The goal of Stage II is to develop formal hypothesis tests for the intuitions you had in Stage I and utilize statistical modeling to prove/disprove them.

Team: Chandan, Chandra Shekhar, Akasha, Sytiva

The first goal of this task is to find the average weekly new cases and deaths across the United States. To achieve this task the dataset is considered from stage-1. Since the merged dataset in stage-1 is not converted into long format it is converted here using the melt function. The dataset is converted to long format to easily perform operations on cases and deaths based on date. Since the data is separated based on the County. The number of new cases and deaths per day are calculated county wise. By using an aggregation function we calculated the mean, median and mode of new cases per week in the United States. The results are mean cases per new week was 20.48, the median was 2, and the mode was 0. The mean deaths was 0.3, median was 0, and mode was 0.

The second goal is to choose the 5 countries whose population is similar to the United States. The United States is the 3rd most populated country in the world and it is 1 billion less than the second most populated country, India. So, we considered the countries that are closely populated as the United States(**331,002,651**). We chose 5 countries Indonesia(**273,523,615**), Pakistan(**220,892,340)**, Brazil(**212,559,417)**, Nigeria(**206,139,589)** and Bangladesh(**164,689,383)**.

The third goal is to plot the daily trends of the US in comparison to these countries. The data is taken from the given single source for all the 6 countries to have the consistency. Since, the rate of cases or death varies from country to country despite showing the same figures considering the difference in their total population. So, we have calculated the number of new cases and deaths per million for each country on a day. The resulting data is plotted as in figure 1 and 2 respectively.

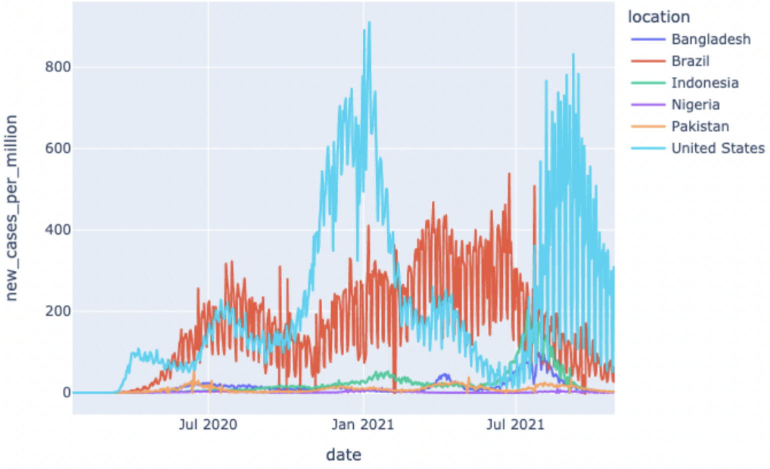


Figure-1. Number of new cases a day per million population.

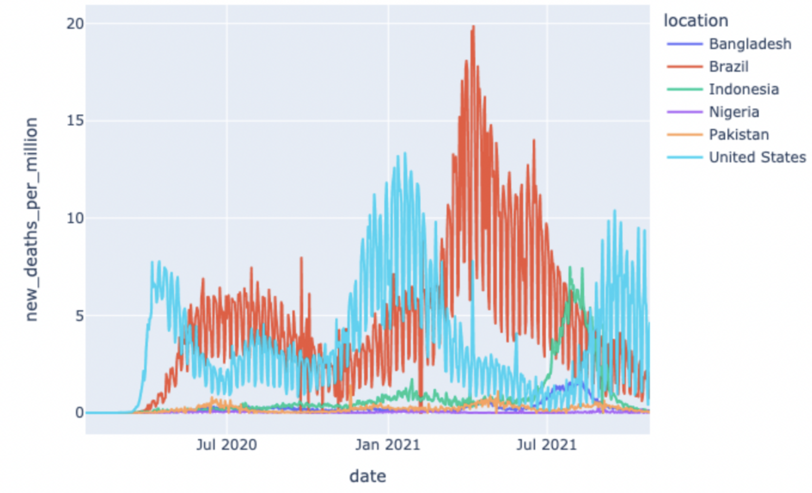


Figure-2. Number of new deaths a day per million population.

The final goal of this task is to identify the peak week of a country that recorded the highest cases and deaths. We have calculated the average number of new cases and deaths per week in each country. The obtained results are as follows:

```
                location          Max_new_cases_in_a_week_per_million
  Max_new_deaths_in_a_week_per_million       New_Max_cases_Week_start_date
                          New_Max_deaths_Week_start_date

0       United States   744.0   10.0    2021-01-10     2021-01-10

1       Brazil          358.0   14.0    2021-03-28     2021-04-11

2       Nigeria         8.0     0.0     2021-01-24     2020-03-01

3       Pakistan        26.0    1.0     2020-06-14     2020-06-21

4       Bangladesh      86.0    1.0     2021-08-01     2021-04-18

5       Indonesia       181.0   6.0     2021-07-18     2021-08-01
```

We also did small background research to find what could be the possible reasons for the spike in cases or deaths. There is a spike of cases in the US around January 2021. There are multiple holidays around this time, such as Christmas and New Years. Also, it is during a time when the weather is colder, which is the perfect weather to catch a cold or some other illness. For Brazil, it is winter during June due to its position in the southern hemisphere, therefore cases and deaths may have peaked because of this. In Pakistan, Indonesia and Bangladesh the relaxation in covid lockdown restrictions caused the increase in cases. All the countries recorded the highest number of deaths when the cases are high and the hospital failed to accommodate everyone to treat. The lack of hospital beds, ventilators and proper treatment led to the peak in deaths.

Akasha

Chose to analyze statistics for the state of Pennsylvania

CASES mean 1.2, median 1.0, mode 0

DEATHS mean 0, median 0, mode 0

compared Pennslyvania to North Carolina.

