Data Science Stage- 2

Task 1: Team:

The main agenda of this stage is to statistically analyze the dataset and understand it's distribution. We have calculated the mean, median and mode of new cases in United states as stated below.

Mean: 62329
Median: 45678
Mode: 1

Death cases:

Mean: 1072
Median: 817
Mode: 0

Reading the world covid data and choosing 5 different countries of our choice. We chose Indonesia, Brazil, Nigeria, Russia, Japan calculated the statistics for each country and compared it with united states. This statistical analysis is done on both new covid cases and new covid deaths by normalizing them. Below values shows the country per day mean

location	new_cases	new_deaths
Brazil	38890.0	1099.0
Indonesia	7389.0	227.0
Japan	2220.0	29.0
Nigeria	349.0	4.0
Russia	12465.0	326.0
United States	70573.0	1188.0

Weekly mean of all countries:

	location	new_cases	new_deaths
0	Brazil	1814.0	51.0
1	Indonesia	267.0	8.0
2	Japan	176.0	2.0
3	Nigeria	16.0	0.0
4	Russia	853.0	22.0
5	United States	2116.0	36.0

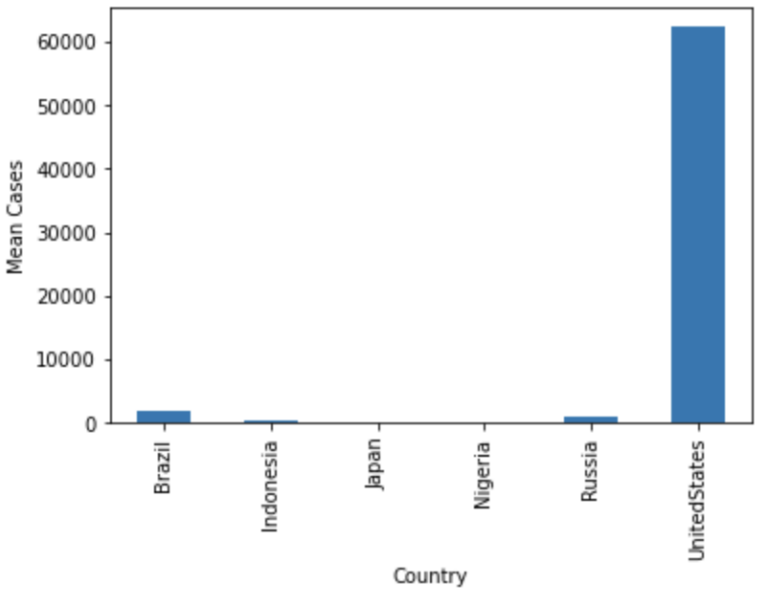
Weekly Median of all countries

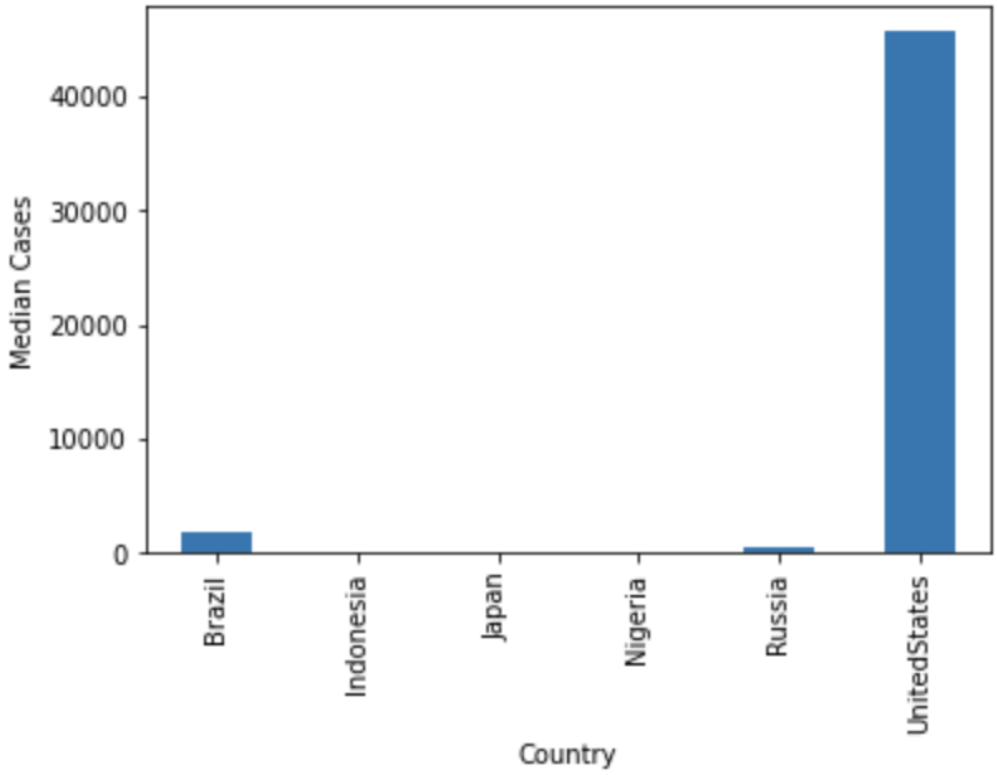
	location	new_cases	new_deaths
0	Brazil	1758.0	45.0
1	Indonesia	163.0	4.0
2	Japan	102.0	1.0
3	Nigeria	8.0	0.0
4	Russia	617.0	25.0
5	United States	1527.0	26.0

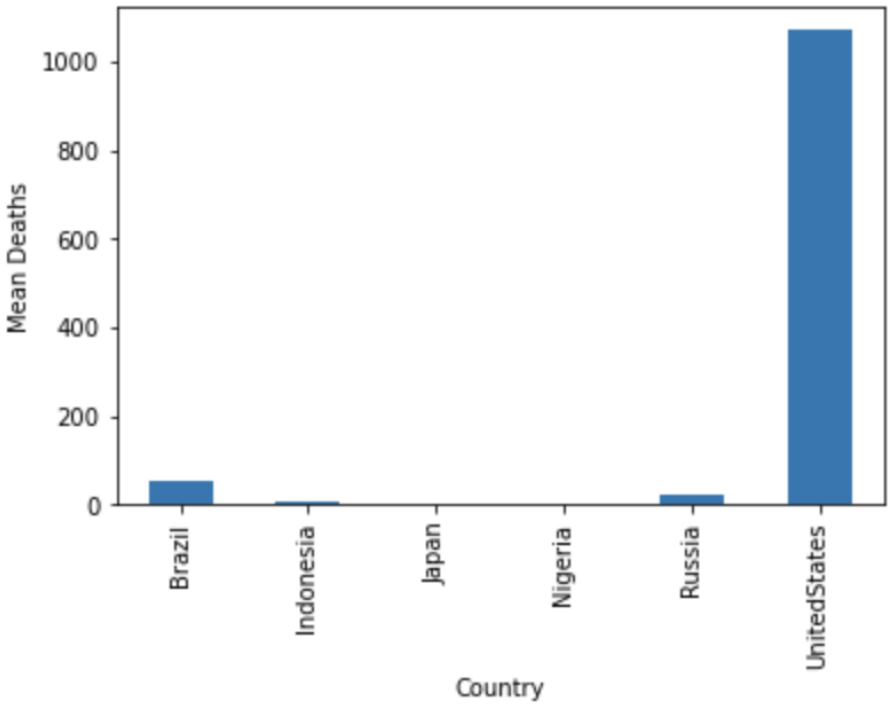
Weekly Mode of all countries

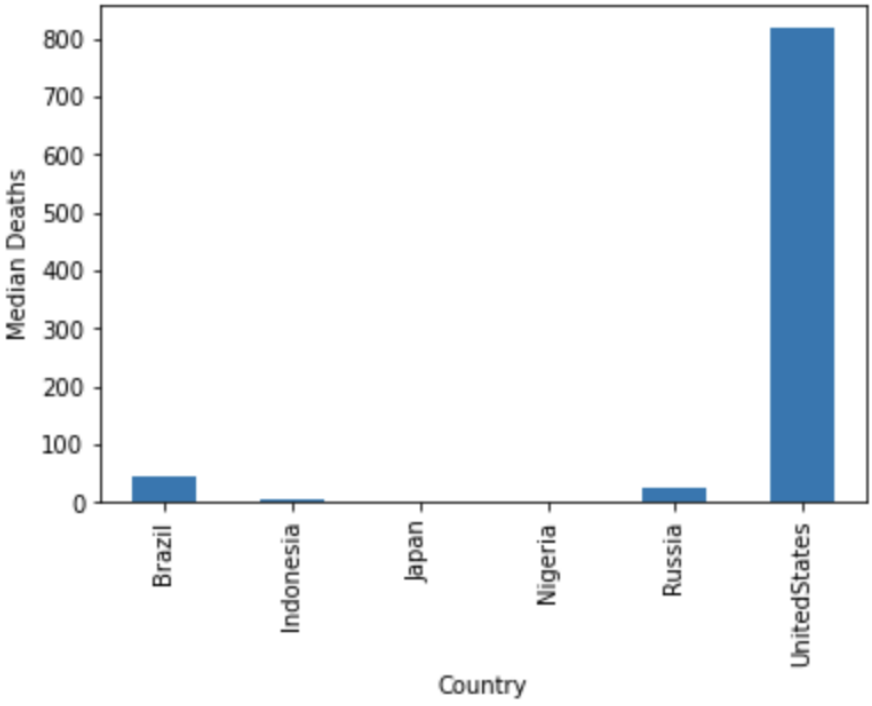
```
Mode of New Cases 3.4113231621321933
Mode of New Deaths 0.0
```

Task- 1 Team Plot graphs (Shipra)


Mean Cases
60000
50000
40000
30000
20000
10000
0
Brazil
Indonesia
Japan
Nigeria
Russia
UnitedStates
Country


Median Cases
40000
30000
20000
10000
0
Brazil
Indonesia
Japan
Nigeria
Russia
UnitedStates
Country



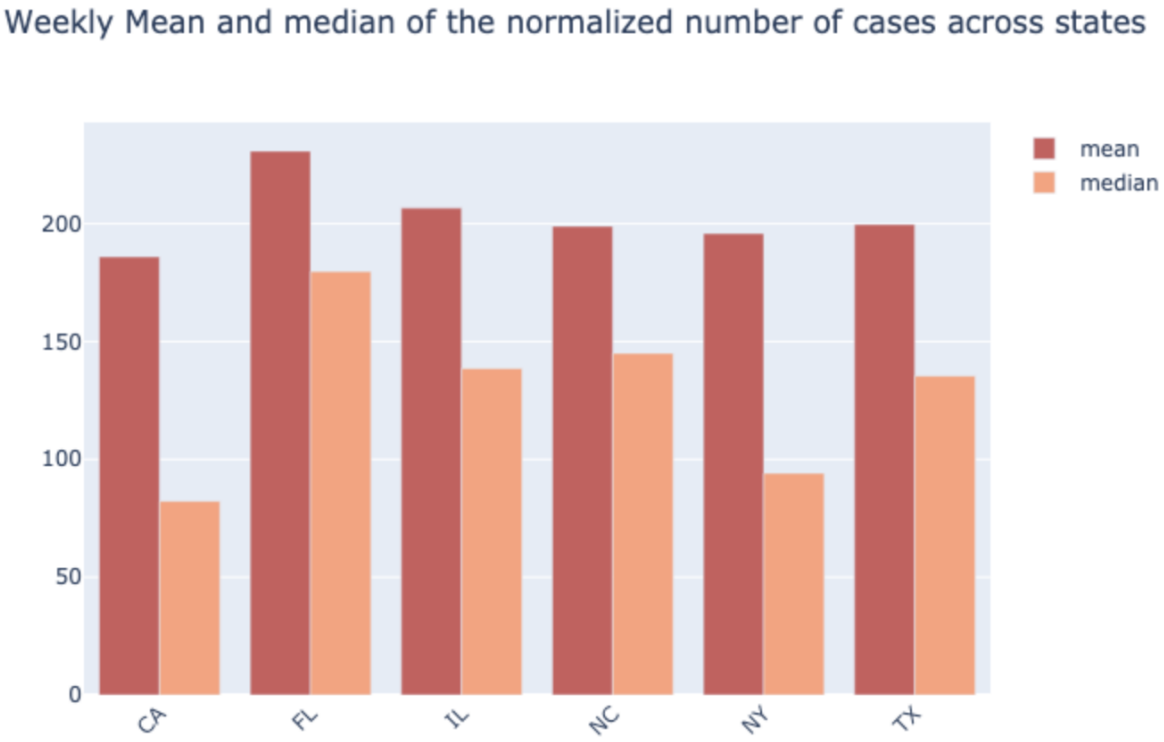


Task-2 Member (Shipra)

Project Stage - II (Data Modeling and Hypothesis Testing)

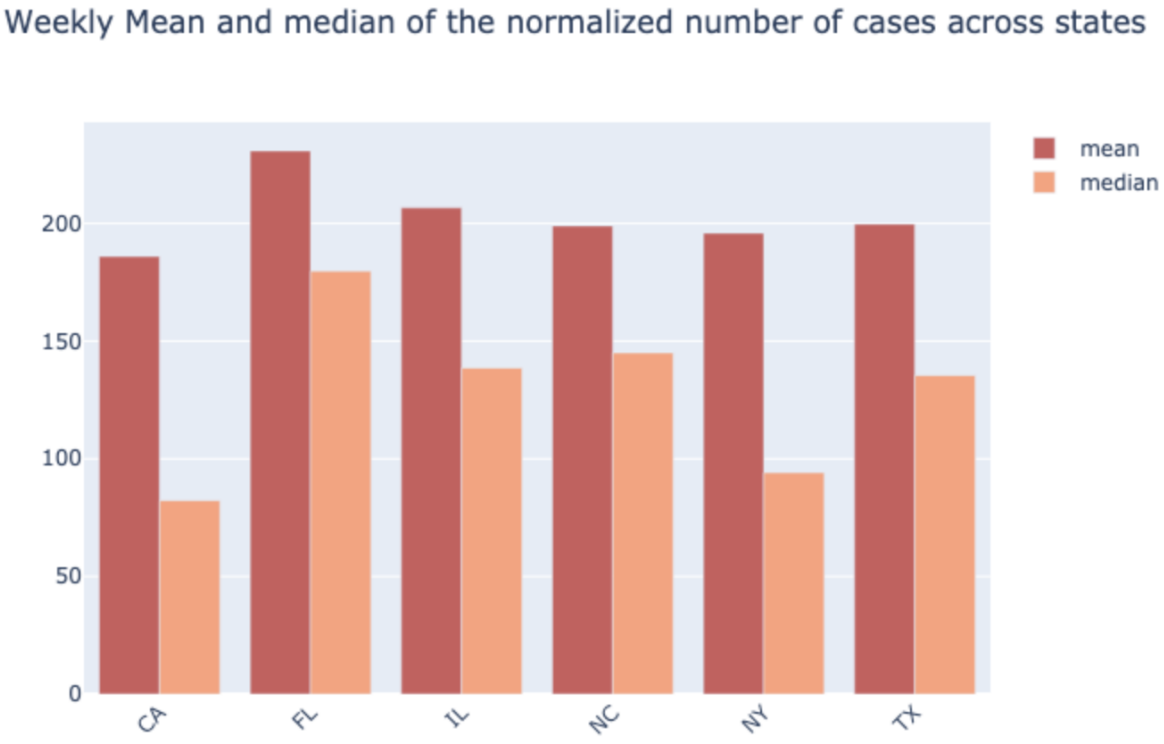
1. Generated weekly statistics (mean, median, mode) for number of new cases and deaths across NC and compared with below Selected states (normalized by population)

- California
- Florida
- Texas
- New York
- Illinois



Comparison

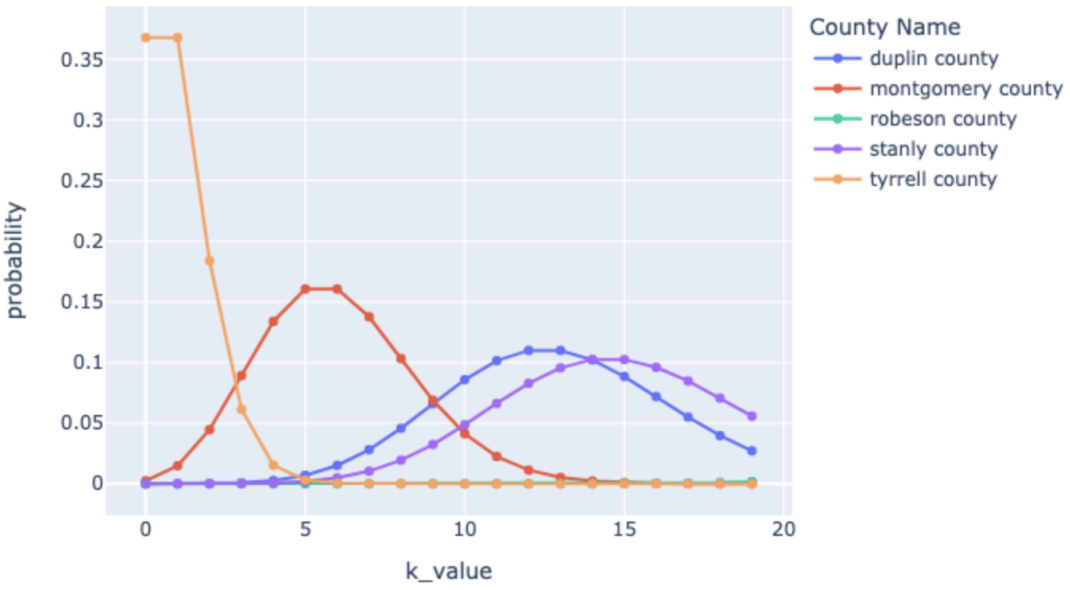
- From the above figure we observe that the Weekly mean number of cases for Florida is the highest and the weekly mean number of cases for CA is the lowest among the 6 states.
- Weekly Median number of cases is the highest for Florida and lowest for California



