UNCG-CSC 405-605 (Data Science) Spring 2022

PROJECT - COVID DATA ANALYSIS

Team 3

We computed the number of new cases and deaths reported in the United States in full weeks since the epidemic began in this step of our investigation. For both new numbers of deaths and cases, we generated a measure of centre (mean) for each week separately and all the observed weeks combined. Using this data, we could compare the United States against five other countries we chose. We were also able to observe the virus's spread daily using the weekly data by plotting the data. The team's final goal was to determine peak weeks for both cases and deaths in the United States and the other countries they were comparing.

Each member replicated the above experiment for a state of their choosing in the United States. We compared the rate of new cases and deaths to five other states of our choosing. We chose the top five counties in our selected state with the most significant rate of new cases and deaths. By charting the daily trends of the county data, we can then examine them. One of the critical goals of this part of the research, as previously said, was to explore our data and fit it into one of the numerous statistical models that we had studied.

TEAM TASK

Compare the weekly statistics (mean, median, mode) for number of new cases and deaths across US.

In this task, we have compared the weekly statistics of new cases and deaths across the US. The mean, median and mode for the weekly statistics is as below:

For new cases across the US

Mean	63221
Median	45978
Mode	1

For deaths across the US

Mean	1065
Median	805
Mode	0

Compare the data against other countries of the world. Plot weekly trends of US and compare other countries.

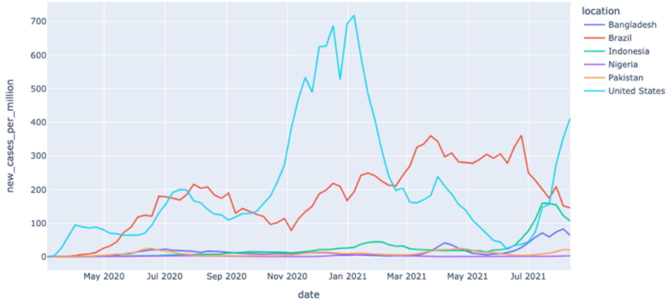


Fig1: Plot weekly trends (cases) of US and for the other five countries

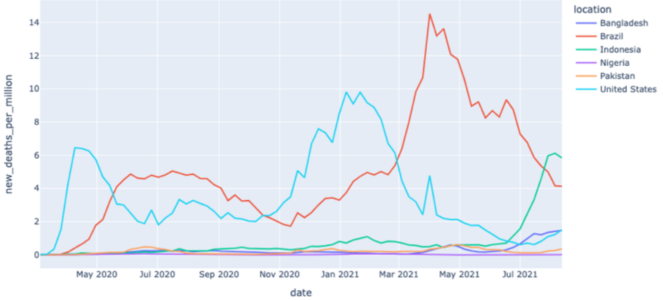
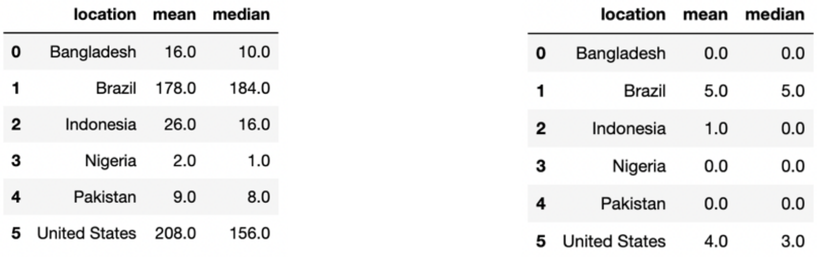


Fig2: Plot weekly trends (deaths) of US and for the other five countries

Here, we have chosen five other countries, namely Indonesia, Pakistan, Brazil, Nigeria, Bangladesh, to compare the statistics of these countries with the US. While comparing the US with the other five countries, we observed that the US follow different pattern from the other five countries. Among the selected five countries, Brazil follows different pattern while all the other four countries follow the similar pattern though the numbers are different.

The mean and median of the cases and deaths across five different countries along with the US is as below:

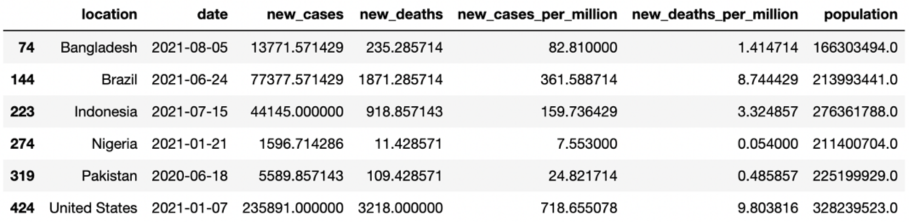


	location	mean	median
0	Bangladesh	16.0	10.0
1	Brazil	178.0	184.0
2	Indonesia	26.0	16.0
3	Nigeria	2.0	1.0
4	Pakistan	9.0	8.0
5	United States	208.0	156.0

	location	mean	median
0	Bangladesh	0.0	0.0
1	Brazil	5.0	5.0
2	Indonesia	1.0	0.0
3	Nigeria	0.0	0.0
4	Pakistan	0.0	0.0
5	United States	4.0	3.0

Fig3: The mean and median of the covid cases and deaths after normalizing

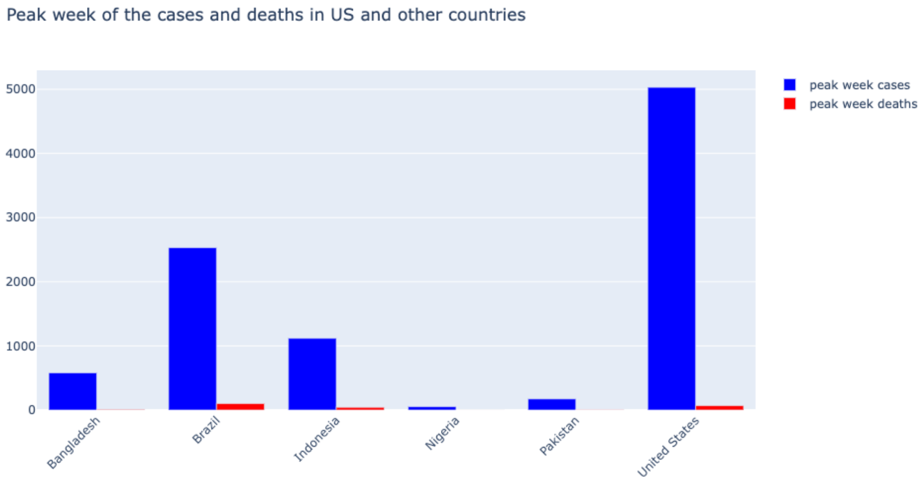
Peak weeks of the cases and deaths in US and other countries:



	location	date	new_cases	new_deaths	new_cases_per_million	new_deaths_per_million	population
74	Bangladesh	2021-08-05	13771.571429	235.285714	82.810000	1.414714	166303494.0
144	Brazil	2021-06-24	77377.571429	1871.285714	361.588714	8.744429	213993441.0
223	Indonesia	2021-07-15	44145.000000	918.857143	159.736429	3.324857	276361788.0
274	Nigeria	2021-01-21	1596.714286	11.428571	7.553000	0.054000	211400704.0
319	Pakistan	2020-06-18	5589.857143	109.428571	24.821714	0.485857	225199929.0
424	United States	2021-01-07	235891.000000	3218.000000	718.655078	9.803816	328239523.0

Fig4: Peak weeks of the cases and deaths in US and other countries

This task shows that covid cases and deaths are high in January for the United States. This rise maybe because of the holiday season. Brazil has the highest death rates than the other five countries. It has the highest peak in the summer season. This might be because of the summer holidays and the population it has. And, in the rest of the three countries have the highest peak of the cases during summer break.



The above plot showed that the United States has the highest rise even in the peak week of the cases and deaths followed by Brazil.

MEMBER TASK

Aman Tej Vidapu, Graduate student

Generate weekly statistics:

I took New York state to do analysis. Mean and median values of mean weekly cases across new york 3835, 1756. Mean and median values of mean weekly deaths across new york 93,

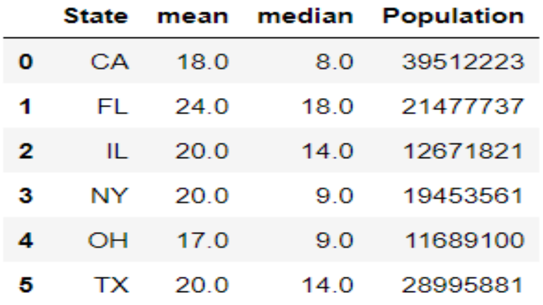
33. (19 million population) - 0.0002. Mode is '0'. I also found mean, median and mode values for the total sum of cases/deaths across New York.

Compare the data against other states and plot trends:

As my task 1 state is New York, I considered taking states which are closely populated as new york or more populated than new York. They are California, Florida, Illinois, Ohio and Texas. So, in order to compare we have to normalise the cases/deaths according to population per each state. Here the selected states population ranges between 12-39 million. So, we are normalising cases/deaths per every 100000 people in the state. Also, we are performing analysis on weekly generated mean cases/deaths across the selected states.

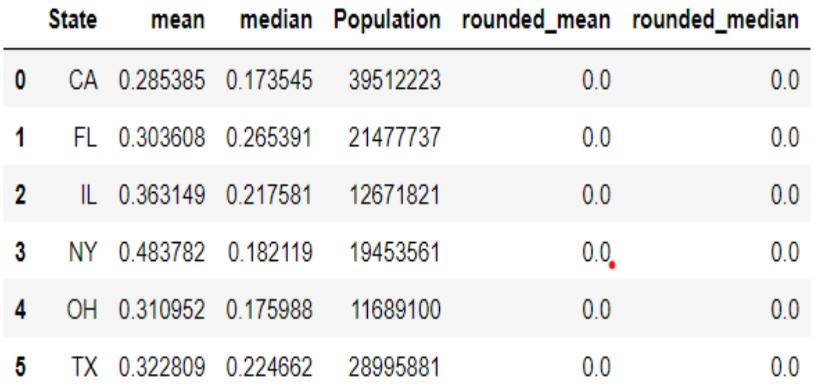
After grouping the data by weeks and normalising. We are calculating mean and median values for all the selected states cases/deaths and making comparisons.

Cases:



	State	mean	median	Population
0	CA	18.0	8.0	39512223
1	FL	24.0	18.0	21477737
2	IL	20.0	14.0	12671821
3	NY	20.0	9.0	19453561
4	OH	17.0	9.0	11689100
5	TX	20.0	14.0	28995881

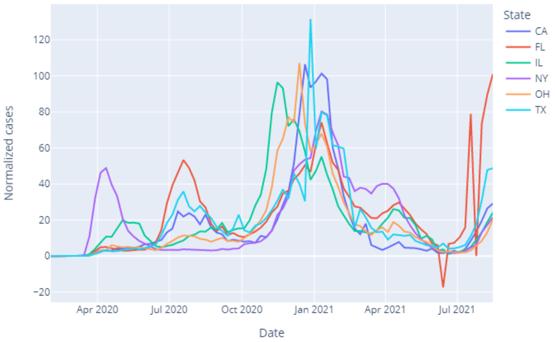
Deaths:

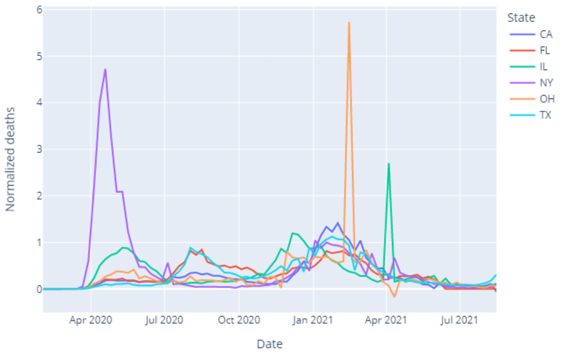


	State	mean	median	Population	rounded_mean	rounded_median
0	CA	0.285385	0.173545	39512223	0.0	0.0
1	FL	0.303608	0.265391	21477737	0.0	0.0
2	IL	0.363149	0.217581	12671821	0.0	0.0
3	NY	0.483782	0.182119	19453561	0.0	0.0
4	OH	0.310952	0.175988	11689100	0.0	0.0
5	TX	0.322809	0.224662	28995881	0.0	0.0

As the data is already normalised, we cannot compare correctly on deaths. But, as of cases we can see states with high populations even have moderate mean, median values. Whereas some states with less population have more cases. We can make hypotheses like less Medicare, educated people.., other factors for these differences.

Here are the visual plots to understand peaks and dip trends of cases/deaths for selected states.





Ohio has more death rates and Texas has the most number of confirmed cases. We can also see 2-3 peaks for every state, most common around December January timeline which is a holiday period. New York has one surge with respect to deaths but normalised later. Illinois and Ohio had a surge for deaths during summer and the rest of the states maintained simple trends. At initial hit with covid in New York, there is no medical knowledge about covid. So, the death rate is high. Again with seasonal changes we can observe the peaks in both trends. One more observation that can be made is after a new_cases surge during dec-jan 2020(holiday season) also influenced the death rate by feb-march 2021.

We can observe the case trends similar to the complete United states trend which is present in Team task (satge_2). Simply, these states push trends across the United states.

Death's trend has the first peak match but, gradually trends do not match as exactly because we have different normalisation factors.

Identify five counties within a state of your choice with high cases and death rates:

Top 5 Infected counties of New York in terms of total cases count in that county:

- Bronx County
- Nassau County
- Suffolk County
- Queens County
- Kings County

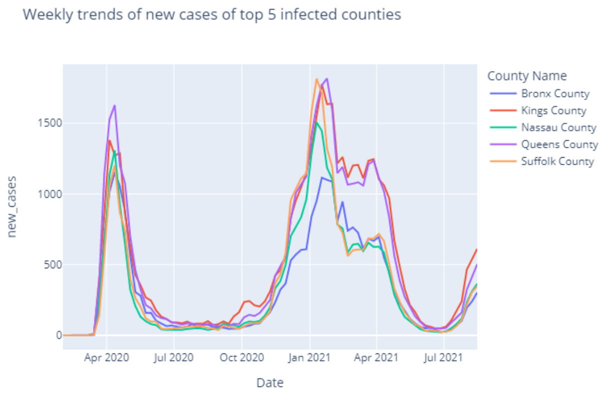
Top 5 Infected counties of New York in terms of total cases count in that county:

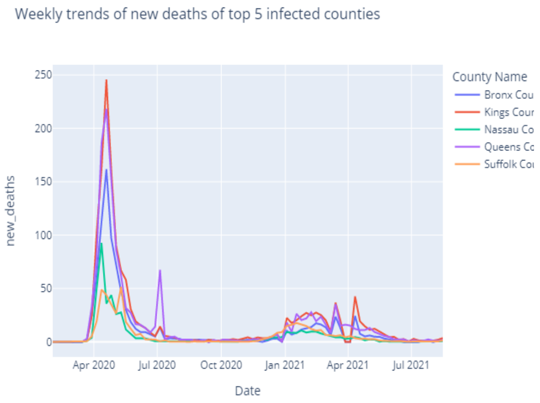
- Bronx County
- New York County
- Suffolk County
- Queens County
- Kings County

Only New York county seems to be having more deaths than the confirmed cases rate. This is an interesting observation that can be more analysed with proper enrichment data.

Trends:

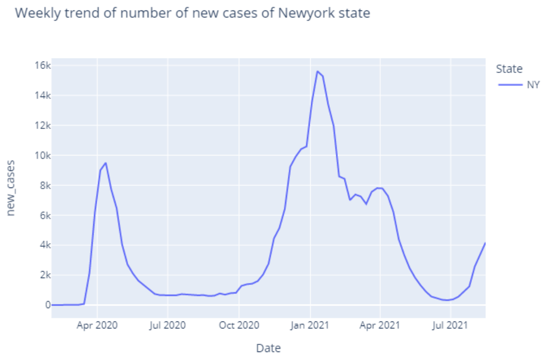
Here are the plots of new_cases, new_deaths per week in these top 5 infected counties. We can understand how one county can influence other counties, as all of them fall into one state. We can also understand the populations and trends of each county. We are normalising the cases or deaths per 10000 people for conty.

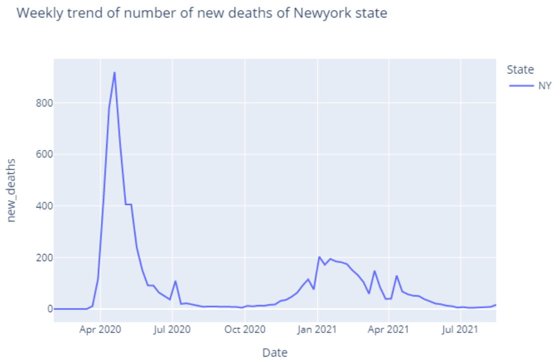




We can see three peaks, the first wave is where covid virus is pretty new and proper knowledge of how to treat it. So, the death rate is high. Later two peaks are in dec-jan 2020, which is an effect of holidays and same lead to hike in death rate. Due to variant in covid, we can see new peak after april 2021.

These county trends match overall New York cases and deaths trends. We can say that these counties are like driving forces for that actual state trend.





Lahari Chilakuri, Graduate Student

Generate weekly statistics:

In the task-1 of the member task, I have chosen the California (CA) state and analyse its weekly statistics regarding the number of new cases and death reported on the week-to-week basis.

The mean, median and mode of the number of cases and deaths per week in the California state have been calculated.

The mean, median and mode for the number of cases per week are

```
mean      127.506024
median     56.000000
mode       0
```

The mean, median and mode for the number of cases per week are

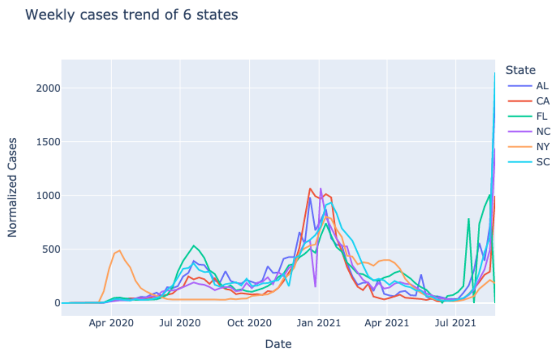
```
mean      2.012048
median    1.000000
mode      1
```

Compare the data against other states:

The statistical results which were obtained for the California state are being comparted with five other states which are North Carolina (NC), South Carolina (SC), New York (NY), Florida (FL) and Alabama (AL). To make these comparison's, I have calculated the number of cases and deaths per day, then these cases and deaths are grouped to find out the number of cases and the deaths per week for all the six states.

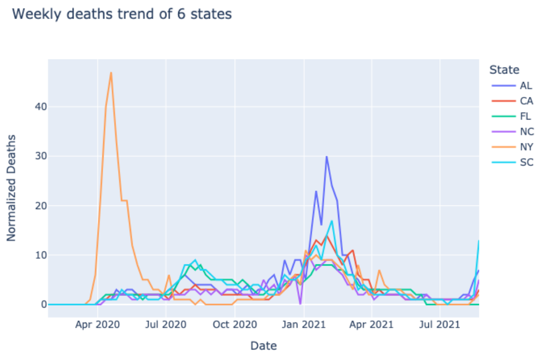
For these data, normalization is performed as the population is not same in every state. The normalization of the number of cases and deaths are done with respective to the population i.e., the number of cases and deaths are represented for every 10000 people in the state.

Below figure shows the plot of the cases for the 6 states:



Here, I observed that all the states follow the same pattern though the numbers are different. The pattern in which the cases were reported are almost similar for all the six states.

Below figure shows the plot of the deaths for the 6 states:



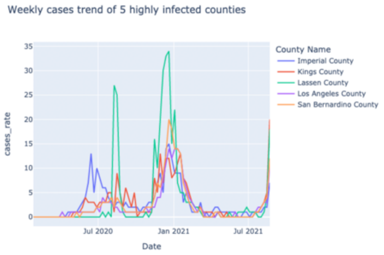
But in the plot of deaths among the six states, I have observed that New York state has highest number of death rates followed by Alabama when we compared to the other states. While the rest four states follow the similar pattern among which almost have their highest peak of deaths in between February and March.

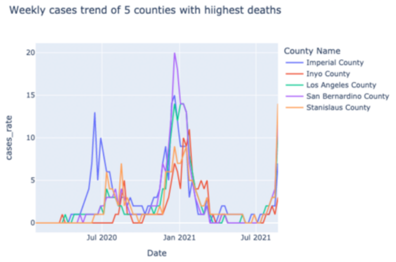
Do they compare with the US pattern?

When we compare the above plots with the US plots, it can be said that the cases follow the similar pattern, where both have the highest peak point in the month January. But the plots of the deaths don’t have any similar pattern as they have different peak points.

Identify five counties within a state of your choice with high cases and death rates:

Here, I did the county level assessment to find five counties within a state with high cases and death rates. Th population of the counties is normalized. The top five counties within California state with the highest number of cases are Lassen, Imperial, Kings, San Bernardino and Los Angeles. The top five counties within California state with the highest number of deaths are Imperial, San Bernardino, Stanislaus, Los Angeles and Inyo counties.





The above two plots show the plots for top five counties with high number of cases and deaths. Here, the counties follow almost the similar pattern in the number of cases and deaths except at some points.

The Lassen County has peak point in the cases San Bernardino County has the peak point in deaths. This is because of the population and the count of hospitals the counties had. All the five counties follow almost the similar patter though there is a difference in the numbers.

Varsha Veeramaneni, Graduate Student

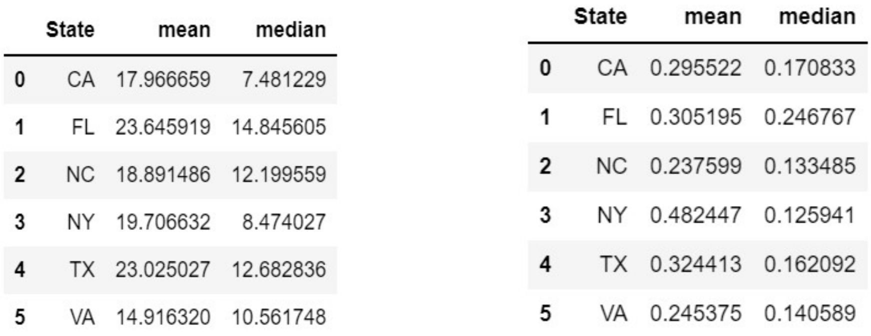
Generate weekly statistics:

As part of the member task, I chose the state of Virginia (VA) and analyzed its weekly statistics on the number of new cases and deaths reported weekly. We determined the mean, median, and mode of cases and deaths per week in Virginia.