NC had lower data across the board and had equal deaths stats with the differences being:

CASES 0.01, median 0

Sytiva

Task 1

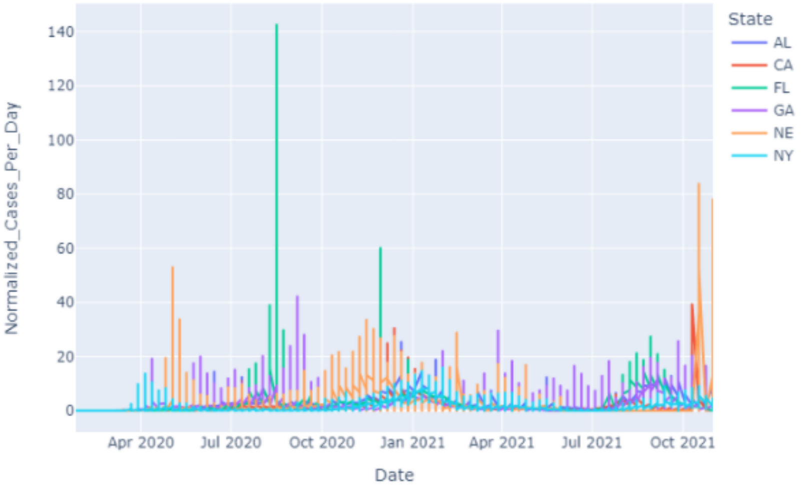
The State I chose to analyze for Task 1 was New York. The weekly statistics were as follows:

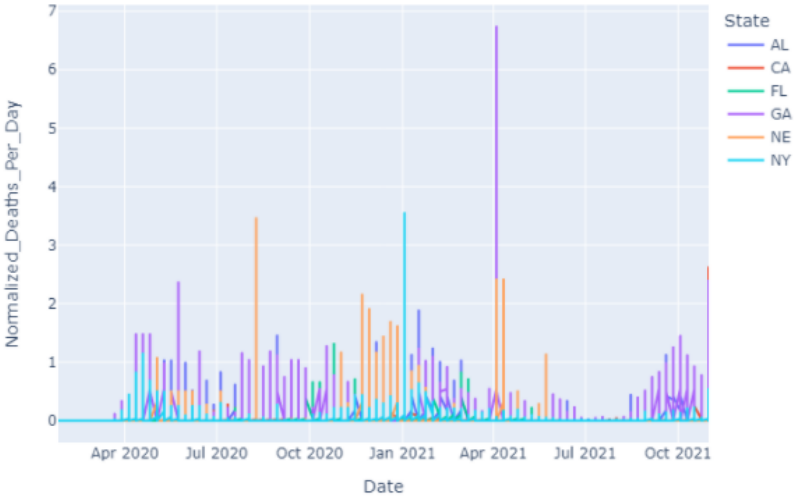
Weekly Cases Mean: 62.657995143947275
Weekly Cases Median: 9.0
Weekly Cases Mode: 0.0

Weekly Deaths Mean: 1.357613596947624
Weekly Deaths Median: 0.0
Weekly Deaths Mode: 0.0

The highest infected counties of New York were Hamilton, Herkimer, Lewis, Genesee, and Rockland counties.

I decided to compare New York to the states of Alaska, California, Florida, Georgia, and Nevada. The comparison graph for both cases and deaths respectively can be found below:





From the data I found here, it seemed like Florida had a very high amount of deaths. I believed this was due to many people in Florida refusing to wear masks, get vaccinated and abide by lockdown mandates. Despite this, Georgia seems to have a much higher death rate than Florida. This could be due to Georgia being one of the worst states in terms of healthcare.

Task 2

The state I chose to analyze for Task 2 was the State of California.
The statistics for the new cases of Covid-19 per day were as follows:

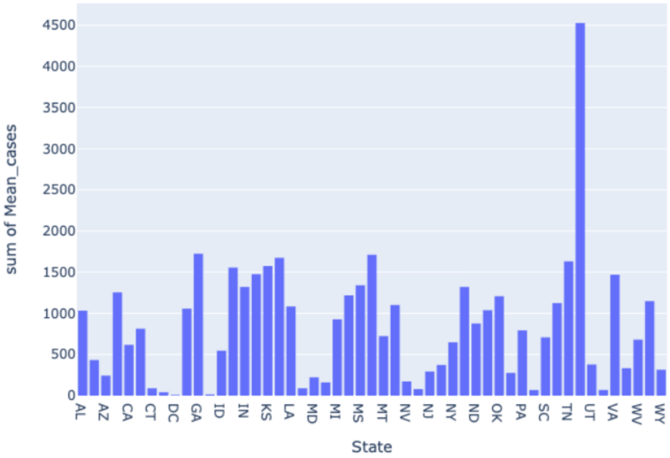
Variance: 380744.90
Skew: 18.08
Kurtosis: 455.21
Center:17778.47

The statistics for the new deaths of Covid-19 per day were as follows:

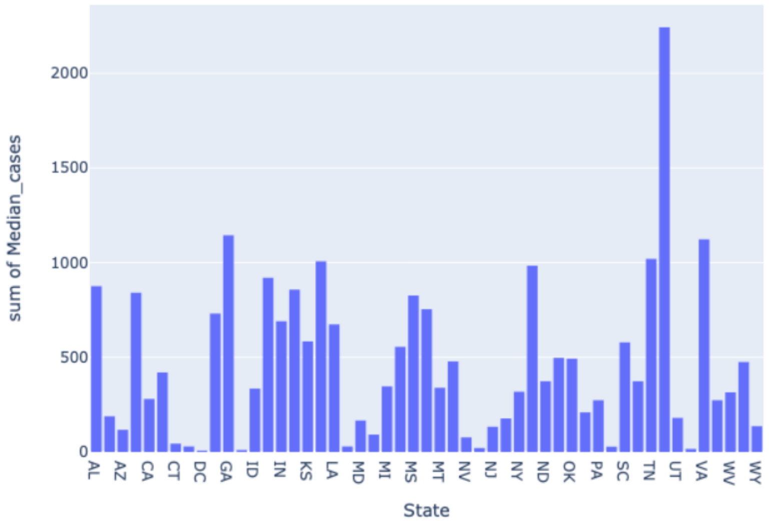
Variance:109.97
Skew: 17.10
Kurtosis: 383.16
Center:17147.13

From the data I found the hypothesis I came up with was that none of my enrichment variables, housing, social, nor economic, correlated with the Covid-19 data. I came up with this hypothesis after I noticed that for all of my graphs, the plotted line and the scattered points never looked like a correlated graph since the points were scattered everywhere.

Chandan:
I have taken Texas as my state of choice to perform statistical analysis .
The mean of weekly cases in Texas is:46010.22891566265
The median of weekly cases in Texas is:27483.0
The mode of weekly cases in Texas is:0.0
Similarly I have calculated the deaths per week
The mean of weekly deaths in Texas is:648.2650602409639
The median of weekly deaths in Texas is:405.0
The mode of weekly deaths in Texas is:0.0
Calculated cases and deaths for all states by normalizing them to 10000 .
Using the weekly data of cases calculated mean and median for each state plotted them as in figures below.



From the above histogram plot we can see that the state of texas has the highest mean number of cases per week and the median per week cases of texas are also high .
My observation here is that Texas has more cases compared to other states.While other states are almost having similar mean number of cases.