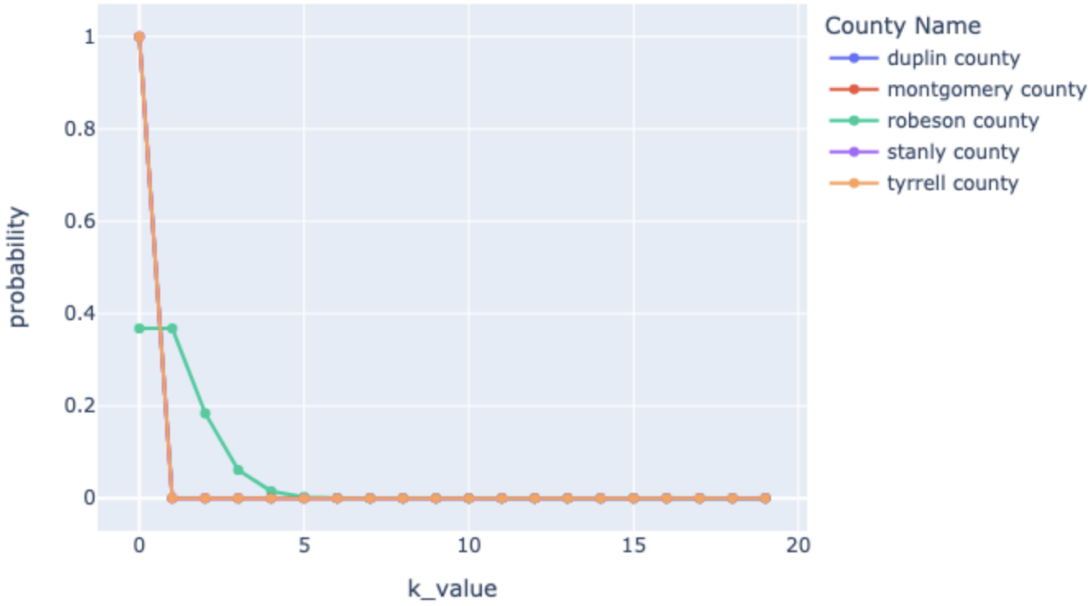
Comparison

- From the above figure we observe that the Weekly mean number of cases for New York is the highest and the weekly mean number of cases for NC is the lowest among the 6 states.
- Weekly Median number of cases is the highest for Florida and lowest for NC

Poisson Distribution for Number of cases across 5 counties in in NC



Poisson Distribution for Number of deaths across 5 counties in in NC



Task – 2: Member

Member Tasks - Deepa Jayanna

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

I have chosen North Carolina state for my analysis because I wanted to get an insight about the state we live in. Everyday cases are calculated and then batched into weekly data and then statistics is measured.

The new cases weekly statistics for North Carolina state is described below. The per weekly cases in North Carolina mostly revolve around the mean value.

Mean -- 13517

Median – 10815

Mode – 0

The new death cases weekly statistics for North Carolina state is described below. The per weekly death cases of North Carolina mostly revolve around 170.

Mean – 170

Median – 125

Mode – 0

Mode value is 0 because at the initial stages of outbreak there were 0 cases and that is the most repeated value for initial few weeks.

Compare the data against other states. (Normalize by population)

For this task, I am choosing California, Texas, Vermont, Newyork and Indiana to compare with North Carolina. Below data is for everyday cases normalized for 100000 population. Texas is the most infected State and Vermont is the least infected state. Indiana, Newyork and California have almost similar range of cases.

State	Mean	Median	Mode
North Carolina	12710	9809	0
California	5894	2741	0
Texas	35793	21605	0
Newyork	6396	2885	0
Indiana	5370	3414	0
Vermont	655	147	0

Death cases comparison of North Carolina state with other states:

Texas has seen the most death cases overall and Vermont has the least number of death cases. When we look at cases per day, even though North Carolina had more cases than Indiana, the death cases of NC is less than Indiana.

State	Mean	Median	Mode
North Carolina	161	100	0
California	45	22	0
Texas	776	393	0
Newyork	82	20	0
Indiana	212	96	0

Vermont	4	0	0

<u>Identify counties within the previous state with high case and death rates. (normalize by population)</u>

Top five counties with high case rates are

Robeson – 176, Hyde – 169, Sampson- 166, Stanly - 164, Columbus – 162

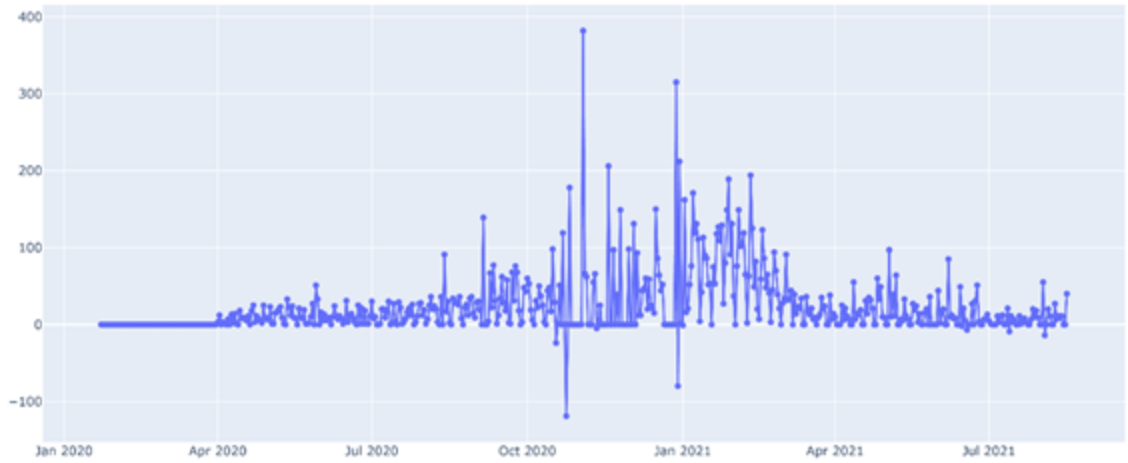
Top five counties with high death rates are

Montgomery - 4, Northampton – 4, Jones – 4, Duplin -- 3, Chowan – 3

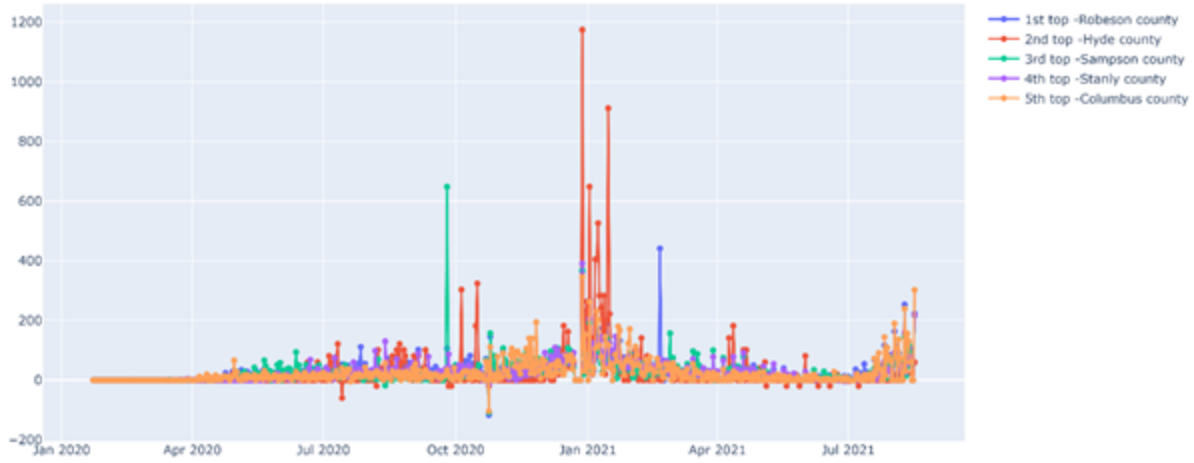
Plots Daily trends:

Daily Trends for North Carolina new cases: Covid started around January and slowly ramped up with cases and reached it's highest peak between January and March 2021. We can see that cases went as high as 30,000. Slowly it started reducing because people started taking vaccinations. And in August, there is an increase of cases again because of the delta variant and the cases still follow the same trend.

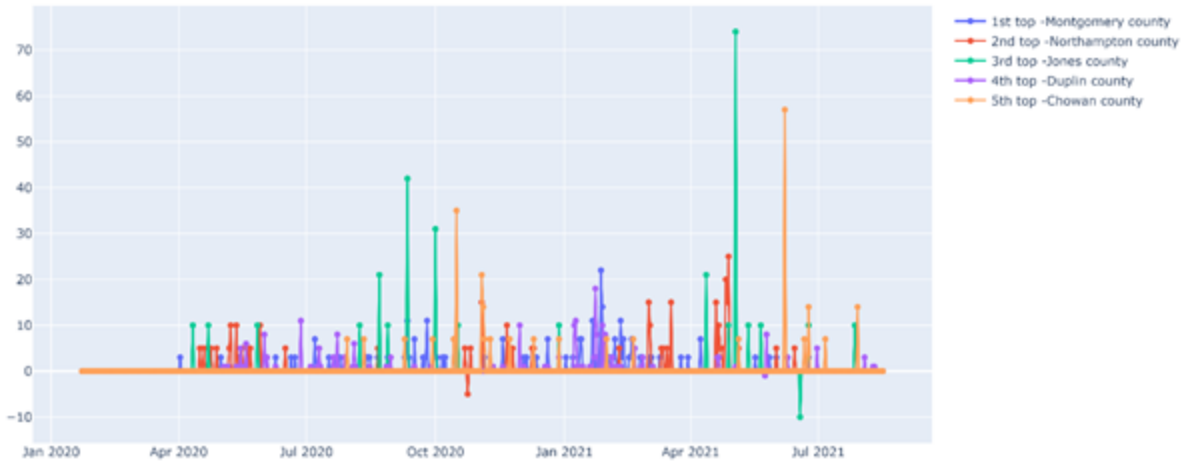
Daily Trends for North Carolina new death cases: Death cases also follow a similar trend as per everyday new cases.



