COVID-19 Data Analysis Stage II Report

Jason Manning
Francis Perez
Jamison Valentine
Raiana Zaman
Xinrui (Sam) Zhang

CSC 405-01: Data Science

University of North Carolina Greensboro

Fall: 2020

Task 1:

Team:

Part 1 (Jamison) .

- *Compare the weekly statistics (mean, median, mode) for number of new cases and deaths across US. You are calculating mean (rounded to integer value) number of new cases and per week and then calculating (mean, median, mode) for all week taken together.*

```
Confirmed Cases Mean: 29,450
Confirmed Cases Median: 28,011
Confirmed Cases Mode: 1
Deaths Mean: 794
Deaths Median: 741
Deaths Mode: 0
```

The mean and median number of new cases were fairly close, with the median being less than the mean, indicating a slight right-skew in the data. Such was also the case with the number of new reported deaths. The respective modes the most frequently occuring observations were 1 new confirmed case and 0 deaths across the US.

Part 2 (Raiana).

list of countries that we are going to compare

1. Brazil in South America
2. Japan in Asia
3. Mexico in North America
4. Nigeria in Africa
5. Russia in Europe

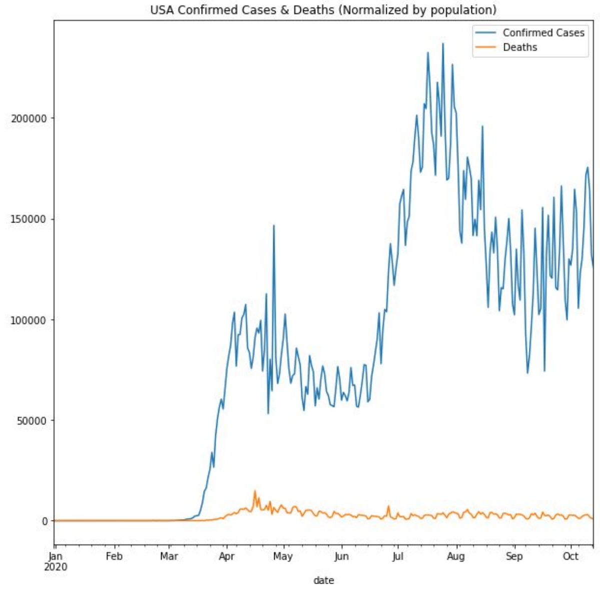
These countries were chosen as a way to represent each of the world land area & similar populations

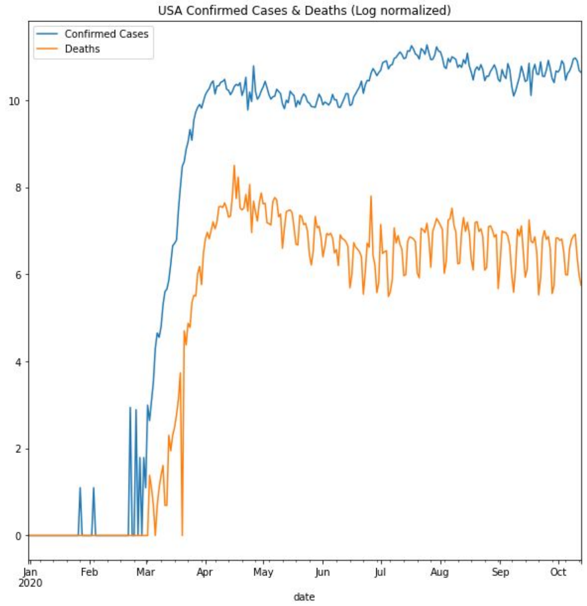
	Confirmed Cases Mean	Confirmed Cases Median	Confirmed Cases Mode	Deaths Mean	Deaths Cases Median	Deaths Cases Mode
Brazil	82,495	64,338	0	2,431	2,954	0
Japan	2,463	1,152	0	44	24	0
Mexico	22,103	22,364	0	2,235	2,120	0
Nigeria	1,011	774	0	18	10	0
Russia	32,173	36,198	0	545	698	0

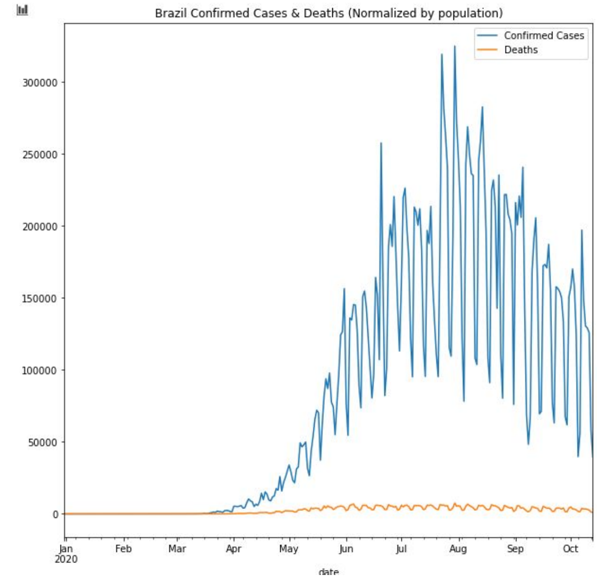
Part 3 ***(Xinrui Zhang)***.

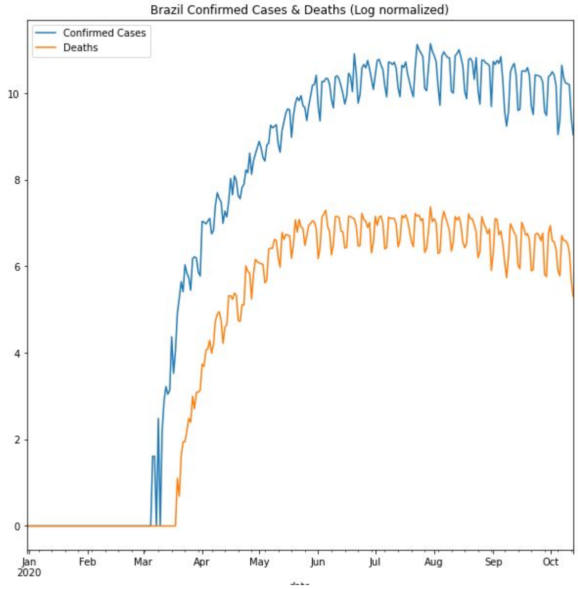
These are the confirmed cases and deaths data with normalized and log normalized figures.

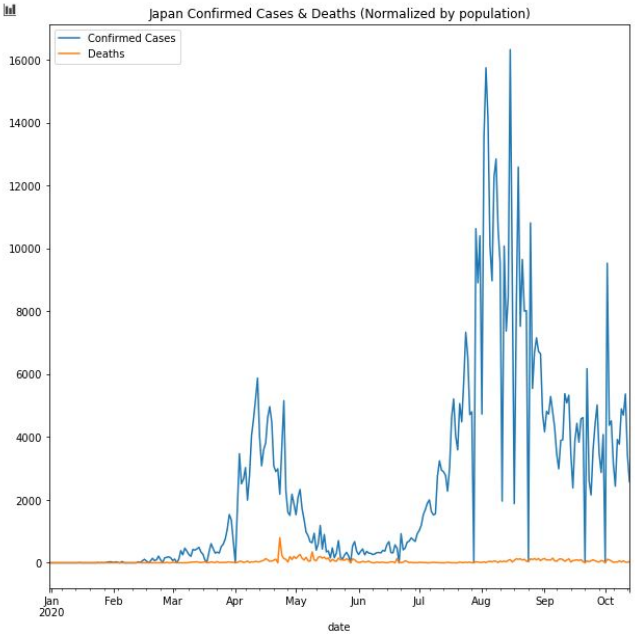
From the figure we can detect that the USA, Brazil and Russia have relatively larger increases in both confirmed cases and deaths. And for Mexico, there is a sudden increase and decrease in October. Japan is doing a really good job on controlling the spreading of diseases. And for Nigeria, there might be missing data which caused the confirmed cases and deaths to be incredibly low.

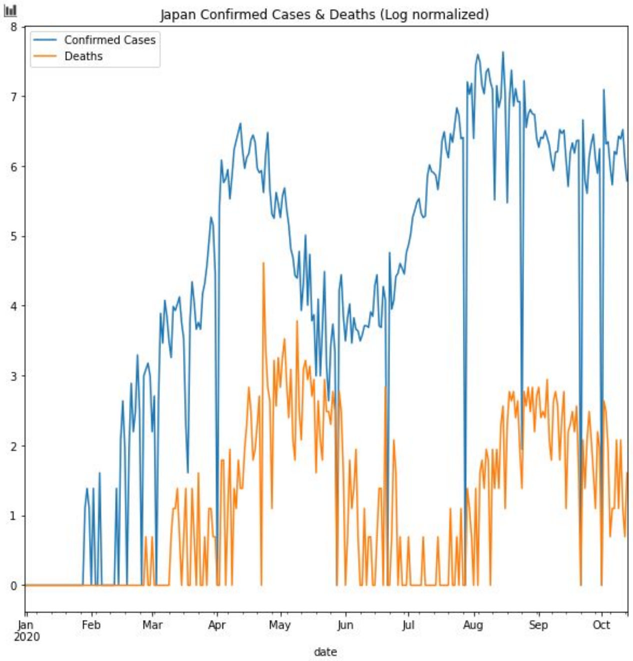

USA Confirmed Cases & Deaths (Normalized by population)

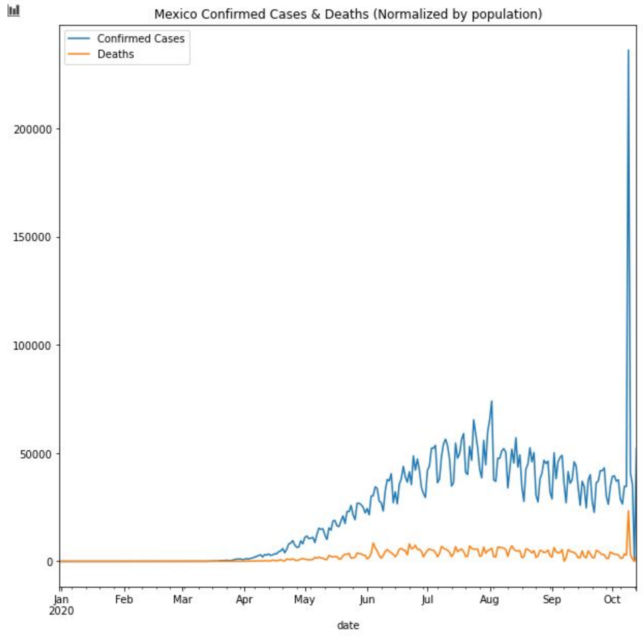

USA Confirmed Cases & Deaths (Log normalized)

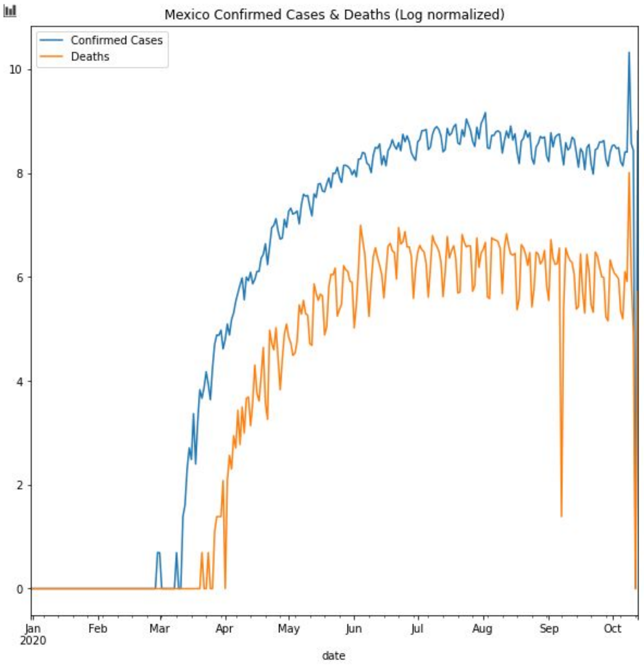

Brazil Confirmed Cases & Deaths (Normalized by population)

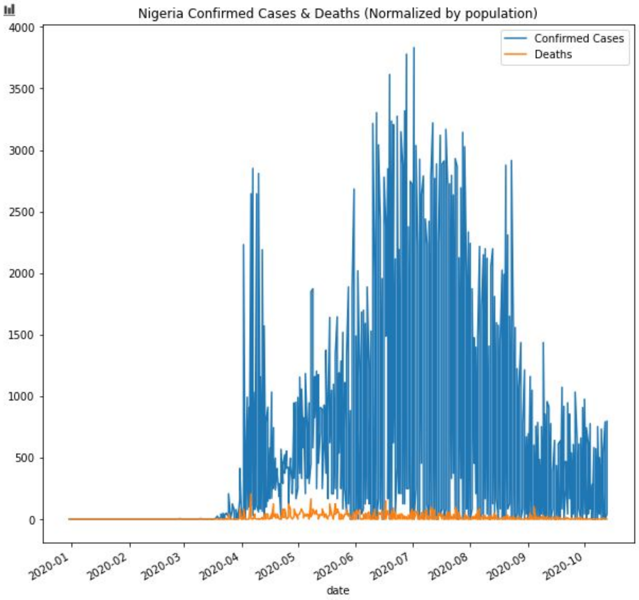

Brazil Confirmed Cases & Deaths (Log normalized)

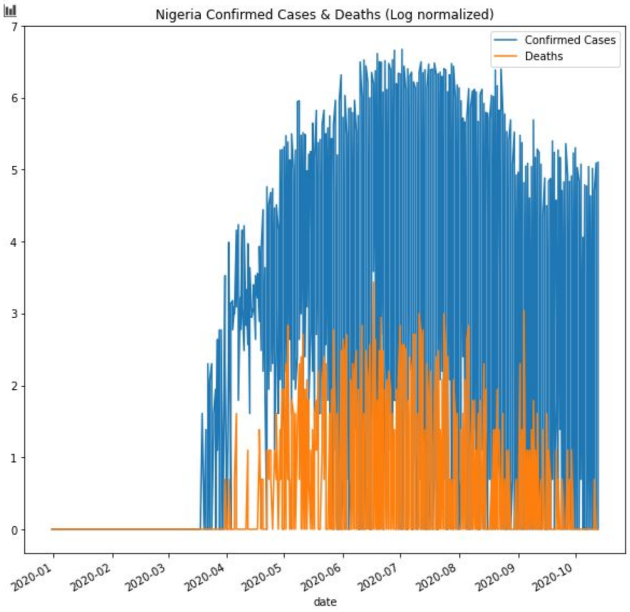

Japan Confirmed Cases & Deaths (Normalized by population)
Confirmed Cases
Deaths
16000
14000
12000
10000
8000
6000
4000
2000
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
date

