DATA SCIENCE COHORT 2 FINAL PROJECT

Capstone Project: Predictive Modelling for COVID-19 Using The Country Wise Dataset

NAME: OPARAUGO IHEOMA A

NO: FE/24/6305992587

In [3]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import plotly.express as px
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings('ignore')
```

In [4]:
```
 # import the dataset
data = pd.read_csv(r"C:\Users\iwund\Desktop\3MTT_Final_Project\country_wise_latest.csv")
```

In [5]:
```
# Shape of dataset
# print(f"The dataset has {data.shape[1]} columns and {data.shape[0]} rows")
print("The number of columns is: ", data.shape[1])
print("The number of rows is: ", data.shape[0])
```

```
The number of columns is:  15
The number of rows is:  187
```

In [6]:
```
# Overview of the dataset
print(data.info())  # Data types and non-null counts
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Country/Region          187 non-null    object 
 1   Confirmed               187 non-null    int64  
 2   Deaths                  187 non-null    int64  
 3   Recovered               187 non-null    int64  
 4   Active                  187 non-null    int64  
 5   New cases               187 non-null    int64  
 6   New deaths              187 non-null    int64  
 7   New recovered           187 non-null    int64  
 8   Deaths / 100 Cases      187 non-null    float64
 9   Recovered / 100 Cases   187 non-null    float64
 10  Deaths / 100 Recovered  187 non-null    float64
 11  Confirmed last week     187 non-null    int64  
 12  1 week change           187 non-null    int64  
 13  1 week % increase       187 non-null    float64
 14  WHO Region              187 non-null    object 
dtypes: float64(4), int64(9), object(2)
memory usage: 22.0+ KB
None
```

In [7]:
```
data.describe() # Summary statistics
```

Out[7]:

	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	Deaths / 100 Recovered	Confirr last w
count	1.870000e+02	187.000000	1.870000e+02	1.870000e+02	187.000000	187.000000	187.000000	187.000000	187.000000	187.00	1.870000e
mean	8.813094e+04	3497.518717	5.063148e+04	3.400194e+04	1222.957219	28.957219	933.812834	3.019519	64.820535	inf	7.868248e
std	3.833187e+05	14100.002482	1.901882e+05	2.133262e+05	5710.374790	120.037173	4197.719635	3.454302	26.287694	NaN	3.382737e
min	1.000000e+01	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	1.000000e
25%	1.114000e+03	18.500000	6.265000e+02	1.415000e+02	4.000000	0.000000	0.000000	0.945000	48.770000	1.45	1.051500e
50%	5.059000e+03	108.000000	2.815000e+03	1.600000e+03	49.000000	1.000000	22.000000	2.150000	71.320000	3.62	5.020000e
75%	4.046050e+04	734.000000	2.260600e+04	9.149000e+03	419.500000	6.000000	221.000000	3.875000	86.885000	6.44	3.708050e
max	4.290259e+06	148011.000000	1.846641e+06	2.816444e+06	56336.000000	1076.000000	33728.000000	28.560000	100.000000	inf	3.834677e

Key Insights from the Analysis

1. Top 10 Countries by Confirmed Cases The top 10 countries dominate the total confirmed cases globally, with significant disparities between them. Leading countries such as the U.S., Brazil, and India show much higher confirmed case counts compared to others.

2. Distribution of Death Rates Death rates range from 0% to a maximum of approximately 28.5%. The majority of countries have a death rate below 5%, indicating that fatality rates are concentrated within this range. Some outliers with higher fatality rates warrant further investigation, potentially related to healthcare system quality or population demographics.

3. Distribution of Recovery Rates Recovery rates vary widely, with most countries having rates above 60%. A significant proportion of countries approach 100% recovery rates, suggesting effective healthcare responses in these regions.

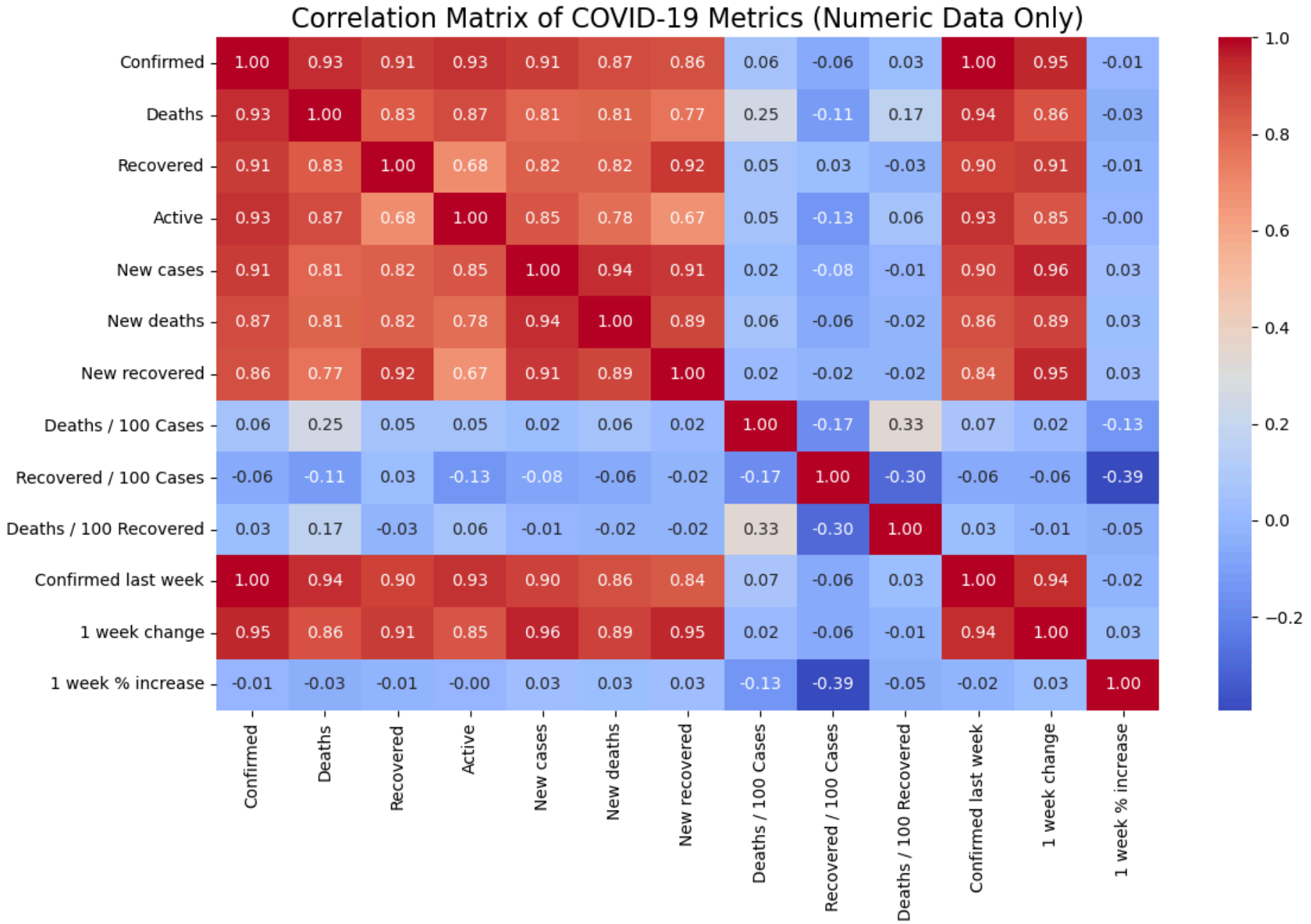
4. Total Cases by WHO Region The Americas region accounts for the majority of confirmed cases globally, followed by Europe and South-East Asia. Regions such as Africa and the Western Pacific report relatively lower confirmed case counts, possibly due to lower transmission rates or underreporting.

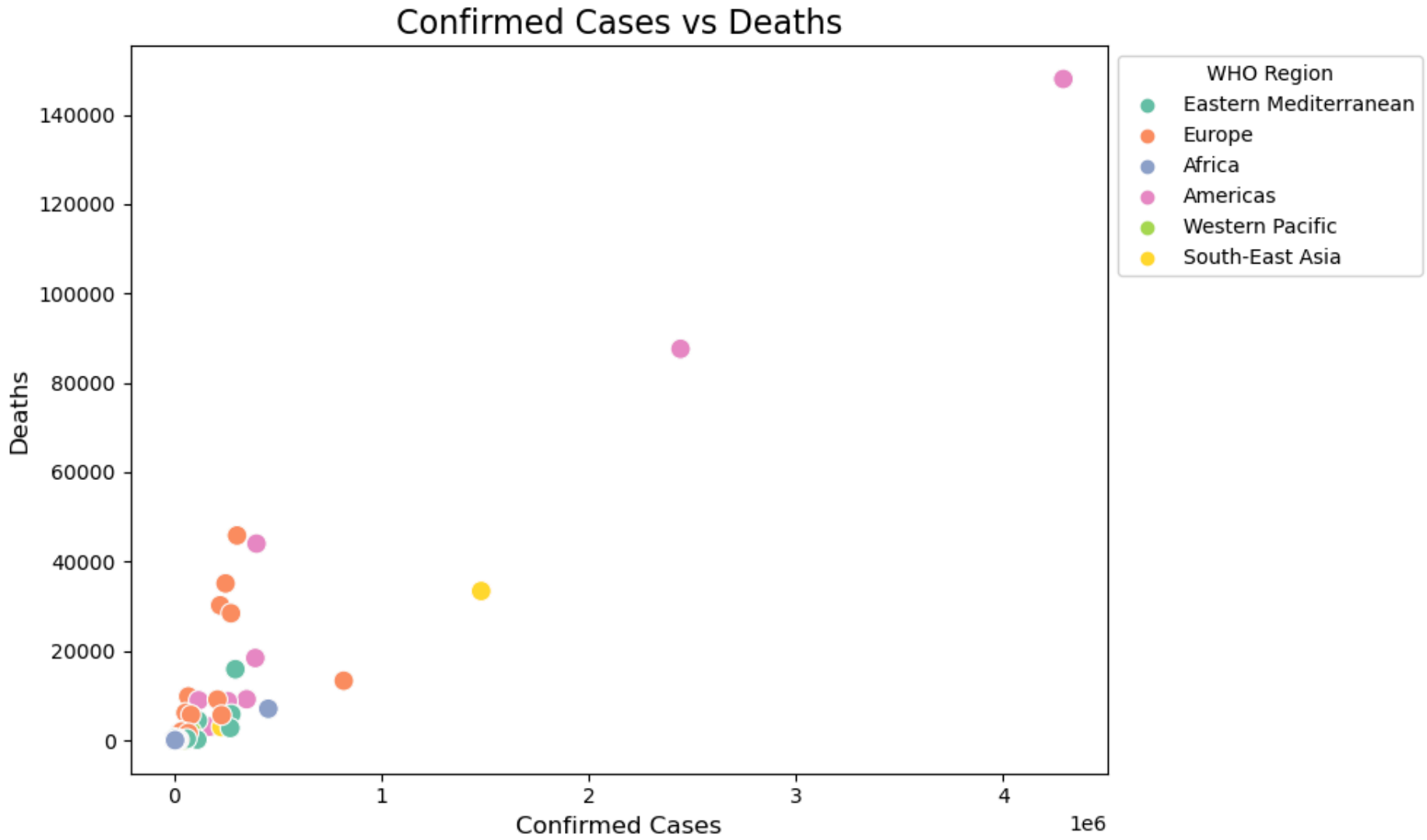
   General Observations

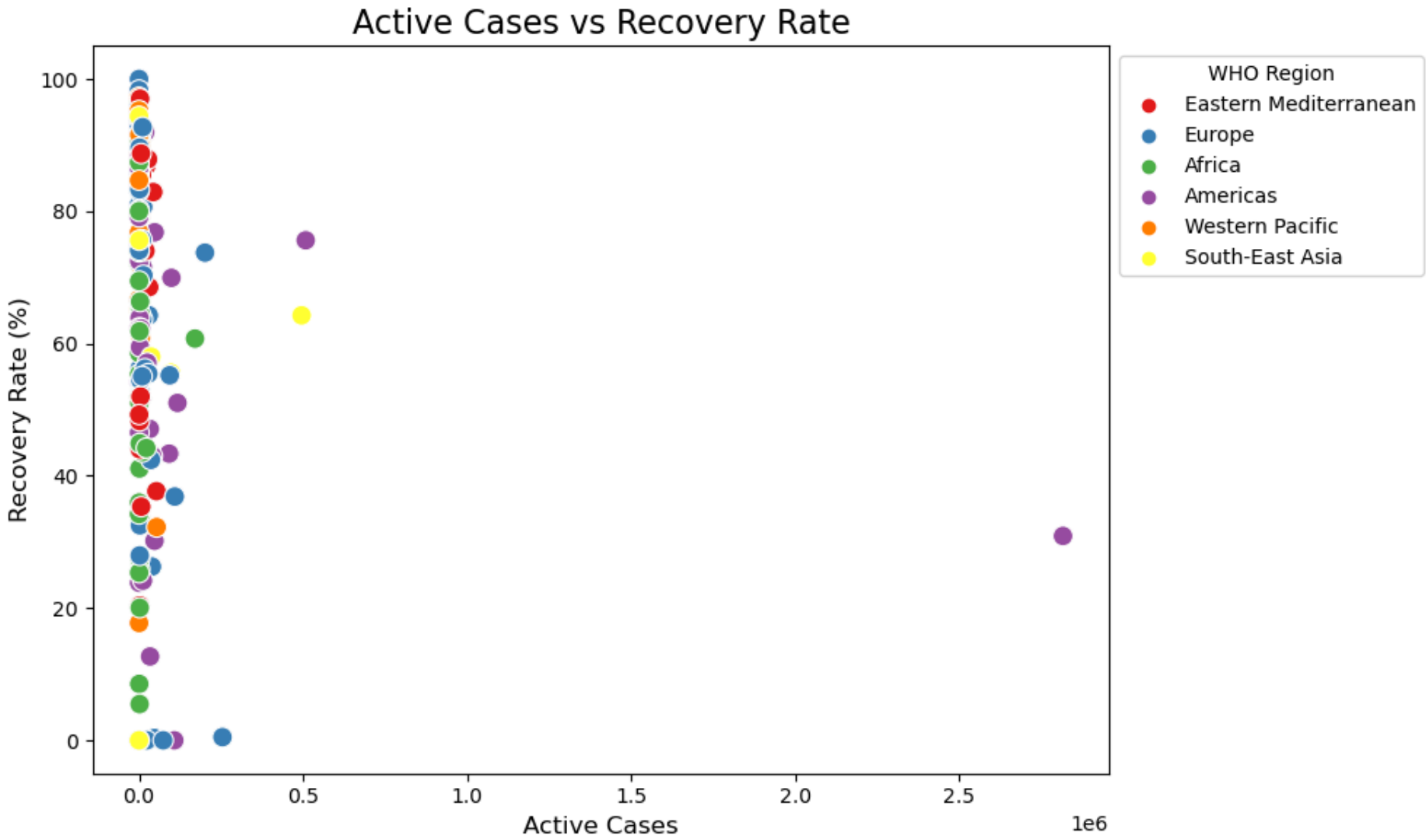
The dataset highlights stark contrasts in COVID-19 impacts across regions and countries. Higher case and death rates are often concentrated in regions with large populations and dense urban centers (e.g., Americas and Europe). Active cases remain substantial, suggesting the ongoing need for public health measures and vaccinations in many regions. The rate of weekly increase in cases provides insight into countries still experiencing significant outbreaks

In [11]:

```
# Drop non-numeric columns for correlation computation
numeric_data = data.select_dtypes(include=["float64", "int64"])
correlation_matrix_fixed = numeric_data.corr()

# Replot the heatmap with fixed data
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix_fixed, annot=True, fmt=".2f", cmap="coolwarm", cbar=True)
plt.title("Correlation Matrix of COVID-19 Metrics (Numeric Data Only)", fontsize=16)
plt.tight_layout()
plt.show()

correlation_matrix_fixed

# Scatter plot: Confirmed vs Deaths
plt.figure(figsize=(10, 6))
sns.scatterplot(x="Confirmed", y="Deaths", data=data, hue="WHO Region", palette="Set2", s=100)
plt.title("Confirmed Cases vs Deaths", fontsize=16)
plt.xlabel("Confirmed Cases", fontsize=12)
plt.ylabel("Deaths", fontsize=12)
plt.legend(title="WHO Region", loc="upper left", bbox_to_anchor=(1, 1))
plt.tight_layout()
plt.show()

# Scatter plot: Active cases vs Recovery rate
plt.figure(figsize=(10, 6))
sns.scatterplot(x="Active", y="Recovered / 100 Cases", data=data, hue="WHO Region", palette="Set1", s=100)
plt.title("Active Cases vs Recovery Rate", fontsize=16)
plt.xlabel("Active Cases", fontsize=12)
plt.ylabel("Recovery Rate (%)", fontsize=12)
plt.legend(title="WHO Region", loc="upper left", bbox_to_anchor=(1, 1))
plt.tight_layout()
plt.show()
```


Correlation Matrix of COVID-19 Metrics (Numeric Data Only)
Confirmed 1.00 0.93 0.91 0.93 0.91 0.87 0.86 0.06 -0.06 0.03 1.00 0.95 -0.01
Deaths 0.93 1.00 0.83 0.87 0.81 0.81 0.77 0.25 -0.11 0.17 0.94 0.86 -0.03
Recovered 0.91 0.83 1.00 0.68 0.82 0.82 0.92 0.05 0.03 -0.03 0.90 0.91 -0.01
Active 0.93 0.87 0.68 1.00 0.85 0.78 0.67 0.05 -0.13 0.06 0.93 0.85 -0.00
New cases 0.91 0.81 0.82 0.85 1.00 0.94 0.91 0.02 -0.08 -0.01 0.90 0.96 0.03
New deaths 0.87 0.81 0.82 0.78 0.94 1.00 0.89 0.06 -0.06 -0.02 0.86 0.89 0.03
New recovered 0.86 0.77 0.92 0.67 0.91 0.89 1.00 0.02 -0.02 -0.02 0.84 0.95 0.03
Deaths / 100 Cases 0.06 0.25 0.05 0.05 0.02 0.06 0.02 1.00 -0.17 0.33 0.07 0.02 -0.13
Recovered / 100 Cases -0.06 -0.11 0.03 -0.13 -0.08 -0.06 -0.02 -0.17 1.00 -0.30 -0.06 -0.06 -0.39
Deaths / 100 Recovered 0.03 0.17 -0.03 0.06 -0.01 -0.02 -0.02 0.33 -0.30 1.00 0.03 -0.01 -0.05
Confirmed last week 1.00 0.94 0.90 0.93 0.90 0.86 0.84 0.07 -0.06 0.03 1.00 0.94 -0.02
1 week change 0.95 0.86 0.91 0.85 0.96 0.89 0.95 0.02 -0.06 -0.01 0.94 1.00 0.03
1 week % increase -0.01 -0.03 -0.01 -0.00 0.03 0.03 0.03 -0.13 -0.39 -0.05 -0.02 0.03 1.00
Confirmed
Deaths
Recovered
Active
New cases
New deaths
New recovered
Deaths / 100 Cases
Recovered / 100 Cases
Deaths / 100 Recovered
Confirmed last week
1 week change
1 week % increase
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Confirmed Cases vs Deaths
Deaths
140000
120000
100000
80000
60000
40000
20000
0
0
1
2
3
4
Confirmed Cases
1e6
WHO Region
Eastern Mediterranean
Europe
Africa
Americas
Western Pacific
South-East Asia



Insights from Correlation and Plots

Correlation Analysis

Confirmed Cases: Strong positive correlation with Deaths (0.93) and Recovered (0.91), indicating that higher case counts are associated with both outcomes. Strong correlation with Active Cases (0.93), showing consistent patterns in countries with high case loads. Death Rate (Deaths / 100

Cases): Weak or negligible correlation with total confirmed cases, suggesting that other factors (e.g., healthcare quality, population demographics) have a more significant impact on fatality outcomes.

Scatter Plots

Confirmed Cases vs Deaths The scatterplot reveals a proportional relationship, with some outliers where deaths are disproportionately high relative to confirmed cases. Regions: The Americas and Europe dominate the higher ends of the spectrum, indicating higher case loads and fatalities.

Active Cases vs Recovery Rates

Trend: Countries with fewer active cases tend to have higher recovery rates, highlighting the effectiveness of containment and healthcare systems. Outliers: Some countries exhibit high active cases with moderate recovery rates, which may point to ongoing outbreaks or challenges in healthcare management.

In [ ]: