COVID 19 in Australia Analysis and Visualization

Author: Cynthia Mueni Kiiti

Introduction

This analysis and visualization of COVID-19 data aims to shed light on key aspects of the pandemic in Australia, from tracking a total of ten datasets pertaining to COVID-19 cases and deaths within five prominent Australian states, specifically referred to as 'New South Wales,' 'Queensland,' 'South Australia,' 'Victoria,' and 'Western Australia.' These datasets contain information regarding the daily COVID-19 cases and fatality counts up to September 9, 2022. However, due to alterations in government reporting regulations, both CASES and DEATHS datasets transition from daily figures to weekly figures. This analysis is conducted on weekly basis to allow for easier comparison.

For the datasets containing new case information, there are structured with an index representing dates and four other columns. The 'NEW' column predominantly contains the daily and weekly reported COVID-19 cases, as such, our primary focus will be directed towards this variable.

Furthermore, the datasets containing daily and weekly deaths exhibit similarities with the case datasets, featuring variables similar to those previously mentioned. Specifically, the 'DEATHS' variable is of significance. However, it is important to note that it maintains a cumulative nature. Consequently, a new column will be generated to transform this data into an unaccumulated format, thereby providing us with weekly fatality figures within these five states. This column is 'NEWDEATHS'.

In [70]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

import os
```

Analysing Distribution of data for NSW.

In [71]:
```
# TO READ THE DATA
# Specifying the directory path and file names
directory_path =  'C:/Users/amwikali/Desktop/MSc. DATA SCIENCE/S.P for Data Science/As

# cases tab
file_name = 'daily_cases_nsw.tsv'

file_path = directory_path + file_name
```

```
cases_nsw = pd.read_csv(file_path, delimiter='\t')

# deaths tab
file_name2 = 'daily_death_nsw.tsv'

file_path2 = directory_path + file_name2

deaths_nsw = pd.read_csv(file_path2, delimiter='\t')
```

In [72]:
```
# NSW CASES Analysis

## Cleaning the columns NEW, CASES and NET for the analysis

cases_nsw['NEW'].fillna('0', inplace=True) # fill in missing values
cases_nsw[['NEW', 'CASES', 'NET']] = cases_nsw[['NEW', 'CASES', 'NET']].astype(str) #
cases_nsw[['NEW', 'CASES', 'NET']] = cases_nsw[['NEW', 'CASES', 'NET']].replace(',','
cases_nsw[['NEW', 'CASES', 'NET']] = cases_nsw[['NEW', 'CASES', 'NET']].replace('-','(
cases_nsw[['NEW', 'CASES', 'NET']] = cases_nsw[['NEW', 'CASES', 'NET']].astype(int) #(
```

In [73]:
```
## Time indexing on the Dataframe

cases_nsw['DATE'] = pd.to_datetime(cases_nsw['DATE'])
cases_nsw.set_index('DATE', inplace=True)
```

In [74]:
```
## Converting the time-indexed data into weekly intervals, while calculating sum of th

cases_nsw_weekly = cases_nsw.resample('W').sum() # resamples the data into weekly
cases_nsw_weekly["TOTALCASES"] = cases_nsw_weekly["NEW"].cumsum()
cases_nsw_weekly["STATE"] = 'NSW'
```

In [75]:
```
# DEATHS NSW Analysis
## We follow the same steps as we did for the CASES data
## Cleaning the columns DEATHS and NET for the analysis

deaths_nsw[['DEATHS', 'NET']] = deaths_nsw[['DEATHS', 'NET']].astype(str)
deaths_nsw[['DEATHS', 'NET']] = deaths_nsw[['DEATHS', 'NET']].replace(',', '', regex=T
deaths_nsw[['DEATHS', 'NET']] = deaths_nsw[['DEATHS', 'NET']].replace('-', '0')
deaths_nsw[['DEATHS', 'NET']] = deaths_nsw[['DEATHS', 'NET']].astype(int)
```

In [76]:
```
## Time indexing on the Dataframe

deaths_nsw['DATE'] = pd.to_datetime(deaths_nsw['DATE'])
deaths_nsw.set_index('DATE', inplace=True)
```

In [77]:
```
## Converting the time-indexed data into weekly intervals, while calculating sum of th

deaths_nsw_weekly = deaths_nsw.resample('W').sum()
deaths_nsw_weekly['NEWDEATHS'] = deaths_nsw_weekly['DEATHS'].diff().fillna(0)  # calcu
```

In [78]:
```
# Merging the two data sets using the DATE column as the common key

mergednsw = pd.merge(cases_nsw_weekly, deaths_nsw_weekly, on='DATE')
print(mergednsw.describe())
```

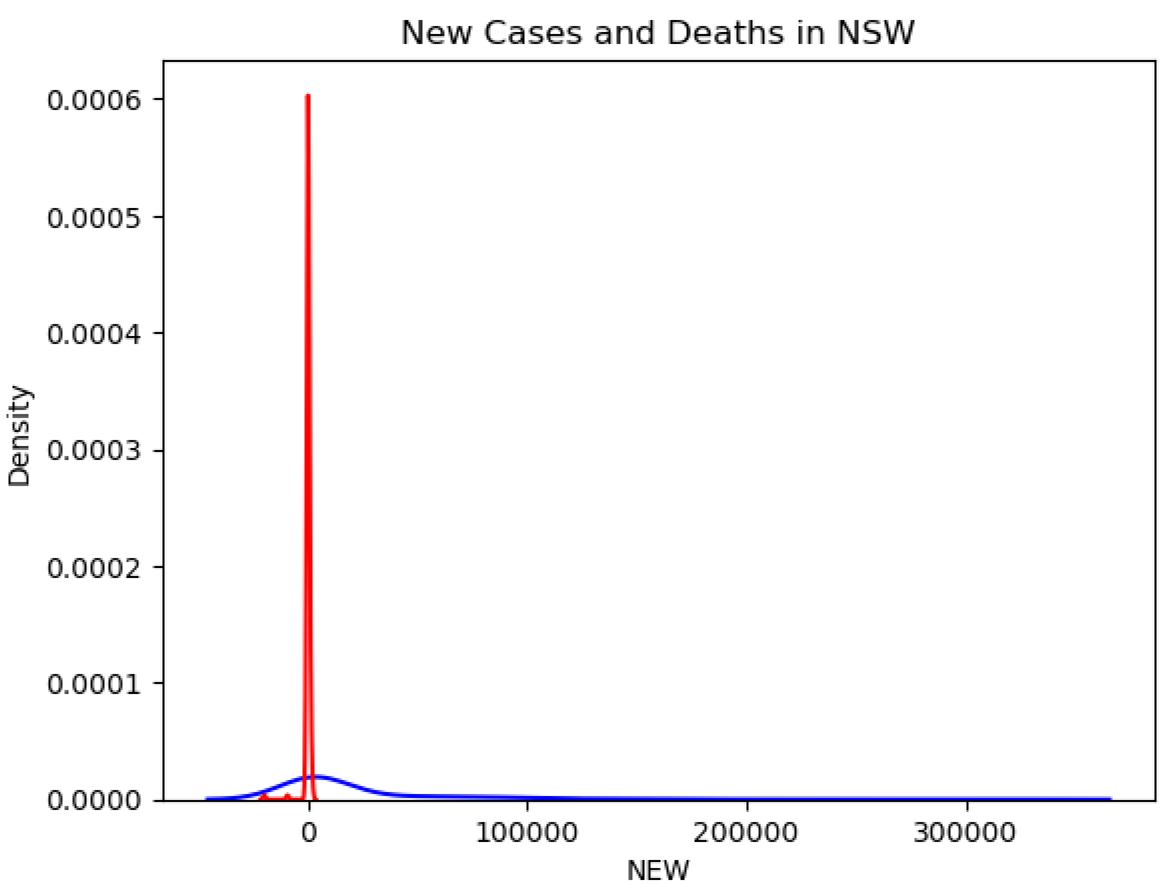
```
                NEW         CASES  VAR_x          NET_x    TOTALCASES  \
count    187.000000  1.870000e+02  187.0     187.000000  1.870000e+02
mean   22084.283422  4.013615e+06    0.0   21907.475936  1.467097e+06
std    43309.511449  6.298968e+06    0.0   43000.925656  1.727579e+06
min        0.000000  6.000000e+00    0.0       0.000000  0.000000e+00
25%       45.500000  3.257700e+04    0.0      44.000000  4.543000e+03
50%     2611.000000  5.325340e+05    0.0    2256.000000  7.784100e+04
75%    16245.000000  4.047765e+06    0.0   16202.500000  3.530475e+06
max   319632.000000  2.406021e+07    0.0  317657.000000  4.129761e+06

             DEATHS  VAR_y       NET_y     NEWDEATHS
count    187.000000  187.0  187.000000    187.000000
mean    5974.358289    0.0   40.657754     40.657754
std     8024.758731    0.0   57.130209   1660.120983
min        0.000000    0.0    0.000000 -19843.000000
25%      371.000000    0.0    0.000000      0.000000
50%     4080.000000    0.0   20.000000     32.000000
75%     6896.000000    0.0   64.000000    138.000000
max    34304.000000    0.0  401.000000   1852.000000
```

In [79]:

```python
# Plotting the kernel density estimate, allowing for smoother distribution of data

KDE_graph = sns.kdeplot(data = mergednsw, x = "NEW", color="blue", label="New Cases")
KDE_graph = sns.kdeplot(data = mergednsw, x = "NEWDEATHS", color="red", label="New Dea
KDE_graph.set_title("New Cases and Deaths in NSW")
```

Out[79]:

```
Text(0.5, 1.0, 'New Cases and Deaths in NSW')
```



Based on the statistical data presented above, from comparing the 'NEW' column in cases_nsw and 'NEWDEATHS' column in deaths_nsw, it is evident that the average number of new COVID-

19 cases per week in New South Wales is 22,084.3, while the average number of deaths is 40.6. Consequently, it can be deduced that out of the 22,084 individuals who contracted COVID-19, approximately 40 individuals succumbed to the disease. This suggests that approximately 22,044 individuals have successfully recovered from the illness in New South Wales.

From the plot of the kernel density estimate, it becomes apparent that the blur curve representing new cases exhibits a broader dispersion around the mean, indicating a higher standard deviation of 43,309.5. Conversely, the red curve representing new deaths appears more concentrated, signifying a lower standard deviation for this variable, which is 1,660.1.

Analysing Distribution of data for QLD.

In [80]:
```
# Reading the data

file_name = 'daily_cases_qld.tsv'

file_path = directory_path + file_name

cases_qld = pd.read_csv(file_path, delimiter='\t')

#deaths tab
file_name2 = 'daily_death_qld.tsv'

file_path2 = directory_path + file_name2

deaths_qld = pd.read_csv(file_path2, delimiter='\t')
```

In [81]:
```
# QLD CASES Analysis
## Cleaning the columns NEW, CASES and NET for the analysis

cases_qld['NEW'].fillna('0', inplace=True)
cases_qld[['NEW', 'CASES', 'NET']] = cases_qld[['NEW', 'CASES', 'NET']].astype(str)
cases_qld[['NEW', 'CASES', 'NET']] = cases_qld[['NEW', 'CASES', 'NET']].replace(',','
cases_qld[['NEW', 'CASES', 'NET']] = cases_qld[['NEW', 'CASES', 'NET']].replace('-','0
cases_qld[['NEW', 'CASES', 'NET']] = cases_qld[['NEW', 'CASES', 'NET']].astype(int)
```

In [82]:
```
## Time indexing on the Dataframe

cases_qld['DATE'] = pd.to_datetime(cases_qld['DATE'])
cases_qld.set_index('DATE', inplace=True)
```

In [83]:
```
## Converting the time-indexed data into weekly intervals, while calculating sum of th

cases_qld_weekly = cases_qld.resample('W').sum()
cases_qld_weekly["TOTALCASES"] = cases_qld_weekly['NEW'].cumsum()
cases_qld_weekly["STATE"] = 'QLD'
```

In [84]:
```
# DEATHS QLD Analysis
## Cleaning the columns DEATHS and NET for the analysis

deaths_qld[['DEATHS', 'NET']] = deaths_qld[['DEATHS', 'NET']].astype(str)
deaths_qld[['DEATHS', 'NET']] = deaths_qld[['DEATHS', 'NET']].replace(',', '', regex=T
```

```
deaths_qld[['DEATHS', 'NET']] = deaths_qld[['DEATHS', 'NET']].replace('-', '0')
deaths_qld[['DEATHS', 'NET']] = deaths_qld[['DEATHS', 'NET']].astype(int)
```

In [85]:

```
## Time indexing on the Dataframe

deaths_qld['DATE'] = pd.to_datetime(deaths_qld['DATE'])
deaths_qld.set_index('DATE', inplace=True)
```

In [86]:

```
## Converting the time-indexed data into weekly intervals, while calculating sum of th

deaths_qld_weekly = deaths_qld.resample('W').sum()
deaths_qld_weekly['NEWDEATHS'] = deaths_qld_weekly['DEATHS'].diff().fillna(0)
```

In [87]:

```
# Merging the two data sets using the DATE column as the common key

mergedqld = pd.merge(cases_qld_weekly, deaths_qld_weekly, on='DATE')
print(mergedqld.describe())
```

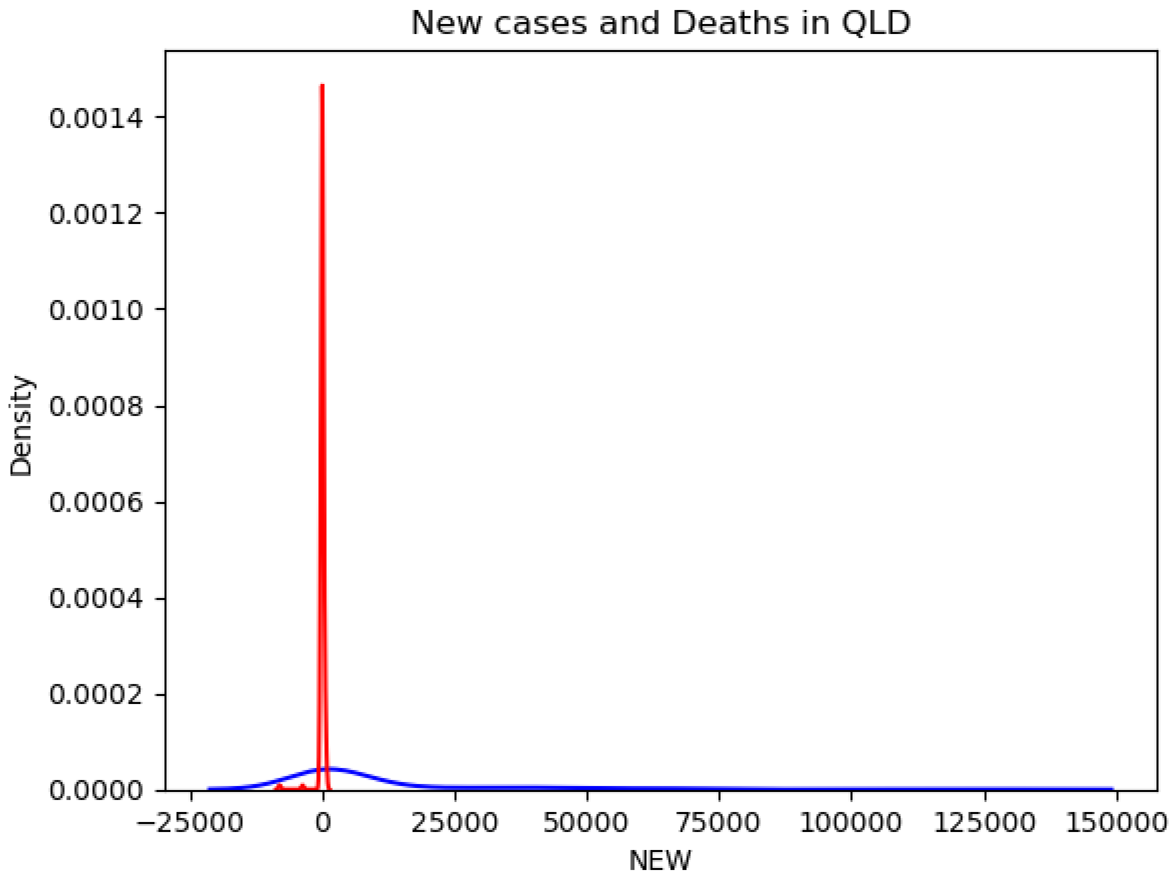
```
                 NEW         CASES  VAR_x            NET_x    TOTALCASES  \
count     187.000000  1.870000e+02  187.0      187.000000  1.870000e+02
mean    10052.620321  1.761785e+06    0.0     9240.743316  6.604294e+05
std     20147.386854  2.899928e+06    0.0    23308.541744  7.940911e+05
min         0.000000  0.000000e+00    0.0 -146726.000000  0.000000e+00
25%         9.000000  8.585500e+03    0.0        9.000000  1.253000e+03
50%        46.000000  1.464800e+04    0.0       39.000000  2.082000e+03
75%      8015.500000  1.727515e+06    0.0     7570.000000  1.634796e+06
max    127914.000000  1.125113e+07    0.0   123179.000000  1.879840e+06

             DEATHS  VAR_y       NET_y   NEWDEATHS
count    187.000000  187.0  187.000000   187.00000
mean    1995.219251    0.0   16.443850    16.44385
std     3001.049243    0.0   26.001464   673.62486
min        0.000000    0.0    0.000000 -8012.00000
25%       42.000000    0.0    0.000000     0.00000
50%       49.000000    0.0    0.000000     0.00000
75%     2901.000000    0.0   22.500000    21.50000
max    13804.000000    0.0  104.000000   702.00000
```

In [88]:

```
# Plotting the kernel density estimate

KDE_graph = sns.kdeplot(data = mergedqld, x = "NEW", color="blue", label="New Cases")
KDE_graph = sns.kdeplot(data = mergedqld, x = "NEWDEATHS", color="red", label="New Dea
KDE_graph.set_title("New cases and Deaths in QLD")
```

Out[88]:

```
Text(0.5, 1.0, 'New cases and Deaths in QLD')
```



Based on the statistical data presented above, from comparing the 'NEW' column in cases_qld and 'NEWDEATHS' column in deaths_qld, it is evident that the average number of new COVID-19 cases per week in Queensland is 10,052.6, while the average number of deaths is 16.4. Consequently, it can be deduced that out of the 10,053 individuals who contracted COVID-19, approximately 16 individuals succumbed to the disease. This suggests that approximately 10,037 individuals have successfully recovered from the illness in Queensland.

From the plot of the kernel density estimate, it becomes apparent that the blur curve representing new cases exhibits a broader dispersion around the mean, indicating a higher standard deviation of 20,147.4. Conversely, the red curve representing new deaths appears more concentrated, signifying a lower standard deviation for this variable, which is 673.6.

Analysing Distribution of data for SA.

In [89]:

```
# Reading the data

file_name = 'daily_cases_sa.tsv'

file_path = directory_path + file_name

cases_sa = pd.read_csv(file_path, delimiter='\t')

#deaths tab
file_name2 = 'daily_death_sa.tsv'
```

```
file_path2 = directory_path + file_name2

deaths_sa = pd.read_csv(file_path2, delimiter='\t')
```

In [90]:
```
# SA CASES Analysis
## Cleaning the columns NEW, CASES and NET for the analysis


cases_sa['NEW'].fillna('0', inplace=True)
cases_sa[['NEW', 'CASES', 'NET']] = cases_sa[['NEW', 'CASES', 'NET']].astype(str)
cases_sa[['NEW', 'CASES', 'NET']] = cases_sa[['NEW', 'CASES', 'NET']].replace(',','',
cases_sa[['NEW', 'CASES', 'NET']] = cases_sa[['NEW', 'CASES', 'NET']].replace('-','0')
cases_sa[['NEW', 'CASES', 'NET']] = cases_sa[['NEW', 'CASES', 'NET']].astype(int)
```

In [91]:
```
## Time indexing on the Dataframe

cases_sa['DATE'] = pd.to_datetime(cases_sa['DATE'])
cases_sa.set_index('DATE', inplace=True)
```

In [92]:
```
## Converting the time-indexed data into weekly intervals, while calculating sum of th

cases_sa_weekly = cases_sa.resample('W').sum()
cases_sa_weekly["TOTALCASES"] = cases_sa_weekly["NEW"].cumsum()
cases_sa_weekly["STATE"] = 'SA'
```

In [93]:
```
# DEATHS SA Analysis
## Cleaning the columns DEATHS and NET for the analysis

deaths_sa[['DEATHS', 'NET']] = deaths_sa[['DEATHS', 'NET']].astype(str)
deaths_sa[['DEATHS', 'NET']] = deaths_sa[['DEATHS', 'NET']].replace(',', '', regex=Tru
deaths_sa[['DEATHS', 'NET']] = deaths_sa[['DEATHS', 'NET']].replace('-', '0')
deaths_sa[['DEATHS', 'NET']] = deaths_sa[['DEATHS', 'NET']].astype(int)
```

In [94]:
```
## Time indexing

deaths_sa['DATE'] = pd.to_datetime(deaths_sa['DATE'])
deaths_sa.set_index('DATE', inplace=True)
```

In [95]:
```
## Converting the time-indexed data into weekly intervals, while calculating sum of th

deaths_sa_weekly = deaths_sa.resample('W').sum()
deaths_sa_weekly['NEWDEATHS'] = deaths_sa_weekly['DEATHS'].diff().fillna(0)
```

In [96]:
```
# Merging the two data sets using DATE

mergedsa = pd.merge(cases_sa_weekly, deaths_sa_weekly, on='DATE')
print(mergedsa.describe())
```

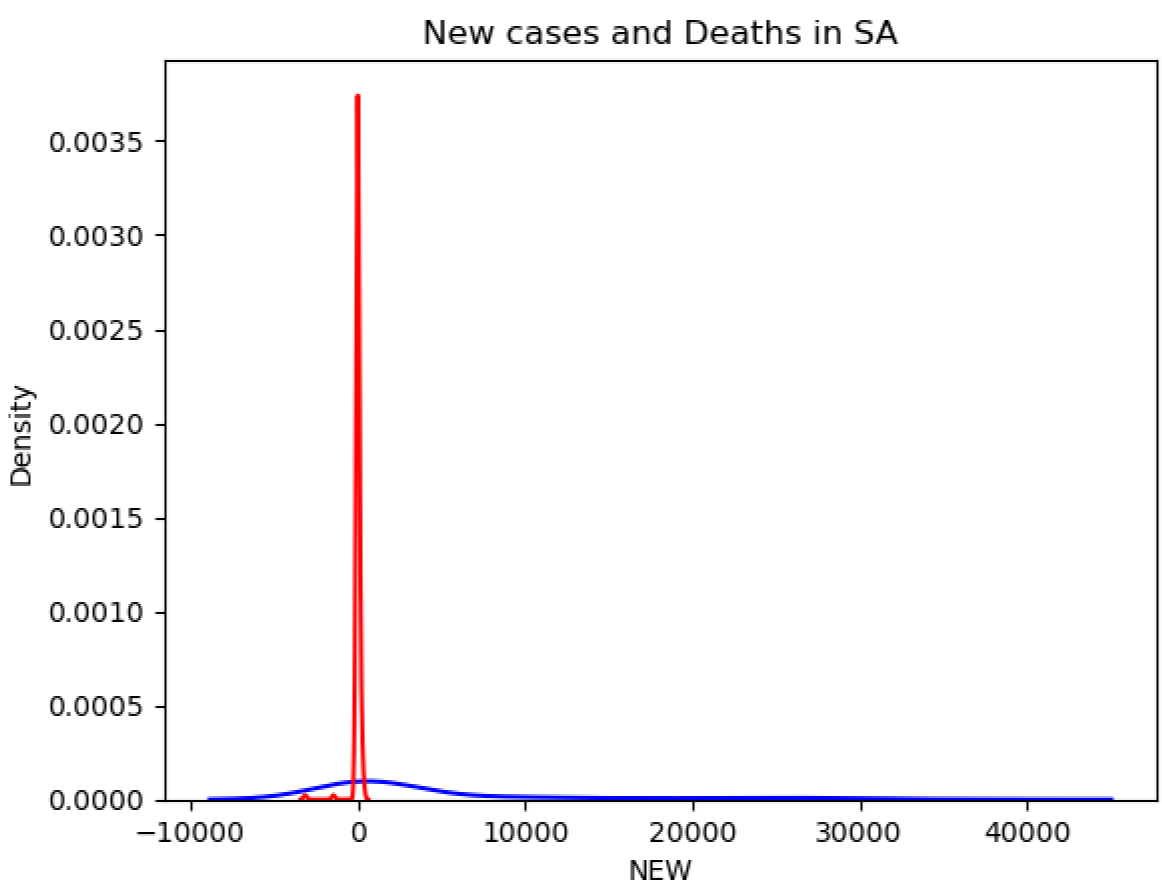
```
                NEW         CASES  VAR_x         NET_x     TOTALCASES  \
count    187.000000  1.870000e+02  187.0    187.000000     187.000000
mean    4699.973262  7.791273e+05    0.0   4972.048128  291214.716578
std     8383.401724  1.298155e+06    0.0   9022.541604  363491.554786
min        0.000000  0.000000e+00    0.0  -3648.000000       0.000000
25%        4.000000  3.939000e+03    0.0      4.000000     566.000000
50%       30.000000  6.426000e+03    0.0     23.000000     936.000000
75%     3762.500000  9.019075e+05    0.0   4355.500000  722212.000000
max    36203.000000  5.278832e+06    0.0  41275.000000  878895.000000

            DEATHS  VAR_y       NET_y    NEWDEATHS
count   187.000000  187.0  187.000000   187.000000
mean    872.609626    0.0    8.411765     8.411765
std    1252.856840    0.0   15.351665   263.607737
min       0.000000    0.0  -11.000000 -3127.000000
25%      28.000000    0.0    0.000000     0.000000
50%      28.000000    0.0    0.000000     0.000000
75%    1370.000000    0.0   13.500000    21.500000
max    5472.000000    0.0   89.000000   369.000000
```

In [97]:

```
# Plotting the kernel density estimate

KDE_graph = sns.kdeplot(data = mergedsa, x = "NEW", color="blue", label="New Cases")
KDE_graph = sns.kdeplot(data = mergedsa, x = "NEWDEATHS", color="red", label="New Deat
KDE_graph.set_title("New cases and Deaths in SA")
```

Out[97]:

```
Text(0.5, 1.0, 'New cases and Deaths in SA')
```



Based on the statistical data presented above, from comparing the 'NEW' column in cases_sa and 'NEWDEATHS' column in deaths_sa, it is evident that the average number of new COVID-19 cases per week in South Australia is 4,699.9, while the average number of deaths is 8.4. Consequently, it can be deduced that out of the 4,700 individuals who contracted COVID-19, approximately 8 individuals succumbed to the disease. This suggests that approximately 4,692 individuals have successfully recovered from the illness in South Australia.

From the plot of the kernel density estimate, it becomes apparent that the blur curve representing new cases exhibits a broader dispersion around the mean, indicating a higher standard deviation of 8,383.4. Conversely, the red curve representing new deaths appears more concentrated, signifying a lower standard deviation for this variable, which is 263.6.

Analysing Distribution of data for VIC.

In [98]:
```
# Reading the data

file_name = 'daily_cases_vic.tsv'

file_path = directory_path + file_name

cases_vic = pd.read_csv(file_path, delimiter='\t')

#deaths tab
file_name2 = 'daily_death_vic.tsv'

file_path2 = directory_path + file_name2

deaths_vic = pd.read_csv(file_path2, delimiter='\t')
```

In [99]:
```
# VIC CASES Analysis
## Cleaning the columns NEW, CASES and NET for the analysis


cases_vic['NEW'].fillna('0', inplace=True)
cases_vic[['NEW', 'CASES', 'NET']] = cases_vic[['NEW', 'CASES', 'NET']].astype(str)
cases_vic[['NEW', 'CASES', 'NET']] = cases_vic[['NEW', 'CASES', 'NET']].replace(',',''
cases_vic[['NEW', 'CASES', 'NET']] = cases_vic[['NEW', 'CASES', 'NET']].replace('-','0
cases_vic[['NEW', 'CASES', 'NET']] = cases_vic[['NEW', 'CASES', 'NET']].astype(int)
```

In [100…
```
## Time indexing

cases_vic['DATE'] = pd.to_datetime(cases_vic['DATE'])
cases_vic.set_index('DATE', inplace=True)
```

In [101…
```
## Converting the time-indexed data into weekly intervals, while calculating sum of th

cases_vic_weekly = cases_vic.resample('W').sum()
cases_vic_weekly["TOTALCASES"] = cases_vic_weekly["NEW"].cumsum()
cases_vic_weekly["STATE"] = 'VIC'
```

In [102…
```
# DEATHS VIC Analysis
## Cleaning the columns DEATHS and NET for the analysis

deaths_vic[['DEATHS', 'NET']] = deaths_vic[['DEATHS', 'NET']].astype(str)
deaths_vic[['DEATHS', 'NET']] = deaths_vic[['DEATHS', 'NET']].replace(',', '', regex=T
```

```
deaths_vic[['DEATHS', 'NET']] = deaths_vic[['DEATHS', 'NET']].replace('-', '0')
deaths_vic[['DEATHS', 'NET']] = deaths_vic[['DEATHS', 'NET']].astype(int)
```

In [103…

```
## Time indexing

deaths_vic['DATE'] = pd.to_datetime(deaths_vic['DATE'])
deaths_vic.set_index('DATE', inplace=True)
```

In [104…

```
## Converting the time-indexed data into weekly intervals, while calculating sum of th

deaths_vic_weekly = deaths_vic.resample('W').sum()
deaths_vic_weekly['NEWDEATHS'] = deaths_vic_weekly['DEATHS'].diff().fillna(0)
```

In [105…

```
# Merging the two data sets on DATE

mergedvic = pd.merge(cases_vic_weekly, deaths_vic_weekly, on='DATE')
print(mergedvic.describe())
```

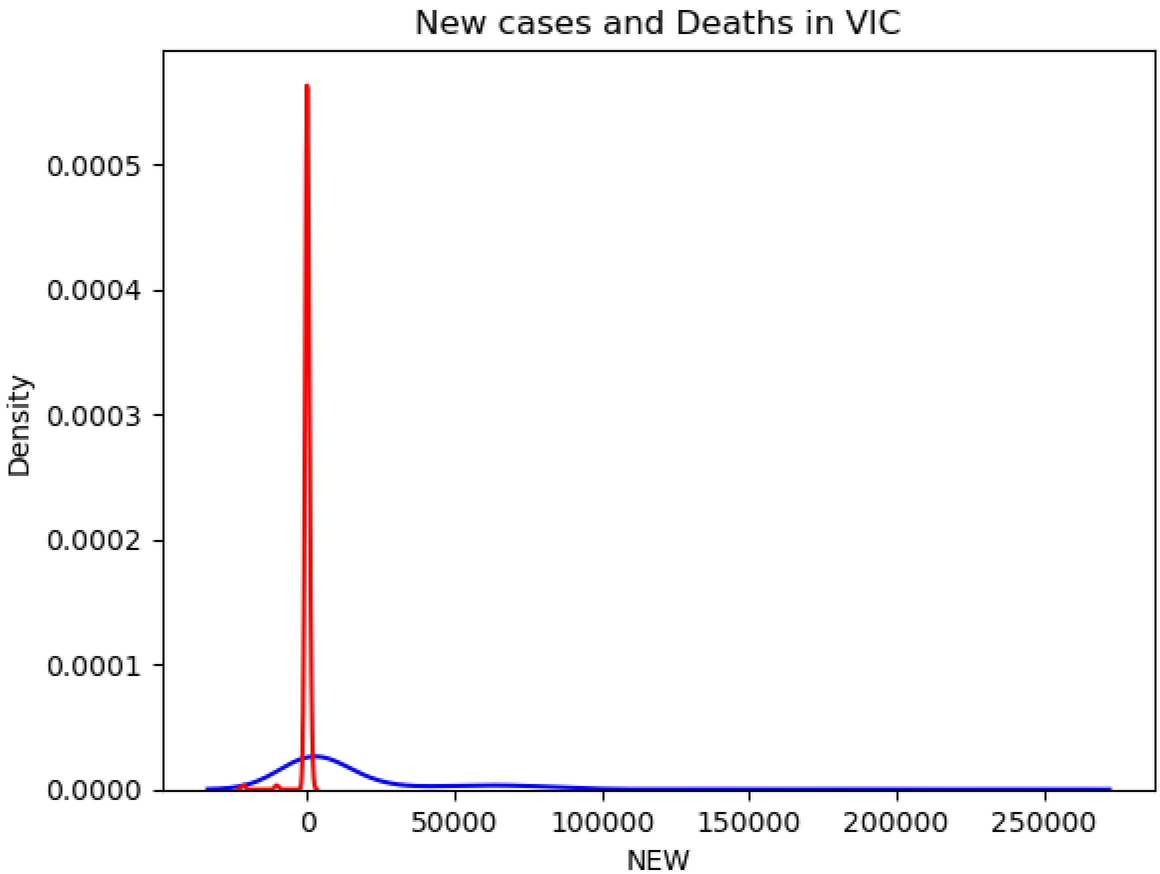
```
                 NEW         CASES  VAR_x          NET_x    TOTALCASES  \
count     187.000000  1.870000e+02  187.0     187.000000  1.870000e+02
mean    16385.032086  3.022377e+06    0.0   15935.561497  1.112760e+06
std     31856.706268  4.638758e+06    0.0   30932.149619  1.292981e+06
min         0.000000  2.000000e+00    0.0      -1.000000  0.000000e+00
25%        40.500000  1.424465e+05    0.0      35.500000  2.181500e+04
50%      3016.000000  6.545720e+05    0.0    2998.000000  1.001340e+05
75%     12653.000000  2.955955e+06    0.0   12032.000000  2.688801e+06
max    238588.000000  1.796496e+07    0.0  234348.000000  3.064001e+06

             DEATHS  VAR_y       NET_y     NEWDEATHS
count    187.000000  187.0  187.000000    187.000000
mean    8650.877005    0.0   44.636364     44.636364
std     8032.237838    0.0   54.787236   1789.655267
min        0.000000    0.0    0.000000 -21507.000000
25%     5733.000000    0.0    0.000000      0.000000
50%     5833.000000    0.0   31.000000     42.000000
75%     8290.500000    0.0   62.000000    303.000000
max    37229.000000    0.0  322.000000   1596.000000
```

In [106…

```
# Plotting the KDE estimate

KDE_graph = sns.kdeplot(data = mergedvic, x = "NEW", color="blue", label="New Cases")
KDE_graph = sns.kdeplot(data = mergedvic, x = "NEWDEATHS",  color="red", label="New De
KDE_graph.set_title("New cases and Deaths in VIC")
```

Out[106]:

```
Text(0.5, 1.0, 'New cases and Deaths in VIC')
```



Based on the statistical data presented above, from comparing the 'NEW' column in cases_vic and 'NEWDEATHS' column in deaths_vic, it is evident that the average number of new COVID-19 cases per week in Victoria is 16,385.0, while the average number of deaths is 44.6. Consequently, it can be deduced that out of the 16,385 individuals who contracted COVID-19, approximately 45 individuals succumbed to the disease. This suggests that approximately 16,340 individuals have successfully recovered from the illness in Victoria.

From the plot of the kernel density estimate, it becomes apparent that the blur curve representing new cases exhibits a broader dispersion around the mean, indicating a higher standard deviation of 31,856.7. Conversely, the red curve representing new deaths appears more concentrated, signifying a lower standard deviation for this variable, which is 1,789.6.

Analysing Distribution of data for WA.

In [107…

```
# Reading the data

file_name = 'daily_cases_wa.tsv'

file_path = directory_path + file_name

cases_wa = pd.read_csv(file_path, delimiter='\t')

#deaths tab
file_name2 = 'daily_death_wa.tsv'
```

```
file_path2 = directory_path + file_name2

deaths_wa = pd.read_csv(file_path2, delimiter='\t')
```

In [108…

```
# WA CASES Analysis
## Cleaning the columns NEW, CASES and NET for the analysis


cases_wa['NEW'].fillna('0', inplace=True)
cases_wa[['NEW', 'CASES', 'NET']] = cases_wa[['NEW', 'CASES', 'NET']].astype(str)
cases_wa[['NEW', 'CASES', 'NET']] = cases_wa[['NEW', 'CASES', 'NET']].replace(',','',
cases_wa[['NEW', 'CASES', 'NET']] = cases_wa[['NEW', 'CASES', 'NET']].replace('-','0')
cases_wa[['NEW', 'CASES', 'NET']] = cases_wa[['NEW', 'CASES', 'NET']].astype(int)
```

In [109…

```
## Time indexing

cases_wa['DATE'] = pd.to_datetime(cases_wa['DATE'])
cases_wa.set_index('DATE', inplace=True)
```

In [110…

```
## Converting the time-indexed data into weekly intervals, while calculating sum of th

cases_wa_weekly = cases_wa.resample('W').sum()
cases_wa_weekly["TOTALCASES"] = cases_wa_weekly["NEW"].cumsum()
cases_wa_weekly["STATE"] = 'WA'
```

In [111…

```
# DEATHS WA Analysis
## Cleaning the columns DEATHS and NET for the analysis

deaths_wa[['DEATHS', 'NET']] = deaths_wa[['DEATHS', 'NET']].astype(str)
deaths_wa[['DEATHS', 'NET']] = deaths_wa[['DEATHS', 'NET']].replace(',', '', regex=Tru
deaths_wa[['DEATHS', 'NET']] = deaths_wa[['DEATHS', 'NET']].replace('-', '0')
deaths_wa[['DEATHS', 'NET']] = deaths_wa[['DEATHS', 'NET']].astype(int)
```

In [112…

```
## Time indexing

deaths_wa['DATE'] = pd.to_datetime(deaths_wa['DATE'])
deaths_wa.set_index('DATE', inplace=True)
```

In [113…

```
## Converting the time-indexed data into weekly intervals, while calculating sum of th

deaths_wa_weekly = deaths_wa.resample('W').sum()
deaths_wa_weekly['NEWDEATHS'] = deaths_wa_weekly['DEATHS'].diff().fillna(0)
```

In [114…

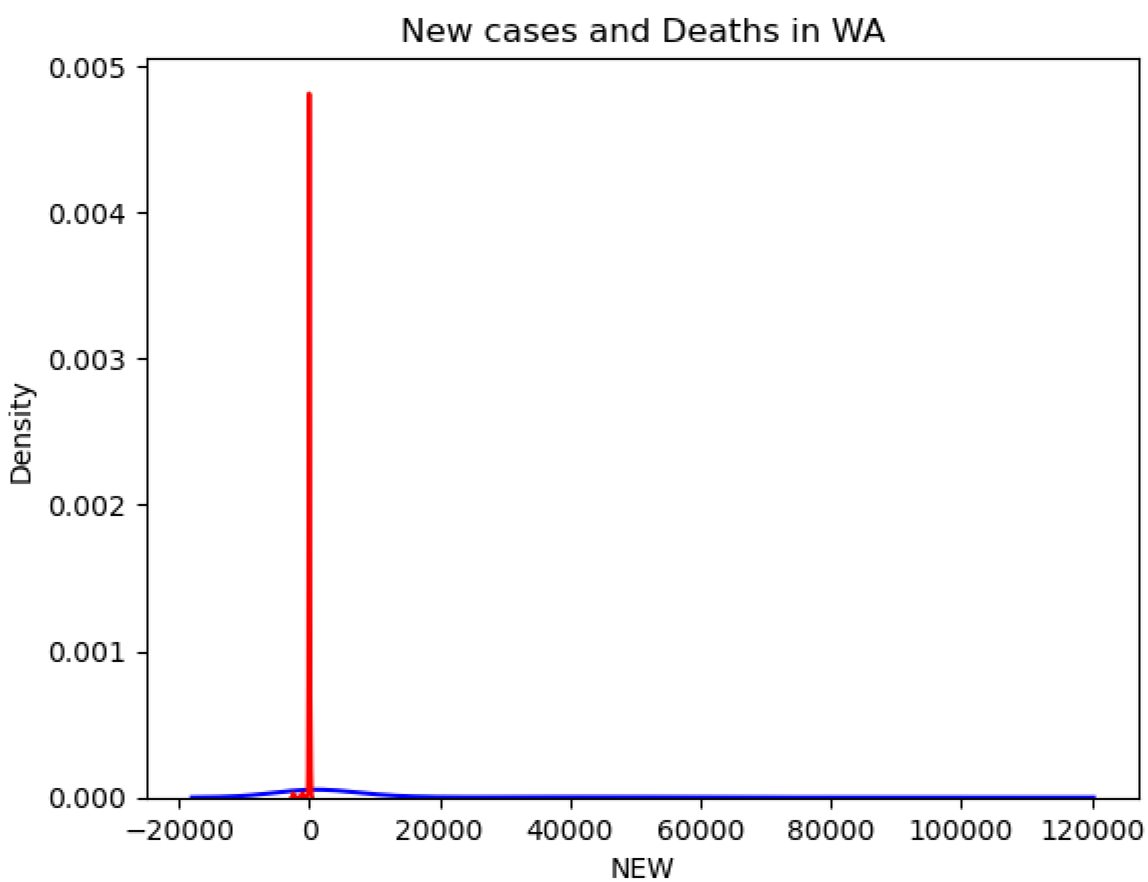
```
# Merging the two data sets on DATE

mergedwa = pd.merge(cases_wa_weekly, deaths_wa_weekly, on='DATE')
print(mergedwa.describe())
```

```
                 NEW         CASES  VAR_x         NET_x    TOTALCASES  \
count     187.000000  1.870000e+02  187.0    187.000000  1.870000e+02
mean     7343.352941  1.024682e+06    0.0   7230.566845  4.446074e+05
std     17059.969704  1.892604e+06    0.0  16809.579413  5.844587e+05
min         0.000000  0.000000e+00    0.0    -27.000000  0.000000e+00
25%         4.000000  5.862000e+03    0.0      4.000000  8.465000e+02
50%        23.000000  7.784000e+03    0.0     23.000000  1.120000e+03
75%      4529.500000  1.294729e+06    0.0   4505.000000  1.172699e+06
max    102305.000000  7.914245e+06    0.0 101731.000000  1.373207e+06

            DEATHS  VAR_y       NET_y    NEWDEATHS
count   187.000000  187.0  187.000000   187.000000
mean    577.106952    0.0    6.582888     6.582888
std     882.404120    0.0   10.509459   206.242960
min       0.000000    0.0    0.000000 -2448.000000
25%      63.000000    0.0    0.000000     0.000000
50%      63.000000    0.0    0.000000     0.000000
75%     940.000000    0.0   11.000000    11.000000
max    4171.000000    0.0   60.000000   246.000000
```

In [115…

```python
# Plotting the KDE estimate

KDE_graph = sns.kdeplot(data = mergedwa, x = "NEW", color="blue", label="New Cases")
KDE_graph = sns.kdeplot(data = mergedwa, x = "NEWDEATHS", color="red", label="New Deat
KDE_graph.set_title("New cases and Deaths in WA")
```

Out[115]: Text(0.5, 1.0, 'New cases and Deaths in WA')



Based on the statistical data presented above, from comparing the 'NEW' column in cases_wa and 'NEWDEATHS' column in deaths_wa, it is evident that the average number of new COVID-19 cases per week in West Australia is 7,343.3, while the average number of deaths is 6.5. Consequently, it can be deduced that out of the 7,343 individuals who contracted COVID-19, approximately 7 individuals succumbed to the disease. This suggests that approximately 7,336 individuals have successfully recovered from the illness in West Australia.

From the plot of the kernel density estimate, it becomes apparent that the blur curve representing new cases exhibits a broader dispersion around the mean, indicating a higher standard deviation of 17059.9. Conversely, the red curve representing new deaths appears more concentrated, signifying a lower standard deviation for this variable, which is 206.2.

Cumulative CASES graph for all the states

Plotting the cumulative history of COVID-19 in different states starting on the week after 1000 cases were reported

In [116…
```
cases_nsw_weekly_after1000 = cases_nsw_weekly[cases_nsw_weekly.TOTALCASES >= 1000] # j
cases_nsw_weekly_after1000 = cases_nsw_weekly_after1000.reset_index(drop=True) #create
cases_nsw_weekly_after1000 = cases_nsw_weekly_after1000.rename_axis('week') #rename to
```

In [117…
```
cases_qld_weekly_after1000 = cases_qld_weekly[cases_qld_weekly.TOTALCASES >= 1000]
cases_qld_weekly_after1000 = cases_qld_weekly_after1000.reset_index(drop=True)
cases_qld_weekly_after1000 = cases_qld_weekly_after1000.rename_axis('week')
```

In [118…
```
cases_sa_weekly_after1000 = cases_sa_weekly[cases_sa_weekly.TOTALCASES >= 1000]
cases_sa_weekly_after1000 = cases_sa_weekly_after1000.reset_index(drop=True)
cases_sa_weekly_after1000 = cases_sa_weekly_after1000.rename_axis('week')
```

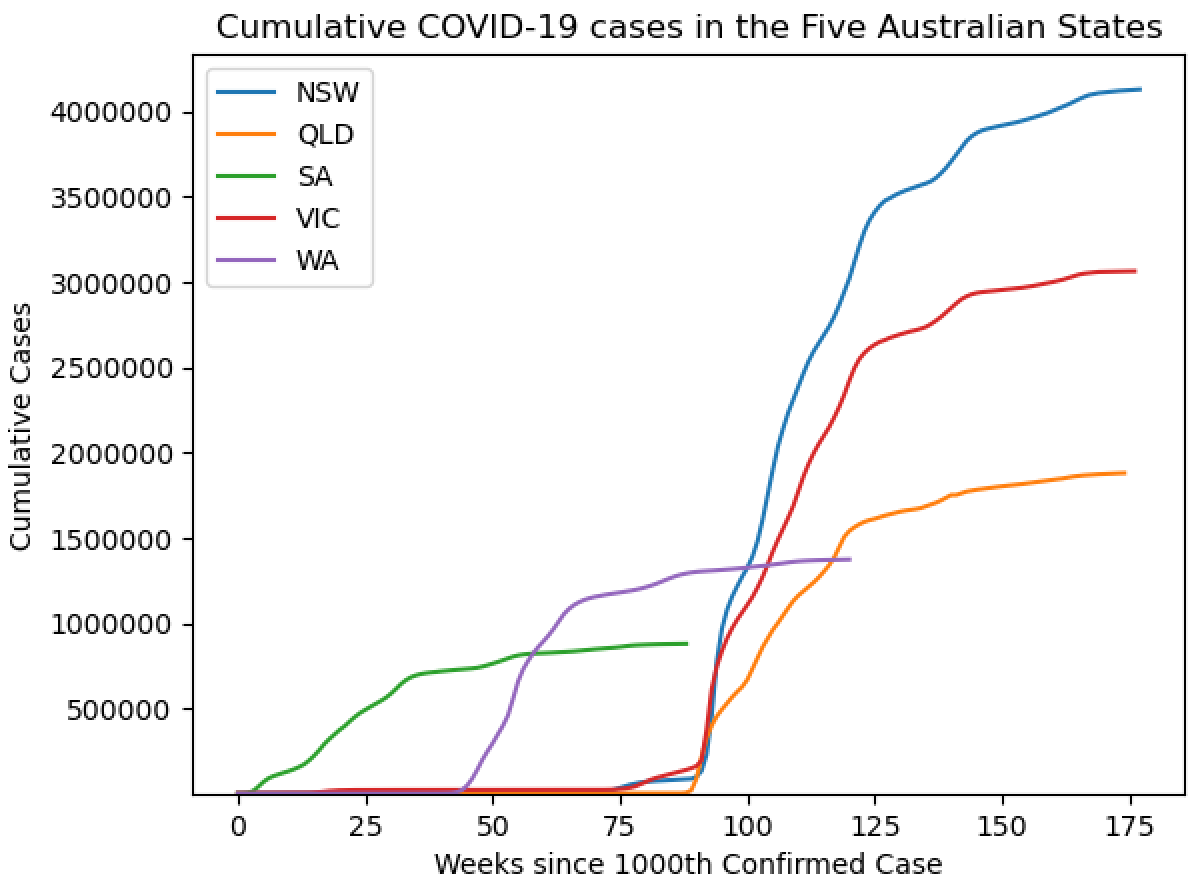
In [119…
```
cases_vic_weekly_after1000 = cases_vic_weekly[cases_vic_weekly.TOTALCASES >= 1000]
cases_vic_weekly_after1000 = cases_vic_weekly_after1000.reset_index(drop=True)
cases_vic_weekly_after1000 = cases_vic_weekly_after1000.rename_axis('week')
```

In [120…
```
cases_wa_weekly_after1000 = cases_wa_weekly[cases_wa_weekly.TOTALCASES >= 1000]
cases_wa_weekly_after1000 = cases_wa_weekly_after1000.reset_index(drop=True)
cases_wa_weekly_after1000 = cases_wa_weekly_after1000.rename_axis('week')
```

In [121…
```
import matplotlib.ticker as mtick
```

In [122…
```
plt.plot (cases_nsw_weekly_after1000.index, cases_nsw_weekly_after1000['TOTALCASES'],
plt.plot (cases_qld_weekly_after1000.index, cases_qld_weekly_after1000['TOTALCASES'],
plt.plot (cases_sa_weekly_after1000.index, cases_sa_weekly_after1000['TOTALCASES'], la
plt.plot (cases_vic_weekly_after1000.index, cases_vic_weekly_after1000['TOTALCASES'],
plt.plot (cases_wa_weekly_after1000.index, cases_wa_weekly_after1000['TOTALCASES'], la


plt.title('Cumulative COVID-19 cases in the Five Australian States')
plt.xlabel('Weeks since 1000th Confirmed Case')
plt.ylabel('Cumulative Cases')


plt.gca().set_ylim(bottom=1000)

plt.gca().yaxis.set_major_formatter(mtick.FormatStrFormatter('%d'))
```

```
plt.legend()

plt.show()
```



The analysis of cumulative COVID-19 cases in five Australian states reveals that, with the exception of South Australia (SA) and Western Australia (WA), the states exhibit similar trends.

In SA, the number of new COVID-19 cases increased significantly at a certain time after the 25th week, nearly doubling the number of cases before the 25th week. In WA, this significant increase occurred after the 50th week where it peaked and remained relatively stable until the 75th week where there was a slight increase.

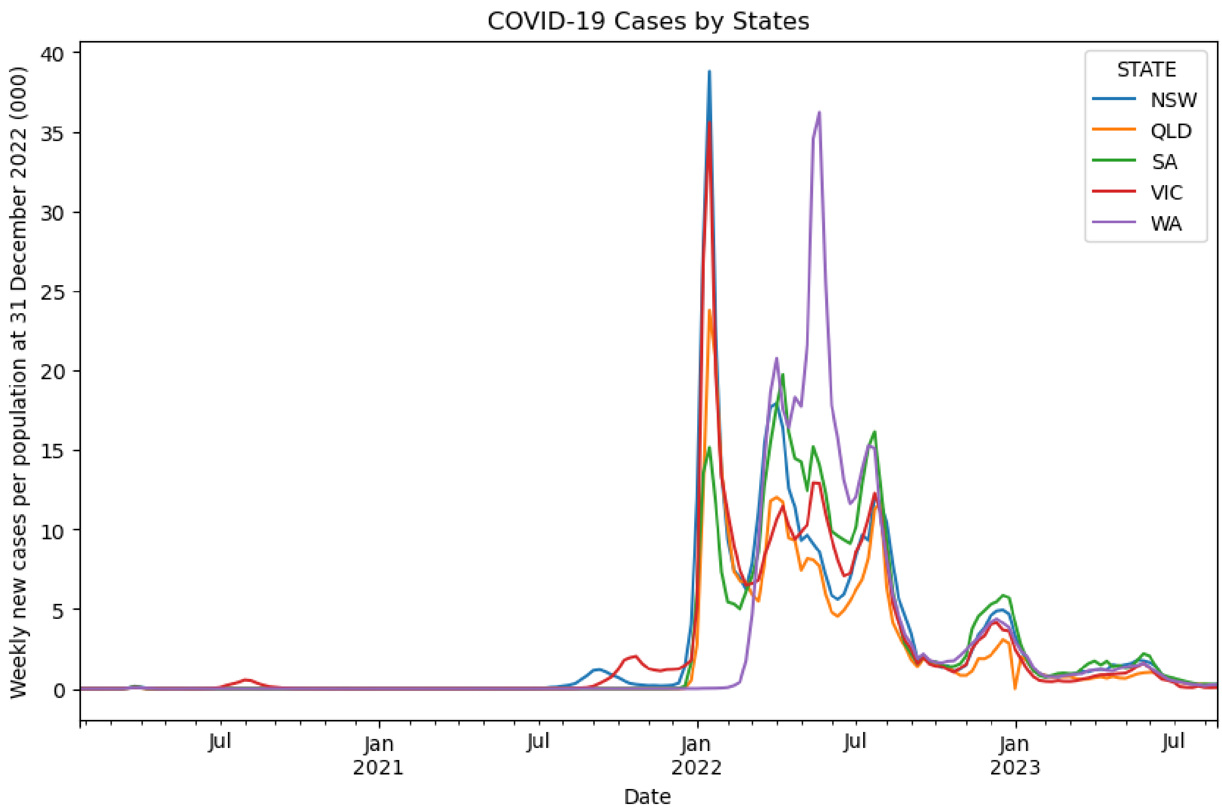
Furthermore, COVID-19 cases remained relatively stable until the 75th week in New South Wales (NSW), Queensland (QLD), and Victoria (VIC). However, after this period, there was a significant surge in cases for all 3 states. In the case of NSW, for instance, the number of cases escalated from around 150,000 between the 75th and 100th week to well over 4,000,000 in subsequent weeks. Similar trends were observed in VIC, where cases gradually increased to a little under 3,000,000. Consequently, in QLD, the cases gradually increased to a little under 2,000,000.

Normalization of cases by the states population.

In [123…

```
#Normalization of cases by the states population.

population_by_state = {
```

```
    'New South Wales': 8238.8,
    'Victoria': 6704.3,
    'Queensland': 5378.3,
    'South Australia': 1834.3,
    'Western Australia': 2825.2
}

cases_nsw_weekly['cases_per_population'] = cases_nsw_weekly['NEW'] / (population_by_st
cases_vic_weekly['cases_per_population'] = cases_vic_weekly['NEW'] / (population_by_st
cases_qld_weekly['cases_per_population'] = cases_qld_weekly['NEW'] / (population_by_st
cases_sa_weekly['cases_per_population'] = cases_sa_weekly['NEW'] / (population_by_stat
cases_wa_weekly['cases_per_population'] = cases_wa_weekly['NEW'] / (population_by_stat
# This normalizes data allowing for easier comparison

#
allstatescases = pd.concat([cases_nsw_weekly, cases_vic_weekly, cases_qld_weekly, case
allstatescasespivot = allstatescases.pivot(columns='STATE', values='cases_per_populati

# This concatenates all data into a single dataframe and uses the column STATE as the

# Creating a line plot, allowing for a comparative analysis of cases per 100,000 popul

fig, ax = plt.subplots(figsize=(10, 6))
allstatescasespivot.plot(ax=ax)
ax.set_xlabel('Date')
ax.set_ylabel('Weekly new cases per population at 31 December 2022 (000)')
ax.set_title('COVID-19 Cases by States')
plt.show()
```



Upon normalizing COVID-19 data by the population of each state, the resulting insights are indeed noteworthy. Similarly, the calendar-based graphical representation makes it easier to understand these findings.

From observing the calendar-based graph, it becomes apparent that all states exhibit a similar trend, except for Western Australia (WA). While the remaining states have considerably higher COVID-19 cases per 100,000 population, WA initially reported very low numbers. However, starting from February-March 2022, WA witnessed a gradual increase, reaching past 2,000 cases per 100,000 people, a notably high figure for that month in comparison to other states. Subsequently, in May-June of the same year, it recorded the highest weekly cases, well over 3,500 per week. Following this peak, WA experienced a gradual decline, aligning its trajectory with that of other states.

In January 2022, all states except WA registered their highest weekly case counts per 100,000 population. New South Wales (NSW) reported the highest incidence, followed by Victoria (VIC) and then Queensland (QLD). Conversely, South Australia (SA) reported the lowest figures. After March 2022, these states witnessed fluctuations in their case numbers, followed by a gradual decline commencing in January 2023.

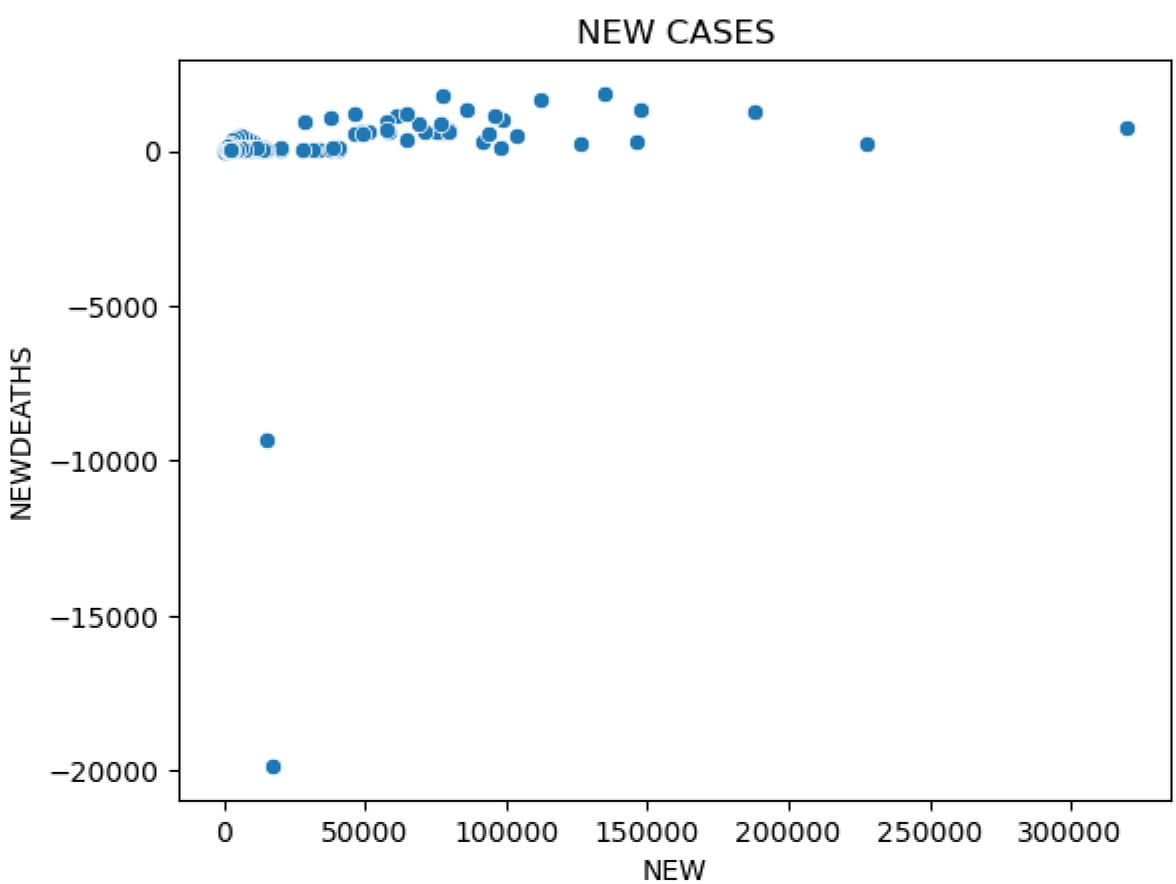
Studying the Relationship between Number of New Cases and Deaths in the five states

In [124…

```
#Relation between Cases and Deaths in NSW.
sns.scatterplot(data = mergednsw, x = "NEW", y = "NEWDEATHS").set(title='NEW CASES')
```

Out[124]:

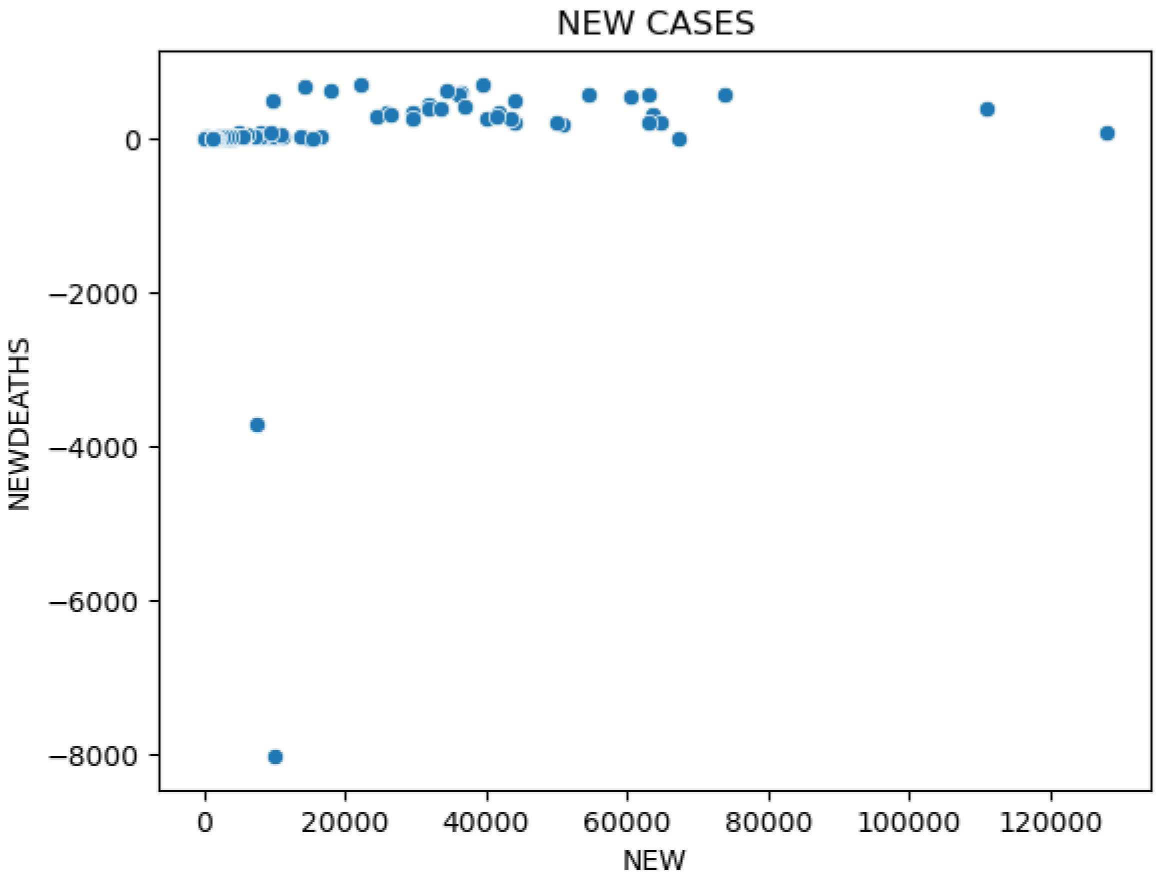
```
[Text(0.5, 1.0, 'NEW CASES')]
```



In [125…

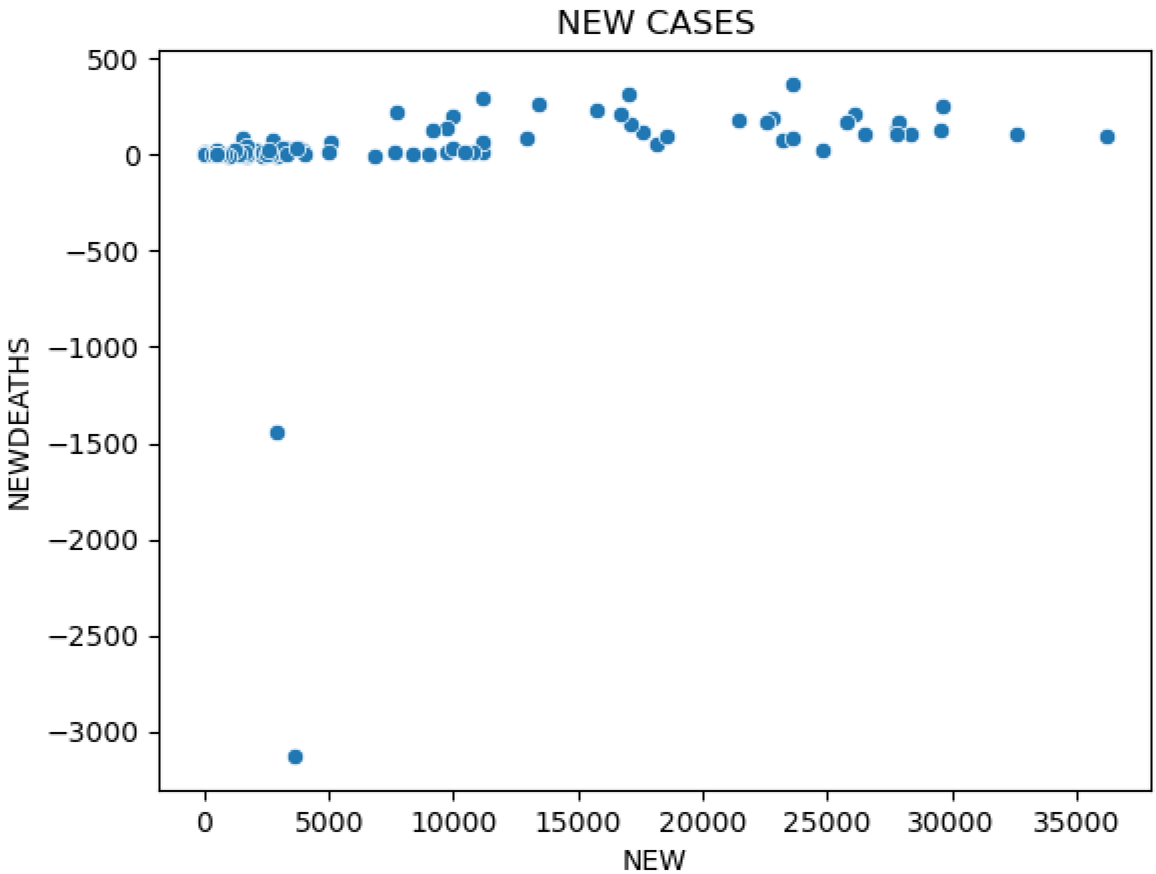
```
sns.scatterplot(data = mergedqld, x = "NEW", y = "NEWDEATHS").set(title='NEW CASES')
```

Out[125]: [Text(0.5, 1.0, 'NEW CASES')]



In [126…
```
sns.scatterplot(data = mergedsa, x = "NEW", y = "NEWDEATHS").set(title='NEW CASES')
```

Out[126]: [Text(0.5, 1.0, 'NEW CASES')]

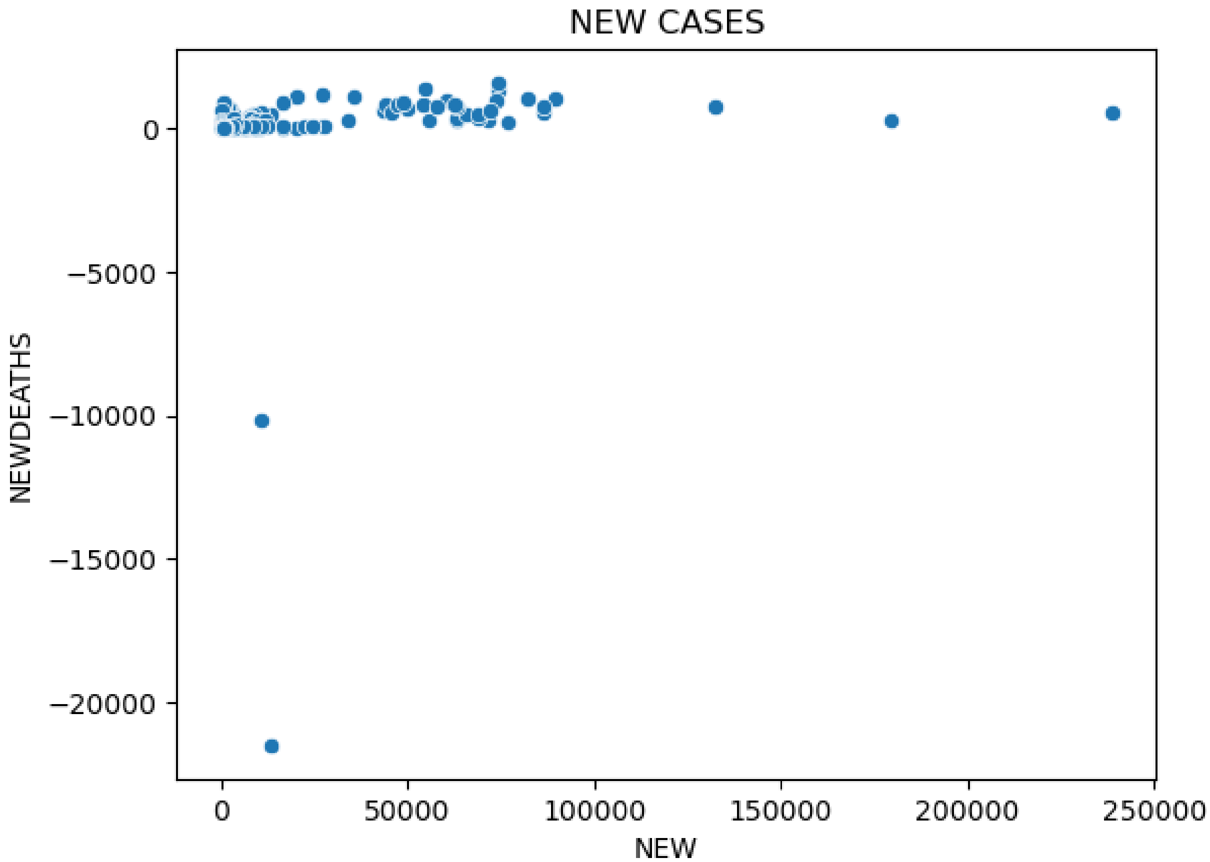


In [127…

```
sns.scatterplot(data = mergedvic, x = "NEW", y = "NEWDEATHS").set(title='NEW CASES')
```

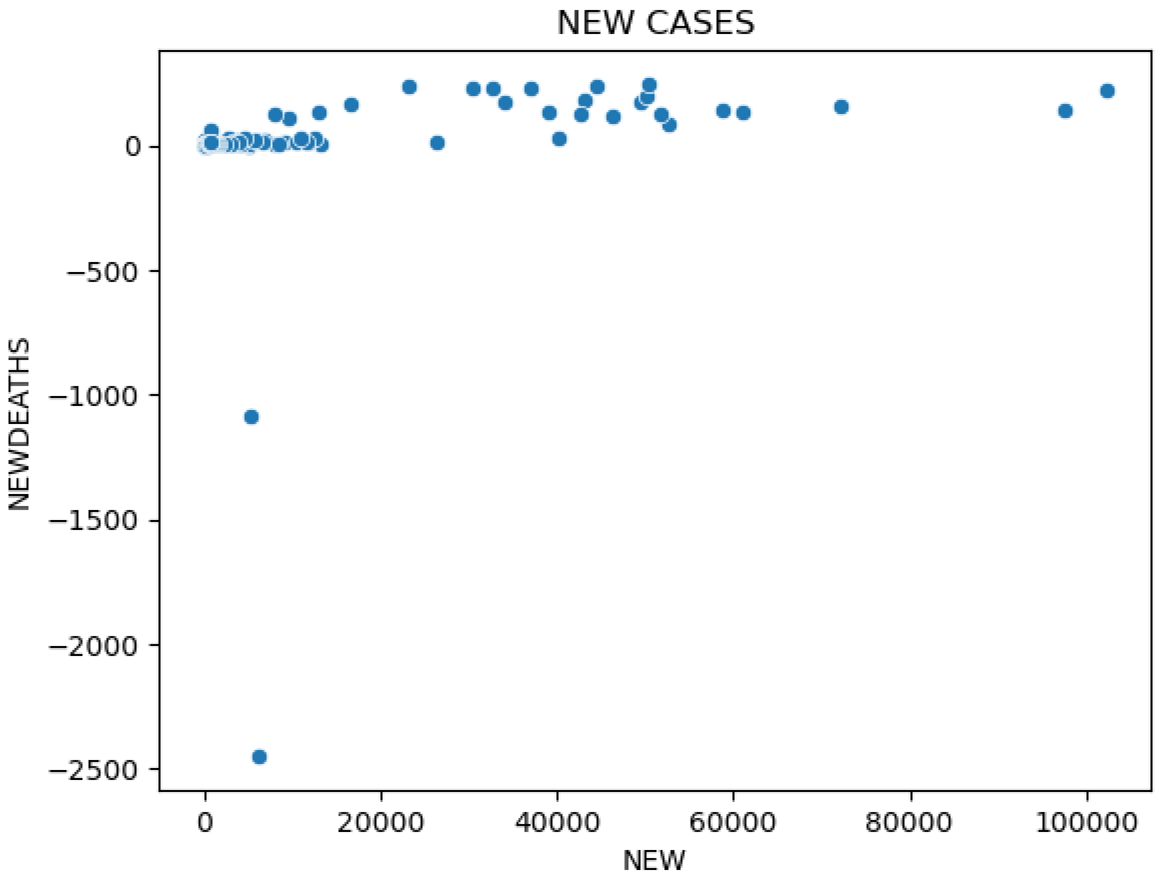
Out[127]:

```
[Text(0.5, 1.0, 'NEW CASES')]
```



In [128…
```
sns.scatterplot(data = mergedwa, x = "NEW", y = "NEWDEATHS").set(title='NEW CASES')
```

Out[128]:
```
[Text(0.5, 1.0, 'NEW CASES')]
```



For all the scatterplots above, the plots are concentrated around NEWDEATHS 0 and NEW from 0 to 100,000. This implies that, for most data points, when the number of new cases is low (0 to 100,000), the number of new deaths tends to be close to zero. However, this doesn't necessarily mean that as new cases increase, new deaths increase proportionally. Therefore this may suggest that there is a lack of a strong linear correlation between the number of new COVID-19 cases and the number of new deaths in the given datasets, the correlation is likely only slightly positive.

It is also important to note that there are outliers for all the scatterplots where NEWDEATHS does not center around 0.

Conclusion

According the summary statistics pertaining to COVID-19 cases, deaths and recovery rates across five Australian states, the highest average case counts were observed in New South Wales (NSW), while the highest average death counts were recorded in Victoria (VIC). This observation implies that the recovery rates in NSW are comparatively more higher than those in VIC. Additionally, the lowest average case numbers were documented in South Australia (SA), and the lowest average death figures were reported in Western Australia (WA).

Over the past three years, it is evident that NSW and VIC have experienced a higher incidence of COVID-19 infections in comparison to South Australia and Western Australia. This disparity can be attributed, in part, to the notably higher populations in the former two states relative to the latter two.

Moreover, the first 3 months of 2022 (Jan-Mar 2022) witnessed a notable increase in COVID-19 infections across all states, except for Western Australia, which experienced a surge five months later, in May-June 2022.

Additionally, the relationship between new cases and deaths is unclear. The correlation between the number of new COVID-19 cases and the number of new deaths is likely only slightly positive. However, there are occasional outliers in the datasets.