Data Science Project Stage 2

Pratik Devkota

CSC 405/605

Generate Weekly Statistics

To generate the weekly statistics, we need to aggregate the dataframe by weeks. To do that I used the date column to convert it into datetime and then extracted the year and week number for each entry. Now using the new year and week values, I grouped the data and calculated sum, mean, median and mode across the US.

Compare data against other states

To compare the data against 4 other states, I defined a function that does the same thing as described above. Then I called the function for 5 defined states to give 5 different dataframe of aggregated values. I, then, merged the dataframes to one and displayed the data.

Identify counties with high cases and death rates

To find the counties with high cases and death rates, I took the difference in the number of new cases between the last two weeks. Since the rate is given by the change in values per unit time, I calculated the difference and sorted them to find the top 5 counties with the highest difference.

Plot daily trends of top 5 infected counties

To quickly get the top 5 infected counties, we can use the normalized dataframe and get the sum across each county. The counties with the largest sum are the top infected counties.

Plot the distributions

The plots are statically presented in the notebook. The descriptions for the remaining part are described within each section of the notebook.