Calculated Top 5 counties in the state of texas with highest rate of cases and highest death rates.

Top 5 counties with highest rate of cases:

1. Hale County-TX
2. Maverick County-TX
3. Childress County-TX
4. Dimmit County-TX
5. Crockett County-TX

Top 5 counties with death rate of cases:

1. Foard County-TX
2. Lamb County-TX
3. Kenedy County-TX
4. Maverick County-TX
5. Brooks County-TX

Statistical Distribution for Daily trends of Texas state.

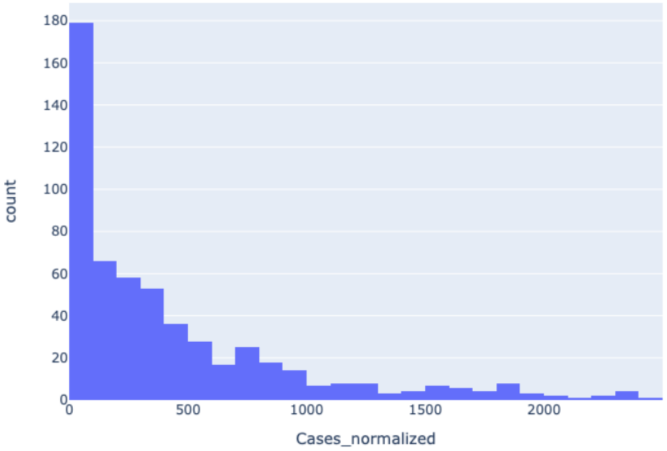
Center of distribution-269.0

Variance-10126184.954732224

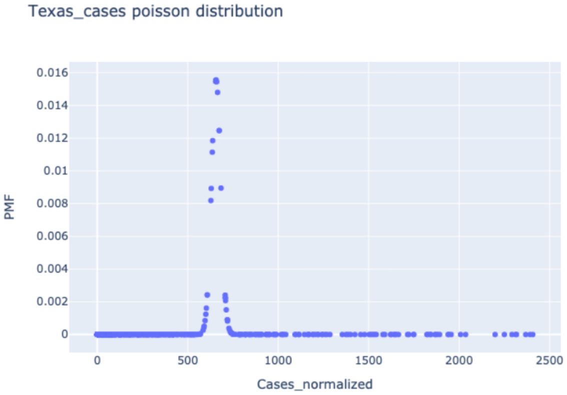
Skewness-18.11817823798287

Kurtosis-358.87937828859407

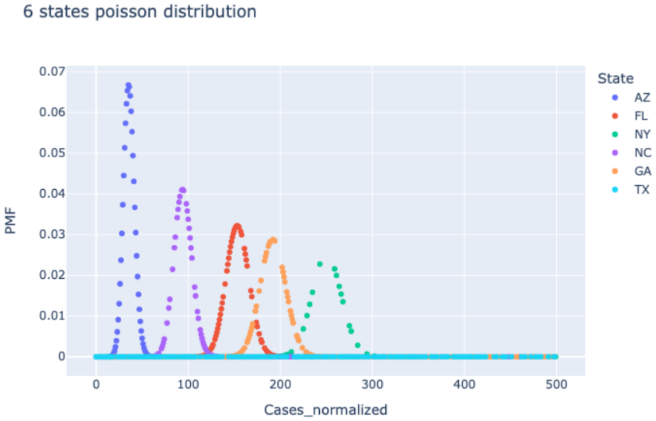
Using histogram I have plotted the daily trends of texas as below.



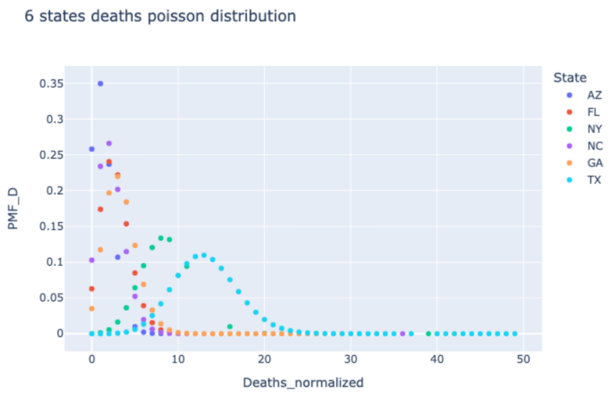
From the plot it is observed that the values are skewed towards left and most of the days have cases as zero.There are few outliers in the data so removed those values for seeing the trends. Poisson distribution plot for cases is as below.



I have selected 5 states along with my texas state. Normalized values by population and plotted its poisson distribution for cases and deaths.

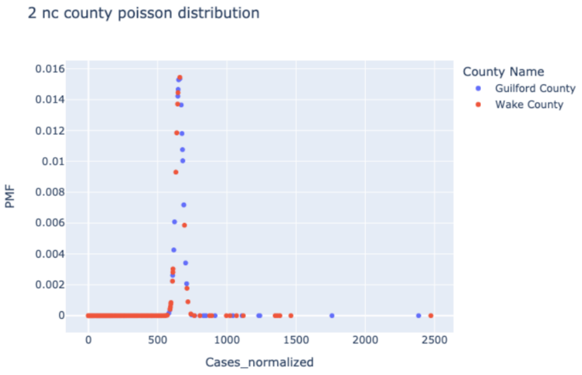


Deaths poisson distribution for 6 states

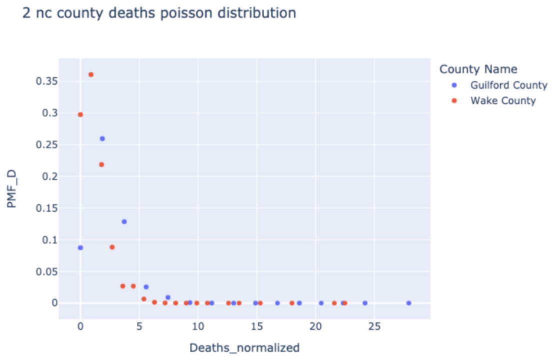

6 states deaths poisson distribution

Calculated NC daily cases for two counties Guliford and Wake county and plotted the poisson distribution for cases and deaths.

Cases poisson distribution


2 nc county poisson distribution

Deaths poisson distribution


2 nc county deaths poisson distribution

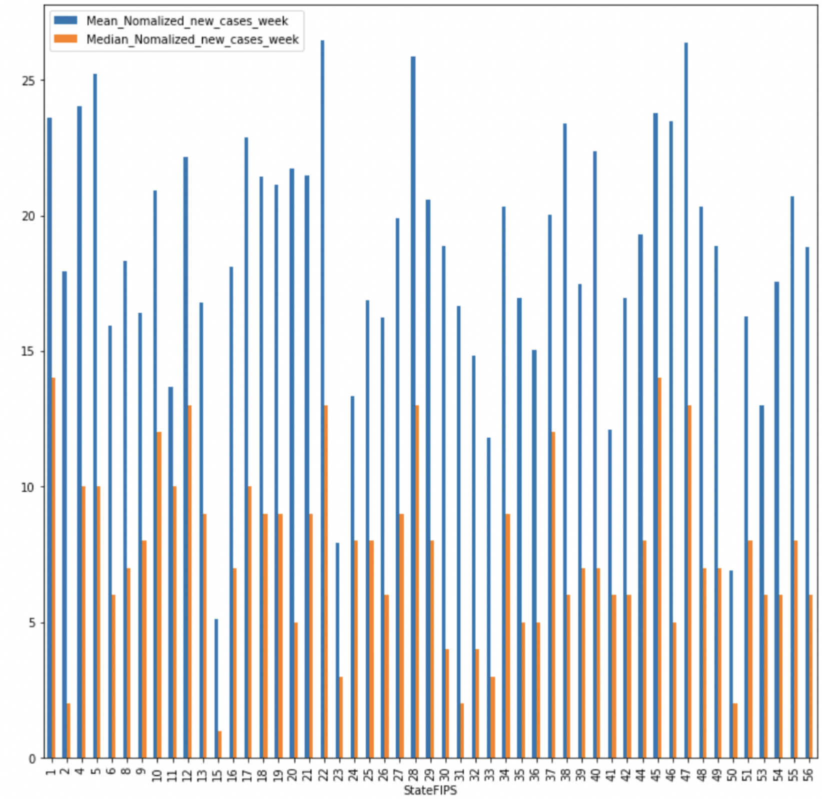
Chandra Shekhar Malgari:

Task-1:

The goal of the first task is to generate weekly statistics for the number of new cases and deaths across a state in the United States and then compare the results with other states in the country. Finally find the counties with highest cases rate and death rate respectively of the selected state and also plot the daily trends of top 5 infected counties in that state.

For this task I have used the merged dataset that was created in the team task of this stage and chose Indiana state (StateFIPS : 18). I have calculated the number of new cases and deaths recorded per day in this state by grouping the data based on the county. The new cases and deaths are calculated using diff(), this will find the difference in current value with the value in the previous row. I also calculated the number of weekly new cases and deaths in each county. The average weekly new cases in Indiana state are 15.83. Whereas the median and mode weekly cases in Indiana state are 3 and 0 respectively. Since there are many records for 0 new cases per day the mode will be 0. The average weekly new deaths in Indiana state are 0.22. Whereas the median and mode weekly cases in Indiana state are 0 and 0 respectively. Since there are many records for 0 new cases per day the mode will be 0.

I have calculated the number of new cases and deaths per day for every county in the United States. Then data is normalised to the population of 100,000 per county. Considering the normalised cases and deaths the average weekly data is calculated. The mean, median and mode for weekly new cases and deaths are calculated for each state. The calculated data is appended to a new data frame. The weekly new cases and deaths values are plotted using histogram for all the states in the United states as shown in below figure.



The mean, median and mode number weekly new cases per week for all states in the US.

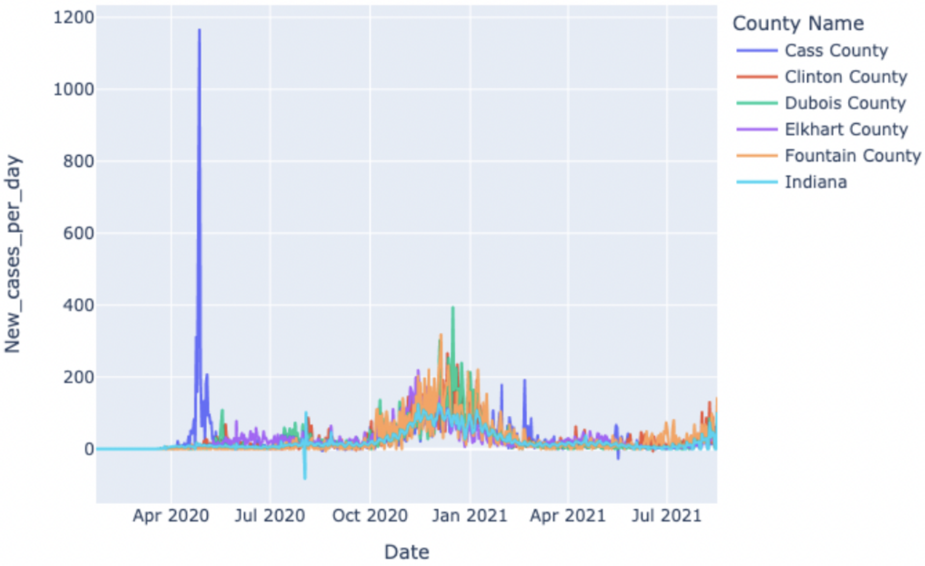
To find the cases and death rates for each county the population of the state is calculated. Then the new cases of county per day is calculated by dividing the new cases per county by the total population of Indiana state. The data is normalised by the total population of the state. The results are as follow:

1. The county with the maximum case rate in Indiana is “Cass County” with a case rate of 1125341.07.
2. The county with the maximum death rate in Indiana is “Pulaski County” with a death rate of 27249.33.

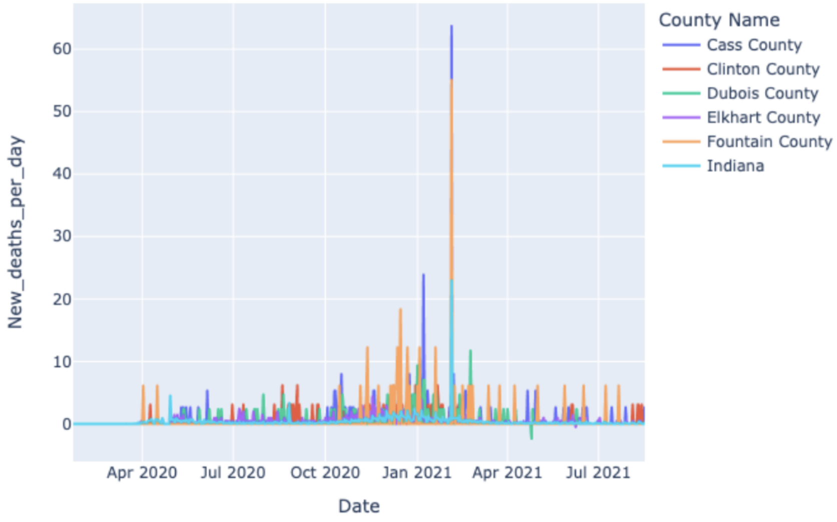
The top 5 infected countries in Indiana state are:

1. Cass County
2. Dubois County
3. Fountain County
4. Clinton County
5. Elkhart County

The comparison of daily new cases and deaths of all the top 5 counties with Indiana state are as shown below respectively.