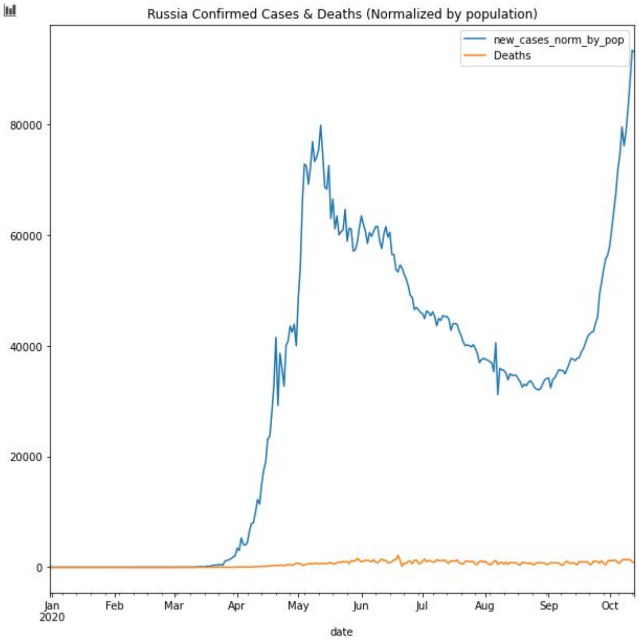

Japan Confirmed Cases & Deaths (Log normalized)
Confirmed Cases
Deaths
8
7
6
5
4
3
2
1
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
date

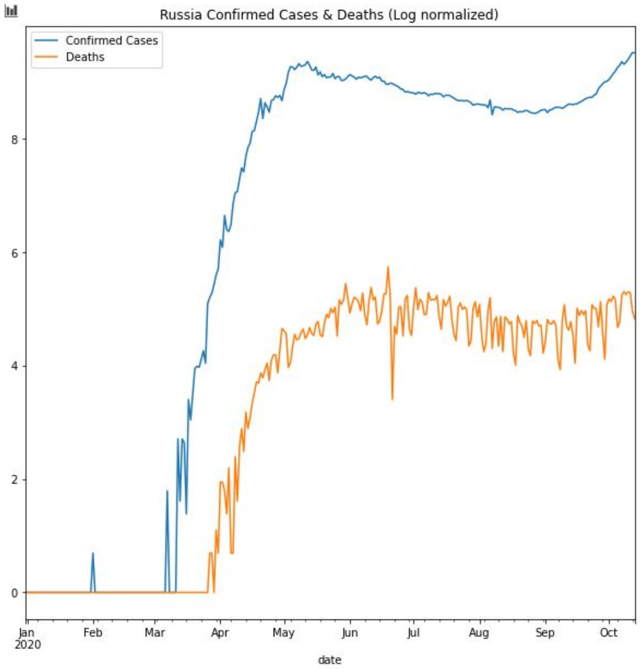

Mexico Confirmed Cases & Deaths (Normalized by population)
Confirmed Cases
Deaths
200000
150000
100000
50000
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
date


Mexico Confirmed Cases & Deaths (Log normalized)
Confirmed Cases
Deaths
10
8
6
4
2
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
date


Nigeria Confirmed Cases & Deaths (Normalized by population)
Confirmed Cases
Deaths
4000
3500
3000
2500
2000
1500
1000
500
0
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
date


Nigeria Confirmed Cases & Deaths (Log normalized)
Confirmed Cases
Deaths
7
6
5
4
3
2
1
0
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
2020-09
2020-10
date


Russia Confirmed Cases & Deaths (Normalized by population)
new_cases_norm_by_pop
Deaths
80000
60000
40000
20000
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
date


Russia Confirmed Cases & Deaths (Log normalized)
Confirmed Cases
Deaths
8
6
4
2
0
Jan
2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
date

Part 4: Identify peak week of the cases and deaths in US and other countries

The following countries were chosen for their similarity in population to the United States. Listed below are the statistics for the peak weeks for each of the following countries: USA, Brazil, Japan, Mexico, Nigeria, Russian. Based on the data below, Russian appears to have the worst week overall for new cases and deaths. Nigeria appears to have fared better on their worst week than the other countries.

```
USA     - Peak Week - New Cases:30  | New Deaths:  16
Brazil  - Peak Week - New Cases:30  | New Deaths:  30
Japan   - Peak Week - New Cases:32  | New Deaths:  17
Mexico  - Peak Week - New Cases:41  | New Deaths:  26
Nigeria - Peak Week - New Cases:26  | New Deaths:  25
Russia  - Peak Week - New Cases:41  | New Deaths:  41
```

<u>Members:</u>

Jamison

Xinrui Zhang

Francis Perez:

Part 1 (Weekly Statistics For North Carolina):

	Mean	Median	Mode
Confirmed New Cases	905	1123	0
Deaths	14	16	0

Given that the Mean and Median are not the same, we can say the neither confirmed cases or deaths follow a normal distribution.

Part 2 (Compare the data against other states):

States: { Georgia, Michigan, New Jersey, Tennessee, Washington }
These states were picked based on their similar population with North Carolina.

		Mean	Median	Mode
Michigan	Confirmed New Cases	591	651	0
	Deaths	27	11	0
Georgia	Confirmed New Cases	1175	786	0
	Deaths	26	29	0
Tennessee	Confirmed New Cases	1227	735	0
	Deaths	16	11	0
Washington	Confirmed New Cases	472	486	0
	Deaths	11	10	0
New Jersey	Confirmed New Cases	906	431	0
	Deaths	67	9	0

Part 3 (Identify counties with high cases and death rates for North Carolina):

Top-5 Counties with largest confirmed cases.
*Confirmed & Deaths have been normalize by population

County	Population	**Confirmed (Hightest)**	Deaths
Duplin County	58,741	**4,549**	92
Scotland County	34,823	**4,414**	63
Robeson County	130,625	**4,301**	66
Montgomery County	27,173	**4,203**	140
Sampson County	63,531	**4,007**	44

Top-5 Counties with largest deaths.
*Confirmed & Deaths have been normalize by population

County	Population	Confirmed	**Deaths (Hightest)**
Jones County	9,419	1,837	**159**
Hertford County	23,677	3,569	**144**
Montgomery County	271,173	4,203	**140**
Columbus County	55,508	2,836	**106**
Vance County	44,535	2,576	**103**

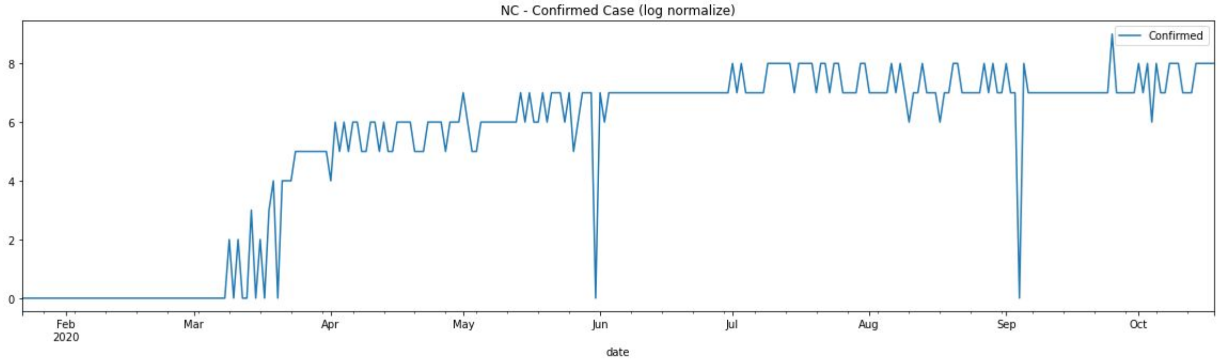
Top-5 Infected Counties in North Carolina.
*These counties were chosen From the previous part of highest cases and deaths listing, round robin from each list until 5 total. (normalize by population)

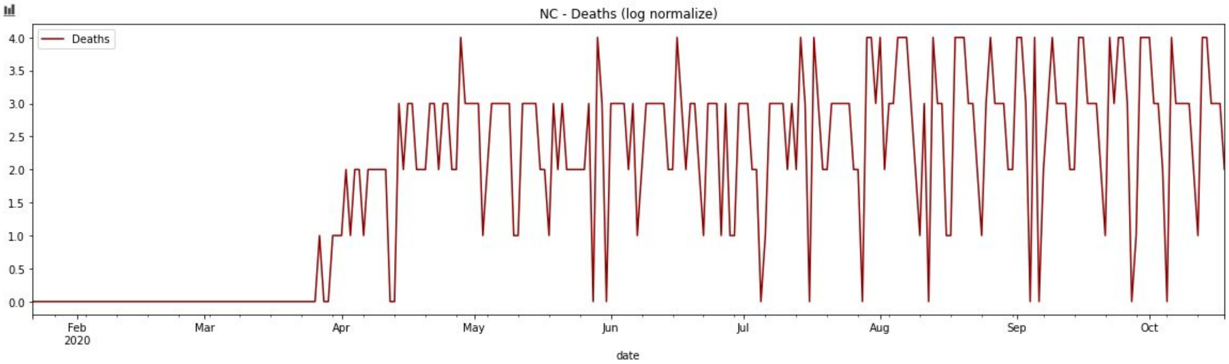
County (alpha order)	Population	Confirmed	Deaths (Hightest)
Duplin County	58,741	4,549	92
Hertford County	23,677	3,569	144
Jones County	9,419	1,837	159
Robeson County	130,625	4,301	66
Scotland County	34,823	4,414	63

**** Looking at the data, counties with the lowest population are the ones with the most infection by population.**

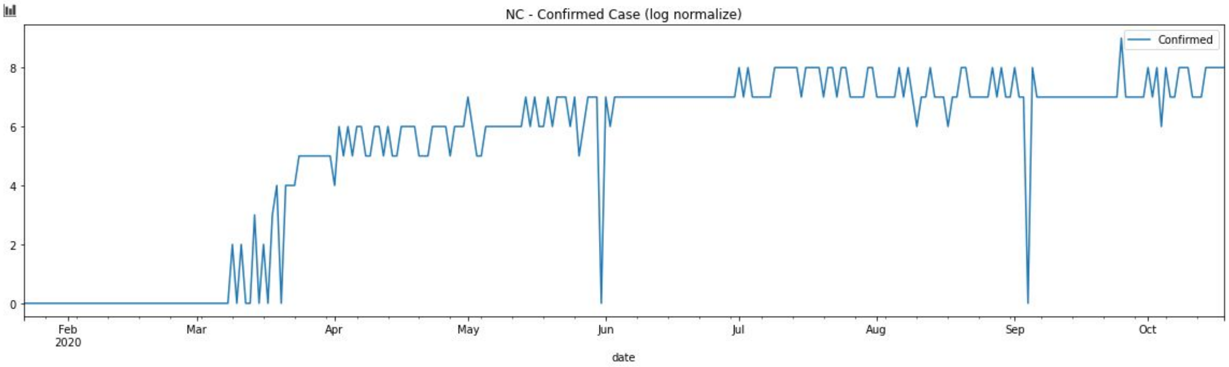
Part 4 Plot daily trends (cases and deaths, new cases) of state and top 5 infected counties.

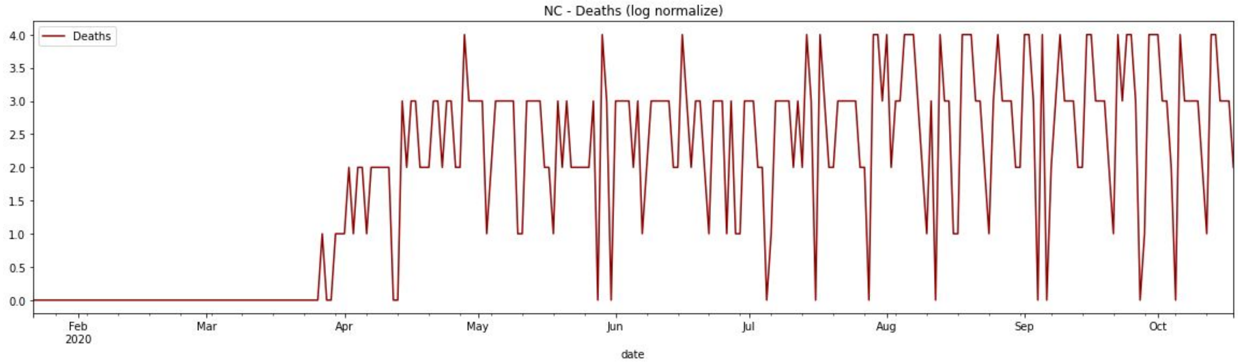
Daily Trends For State - North Carolina (normalize by population)



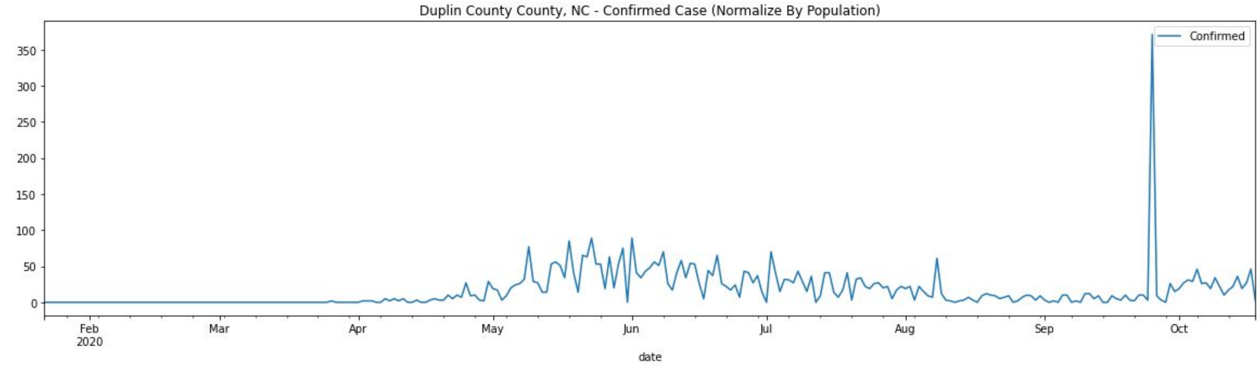


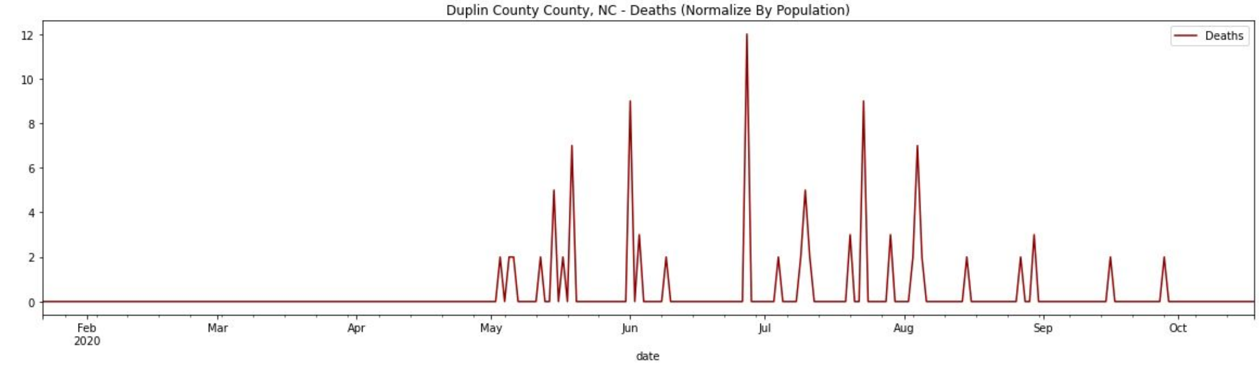
Daily Trends For State - North Carolina (log normalize)

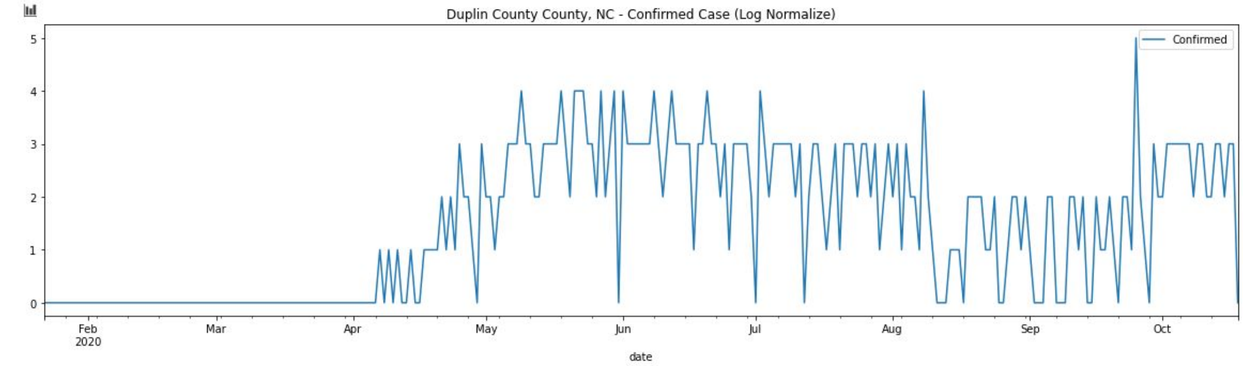


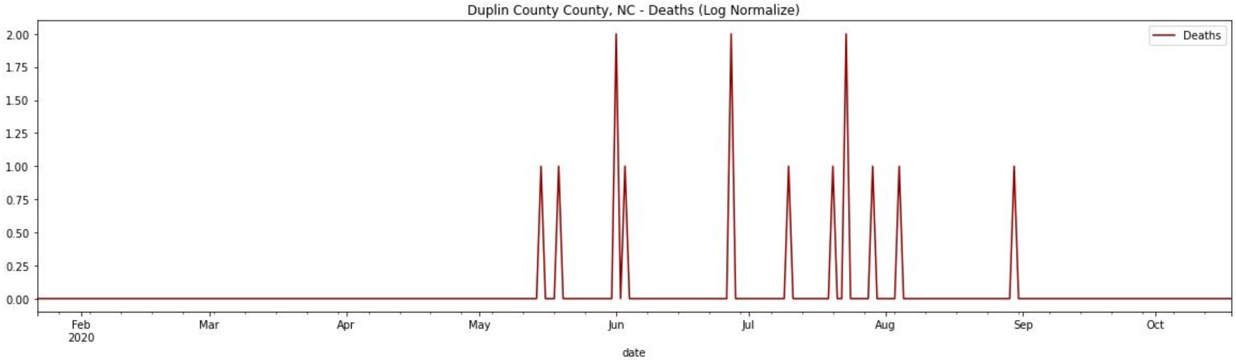


Plot Data For Duplin County

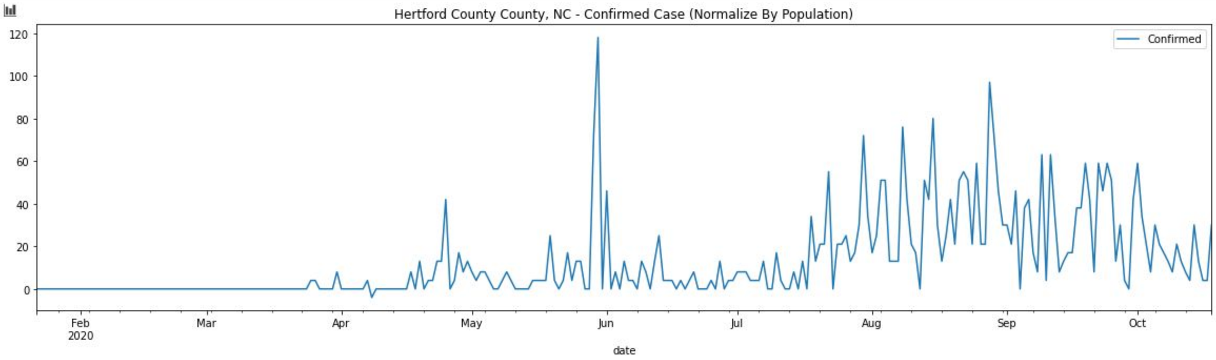


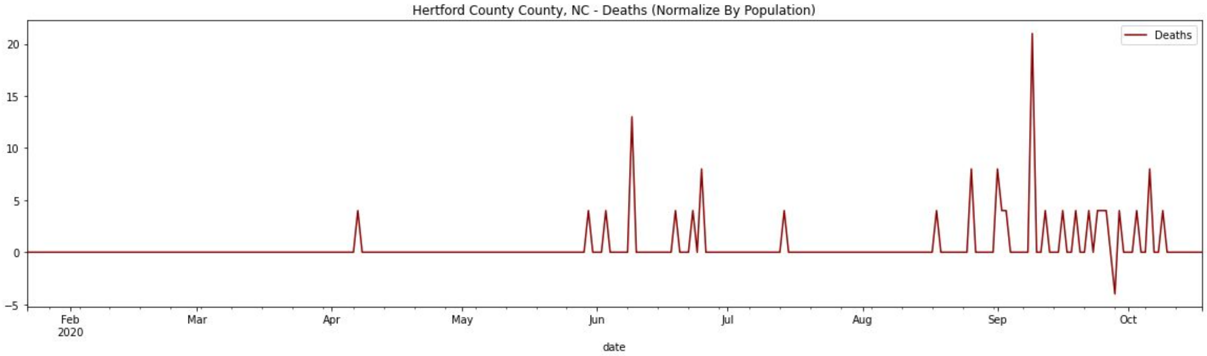


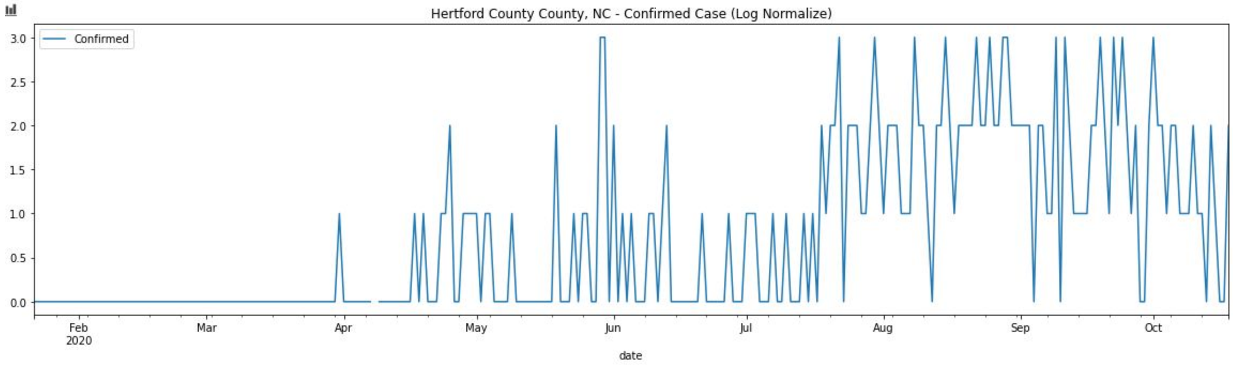


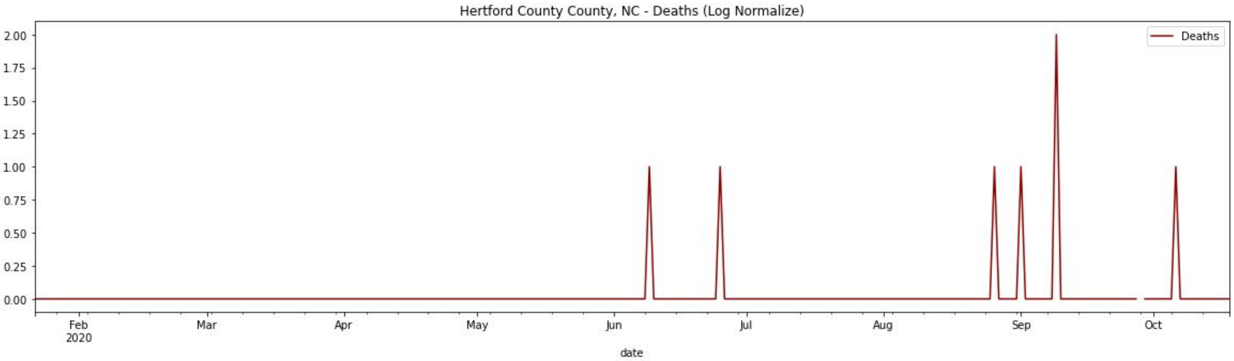


Plot Data For Hertford County

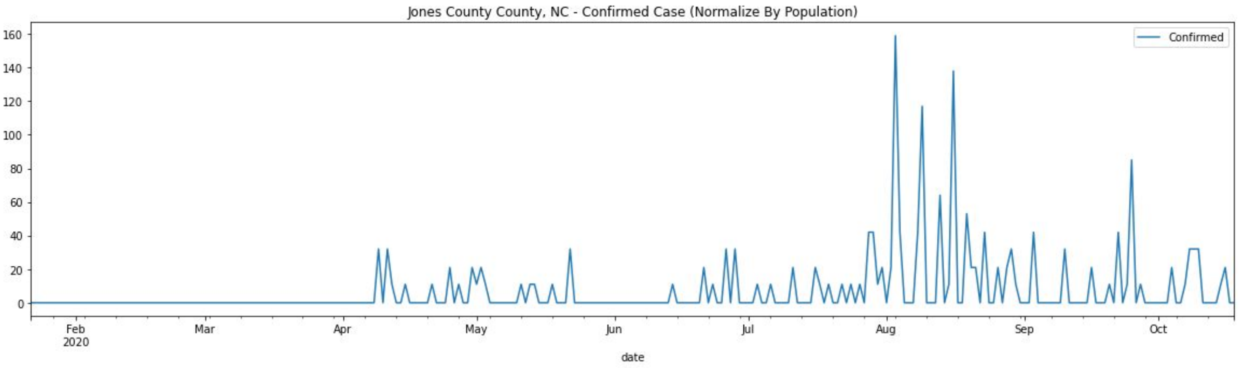

Hertford County County, NC - Confirmed Case (Normalize By Population)

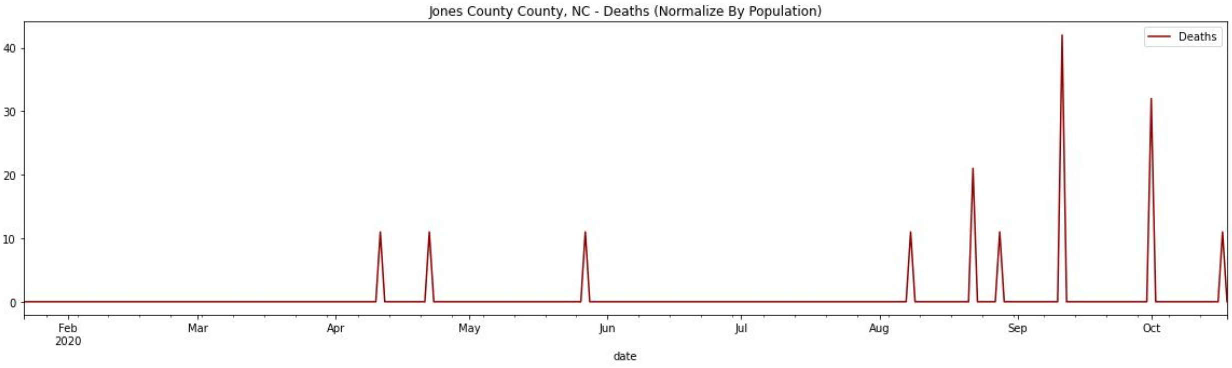

Hertford County County, NC - Deaths (Normalize By Population)