Daily Trends for five most infected counties within North Carolina:



Daily Trends for five most death rate counties within North Carolina:

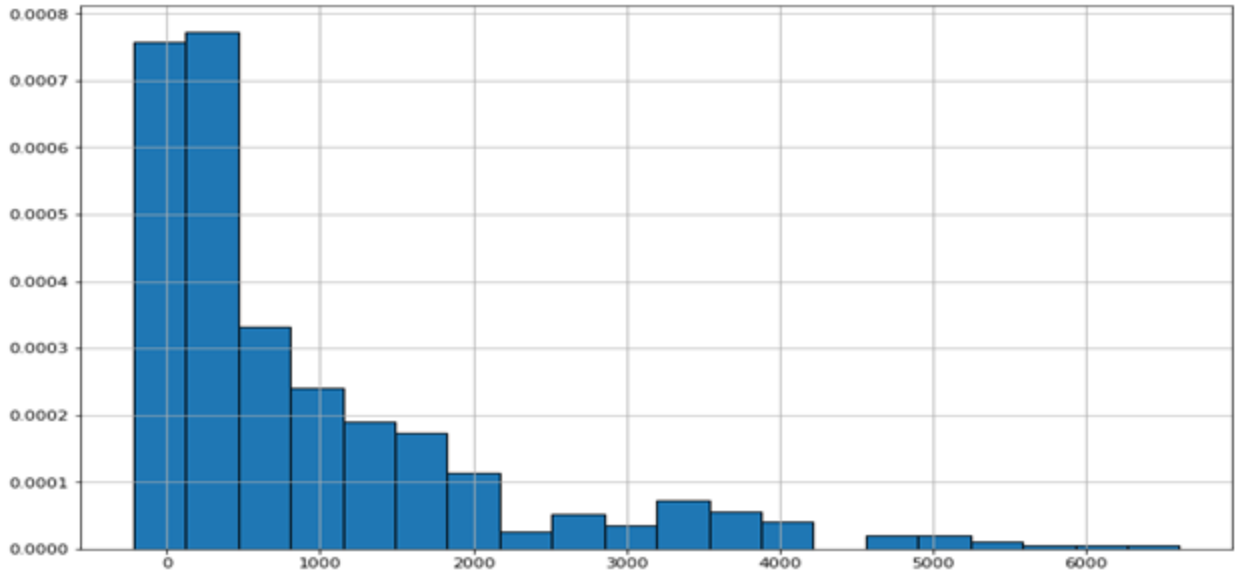


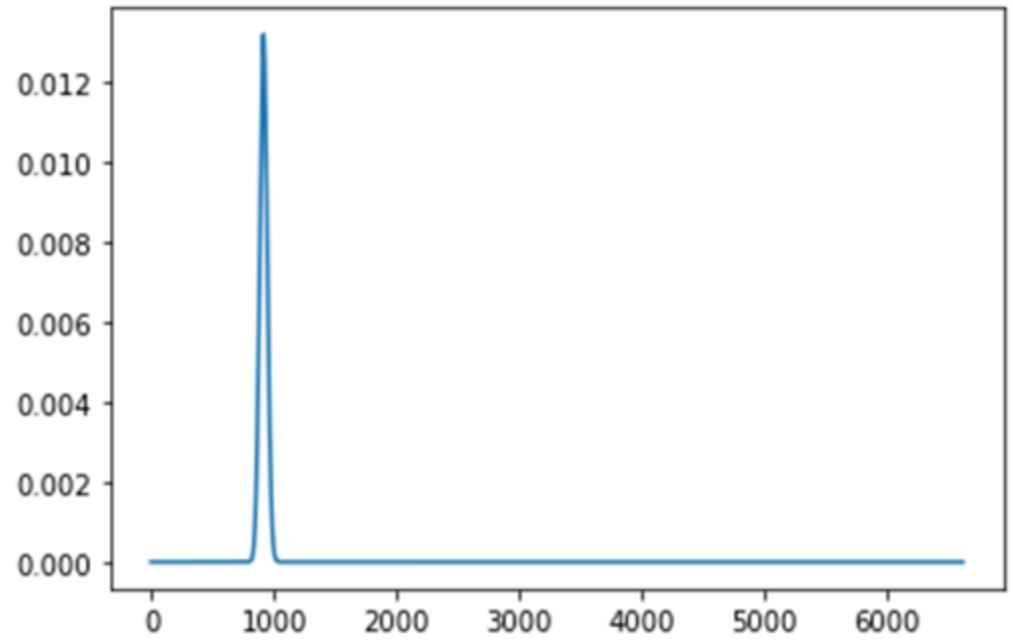
Task 2:

Fit a distribution to the number of COVID-19 new cases of a state of your choosing:

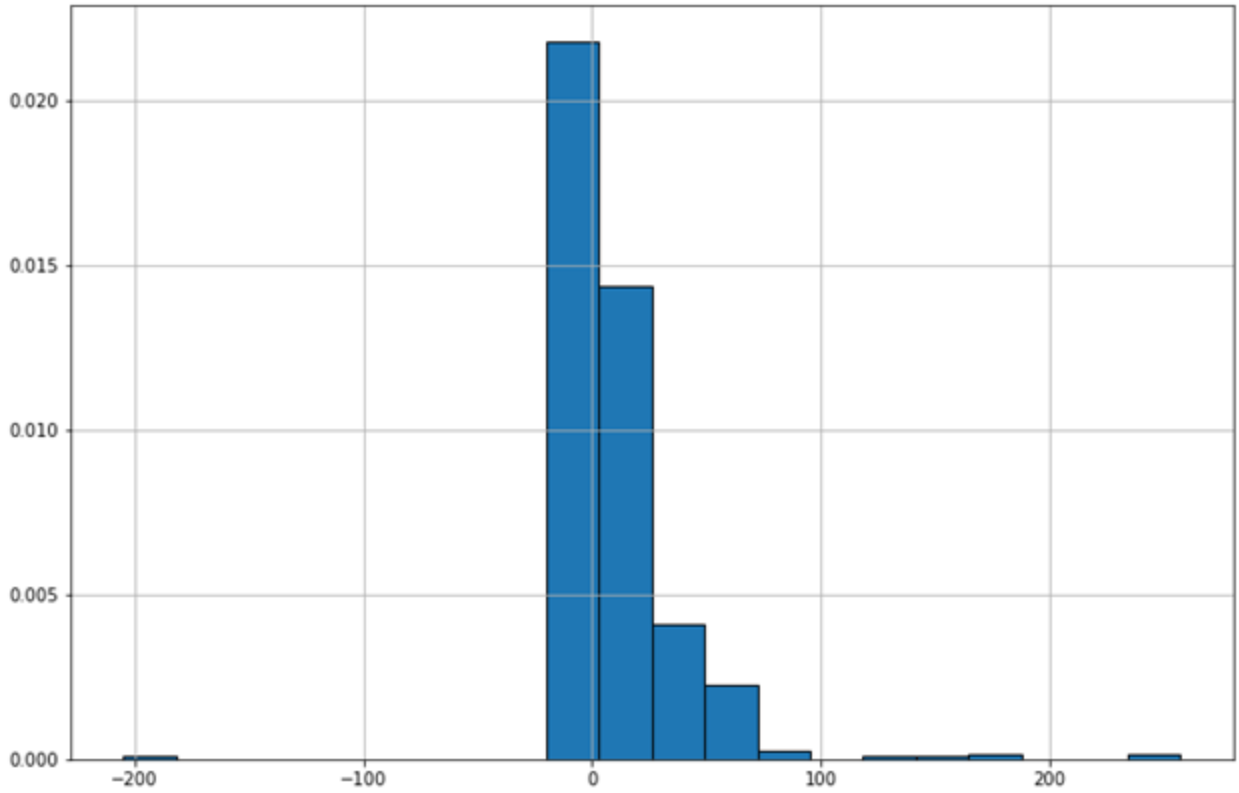
For this task, I am choosing Poisson distribution because and plotting pmf values. Selected state is Newyork

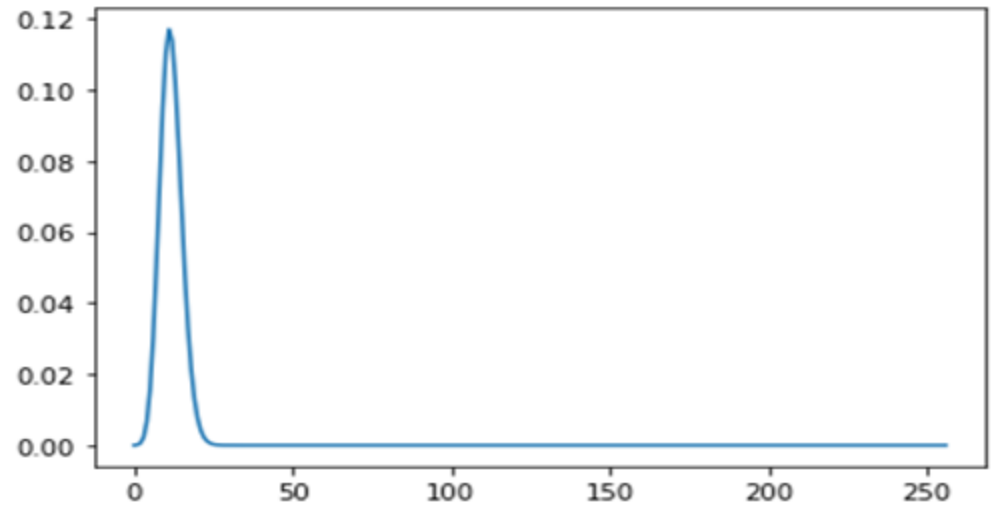
Histogram and pmf plot is shown below. By looking at the histogram and positive Kurtosis our data is skewed.