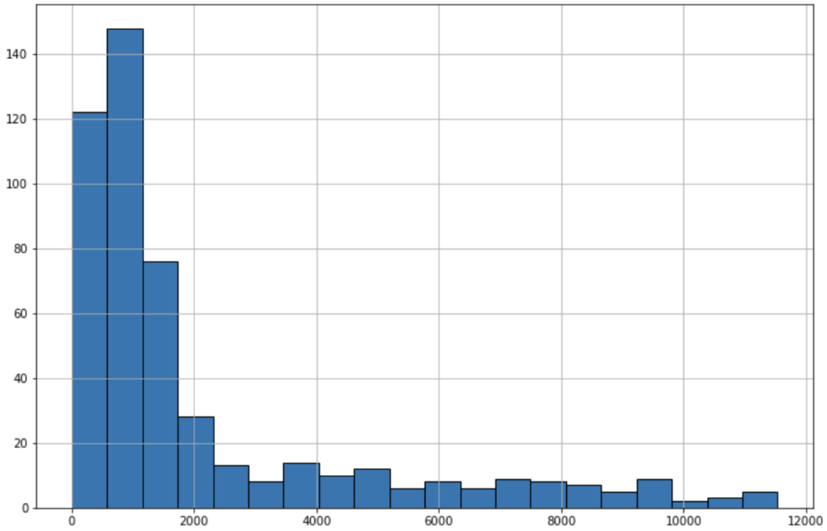
The comparison of new cases per day between the top 5 counties of Indiana and the state.



The comparison of new deaths per day between the top 5 counties of Indiana and the state.

Task-2:

The Indiana state covid dataset is considered from the superset long format data. Then the new cases and new deaths per day are calculated for Indiana state. Then the values are normalised to 100000. All the rows with NaN values are removed to remove the anomalies. Since there are many rows with the value of 0, the distribution of the dataset will be skewed to that value giving the highest peak only at 1. To avoid this the rows where the new cases value is 0 are removed. The daily cases for Indiana state are calculated and plotted as below.



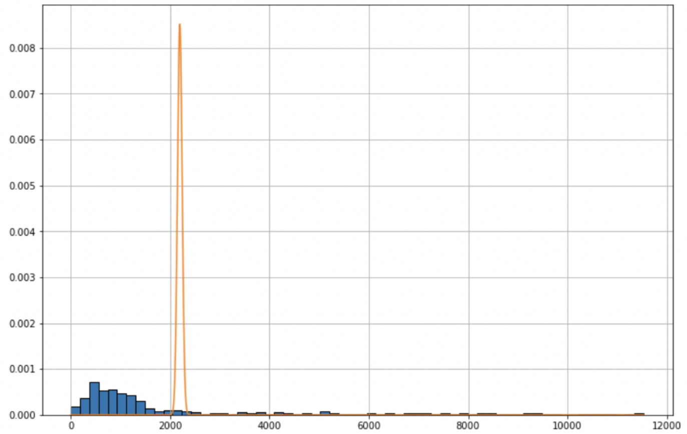
Histogram of Daily new cases count in Indiana state.

The statistical new cases data is calculated for this state and the results are as follows.

1. Center: 2193.81
2. Variance: 6888310.55
3. Skewness: 1.81

4. Kurtosis: 2.34

The probability mass function plot of Indiana state is as follows.

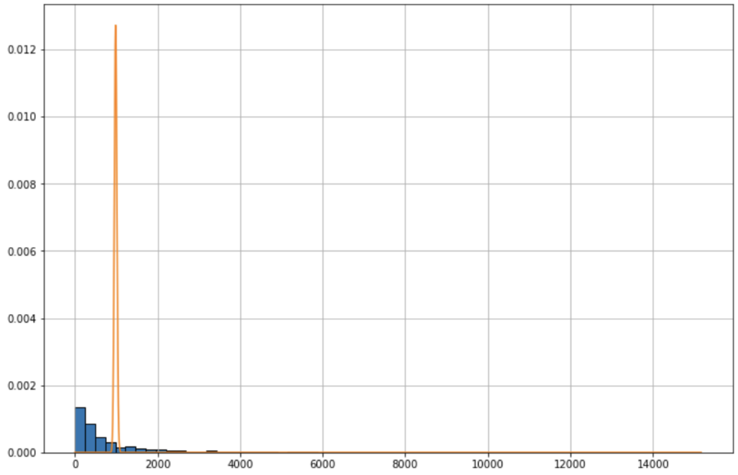


Probability mass function plot of Indiana State

I have considered California, Wyoming, North Carolina, Ohio and New York. The reason is I wanted to see how the most populated and least populated states impacted. The other 3 states are some randomly selected. The data is collected for each state separately and statistical data is calculated first and then the probability mass function is calculated. The statistical data of new cases in California state is as follows:

1. Center: 984.26
2. Variance: 2082453.08
3. Skewness: 3.47
4. Kurtosis: 20.38

The probability mass function of new cases in California is as follows.



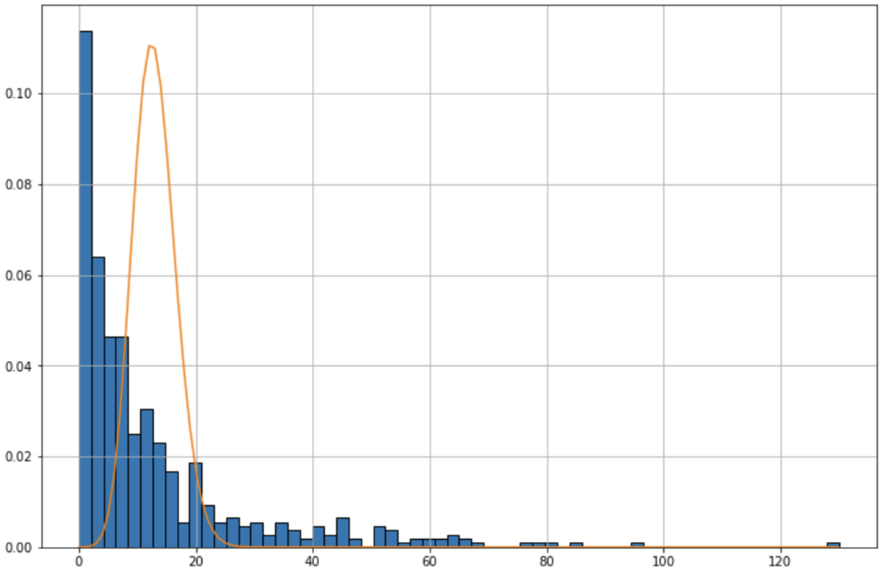
Probability Mass function of new cases in California State.

The statistical data for new deaths in California state is as follows:

1. Center: 984.26
2. Variance: 280.60

3. Skewness: 2.44
4. Kurtosis: 7.86

The probability mass function of new deaths in California is as follows.

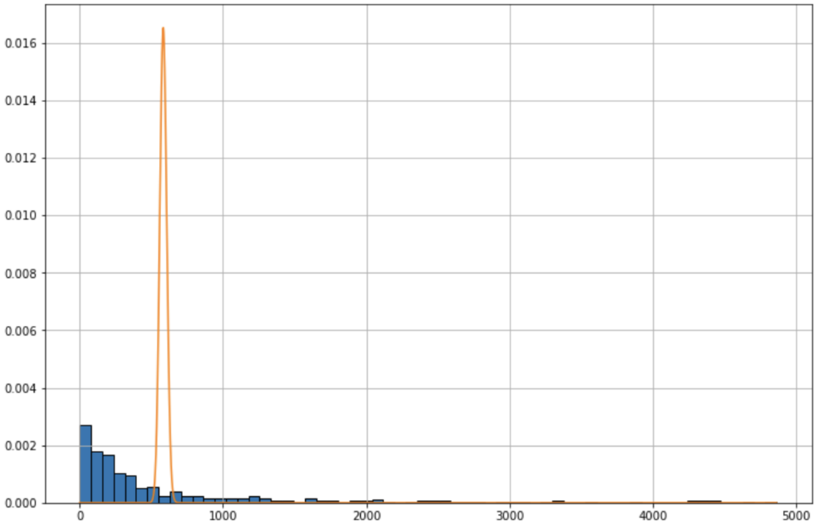


Probability Mass function of new deaths in California State.

The statistical and probability mass function is calculated in Wyoming, the least populated state in the country. The data speaks of the effect of covid, however the results are normalised and the results are as follows. The statistical data of Wyoming for new cases per day are:

1. Center: 583.22
2. Variance: 751425.80
3. Skewness: 2.70
4. Kurtosis: 7.84

The PMF of new cases in Wyoming State is as follows:

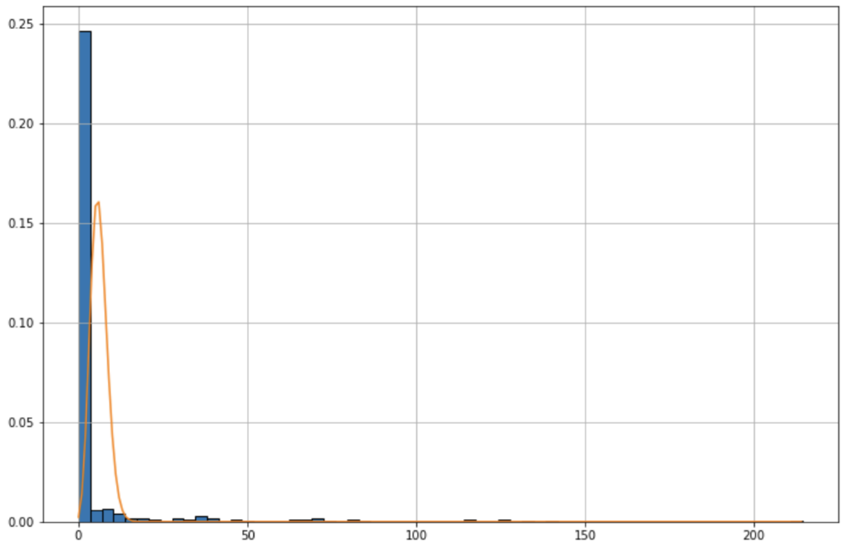


Probability Mass function of new cases in Wyoming State.

The statistical data of Wyoming for new deaths per day are:

1. Center: 583.22
2. Variance: 471.82
3. Skewness: 5.13
4. Kurtosis: 31.66

The probability mass function of new deaths in Wyoming State.

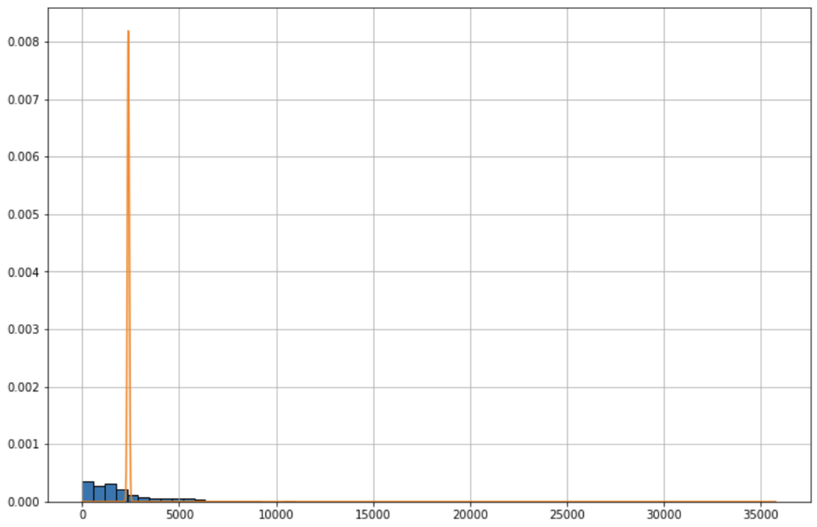


Probability Mass function of new deaths in Wyoming State.

The third state is North Carolina. The probability mass function of North Carolina is skewed to the left with a long tail to the right. The statistical data are:

1. Center: 2375.41
2. Variance: 8728352.21
3. Skewness: 5.14
4. Kurtosis: 45.08

The probability Mass Function of new cases in North Carolina is as follows:

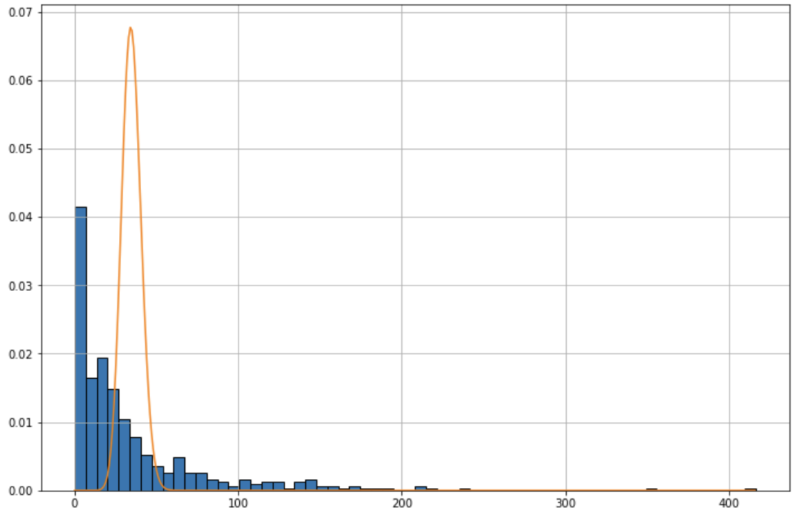


The PMF of new cases in North Carolina.

The statistical data of North Carolina data of new deaths per day are as follows:

1. Center: 2375.41
2. Variance: 2282.33
3. Skewness: 3.07
4. Kurtosis: 14.39

The Probability Mass Function of new deaths in North Carolina state are as follows.

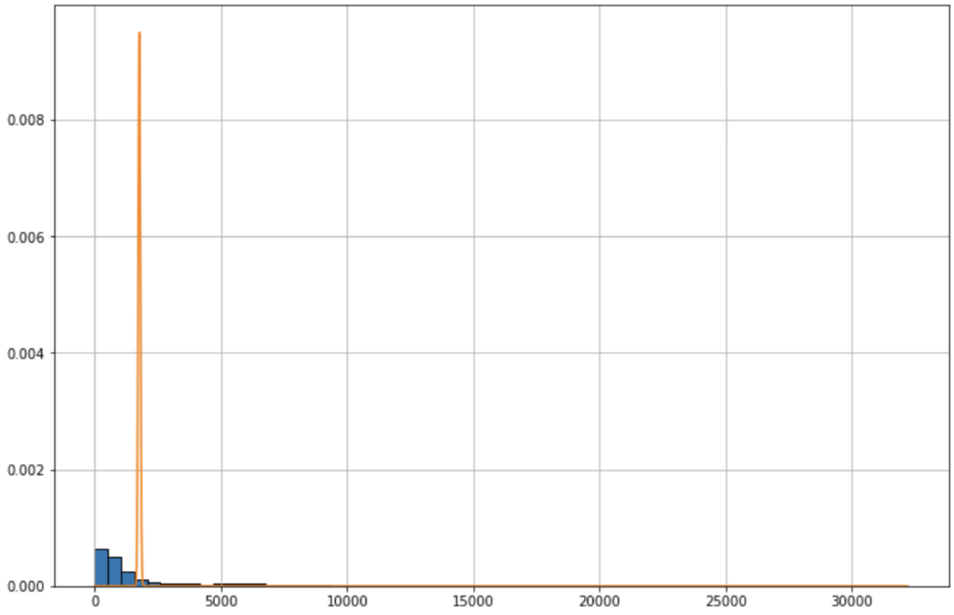


Probability Mass Function of new deaths in North Carolina state

The fourth state is Ohio. The statistical and Probability mass function are calculated on the normalised dataset. The statistical data of new cases per day in Ohio.

1. Center: 1762.92
2. Variance: 7655932.19
3. Skewness: 4.95
4. Kurtosis: 40.04

The probability mass function of new cases in Ohio.

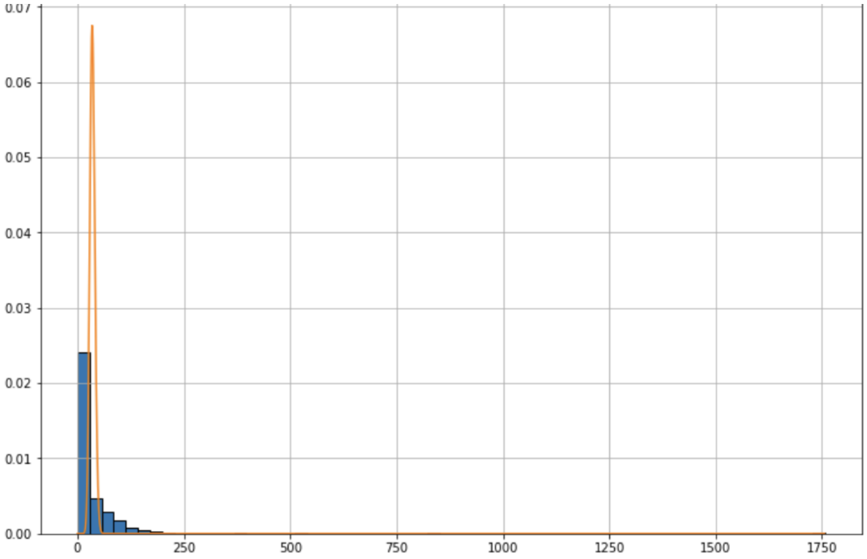


The probability Mass Function of new cases in Ohio.

The statistical data of new deaths in Ohio are as follows.

1. Center: 1762.92
2. Variance: 9600.47
3. Skewness: 12.74
4. Kurtosis: 204.77

The probability mass function of new deaths in Ohio.



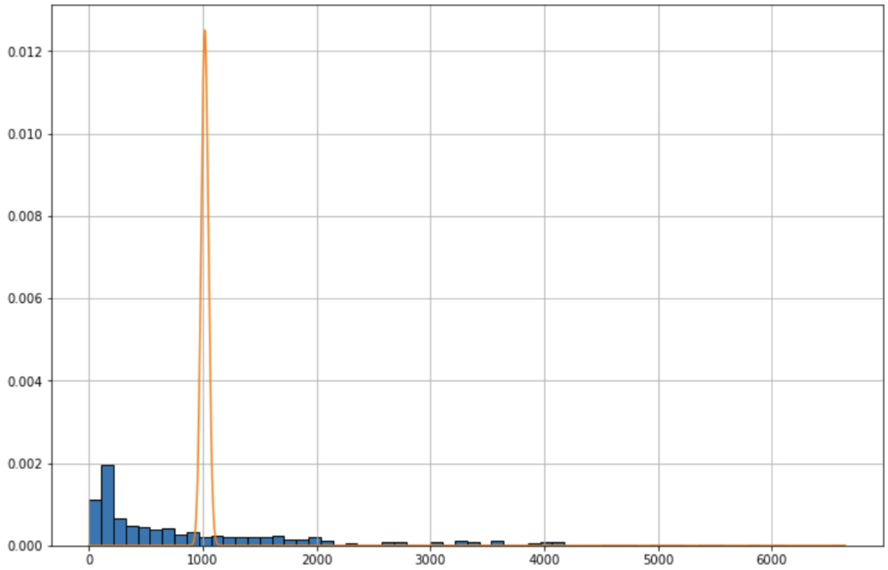
The probability mass function of new deaths in Ohio.

The fifth state is New York. The statistical and probability mass function is calculated on this dataset. The statistical data of new cases in New York is as follows:

1. Center: 1017.795

2. Variance: 1455335.17
3. Skewness: 1.83
4. Kurtosis: 3.23

The probability mass function of New York new cases is as follows:

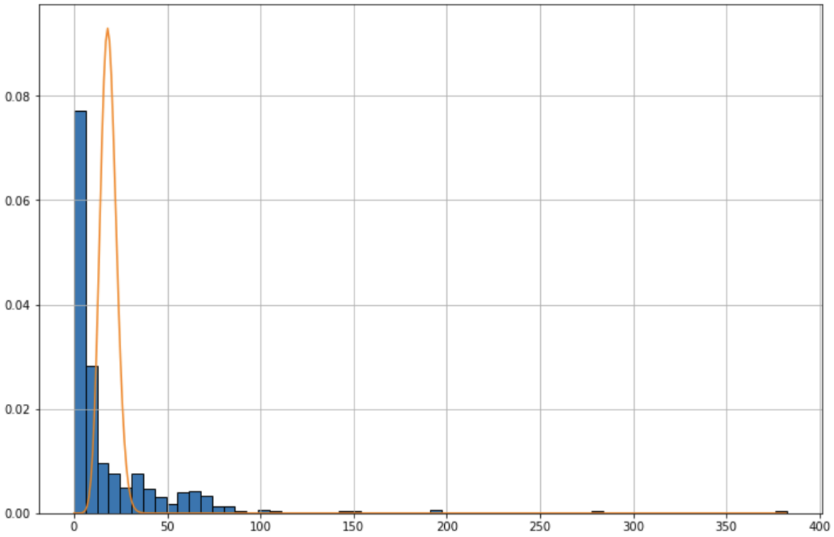


The probability mass function of New York new cases.

The statistical data of new deaths of New York as follows:

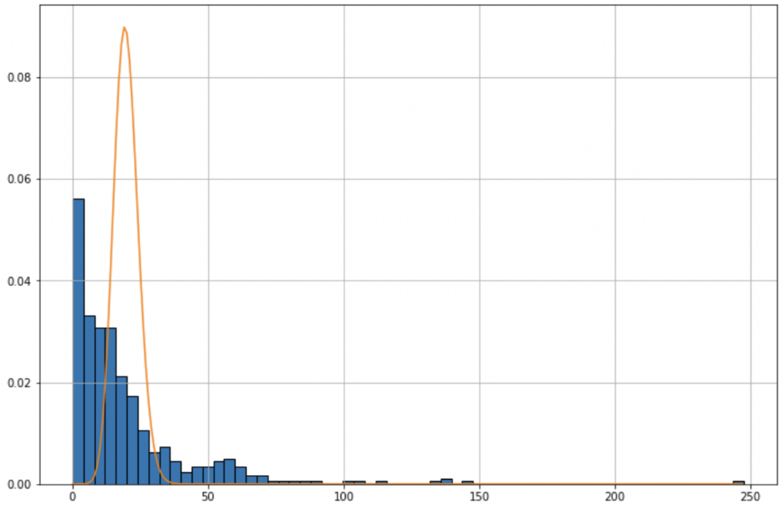
1. Center: 1017.795
2. Variance: 1031.80
3. Skewness: 5.09
4. Kurtosis: 42.10

The probability mass function of New York new deaths is as follows:

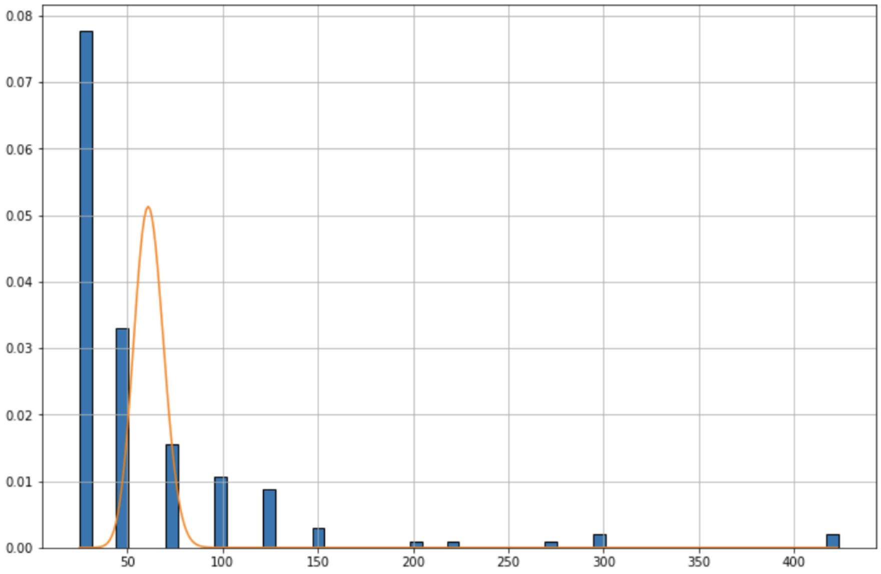


The probability mass function of New York new deaths.

The next task is to calculate the probability mass function of the counties in North Carolina. I have considered the counties with highest and lowest population. The probability mass functions are as follows.



The probability mass function of North Carolina's most populated county.



The probability mass function of North Carolina's least populated county.

The next task is to apply correlation on the covid dataset and the employment dataset. I have merged the data with the long format of the data and applied correlation. It generated the following results. From the above correlation it is found that the Quarterly establishments, Quarterly implement, Quarterly wages, quarterly taxable wages and quarterly contributions are highly correlated to number of cases and deaths. Whereas third month employment change and weekly wages are least impacted by the number of covid cases and deaths