Stage 2 Report

Team Task:

In this stage of the project, basic statistics from the super COVID19 dataframe are derived. To do so, the daily records of the US were centralized by taking the mean across all the data. After this, a calculation on mean, median and mode were performed. These statistics allow the data analysis to determine its center values. The US average value of covid19 cases was about 4571 whereas its median was 4462. Therefore, a conclusion on the shape of the distribution can be drawn as not being a symmetric or a normal distribution. As for the mode, it has a value of 0, and that makes sense because in the beginning of the year, zero cases of COVID-19 were confirmed across the US which makes it the most occurring value in the dataset. Similarly, for the number of registered deaths, the mean was 123 and the median 118; this also confirmed the previous conclusion about the distribution of the dataset.The mode remains zero.

When these statistics are compared against the other countries around the world, there is a significant difference as the table below shows:

	Country	Mean	Median	Mode
0	United States	4571.0	4462.0	0
1	Bangladesh	435.0	431.0	0
2	Indonesia	244.0	133.0	0
3	Pakistan	269.0	144.0	0
4	Brazil	4510.0	5093.0	0
5	Nigeria	55.0	43.0	0

Among these 5 countries, Brazil outperformed the US based on the median value, whereas the other countries have very small numbers of.

For the number of deaths, Brazil has recorded the highest number of deaths based on the last register data which in this case was on October 16th, 2020.

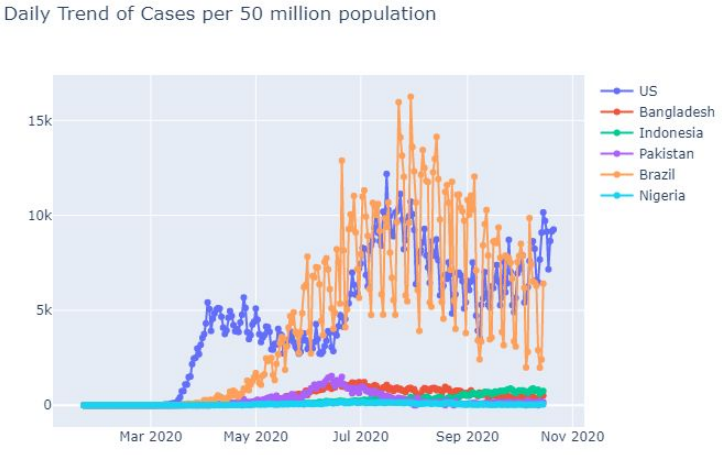
	Country	Mean	Median	Mode
0	United States	123.0	118.0	0
1	Bangladesh	6.0	8.0	0
2	Indonesia	8.0	6.0	0
3	Pakistan	5.0	2.0	0
4	Brazil	133.0	161.0	0
5	Nigeria	1.0	1.0	0

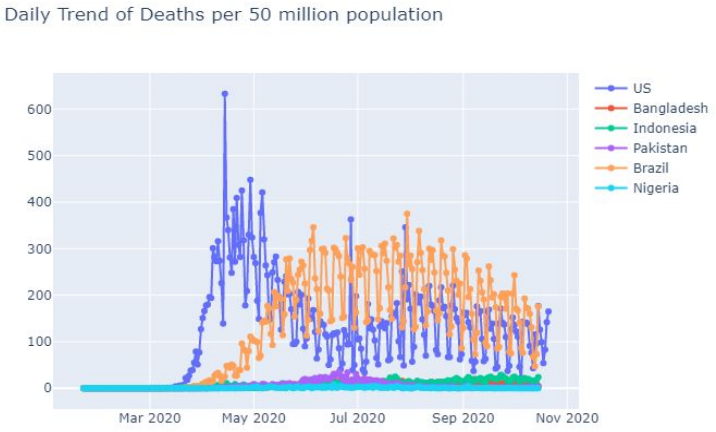
Below is the daily trend of new cases and new deaths for the United States and the other 5 countries per 50 million population. From the graph we can see that the United States cases and deaths are really high alongside with Brazil. Just recently we can see that the United States almost hit the peak amount of new cases which previously was in July. Brazil's new cases are really high one day, but really low another day. When looking At the daily trend of deaths, we can see that peak new deaths of the United States was in April. However, the death trend is higher than other countries.

Member Tasks:

Harinder Badesha :

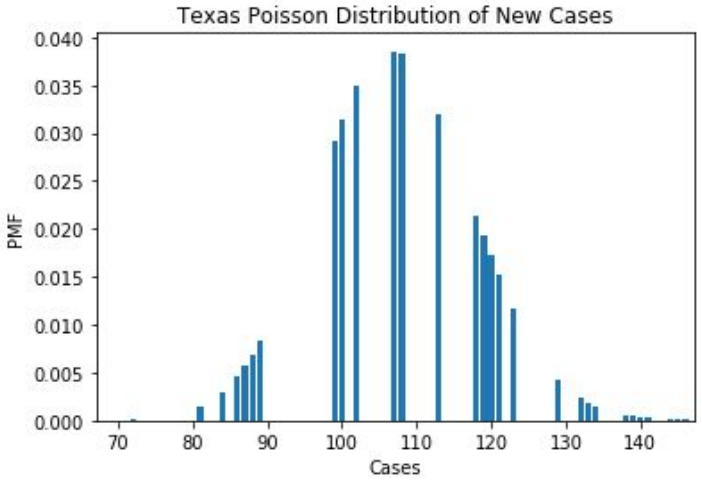
For my member task:Task 1, I chose Texas as my state to generate weekly statistics and compared it with other states with a close population to Texas which included California, Florida, New York, North Carolina and Philadelphia. Texas had the second highest weekly average of new cases out of the other five states listed above right behind Florida, but the weekly average of new deaths in Texas was not as high as other states. Daily trend of new cases in Texas was left skewed with peak days on July 16 and September 22. The daily trend of the top 5 infected counties in Texas was similar to the states with peak days hitting around July and September. Daily trend of new cases and new deaths of the state of Texas and its

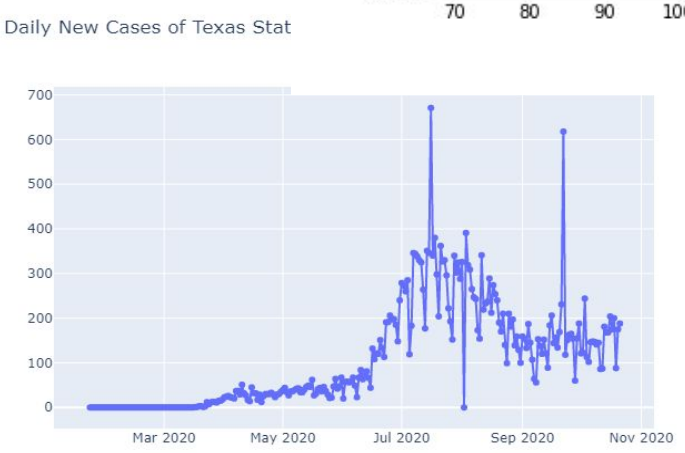


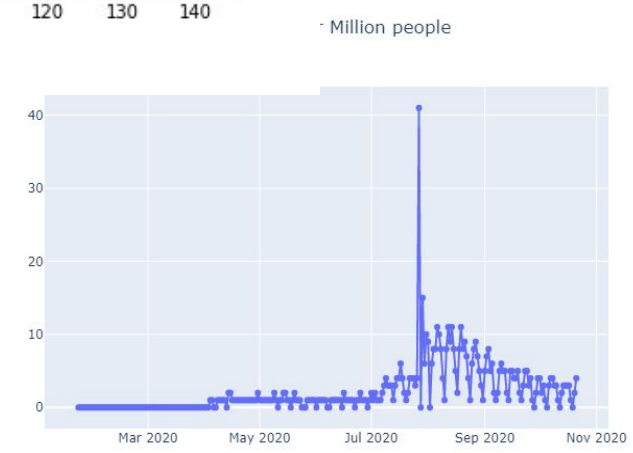


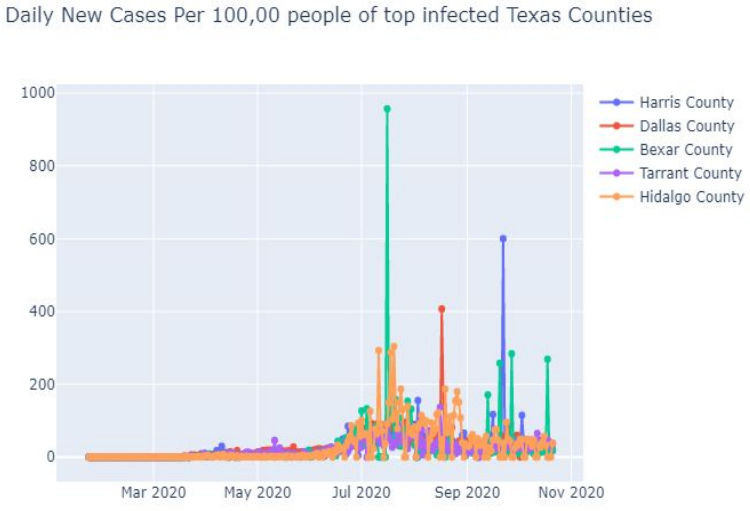
top 5 infected counties are showcased below:

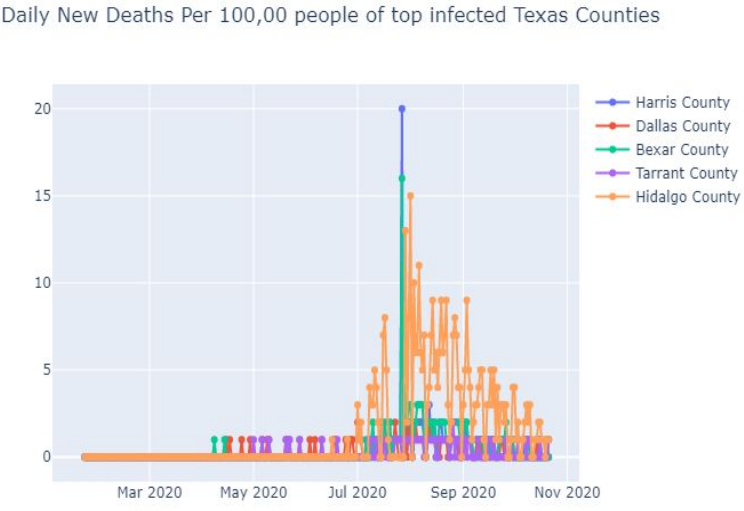
When fitting a distribution to the number of COVID-19 cases of the state of Texas, I used a poisson distribution because we have discrete values of cases(we cannot have a fraction of a case) and for a poisson distribution in this scenario, it is a good analysis to see the probability of the number of new cases happening in a day. Below is the Texas Poisson distribution:



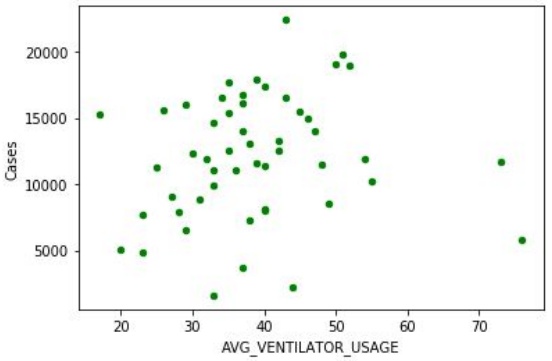






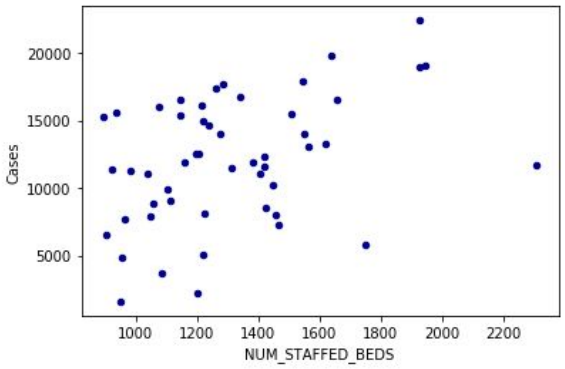


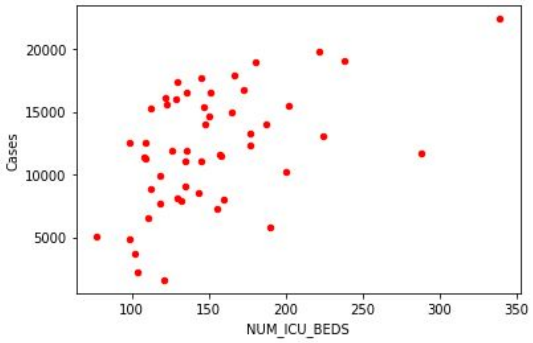
As the graph above shows, we can see that in Texas there is a high probability that 100-120 new cases will happen in a day with 107 new cases being the average/mean. When comparing the Poisson distribution of Texas to other states like California, Florida, North Carolina, New York, and Philadelphia, We see that the Texas interval of cases is close to Florida's where their interval of cases is in the 100's. The new COVID-19 deaths distribution for states is shaped more differently as states like California, North Carolina and California are right skewed.



The five North Carolina counties I chose were Durham, Forsyth, Guilford, MecklenBurg, and Wake county which had similar poisson distributions. The counties are normalized by 200,000 people. All counties have a high probability of intervals 0-30 new cases per 200,000 people. Looking at all the distributions for new cases, It showcases that the average highest probability of new cases happening in a day is around 20 cases per 200,00 population.

The three enrichment data variables that I chose from the hospital bed dataset to find Any correlation with cases was the number of staffed beds, number of ICU beds and average ventilator usage. Below are the graphs corresponding to the three variables.





It can be seen that all variables have a positive correlation towards the number of cases with the number of ICU beds variable having a high correlation with the total number of cases.

Sanam Khalili:

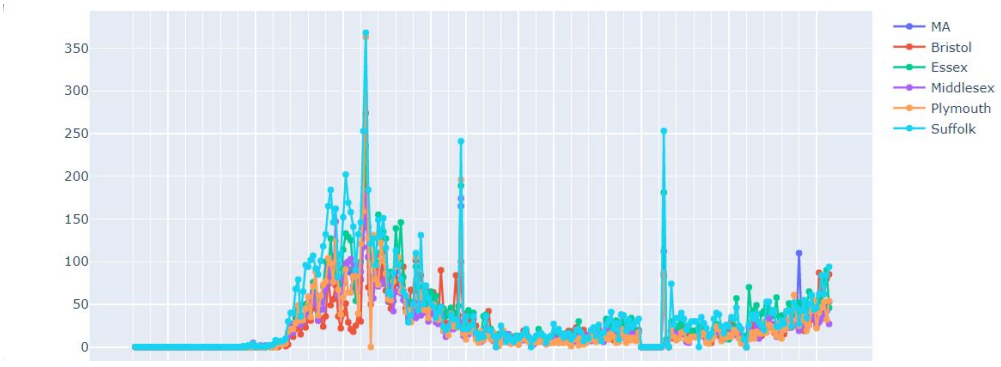
This task demands for weekly statistics i.e. mean, median, and mode containing cases and death numbers in each state. For this reason, I selected MA to generate weekly statistics. Then, I continued with the following states that are similar to MA: AZ, IN, MO, NJ, TN thereby a comparison with primary data (MA data) can be made. Next step was to normalize the cases and deaths based on a number that must be selected. Since the population of the selected states is around 7000000, I selected 1000000 as the base for normalization. The following table shows the normalized values:

	AZ	IN	MA	MO	NJ	TN
Mean	118	90	82	102	93	129
Median	64	76	50	39	46	80
Mode	0	0	0	0	0	0

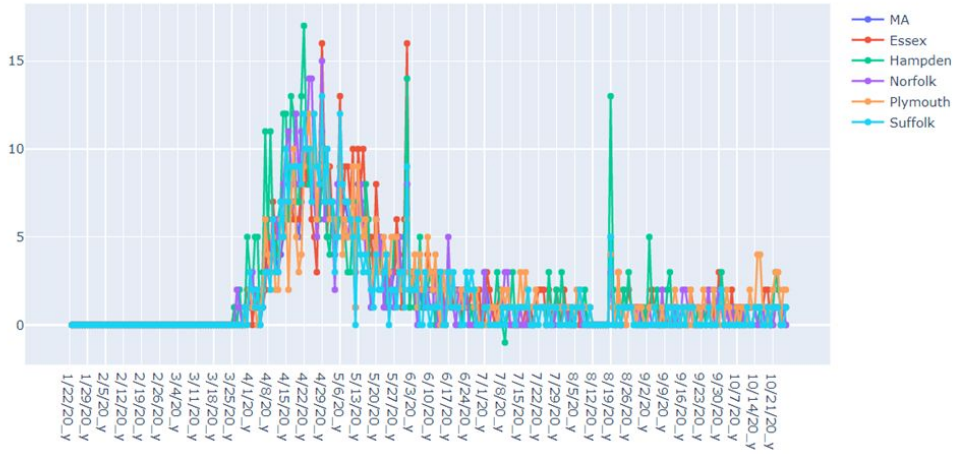
In all states mean is larger than median and mode is zero which means the distribution is right skewed. Larger value of mean shows that there are a few larger values of new cases in the data which drives the mean upward, but they could not affect the median. The difference between mean and median in IN is not very big which shows that the distribution of new cases data can be assumed to be approximately symmetrical compared to other states. TN has the highest mean and MA lowest mean among other states. Since the data is normalized, we can conclude that MA has the lowest new cases and TN has the highest new Cases among other states.

	AZ	IN	MA	MO	NJ	TN
Mean	3	2	5	2	7	2
Median	2	2	2	1	1	1
Mode	0	2	2	1	1	1

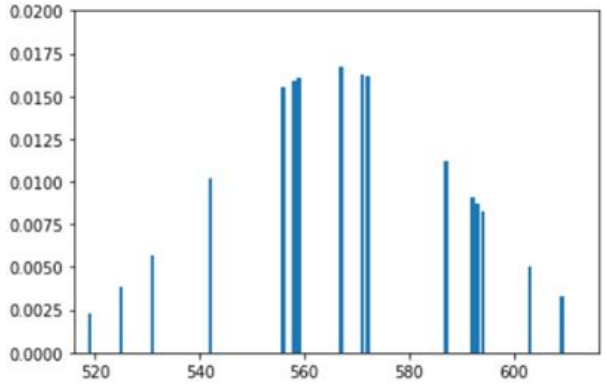
NJ has the highest mean of new deaths and the difference between its median and its mean is remarkable compared to other states which shows the peak is close to the left side of the graph. IN, TN and MO have the lowest Mean but the mean, median and mode of IN are equal which demonstrate that the distribution of new cases in IN as approximately symmetrical. For some period of time mode = 0 indicates that the state was successful to minimize the number of deaths for a greater number of weeks, however it was temporary.



The above image shows that Suffolk county has the highest normalized number of cases.



The above image shows that Hampden had the highest normalized number of deaths. Histogram of the data shows that the data is right skewed. Furthermore, we know that the data is discrete, thus, the poisson distribution can fit our data.



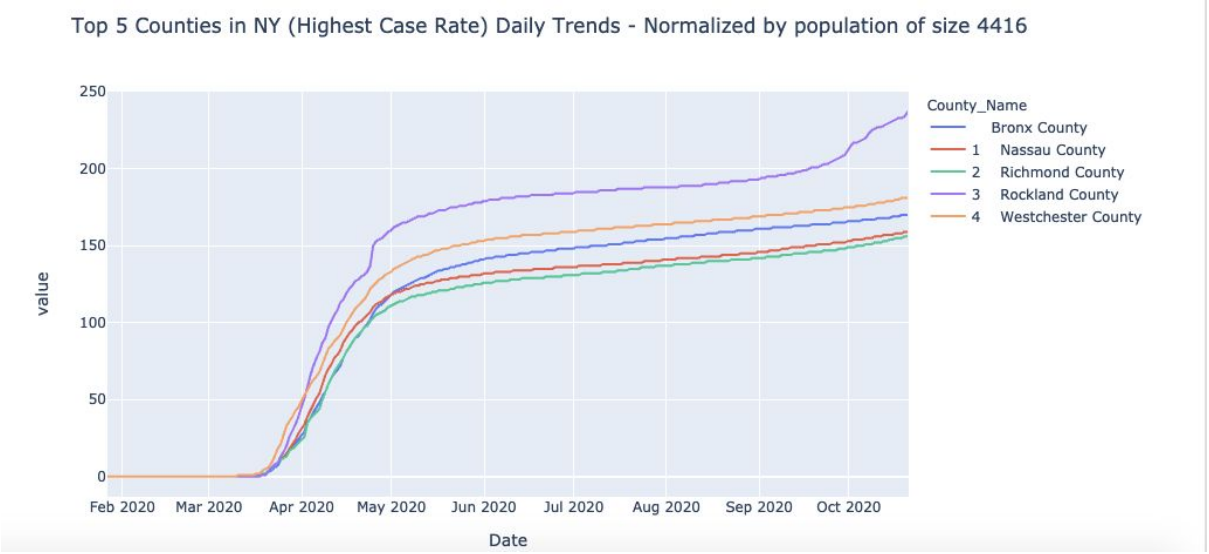
I selected AZ, IN, MA, MO, MD, and TN for correlation analysis. I normalized total cases for states along with values for enrichment data sets. Next, I applied correlation between each column in the enrichment data set and total cases of each state. The columns with high correlation are:

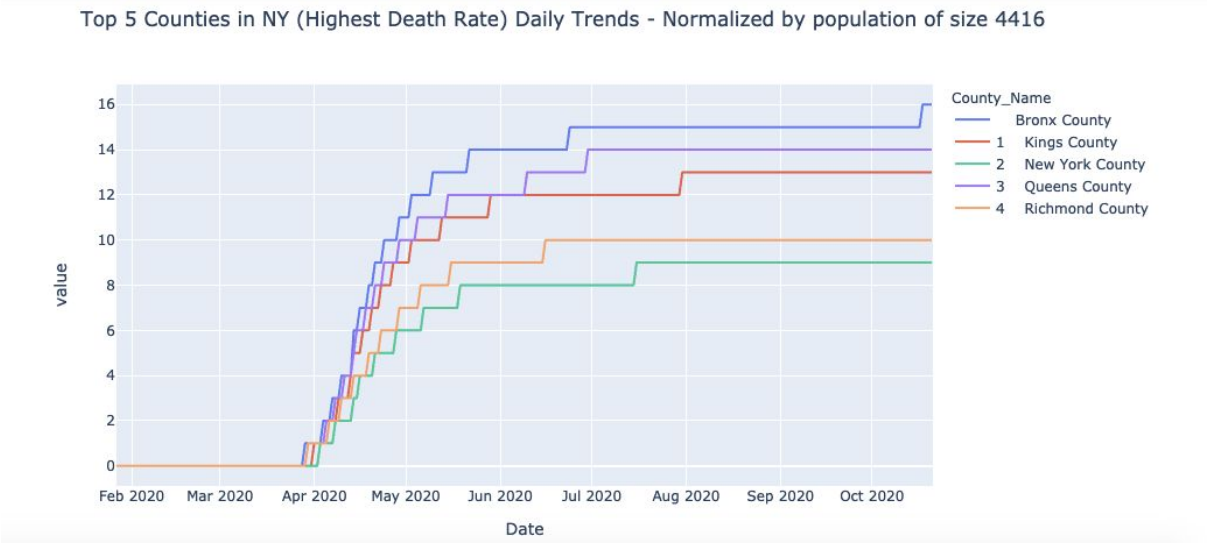
```
[['Estimate!!SEX AND AGE!!Total population!!Median age (years)',
  0.8920077668776693],
 ['Estimate!!SEX AND AGE!!Total population!!Sex ratio (males per 100 females)',
  0.8871059023890782],
 ['Estimate!!SEX AND AGE!!Total population!!18 years and over!!Sex ratio (males per 100 females)',
  0.8866263268109458],
 ['Estimate!!HISPANIC OR LATINO AND RACE!!Total population!!Not Hispanic or Latino',
  0.6522315121971706],
 ['Estimate!!HISPANIC OR LATINO AND RACE!!Total population!!Not Hispanic or Latino!!White alone',
  0.583684863854644],
 ['Estimate!!SEX AND AGE!!Total population!!Under 5 years',
  0.3116931426937923],
 ['Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18 and over population!!Female',
  0.29622394302877697],
 ['Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18 and over population',
  0.28430067396059033],
 ['Estimate!!SEX AND AGE!!Total population!!5 to 9 years',
  0.28169060012203373],
 ['Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18 and over population!!Male',
  0.26601491294280521
```

Hypothesis: from the results it can be inferred that younger ages are less likely to appear in cases, and a large portion of cases arise from older generations. It also seems that not being a hispanic or latino has a slight advantage in probability of appearing in the cases counted.

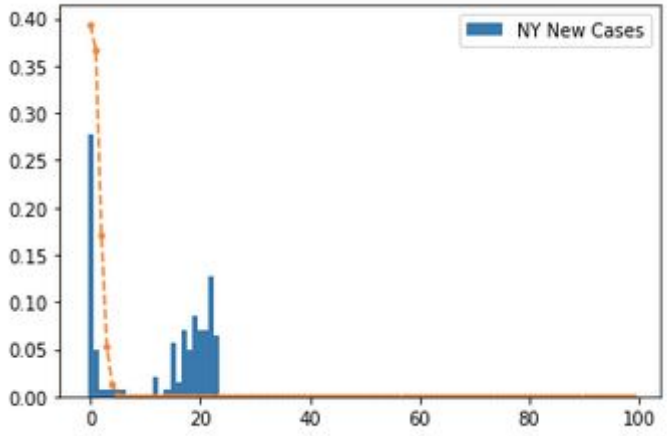
Isaac Taylor:

Member Task 1: For this task, I analyzed the state of NY. By doing this, I found that its weekly mean was 12526, the median was 5101, and the mode was 0 for new cases. For new deaths, I found that the mean weekly mean, median, and mode were 846, 91, and 0 respectively. The mode being 0 was interesting, and it may possibly be due to a lack of test results being available in the first few weeks when covid data was beginning to be collected. Comparing NY to other states, I found that the average new cases for all states was 13999. The average new death was 328. It is important to note that these values are normalized by population per 578,759 individuals because that is the population size of the state with the smallest population. Ultimately I found in my comparisons that NY had both a higher than average case and death rate. Also it was also interesting to find out that the Bronx was in the Top 5 counties in NY with the highest case and death rates. Despite not having the highest case rates, it still had the highest death rates.





Member Task 2: In this task, I fitted a distribution to New York's Data. I found that the poisson Distribution was a good choice given that we are dealing with cases and deaths (discrete values). Furthermore, after being required to model the data with a poisson distribution, I found that NY seemed to have a higher case rate than most other states specifically around certain periods of time which may have to do with city quarantine easing up at times or other factors. In this task, I also merged the Social enrichment dataset with the COVID-19 normalized state cases. I attempted to correlate new cases to non family households, non family households with 65+ age owners, and non family single female owner households. My hypothesis was that certain household types affect the number of cases more than others across states. Interestingly enough, non family single female households have a slightly higher positive correlation with cases than the other variables. The Pearson correlation coefficients for non family households, non family households with 65+ age owners, and non family single female owner households respectively are 0.10763, 0.10975, and 0.1420.



Nadia Doudou:

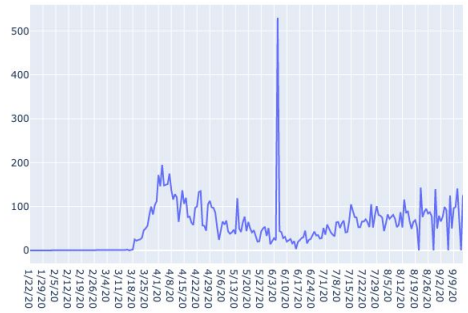
Task1: In this task, MI state was chosen to compute the weekly statistics for number of confirmed cases and number of deaths of COVID19. The mean and median for the number of cases were respectively 53 and 60; this means the dataset does not have a normal distribution. The mode was 0 probably because at the beginning of the year, there were 0 cases identified in the US which is why it is the most frequent value in the dataset. Similarly for the number of deaths the mean and median are different as well.

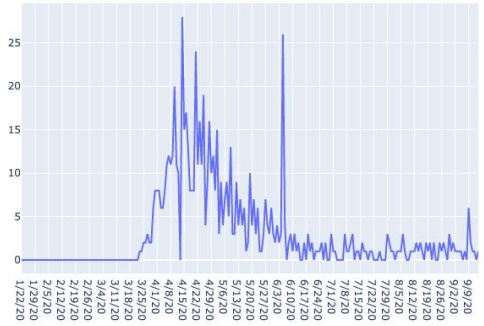
When this state was compared with other states with similar numbers of population, MI appears to have a lower number of cases but the number of deaths was relatively the same as shown in the tables.

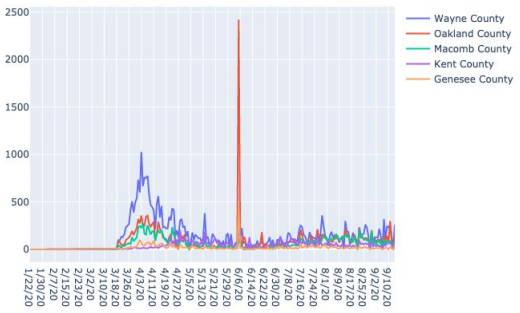
	State	Mean	Median	Mode
0	MI	53.0	60.0	0.0
1	GA	118.0	70.0	0.0
2	IL	88.0	88.0	0.0
3	NC	75.0	60.0	0.0
4	OH	50.0	47.0	0.0

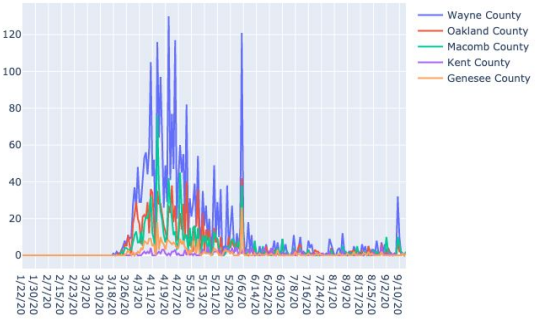
	State	Mean	Median	Mode
0	MI	3.0	1.0	1.0
1	GA	3.0	2.0	1.0
2	IL	3.0	1.0	1.0
3	NC	1.0	1.0	1.0
4	OH	2.0	2.0	1.0

After this, the 5 top counties were identified within MI state and a daily plot for the trend was made for the purpose of visualization. The Graphs respectively represent the daily trend for cases and death in MI state and the top 5 counties .





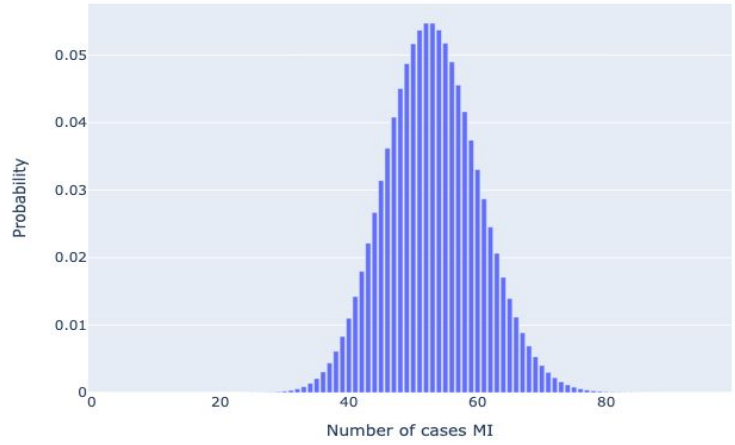




Task2:

A histogram of the daily record of Michigan state number of confirmed cases of COVI19 shows that the distribution was right skewed which means a Gamma distribution can fit this data. However, Gamma distribution works only for continuous variables so it will not work

for our dataset which is discrete. This is the reason for which the Poisson distribution was chosen to fit the data. The statistics for the Poisson distribution of Michigan State are as follows: The mean and the variance are equal to the lambda value E(X) = Var(X) = . The skewness and the kurtosis of the distribution are 0.137 and 0.018 respectively. This result also indicates that the shape of the Poisson distribution will be close to normal which the histogram above confirms because a normal distribution is not skewed so the skewness ratio is 0. Below is a graph for the number of cases of COVID19 in MI state.The mean is equal 53 so the probability around this value is higher (more than 0.05).



Correlation: Employment Dataset and COVID19

An analysis on the employment level and the number of confirmed cases of COVID19 was performed so that a hypothesis can be derived between these two variables. The given dataset has the total number of COVID19 cases per counties and the number of employees per counties for the first quarter of the year: January, February and March. A first analysis was done on the State of Michigan and its level of Employment by calculating the correlation coefficient. A Pearson correlation was calculated between the variables that were cited above. The number of cases and the employment level were found to be strongly correlated. This study indicates that there is a linear relationship between the employment level and the

number of cases of COVID19. In other words, regions with higher employment levels tend to have a higher number of COVID19 cases.

Ali Altamimi:

Task 1:

I choose to analyze Florida's COVID-19 cases and deaths. I first started finding the mean of weekly cases and deaths(weeks start from monday) and normalize by population and per 1M population. After that I calculated the Mean, Median and Mode for all states including Florida.

United States COVID-19 Cases

	State	Mean	Median	Mode
0	AK	55	18	0
1	AL	129	97	0
2	AR	118	99	0
3	AZ	116	63	0
4	CA	81	67	0
5	CO	54	54	0
6	CT	66	38	0
7	DC	85	75	0
8	DE	86	95	0
9	FL	128	52	0
10	GA	112	71	0
11	HI	36	9	0
12	IA	124	116	0
13	ID	109	26	0

United States COVID-19 Deaths

	State	Mean	Median	Mode
0	AK	0	0	0
1	AL	2	2	2
2	AR	2	1	0
3	AZ	3	2	0
4	CA	2	2	2
5	CO	1	1	1
6	CT	5	1	0
7	DC	3	1	1
8	DE	3	1	1
9	FL	3	2	2
10	GA	3	3	3
11	HI	0	0	0
12	IA	2	2	2
13	ID	1	0	0

After that I compared Florida's cases and deaths with other states and plots.

States Weekly Cases

States Weekly Deaths

After comparing Florida with other states I wanted to identify counties within florida' state with high cases and death rates. So, what I did is I normalized the population for the counties of Florida and sorted them based on the highest cases.

Top FL counties cases

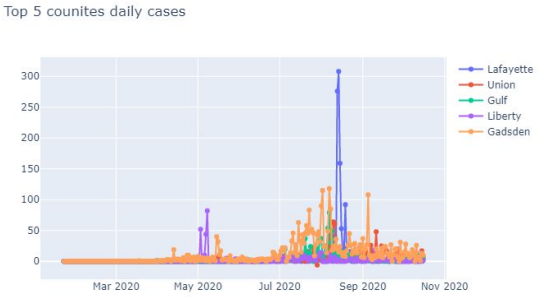
	County Name	State	population	Normalized Cases
32	Lafayette County	FL	8422	138.0
62	Union County	FL	15237	68.0
21	Gulf County	FL	13639	62.0
37	Liberty County	FL	8354	53.0
18	Gadsden County	FL	45660	52.0

Top FL counties deaths

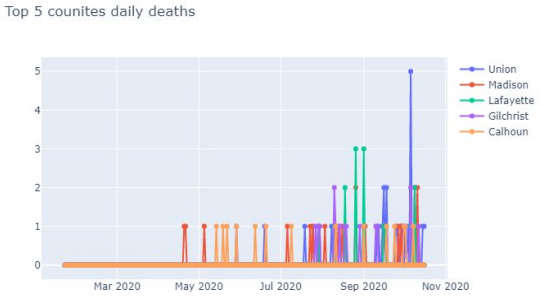
	County Name	State	population	Normalized Deaths
62	Union County	FL	15237	23.0
38	Madison County	FL	18493	21.0
32	Lafayette County	FL	8422	17.0
19	Gilchrist County	FL	18582	15.0
6	Calhoun County	FL	14105	14.0

Then I plot the daily cases trends of Florida top 5 cases and deaths counties.