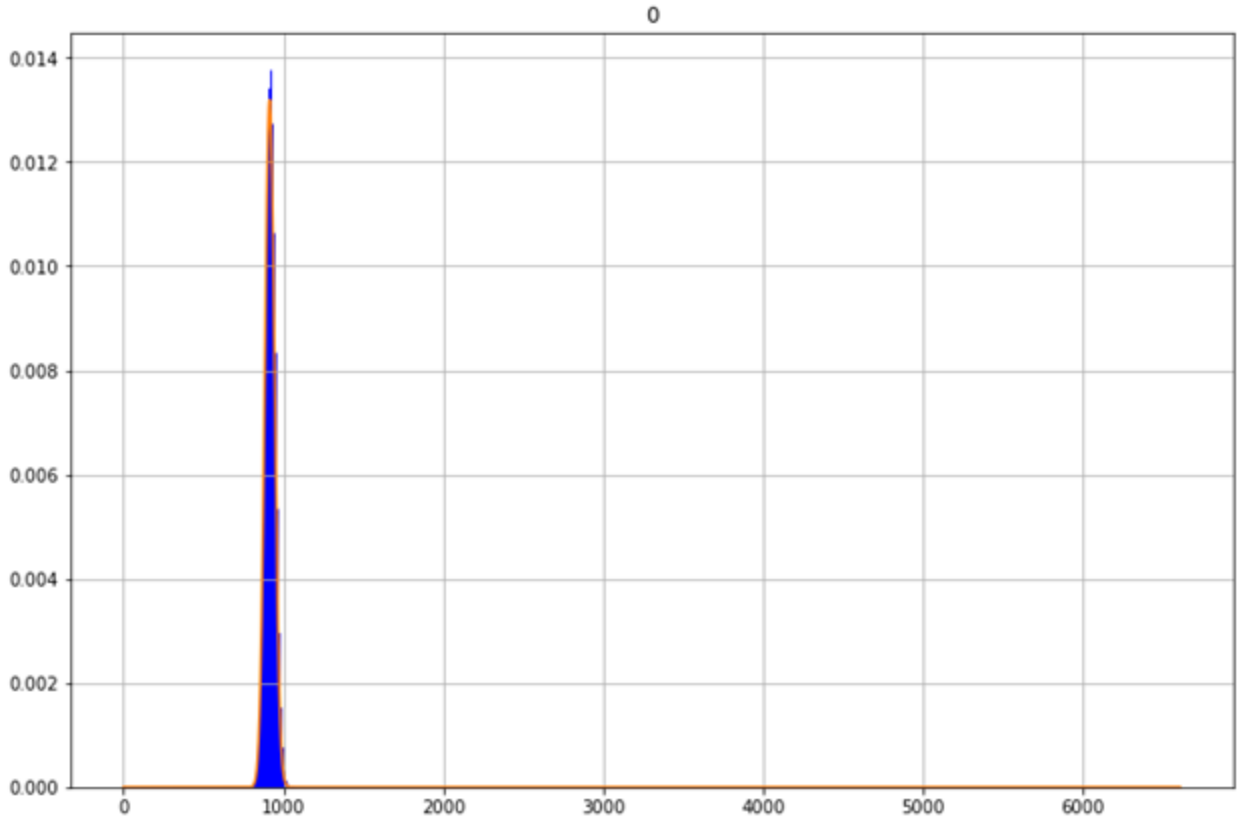


Plotting and distribution for New York Death case

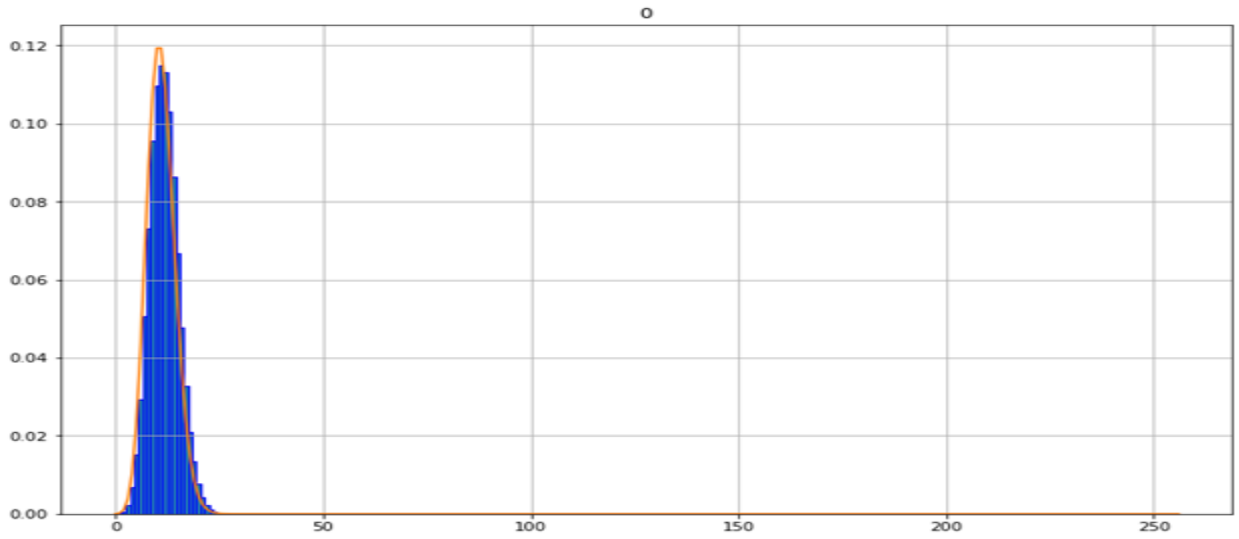




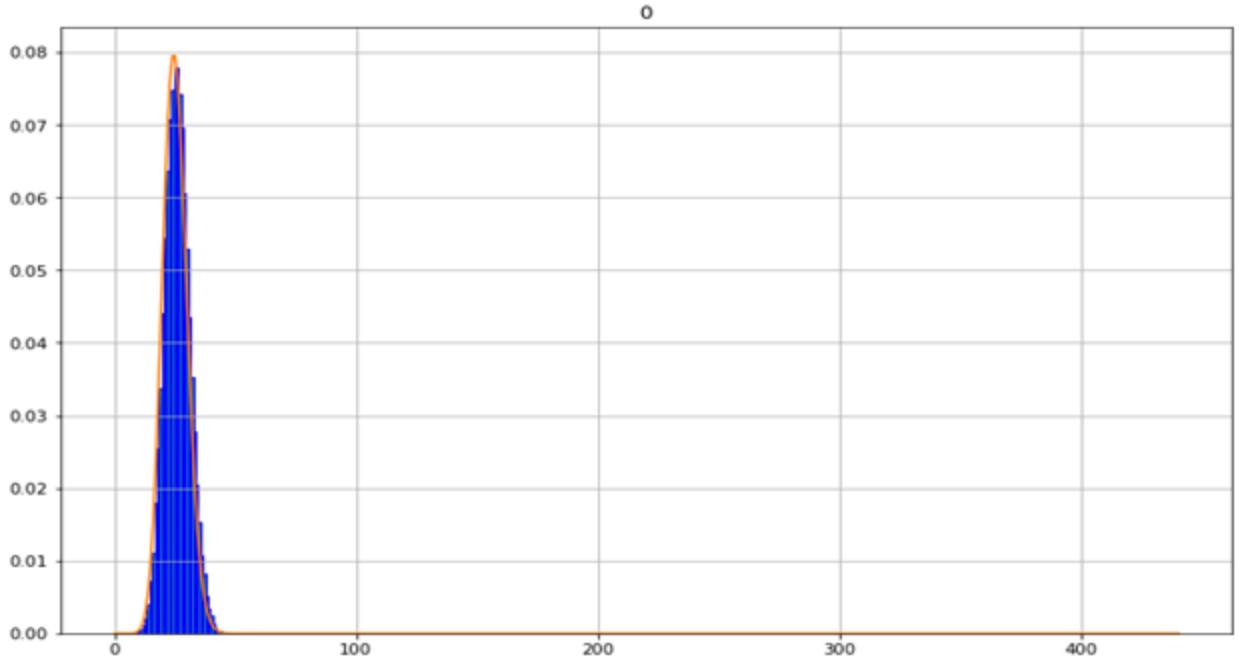
Model Poisson distribution for Newyork:



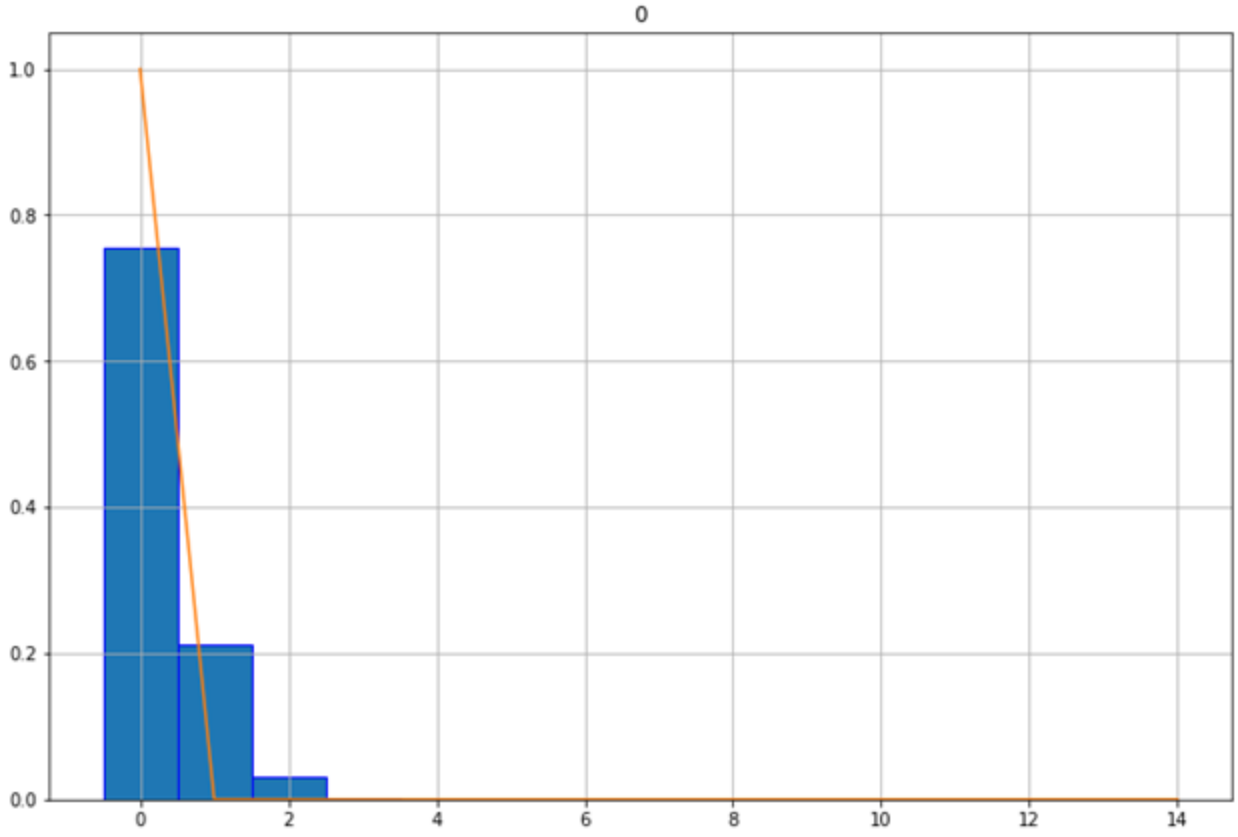
Modelling Poisson Distribution for North Carolina death cases:



Robeson County in NC is the highest infected County. Poisson modelling for new cases is shown below



Robeson County death cases distribution



Perform Correlation with Enrichment Dataset:

In stage 1 I did my analysis on ACS Demographic dataset. I have merged this dataset alongside total cases per day and total deaths per day across a county. I have selected three variables to correlate with i.e Total male population, Total female population and voting age population above 18 years. I have normalized the data per 100000 population and used corr() function from the pandas over the selected variables.

California state Correlation new cases

	Estimate!!SEX AND AGE!!Total population!!Male	Estimate!!SEX AND AGE!!Total population!!Female	Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18 and over population!!Female	newcases
Estimate!!SEX AND AGE!!Total population!!Male	1.000000	-1.000000	-0.655974	0.577321
Estimate!!SEX AND AGE!!Total population!!Female	-1.000000	1.000000	0.655974	-0.577321
Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18 and over population!!Female	-0.655974	0.655974	1.000000	-0.760211
newcases	0.577321	-0.577321	-0.760211	1.000000

When the data is normalized for 100000 population , I obsereved negative correlation between some values. A negative correlation means when one variable increases other variable decreases. If we consider the total cases, with it's increase, there is a reduction in female 18 over population, meaning there are very less people in that age category infected as the cases number increased.

California state death cases

```
ca_corr_death = calculate_corr('CA','death')
ca_corr_death
```

	Estimate!!SEX AND AGE!!Total population!!Male	Estimate!!SEX AND AGE!!Total population!!Female	Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18 and over population!!Female	population	death_cases
Estimate!!SEX AND AGE!!Total population!!Male	1.000000	0.999837	0.999279	0.999958	0.969729
Estimate!!SEX AND AGE!!Total population!!Female	0.999837	1.000000	0.999399	0.999960	0.971916
Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18 and over population!!Female	0.999279	0.999399	1.000000	0.999381	0.966431
population	0.999958	0.999960	0.999381	1.000000	0.970875
death_cases	0.969729	0.971916	0.966431	0.970875	1.000000

We can observe that there is a positive values for non- normalized data, meaning there is dependency of one value of other. If variable raises, other variable also raises. I believe with the increase in new cases and death cases the relation between all age category will increase.

Heng Team task:

4 Russia 28184.714286 789.142857

In [41]:
```
brazil_data = covid_week_mean[covid_week_mean['location']== 'Brazil']
brazil_data = brazil_data[brazil_data['new_cases'] == brazil_data['new_cases'].max()]
brazil_data
```

Out[41]:

	location	date	new_cases	new_deaths	population
53	Brazil	2021-03-22	75416.714286	2305.571429	213993441.0

In [42]:
```
brazil_data = covid_week_mean[covid_week_mean['location']== 'Brazil']
brazil_data = brazil_data[brazil_data['new_deaths'] == brazil_data['new_deaths'].max()]
brazil_data
```

Out[42]:

	location	date	new_cases	new_deaths	population
56	Brazil	2021-04-12	72029.571429	3123.571429	213993441.0

The cases were high due to lack of restrictions in the prior weeks. This week is when strict restrictions like curfews were enforced in Brazil due to high amount of new cases.

The deaths were high that week because it was 1 week after Easter weekend for Brazil.

Calculated all the peak weeks for the US and 5 selected countries. I did some research to see why these weeks were peak weeks for these selected countries. The image above shows an example of Brazil.

The countries we selected were Brazil, Indonesia, Russia, United States, Japan and Nigeria.

Task – 1: Member (Poojitha)

Selected California data to perform statistical analysis and plot them.

California State new cases statistics:

Mean: 7026

Median: 3106

Mode: 0

California State new deaths statistics:

Mean: 69

Median: 8

Mode: 0

Found weekly statistics for other states:

	State	date	high_cases	high_deaths
0	AK	2020-01-20	0.0	0.0
1	AK	2020-01-27	0.0	0.0
2	AK	2020-02-03	0.0	0.0
3	AK	2020-02-10	0.0	0.0
4	AK	2020-02-17	0.0	0.0
...	...	...	...	...
4145	WV	2021-07-19	374.0	116.0
4146	WV	2021-07-26	603.0	37.0
4147	WV	2021-08-02	965.0	144.0
4148	WV	2021-08-09	851.0	0.0
4149	WV	2021-08-16	1333.0	0.0

Highest cases and deaths of all states

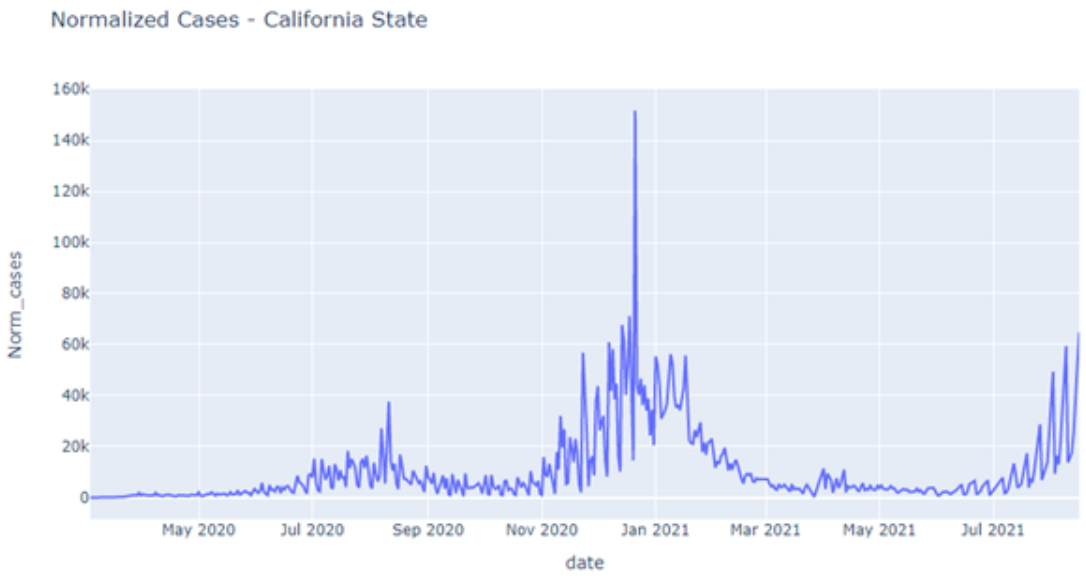
```
Highest Cases :
    State  high_cases  high_deaths
41    SD      1765.0        226.0
Highest Deaths :
   State  high_cases  high_deaths
3    AZ      1580.0        364.0
```

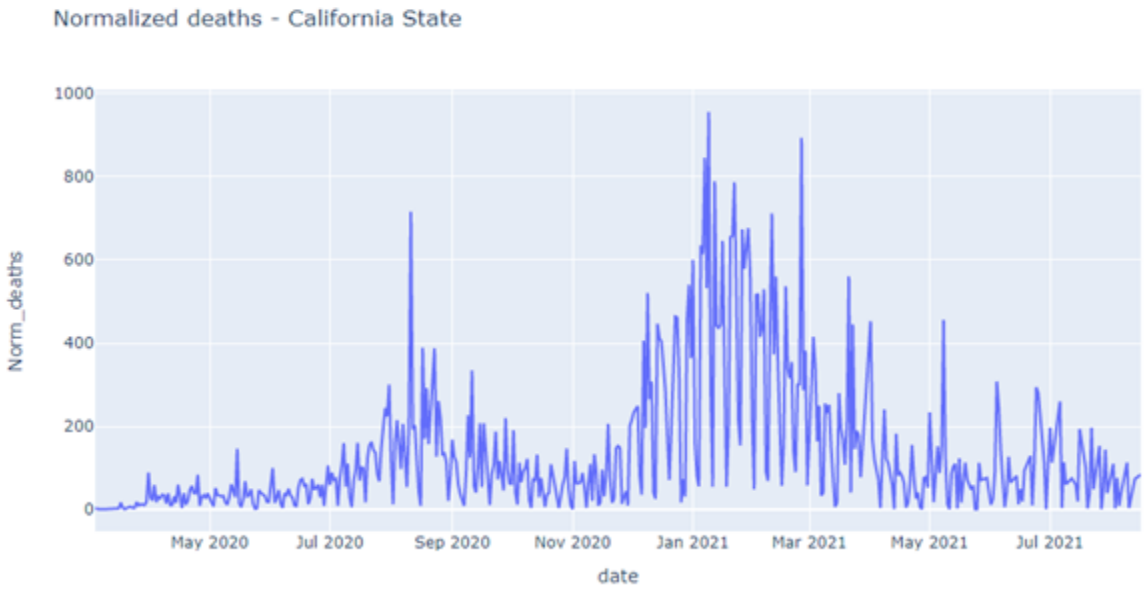
Median of all states

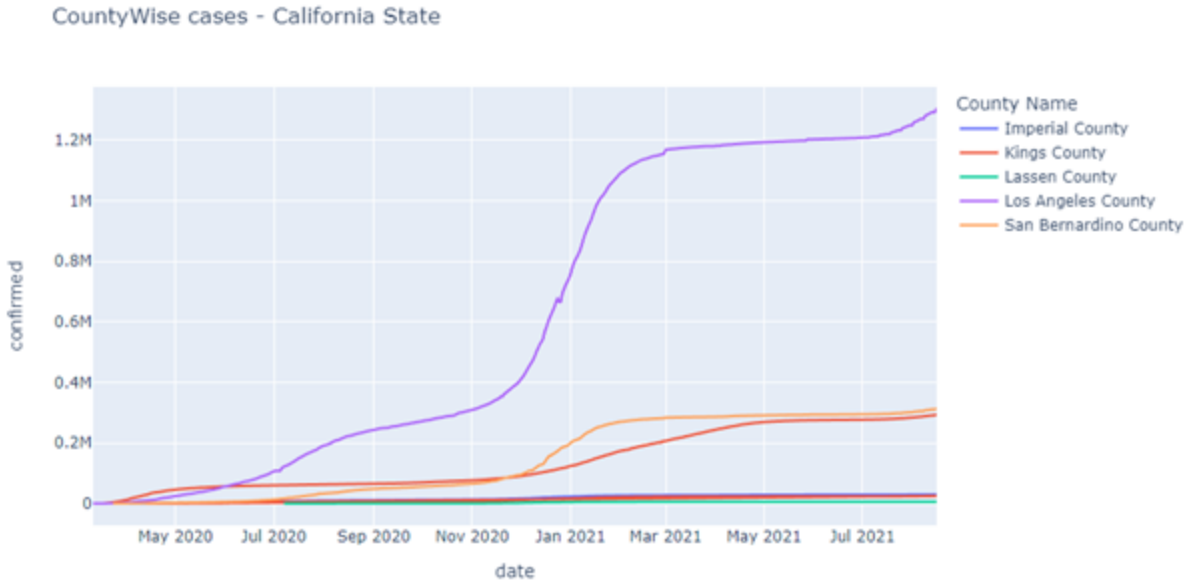
	State	high_cases	high_deaths
0	AK	683.0	0.0
1	AL	978.0	122.0
2	AR	912.0	91.0
3	AZ	678.0	159.0
4	CA	514.0	67.0
5	CO	657.0	70.0
6	CT	503.0	70.0
7	DC	697.0	98.0
8	DE	841.0	128.0
9	FL	952.0	136.0
10	GA	715.0	136.0
11	HI	210.0	0.0
12	IA	930.0	92.0
13	ID	914.0	83.0

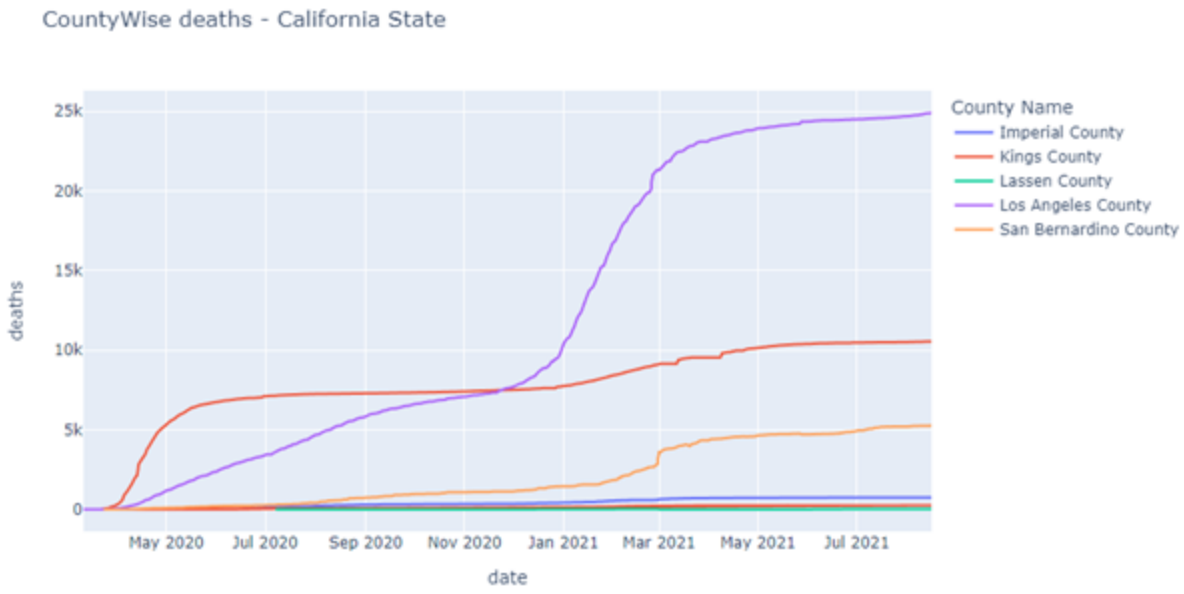
Mode of all states is 0

Since California state has highest no of cases and deaths per day. This can be because of the population in California, it is an IT hub, and the state is also very big so there might be more population because of all these factors. Selecting it and normalizing.




Normalized deaths - California State
Norm_deaths
1000
800
600
400
200
0
May 2020
Jul 2020
Sep 2020
Nov 2020
Jan 2021
Mar 2021
May 2021
Jul 2021
date


CountyWise cases - California State
confirmed
1.2M
1M
0.8M
0.6M
0.4M
0.2M
0
May 2020
Jul 2020
Sep 2020
Nov 2020
Jan 2021
Mar 2021
May 2021
Jul 2021
date
County Name
Imperial County
Kings County
Lassen County
Los Angeles County
San Bernardino County



Task – 2: Member

```
*** Probability Statistics of California State ***
Mean of CA state :  [1708267.95287958]
Variance of CA state :  [1708267.95287958]
skewness of CA state :  [0.00076511]
kurtosis of CA state :  [5.85388257e-07]
Maximum value :  4033071
Minimum value :  0

*** Probability Statistics of Virginia State ***
Mean of VA state :  [301920.21291449]
Variance of VA state :  [301920.21291449]
skewness of VA state :  [0.00181993]
kurtosis of VA state :  [3.31213333e-06]
Maximum value :  723727
Minimum value :  0

*** Probability Statistics of Oregon State ***
Mean of OR state :  [85839.65968586]
Variance of OR state :  [85839.65968586]
skewness of OR state :  [0.00341315]
kurtosis of OR state :  [1.16496268e-05]
Maximum value :  242843
Minimum value :  0

*** Probability Statistics of Newyork State ***
Mean of NY state :  [951133.61780105]
Variance of NY state :  [951133.61780105]
skewness of NY state :  [0.00102537]
kurtosis of NY state :  [1.05137699e-06]
Maximum value :  2192336
Minimum value :  0

*** Probability Statistics of South Carolina State ***
Mean of SC state :  [267457.63176265]
Variance of SC state :  [267457.63176265]
skewness of SC state :  [0.00193363]
kurtosis of SC state :  [3.73890995e-06]
Maximum value :  667352
Minimum value :  0

*** Probability Statistics of Michigan State ***
Mean of MI state :  [390640.69982548]
Variance of MI state :  [390640.69982548]
skewness of MI state :  [0.00159997]
kurtosis of MI state :  [2.55989711e-06]
Maximum value :  1003555
Minimum value :  0
```

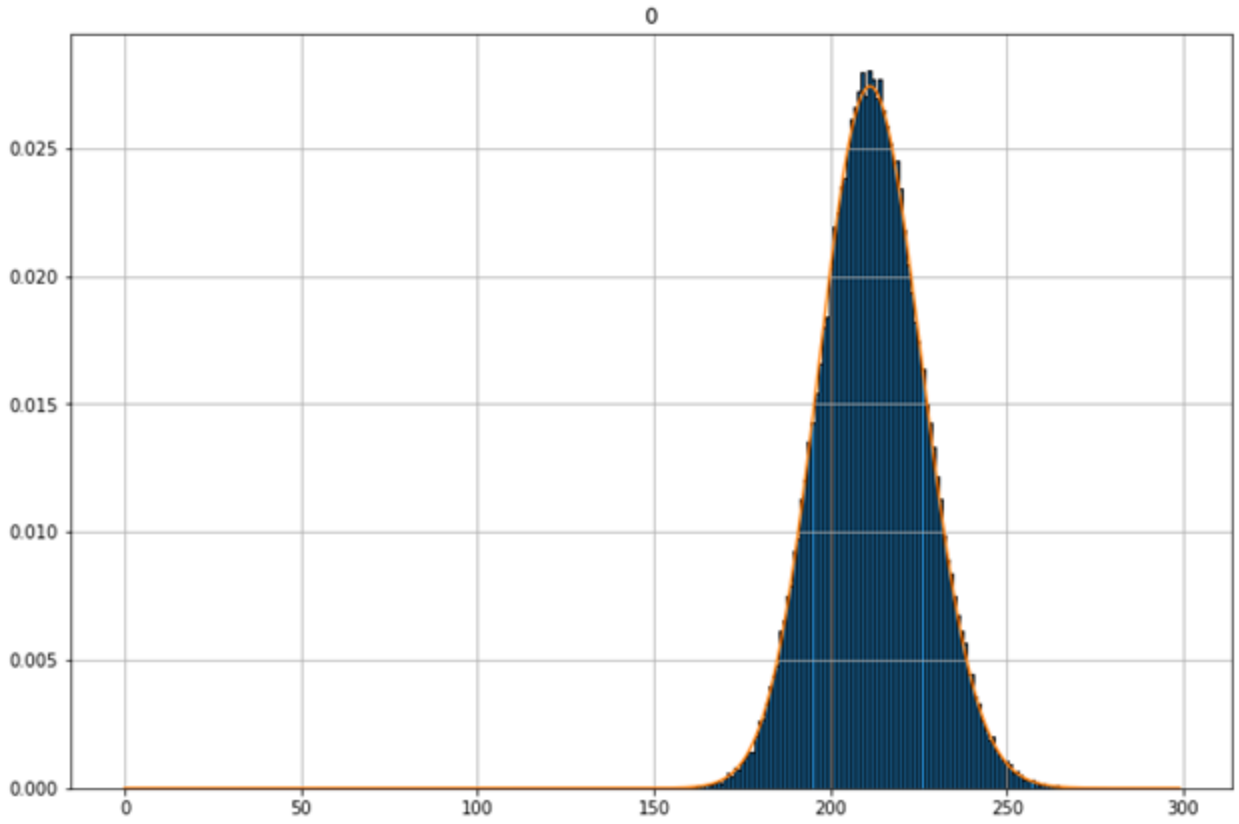
Based on the above data we can see that California has the highest mean, variance and kurtosis than any other state. California is a big state and due to its high population, the number of cases are also high. Since the skew value for all states is near to zero all the states have symmetrical data. Kurtosis shows us whether the data has outliers or not. From the above data we can see that California data has many outliers compared to other states. After California South Carolina has more outliers. Since we have positive kurtosis and are above 0 this data has many outliers in every state. among them California has the highest¶

Plotted Poisson distribution of new cases and deaths for all the above states.

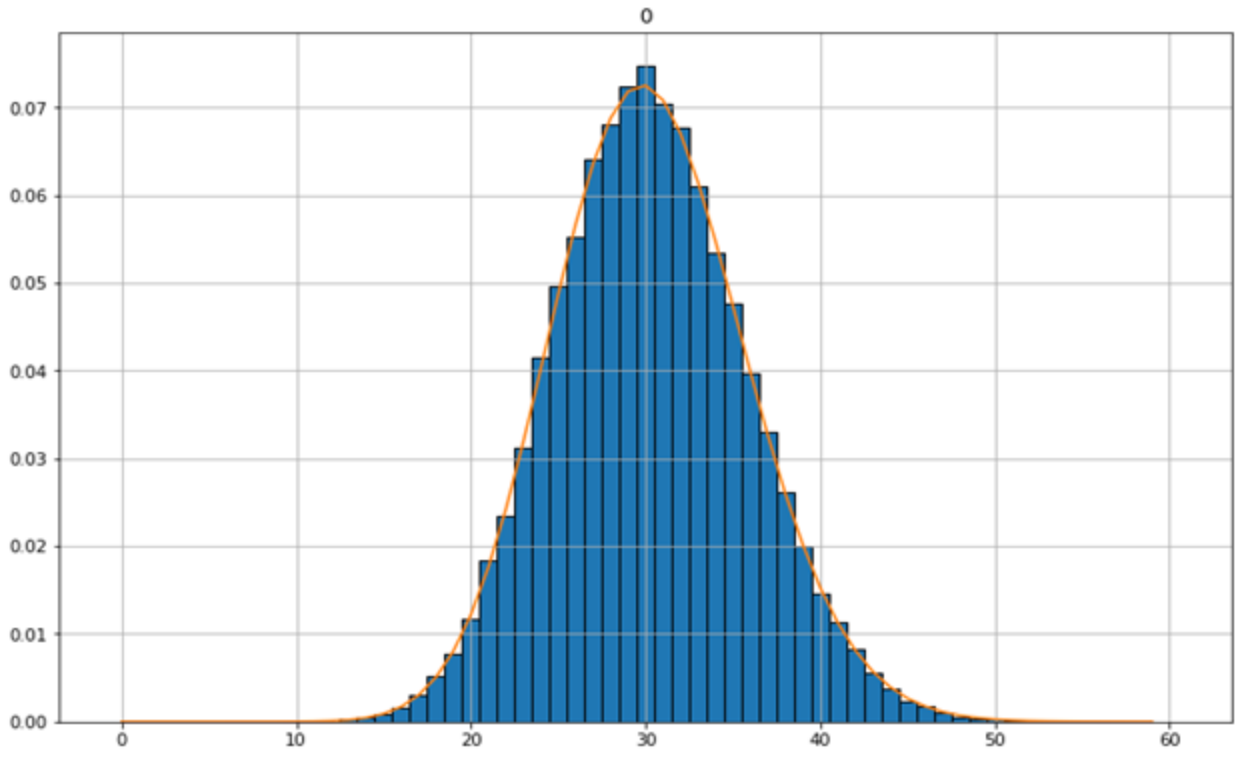
For NC state, took 5 counties and plotted Poisson distribution. One of the county's plots is as below.

Plotted the same graphs to 5 other counties of NC state

Alamance County Poisson distribution – New cases



Alamance County Poisson distribution – New Deaths



Performed correlation between Covid-19 and Presidential Election data.

The three variables that I have chosen to compare correlation are total votes per state, democratic party votes and republic party votes.

	high_cases	high_deaths	total_votes	population	DEM_votes	REP_votes
high_cases	1.000000	0.952932	0.381860	0.429731	0.269147	0.526439
high_deaths	0.952932	1.000000	0.351641	0.398618	0.246115	0.489057
total_votes	0.381860	0.351641	1.000000	0.990703	0.981615	0.967584
population	0.429731	0.398618	0.990703	1.000000	0.972677	0.957427
DEM_votes	0.269147	0.246115	0.981615	0.972677	1.000000	0.906263
REP_votes	0.526439	0.489057	0.967584	0.957427	0.906263	1.000000

Hypothesis 1: Does a higher voting population in a state result in higher covid cases? The thought behind the hypothesis is that states with higher voting populations have citizens who are willing to go out of social distance to perform duties like vote.

Hypothesis 2: Based on the party propaganda during the elections - we could hypothesize that a higher democratic voting population in a state could result in a lower number of cases.

Hypothesis 3: Based on the party propaganda during the elections- we could hypothesize that a higher republican voting population in a state could result in a higher number of cases.

We can see that democratic votes have a very low correlation with increase in cases whereas republican votes have a moderate correlation with high cases in a state. Using this comparison we can supposedly say that a high republican voter presence in a state appears to show some correlation with the increasing cases. From this observation we can supposedly say that our hypotheses 2 and 3 are true with a moderate confidence level.

The total votes in a state do not seem to be well correlated with the increased cases. This tells us that the high voting population in a state does not result in increased cases. This proves that hypothesis 1 is false.