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

NAME: OPARAUGO IHEOMA A.

NO: FE/24/6305992587

This Dataset Offers A Day By Day Overview Of The COVID-19 Virus

Case Scenario In response to the COVID-19 pandemic, public health organizations have faced immense challenges in predicting the spread of the virus and understanding key factors that influence transmission and patient outcomes. Imagine you have been hired as a data scientist by a public health organization, "HealthGuard Analytics," to build a predictive modeling system. The organization requires actionable insights to inform policies, anticipate future outbreaks, and improve health resource allocation. Using historical COVID-19 data, you will conduct data cleaning, perform exploratory data analysis (EDA), and develop predictive models to forecast COVID-19 trends. You will present your findings through visualizations and provide a final report summarizing insights and recommendations for public health responses.

In [28]:
```
# import all needed libraries
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 [41]:
```
# import all required datasets
data = pd.read_csv(r"C:\Users\iwund\Desktop\3MTT_Final_Project\day_wise.csv")
data_covid = pd.read_csv(r"C:\Users\iwund\Desktop\3MTT_Final_Project\covid_19_clear
data_country = pd.read_csv(r"C:\Users\iwund\Desktop\3MTT_Final_Project\country_wise
data.rename(columns={'WHO Region': 'Continent'}, inplace=True)
```

In [5]:
```
# Display the first few rows and summary information of the dataset to understand i
data.head()
```

Out[5]:

	Date	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	R
0	2020-01-22	555	17	28	510	0	0	0	3.06	5.05	
1	2020-01-23	654	18	30	606	99	1	2	2.75	4.59	
2	2020-01-24	941	26	36	879	287	8	6	2.76	3.83	
3	2020-01-25	1434	42	39	1353	493	16	3	2.93	2.72	
4	2020-01-26	2118	56	52	2010	684	14	13	2.64	2.46	

In [6]:
```
# Display the first few rows and summary information of the dataset to understand i
data.tail()
```

Out[6]:

	Date	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recover / 1 Cas
183	2020-07-23	15510481	633506	8710969	6166006	282756	9966	169714	4.08	56.
184	2020-07-24	15791645	639650	8939705	6212290	281164	6144	228736	4.05	56.
