COVID-19 in Australia

Introduction

The primary objective of this assignment is to analyze COVID-19 in Australia across five different states: NSW, QLD, VIC, SA and WA. The data used for analysis is taken from **COVID Live** (https://covidlive.com.au/). The data taken from this source contains two primary datasets: daily cases and daily deaths. For this analysis, the DATE & NEW columns are considered for daily cases and the DATE & DEATHS columns for daily deaths. It's important to note that the data available is on a daily basis until Sept 9, 2022 and after Sept 9, 2022, the data is on a weekly basis. To ensure meaningful comparison, the data has to be aggregated on a **weekly basis**, as a result the number of observations will vary. The types of analysis planned to examine are distribution analysis of weekly cases and deaths for five states, historical cumulative analysis of new cases for five states, normalized analysis by plotting a historical graph of normalized new cases by population for five states and study the relationship between new cases and deaths in five states.

In [1]:

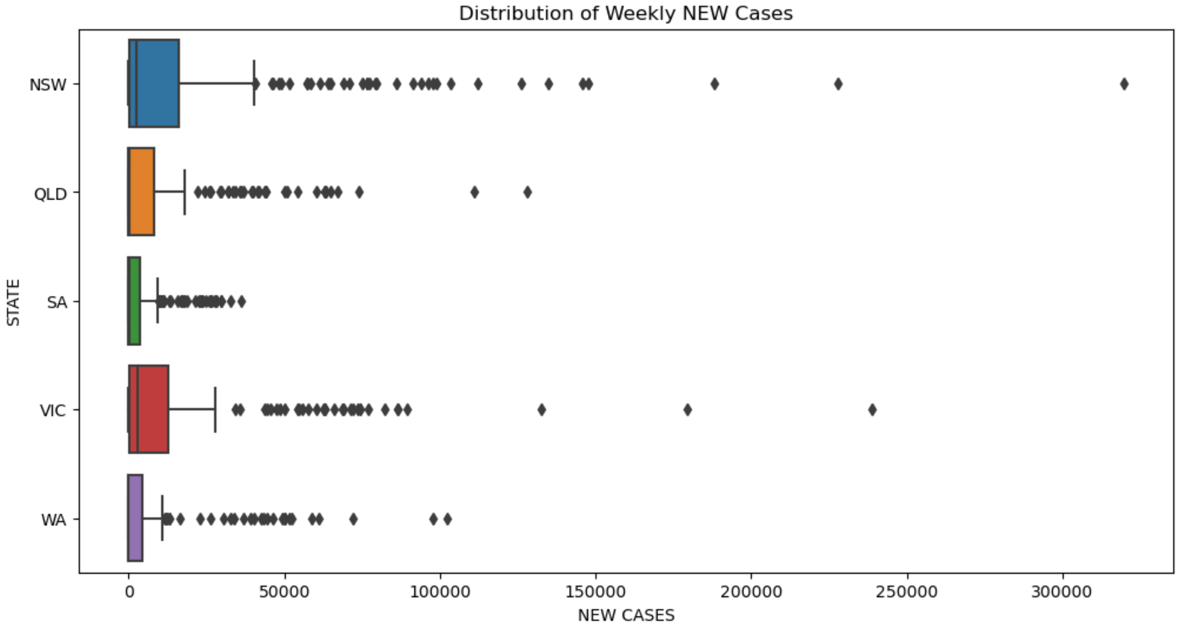
```
import numpy as np
import pandas as pd
import zipfile
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

weekly_cases_data = pd.DataFrame()
weekly_deaths_data = pd.DataFrame()

# Function to load and process covid data from each tsv file
def load_data(file_name, column_name):
    covid_data = pd.read_csv(file_name, sep='\t', usecols=['DATE', column_name], pa
    state = file_name.split('_')[-1].split('.')[0].upper()
    covid_data['STATE'] = state
    covid_data[column_name] = pd.to_numeric(covid_data[column_name].str.replace(',
    covid_data = covid_data.dropna(subset=[column_name])
    # Aggregating on weekly basis
    if column_name == 'NEW':
        covid_data = covid_data.groupby(['STATE', pd.Grouper(key='DATE', freq='W')
    elif column_name == 'DEATHS':
        covid_data = covid_data.groupby(['STATE', pd.Grouper(key='DATE', freq='W')
    return covid_data

# Loading data from zip file
with zipfile.ZipFile('covid_data.zip', 'r') as zip_file:
    for file_name in zip_file.namelist():
        if file_name.startswith('daily_cases'):
            cases_data = load_data(file_name, 'NEW')
            cases_data = cases_data[cases_data['NEW'] >= 0] # removing negative va
            cases_data['CUMULATIVE'] = cases_data.groupby('STATE')['NEW'].cumsum()
            weekly_cases_data = pd.concat([weekly_cases_data, cases_data], ignore_i

        elif file_name.startswith('daily_death'):
            deaths_data = load_data(file_name, 'DEATHS')
            weekly_deaths_data = pd.concat([weekly_deaths_data, deaths_data], ignor
```

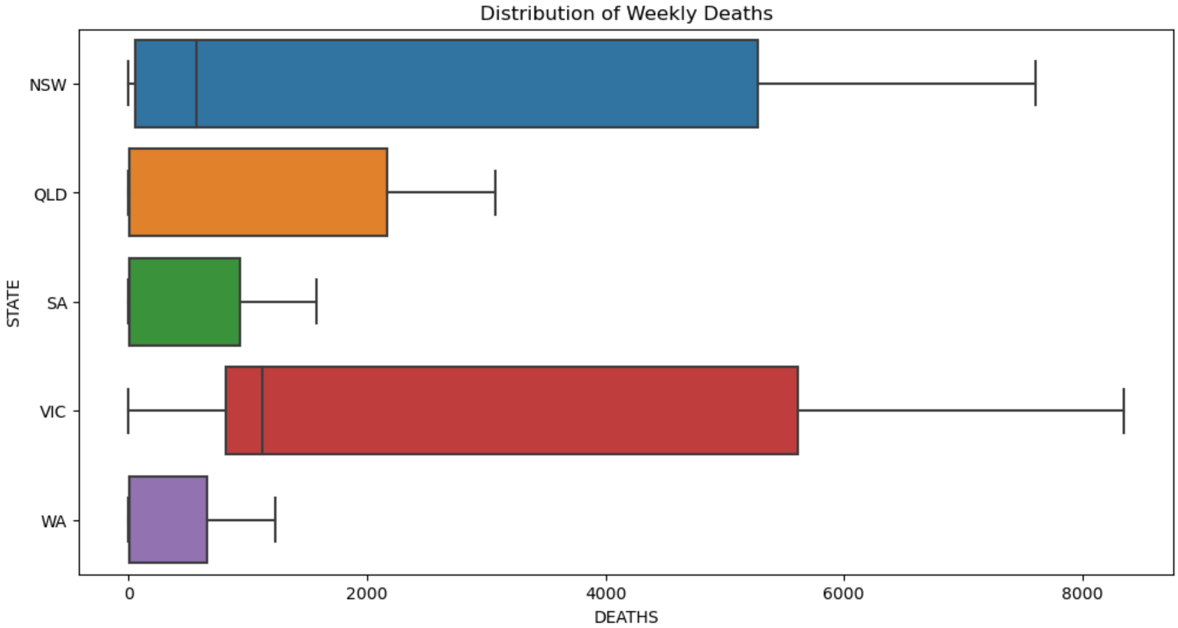
Distributions of new cases and deaths weekly numbers in five states.

In [2]:

```
# Fuction to plot boxplot and to get descriptive statistics for weekly cases and de
def plot_and_describe(data, column_name, xlabel, title):
    plt.figure(figsize=(12, 6))
    sns.boxplot(data=data, x=column_name, y="STATE").set(title=title, xlabel=xlabe

    statistics = data.groupby("STATE")[column_name].agg([np.min, np.max, np.mean,
    statistics['IQR'] = (data.groupby("STATE")[column_name].quantile(0.75) - data.
    statistics['Skewness'] = data.groupby("STATE")[column_name].skew().round(2)

    return statistics

# Plot and describe weekly new cases
weekly_cases_statistics = plot_and_describe(weekly_cases_data, "NEW", "NEW CASES",
print(f'Descriptive Statistics for weekly new cases\n')
print(weekly_cases_statistics)

# Plot and describe weekly deaths
weekly_deaths_statistics = plot_and_describe(weekly_deaths_data, "DEATHS", "DEATHS
print(f'\nDescriptive Statistics for weekly deaths\n')
print(weekly_deaths_statistics)
```

```
Descriptive Statistics for weekly new cases

       amin    amax   mean  median    std    IQR  Skewness
STATE
NSW       0  319632  22084    2611  43309  16199      3.36
QLD       0  127914  10052      46  20147   8006      2.87
SA        0   36203   4699      30   8383   3758      1.98
VIC       0  238588  16385    3016  31856  12612      3.41
WA        0  102305   7343      23  17059   4525      3.14

Descriptive Statistics for weekly deaths

       amin  amax  mean  median   std   IQR  Skewness
STATE
NSW       0  7603  2282     572  2734  5220      0.73
QLD       0  3075   885       7  1167  2157      0.84
SA        0  1573   414       4   563   933      0.95
VIC       0  8347  2817    1125  2823  4798      0.76
WA        0  1231   292       9   414   644      1.03
```





Weekly New Cases

The descriptive statistics for weekly cases represent the minimum value of zero across all states, indicating weeks with no reported cases, as expected at the start of the pandemic. However, the maximum values vary significantly, with NSW reporting the highest weekly count of 319,632, followed by VIC (238,588), QLD (127,914), WA (102,305) and SA (36,203). Since the skewness value is greater than 1 for all states, indicating positive skewness. Hence, to measure central tendency and dispersion, the median and Interquartile range in the descriptive statistics are considered. The distribution of weekly cases also demonstrated that it is right skewed and for some states, median is overlapped with first quartile. The presence of outliers for all states, emphasizes the variability in new cases across the weeks.

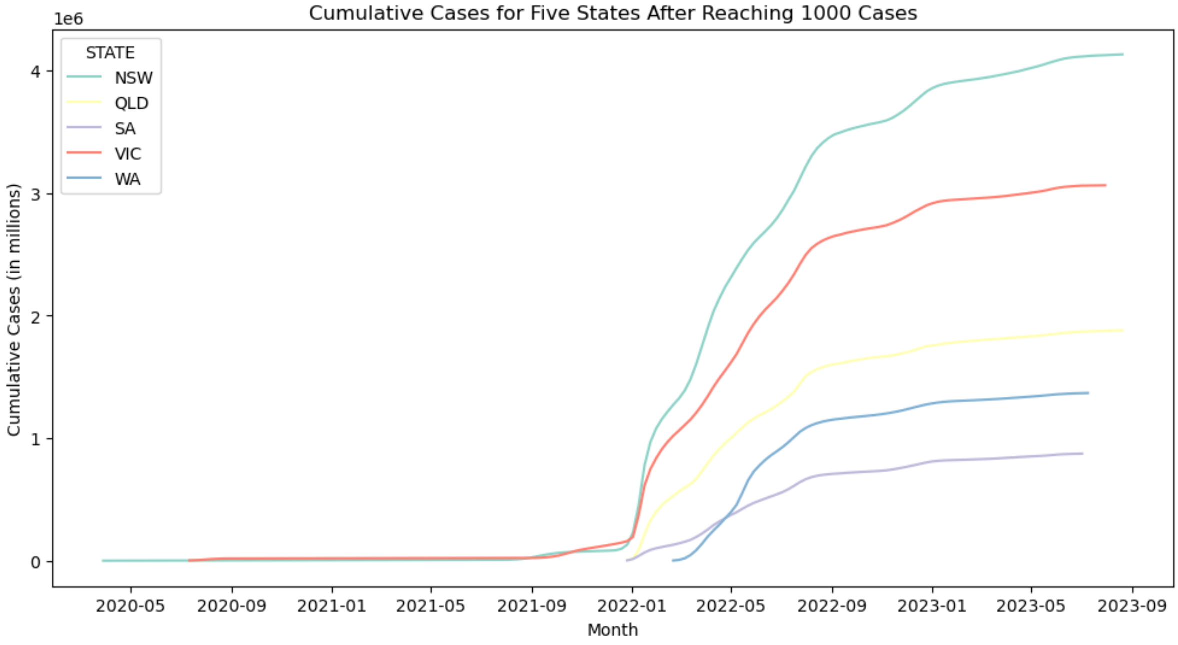
Weekly Deaths

The descriptive statistics for weekly deaths represent a minimum value of zero similar to cases, indicating no deaths reported. The maximum values are highest for VIC (8347) and NSW (7603), followed by QLD (3075), SA (1573) and WA (1231). For central tendency and dispersion, median and Interquartile range in descriptive statistics are considered, as skewness is close to 1 for all states. The distribution of weekly deaths shows it is right skewed for all states and and for some states median is overlapped with first quartile.

History of COVID-19 in different states

```
In [3]: # To create cumulative values and plot line chart for five states after 1000 cases
        weekly_cases_data_after_1000_cases = weekly_cases_data[weekly_cases_data['NEW'] >=
        plt.figure(figsize=(12, 6))
        sns.lineplot(weekly_cases_data_after_1000_cases, x='DATE', y='CUMULATIVE', hue='ST/
        plt.title('Cumulative Cases for Five States After Reaching 1000 Cases')
        plt.xlabel('Month')
        plt.ylabel('Cumulative Cases (in millions)')
        cumulative_cases = weekly_cases_data.groupby("STATE")['NEW'].agg(Cumulative_Count=
        cumulative_cases.columns = ['Cumulative cases for each state']
        print(cumulative_cases)
```

```
       Cumulative cases for each state
STATE
NSW                            4129761
QLD                            1879840
SA                              878895
VIC                            3064001
WA                             1373207
```



Based on the history of COVID-19 cases in Australia, it is observed that NSW has the highest cumulative of around 4 million, with 1000 cases reported from March 2020. Over time, the cumulative case count increased significantly for NSW. On the other hand, SA has the lowest cumulative cases among other states which is less than a million and the first 1000 cases were reported in March 2022. Following NSW, VIC ranks second with a cumulative case count of over 3 million, reporting its first 1000 cases from July 2020, QLD has a cumulative count approaching 2 million and WA has a cumulative count exceeding 1.3 million, both 1000 cases reporting began around Jan 2021.

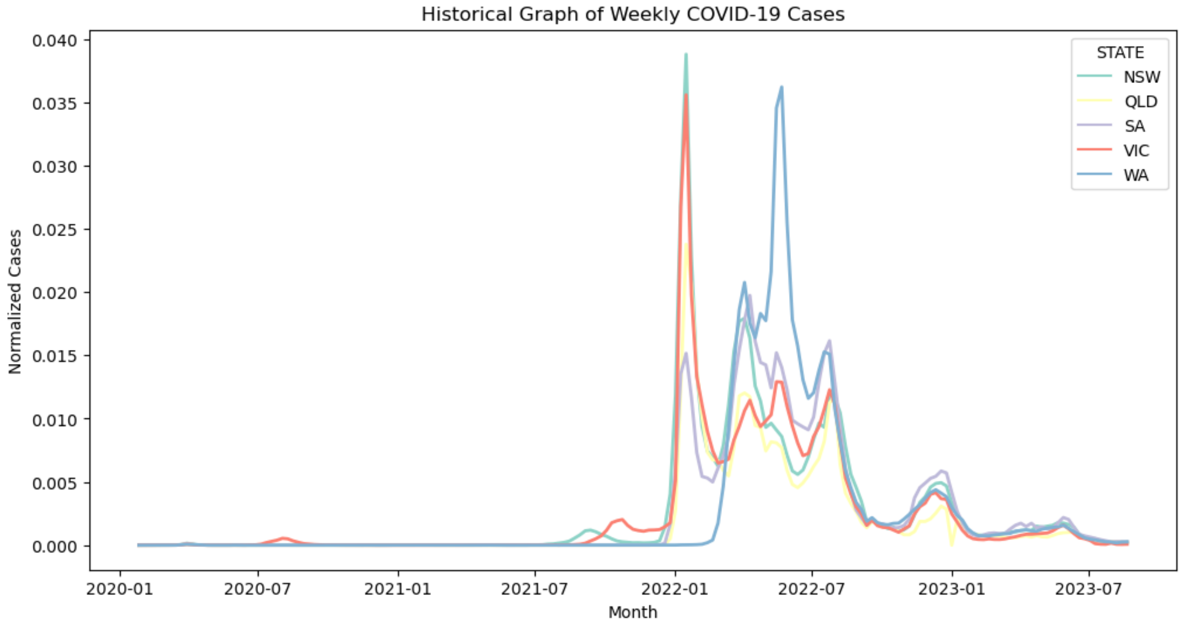
Historical graph of normalised weekly cases.

In [4]:

```
# Creating population data dataframe
population_data = pd.DataFrame({
    'STATE': ['NSW', 'VIC', 'SA', 'QLD', 'WA'],
    'Population at 31 December 2022 (\'000)': [8238.8, 6704.3, 1834.3, 5378.3, 282
})
# To plot a historical graph of normalized weekly cases.
weekly_cases_data = pd.merge(weekly_cases_data, population_data, on='STATE', how='
weekly_cases_data['NORMALISED'] = weekly_cases_data['NEW'] / (weekly_cases_data["P
plt.figure(figsize=(12, 6))

sns.lineplot(data=weekly_cases_data, x='DATE', y='NORMALISED', hue='STATE', palett
plt.title('Historical Graph of Weekly COVID-19 Cases')
plt.xlabel('Month')
plt.ylabel('Normalized Cases')
```

Out[4]:

```
Text(0, 0.5, 'Normalized Cases')
```

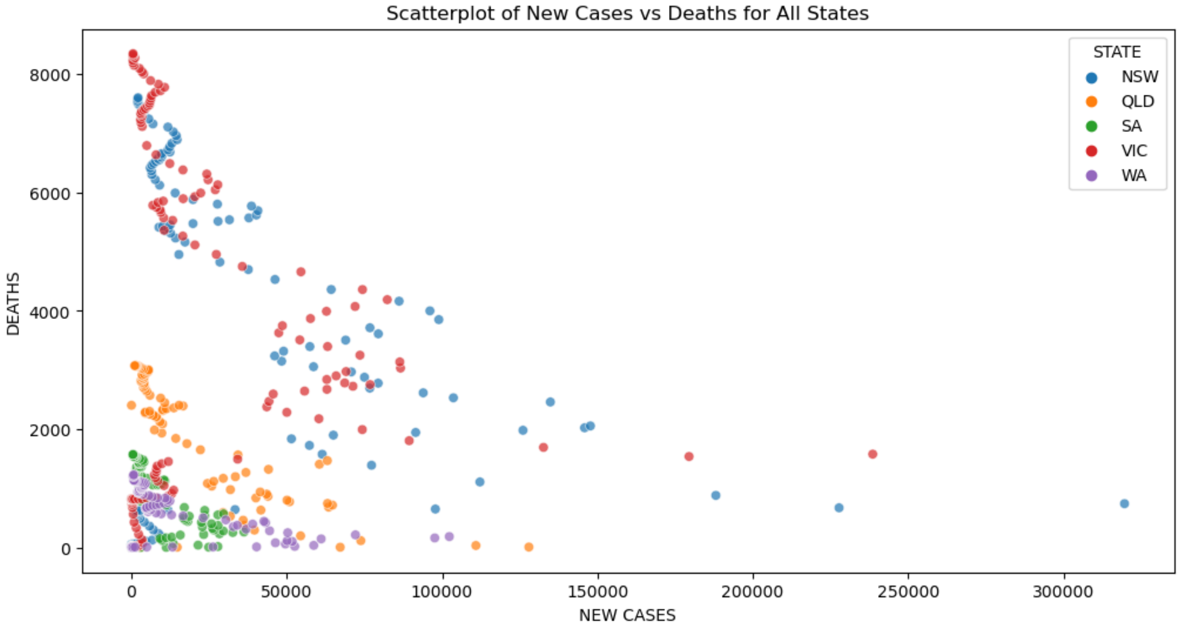


In this analysis, weekly cases for five states are normalized by the population of each state. NSW exhibits the highest peak around 0.039 on Jan 2022, indicating a higher case rate per population. VIC and WA have similar levels of impact, with their peaks around 0.036 in Jan 2022 and July 2022 respectively. Even though QLD has more population than WA, its peak is around 0.025, indicating a comparatively lower impact per population on Jan 2022. SA maintains the lowest peak normalized case rate among all other states, having a peak of around 0.020 in March 2022.

Relationship between number of new cases and deaths in five states.

In [5]:

```
# To find the relationship between NEW and DEATH cases
merged_data = pd.merge(weekly_cases_data, weekly_deaths_data, on=['DATE', 'STATE']
plt.figure(figsize=(12, 6))
# To plot scatterplot with NEW column in x-axis and DEATHS column in y-axis
sns.scatterplot(data=merged_data, x='NEW', y='DEATHS', hue='STATE', alpha=0.7)
plt.title('Scatterplot of New Cases vs Deaths for All States')
plt.ylabel('DEATHS')
plt.xlabel('NEW CASES')
plt.legend(title='STATE')
plt.show()
```



In the above analysis, to find the relationship between new cases and deaths, the scatterplot is plotted with NEW on the x-axis and DEATHS on the y-axis. It can be observed that new cases in NSW and VIC reported are high compared to other states, hence the deaths are also high for these two states. This indicates that **higher new cases are associated with more deaths**. For QLD, the new cases and deaths are moderate when compared to the other four states. SA and WA have comparatively less cases and deaths when compared to other states. So, it can be concluded that 'Higher the new cases, higher the death count and lower the new cases, lower death count is reported.

Conclusion

To analyze COVID-19 in Australia in five states, 4 different analyses were performed. Firstly, the distribution of new weekly cases and deaths is studied, which indicates that the distribution of weekly cases for five states are right skewed and has outliers and the distribution of weekly deaths for five states are also right skewed. It was observed from skewness value(greater than 1). Basic statistics were calculated to measure central tendency and dispersion. NSW reported high cases, followed by VIC, QLD, WA and SA. Next, the history of COVID-19 in five different states starting from the week after 1000 cases were reported. The line chart was plotted with cumulative weekly numbers, demostrating that NSW has the highest cumulative cases of around 4 million and SA has the lowest cumulative case below 1 million. Furthermore, weekly new cases are normalized by the population of each state in the next analysis. NSW exhibited a higher case rate per population. VIC and WA had similar impacts, followed by QLD, while SA had the lowest case rate per population. Finally, the relationship between new cases and deaths across five states is studied. The scatterplot revealed that higher new cases are associated with more deaths, indicating a connection between case count and mortality.