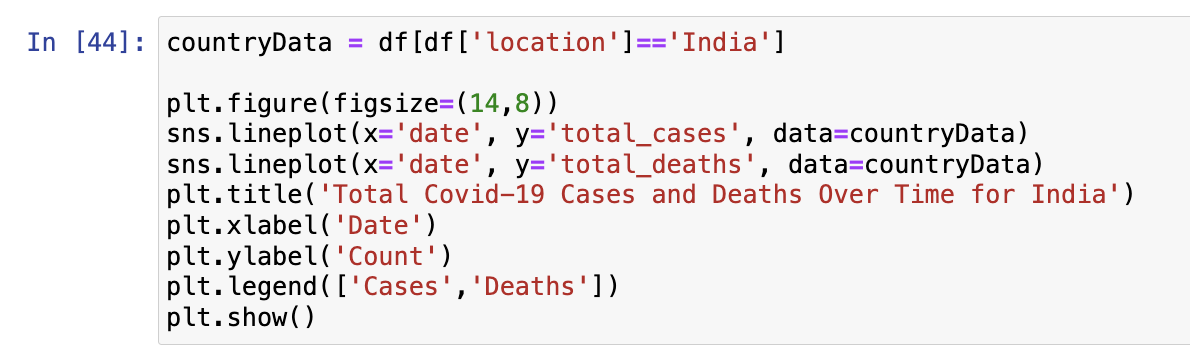


Both total tests and positive rate showed an increasing trend from 2020 to 2022 before peaking in the early half of 2022 and then having a sharp decline during June 2022. After which, the total tests and positive rate remained constant. The decline could be due to the discontinuation of formal testing for the diagnosis of COVID-19 as it became more prevalent and treatable. Patients could possibly use a self-test kit for diagnosis instead of going for swabs and then get treated at clinics instead of hospitals.

**8. Activity 6:** List the problem and screenshot for the same.

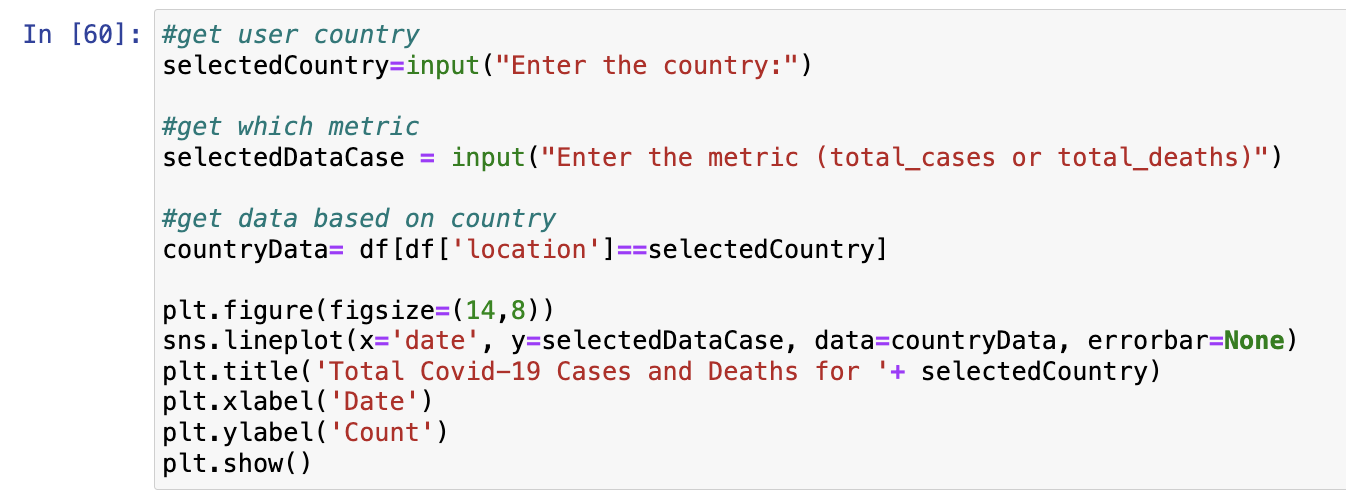
**Problem (In-depth Country Analysis): There is a need to gather deeper insights on specific countries graphically, to understand the correlation between cases and deaths globally, and to explore continental variations in COVID-19 case distribution.**