Top counties daily cases

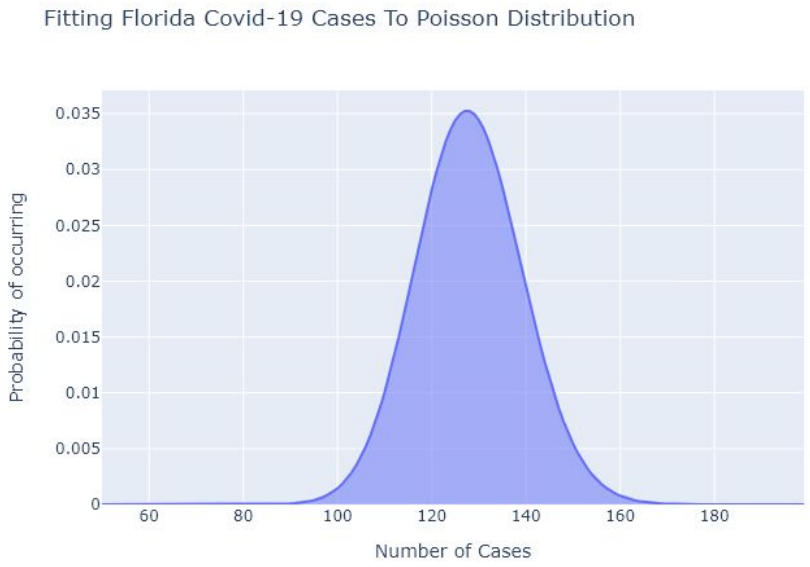


Top counties daily death



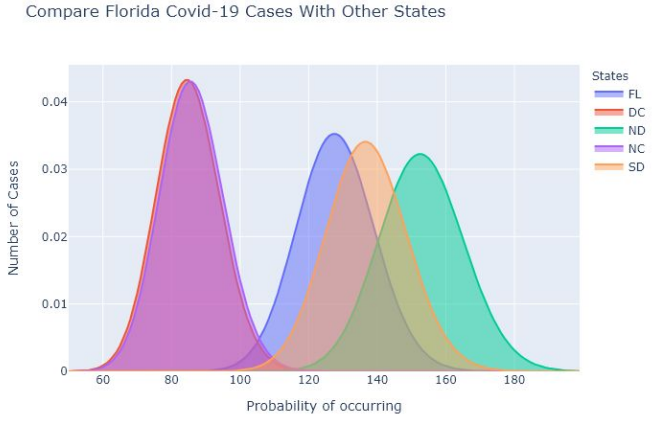
Task 2:

I think poisson distribution is the best choice here to distribute COVID-19 data since it's discrete. So, I fit the distribution to a Florida state and this what i got after normalization per 1M:

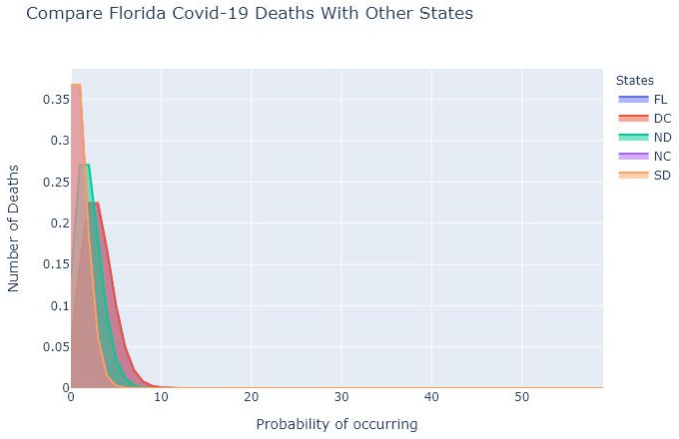


I got this shape by using probability mass function and I tested all the possibilities between 750-1600 Covid-19 cases. This graph shows the probability of having that amount in a specific amount of cases. For example, the chances to have 124 covid cases is 0.005837% which is low compared to the mean probability. I then compared the probabilities for four other states and compared them with Florida state (after normalization).

Compare Florida Covid-19 Cases With Other States

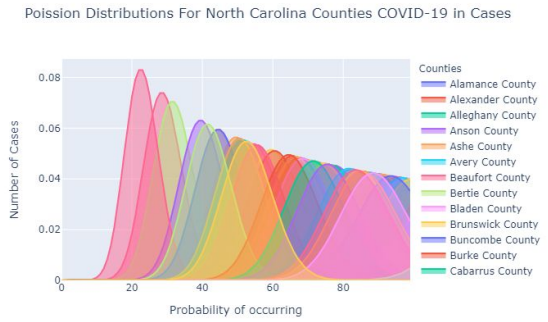


Compare Florida Covid-19 Deaths With Other States

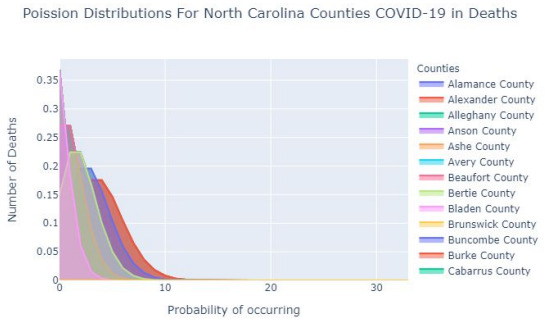


Then I Model poisson distributions for all North Carolina counties COVID-19 cases and deaths.

Poisson Distributions For North Carolina Counties COVID-19 in Cases



Poisson Distributions For North Carolina Counties COVID-19 in Deaths



I then performed a correlation between Enrichment data variables and COVID-19 cases for Florida. The variables that I selected are from hospital data because I was curious about the correlation between number of beds and deaths.

```
Corralation between new Death and number of ICU beds: -0.019879705854529493
Corralation between new Death and avg ventilator usage: -0.04101837630256754
Corralation between new Death and number of staffed beds: -0.04454985654967586

Corralation between new Cases and number of ICU beds: 0.020085438865893222
Corralation between new Cases and avg ventilator usage: -0.016252742032337897
Corralation between new Cases and number of staffed beds: -0.01939198087606299
```

The Correlation between Florida's Deaths and number of ICU beds is negative, meaning that as more the number of ICU beds increase the less deaths happen in the state. That also true in case of avg ventilator usage and staffed beds. On the other hand, the correlation between new cases and number of ICU beds is positive. I think it's positive because the more beds in the hospitals the more people who don't have covid visit their family and they get covid from other people who are in the other rooms.