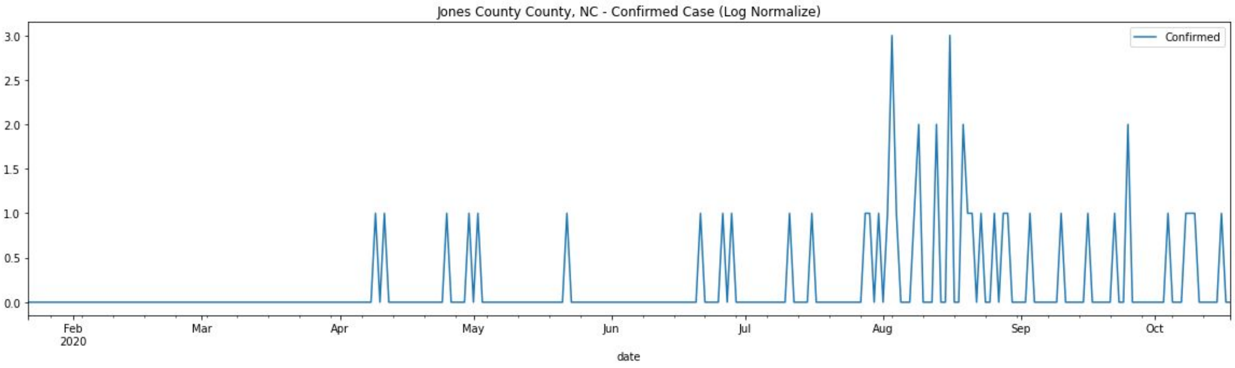

Hertford County County, NC - Confirmed Case (Log Normalize)

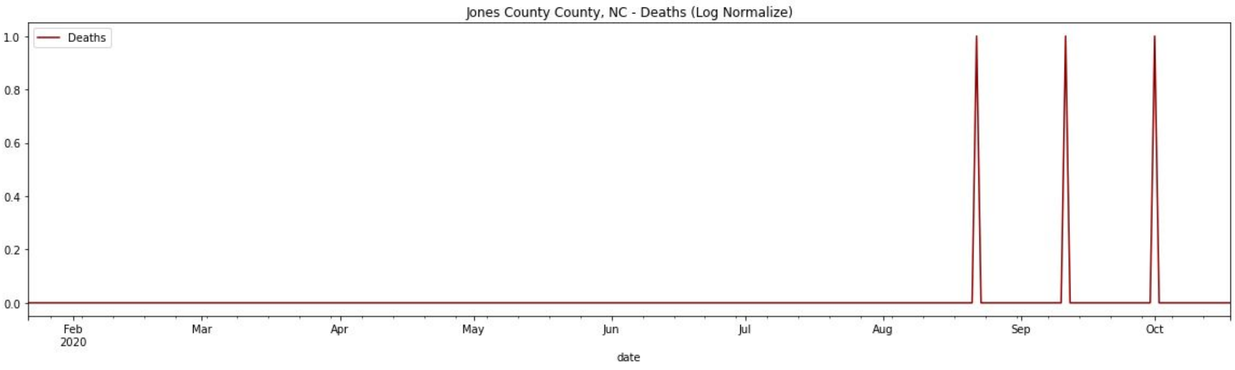

Hertford County County, NC - Deaths (Log Normalize)

Plot Data For Jones County

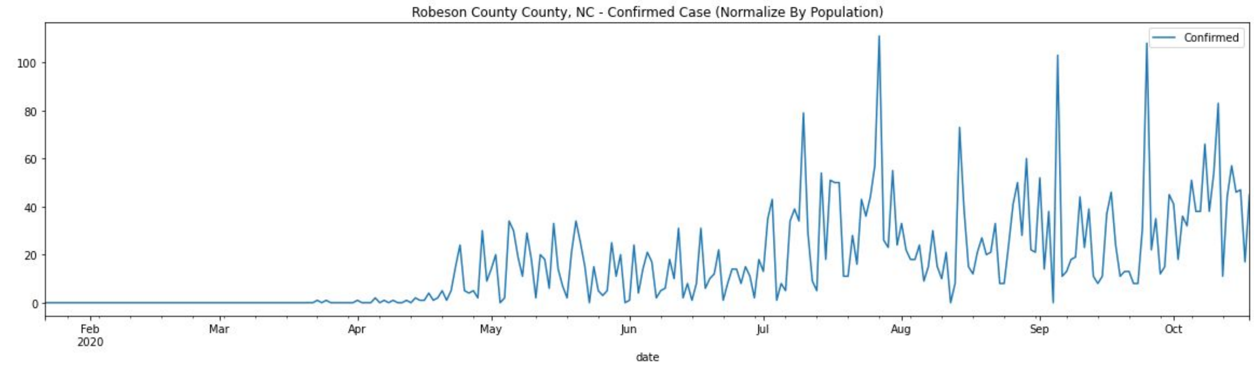


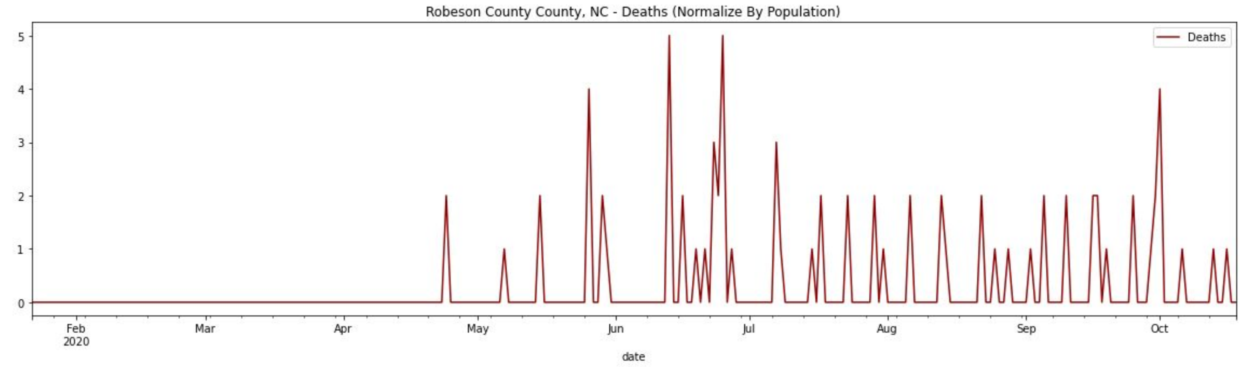


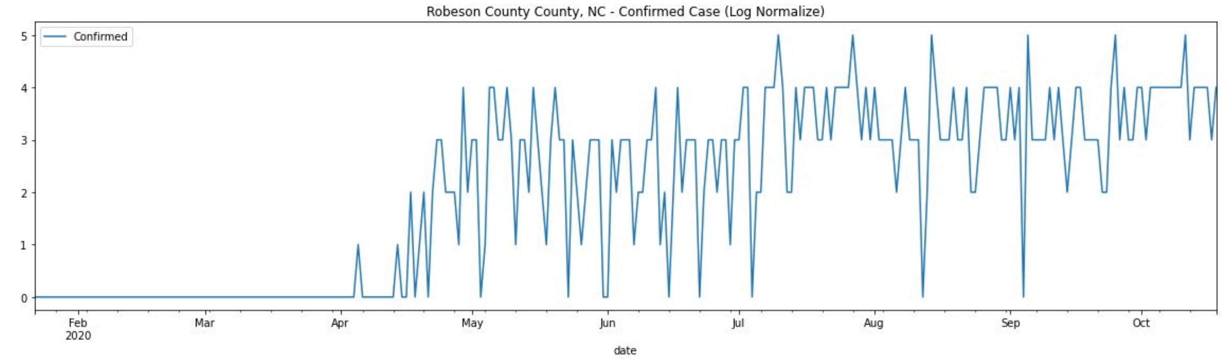


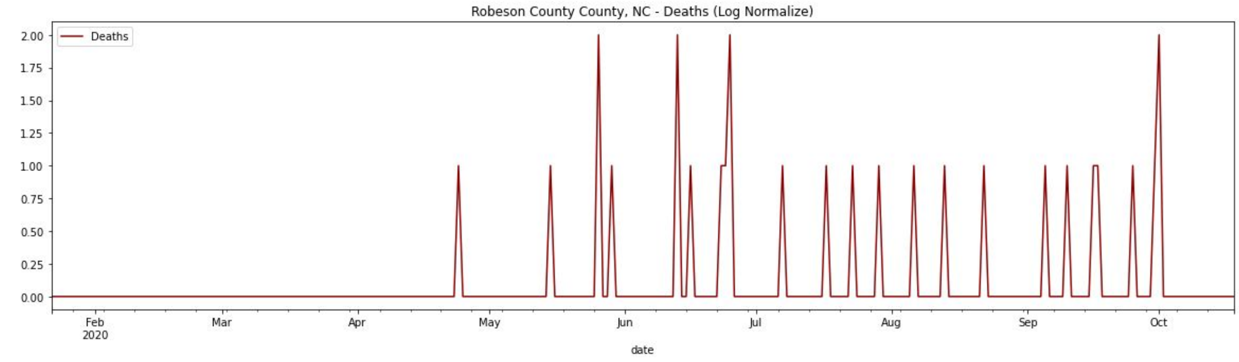


Plot Data For Robeson County

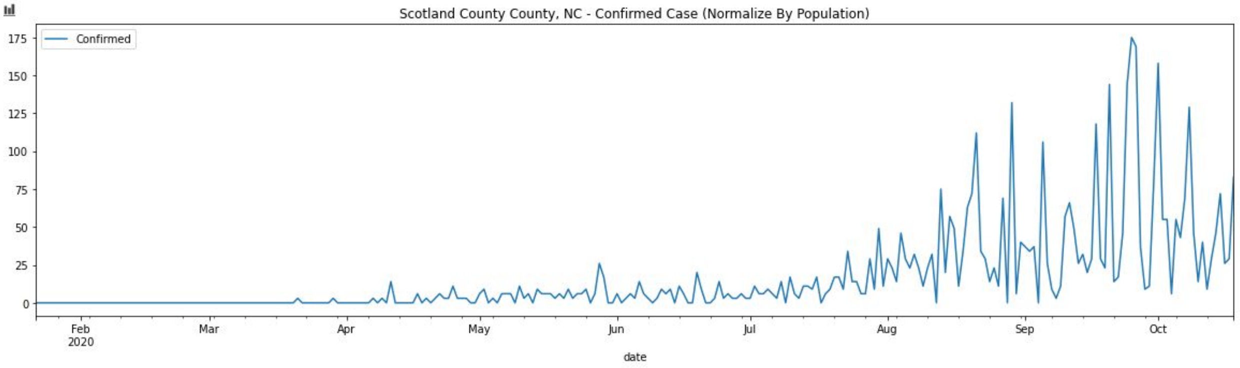




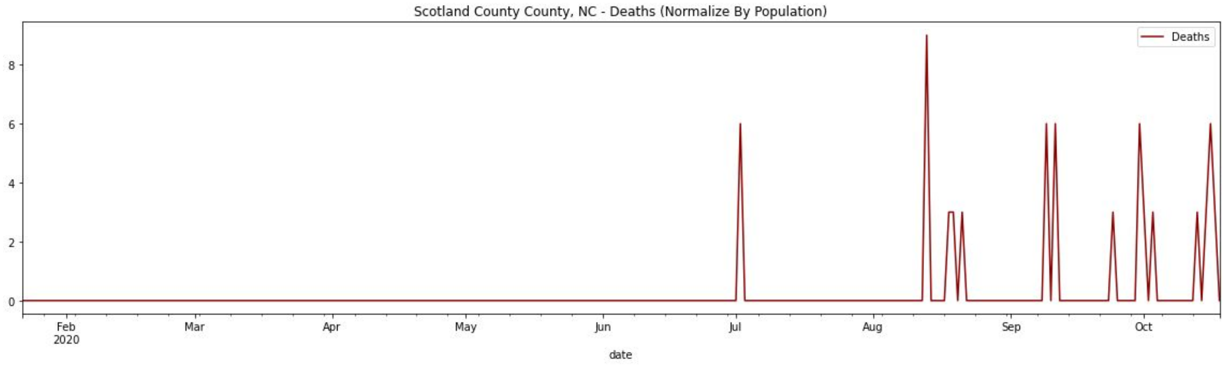




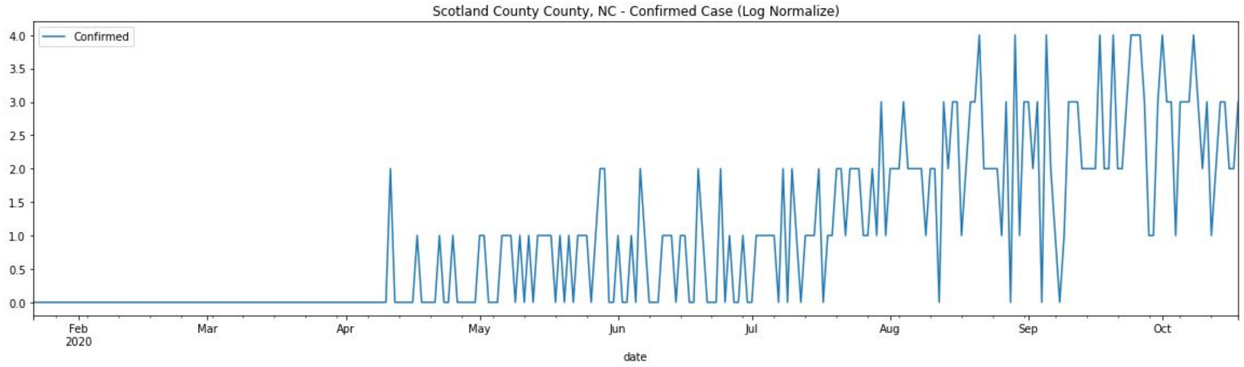
Plot Data For Scotland County



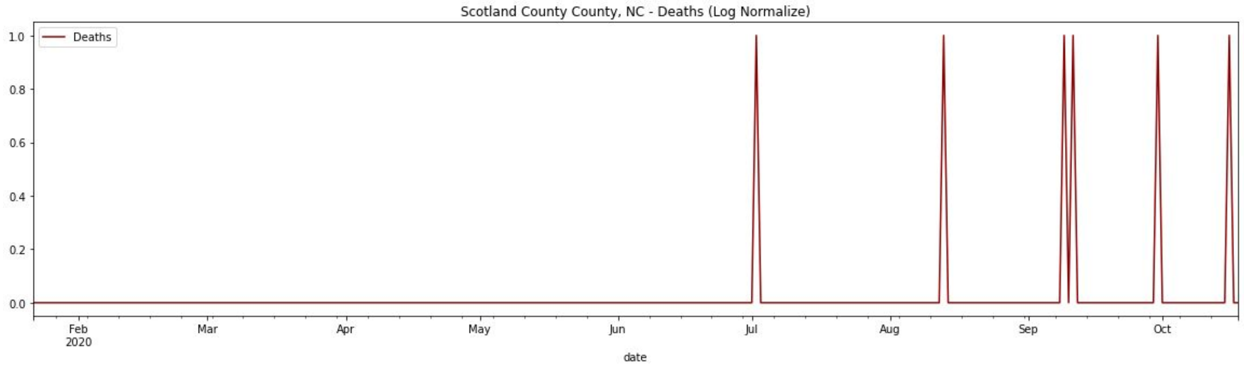
Scotland County County, NC - Confirmed Case (Normalize By Population)



Scotland County County, NC - Deaths (Normalize By Population)

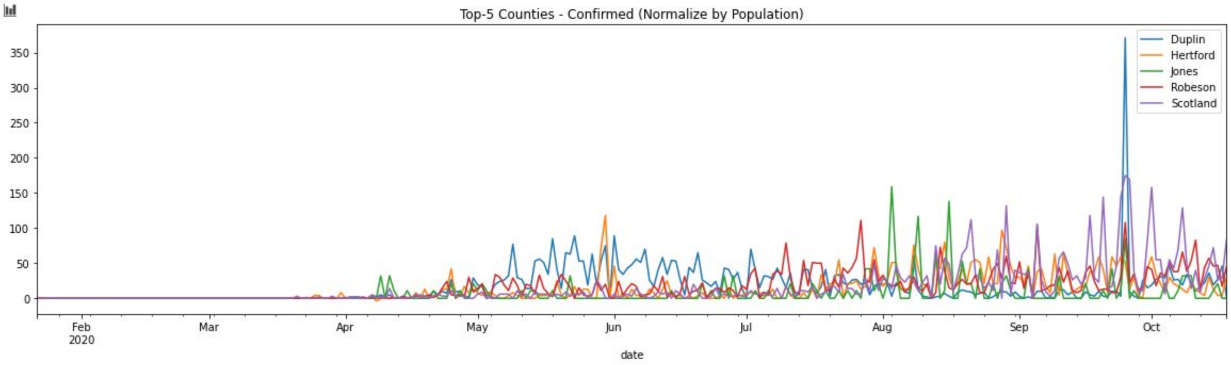


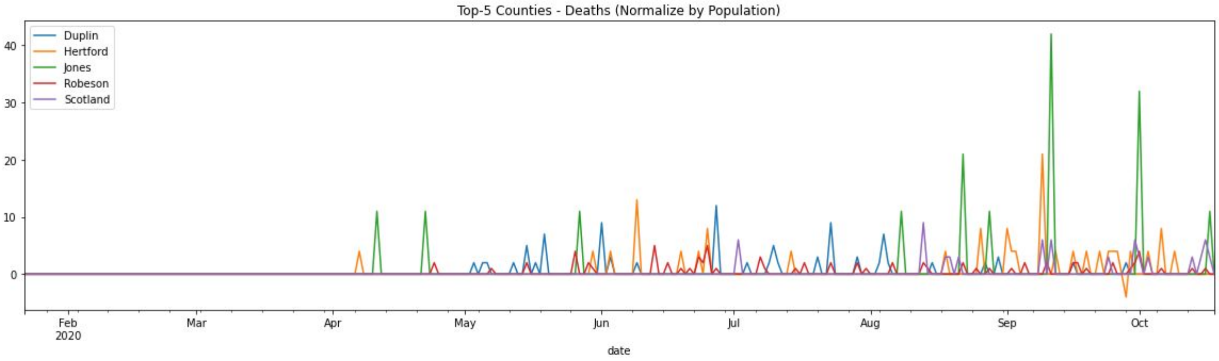
Scotland County County, NC - Confirmed Case (Log Normalize)

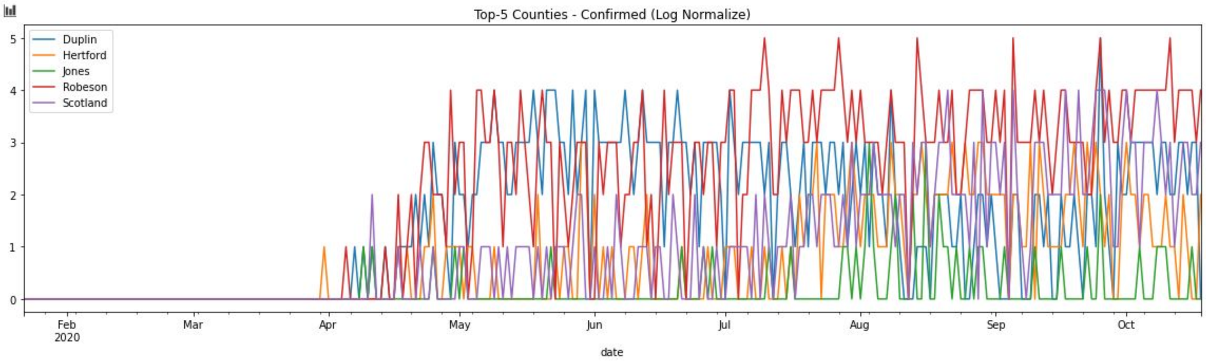


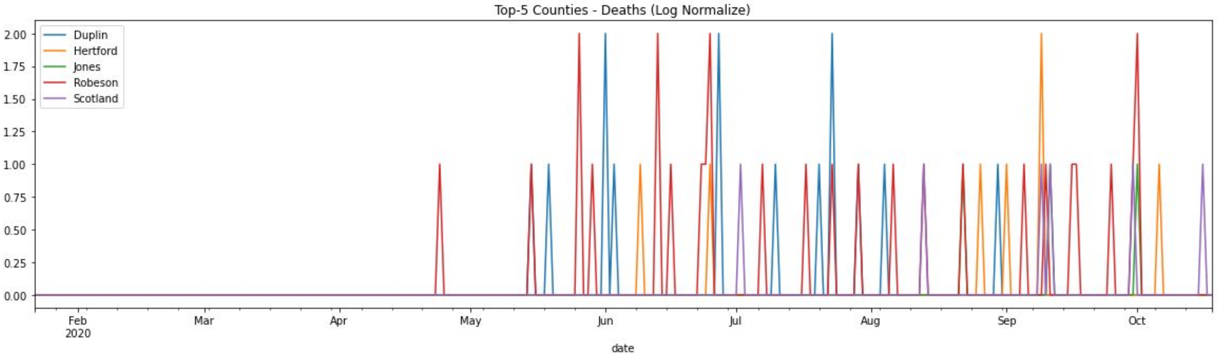
Scotland County County, NC - Deaths (Log Normalize)

Plot Top-5 Counties Together - Confirmed Cases & Deaths









Francis Pere - END OF Task 1

Jason Manning

Part 1 Weekly Statistics for Washington per 100000:

	Mean	Median	Mode
Confirmed Cases	232	233	0
Deaths	5	5	0

The new cases have a slightly higher median than mean, meaning the data has a left skewed distribution. The deaths appear to have a normal distribution since the values are the same.

Part 2 Compare the data against other states per 100000:

The following states were chosen for their proximity in population to Washington:

Massachusetts, New Jersey, Arizona, Tennessee, and Virginia.

		Mean	Median	Mode
Massachusetts	Confirmed Cases	232	233	0
	Deaths	26	11	0

New Jersey	**Confirmed Cases**	**398**	**241**	**0**
	Deaths	**34**	**4**	**0**
Virginia	**Confirmed Cases**	**352**	**427**	**0**
	Deaths	**7**	**7**	**0**
Arizona	**Confirmed Cases**	**579**	**318**	**0**
	Deaths	**15**	**13**	**0**
Tennessee	**Confirmed Cases**	**600**	**347**	**0**
	Deaths	**8**	**5**	**0**

Part 3 Identify counties with high cases and death rates for Washington:

Top 5 counties with most confirmed cases per 10,000 people

County Name	**Population**	**Total Cases**	**Cases per 10,000**
Franklin County	**95222**	**4637**	**487**
Yakima County	**250873**	**11708**	**467**
Adams County	**19983**	**908**	**454**

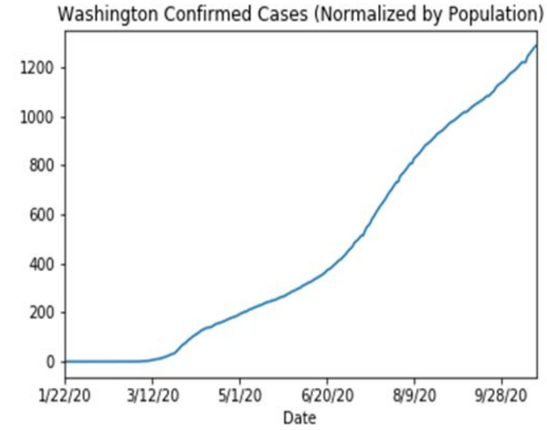
Whitman County	**50104**	**1670**	**333**
Grant County	**97733**	**3257**	**333**

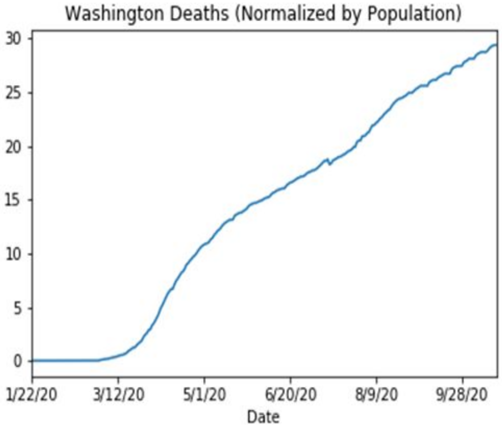
Top 5 counties with most deaths per 10,000 people

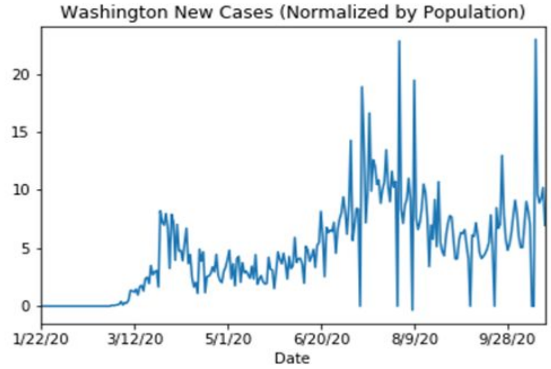
County Name	**Population**	**Total Deaths**	**Deaths per 10,000**
Yakima County	**250873**	**267**	**11**
Franklin County	**95222**	**67**	**7**
Benton County	**204390**	**132**	**6**
Kittitas County	**47935**	**22**	**5**
Adams county	**19983**	**10**	**5**

Part 4 Plot daily trends (cases and deaths, new cases) of state and to 5 infected counties

Daily trends for state: Washington

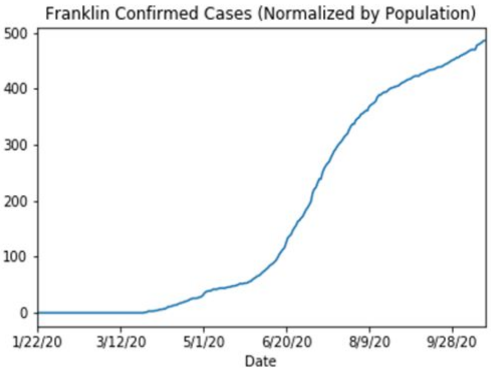


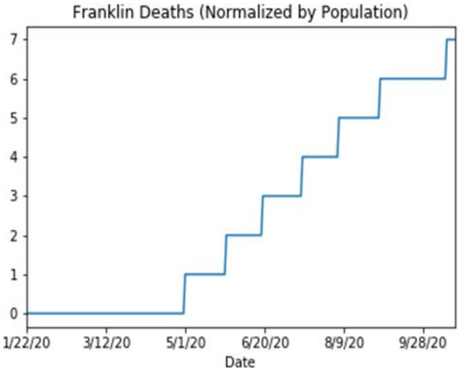


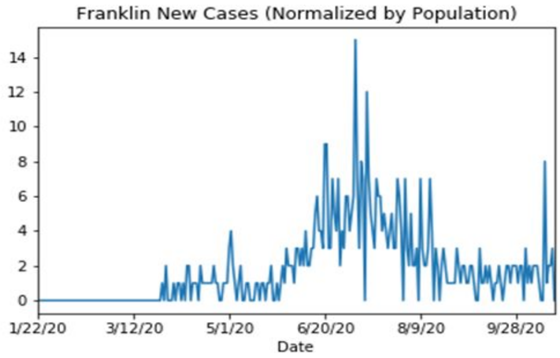


Plot county data:

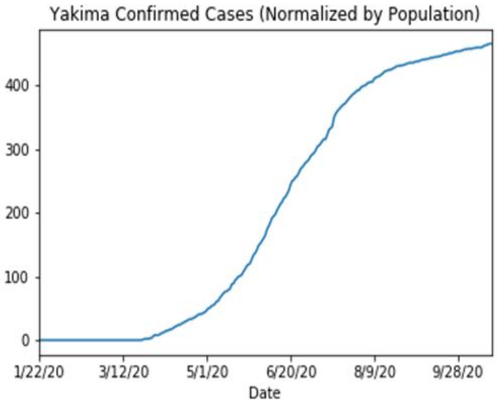
Franklin County

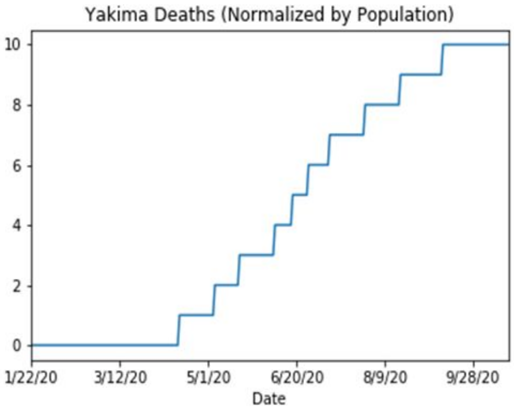


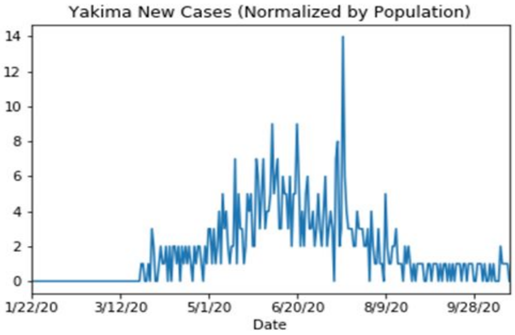




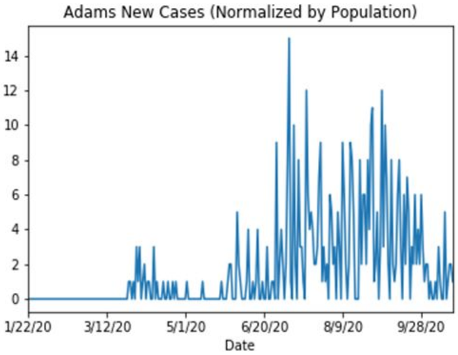
Yakima County:

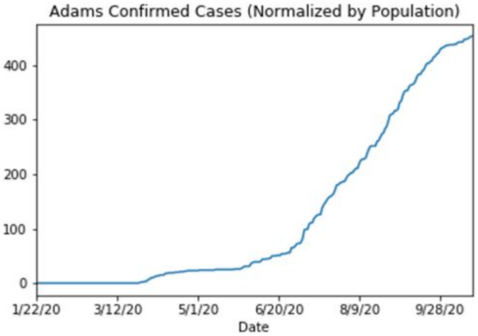


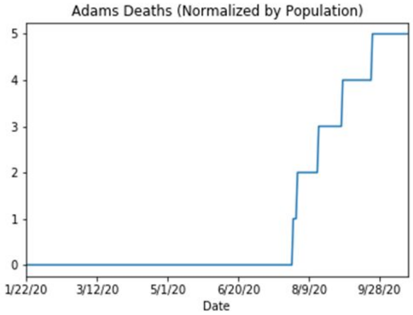




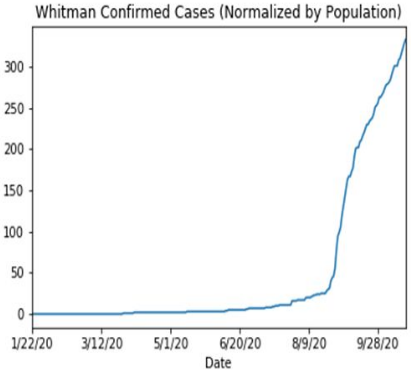
Adams County:

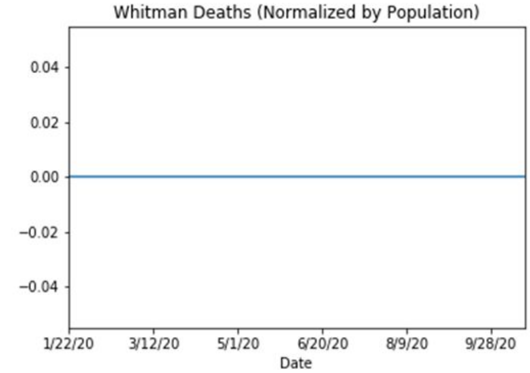


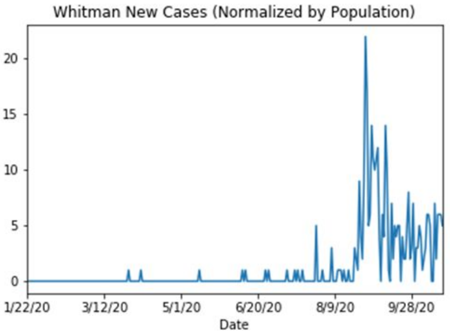




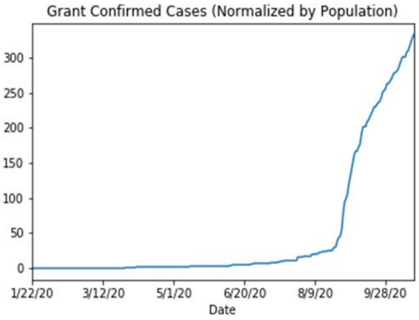
Whitman County:

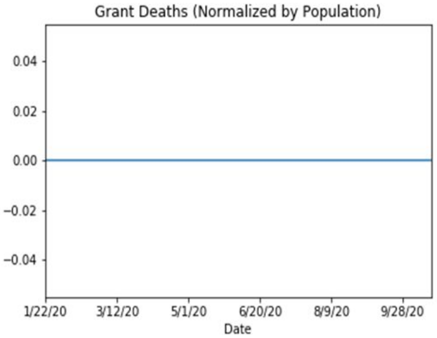


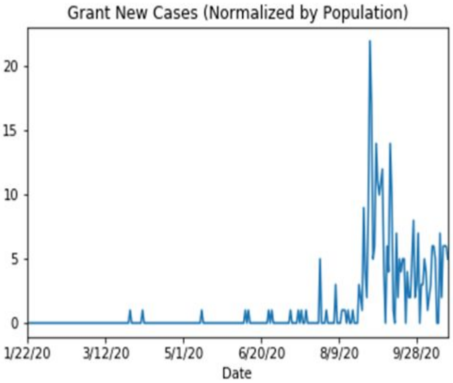




Grant County:







End Task 1

Raiana Zaman:

Task 2:

Task 2:

Members:

Jamison

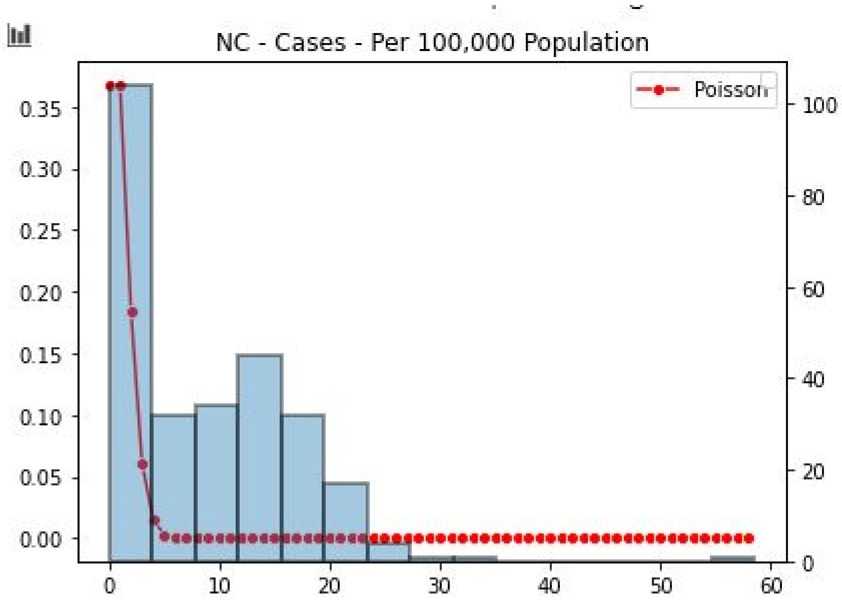
Xinrui Zhang

.

Francis Perez:

Part 1 (Fit a distribution to the number of COVID-19 cases of a state):

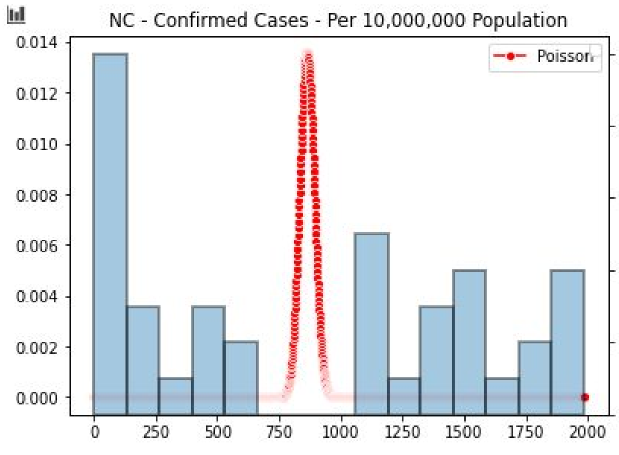
The poisson distribution was suggested as a discrete distribution by Dr. Mohanty. The plot below has two data plots, one is the histogram of the covid new cases in North Carolina and the poisson distribution with a mu of 1. The poisson's PMF function was used to generate the "Poisson" red line. A "mu" of 1 was selected because it best fit the histogram data for the new confirmed cases.

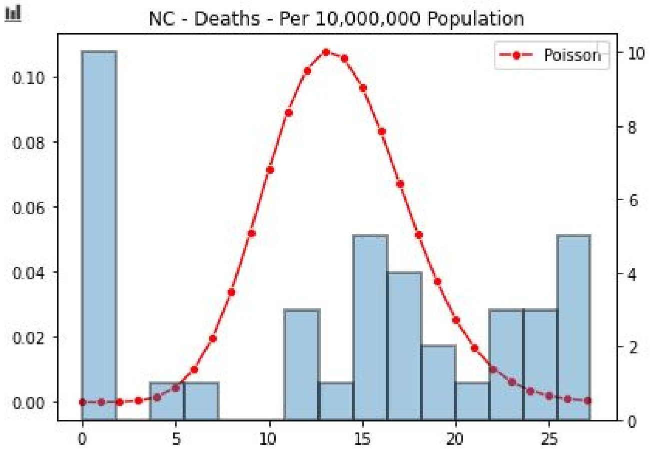


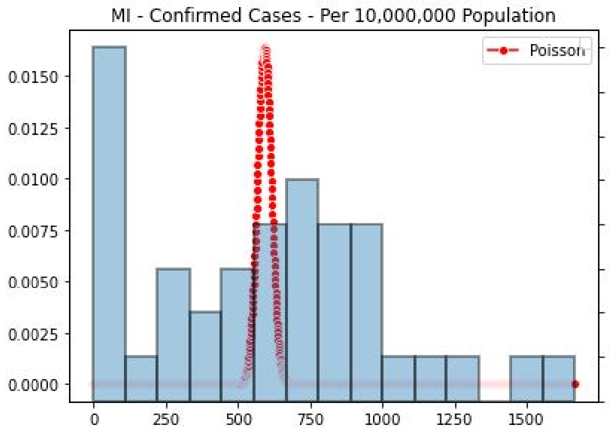
Part 2 (Model a poisson distribution of COVID-19 cases & deaths of a state and compare to other 5 states):

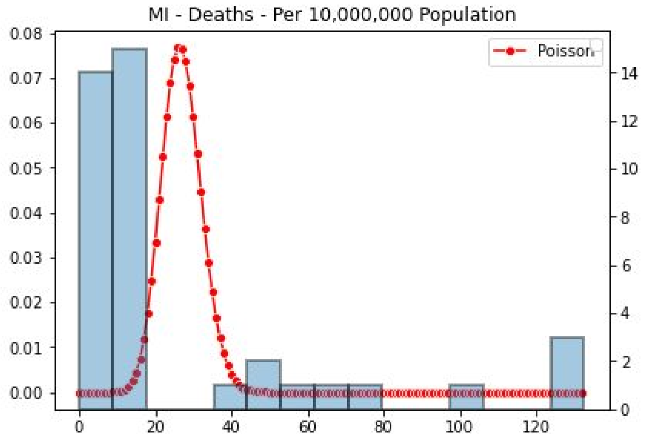
The states used, based on a similar population to my state of North Carolina.

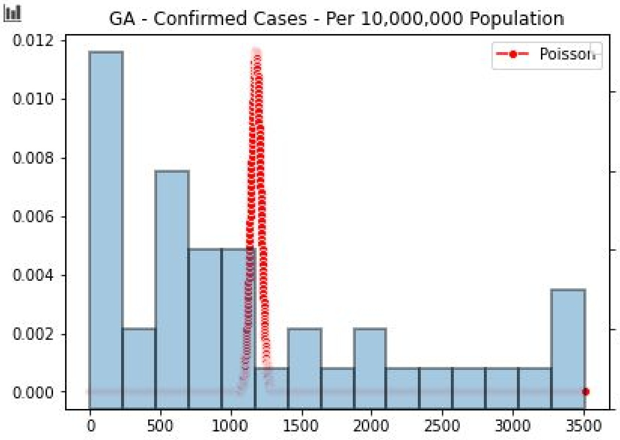
Michigan, Georgia, Tennessee, Washington, New Jersey

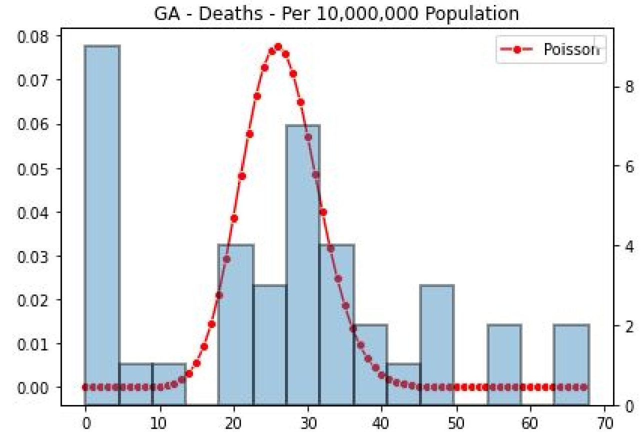


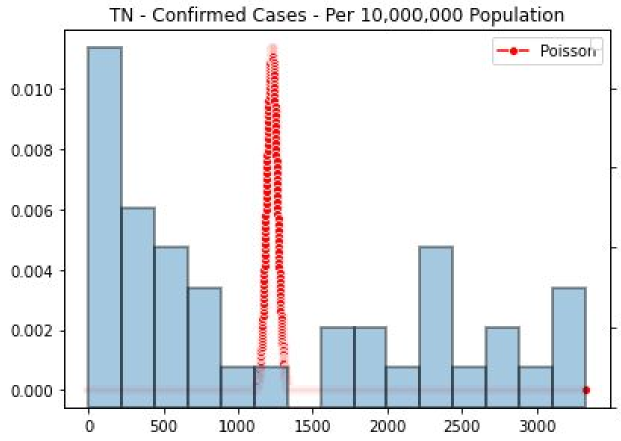


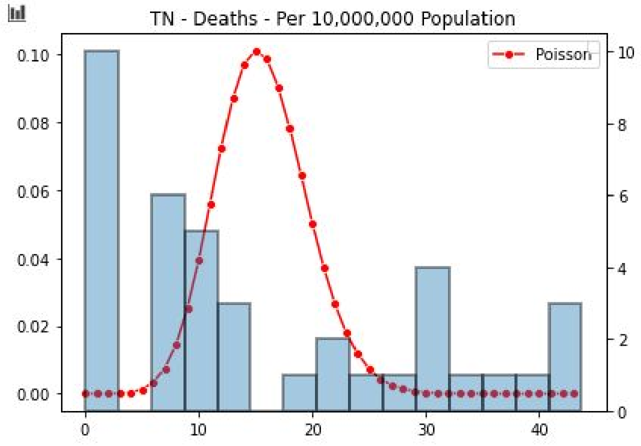

MI - Confirmed Cases - Per 10,000,000 Population

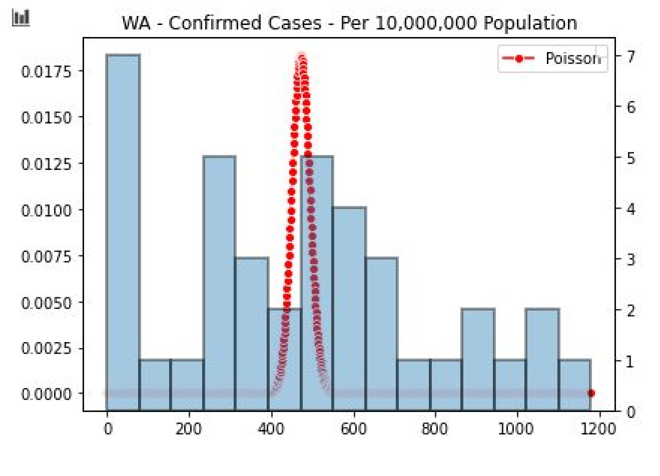

MI - Deaths - Per 10,000,000 Population

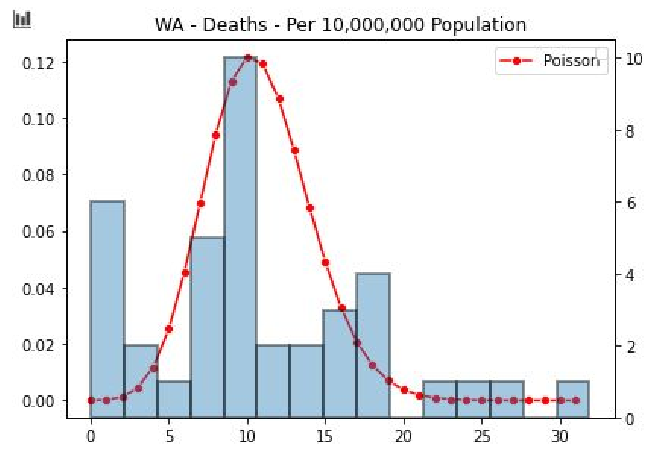

GA - Confirmed Cases - Per 10,000,000 Population


GA - Deaths - Per 10,000,000 Population


TN - Confirmed Cases - Per 10,000,000 Population


TN - Deaths - Per 10,000,000 Population

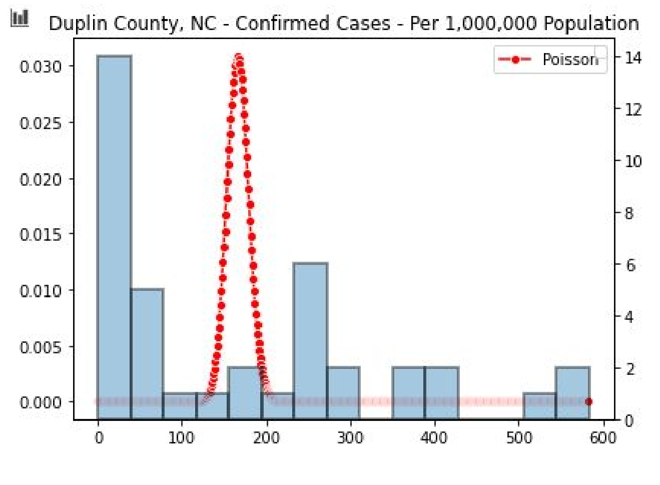

WA - Confirmed Cases - Per 10,000,000 Population

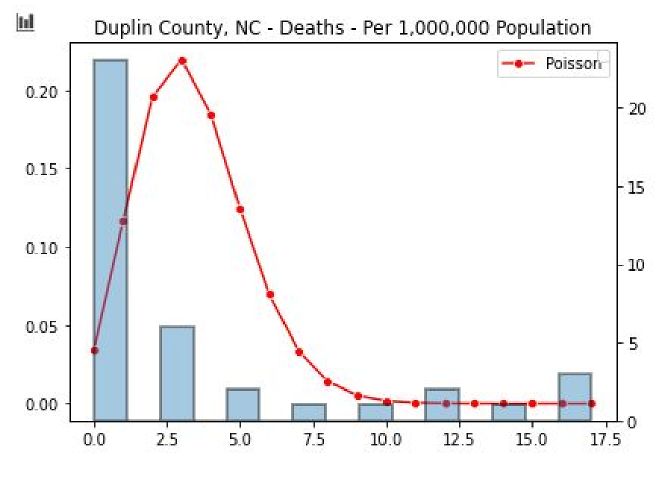

WA - Deaths - Per 10,000,000 Population

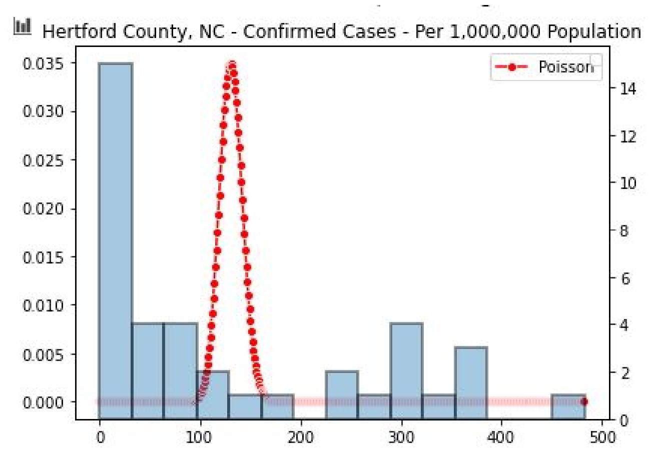
Part 3 (Model poisson distributions for North Carolina counties COVID-19 in cases & deaths):

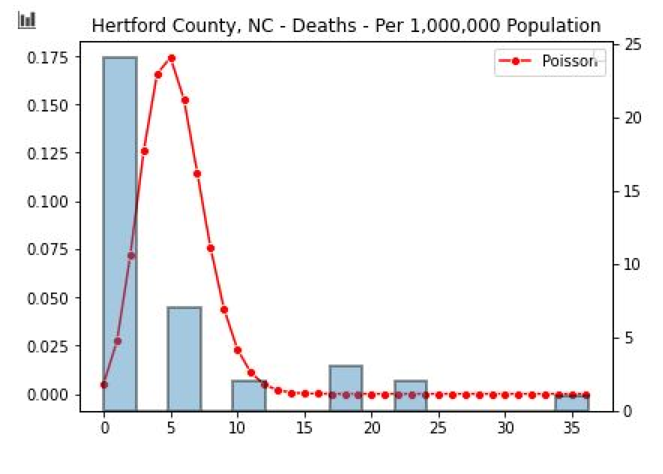
These counties were chosen based on the Top - 5 Infected Counties In North Carolina. From Task 1.

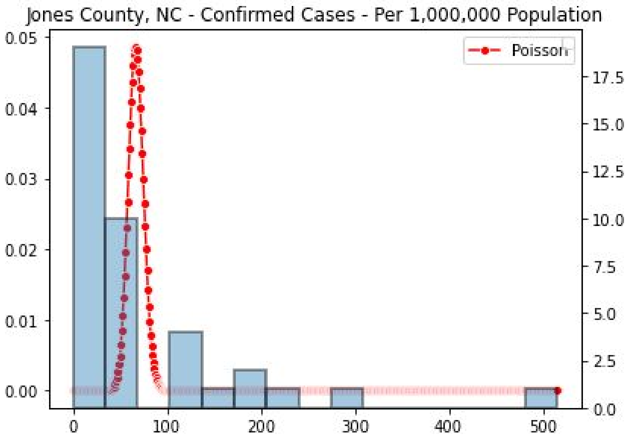
It does seem that the counties with the lowest population were the hardest hit.


Duplin County, NC - Confirmed Cases - Per 1,000,000 Population

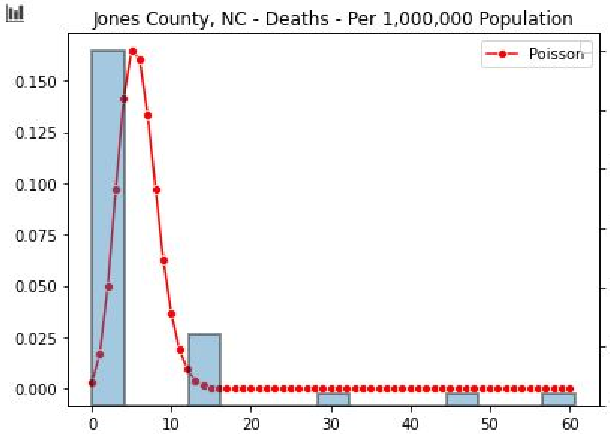

Duplin County, NC - Deaths - Per 1,000,000 Population


Hertford County, NC - Confirmed Cases - Per 1,000,000 Population

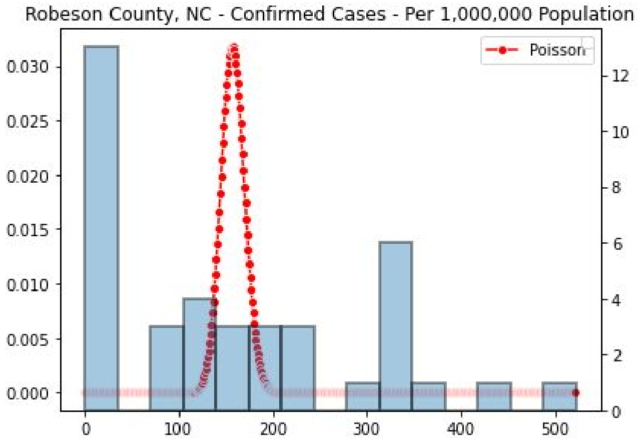

Hertford County, NC - Deaths - Per 1,000,000 Population



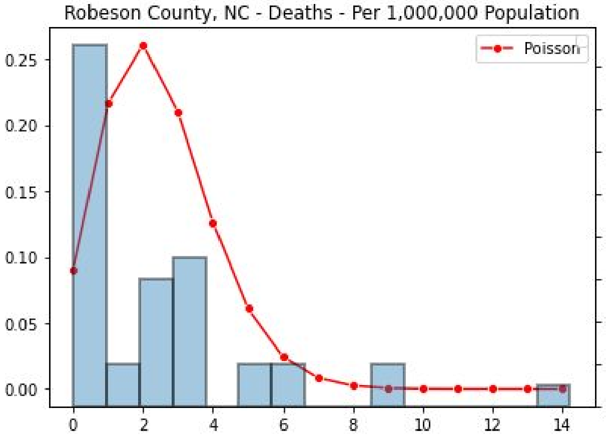
Jones County, NC - Confirmed Cases - Per 1,000,000 Population



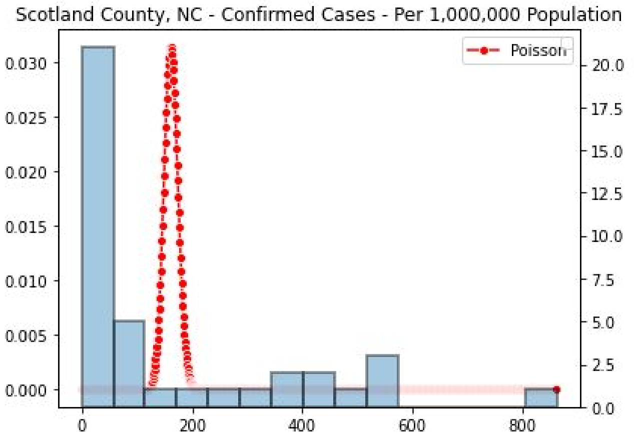
Jones County, NC - Deaths - Per 1,000,000 Population



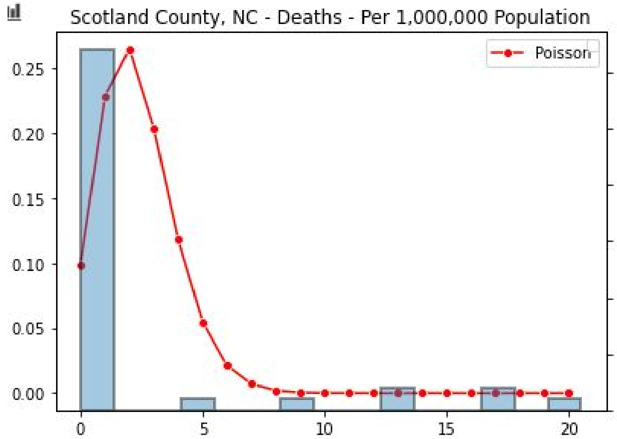
Robeson County, NC - Confirmed Cases - Per 1,000,000 Population



Robeson County, NC - Deaths - Per 1,000,000 Population



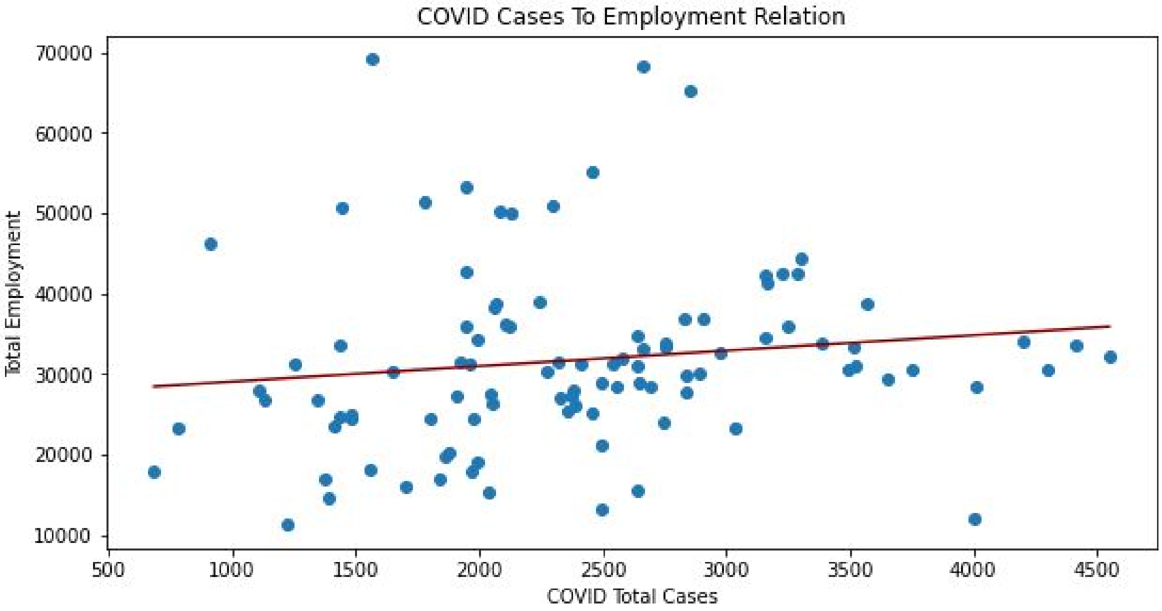
Scotland County, NC - Confirmed Cases - Per 1,000,000 Population



Scotland County, NC - Deaths - Per 1,000,000 Population

Part 4 (Perform correlation between Enrichment data variables and COVID-19 cases for patterns):

In comparing North Carolina Counties with Enrichment data (employment numbers). A scatter plot with x - Covid Cases & y - Total March Employment numbers was created. There seems to be a correlation between these values. The graph below a line of regression was computed, based on least-squares. This line of regression, in red below, has positive slope with a R and P values of .140 and .164 respectively. Given the positive slope of the line it can be said that there might exist a pattern of correlation, in respect to the higher the employment to a higher number of cases. However, we could not say the higher number of employment causes the confirmed cases to increase without further study.



Statistics	
R	0.140
P	0.164
Std Error	1.37

Part 5 (Formulate hypothesis between Enrichment data & number of cases to be compared against states): Based on the data above, it can be said that there seems to be a correlation between Employment totals and the number of cases. It is possible that the different types of employment, government, manufacturing, and service jobs might have a higher correlation.

Hypothesis:
Does a larger employment of government, manufacturing, or service jobs in an area cause more confirmed cases of covid?

End Francis Perez:

Raiana Zaman:

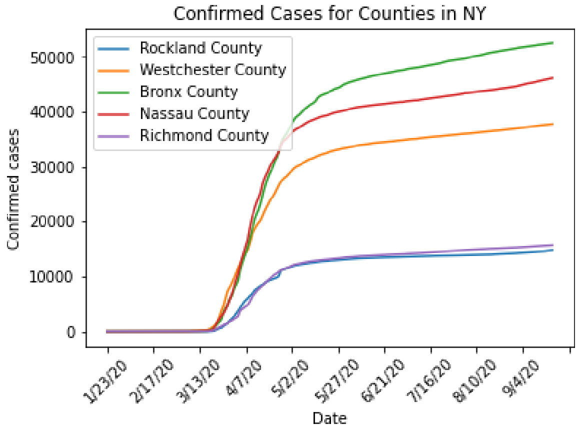
1.Generate weekly statistics (mean, median, mode) for number of new cases and deaths across a NY state

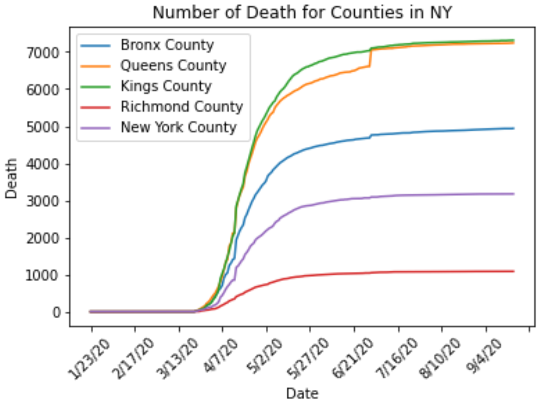
```
Weekly Confirmed
==============================
        Mean  | Median |  Mode
  NY    1479  |  565   |   0


Weekly Death
==============================
        Mean  | Median |  Mode
  NY    103   |   11   |   0
```

2.Compare the data against other states.

```
Weekly Confirmed
==============================
        Mean  | Median |  Mode
  GA    990   |  822   |   0
  IL    909   |  990   |   0
  NC    617   |  536   |   0
  NY    1479  |  565   |   0
  OH    538   |  543   |   0
  PA    482   |  474   |   0
Weekly Death
==============================
        Mean  | Median |  Mode
  GA     23   |   17   |   0
  IL     14   |   7    |   0
  NC     4    |   3    |   0
  NY    103   |   11   |   0
  OH     12   |   8    |   0
  PA     18   |   3    |   0
```

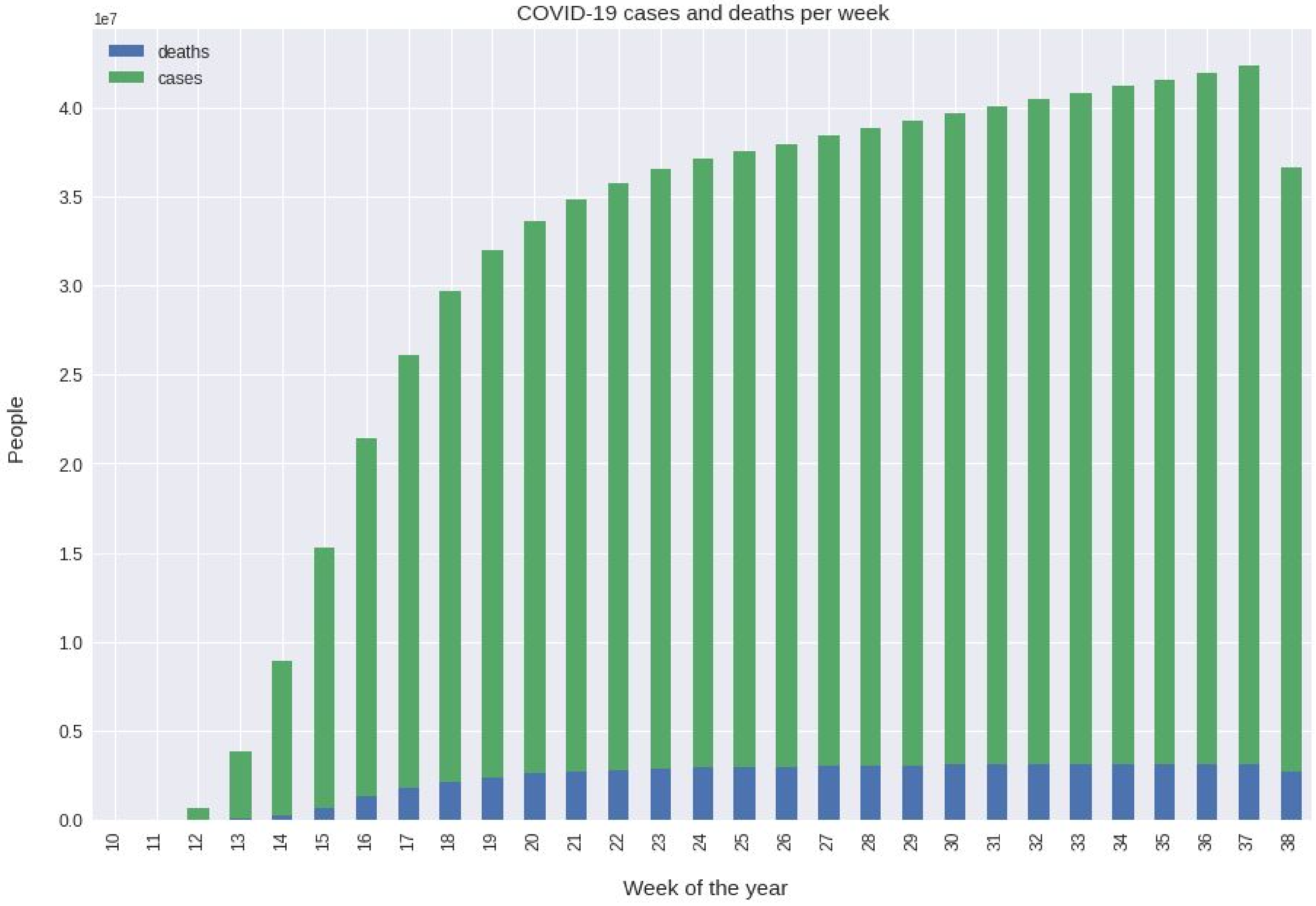
3.Identify counties in NY state with high case and death rates





Part 2

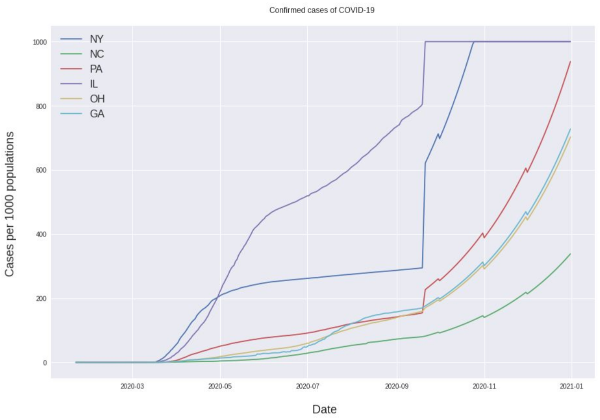
Fit a distribution to the number of COVID-19 cases of New York state.

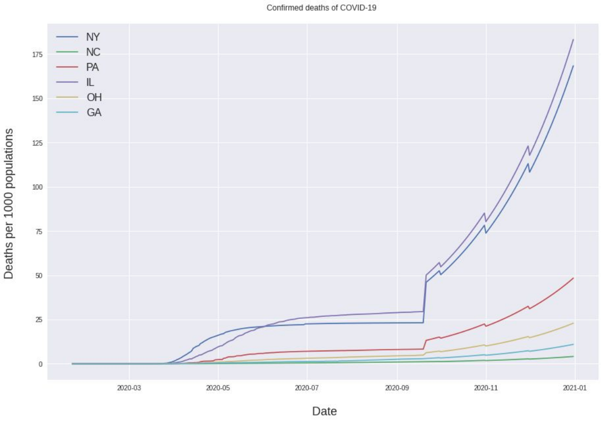


This graph shows how since the first COVID cases were confirmed in New York until now, the number of cases and deaths has been growing. As we can see, the number of confirmed cases grew at a very high rate, while deaths grew a lot at first, but from week 20, they stabilized. This may be due to the turnover of people occupying beds and also to the fact that the number of beds in hospitals increased.I decided to show the distribution by weeks because it is a considerable time to see the pandemic advanze and alsits the time in the majority of people in which they leave the hospitals.

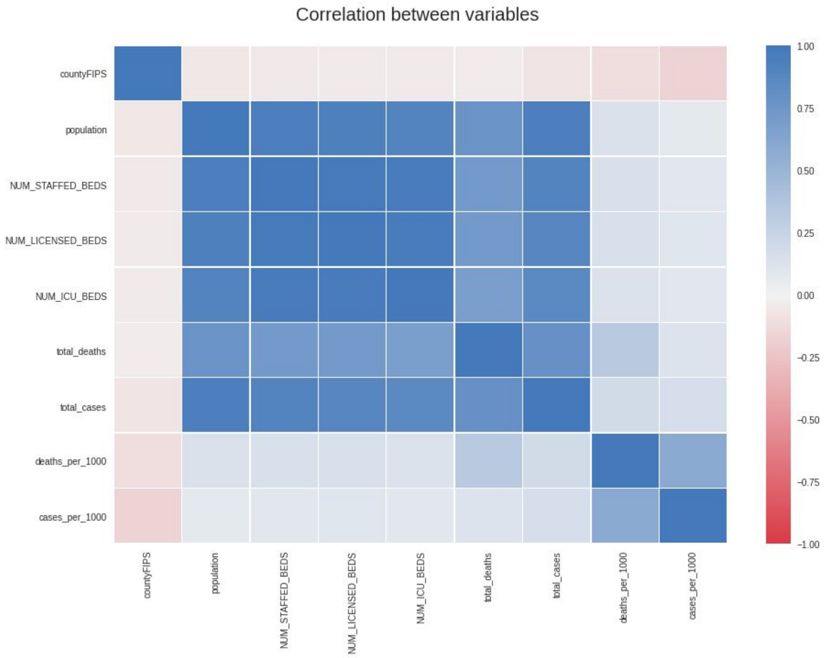
Model a poission distribution of COVID-19 cases and deaths of a state and compare to other 5 states.

Here, using poisson distribution to predict the confirmed cases of the next months, we can see that NY and IL will probabbly have all their people infected, and the other states goes for the same way but slowly. Deaths will increase but slower than the confirmed cases.





Perform corelation between Enrichment data valiables and COVID-19 cases to observe any patterns.



Watching this heatmap, we can see that cases and deaths are directly related with population. Population is really related with beds. So, the four attributes (deaths, cases, beds and population) have a strong correlation. This means that when one of this grows up, the others too.