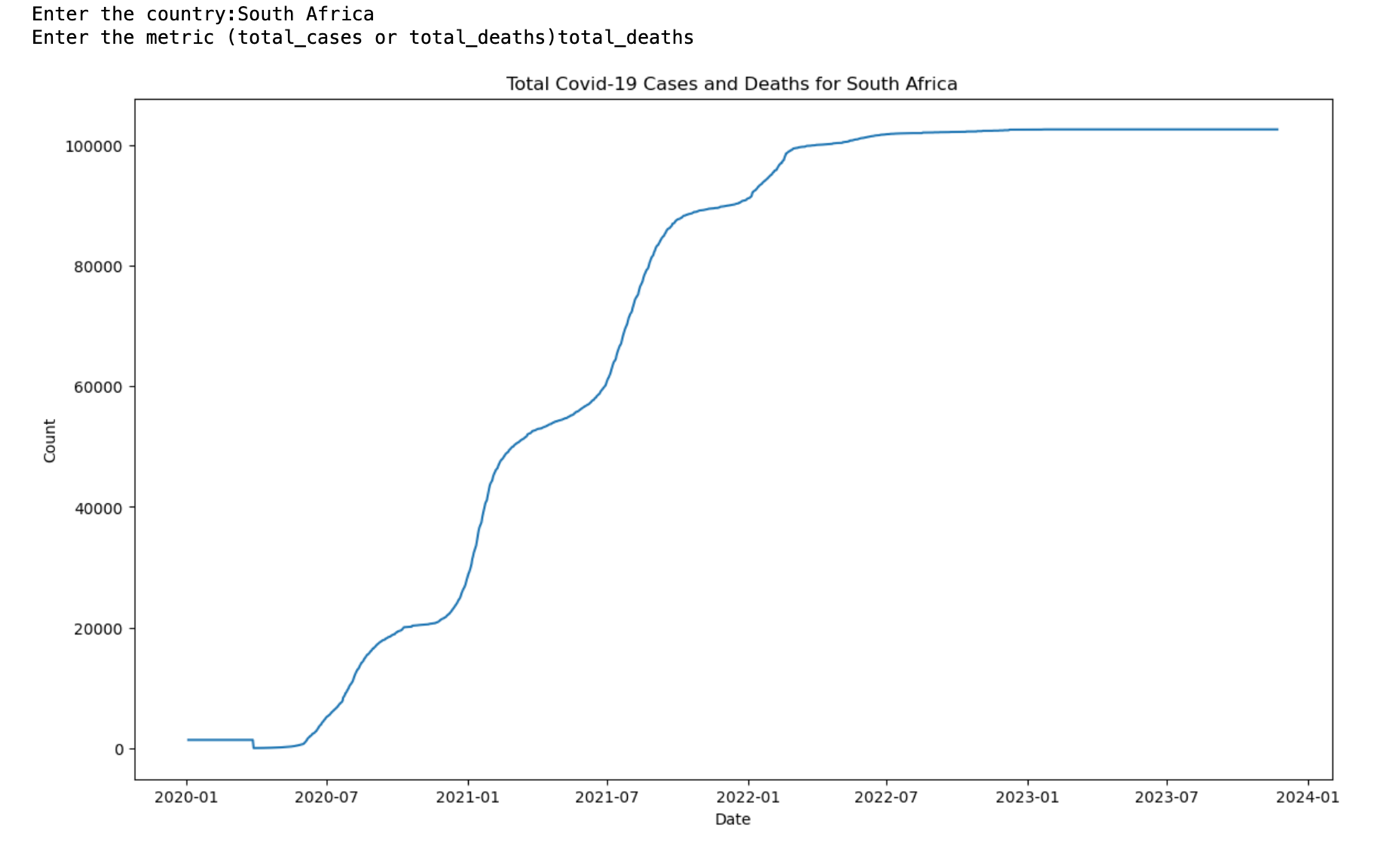
1. **Visualize the total COVID-19 cases and deaths over time for a specific country.**



For India, COVID-19 cases increased over time from 2020 to 2022 before sharply declining towards the end of 2023. Meanwhile, death rates have remained relatively constant over the same time period. This showed that the India healthcare system has been able to cope with their COVID-19 case load and the healthcare provided is good enough that COVID-19 deaths remain relatively low.

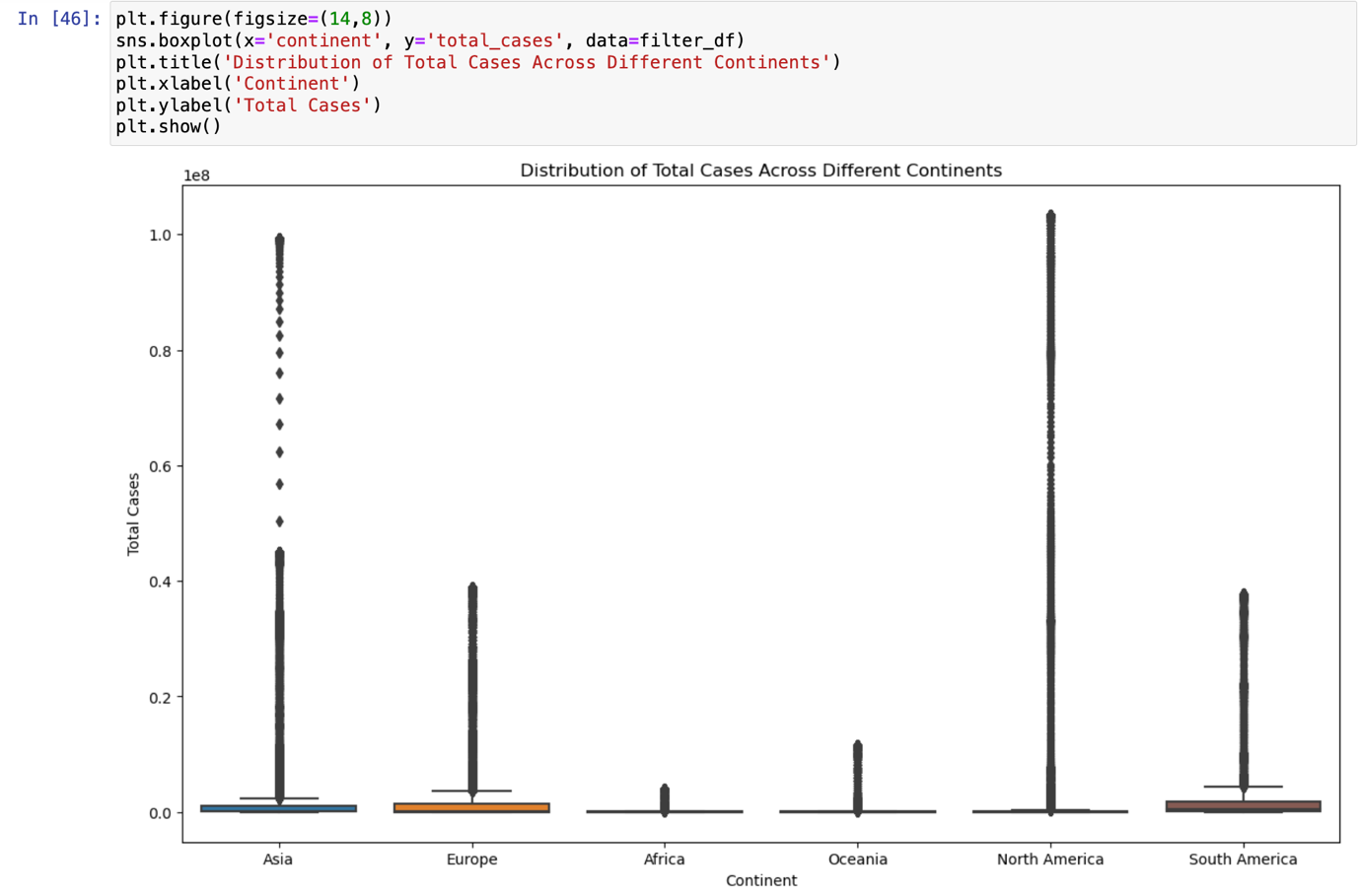
**2. Get user input for the country and metric (total\_cases or total\_deaths) and plot a line chart for the user selected values.**



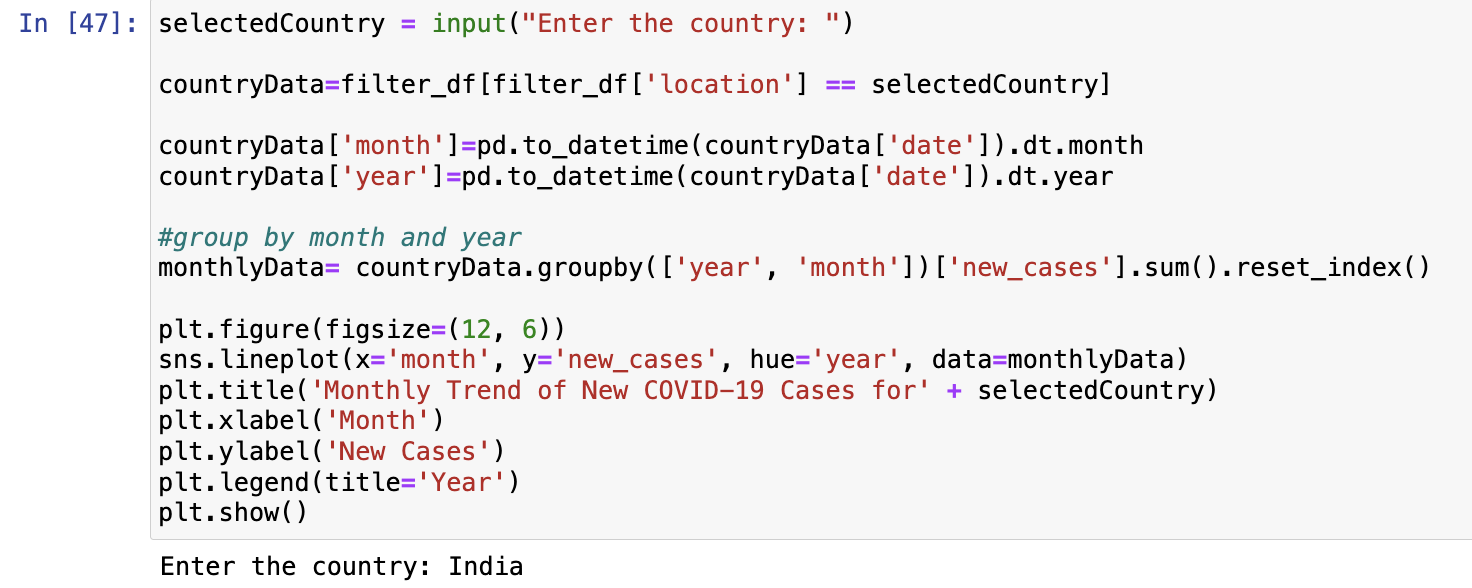


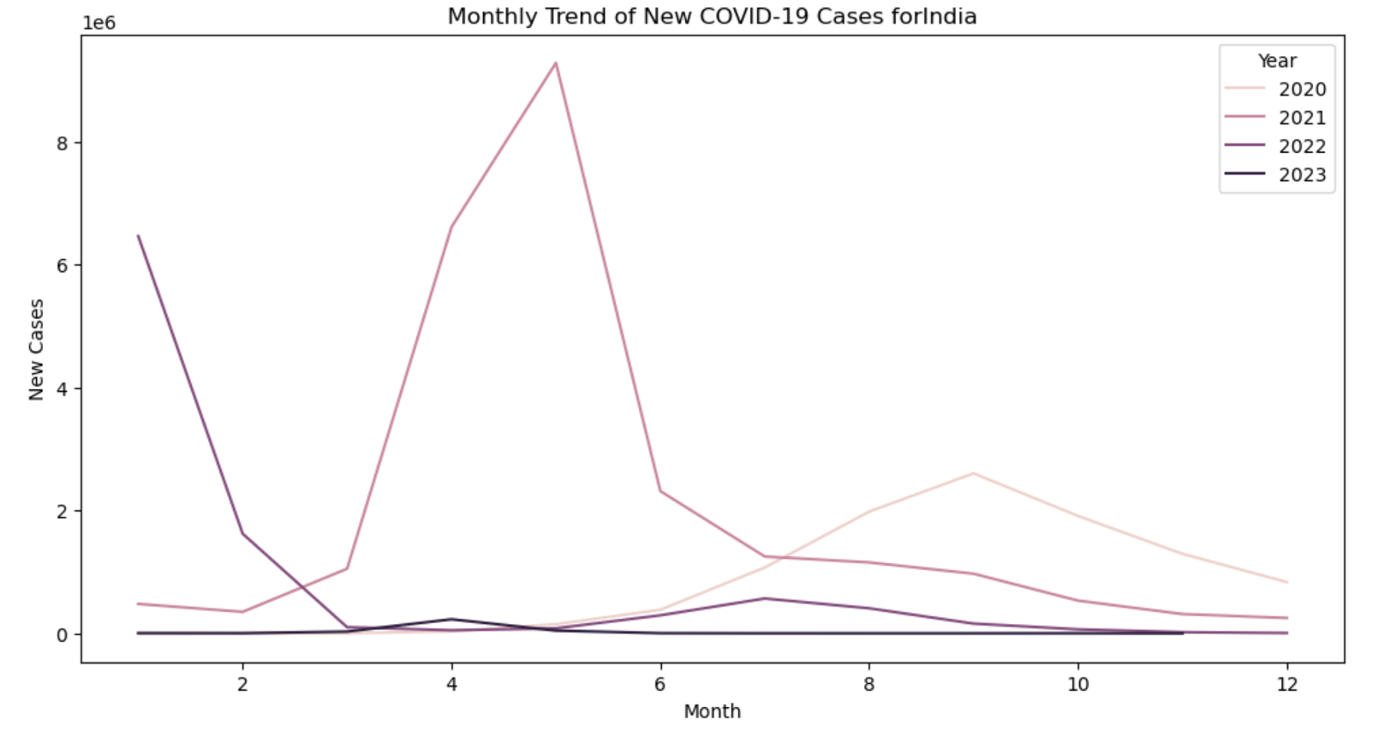
The total deaths for South Africa increased from 2020 to the first half of 2022 before it stabilises. This showed that for South Africa, they became more ready to cope with the COVID-19 pandemic from 2022 onwards.

3. **Analyze the distribution of total cases across different continents using a box plot.**



Both North America and Asia had high total cases while Africa has the lowest total cases.

**4. Visualize the Year wise monthly trend of new COVID-19 cases for user selected country**.

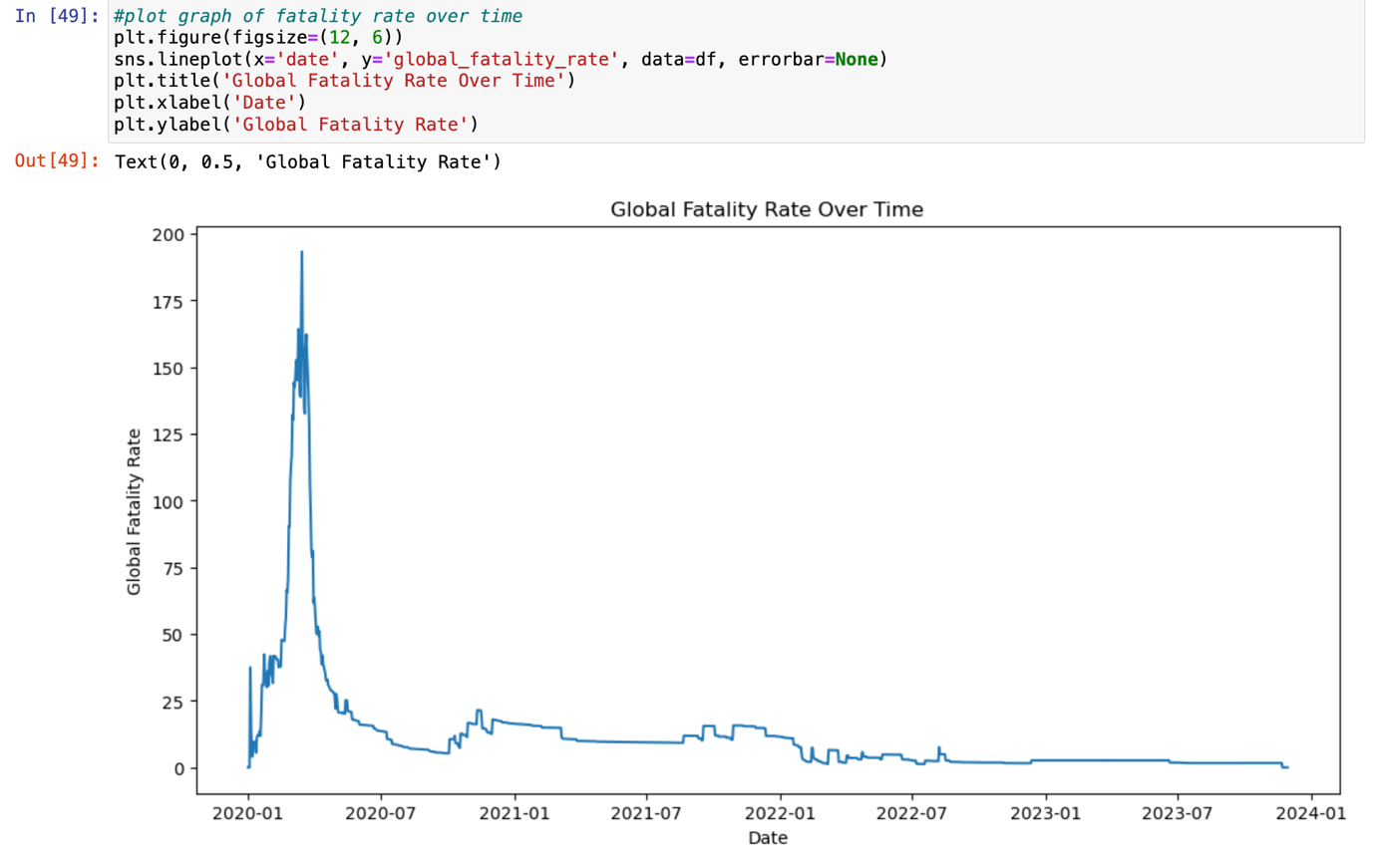
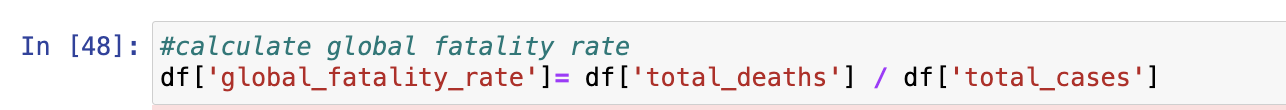


For India, it showed that the year 2021 was a difficult year for the country with many new COVID-19 cases but the country has managed to manage it well such that over the next year 2022 onwards, new cases have remained low.

1. **Activity 7:** List the problem and screenshot for the same.

**Problem (Additional Insights): There is a need to extract additional insights, such as examining the influence of external factors, and evaluating regional disparities for a more holistic understanding of the COVID-19 landscape.**

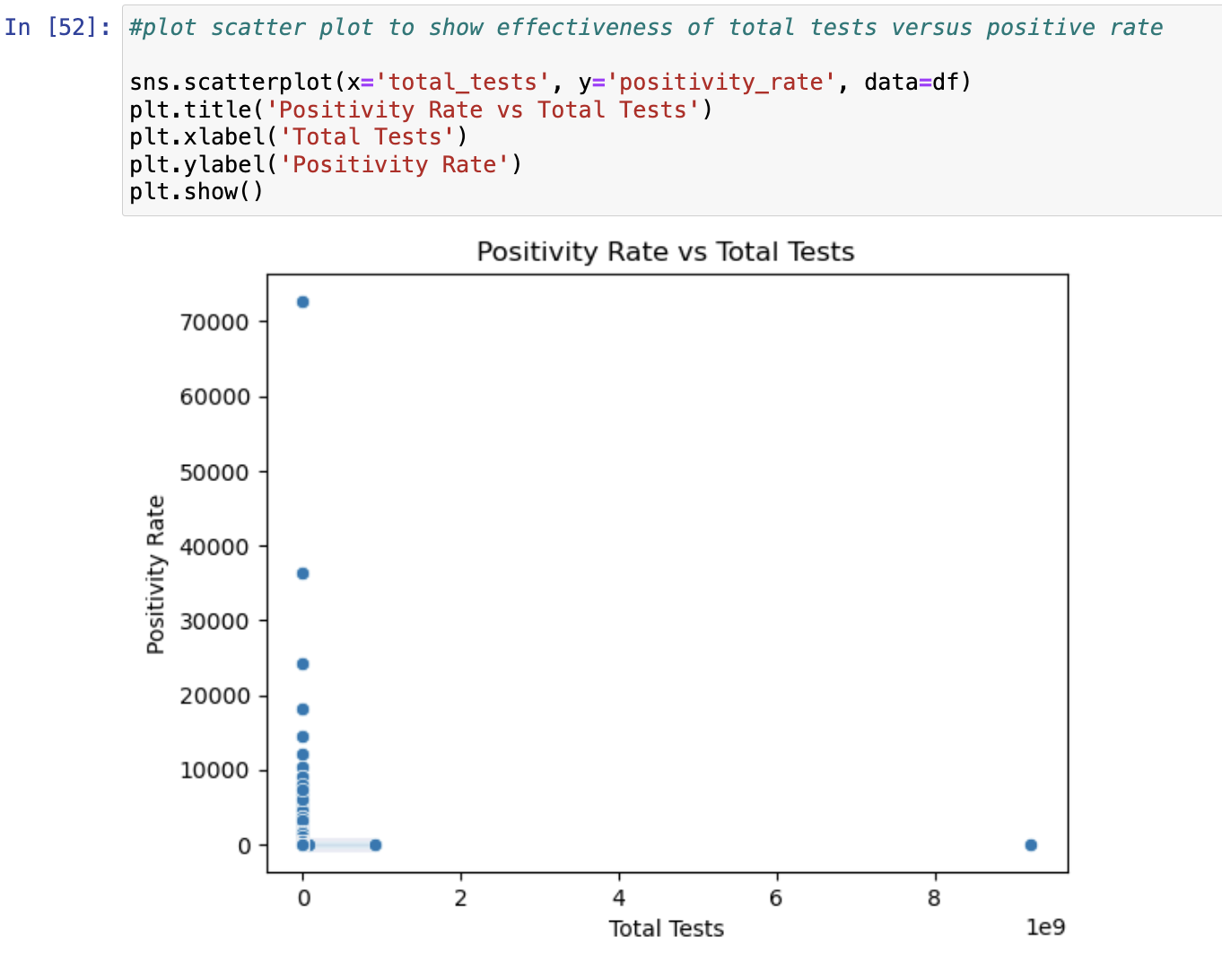
1. **Visualize the fatality rate (total deaths / total cases) over time globally.**



The fatality rate increased in 2020 and rose sharply around March of 2020 before declining sharply. After which, the fatality rate has some rise and falls but remains at a much lower rate than in 2020. This showed that the first wave of COVID-19 infections were more life threatening and resulted in more deaths.

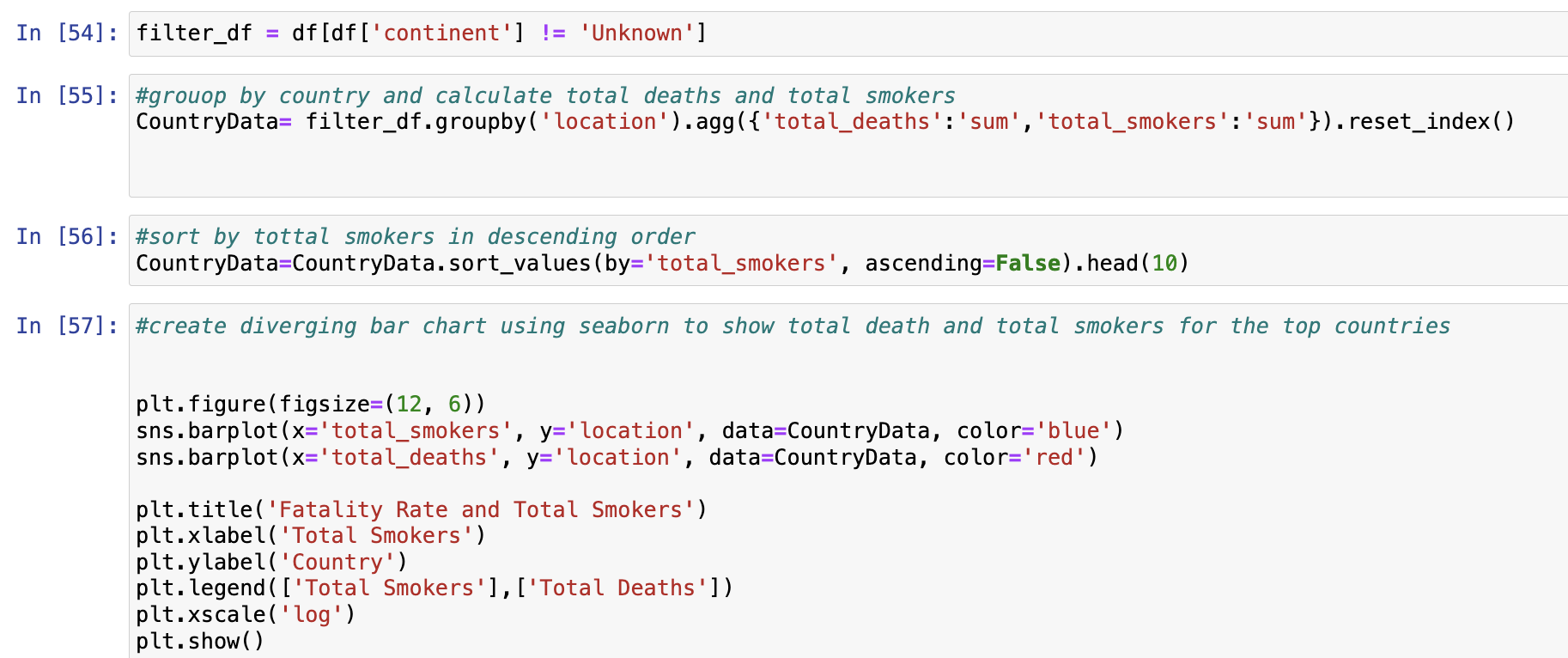
1. **Explore positivity rate (total\_cases/ total tests) verses total tests conducted to analyse testing effectiveness use x axis as logarithmic scale for better visualization.**

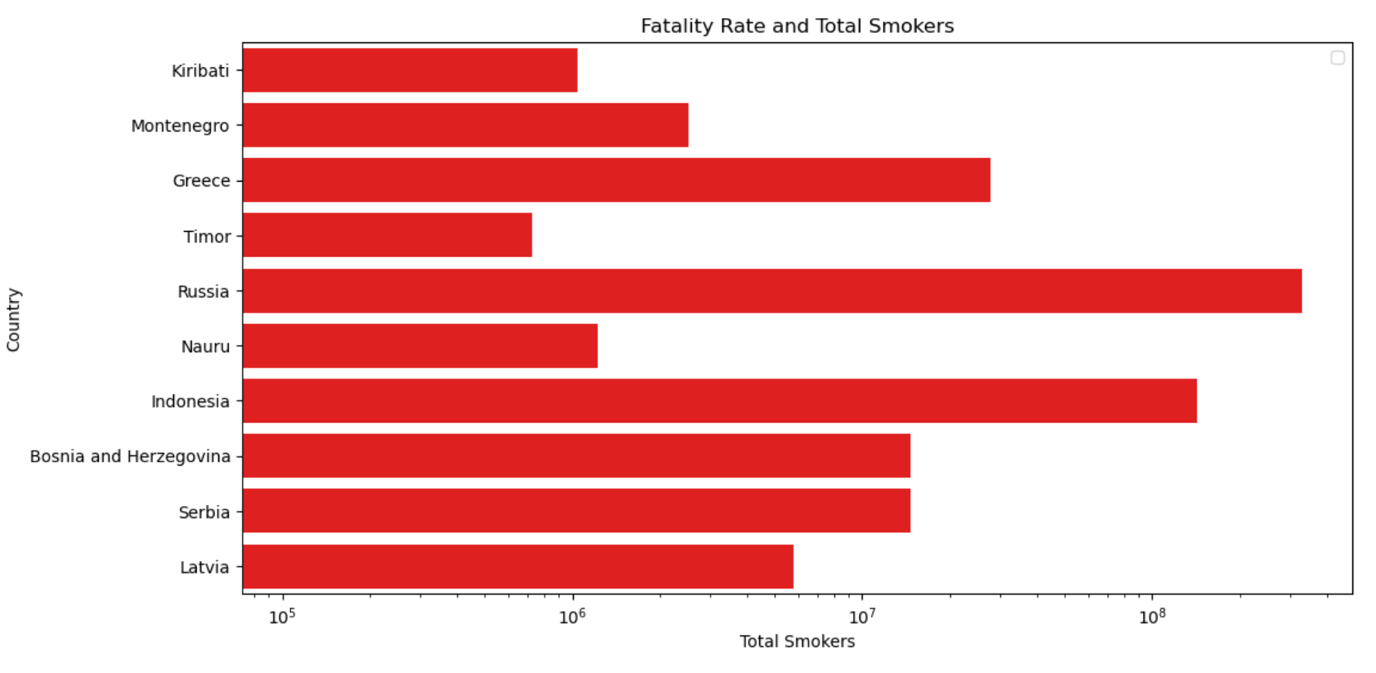




There is no strong correlation between the positivity rate and total tests. This showed that there is not much testing effectiveness perhaps due to the lack of resources in tests available or that not all COVID-19 patients are formally tested.

1. **Analyze fatality rate and its relationship with smoking (Use male\_smokers and female\_smokers columns)**

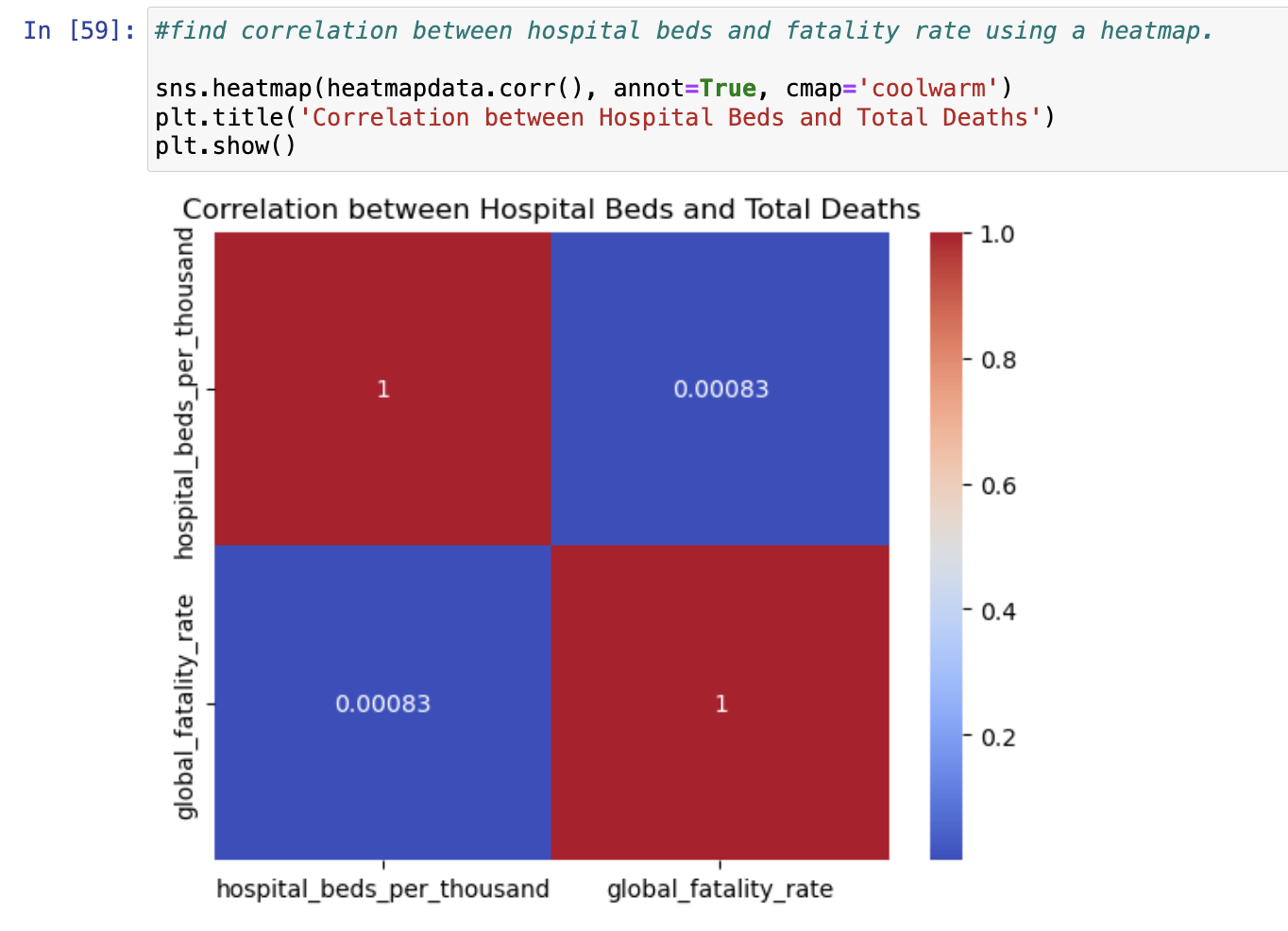




Russia has the most smokers and fatality rate among the top 10 countries. This graph alone is not enough to establish the relationship between whether smoking has an effect on the fatality rate of COVID-19 patients. Smoking is a long term habit and may have affected the lungs of the patients hence it will be better to compare smokers’ fatality rate with non-smokers’ fatality rate to get a better understanding on the relationship between smoking and the fatality rate.

1. **Create a heatmap to analyze the relationship between hospital beds per thousand and fatality rate.**





There is a positive but low correlation between the hospital beds per thousand and the global fatality rate. This showed that the ‘hard’ capacity of the hospital—such as beds available, has not much impact on the death rates. Rather, it could be the other reasons—like the availability of medicine, access to treatments etc that could have a stronger impact on the fatality rate.

1. **Conclusion and Future Improvements**

The graphs depicted generally showed that the COVID-19 pandemic was serious when it first happened in 2020 with higher death rates but gradually the subsequent waves of COVID may have high case numbers but lower death rates. This showed that globally, with time, countries became more able to cope with the COVID-19 pandemic.

For future studies, it will be good to look at death rates to analyse trends at a macro level. The impact of COVID-19 on different continents were shown on graphs, but at the micro level, in a same continent, different countries may have been affected by COVID-19 differently depending on the economic and healthcare resources of the country.

For instance, a first-world country and a developing country, both in the same continent Asia, will have a different death rate due to the disparity in accessing healthcare treatments by the people. However, in this analysis, such details are overlooked and the total deaths are grouped together and analysed as one continent.

Similarly, to establish and look at whether smoking has an impact on the fatality rate, the analysis in this study is not sufficient.

Smoking is a long term habit and may have affected the lungs of the patients hence it will be better to compare smokers’ fatality rate with non-smokers’ fatality rate, at a micro level such as in one country, to get a better understanding on the relationship between smoking and the fatality rate.