185	2020-07-25	16047190	644517	9158743	6243930	255545	4867	219038	4.02	57.
186	2020-07-26	16251796	648621	9293464	6309711	204606	4104	134721	3.99	57.
187	2020-07-27	16480485	654036	9468087	6358362	228693	5415	174623	3.97	57.

In [7]:
```
# Check dataset info
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 188 entries, 0 to 187
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Date                    188 non-null    object 
 1   Confirmed               188 non-null    int64  
 2   Deaths                  188 non-null    int64  
 3   Recovered               188 non-null    int64  
 4   Active                  188 non-null    int64  
 5   New cases               188 non-null    int64  
 6   New deaths              188 non-null    int64  
 7   New recovered           188 non-null    int64  
 8   Deaths / 100 Cases      188 non-null    float64
 9   Recovered / 100 Cases   188 non-null    float64
 10  Deaths / 100 Recovered  188 non-null    float64
 11  No. of countries        188 non-null    int64  
dtypes: float64(3), int64(8), object(1)
memory usage: 17.8+ KB
```

In [8]:
```
# Statistical Summary
data.describe()
```

Out[8]:

	Confirmed	Deaths	Recovered	Active	New cases	New deaths	
count	1.880000e+02	188.000000	1.880000e+02	1.880000e+02	188.000000	188.000000	
mean	4.406960e+06	230770.760638	2.066001e+06	2.110188e+06	87771.021277	3478.824468	50
std	4.757988e+06	217929.094183	2.627976e+06	1.969670e+06	75295.293255	2537.735652	56
min	5.550000e+02	17.000000	2.800000e+01	5.100000e+02	0.000000	0.000000	
25%	1.121910e+05	3935.000000	6.044125e+04	5.864175e+04	5568.500000	250.750000	2
50%	2.848733e+06	204190.000000	7.847840e+05	1.859759e+06	81114.000000	4116.000000	30
75%	7.422046e+06	418634.500000	3.416396e+06	3.587015e+06	131502.500000	5346.000000	79
max	1.648048e+07	654036.000000	9.468087e+06	6.358362e+06	282756.000000	9966.000000	284

In [9]:
```
# Date range of the dataset
print("Date Range: ", data['Date'].min(), " to ", data['Date'].max())
```

```
Date Range:  2020-01-22  to  2020-07-27
```

In [10]:
```
# Check for mising values
data.isnull().sum()
```

Out[10]:
```
Date                    0
Confirmed               0
Deaths                  0
Recovered               0
Active                  0
New cases               0
New deaths              0
New recovered           0
Deaths / 100 Cases      0
Recovered / 100 Cases   0
Deaths / 100 Recovered  0
No. of countries        0
dtype: int64
```

In [11]:
```
date_column = data['Date']
# Convert 'Date' column to datetime format
date_column = pd.to_datetime(date_column, errors='coerce')
data['Date'] = date_column
# Set the date column as the index
data.set_index('Date', inplace=True)
```

In [12]:
```
# Check data info again to confirm Date format
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 188 entries, 2020-01-22 to 2020-07-27
Data columns (total 11 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Confirmed              188 non-null    int64  
 1   Deaths                 188 non-null    int64  
 2   Recovered              188 non-null    int64  
 3   Active                 188 non-null    int64  
 4   New cases              188 non-null    int64  
 5   New deaths             188 non-null    int64  
 6   New recovered          188 non-null    int64  
 7   Deaths / 100 Cases     188 non-null    float64
 8   Recovered / 100 Cases  188 non-null    float64
 9   Deaths / 100 Recovered  188 non-null    float64
 10  No. of countries       188 non-null    int64  
dtypes: float64(3), int64(8)
memory usage: 17.6 KB
```

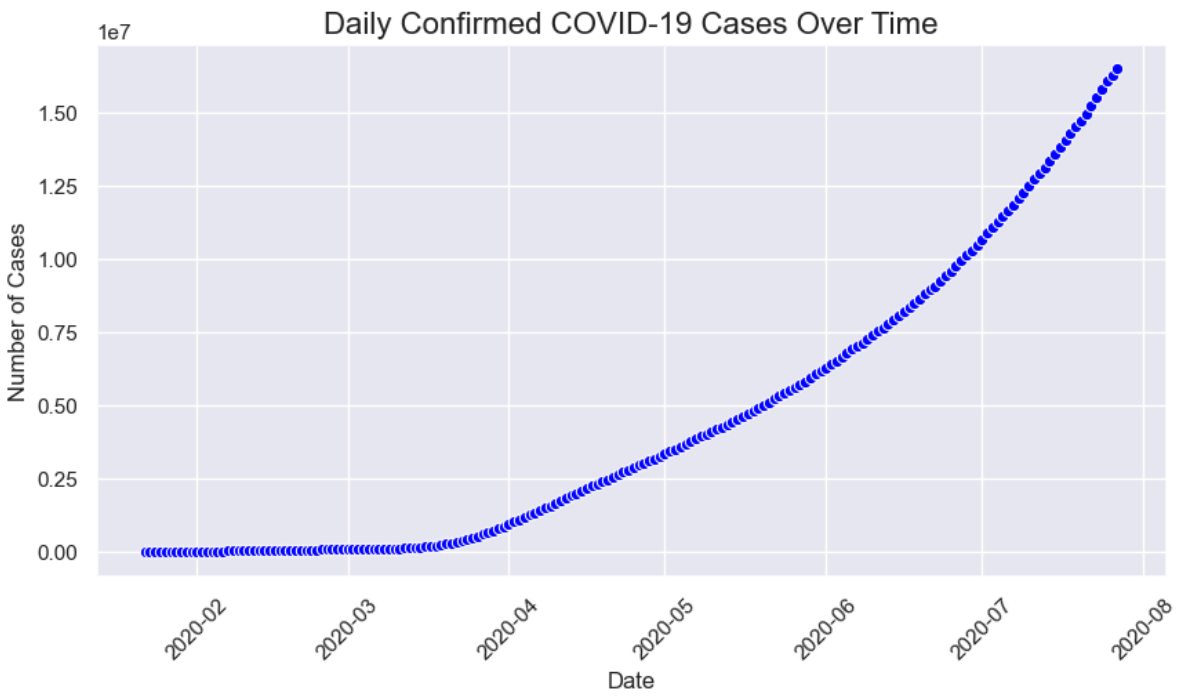
In [13]:
```
# Data head
data.head()
```

Out[13]:

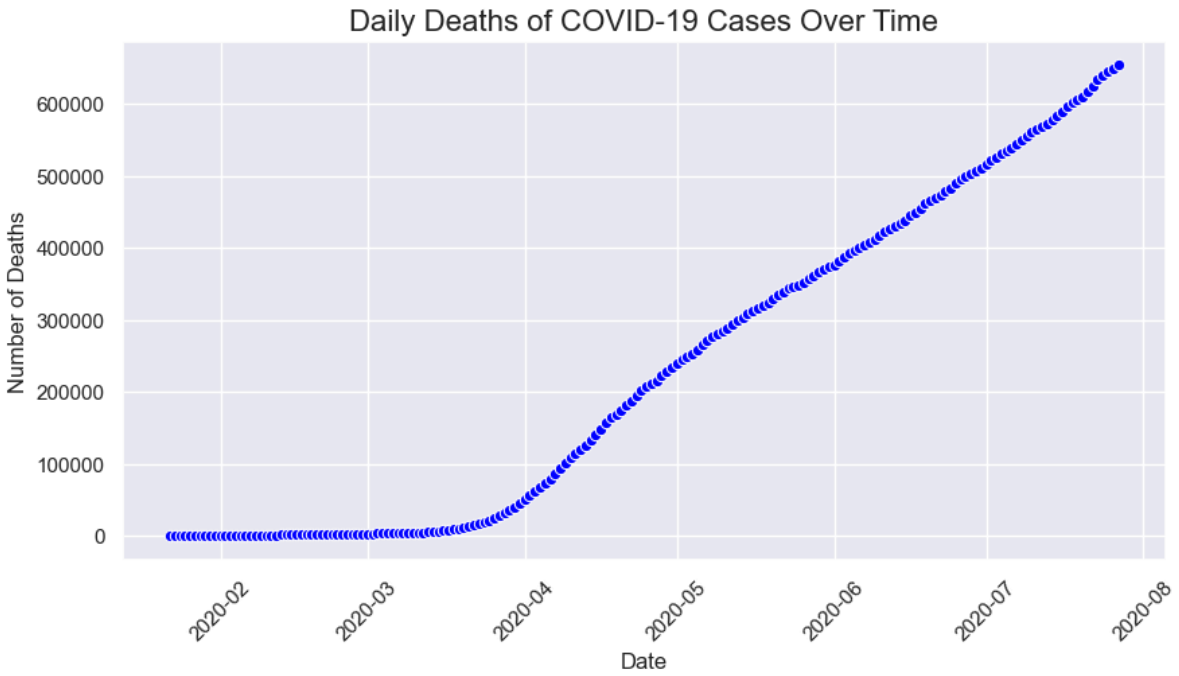
Date	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	De Reco
2020-01-22	555	17	28	510	0	0	0	3.06	5.05	
2020-01-23	654	18	30	606	99	1	2	2.75	4.59	
2020-01-24	941	26	36	879	287	8	6	2.76	3.83	
2020-01-25	1434	42	39	1353	493	16	3	2.93	2.72	
2020-01-26	2118	56	52	2010	684	14	13	2.64	2.46	

In [14]:
```
# Daily Trends: Confirmed Cases Over Time
daily_trends = data.groupby("Date")["Confirmed"].sum()
plt.figure(figsize=(10, 5))
sns.set_theme(style="darkgrid")
sns.lineplot(data=daily_trends, marker="o", color="blue")
plt.title("Daily Confirmed COVID-19 Cases Over Time", fontsize=16)
plt.xlabel("Date", fontsize=12)
plt.ylabel("Number of Cases", fontsize=12)
```

```
plt.xticks(rotation=45)
plt.show()
```

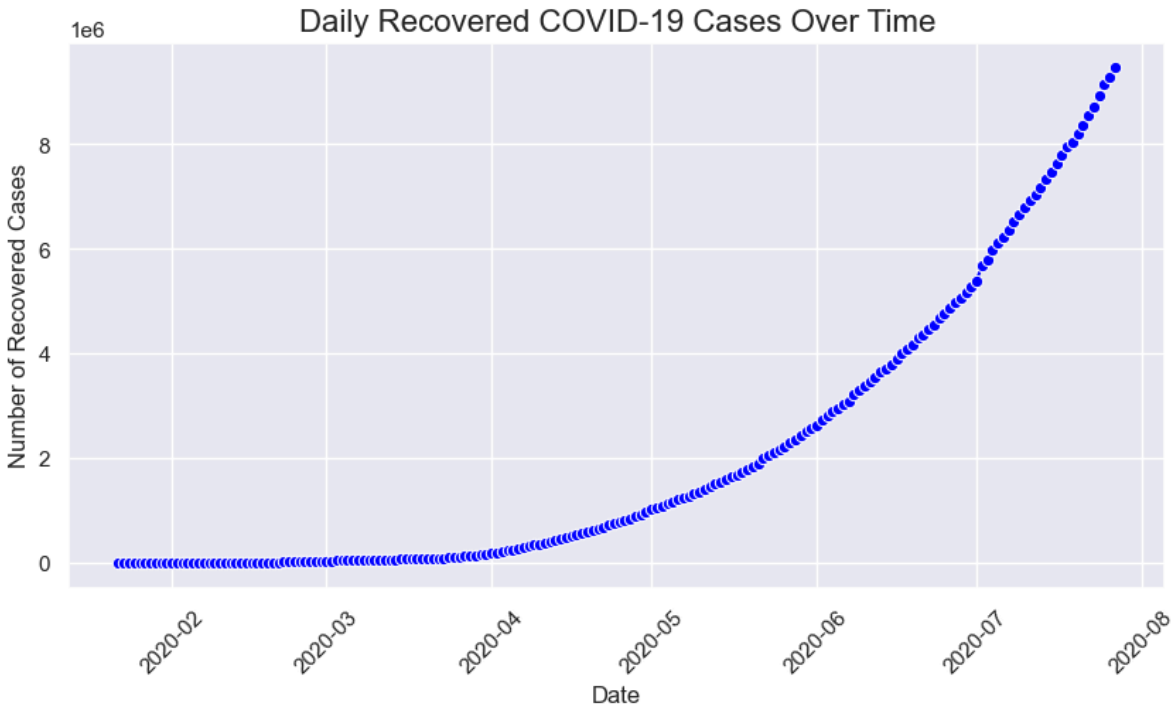


In [15]:
```
# Daily Trends: Confirmed Cases Over Time
daily_trends = data.groupby("Date")["Deaths"].sum()
plt.figure(figsize=(10, 5))
sns.set_theme(style="darkgrid")
sns.lineplot(data=daily_trends, marker="o", color="blue")
plt.title("Daily Deaths of COVID-19 Cases Over Time", fontsize=16)
plt.xlabel("Date", fontsize=12)
plt.ylabel("Number of Deaths", fontsize=12)
plt.xticks(rotation=45)
plt.show()
```

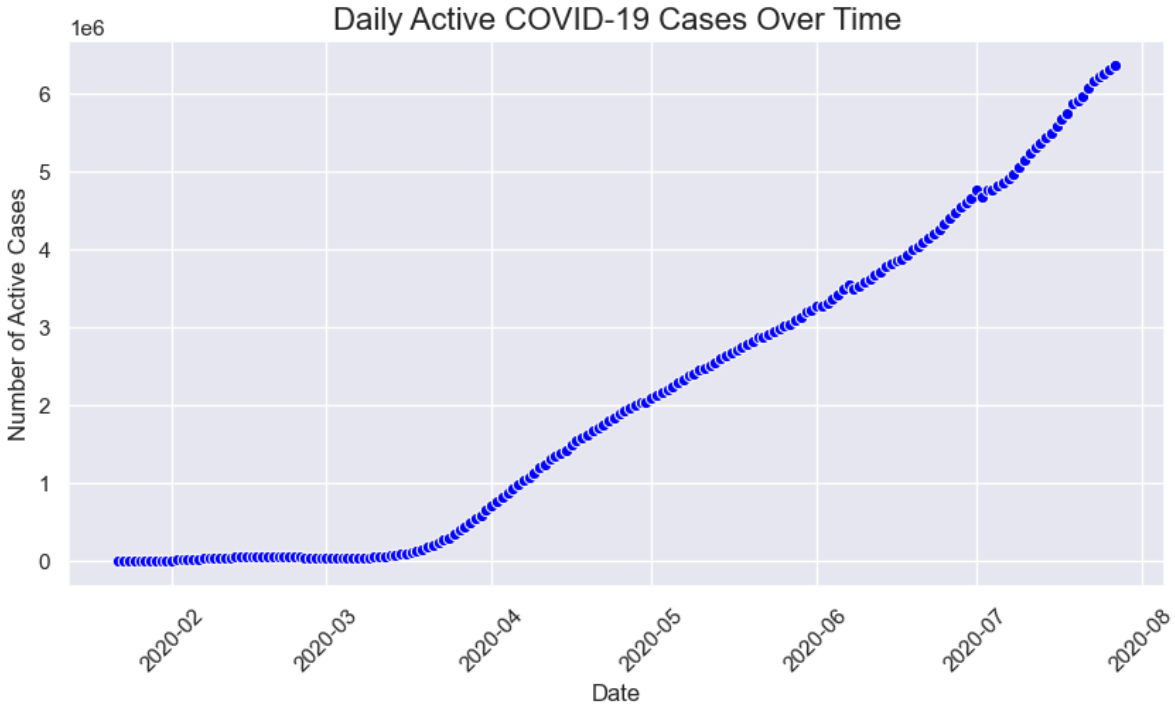


In [16]:
```
# Daily Trends: Confirmed Cases Over Time
daily_trends = data.groupby("Date")["Recovered"].sum()
plt.figure(figsize=(10, 5))
sns.set_theme(style="darkgrid")
sns.lineplot(data=daily_trends, marker="o", color="blue")
```

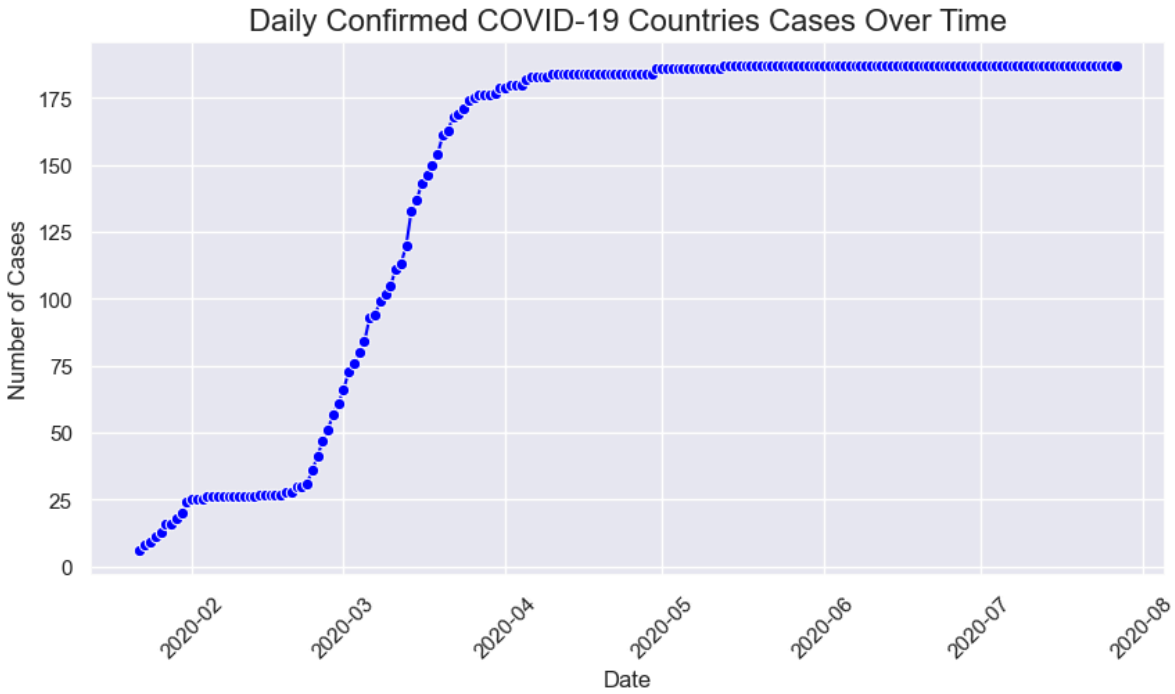
```
plt.title("Daily Recovered COVID-19 Cases Over Time", fontsize=16)
plt.xlabel("Date", fontsize=12)
plt.ylabel("Number of Recovered Cases", fontsize=12)
plt.xticks(rotation=45)
plt.show()
```



In [17]:
```
# Daily Trends: Confirmed Cases Over Time
daily_trends = data.groupby("Date")["Active"].sum()
plt.figure(figsize=(10, 5))
sns.set_theme(style="darkgrid")
sns.lineplot(data=daily_trends, marker="o", color="blue")
plt.title("Daily Active COVID-19 Cases Over Time", fontsize=16)
plt.xlabel("Date", fontsize=12)
plt.ylabel("Number of Active Cases", fontsize=12)
plt.xticks(rotation=45)
plt.show()
```



In [18]:
```
# Daily Trends: Number of Countries Affected Per Day
daily_trends = data.groupby("Date")["No. of countries"].sum()
plt.figure(figsize=(10, 5))
sns.set_theme(style="darkgrid")
sns.lineplot(data=daily_trends, marker="o", color="blue")
plt.title("Daily Confirmed COVID-19 Countries Cases Over Time", fontsize=16)
plt.xlabel("Date", fontsize=12)
plt.ylabel("Number of Cases", fontsize=12)
plt.xticks(rotation=45)
plt.show()
```



Insights The plots above highlight a significant trend in the progression of Confirmed, Active, Recovered, and Death cases of the COVID-19 pandemic, with notable exponential increases beginning in April 2020. A closer look at the data reveals several key observations:

Active Cases: The number of Active cases began to rise earlier, specifically from mid-March 2020. This early surge can likely be attributed to the delayed distribution of vaccines in several countries, which allowed the virus to spread more rapidly before containment measures could be fully implemented.

Confirmed Cases: The overall number of Confirmed cases followed a similar trajectory, with a sharp increase seen in April 2020. This aligns with the general global spread of the virus and indicates the period when testing capacities and case reporting began to expand in many regions.

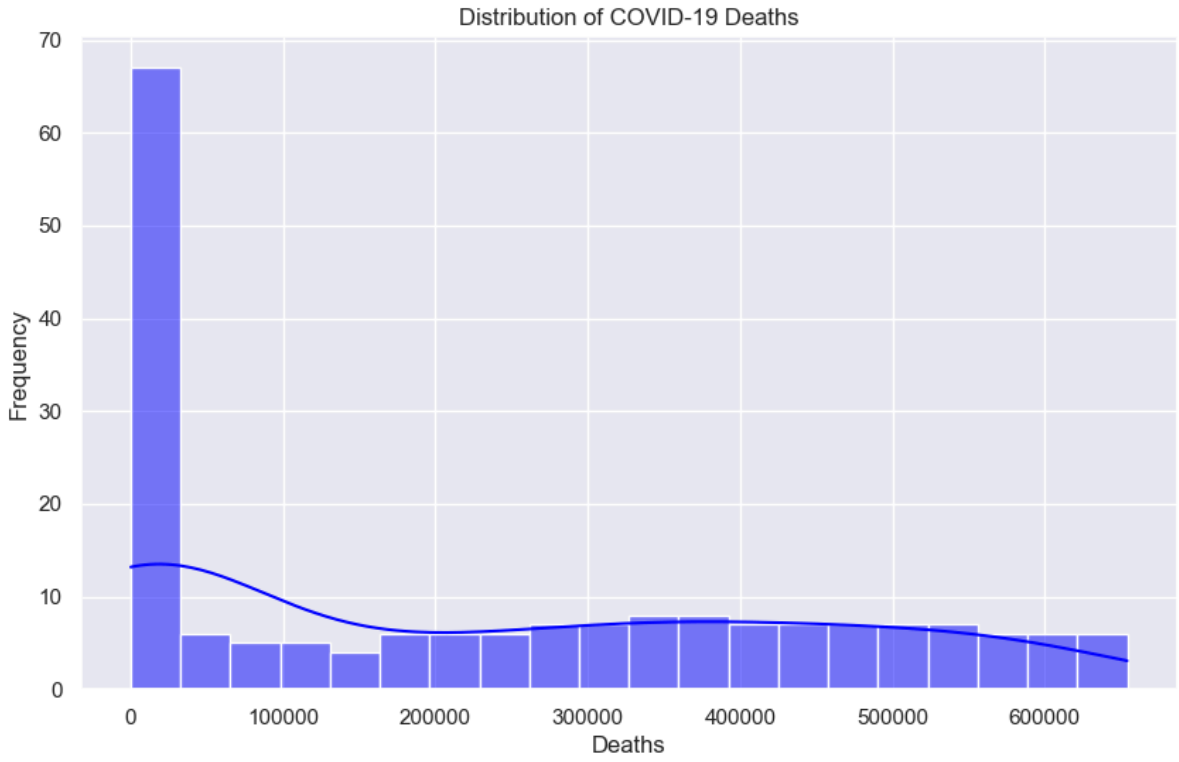
Deaths: Deaths also started to rise around mid-March 2020, coinciding with the increase in Active cases. This suggests a direct relationship between the two metrics, where a surge in Active cases often led to an increase in the number of fatalities. The trend emphasizes the severity of the virus, particularly when healthcare systems were overwhelmed, and treatments were not yet fully effective.

Correlation Between Active Cases and Deaths: The data shows a strong correlation between Active cases and Deaths. As Active cases surged, there was a parallel rise in the number of Deaths, indicating that the more individuals who were actively infected, the higher the

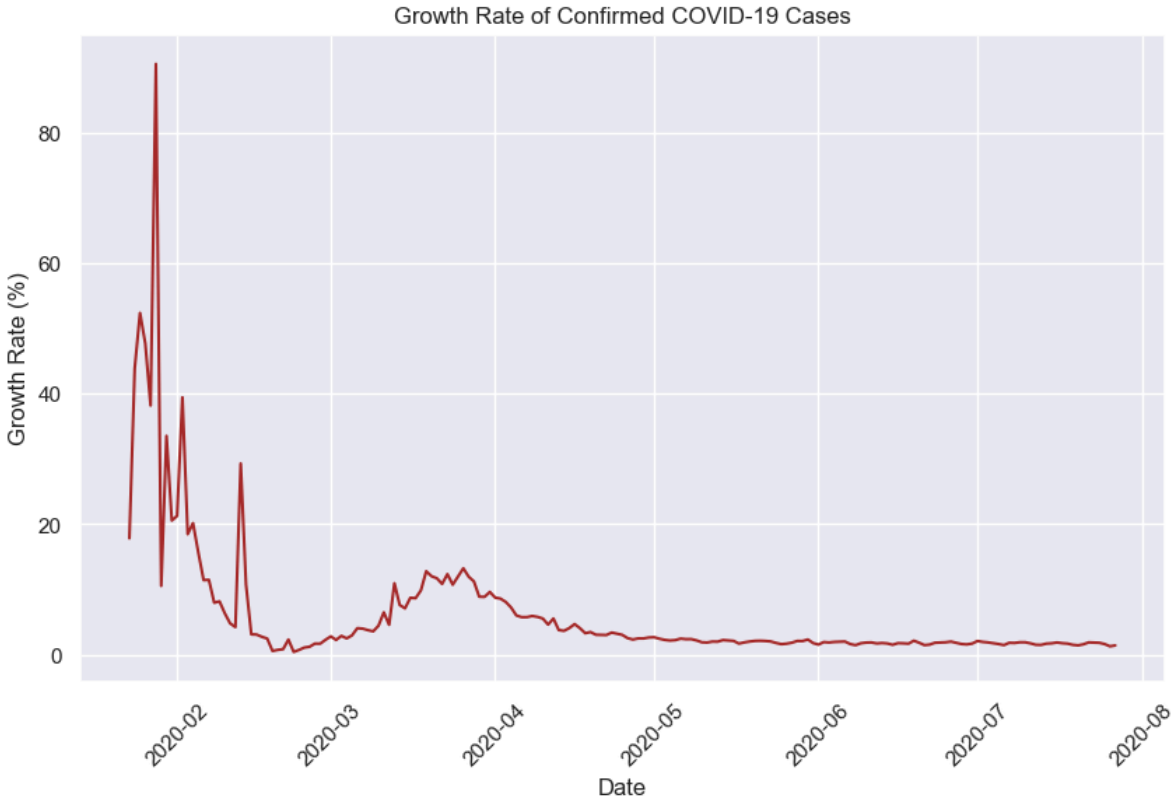
likelihood of fatal outcomes. This could be due to factors such as the strain on healthcare resources, the severity of infections, and the lack of widespread vaccinations during the early stages

Date with No. od Countries Been Affected: We can see that late february, there was an exponential rise in the number of countries been affected around the world. the pandemic. These insights reinforce the critical importance of timely intervention, vaccination efforts, and healthcare preparedness to mitigate the impact of future waves of infection. The data suggests that controlling the spread of the virus and managing the number of Active cases is key to reducing mortality rates during pandemics.

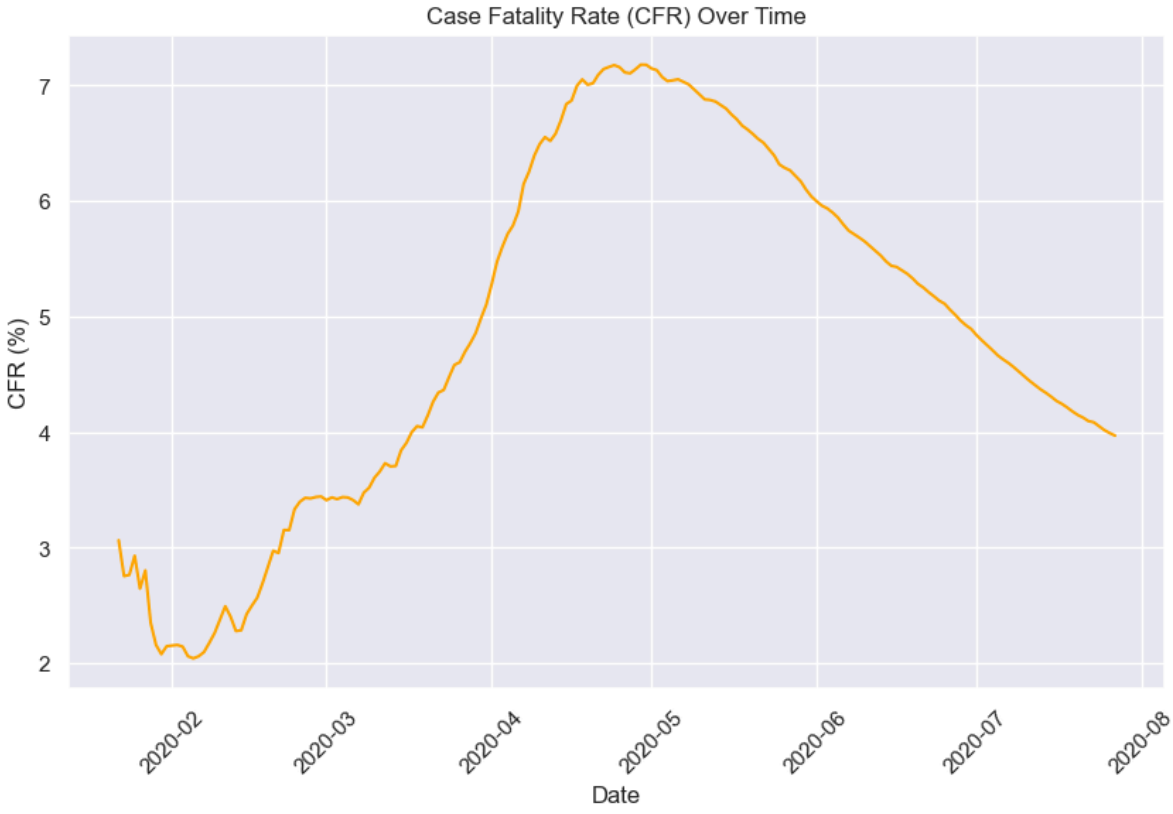
In [19]:
```
# Univariate Analysis for Deaths
plt.figure(figsize=(10, 6))
sns.histplot(data['Deaths'], bins=20, kde=True, color='blue')
plt.title('Distribution of COVID-19 Deaths')
plt.xlabel('Deaths')
plt.ylabel('Frequency')
plt.show()
```



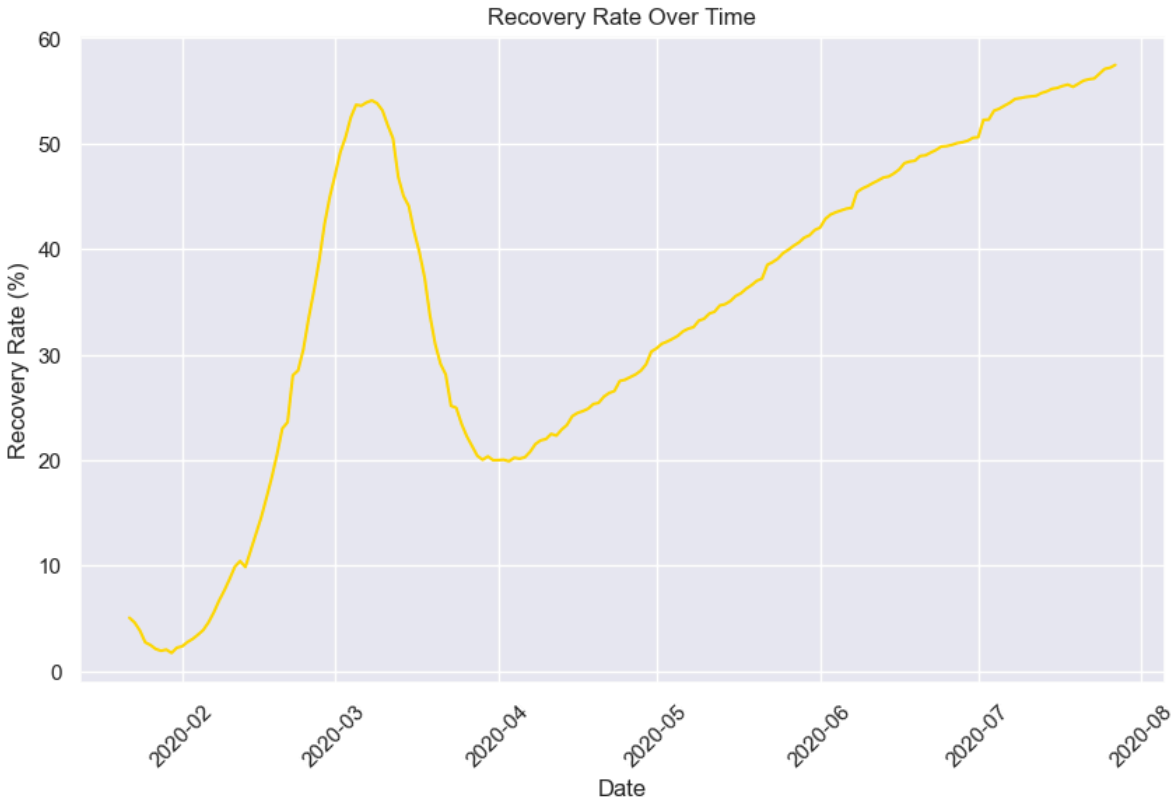
In [20]:
```
# Growth Rate for Confirmed Cases (Percent change per day)
data['Confirmed_Growth_Rate'] = data['Confirmed'].pct_change() * 100
plt.figure(figsize=(10, 6))
sns.lineplot(x=data.index, y=data['Confirmed_Growth_Rate'], color='brown')
plt.title('Growth Rate of Confirmed COVID-19 Cases')
plt.xlabel('Date')
plt.ylabel('Growth Rate (%)')
plt.xticks(rotation=45)
plt.show()
```



In [21]:
```
# Calculate Case Fatality Rate (CFR)
data['CFR'] = (data['Deaths'] / data['Confirmed']) * 100
plt.figure(figsize=(10, 6))
sns.lineplot(x=data.index, y=data['CFR'], color='orange')
plt.title('Case Fatality Rate (CFR) Over Time')
plt.xlabel('Date')
plt.ylabel('CFR (%)')
plt.xticks(rotation=45)
plt.show()
```



In [22]:
```
# Calculate Recovery Rate
data['Recovery_Rate'] = (data['Recovered'] / data['Confirmed']) * 100
plt.figure(figsize=(10, 6))
sns.lineplot(x=data.index, y=data['Recovery_Rate'], color='gold')
plt.title('Recovery Rate Over Time')
plt.xlabel('Date')
plt.ylabel('Recovery Rate (%)')
plt.xticks(rotation=45)
plt.show()
```

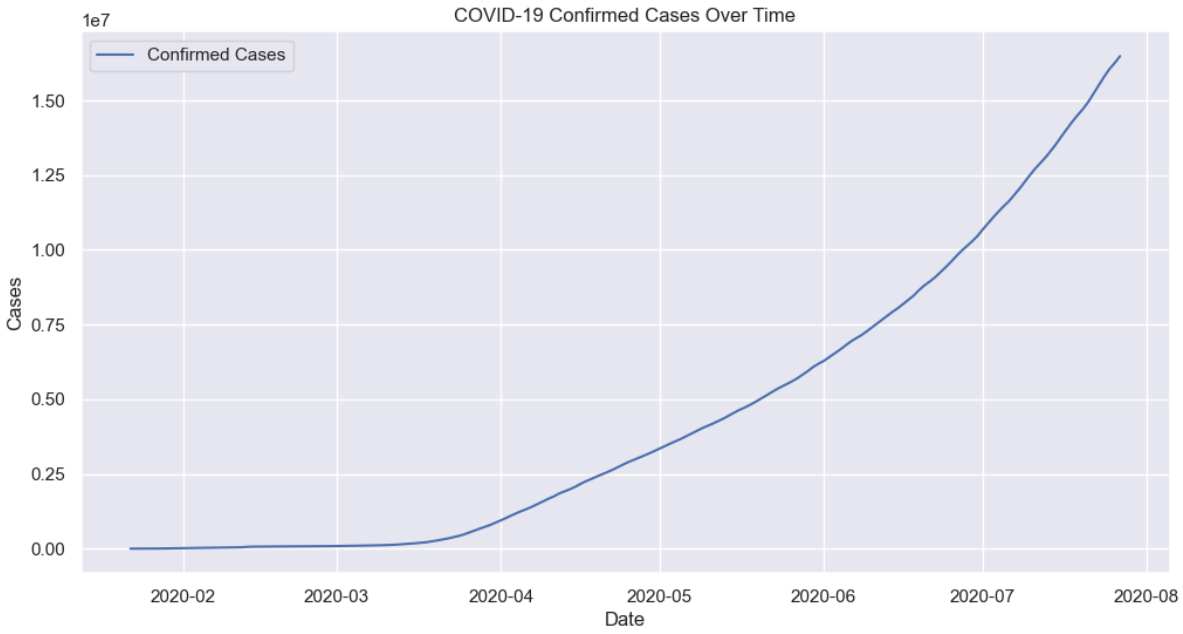


Univariate Analysis Insights on COVID-19 Data This analysis explores the distribution and trends of key COVID-19 variables, including confirmed cases, deaths, recoveries, and active cases. The insights are derived from the following visualizations:

Distribution of Confirmed COVID-19 Cases (Histogram) Insight: The histogram shows the distribution of confirmed COVID-19 cases over time. If the distribution is skewed to the right (positively skewed), it suggests that while most days had relatively low numbers of confirmed cases, some days had very high counts. A more uniform or bell-shaped distribution may suggest an even spread of cases. Interpretation: A right-skewed distribution often indicates an exponential or rapid increase in cases, common during outbreaks or peak periods of the pandemic. Growth Rate of Confirmed COVID-19 Cases (Line Plot) Insight: The line plot displays the daily growth rate of confirmed COVID-19 cases as a percentage change from the previous day. Sharp upward spikes indicate rapid growth on certain days, while sharp downward trends suggest slow or stable periods. Interpretation: High positive growth rates indicate a rapid increase in cases, usually observed during pandemic surges. A downward or flat growth rate indicates stabilization or reduction in case numbers, potentially due to control measures like lockdowns. Distribution of COVID-19 Deaths (Histogram) Insight: This histogram shows the distribution of deaths throughout the dataset period. A right-skewed distribution suggests that most deaths occurred on certain days, with some days having higher death tolls. A more evenly spread distribution would suggest a

consistent number of deaths across the timeline. Interpretation: A skewed distribution might indicate a concentrated period of high mortality, often during waves of infections. A uniform distribution indicates deaths were more evenly distributed throughout the time period. Case Fatality Rate (CFR) Over Time (Line Plot) Insight: This plot shows the Case Fatality Rate (CFR), which is the ratio of deaths to confirmed cases, over time. Fluctuations in the line suggest varying fatality rates during different periods. A rising CFR could signal worsening medical conditions or healthcare system strain. Interpretation: Fluctuations in CFR may reflect the varying severity of the disease or healthcare capacity challenges. A decrease in CFR could indicate improvements in medical treatments or interventions, while an increase might highlight overwhelmed healthcare systems. Distribution of COVID-19 Recoveries (Histogram) Insight: The histogram displays the distribution of recoveries over time. A left-skewed distribution suggests that recoveries were fewer than confirmed cases, while a right-skewed distribution indicates that most confirmed cases recovered. Interpretation: A large number of recoveries relative to deaths is a positive indicator of effective treatment and recovery processes. Conversely, a lower recovery count could signal a healthcare system under strain. Recovery Rate Over Time (Line Plot) Insight: The recovery rate, calculated as the ratio of recoveries to confirmed cases, is shown over time. A high recovery rate indicates good healthcare management and patient outcomes. Fluctuations in the line suggest variability in recovery rates over time, possibly due to changing treatment efficacy or healthcare system strain. Interpretation: A consistently high recovery rate signals effective treatment and management. A low or fluctuating recovery rate may indicate challenges in the healthcare system or emerging issues with treatment. Distribution of Active COVID-19 Cases (Histogram) Insight: This histogram shows the distribution of active cases, or cases that have not yet recovered or resulted in death. A high concentration of active cases suggests ongoing transmission or slow recovery rates. A lower number of active cases could indicate fewer infections or improvements in the recovery rate. Interpretation: A high number of active cases during certain periods indicates sustained transmission and a heavy ongoing burden on the healthcare system. A low count suggests the situation may be improving. Summary of Insights: Confirmed Cases: The distribution and growth rate provide insights into how quickly the pandemic spread. A right-skewed distribution suggests a rapid rise in cases, while a decreasing growth rate indicates a reduction in new cases. Deaths: The distribution and CFR help understand the mortality trends and the effectiveness of public health measures. Recoveries: The distribution of recoveries and recovery rate indicate how effectively patients are recovering, with a high recovery rate being a positive sign of medical management. Active Cases: The active case distribution and boxplot reveal ongoing transmission trends and potential healthcare strain, with fluctuations pointing to periods of significant outbreak.

In [23]:
```
confirmed_cases = data['Confirmed']
plt.figure(figsize=(12, 6))
plt.plot(confirmed_cases, label="Confirmed Cases")
plt.title("COVID-19 Confirmed Cases Over Time")
plt.xlabel("Date")
plt.ylabel("Cases")
plt.legend()
plt.show()
```



In [24]:
```
data.head()
```

Out[24]:

Date	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	De Reco
2020-01-22	555	17	28	510	0	0	0	3.06	5.05	
2020-01-23	654	18	30	606	99	1	2	2.75	4.59	
2020-01-24	941	26	36	879	287	8	6	2.76	3.83	
2020-01-25	1434	42	39	1353	493	16	3	2.93	2.72	
2020-01-26	2118	56	52	2010	684	14	13	2.64	2.46	

In [ ]: