Importing Modules

In [1]:
```python
import numpy as np
import pandas as pd
import datetime as dt
import plotly.express as px
from scipy.signal import find_peaks
```

Reading Super Covid Data

In [2]:
```python
# Reading super covid dataset
super_covid = pd.read_csv('../../Team/STAGE1/superCovidDS.CSV')
super_covid.head()
```

Out[2]:

	countyFIPS	County Name	State	StateFIPS	2020-01-22_x	2020-01-23_x	2020-01-24_x	2020-01-25_x	2020-01-26_x	2020-01-27_x	...	2023-01-08_y	2023-01-09_y	2023-01-10_y	2023-01-11_y	2023-01-12_y	2023-01-13_y	2023-01-14_y	2023-01-15_y	202
0	1001	Autauga County	AL	1	0	0	0	0	0	0	...	230	230	230	230	230	230	230	230	2
1	1003	Baldwin County	AL	1	0	0	0	0	0	0	...	719	719	719	719	721	721	721	721	7
2	1005	Barbour County	AL	1	0	0	0	0	0	0	...	103	103	103	103	103	103	103	103	1
3	1007	Bibb County	AL	1	0	0	0	0	0	0	...	108	108	108	108	108	108	108	108	1
4	1009	Blount County	AL	1	0	0	0	0	0	0	...	260	260	260	260	261	261	261	261	2

5 rows × 2187 columns

In [3]:
```python
#Renaming County Name column to use it for ease of access
super_covid_column_names = list(super_covid.columns)
super_covid_column_names[super_covid_column_names.index('County Name')] = "County_Name"
super_covid.columns=super_covid_column_names
```

Creating a Transformed Dataset

In [4]:
```python
transformed_df = pd.DataFrame(columns=['Date','Week','countyFIPS','County_Name', 'State', 'StateFIPS', 'population', 'Cases', 'N
transformed_df.head()
```

Out[4]:

Date	Week	countyFIPS	County_Name	State	StateFIPS	population	Cases	New_Cases	Deaths	New_Deaths

In [5]:
```python
# Process to transform data from June 2022 to December 2022
start_date = dt.datetime(2022,6,1)
end_date = dt.datetime(2022,12,31)
date_series = pd.date_range(start_date, end_date, freq='d')
date_delta = dt.timedelta(days=1)
for date in date_series:
    data = []
    for _ , row in super_covid.iterrows():
        temp = [date, date.isocalendar()[1], getattr(row, 'countyFIPS'), getattr(row, 'County_Name'),
                getattr(row, 'State'), getattr(row, 'StateFIPS'), getattr(row, 'population')]
        cases_column = date.strftime('%Y-%m-%d_x')
        temp.append(getattr(row, cases_column))
        temp.append(getattr(row, cases_column) - getattr(row, (date-date_delta).strftime('%Y-%m-%d_x')))
        deaths_column = date.strftime('%Y-%m-%d_y')
        temp.append(getattr(row, deaths_column))
        temp.append(getattr(row, deaths_column) - getattr(row, (date-date_delta).strftime('%Y-%m-%d_y')))
        data.append(temp)
    transformed_df = pd.concat([transformed_df, pd.DataFrame(data, columns=transformed_df.columns)])
transformed_df.head()
```

Out[5]:

	Date	Week	countyFIPS	County_Name	State	StateFIPS	population	Cases	New_Cases	Deaths	New_Deaths
0	2022-06-01	22	1001	Autauga County	AL	1	55869	15969	6	216	0
1	2022-06-01	22	1003	Baldwin County	AL	1	223234	56580	68	683	0
2	2022-06-01	22	1005	Barbour County	AL	1	24686	5710	3	99	0
3	2022-06-01	22	1007	Bibb County	AL	1	22394	6508	8	105	0
4	2022-06-01	22	1009	Blount County	AL	1	57826	15077	4	244	0

In [6]:
```python
transformed_df.shape
```

Out[6]:
```
(672388, 11)
```

In [7]:

```
transformed_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 672388 entries, 0 to 3141
Data columns (total 11 columns):
 #   Column       Non-Null Count   Dtype         
---  ------       --------------   -----         
 0   Date         672388 non-null  datetime64[ns]
 1   Week         672388 non-null  object        
 2   countyFIPS   672388 non-null  object        
 3   County_Name  672388 non-null  object        
 4   State        672388 non-null  object        
 5   StateFIPS    672388 non-null  object        
 6   population   672388 non-null  object        
 7   Cases        672388 non-null  object        
 8   New_Cases    672388 non-null  object        
 9   Deaths       672388 non-null  object        
 10  New_Deaths   672388 non-null  object        
dtypes: datetime64[ns](1), object(10)
memory usage: 61.6+ MB
```

In [8]:

```
#Changing datatypes
transformed_df = transformed_df.astype({'population':int,'Cases':int,'New_Cases':int,'Deaths':int,'New_Deaths':int})
```

In [9]:

```
transformed_df[['population','Cases','New_Cases','Deaths','New_Deaths']].describe()
```

Out[9]:

	population	Cases	New_Cases	Deaths	New_Deaths
count	6.723880e+05	6.723880e+05	672388.000000	672388.000000	672388.000000
mean	1.044683e+05	2.830715e+04	18.528699	308.897577	0.068233
std	3.334039e+05	9.683983e+04	816.352474	1006.922337	12.226447
min	8.600000e+01	0.000000e+00	-546013.000000	0.000000	-7980.000000
25%	1.090100e+04	2.761000e+03	0.000000	42.000000	0.000000
50%	2.572600e+04	6.981000e+03	0.000000	101.000000	0.000000
75%	6.809800e+04	1.876200e+04	0.000000	239.000000	0.000000
max	1.003911e+07	3.420119e+06	167919.000000	34356.000000	3162.000000

We can see negative numbers in the New Cases and Deaths which could be beacuse of data inconsistency. Let us verify the data for those rows

In [10]:

```
transformed_df.query('New_Cases < 0')
```

Out[10]:

	Date	Week	countyFIPS	County_Name	State	StateFIPS	population	Cases	New_Cases	Deaths	New_Deaths
387	2022-06-01	22	13001	Appling County	GA	13	18386	3558	-1	128	0
389	2022-06-01	22	13005	Bacon County	GA	13	11164	2666	-3	78	0
391	2022-06-01	22	13009	Baldwin County	GA	13	44890	7347	-10	240	0
392	2022-06-01	22	13011	Banks County	GA	13	19234	3432	-9	94	0
393	2022-06-01	22	13013	Barrow County	GA	13	83240	19650	-56	257	0
...	...	...	...	...	...	...	...	...	...	...	...
1602	2022-12-30	52	30009	Carbon County	MT	30	10725	2418	-2	29	0
2178	2022-12-30	52	40095	Marshall County	OK	40	16931	2571	-2837	48	0
2380	2022-12-30	52	46039	Deuel County	SD	46	4351	1172	-1	12	0
2400	2022-12-30	52	46079	Lake County	SD	46	12797	2588	-2	28	0
1178	2022-12-31	52	23003	Aroostook County	ME	23	67055	17709	-2	192	0

2226 rows × 11 columns

Filtered the data for negative New Cases, The first row is for Appling County shows negative New_Cases Let us verify the data for couple of Counties and those dates.

Picking the below entries for analysis Appling County with Date Jun 1st. Barrow County with Date Jun1st Marshall County with Date Dec30 Carbon County with Date Dec30

In [11]:

```
# Fetching cases data for above selected counties and dates(along with before and after) from super_covid
countyIds = [13001, 13013, 30009, 40095]
columns = ['County_Name','2022-05-30_x','2022-05-31_x','2022-06-01_x','2022-06-02_x','2022-12-28_x','2022-12-29_x','2022-12-30_x
super_covid.query(f'countyFIPS in {countyIds}')[columns]
```

Out[11]:

	County_Name	2022-05-30_x	2022-05-31_x	2022-06-01_x	2022-06-02_x	2022-12-28_x	2022-12-29_x	2022-12-30_x	2022-12-31_x
387	Appling County	3559	3559	3558	3558	3757	3757	3794	3794
393	Barrow County	19706	19706	19650	19650	22164	22164	22737	22737
1602	Carbon County	2101	2101	2101	2101	2420	2420	2418	2418
2178	Marshall County	4608	4608	4608	4608	5408	5408	2571	2571

Here we can see there are inconsistencies in the data, the cases/deaths should not decrease as they are total values. So to eliminate these negative values in plotting and mean values. Making those values to zeros

In [12]:

```
transformed_df['New_Cases'] = transformed_df['New_Cases'].apply(lambda x: 0 if x<0 else x)
transformed_df['New_Deaths'] = transformed_df['New_Deaths'].apply(lambda x: 0 if x<0 else x)
```

In [13]:

```
transformed_df.head()
```

Out[13]:

	Date	Week	countyFIPS	County_Name	State	StateFIPS	population	Cases	New_Cases	Deaths	New_Deaths
0	2022-06-01	22	1001	Autauga County	AL	1	55869	15969	6	216	0
1	2022-06-01	22	1003	Baldwin County	AL	1	223234	56580	68	683	0
2	2022-06-01	22	1005	Barbour County	AL	1	24686	5710	3	99	0
3	2022-06-01	22	1007	Bibb County	AL	1	22394	6508	8	105	0
4	2022-06-01	22	1009	Blount County	AL	1	57826	15077	4	244	0

1. Generate weekly statistics (mean, median, mode) for number of new cases and deaths across a specific state.

In [14]:

```
states_list = list(transformed_df['State'].unique())

def get_week_range_string(weekNumber):
    """
    Function to return Week StartDate EndDate (In range of Jun2022 to Dec 2022) string for a given weekNumber in 2022
    """
    week_start = dt.datetime.strptime(f'2022-W{weekNumber}-1', "%Y-W%W-%w")
    week_end = dt.datetime.strptime(f'2022-W{weekNumber}-0', "%Y-W%W-%w")
    start_date = dt.datetime(2022, 6, 1)
    end_date = dt.datetime(2022, 12, 31)
    output_format = '%b-%d'
    if week_start < start_date:
        week_start = start_date
    if week_end > end_date:
        week_end = end_date
    return ' to '.join([week_start.strftime(output_format), week_end.strftime(output_format)])

def state_stats_df(state):
    """
    Function to return a Mean Median Mode statistics Dataframe for a given state
    """
    if state in states_list:
        State_Covid = transformed_df.query(f"State=='{state}'").copy()
        State_aggregate_df = State_Covid.groupby(by=['State','Date','Week']).sum(numeric_only=True).reset_index()
        #State_aggregate_df.drop(columns=['countyFIPS', 'StateFIPS', 'County_Name'], inplace=True)
        aggregations = ['mean', 'median', pd.Series.mode]
        State_Covid_Statistics = State_aggregate_df.groupby(by='Week').agg({'New_Cases': aggregations, 'New_Deaths': aggregations
        State_Covid_Statistics.columns = ['_'.join(col) for col in State_Covid_Statistics.columns.values]
        cols = list(State_Covid_Statistics.columns)
        cols[cols.index('Week_')] = 'Week_Number'
        State_Covid_Statistics.columns = cols
        State_Covid_Statistics['Week_Dates'] = State_Covid_Statistics['Week_Number'].apply(get_week_range_string)
        State_Covid_Statistics['State'] = state
        State_Covid_Statistics['Population'] = State_aggregate_df['population'].unique()[0]
        return State_Covid_Statistics
```

In [15]:

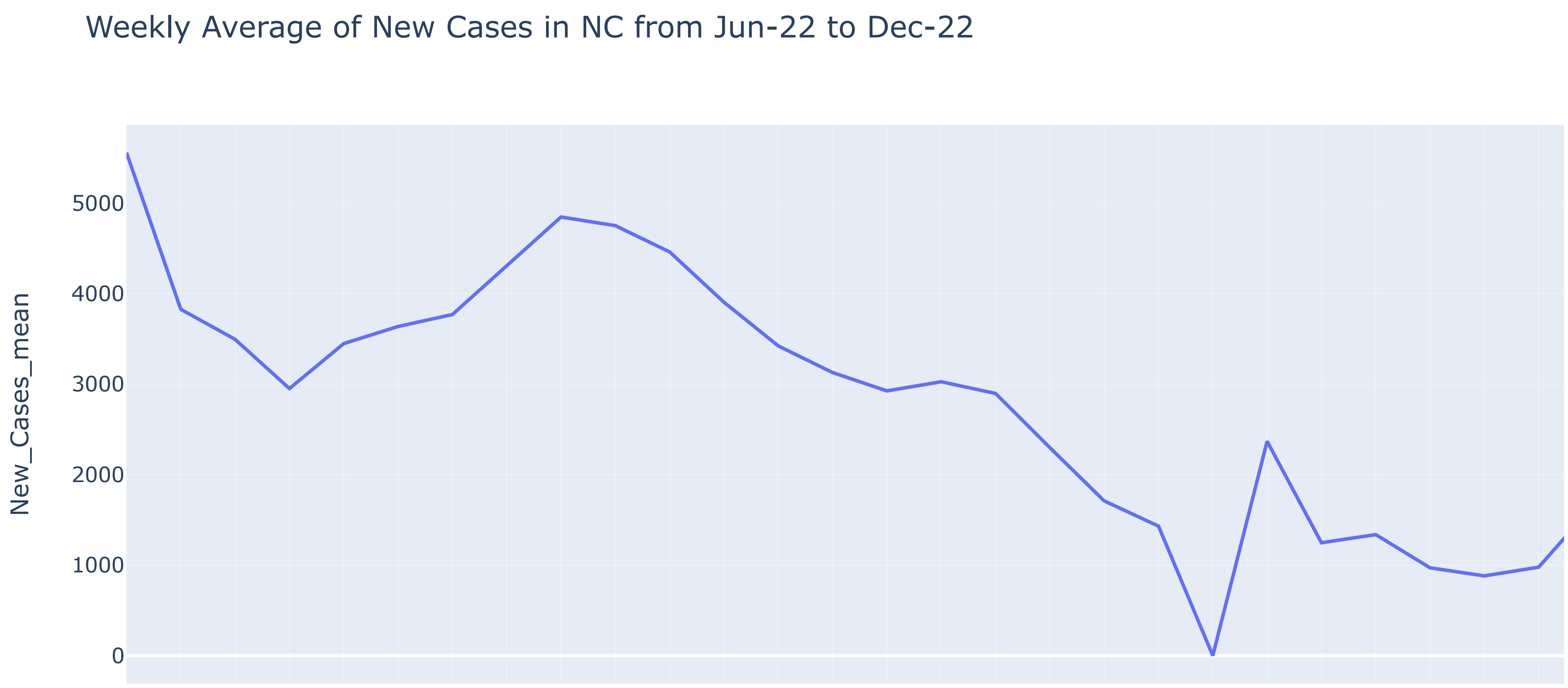
```
#Calculating Stats for NC
NC_Covid_Statistics = state_stats_df('NC')
NC_Covid_Statistics.head()
```

Out[15]:

	Week_Number	New_Cases_mean	New_Cases_median	New_Cases_mode	New_Deaths_mean	New_Deaths_median	New_Deaths_mode	Week_Dates	Sta
0	22	5558.000000	0.0	0	3.000000	0.0	0	Jun-01 to Jun-05	N
1	23	3827.142857	0.0	0	63.000000	0.0	0	Jun-06 to Jun-12	N
2	24	3494.857143	0.0	0	7.142857	0.0	0	Jun-13 to Jun-19	N
3	25	2951.571429	0.0	0	3.142857	0.0	0	Jun-20 to Jun-26	N
4	26	3448.714286	0.0	0	8.428571	0.0	0	Jun-27 to Jul-03	N

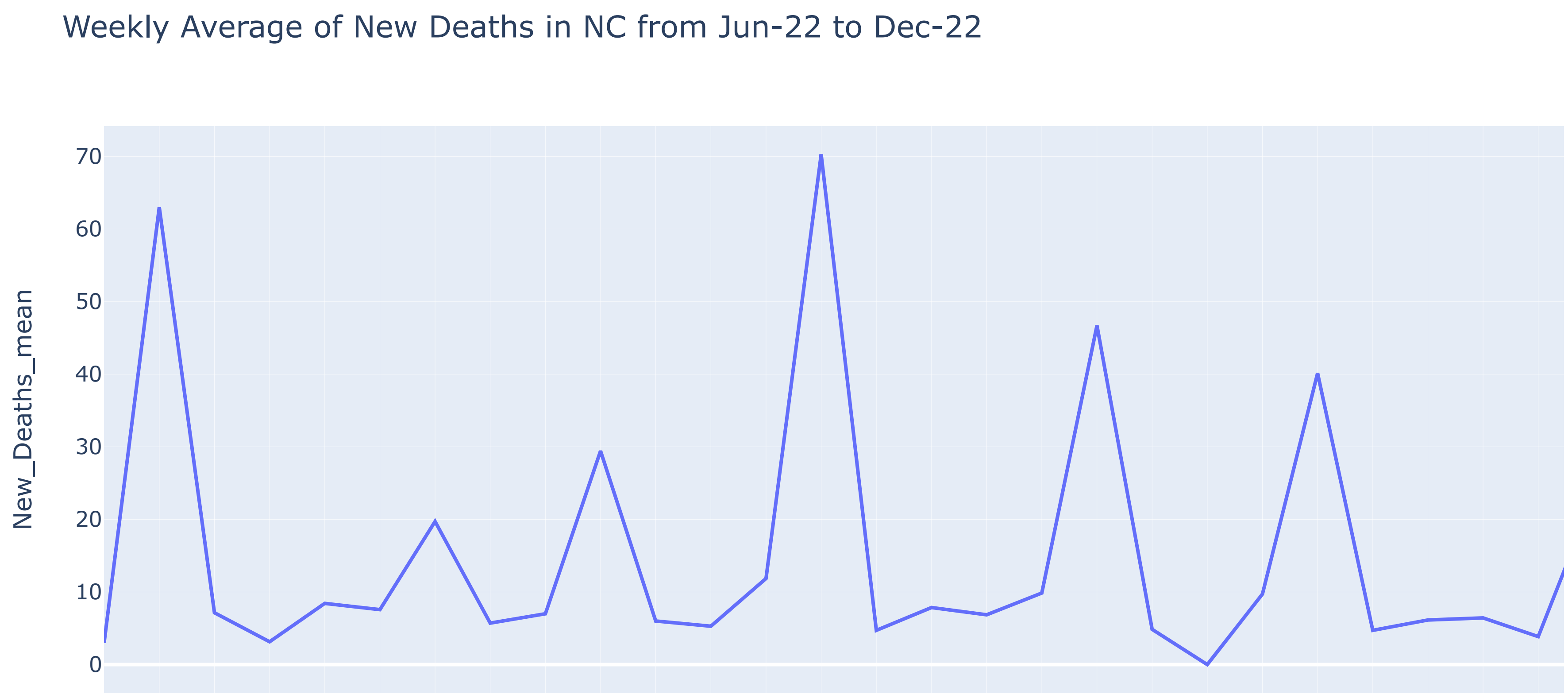
In [16]:

```
#Plotting Weekly Average of New cases for NC
px.line(NC_Covid_Statistics, x='Week_Dates', y='New_Cases_mean', title='Weekly Average of New Cases in NC from Jun-22 to Dec-22'
```



In [17]:

```
#Plotting Weekly Average of New Deaths in NC
px.line(NC_Covid_Statistics, x='Week_Dates', y='New_Deaths_mean', title='Weekly Average of New Deaths in NC from Jun-22 to Dec-2
```



2. Compare the data against 3 other states. Normalize by population, use a normalization factor which is able to identify cases and deaths, for example try per 10,000 or 100,000 (this depends on the population). Plot the values across the weeks in a line plot for the 3 states in a single graph. Describe why the rates differ across these states in the notebook. Identify the peaks, are they consistent with the US pattern?

In [18]:

```
#Calculating Statistics of California
CA_Covid_Statistics = state_stats_df('CA')
CA_Covid_Statistics.head()
```

Out[18]:

	Week_Number	New_Cases_mean	New_Cases_median	New_Cases_mode	New_Deaths_mean	New_Deaths_median	New_Deaths_mode	Week_Dates	Sta
0	22	2657.000000	0.0	0	0.200000	0.0	0	Jun-01 to Jun-05	C
1	23	25480.285714	0.0	0	42.000000	0.0	0	Jun-06 to Jun-12	C
2	24	15981.714286	8637.0	[335, 1023, 3448, 8637, 21846, 27526, 49057]	43.714286	4.0	0	Jun-13 to Jun-19	C
3	25	13680.142857	10147.0	[1867, 5252, 6898, 10147, 12709, 14015, 44873]	15.428571	1.0	[0, 1, 3]	Jun-20 to Jun-26	C
4	26	21643.285714	17789.0	[0, 1368, 11346, 17789, 18047, 19695, 83258]	132.857143	33.0	0	Jun-27 to Jul-03	C

In [19]:

```
# Calculating Statistics of NewYork
NY_Covid_Statistics = state_stats_df('NY')
NY_Covid_Statistics.head()
```

Out[19]:

	Week_Number	New_Cases_mean	New_Cases_median	New_Cases_mode	New_Deaths_mean	New_Deaths_median	New_Deaths_mode	Week_Dates	Sta
0	22	6114.200000	5812.0	0	26.200000	28.0	0	Jun-01 to Jun-05	N
1	23	5500.000000	5663.0	0	21.428571	27.0	0	Jun-06 to Jun-12	N
2	24	4811.428571	5176.0	0	23.000000	30.0	[0, 30]	Jun-13 to Jun-19	N
3	25	5083.285714	4211.0	0	18.142857	13.0	0	Jun-20 to Jun-26	N
4	26	5674.571429	0.0	0	17.857143	20.0	0	Jun-27 to Jul-03	N

In [20]:

```
#Calculating Statistics of Washington
WA_Covid_Statistics = state_stats_df('WA')
WA_Covid_Statistics.head()
```

Out[20]:

	Week_Number	New_Cases_mean	New_Cases_median	New_Cases_mode	New_Deaths_mean	New_Deaths_median	New_Deaths_mode	Week_Dates	Sta
0	22	3599.200000	0.0	0	10.000000	0.0	0	Jun-01 to Jun-05	W
1	23	2734.857143	0.0	0	11.571429	0.0	0	Jun-06 to Jun-12	W
2	24	2512.571429	0.0	0	12.714286	0.0	0	Jun-13 to Jun-19	W
3	25	2610.000000	0.0	0	10.142857	0.0	0	Jun-20 to Jun-26	W
4	26	2803.714286	0.0	0	13.571429	0.0	0	Jun-27 to Jul-03	W

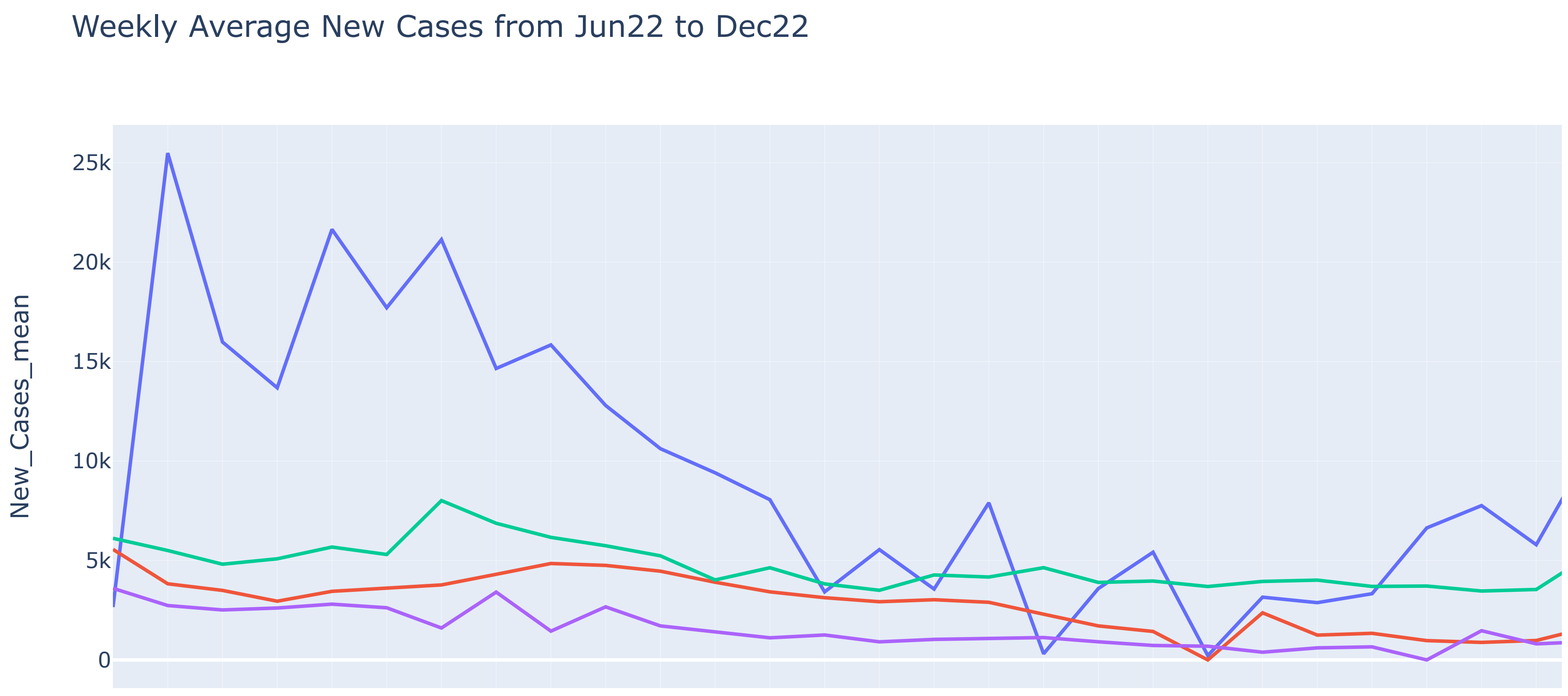
In [21]:

```
# Merging the new 3 states with inital NC for comparison
Four_states_covid_stats = pd.concat([CA_Covid_Statistics,NC_Covid_Statistics,NY_Covid_Statistics,WA_Covid_Statistics],axis=0)
Four_states_covid_stats.head()
```

Out[21]:

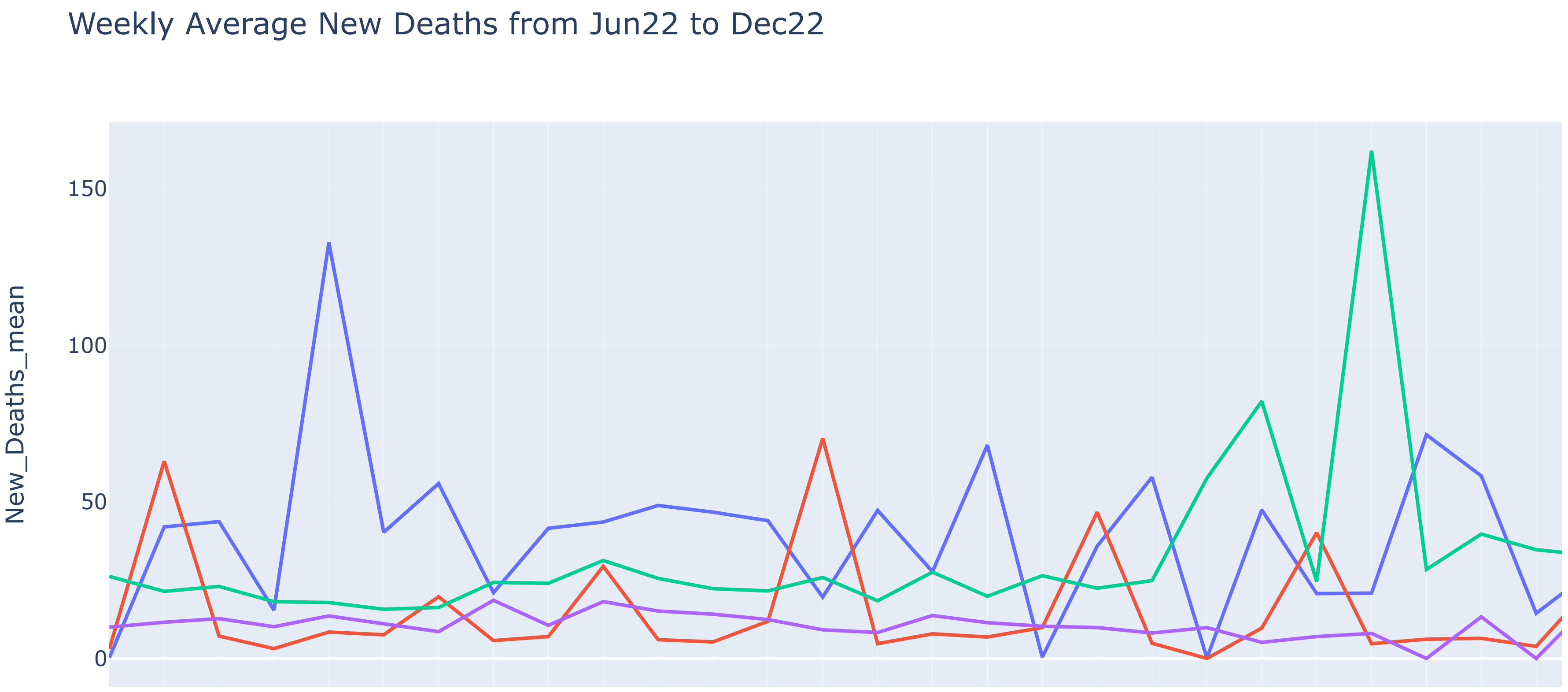
	Week_Number	New_Cases_mean	New_Cases_median	New_Cases_mode	New_Deaths_mean	New_Deaths_median	New_Deaths_mode	Week_Dates	Sta
0	22	2657.000000	0.0	0	0.200000	0.0	0	Jun-01 to Jun-05	C
1	23	25480.285714	0.0	0	42.000000	0.0	0	Jun-06 to Jun-12	C
2	24	15981.714286	8637.0	[335, 1023, 3448, 8637, 21846, 27526, 49057]	43.714286	4.0	0	Jun-13 to Jun-19	C
3	25	13680.142857	10147.0	[1867, 5252, 6898, 10147, 12709, 14015, 44873]	15.428571	1.0	[0, 1, 3]	Jun-20 to Jun-26	C
4	26	21643.285714	17789.0	[0, 1368, 11346, 17789, 18047, 19695, 83258]	132.857143	33.0	0	Jun-27 to Jul-03	C

In [22]:
```
#Plotting the Weekly Average New Cases for the 4 states
px.line(Four_states_covid_stats,x='Week_Dates',y='New_Cases_mean',color='State', title = 'Weekly Average New Cases from Jun22 to
```



- In the above graph we can see that for most of the graph area, California has more number of cases. This Could be because of highest population in California
- In the Week of Jun 6th to Jun 12th above graph we can see California has highest number of cases, where as in other states there was a decline trend for that week.
- Starting the week "Jul-25 to Jul-31" there was either decline or stable trend in almost all states For a couple of months. Again in the end, we can see there is raise in cases. This could be because of Holiday season (Halloween, Thanksgiving, Christmas)
- Since there is difference in the population size, it will be difficult to clearly compare the patterns in this graph. So let us compare this again in a normalized chart

In [23]:
```
# Plotting the Weekly Average New Deaths for the 4 states
px.line(Four_states_covid_stats,x='Week_Dates',y='New_Deaths_mean',color='State', title='Weekly Average New Deaths from Jun22 to
```


Weekly Average New Deaths from Jun22 to Dec22

- From the above graph we can see that the covid deaths data has no particular pattern
- However the highest peak for CA is in Jun27 to Jul03 week i.e. mid of the year and highest peak for NY is in the end of the year for the week Nov07 to Nov13.
- Just like the cases, we cannot predict if some event has complete correlation with Deaths. As deaths occur either immediately or some time later once a person is infected.
- Let us see if we can identify some pattern in normalized data

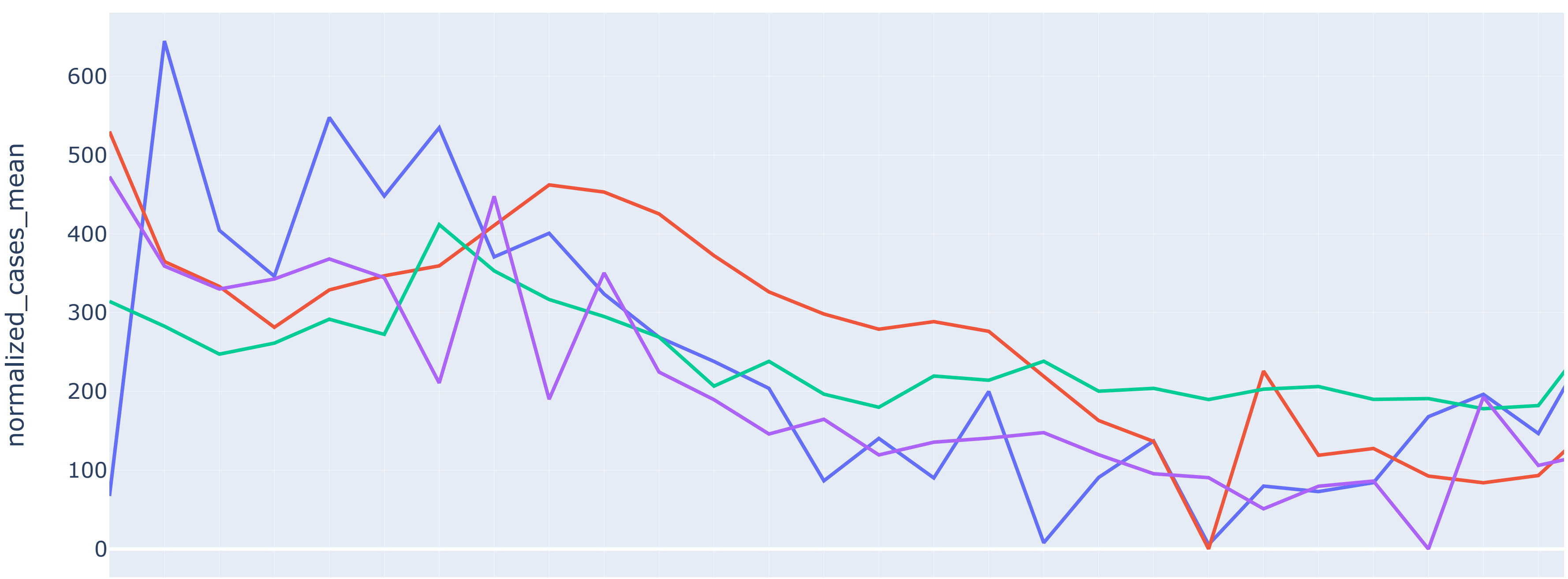
In [24]:
```
#Normalizing the values per 1M population
Four_states_covid_stats['normalized_cases_mean'] =  1000000 * Four_states_covid_stats['New_Cases_mean']/Four_states_covid_stats[
Four_states_covid_stats['normalized_deaths_mean'] =  1000000 * Four_states_covid_stats['New_Deaths_mean']/Four_states_covid_stat
Four_states_covid_stats.head()
```

Out[24]:

	Week_Number	New_Cases_mean	New_Cases_median	New_Cases_mode	New_Deaths_mean	New_Deaths_median	New_Deaths_mode	Week_Dates	Sta
0	22	2657.000000	0.0	0	0.200000	0.0	0	Jun-01 to Jun-05	C
1	23	25480.285714	0.0	0	42.000000	0.0	0	Jun-06 to Jun-12	C
2	24	15981.714286	8637.0	[335, 1023, 3448, 8637, 21846, 27526, 49057]	43.714286	4.0	0	Jun-13 to Jun-19	C
3	25	13680.142857	10147.0	[1867, 5252, 6898, 10147, 12709, 14015, 44873]	15.428571	1.0	[0, 1, 3]	Jun-20 to Jun-26	C
4	26	21643.285714	17789.0	[0, 1368, 11346, 17789, 18047, 19695, 83258]	132.857143	33.0	0	Jun-27 to Jul-03	C

In [25]:
```
# Plotting Weekly Average Cases per 1M population for 4 states
px.line(Four_states_covid_stats,x='Week_Dates',y='normalized_cases_mean',color='State', title="Weekly Average New Cases per 1M Po
```

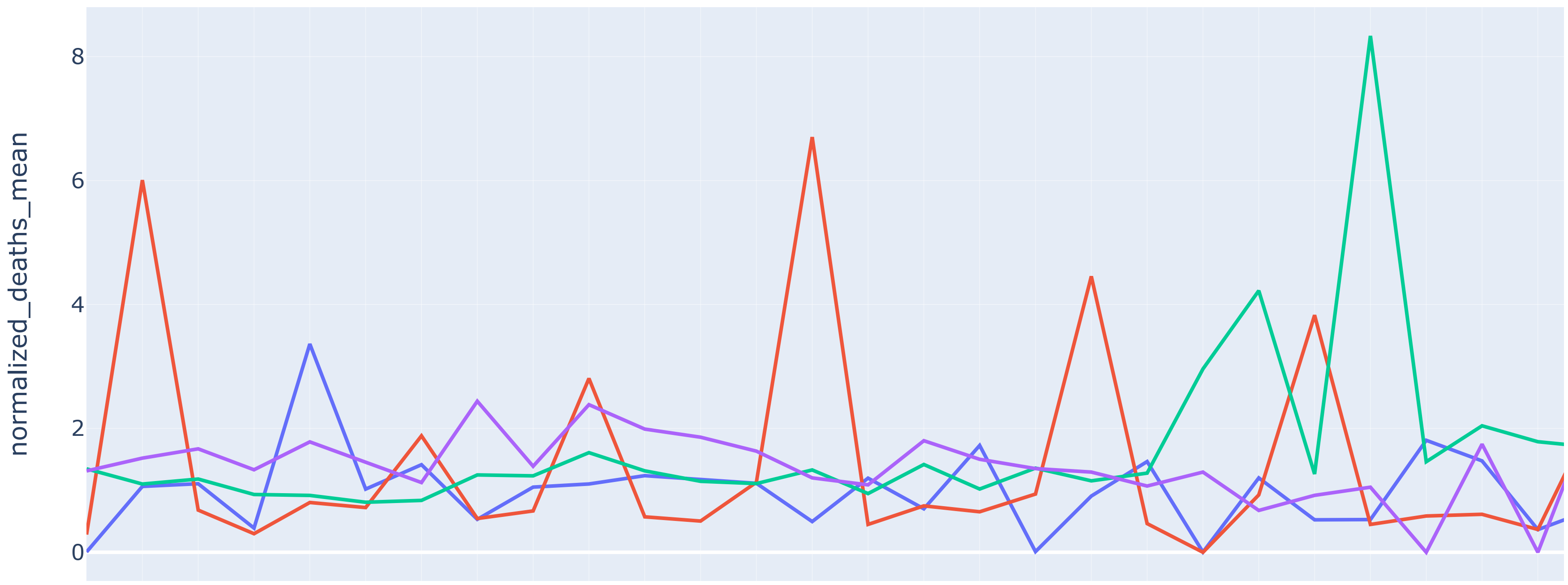
Weekly Average New Cases per 1M Population from Jun22 to Dec22



- In the above graph after normalizing the pattern is clearly evident where the weekly average of new cases is high in the mid of the year 2022 and gradually decreasing till November and then increased.
- This pattern can be highly correlated with holidays.
- In summer people would have gone for vacations which resulted in high number of new cases in the mid of the year.
- Once everyone gets back to their daily routine, slowly the new cases decreased and raised sharply in the winter holiday season
- NY and CA being more populated, were infected more during the winter holidays and has more number of cases per 1M population

In [26]:
```
# Plotting Weekly Average New Deaths per 1M population for 4 States
px.line(Four_states_covid_stats,x='Week_Dates',y='normalized_deaths_mean',color='State', title='Weekly Average New Deaths per 1M
```

Weekly Average New Deaths per 1M population from Jun22 to Dec22



- In the above plot, we can see for the most part all the states had similar range of Deaths with NC having some peaks at many stages
- In this link we can see NC has more number of hospitalizations when compared to other states in the analysis. https://www.nbcnews.com/data-graphics/covid-hospitalizations-see-latest-trend-current-count-rcna61053
- Hence NC could have more number of death peaks than other states

In [27]:
```
CA_peaks = CA_Covid_Statistics[CA_Covid_Statistics.index.isin(find_peaks(CA_Covid_Statistics['New_Cases_mean'],width=1)[0])]
```

In [28]:
```
NC_peaks = NC_Covid_Statistics[NC_Covid_Statistics.index.isin(find_peaks(NC_Covid_Statistics['New_Cases_mean'],width=1)[0])]
```

In [29]:
```
NY_peaks = NY_Covid_Statistics[NY_Covid_Statistics.index.isin(find_peaks(NY_Covid_Statistics['New_Cases_mean'],width=1)[0])]
```

In [30]:
```
WA_peaks = WA_Covid_Statistics[WA_Covid_Statistics.index.isin(find_peaks(WA_Covid_Statistics['New_Cases_mean'], width=1)[0])]
```

In [31]:
```
#fetching US level metrics using transformed_df
aggregated_super_covid = transformed_df.groupby(by=['Date','Week']).sum(numeric_only=True).reset_index()
aggregated_super_covid.head()
```

Out[31]:

	Date	Week	population	Cases	New_Cases	Deaths	New_Deaths
0	2022-06-01	22	328239523	81427445	169355	946824	498
1	2022-06-02	22	328239523	81494654	68697	947016	196
2	2022-06-03	22	328239523	81701504	206870	947235	227
3	2022-06-04	22	328239523	81712058	10554	947279	46
4	2022-06-05	22	328239523	81737066	25008	947279	0

In [32]:
```
US_covid_statistics = aggregated_super_covid.groupby(by=['Week','population']).agg({'New_Cases': 'mean', 'New_Deaths': 'mean'}).
US_covid_statistics['Week_Dates'] = US_covid_statistics['Week'].apply(get_week_range_string)
US_covid_statistics = US_covid_statistics.rename(columns={'New_Cases':'New_Cases_mean','New_Deaths':'New_Deaths_mean'})
US_covid_statistics.head()
```

Out[32]:

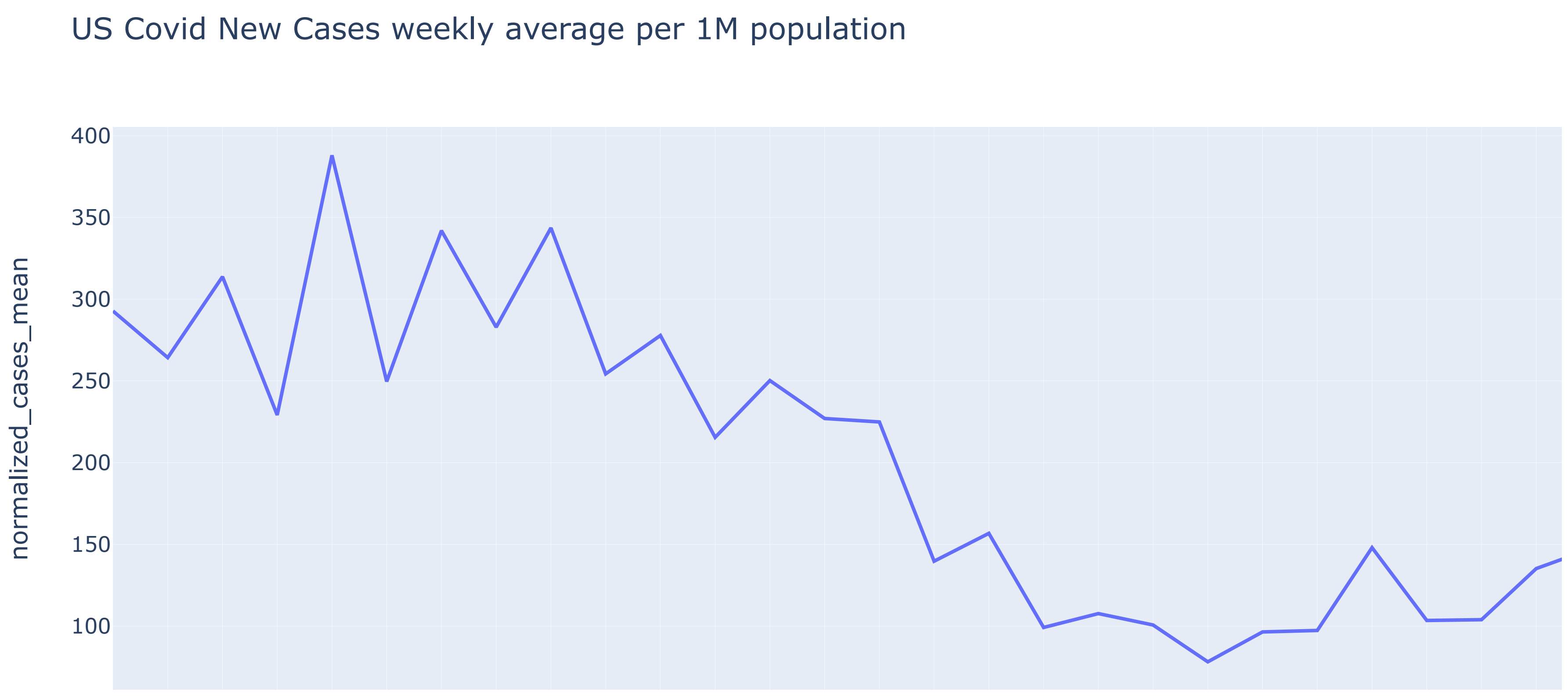
	Week	population	New_Cases_mean	New_Deaths_mean	Week_Dates
0	22	328239523	96096.800000	193.400000	Jun-01 to Jun-05
1	23	328239523	86738.285714	301.142857	Jun-06 to Jun-12
2	24	328239523	102986.571429	262.285714	Jun-13 to Jun-19
3	25	328239523	75216.571429	228.000000	Jun-20 to Jun-26
4	26	328239523	127359.428571	841.000000	Jun-27 to Jul-03

In [33]:
```
US_covid_statistics['normalized_cases_mean'] =  1000000 * US_covid_statistics['New_Cases_mean']/US_covid_statistics['population'
US_covid_statistics['normalized_deaths_mean'] =  1000000 * US_covid_statistics['New_Deaths_mean']/US_covid_statistics['populatio
US_covid_statistics.head()
```

Out[33]:

	Week	population	New_Cases_mean	New_Deaths_mean	Week_Dates	normalized_cases_mean	normalized_deaths_mean
0	22	328239523	96096.800000	193.400000	Jun-01 to Jun-05	292.764257	0.589204
1	23	328239523	86738.285714	301.142857	Jun-06 to Jun-12	264.253021	0.917448
2	24	328239523	102986.571429	262.285714	Jun-13 to Jun-19	313.754329	0.799068
3	25	328239523	75216.571429	228.000000	Jun-20 to Jun-26	229.151477	0.694615
4	26	328239523	127359.428571	841.000000	Jun-27 to Jul-03	388.007597	2.562153

In [34]:
```
px.line(US_covid_statistics,x='Week_Dates',y='normalized_cases_mean',title='US Covid New Cases weekly average per 1M population'
```



In [35]:
```
US_peaks = US_covid_statistics[US_covid_statistics.index.isin(find_peaks(US_covid_statistics['New_Cases_mean'], width=1)[0])]
US_peaks
```

Out[35]:

	Week	population	New_Cases_mean	New_Deaths_mean	Week_Dates	normalized_cases_mean	normalized_deaths_mean
4	26	328239523	127359.428571	841.000000	Jun-27 to Jul-03	388.007597	2.562153
8	30	328239523	112782.142857	275.857143	Jul-25 to Jul-31	343.597084	0.840414
12	34	328239523	82138.571429	355.000000	Aug-22 to Aug-28	250.239736	1.081527
18	40	328239523	35346.714286	334.285714	Oct-03 to Oct-09	107.685735	1.018420
28	50	328239523	50453.714286	345.142857	Dec-12 to Dec-18	153.710052	1.051497

In [36]:

```python
# Adding the US data to the previous 4 states data
US_covid_statistics['State'] = 'US'
US_covid_statistics.head()
```

Out[36]:

	Week	population	New_Cases_mean	New_Deaths_mean	Week_Dates	normalized_cases_mean	normalized_deaths_mean	State
0	22	328239523	96096.800000	193.400000	Jun-01 to Jun-05	292.764257	0.589204	US
1	23	328239523	86738.285714	301.142857	Jun-06 to Jun-12	264.253021	0.917448	US
2	24	328239523	102986.571429	262.285714	Jun-13 to Jun-19	313.754329	0.799068	US
3	25	328239523	75216.571429	228.000000	Jun-20 to Jun-26	229.151477	0.694615	US
4	26	328239523	127359.428571	841.000000	Jun-27 to Jul-03	388.007597	2.562153	US

In [37]:

```python
cols = ['Week_Dates','normalized_cases_mean','normalized_deaths_mean','State']
us_and_states_merged_stats = pd.concat([Four_states_covid_stats[cols],US_covid_statistics[cols]],axis=0)
us_and_states_merged_stats.head()
```

Out[37]:

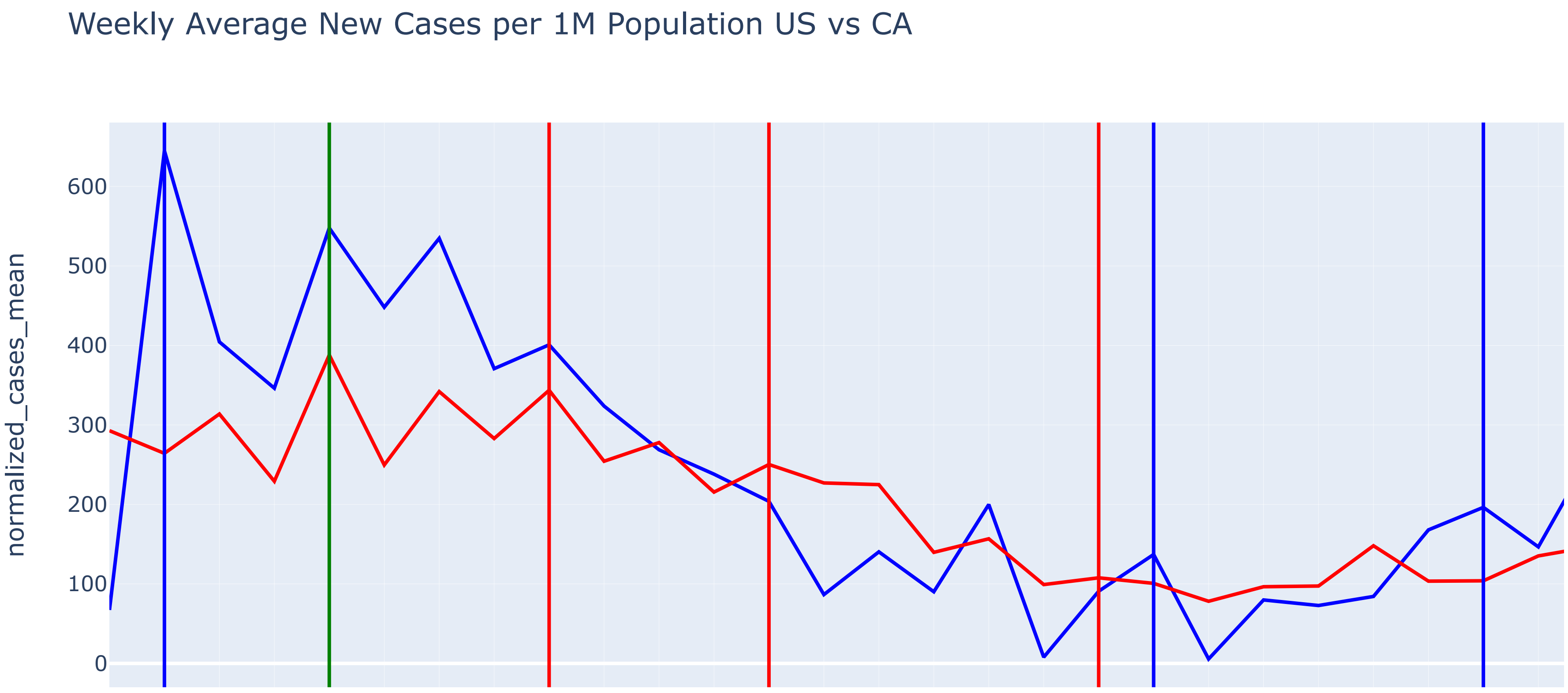
	Week_Dates	normalized_cases_mean	normalized_deaths_mean	State
0	Jun-01 to Jun-05	67.245014	0.005062	CA
1	Jun-06 to Jun-12	644.870974	1.062962	CA
2	Jun-13 to Jun-19	404.475200	1.106348	CA
3	Jun-20 to Jun-26	346.225593	0.390476	CA
4	Jun-27 to Jul-03	547.761783	3.362431	CA

In [38]:

```python
peaks = {'CA': CA_peaks['Week_Dates'].to_list(),
         'NC': NC_peaks['Week_Dates'].to_list(),
         'WA': WA_peaks['Week_Dates'].to_list(),
         'NY': NY_peaks['Week_Dates'].to_list(),
         'US': US_peaks['Week_Dates'].to_list()
        }
def plot_state_vs_US_peaks(state):
    state_peaks_list = peaks[state]
    US_peaks_list = peaks['US']
    matching_peaks = list(set(state_peaks_list).intersection(set(US_peaks_list)))
    unmatched_state_peaks = list(set(state_peaks_list).difference(set(US_peaks_list)))
    unmatched_US_peaks = list(set(US_peaks_list).difference(set(state_peaks_list)))
    fig = px.line(us_and_states_merged_stats.query(f"State in ['{state}','US']"),
                  x='Week_Dates',y='normalized_cases_mean',color='State',
                  title=f"Weekly Average New Cases per 1M Population US vs {state}",
                  color_discrete_map={
                 "US": "red",
                 state: "blue"
                 })
    for week in unmatched_state_peaks:
        fig.add_vline(x=week, line_color='blue')
    for week in unmatched_US_peaks:
        fig.add_vline(x=week, line_color='red')
    for week in matching_peaks:
        fig.add_vline(x=week, line_color='green')
    return fig
```

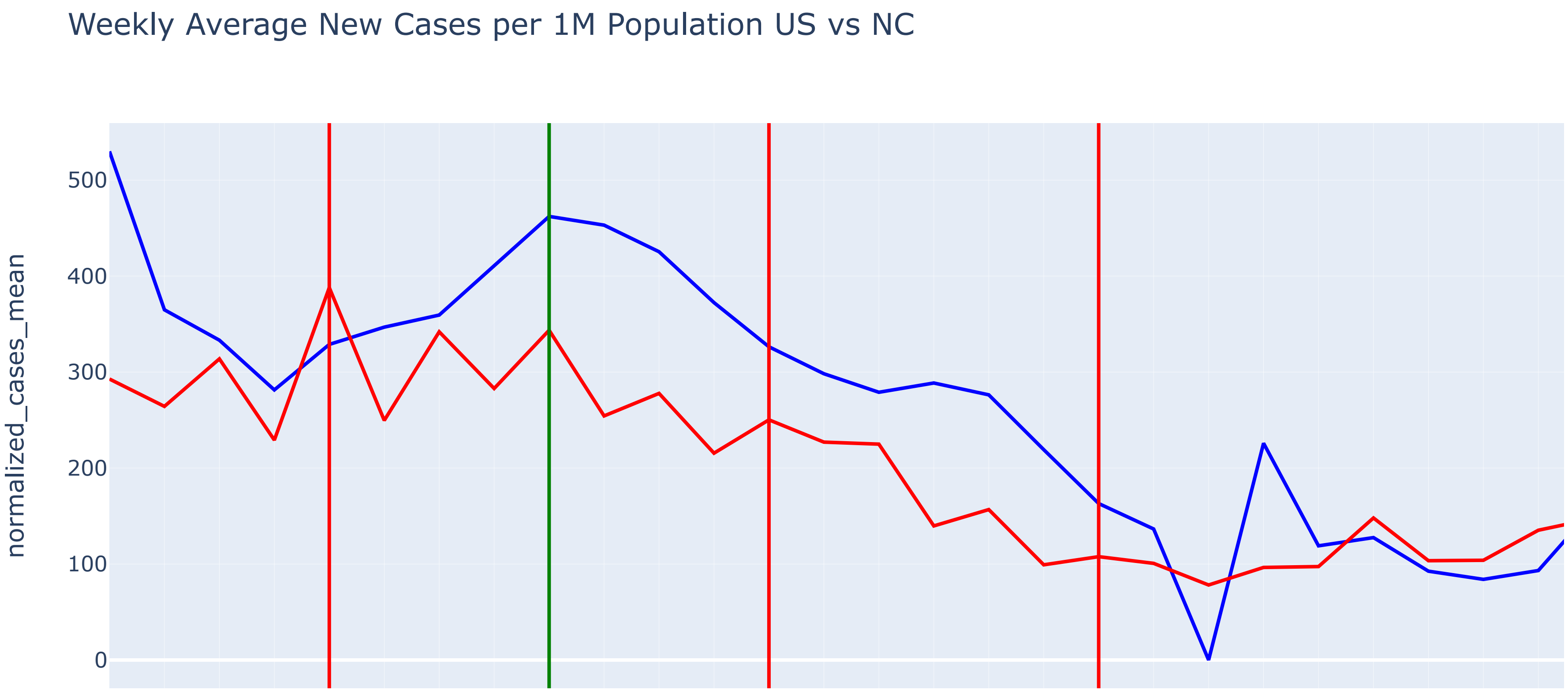
In [39]:

```python
plot_state_vs_US_peaks('CA').show()
```



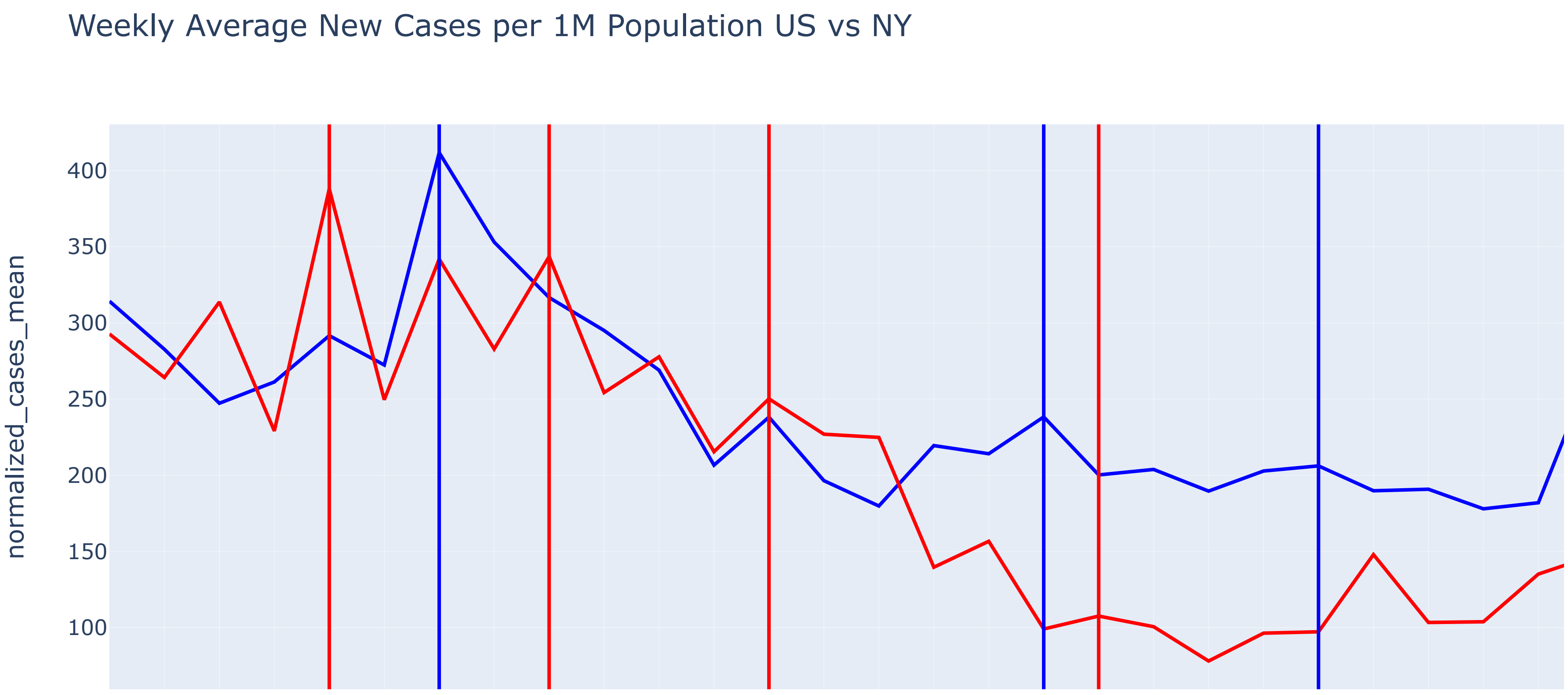
- In the above graph, the vertical blue lines identifies state peaks, red lines identifies US peaks and the green lines identify peaks that are matching for both US and the state
- In the above graph we can see that the new cases weekly average has only one matching peak for CA and US and that is the highest peak for the US (Week Jun27 to Jul03)
- The highest peak for CA is in Week Jun06 to Jun12 but there is no increase in cases for US during that week
- However overall the pattern for US and CA state matches roughly with high cases in month of July and decreased the following months until November and there is increase in the cases

In [40]:
```
plot_state_vs_US_peaks('NC').show()
```



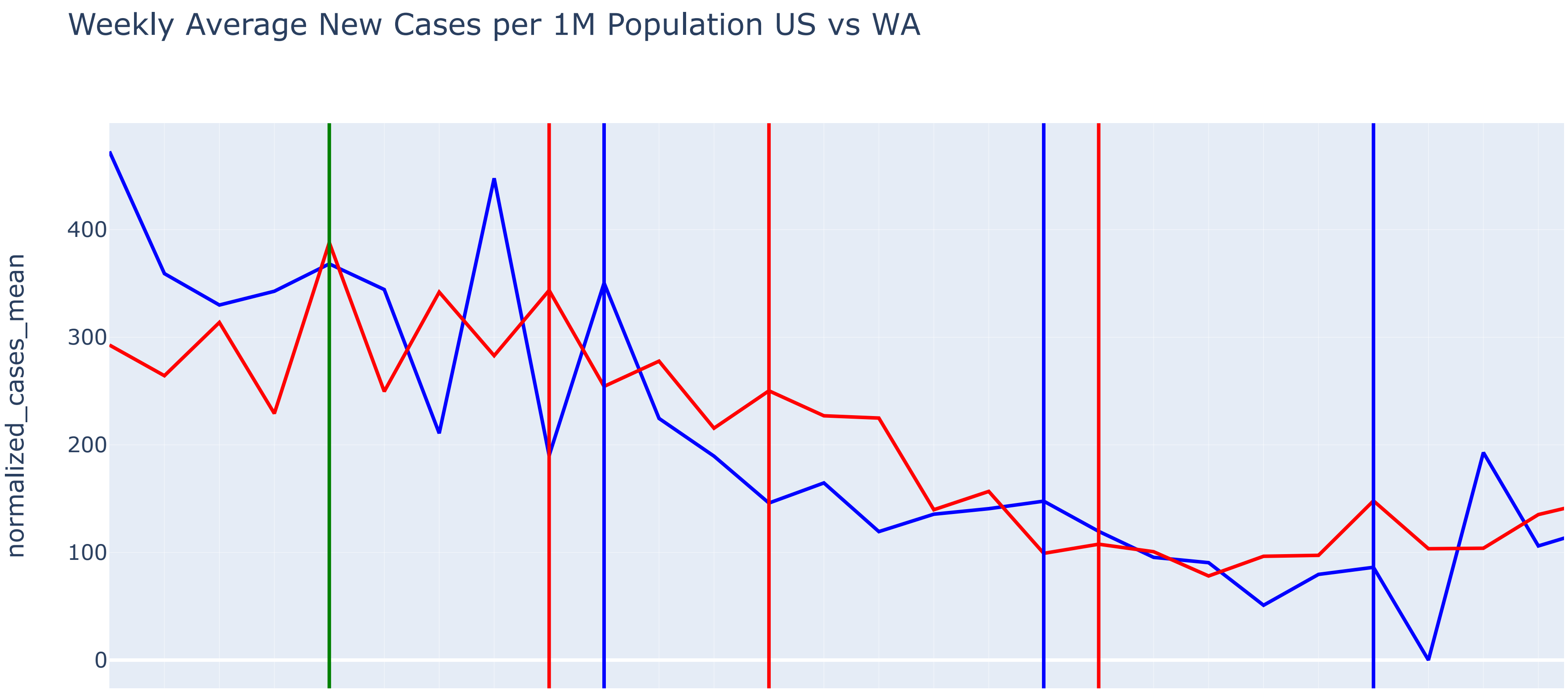
- In the above graph, the vertical blue lines identifies state peaks, red lines identifies US peaks and the green lines identify peaks that are matching for both US and the state
- In the above graph we can see that the new cases weekly average has only one matching peak for NC and US and that is the second highest peak for the NC (Week JuL25 to Jul31)
- The highest peak for NC is in the first Week of June where the new cases were also higher than the subsequent week for US
- At the highest peak of US we can se that there was raise in cases of NC
- Overall the pattern for US and NC state matches roughly with high cases in month of July and decreased the following months until November and there is increase in the cases. However in December the cases fell drastically for NC

```
In [41]: plot_state_vs_US_peaks('NY').show()
```



- In the above graph, the vertical blue lines identifies state peaks, red lines identifies US peaks and the green lines identify peaks that are matching for both US and the state
- In the above graph we can see that the new cases weekly average has only one matching peak for NY and US and that is in the end of the year (week Dec12 to Dec18)
- The highest peak for NY is in Week Jul11 to Jul17 and there is also increase in cases for US during that week
- The US highest peak is the week of Jun27 to Jul03 and there is also increase in cases for NY during that week
- However overall the pattern for US and NY state matches roughly with high cases in month of July and decreased the following months until November and there is increase in the cases

```
In [42]: plot_state_vs_US_peaks('WA').show()
```



- In the above graph, the vertical blue lines identifies state peaks, red lines identifies US peaks and the green lines identify peaks that are matching for both US and the state
- In the above graph we can see that the new cases weekly average has two matching peaks for WA and US and that is in the mid and end of the year (weeks "Jun27 to Jul03" "Dec12 to Dec18")
- The highest peak for WA is in Week Jul18 to Jul24 but there is decrease in cases for US during that week

- The US highest peak is the week of Jun27 to Jul03 and there is also increase in cases for WA during that week
- However overall the pattern for US and NY state matches roughly with high cases in month of July and decreased the following months until November and there is increase in the cases. But in the December WA cases started decreasing whereas the US cases increases

In [43]:
```
CA_death_peaks = CA_Covid_Statistics[CA_Covid_Statistics.index.isin(find_peaks(CA_Covid_Statistics['New_Deaths_mean'],width=1)[0
CA_death_peaks
```

Out[43]:

	Week_Number	New_Cases_mean	New_Cases_median	New_Cases_mode	New_Deaths_mean	New_Deaths_median	New_Deaths_mode	Week_Dates	St
2	24	15981.714286	8637.0	[335, 1023, 3448, 8637, 21846, 27526, 49057]	43.714286	4.0	0	Jun-13 to Jun-19	
4	26	21643.285714	17789.0	[0, 1368, 11346, 17789, 18047, 19695, 83258]	132.857143	33.0	0	Jun-27 to Jul-03	
10	32	10619.142857	6898.0	[981, 1381, 3599, 6898, 7220, 24774, 29481]	48.857143	1.0	0	Aug-08 to Aug-14	
19	41	5414.714286	0.0	0	57.857143	0.0	0	Oct-10 to Oct-16	
24	46	6636.714286	0.0	0	71.428571	0.0	0	Nov-14 to Nov-20	
29	51	13517.857143	7035.0	[2686, 5394, 6171, 7035, 8105, 8914, 56320]	78.714286	8.0	[0, 1, 5, 8, 10, 15, 512]	Dec-19 to Dec-25	

In [44]:
```
NC_death_peaks = NC_Covid_Statistics[NC_Covid_Statistics.index.isin(find_peaks(NC_Covid_Statistics['New_Deaths_mean'],width=1)[0
NC_death_peaks
```

Out[44]:

	Week_Number	New_Cases_mean	New_Cases_median	New_Cases_mode	New_Deaths_mean	New_Deaths_median	New_Deaths_mode	Week_Dates	St
1	23	3827.142857	0.0	0	63.000000	0.0	0	Jun-06 to Jun-12	
6	28	3769.571429	0.0	0	19.714286	0.0	0	Jul-11 to Jul-17	
9	31	4752.142857	0.0	0	29.428571	0.0	0	Aug-01 to Aug-07	
13	35	3128.285714	0.0	0	70.285714	0.0	0	Aug-29 to Sep-04	
18	40	1710.142857	0.0	0	46.714286	0.0	0	Oct-03 to Oct-09	
22	44	1247.857143	0.0	0	40.142857	0.0	0	Oct-31 to Nov-06	
25	47	881.428571	0.0	0	6.428571	0.0	0	Nov-21 to Nov-27	
27	49	1657.857143	0.0	0	23.142857	0.0	0	Dec-05 to Dec-11	

In [45]:
```
WA_death_peaks = WA_Covid_Statistics[WA_Covid_Statistics.index.isin(find_peaks(WA_Covid_Statistics['New_Deaths_mean'],width=1)[0
WA_death_peaks
```

Out[45]:

	Week_Number	New_Cases_mean	New_Cases_median	New_Cases_mode	New_Deaths_mean	New_Deaths_median	New_Deaths_mode	Week_Dates	St
2	24	2512.571429	0.0	0	12.714286	0.0	0	Jun-13 to Jun-19	
4	26	2803.714286	0.0	0	13.571429	0.0	0	Jun-27 to Jul-03	
7	29	3409.428571	0.0	0	18.571429	0.0	0	Jul-18 to Jul-24	
9	31	2668.571429	0.0	0	18.142857	0.0	0	Aug-01 to Aug-07	
15	37	1032.428571	0.0	0	13.714286	0.0	0	Sep-12 to Sep-18	
23	45	656.428571	0.0	0	8.000000	0.0	0	Nov-07 to Nov-13	
25	47	1468.857143	0.0	0	13.285714	0.0	0	Nov-21 to Nov-27	

In [46]:
```
NY_death_peaks = NY_Covid_Statistics[NY_Covid_Statistics.index.isin(find_peaks(NY_Covid_Statistics['New_Deaths_mean'],width=1)[0
NY_death_peaks
```

Out[46]:

	Week_Number	New_Cases_mean	New_Cases_median	New_Cases_mode	New_Deaths_mean	New_Deaths_median	New_Deaths_mode	Week_Dates	St
9	31	5740.571429	6502.0	0	31.285714	29.0	0	Aug-01 to Aug-07	
21	43	3946.571429	4390.0	0	82.142857	36.0	0	Oct-24 to Oct-30	
25	47	3463.571429	3931.0	0	39.714286	16.0	0	Nov-21 to Nov-27	

In [47]:

```
US_death_peaks = US_covid_statistics[US_covid_statistics.index.isin(find_peaks(US_covid_statistics['New_Deaths_mean'], width=1)[
US_death_peaks
```

Out[47]:

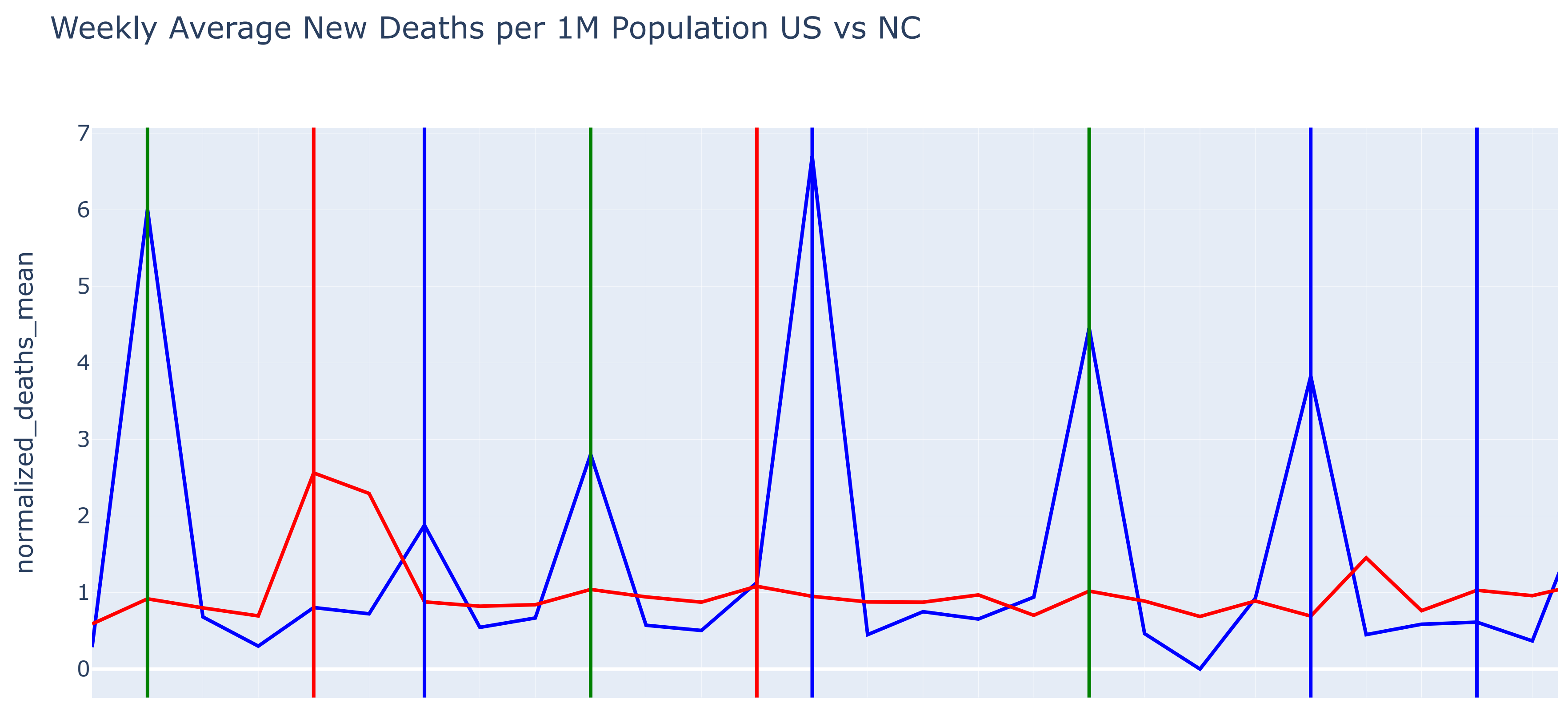
	Week	population	New_Cases_mean	New_Deaths_mean	Week_Dates	normalized_cases_mean	normalized_deaths_mean	State
1	23	328239523	86738.285714	301.142857	Jun-06 to Jun-12	264.253021	0.917448	US
4	26	328239523	127359.428571	841.000000	Jun-27 to Jul-03	388.007597	2.562153	US
9	31	328239523	83478.428571	341.428571	Aug-01 to Aug-07	254.321685	1.040181	US
12	34	328239523	82138.571429	355.000000	Aug-22 to Aug-28	250.239736	1.081527	US
18	40	328239523	35346.714286	334.285714	Oct-03 to Oct-09	107.685735	1.018420	US
27	49	328239523	48360.142857	369.714286	Dec-05 to Dec-11	147.331870	1.126355	US

In [48]:

```
death_peaks = {'CA': CA_death_peaks['Week_Dates'].to_list(),
          'NC': NC_death_peaks['Week_Dates'].to_list(),
          'WA': WA_death_peaks['Week_Dates'].to_list(),
          'NY': NY_death_peaks['Week_Dates'].to_list(),
          'US': US_death_peaks['Week_Dates'].to_list()
         }
def plot_state_vs_US_death_peaks(state):
    state_peaks_list = death_peaks[state]
    US_peaks_list = death_peaks['US']
    matching_peaks = list(set(state_peaks_list).intersection(set(US_peaks_list)))
    unmatched_state_peaks = list(set(state_peaks_list).difference(set(US_peaks_list)))
    unmatched_US_peaks = list(set(US_peaks_list).difference(set(state_peaks_list)))
    fig = px.line(us_and_states_merged_stats.query(f"State in ['{state}','US']"),
                  x='Week_Dates',y='normalized_deaths_mean',color='State',
                  title=f"Weekly Average New Deaths per 1M Population US vs {state}",
                  color_discrete_map={
                 "US": "red",
                 state: "blue"
                 })
    for week in unmatched_state_peaks:
        fig.add_vline(x=week, line_color='blue')
    for week in unmatched_US_peaks:
        fig.add_vline(x=week, line_color='red')
    for week in matching_peaks:
        fig.add_vline(x=week, line_color='green')
    return fig
```

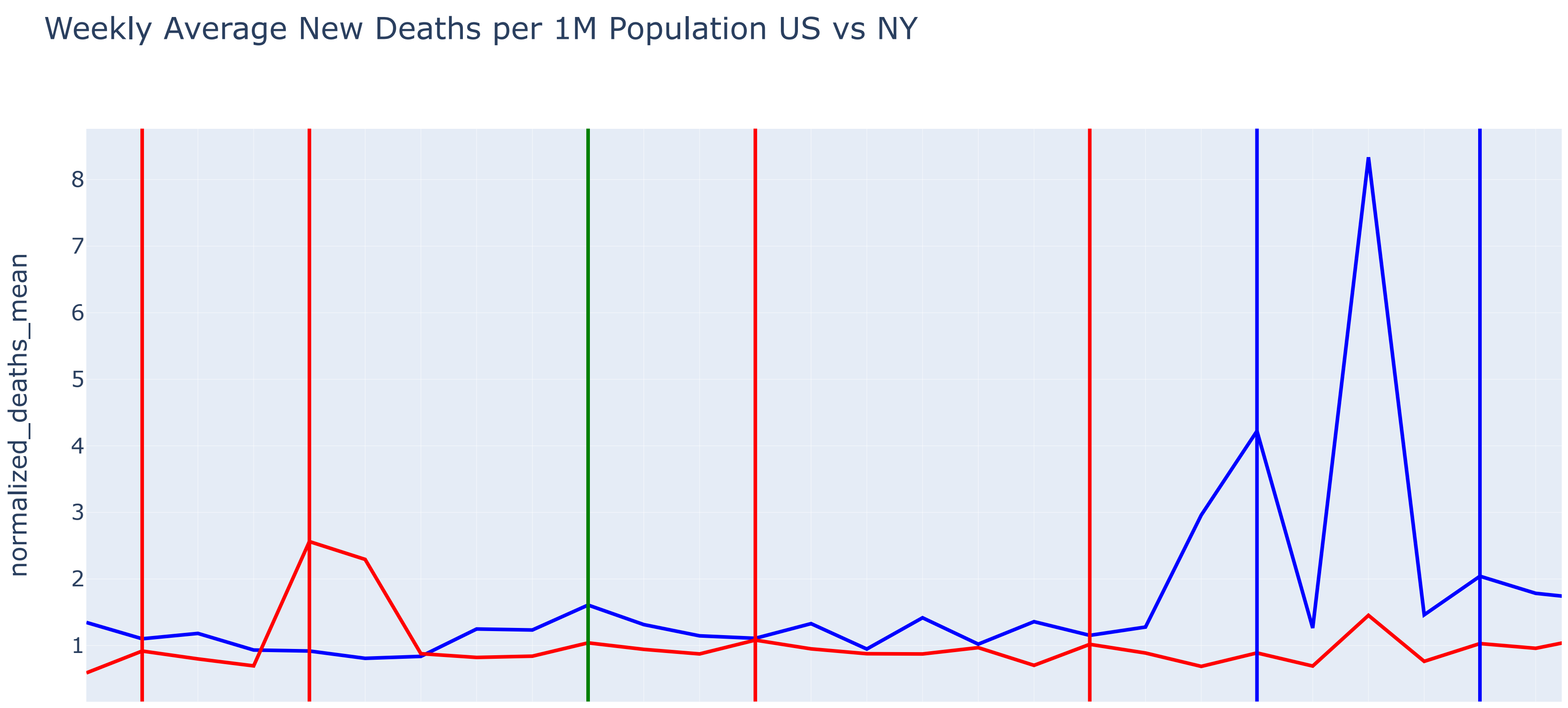
In [49]:

```
plot_state_vs_US_death_peaks('NC')
```


Weekly Average New Deaths per 1M Population US vs NC

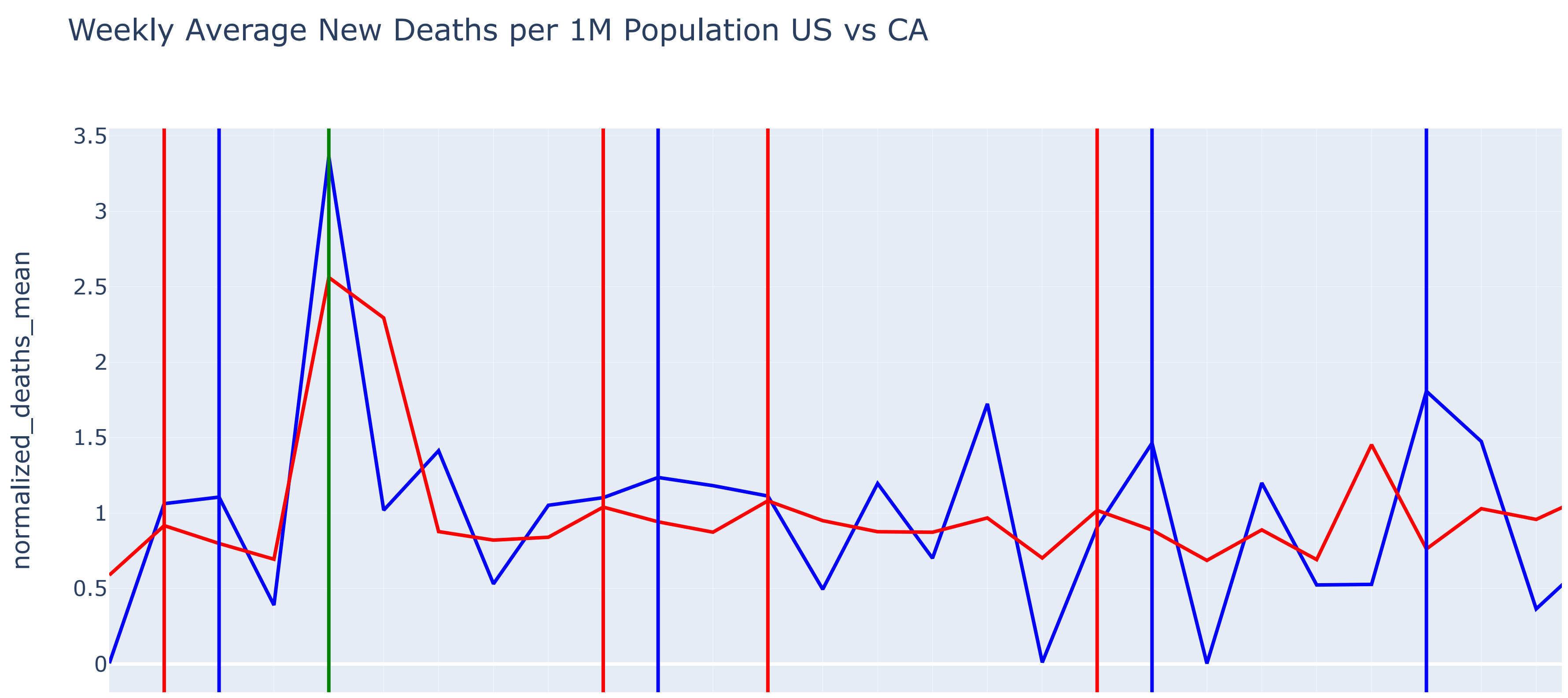
- In the above graph, the vertical blue lines identifies state peaks, red lines identifies US peaks and the green lines identify peaks that are matching for both US and the state
- In the above graph we can see that the new deaths weekly average has many matching peaks for NC and US throughout the year.
- The highest peak for NC is in Week Aug29 to Sep04 but there is decrease in cases for US during that week
- The US highest peak is the week of Jun27 to Jul03 and there is also increase in cases for NC during that week
- The overall pattern for US and the state doesn't match. State has more number of large peaks when compared to the country. The countries peaks might be subsided due to the average values of all states

In [50]:
```
plot_state_vs_US_death_peaks('NY')
```



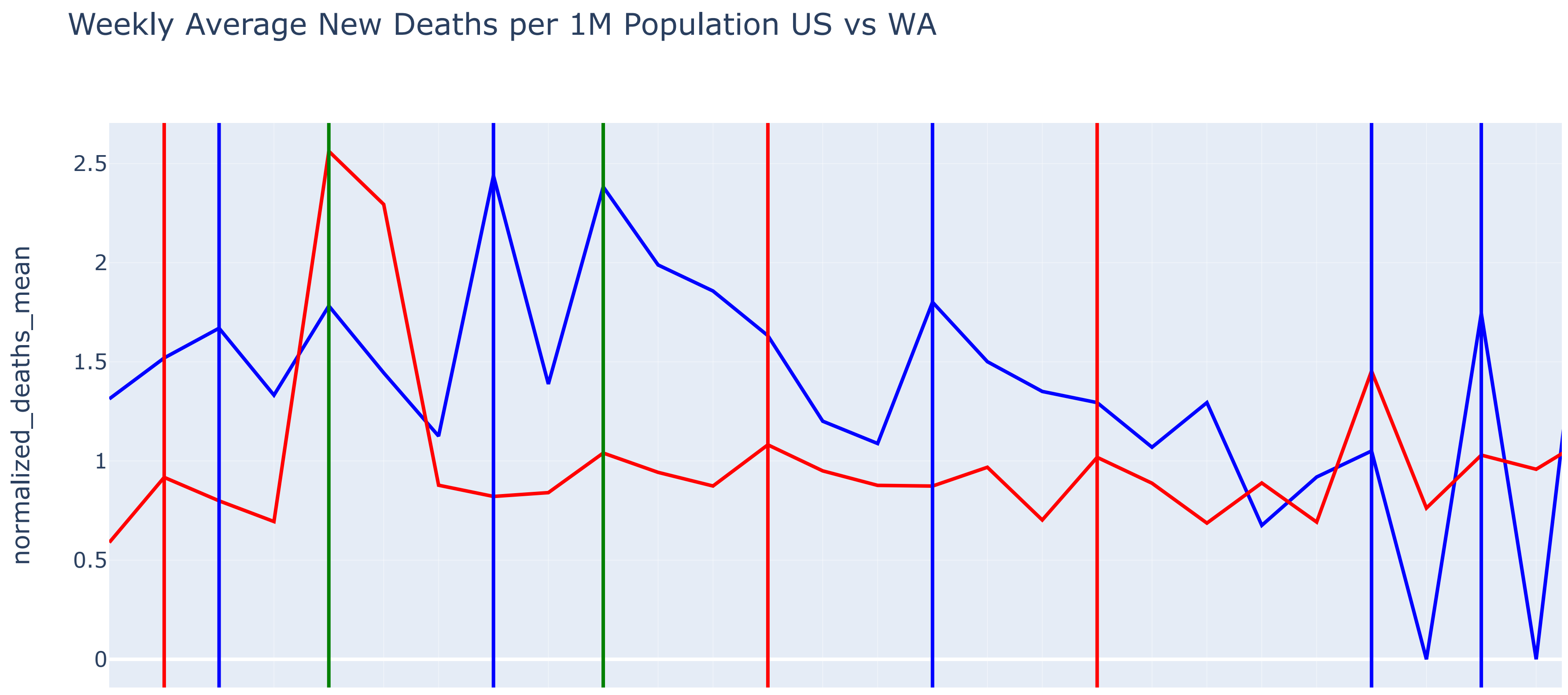
- In the above graph, the vertical blue lines identifies state peaks, red lines identifies US peaks and the green lines identify peaks that are matching for both US and the state
- In the above graph we can see that the new deaths weekly average has one matching peaks for NY and US for the week Aug01 to Aug07.
- The highest peak for NY is in Week Nov07 to Nov13 and there is also increase in cases for US during that week. This peak is not identified by the model since the distance between the adjacent peaks is set to 1.
- The US highest peak is the week of Jun27 to Jul03 and there was decrease in cases for NY during that week
- The overall pattern for US and the state matches slightly with steady number of deaths for most part of the time period and raise in deaths during November.

In [51]:
```
plot_state_vs_US_death_peaks('CA')
```



- In the above graph, the vertical blue lines identifies state peaks, red lines identifies US peaks and the green lines identify peaks that are matching for both US and the state
- In the above graph we can see that the new deaths weekly average has one matching peaks for CA and US for the week Jun27 to Jul03. Which is highest peak for both US and CA
- The overall pattern for US and the state matches approximately with high number of deaths during the mid of the year and then steady number of deaths for most part of the time period.
- In the end during December cases decreased for CA where as it increase for US

```
In [52]: plot_state_vs_US_death_peaks('WA')
```



- In the above graph, the vertical blue lines identifies state peaks, red lines identifies US peaks and the green lines identify peaks that are matching for both US and the state
- In the above graph we can see that the new deaths weekly average has two matching peaks for WA and US for the week Jun27 to Jul03 and Aug01 to Aug07. Which is highest peak each of them separately
- The overall pattern for US and the state matches approximately with high number of deaths during the mid of the year and then steady number of deaths for most part of the time period till November
- In the end of December cases decreased for WA where as it increase for US

3. Identify 3 counties within a state of your choice with high cases and death rates.

In [53]:

```
#Identifying 3 counties within CA state of high cases and death rates
CA_covid = transformed_df.query("State=='CA'").reset_index().drop(columns=['index','State','StateFIPS'])
CA_covid['case_rate'] = CA_covid['New_Cases']/CA_covid['population']
CA_covid['death_rate'] = CA_covid['New_Deaths']/CA_covid['population']
CA_covid.head()
```

Out[53]:

	Date	Week	countyFIPS	County_Name	population	Cases	New_Cases	Deaths	New_Deaths	case_rate	death_rate
0	2022-06-01	22	6001	Alameda County	1671329	285709	658	1870	0	0.000394	0.0
1	2022-06-01	22	6003	Alpine County	1129	128	0	0	0	0.000000	0.0
2	2022-06-01	22	6005	Amador County	39752	8820	3	87	0	0.000075	0.0
3	2022-06-01	22	6007	Butte County	219186	34122	17	427	0	0.000078	0.0
4	2022-06-01	22	6009	Calaveras County	45905	7522	8	121	0	0.000174	0.0

In [54]:

```
# Top 3 Counties with High Case Rate
CA_covid_top_3_case_rate_county = CA_covid.groupby(['County_Name','population','countyFIPS']).agg({'New_Cases': sum,'New_Deaths'
CA_covid_top_3_case_rate_county
```

Out[54]:

	County_Name	population	countyFIPS	New_Cases	New_Deaths	case_rate	death_rate
13	Imperial County	181215	6025	12444	32	0.068670	0.000177
16	Kings County	152940	6031	9222	25	0.060298	0.000163
19	Los Angeles County	10039107	6037	604454	2715	0.060210	0.000270

In [55]:

```
# Top 3 Counties with High Death Rate
CA_covid_top_3_death_rate_county = CA_covid.groupby(['County_Name','population','countyFIPS']).agg({'New_Cases': sum,'New_Deaths
CA_covid_top_3_death_rate_county
```

Out[55]:

	County_Name	population	countyFIPS	New_Cases	New_Deaths	case_rate	death_rate
56	Yolo County	220500	6113	9692	175	0.043955	0.000794
11	Glenn County	28393	6021	712	21	0.025077	0.000740
54	Tuolumne County	54478	6109	3236	27	0.059400	0.000496

4. Plot weekly trends (new cases and deaths) for the top 3 infected counties. Show plots by raw values and log normalized values. Describe what is causing them and what were the peaks. Do the counties follow state pattern.

In [56]:

```
top_3_case_rate_counties = CA_covid_top_3_case_rate_county['County_Name'].to_list()
top_3_death_rate_counties = CA_covid_top_3_death_rate_county['County_Name'].to_list()
```

In [57]:

```
CA_covid["Week_Dates"] = CA_covid['Week'].apply(get_week_range_string)
CA_covid.head()
```

Out[57]:

	Date	Week	countyFIPS	County_Name	population	Cases	New_Cases	Deaths	New_Deaths	case_rate	death_rate	Week_Dates
0	2022-06-01	22	6001	Alameda County	1671329	285709	658	1870	0	0.000394	0.0	Jun-01 to Jun-05
1	2022-06-01	22	6003	Alpine County	1129	128	0	0	0	0.000000	0.0	Jun-01 to Jun-05
2	2022-06-01	22	6005	Amador County	39752	8820	3	87	0	0.000075	0.0	Jun-01 to Jun-05
3	2022-06-01	22	6007	Butte County	219186	34122	17	427	0	0.000078	0.0	Jun-01 to Jun-05
4	2022-06-01	22	6009	Calaveras County	45905	7522	8	121	0	0.000174	0.0	Jun-01 to Jun-05

In [58]:

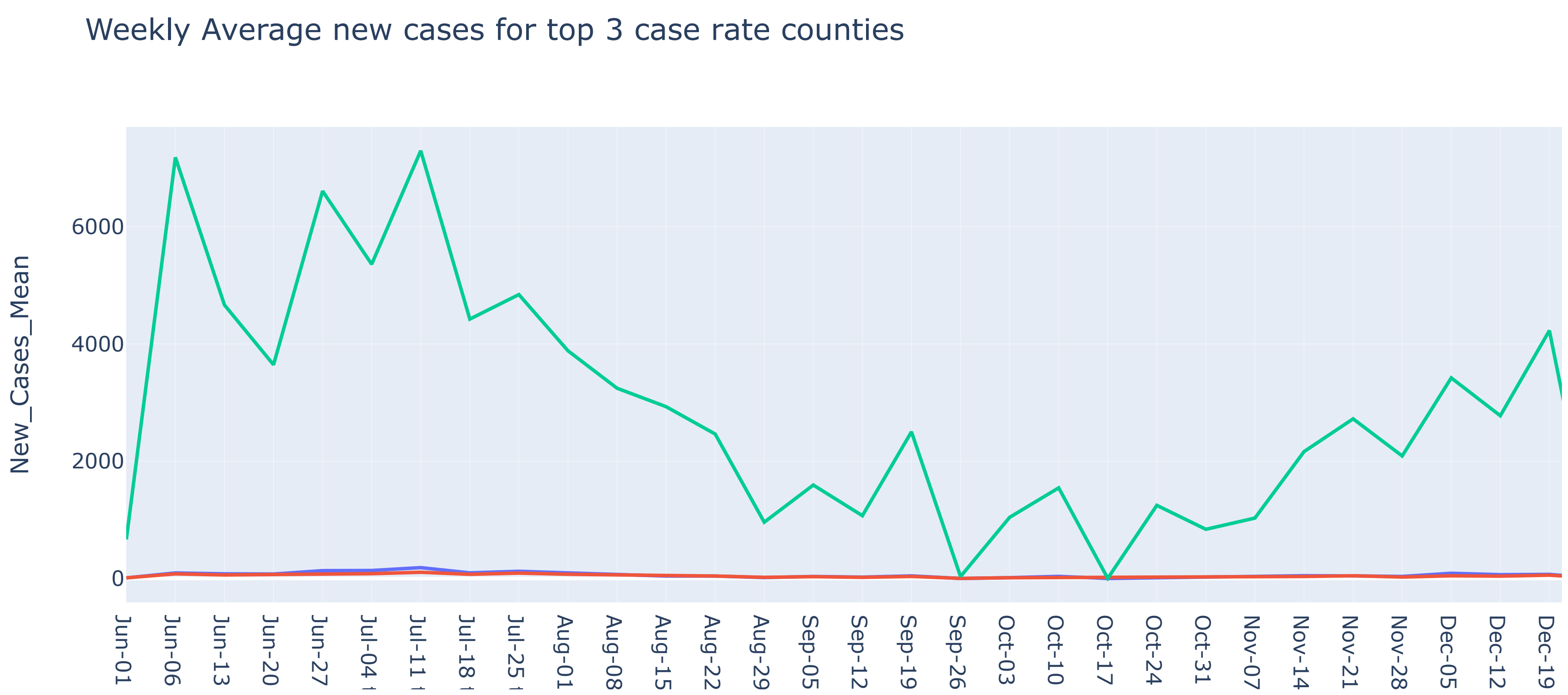
```
CA_top3_case_rate_counties_weekly_mean = CA_covid.query(f"County_Name in {top_3_case_rate_counties}").groupby(by=['Week', 'Week_I
CA_top3_case_rate_counties_weekly_mean = CA_top3_case_rate_counties_weekly_mean.rename(columns={"New_Cases":"New_Cases_Mean"})
CA_top3_case_rate_counties_weekly_mean.head()
```

Out[58]:

	Week	Week_Dates	County_Name	New_Cases_Mean
0	22	Jun-01 to Jun-05	Imperial County	11.800000
1	22	Jun-01 to Jun-05	Kings County	11.800000
2	22	Jun-01 to Jun-05	Los Angeles County	666.600000
3	23	Jun-06 to Jun-12	Imperial County	95.714286
4	23	Jun-06 to Jun-12	Kings County	80.000000

In [59]:

```
px.line(CA_top3_case_rate_counties_weekly_mean, x='Week_Dates', y='New_Cases_Mean', color='County_Name', title='Weekly Average ne
```



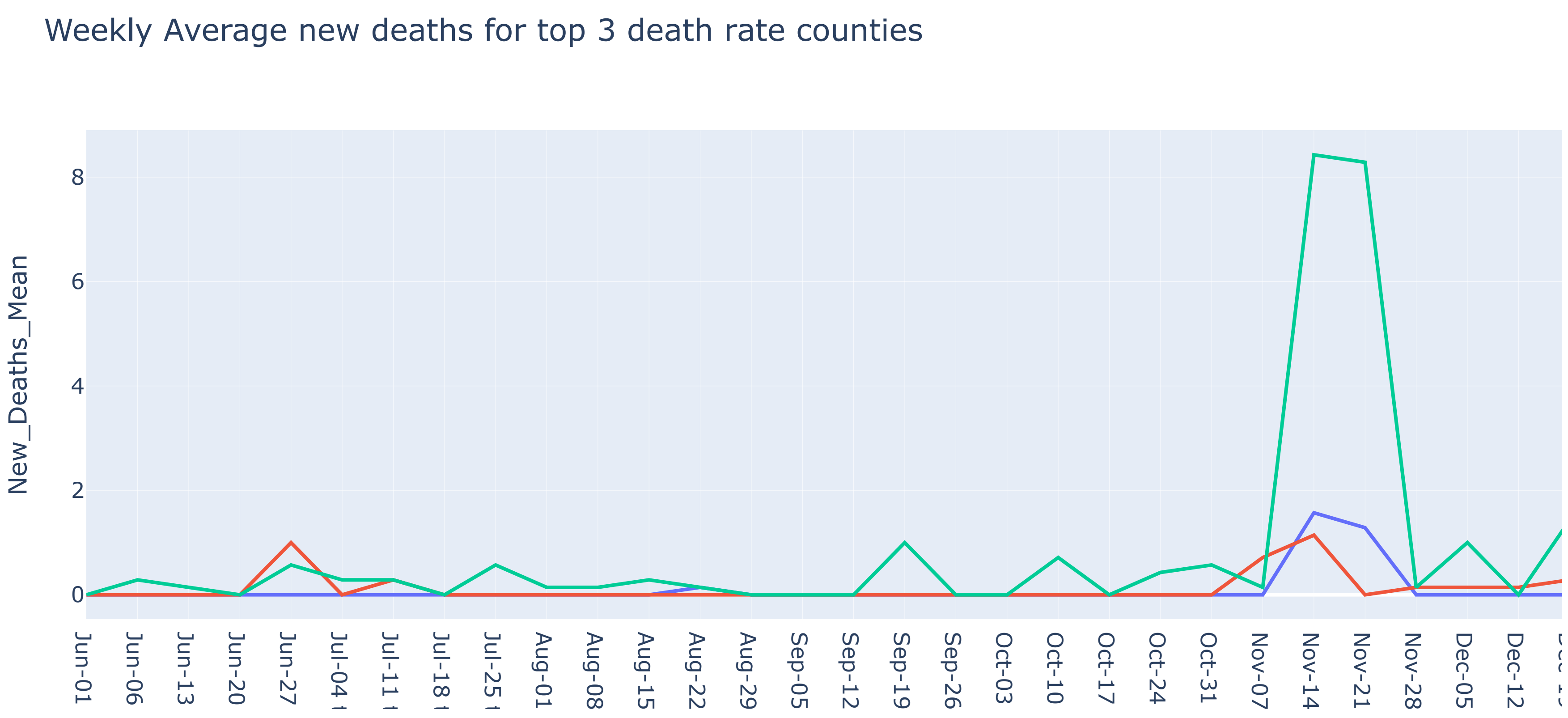
In the above plot we cannot identify any correlation between both states since the numbers are varying by large number due to population difference. Let us see if we can find any insights in normalized plot

```
In [60]: CA_top3_death_rate_counties_weekly_mean = CA_covid.query(f"County_Name in {top_3_death_rate_counties}").groupby(by=['Week','Week_
         CA_top3_death_rate_counties_weekly_mean = CA_top3_death_rate_counties_weekly_mean.rename(columns={"New_Deaths":"New_Deaths_Mean"
         CA_top3_death_rate_counties_weekly_mean.head()
```

Out[60]:

	Week	Week_Dates	County_Name	New_Deaths_Mean
0	22	Jun-01 to Jun-05	Glenn County	0.0
1	22	Jun-01 to Jun-05	Tuolumne County	0.0
2	22	Jun-01 to Jun-05	Yolo County	0.0
3	23	Jun-06 to Jun-12	Glenn County	0.0
4	23	Jun-06 to Jun-12	Tuolumne County	0.0

```
In [61]: px.line(CA_top3_death_rate_counties_weekly_mean, x='Week_Dates', y='New_Deaths_Mean', color='County_Name', title='Weekly Average
```



In the above plot we can see for the most part the deaths were zero where as the deaths raised for all the counties during the month of November. Let us see if we can find any insights in the normalized plot

In [62]:

```
# Plotting values per 1M population
top_3_case_rate_county_population = {county:CA_covid.query(f"County_Name=='{county}'")['population'].unique()[0] for county in to
top_3_death_rate_county_population = {county:CA_covid.query(f"County_Name=='{county}'")['population'].unique()[0] for county in t
CA_top3_case_rate_counties_weekly_mean['New_Cases_Mean_Per_1M'] = CA_top3_case_rate_counties_weekly_mean.apply(lambda x: 1000000
CA_top3_death_rate_counties_weekly_mean['New_Deaths_Mean_Per_1M'] = CA_top3_death_rate_counties_weekly_mean.apply(lambda x: 10000
```

In [63]:

```
CA_top3_case_rate_counties_weekly_mean.head()
```

Out[63]:

	Week	Week_Dates	County_Name	New_Cases_Mean	New_Cases_Mean_Per_1M
0	22	Jun-01 to Jun-05	Imperial County	11.800000	65.116022
1	22	Jun-01 to Jun-05	Kings County	11.800000	77.154440
2	22	Jun-01 to Jun-05	Los Angeles County	666.600000	66.400328
3	23	Jun-06 to Jun-12	Imperial County	95.714286	528.180811
4	23	Jun-06 to Jun-12	Kings County	80.000000	523.080947

In [64]:

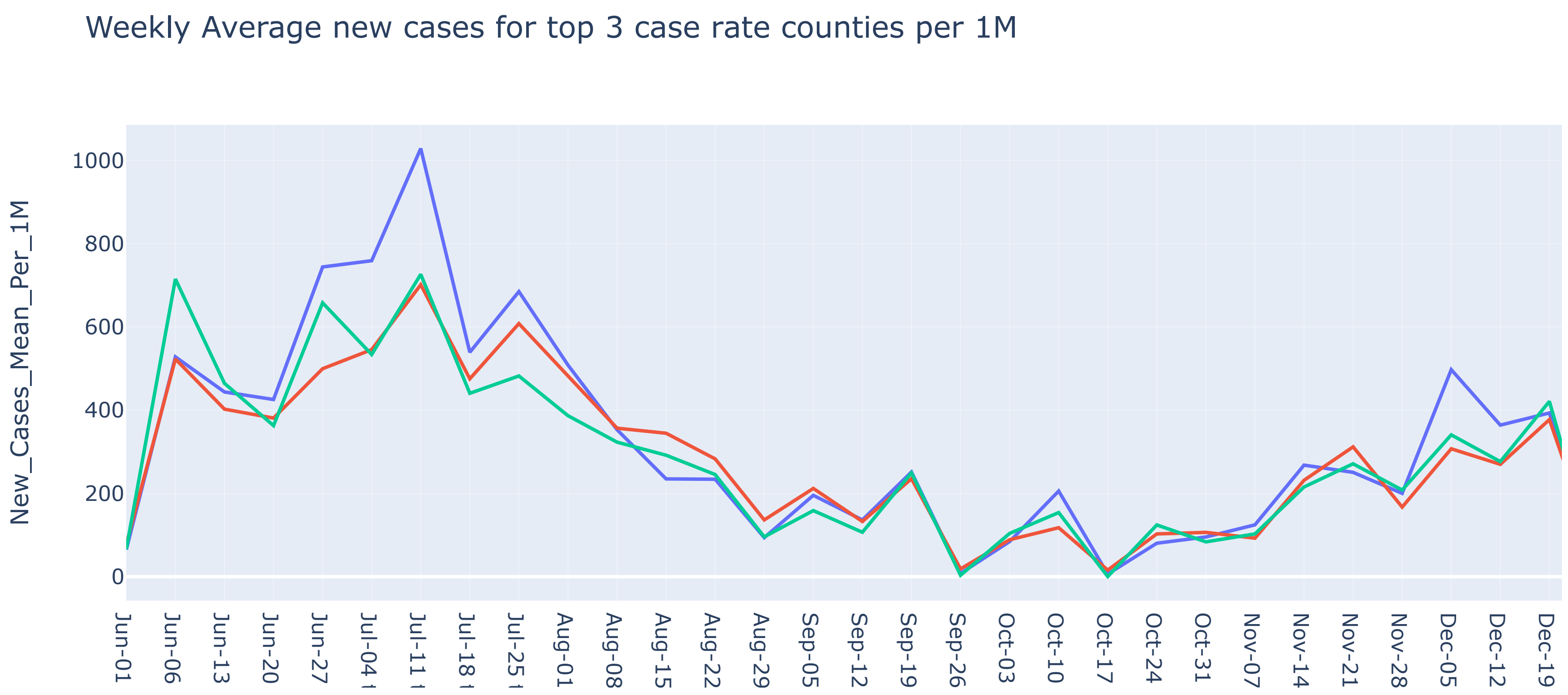
```
CA_top3_death_rate_counties_weekly_mean.head()
```

Out[64]:

	Week	Week_Dates	County_Name	New_Deaths_Mean	New_Deaths_Mean_Per_1M
0	22	Jun-01 to Jun-05	Glenn County	0.0	0.0
1	22	Jun-01 to Jun-05	Tuolumne County	0.0	0.0
2	22	Jun-01 to Jun-05	Yolo County	0.0	0.0
3	23	Jun-06 to Jun-12	Glenn County	0.0	0.0
4	23	Jun-06 to Jun-12	Tuolumne County	0.0	0.0

In [65]:

```
px.line(CA_top3_case_rate_counties_weekly_mean, x='Week_Dates', y='New_Cases_Mean_Per_1M', color='County_Name', title='Weekly Ave
```

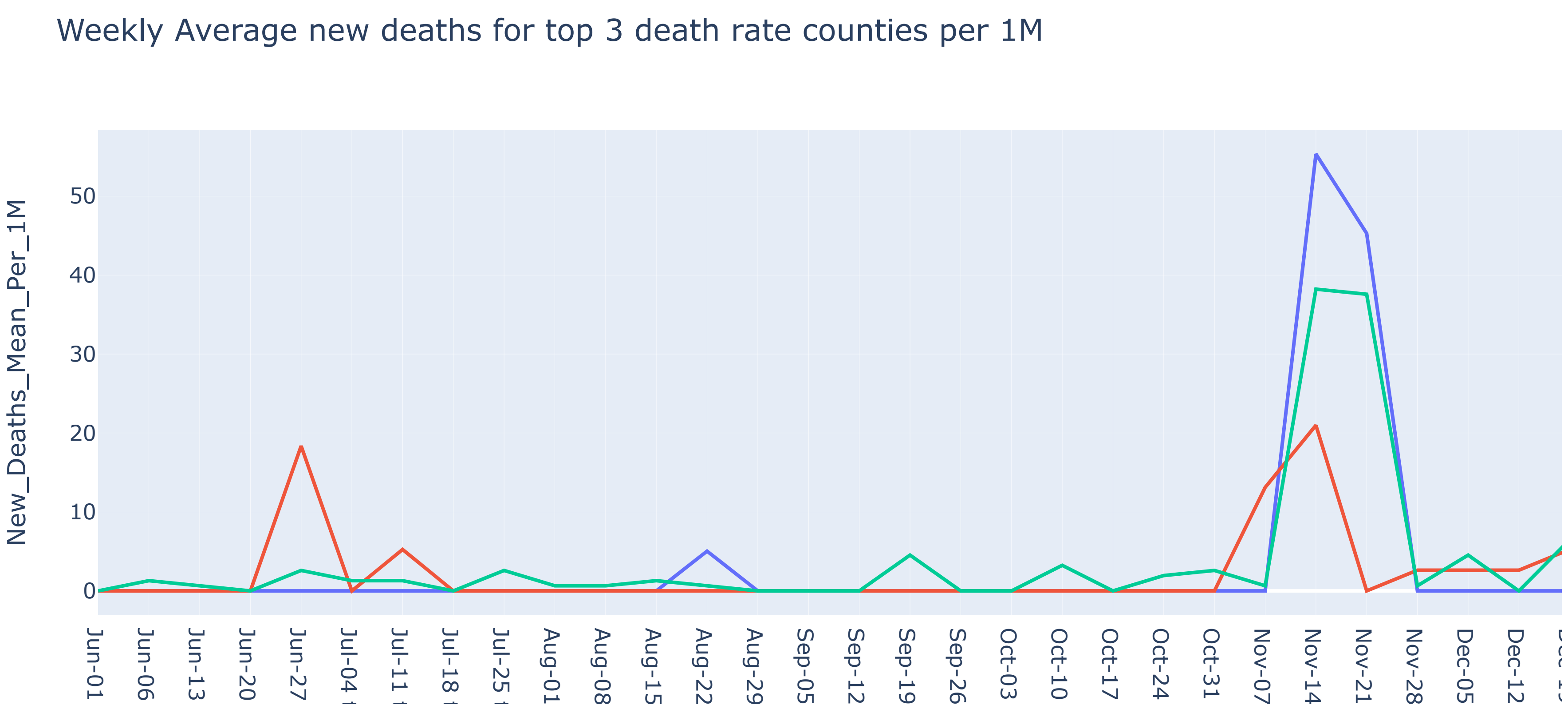


In the normalized plot for 1M population we can see that the weekly average of new case patterns matched exactly. All the counties are close by so the trends could be similar.

The new Cases were high during July and decreaased until November and then the cases increased by the end of the year. In Dec end, the new cases dropped to 0

In [66]:

```
px.line(CA_top3_death_rate_counties_weekly_mean, x='Week_Dates', y='New_Deaths_Mean_Per_1M', color='County_Name', title='Weekly A
```



In the normalized plot for 1M population we can see that the weekly average of new case patterns resemble the same as the raw values plot. The Weekly average of new Deaths were low initially and increased in November and then the cases decreased by the end of the year. In Dec end, the new cases dropped to 0

```
In [67]: # Plotting Log Normal values
         CA_top3_case_rate_counties_weekly_mean['New_Cases_Mean_log_normal'] = np.log(CA_top3_case_rate_counties_weekly_mean['New_Cases_Me
         # Adding +1 for deaths values to prevent divide by 0 errors in log transformation
         CA_top3_death_rate_counties_weekly_mean['New_Deaths_Mean_log_normal'] = np.log(CA_top3_death_rate_counties_weekly_mean['New_Deatl
```

```
In [68]: CA_top3_case_rate_counties_weekly_mean.head()
```

Out[68]:

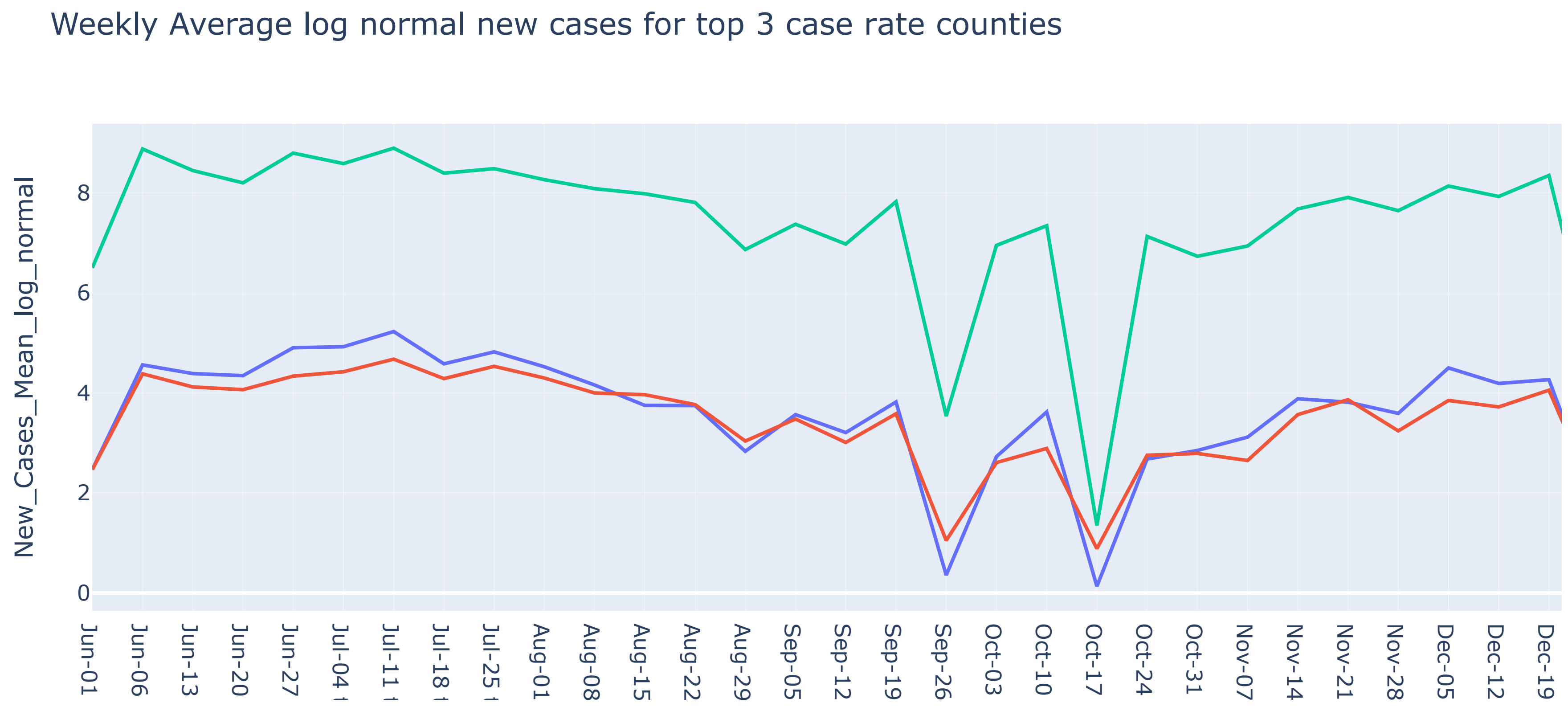
	Week	Week_Dates	County_Name	New_Cases_Mean	New_Cases_Mean_Per_1M	New_Cases_Mean_log_normal
0	22	Jun-01 to Jun-05	Imperial County	11.800000	65.116022	2.468100
1	22	Jun-01 to Jun-05	Kings County	11.800000	77.154440	2.468100
2	22	Jun-01 to Jun-05	Los Angeles County	666.600000	66.400328	6.502190
3	23	Jun-06 to Jun-12	Imperial County	95.714286	528.180811	4.561368
4	23	Jun-06 to Jun-12	Kings County	80.000000	523.080947	4.382027

```
In [69]: CA_top3_death_rate_counties_weekly_mean.head()
```

Out[69]:

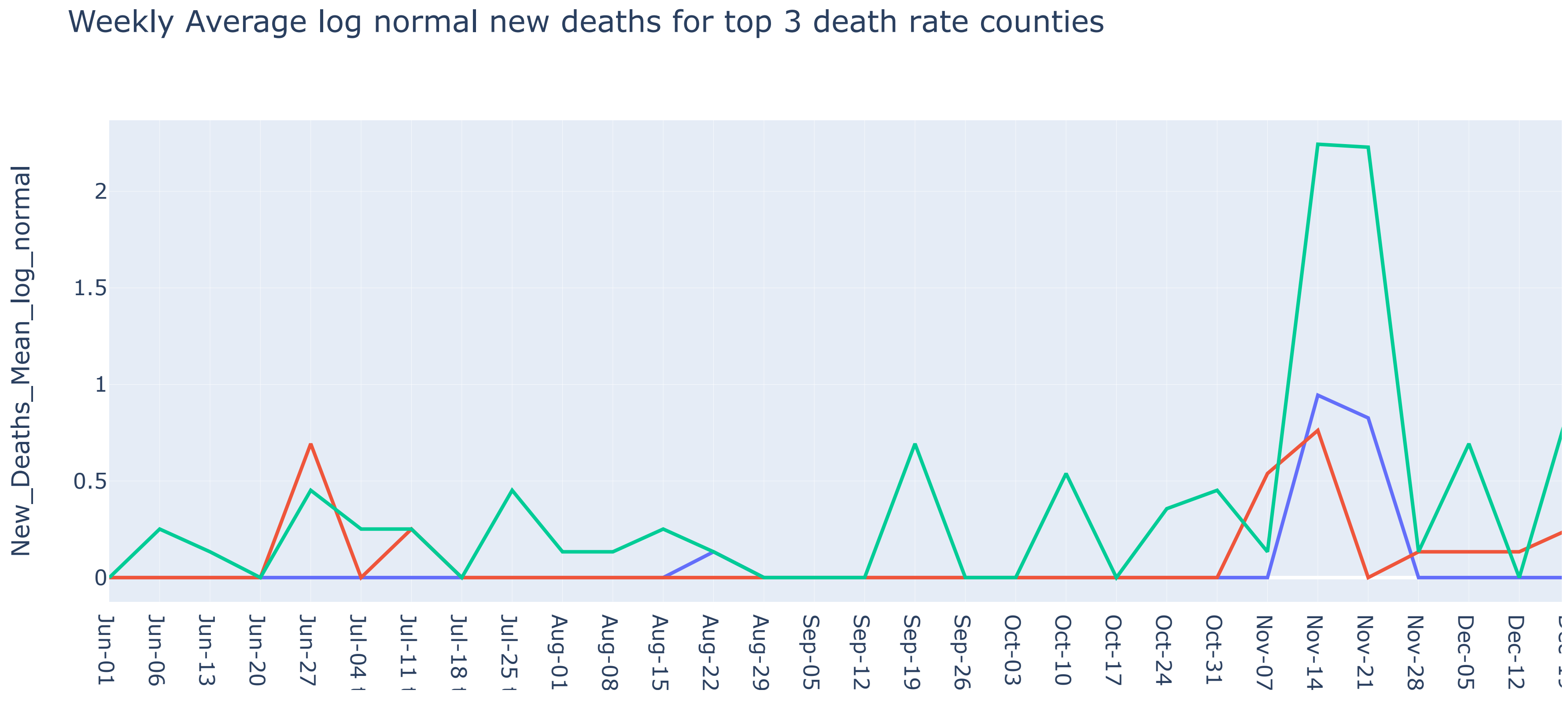
	Week	Week_Dates	County_Name	New_Deaths_Mean	New_Deaths_Mean_Per_1M	New_Deaths_Mean_log_normal
0	22	Jun-01 to Jun-05	Glenn County	0.0	0.0	0.0
1	22	Jun-01 to Jun-05	Tuolumne County	0.0	0.0	0.0
2	22	Jun-01 to Jun-05	Yolo County	0.0	0.0	0.0
3	23	Jun-06 to Jun-12	Glenn County	0.0	0.0	0.0
4	23	Jun-06 to Jun-12	Tuolumne County	0.0	0.0	0.0

```
In [70]: px.line(CA_top3_case_rate_counties_weekly_mean, x='Week_Dates', y='New_Cases_Mean_log_normal', color='County_Name', title='Weekly
```


Weekly Average log normal new cases for top 3 case rate counties

Just like normalized data for 1M population, this log normalized plot also shows clear matching trend among all the three counties with high case rate.

```
In [71]: px.line(CA_top3_death_rate_counties_weekly_mean, x='Week_Dates', y='New_Deaths_Mean_log_normal', color='County_Name', title='Weel
```


Weekly Average log normal new deaths for top 3 death rate counties

Just like normalized data for 1M population, this log normalized plot also shows similar trend among all the three counties with high death average during November. Yolo county has more deaths when compared to the other two counties.

The interesting factor is The counties with high death rate are close together and the counties with high case rate are close together.

High population density might be the reason for high case rate in those three counties (Imperial, Kings, Los Angeles) whereas poor medical facilities in the other three counties might be the reason for high death rate in the counties (Glenn, Tuolumne, Yolo)

```
In [72]: #Identifying Peaks
         print(f"Top 3 Case rate counties: {top_3_case_rate_counties}")
         print(f"Top 3 Death rate counties: {top_3_death_rate_counties}")
```

```
Top 3 Case rate counties: ['Imperial County ', 'Kings County ', 'Los Angeles County ']
Top 3 Death rate counties: ['Yolo County ', 'Glenn County ', 'Tuolumne County ']
```

```
In [73]: Imperial_County_df = CA_top3_case_rate_counties_weekly_mean.query("County_Name=='Imperial County '").reset_index().drop(columns=
         Imperial_County_Peaks_indices = find_peaks(Imperial_County_df['New_Cases_Mean'],width=1)[0]
```

```
Imperial_County_Case_Peaks = Imperial_County_df[Imperial_County_df.index.isin(Imperial_County_Peaks_indices)]
Imperial_County_Case_Peaks
```

Out[73]:

	Week	Week_Dates	County_Name	New_Cases_Mean	New_Cases_Mean_Per_1M	New_Cases_Mean_log_normal
6	28	Jul-11 to Jul-17	Imperial County	186.428571	1028.770088	5.228048
16	38	Sep-19 to Sep-25	Imperial County	45.571429	251.477133	3.819281
19	41	Oct-10 to Oct-16	Imperial County	37.285714	205.754018	3.618610
24	46	Nov-14 to Nov-20	Imperial County	48.571429	268.032053	3.883035
27	49	Dec-05 to Dec-11	Imperial County	90.142857	497.435958	4.501396

In [74]:

```
Kings_County_df = CA_top3_case_rate_counties_weekly_mean.query("County_Name=='Kings County '").reset_index().drop(columns='index
Kings_County_Peaks_indices = find_peaks(Kings_County_df['New_Cases_Mean'],width=1)[0]
Kings_County_Case_Peaks = Kings_County_df[Kings_County_df.index.isin(Kings_County_Peaks_indices)]
Kings_County_Case_Peaks
```

Out[74]:

	Week	Week_Dates	County_Name	New_Cases_Mean	New_Cases_Mean_Per_1M	New_Cases_Mean_log_normal
6	28	Jul-11 to Jul-17	Kings County	107.285714	701.488913	4.675496
8	30	Jul-25 to Jul-31	Kings County	93.000000	608.081601	4.532599
19	41	Oct-10 to Oct-16	Kings County	18.000000	117.693213	2.890372
22	44	Oct-31 to Nov-06	Kings County	16.285714	106.484336	2.790288
25	47	Nov-21 to Nov-27	Kings County	47.714286	311.980422	3.865231
29	51	Dec-19 to Dec-25	Kings County	57.714286	377.365540	4.055505

In [75]:

```
LosAngeles_County_df = CA_top3_case_rate_counties_weekly_mean.query("County_Name=='Los Angeles County '").reset_index().drop(col
LosAngeles_County_Peaks_indices = find_peaks(LosAngeles_County_df['New_Cases_Mean'],width=1)[0]
LosAngeles_County_Case_Peaks = LosAngeles_County_df[LosAngeles_County_df.index.isin(LosAngeles_County_Peaks_indices)]
LosAngeles_County_Case_Peaks
```

Out[75]:

	Week	Week_Dates	County_Name	New_Cases_Mean	New_Cases_Mean_Per_1M	New_Cases_Mean_log_normal
6	28	Jul-11 to Jul-17	Los Angeles County	7293.000000	726.459037	8.894670
19	41	Oct-10 to Oct-16	Los Angeles County	1546.285714	154.026221	7.343611
25	47	Nov-21 to Nov-27	Los Angeles County	2722.857143	271.225035	7.909437
29	51	Dec-19 to Dec-25	Los Angeles County	4228.000000	421.152997	8.349484

In [76]:

```
CA_normalized_case_data = us_and_states_merged_stats.query("State=='CA'")[['Week_Dates','State','normalized_cases_mean']].rename
CA_and_top_case_rate_counties_merged = pd.concat([CA_normalized_case_data,CA_top3_case_rate_counties_weekly_mean[['Week_Dates','
CA_and_top_case_rate_counties_merged.head()
```

Out[76]:

	Week_Dates	County_Name	New_Cases_Mean_Per_1M
0	Jun-01 to Jun-05	CA	67.245014
1	Jun-06 to Jun-12	CA	644.870974
2	Jun-13 to Jun-19	CA	404.475200
3	Jun-20 to Jun-26	CA	346.225593
4	Jun-27 to Jul-03	CA	547.761783

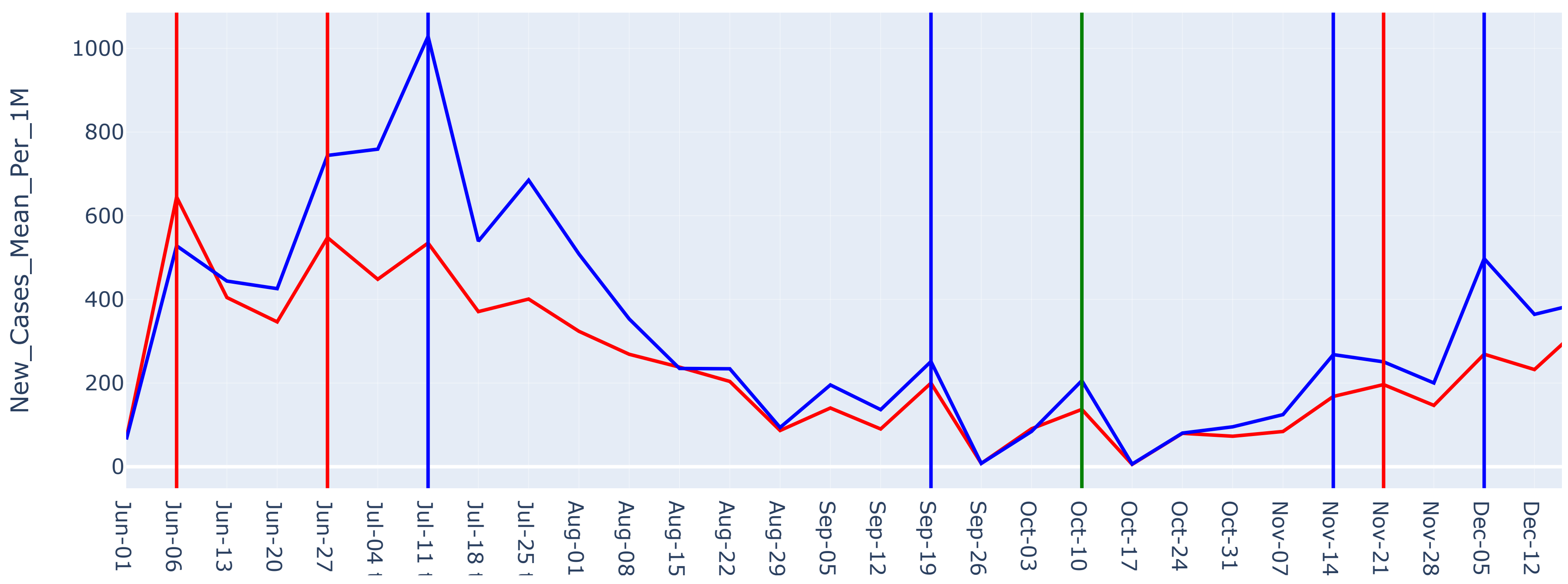
In [77]:

```
county_peaks = {'CA': CA_peaks['Week_Dates'].to_list(),
         'Imperial County ': Imperial_County_Case_Peaks['Week_Dates'].to_list(),
         'Kings County ': Kings_County_Case_Peaks['Week_Dates'].to_list(),
         'Los Angeles County ': LosAngeles_County_Case_Peaks['Week_Dates'].to_list()
        }
def plot_county_vs_CA_peaks(county):
    county_peaks_list = county_peaks[county]
    CA_peaks_list = county_peaks['CA']
    matching_peaks = list(set(county_peaks_list).intersection(set(CA_peaks_list)))
    unmatched_county_peaks = list(set(county_peaks_list).difference(set(CA_peaks_list)))
    unmatched_CA_peaks = list(set(CA_peaks_list).difference(set(county_peaks_list)))
    fig = px.line(CA_and_top_case_rate_counties_merged.query(f"County_Name in ['{county}','CA']"),
              x='Week_Dates',y='New_Cases_Mean_Per_1M',color='County_Name',
              title=f"Weekly Average New Cases per 1M Population CA vs {county}",
             color_discrete_map={
             "CA": "red",
             county: "blue"
         })
    for week in unmatched_county_peaks:
        fig.add_vline(x=week, line_color='blue')
    for week in unmatched_CA_peaks:
        fig.add_vline(x=week, line_color='red')
    for week in matching_peaks:
        fig.add_vline(x=week, line_color='green')
    return fig
```

In [78]:

```
plot_county_vs_CA_peaks('Imperial County ')
```

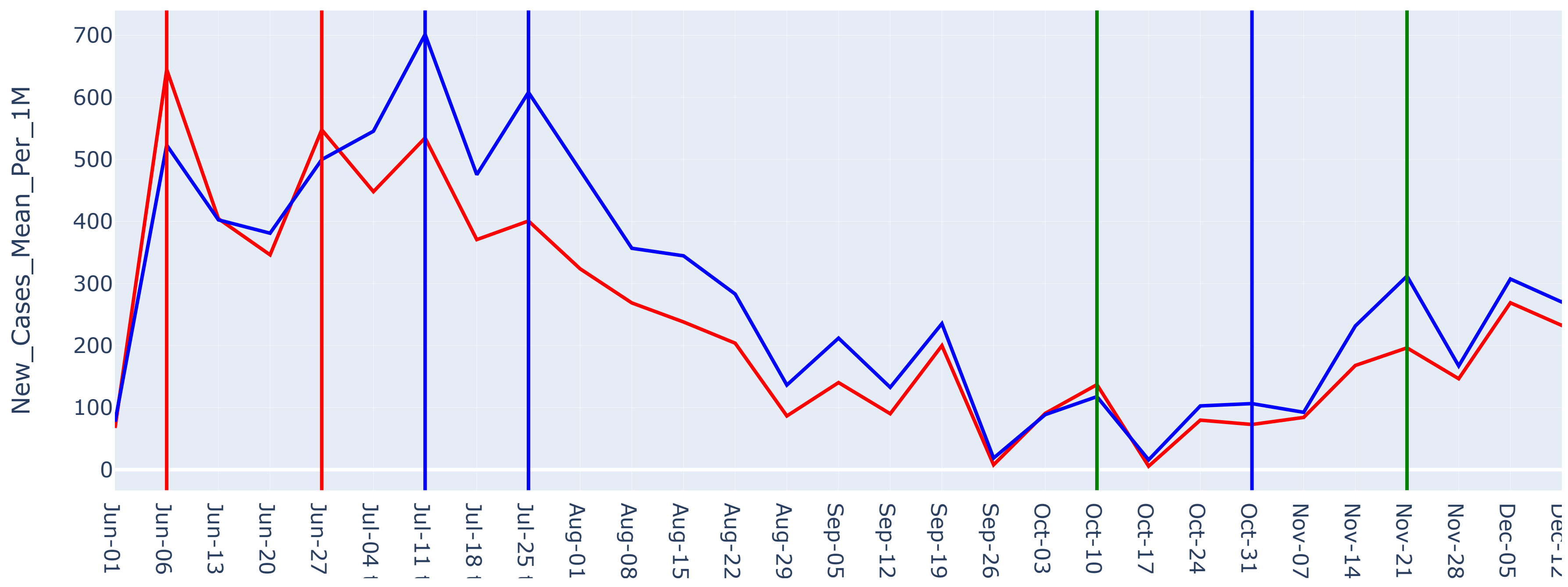
Weekly Average New Cases per 1M Population CA vs Imperial County



From the above plot we can see that the California weekly case average matches with Imperial County all the peaks of CA state has similar peaks for the Imperial county as well

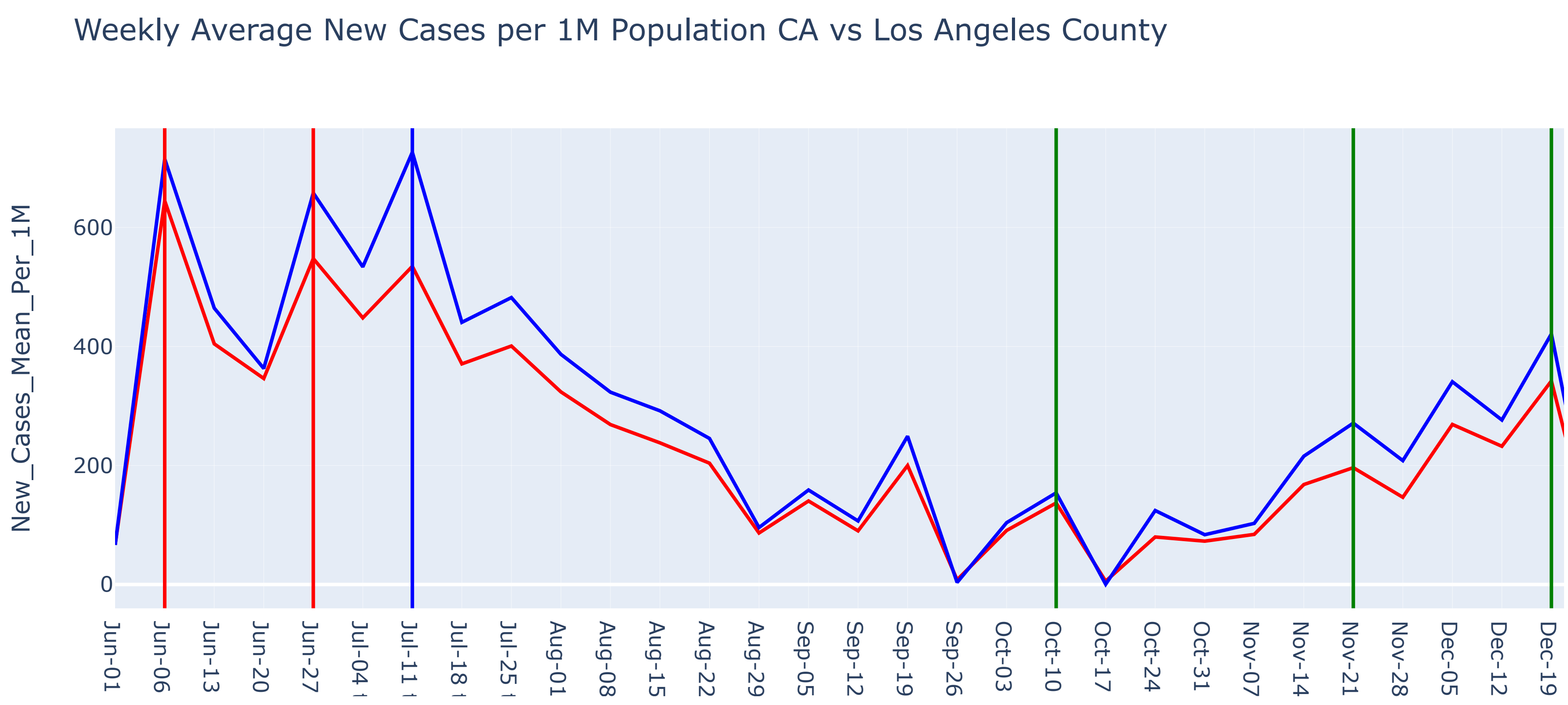
In [79]:
```
plot_county_vs_CA_peaks('Kings County ')
```

Weekly Average New Cases per 1M Population CA vs Kings County



From the above plot we can see that the California weekly case average matches with Imperial County all the peaks of CA state has similar peaks for the Kings county as well

In [80]:
```
plot_county_vs_CA_peaks('Los Angeles County ')
```


Weekly Average New Cases per 1M Population CA vs Los Angeles County

From the above plot we can see that the California weekly case average matches with Imperial County all the peaks of CA state has similar peaks for the Los Angeles county as well

In [81]:
```
Yolo_County_df = CA_top3_death_rate_counties_weekly_mean.query("County_Name=='Yolo County '").reset_index().drop(columns='index'
Yolo_County_Peaks_indices = find_peaks(Yolo_County_df['New_Deaths_Mean'],width=1)[0]
Yolo_County_Case_Peaks = Yolo_County_df[Yolo_County_df.index.isin(Yolo_County_Peaks_indices)]
Yolo_County_Case_Peaks
```

Out[81]:

	Week	Week_Dates	County_Name	New_Deaths_Mean	New_Deaths_Mean_Per_1M	New_Deaths_Mean_log_normal
1	23	Jun-06 to Jun-12	Yolo County	0.285714	1.295756	0.251314
4	26	Jun-27 to Jul-03	Yolo County	0.571429	2.591513	0.451985
8	30	Jul-25 to Jul-31	Yolo County	0.571429	2.591513	0.451985
11	33	Aug-15 to Aug-21	Yolo County	0.285714	1.295756	0.251314
16	38	Sep-19 to Sep-25	Yolo County	1.000000	4.535147	0.693147
19	41	Oct-10 to Oct-16	Yolo County	0.714286	3.239391	0.538997
22	44	Oct-31 to Nov-06	Yolo County	0.571429	2.591513	0.451985
24	46	Nov-14 to Nov-20	Yolo County	8.428571	38.224814	2.243745
29	51	Dec-19 to Dec-25	Yolo County	1.428571	6.478782	0.887303

In [82]:
```
Glenn_County_df = CA_top3_death_rate_counties_weekly_mean.query("County_Name=='Glenn County '").reset_index().drop(columns='inde
Glenn_County_Peaks_indices = find_peaks(Glenn_County_df['New_Deaths_Mean'],width=1)[0]
Glenn_County_Case_Peaks = Glenn_County_df[Glenn_County_df.index.isin(Glenn_County_Peaks_indices)]
Glenn_County_Case_Peaks
```

Out[82]:

	Week	Week_Dates	County_Name	New_Deaths_Mean	New_Deaths_Mean_Per_1M	New_Deaths_Mean_log_normal
12	34	Aug-22 to Aug-28	Glenn County	0.142857	5.031421	0.133531
24	46	Nov-14 to Nov-20	Glenn County	1.571429	55.345633	0.944462

In [83]:
```
Tuolumne_County_df = CA_top3_death_rate_counties_weekly_mean.query("County_Name=='Tuolumne County '").reset_index().drop(columns
Tuolumne_County_Peaks_indices = find_peaks(Tuolumne_County_df['New_Deaths_Mean'],width=1)[0]
Tuolumne_County_Case_Peaks = Tuolumne_County_df[Tuolumne_County_df.index.isin(Tuolumne_County_Peaks_indices)]
Tuolumne_County_Case_Peaks
```

Out[83]:

	Week	Week_Dates	County_Name	New_Deaths_Mean	New_Deaths_Mean_Per_1M	New_Deaths_Mean_log_normal
4	26	Jun-27 to Jul-03	Tuolumne County	1.000000	18.356034	0.693147
6	28	Jul-11 to Jul-17	Tuolumne County	0.285714	5.244581	0.251314
24	46	Nov-14 to Nov-20	Tuolumne County	1.142857	20.978324	0.762140
29	51	Dec-19 to Dec-25	Tuolumne County	0.285714	5.244581	0.251314

In [84]:
```
CA_normalized_death_data = us_and_states_merged_stats.query("State=='CA'")[['Week_Dates','State','normalized_deaths_mean']].rena
CA_and_top_death_rate_counties_merged = pd.concat([CA_normalized_death_data,CA_top3_death_rate_counties_weekly_mean[['Week_Dates
CA_and_top_death_rate_counties_merged.head()
```

Out[84]:

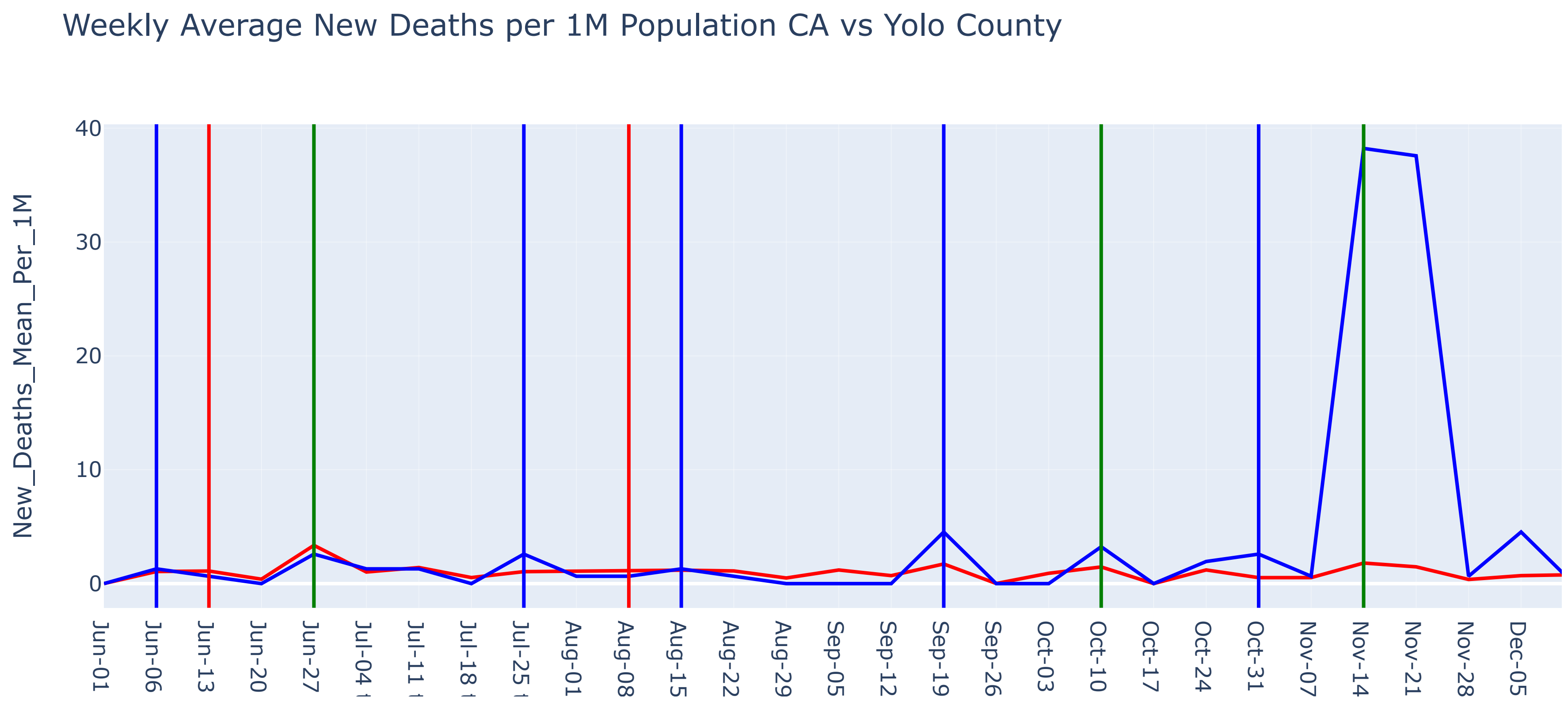
	Week_Dates	County_Name	New_Deaths_Mean_Per_1M
0	Jun-01 to Jun-05	CA	0.005062
1	Jun-06 to Jun-12	CA	1.062962
2	Jun-13 to Jun-19	CA	1.106348
3	Jun-20 to Jun-26	CA	0.390476
4	Jun-27 to Jul-03	CA	3.362431

In [85]:

```
county_death_peaks = {'CA': CA_death_peaks['Week_Dates'].to_list(),
         'Yolo County ': Yolo_County_Case_Peaks['Week_Dates'].to_list(),
         'Glenn County ': Glenn_County_Case_Peaks['Week_Dates'].to_list(),
         'Tuolumne County ': Tuolumne_County_Case_Peaks['Week_Dates'].to_list()
        }
def plot_county_vs_CA_death_peaks(county):
    county_peaks_list = county_death_peaks[county]
    CA_peaks_list = county_death_peaks['CA']
    matching_peaks = list(set(county_peaks_list).intersection(set(CA_peaks_list)))
    unmatched_county_peaks = list(set(county_peaks_list).difference(set(CA_peaks_list)))
    unmatched_CA_peaks = list(set(CA_peaks_list).difference(set(county_peaks_list)))
    fig = px.line(CA_and_top_death_rate_counties_merged.query(f"County_Name in ['{county}','CA']"),
                  x='Week_Dates',y='New_Deaths_Mean_Per_1M',color='County_Name',
                  title=f"Weekly Average New Deaths per 1M Population CA vs {county}",
                 color_discrete_map={
                 "CA": "red",
                 county: "blue"
            })
    for week in unmatched_county_peaks:
        fig.add_vline(x=week, line_color='blue')
    for week in unmatched_CA_peaks:
        fig.add_vline(x=week, line_color='red')
    for week in matching_peaks:
        fig.add_vline(x=week, line_color='green')
    return fig
```

In [86]:

```
plot_county_vs_CA_death_peaks('Yolo County ')
```

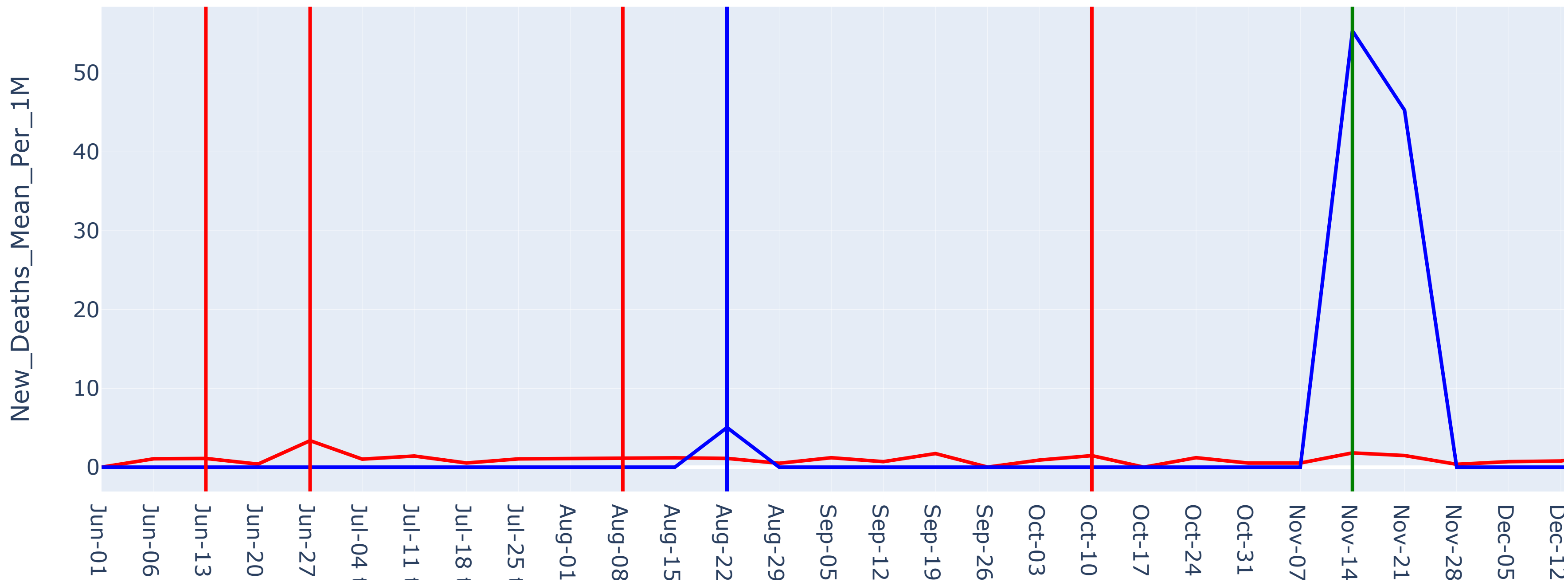


The pattern of New Average Deaths of Yolo County matches with CA except that the in the month of november Yolo county has large peak of Average Deaths

In [87]:

```
plot_county_vs_CA_death_peaks('Glenn County ')
```

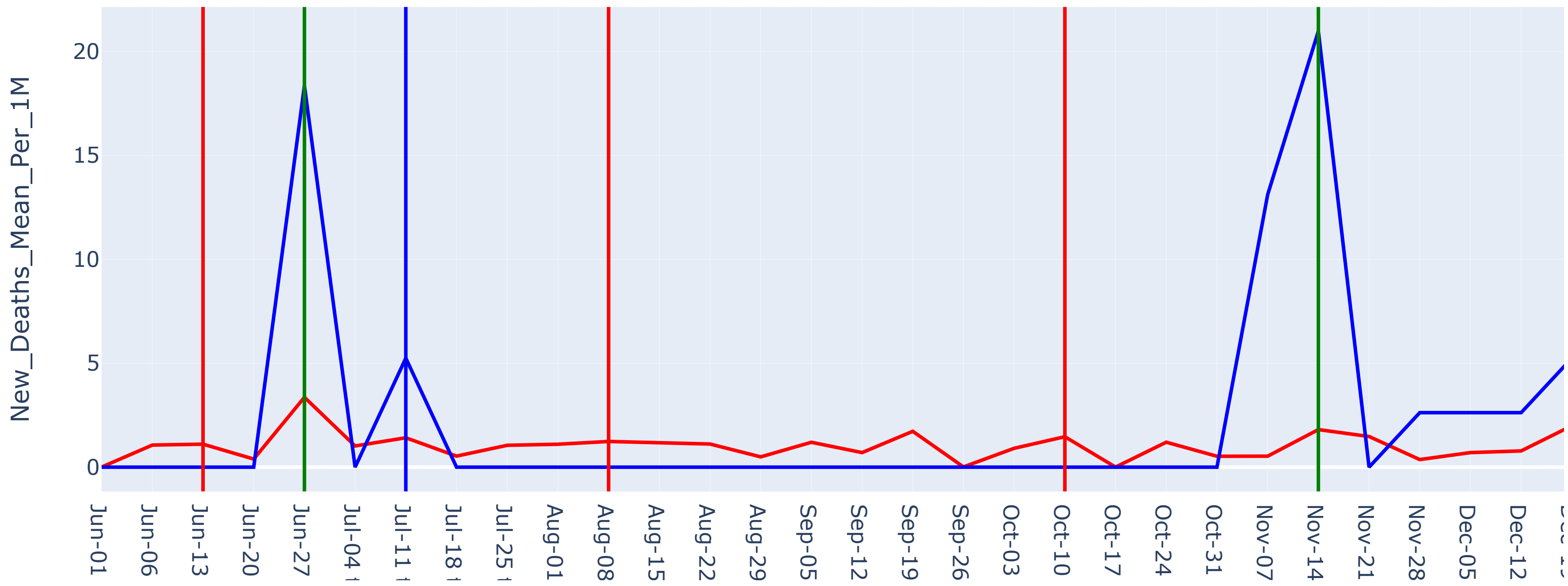
Weekly Average New Deaths per 1M Population CA vs Glenn County



The pattern of New Average Deaths of Glenn County didn't with CA being close to zero for most of the time period. But in the month of november Glenn county has large peak of Average Deaths.

In [88]:
```
plot_county_vs_CA_death_peaks('Tuolumne County ')
```

Weekly Average New Deaths per 1M Population CA vs Tuolumne County



The pattern of New Average Deaths of Tuolumne County didn't with CA being close to zero for most of the time period. But the in the month of november Glenn county has large peak of Average Deaths