Compare the data against other states:

The statistical findings for Virginia are compared to those for five other states: North Carolina (NC), Florida (FL), New York (NY), Texas (TX), and California (CA). To make these comparisons, I estimated the number of cases and deaths per day to determine the number of cases and deaths per week for each of the six states. The data is normalized by the population because the population in each state is different.

Below figure shows the plot of the cases for the 6 states:



	State	mean	median
0	CA	17.966659	7.481229
1	FL	23.645919	14.845605
2	NC	18.891486	12.199559
3	NY	19.706632	8.474027
4	TX	23.025027	12.682836
5	VA	14.916320	10.561748

	State	mean	median
0	CA	0.295522	0.170833
1	FL	0.305195	0.246767
2	NC	0.237599	0.133485
3	NY	0.482447	0.125941
4	TX	0.324413	0.162092
5	VA	0.245375	0.140589

Fig1: The mean and median of the covid cases and deaths after normalizing

From the above data, I observed that the Weekly mean number of cases for Florida is the highest and the weekly mean number of cases for Virginia is the lowest among the 6 states. And also, Weekly Median number of cases is the highest for Florida and lowest for California.

Weekly mean number of deaths for New York is the highest and the weekly mean number of deaths for NC is the lowest among the 6 states. Weekly Median number of deaths is the highest for Florida and lowest for New York

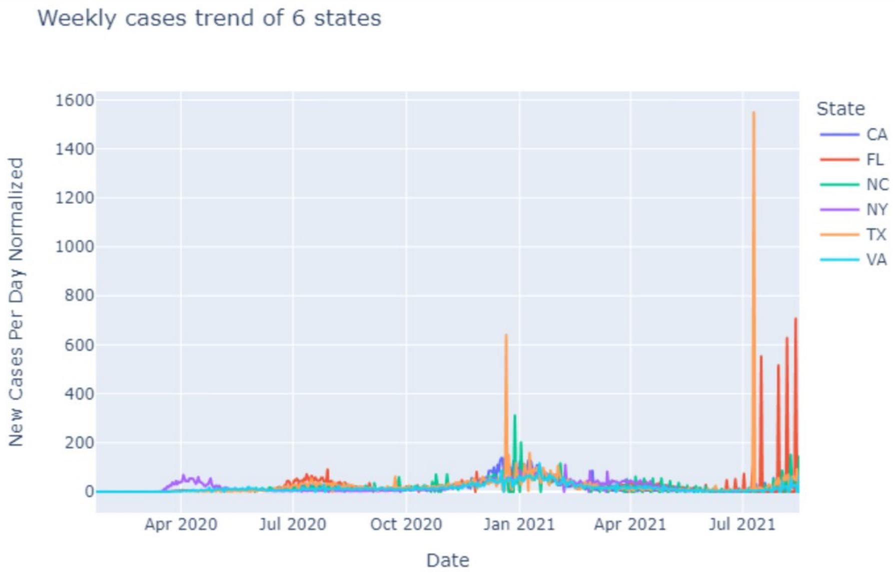


Fig2: shows the plot of the deaths for the 6 states

From the above figure, I observed that though Texas state has less Weekly mean number of cases than Florida, at some point, the rise in cases is very high than in Florida. And remaining states follow the same pattern though the numbers are different.

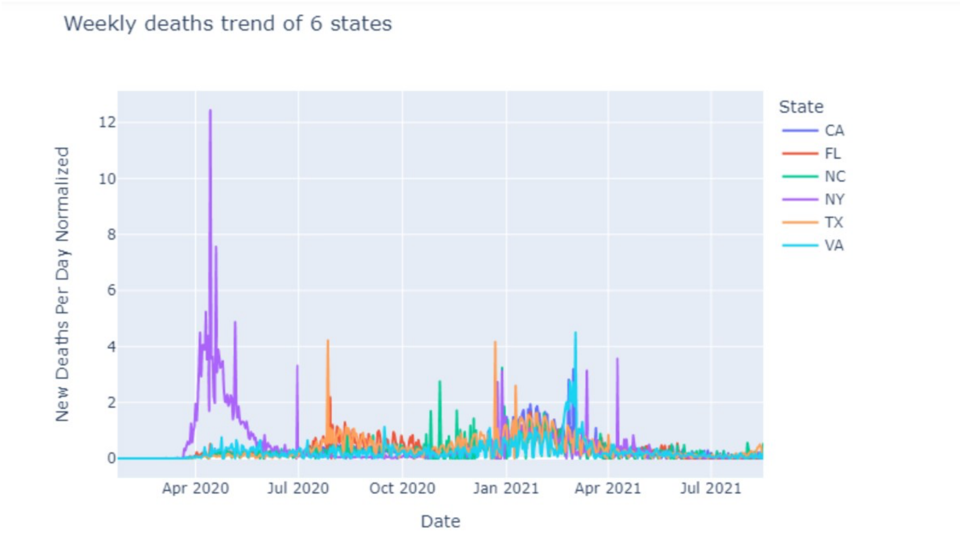


Fig3: shows the plot of the deaths for the 6 states

But in the plot of deaths among the six states, I have observed that New York state has highest number of death rates followed by Texas when we compared to the other states. While the rest four states follow the similar pattern among which almost have their highest peak of deaths in between January and April.

Do they compare with the US pattern?

When we compare the above plots with the US plots, it can be said that the cases follow a similar pattern, where both have the highest peak point in January, and after that, there is a fall in the graph, and again from June, the cases increased slightly.

In the plots of the deaths, both figures follow the almost same pattern, where the number of deaths is high in May and later in January of the following year.

Identify five counties within a state of your choice with high cases and death rates:

In this task, to find five counties within a state with high cases and deaths I did the county level assessment. The population of the counties is normalized. The top five counties within Virginia state with the highest number of cases are City of Galax, City of Lexington, City of Franklin, Richmond County and City of Buena Vista. The top five counties within Virginia state with the highest number of deaths are City of Galax, City of Emporia, City of Martinsville, City of Lexington and City of Franklin.

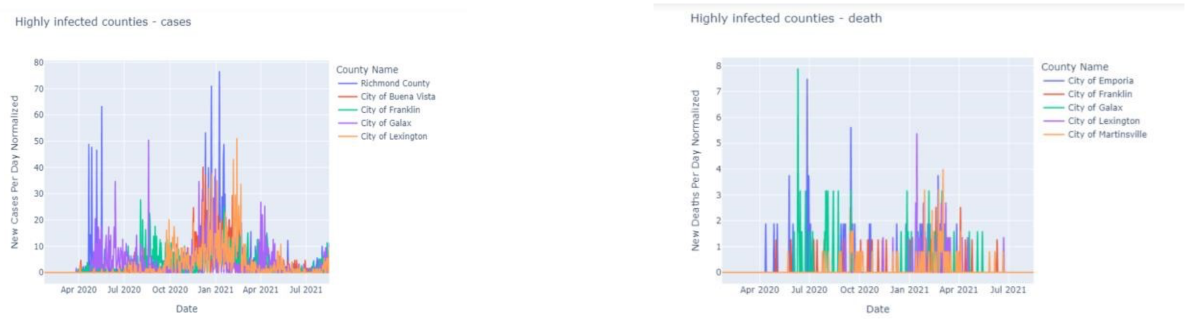


Fig4: Top five infected counties(cases and deaths)

The above two plots show the top five counties with a high number of cases and deaths. After normalization by population, it can be observed that Richmond county has been highly affected in some points. This may be due to the greater population density compared with other counties. And also, we can say that the deaths in the City of Galax have a peak point; maybe this is due to the lack of medical facilities in the initial days of the Pandemic. Here, all the counties follow almost a similar pattern of state pattern in the number of cases and deaths except at some points.

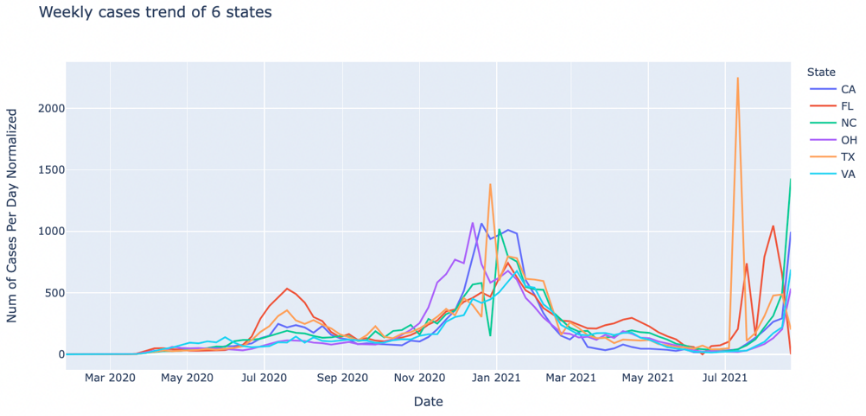
VIVEKA REDDY ERRAM, Graduate Student

For the individual task, I have chosen Florida state to analyze its weekly statistics for the number of covid-19 cases and the deaths. Here I have chosen Florida state because there are many elderly people in FL, So I just want to see the how many numbers of covid-19 cases and deaths occurred in FL.

First, I have calculated the weekly statistics for FL state. Then, from the results obtained from the Florida (FL) state are compared with five other states weekly statistics which are North Carolina (NC), Texas (TX), California (CA), Virginia (VA), Ohio (OH). To make these comparisons, first I have calculated the number of new cases and deaths reported per day. Then I have normalized, as every state doesn't have the same population. So, the number of covid cases and deaths are normalized by the population so that numbers are representative per 100,00 people. The results for these comparisons are reflected in detail in my project notebook.

Later I have found the top 5 counties within Florida with the highest normalized means of covid-19 cases and deaths. The top five counties for new cases per week are calculated based on population per 1000 people. After getting the normalized weekly covid cases, deaths mean based on population the top five counties with high covid cases are statewide unallocated, Lafayette County, Miami-Dade County, Gulf County, Liberty County. Five counties with highest covid death cases are Union County, Calhoun County, Jackson County, Highlands County, Suwannee County.

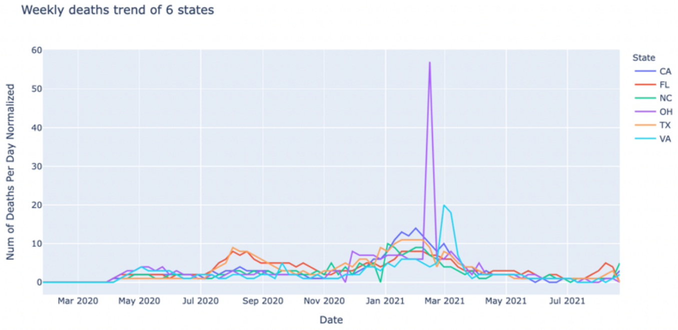
Weekly number of covid cases:



Analysis

- We can see that by July 2021 the cases in TX have increased there are around 2600 cases for every 100,000 people.
- In Jan 2021 in Texas there are around 1800 covid cases per 100,000 people.
- In May 2020 ,50 to 100 cases in all the 6 states for every 100,000 people.
- From Jan 2021 the cases started to fall from 500 cases to around 50 cases till July 2021
- We can see the cases in July are high because of the long weekend (July 4th). This might be one of the reasons for the increase in cases.
- We could also see that the covid-19 cases started to increase from Nov 2021 till Jan 2021 end, this might be because of Thanksgiving gathering, Christmas and new year celebrations.
- Finally, we can analyze and say the covid- 19 cases are high during the holiday season.

Weekly number of covid deaths:



Analysis

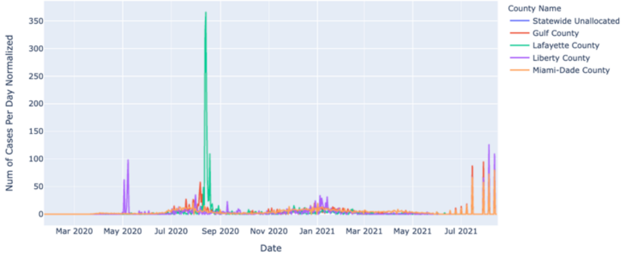
- Actually, I covid-19 deaths in all the states which I have chosen have similar weekly death pattern, except California.
- In California the covid-19 deaths are high from around mid of Feb 2021 to March 2021.
- In Virginia, there are many covid-19 deaths from Feb 2021 till April 2021, till then the death cases are very low in VA.
- Again, the death cases have increased after July (might be because of July-4th long weekend) there are more spread and death cases
- I could see the death cases are probably increasing after a month of long weekends that might be because of vacations.

Do they compare with the US patterns?

When we compare the above plots with the US plots, it can be said that the cases follow the similar pattern, where both have the highest peak point in the month January. But the plots of the deaths don’t have any similar pattern as they have different peak points.

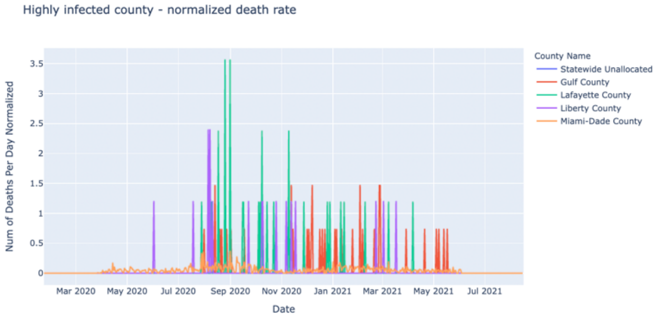
The below figure shows the comparison between the data for these counties for new cases per week.

Top 5 highly infected county- normalized cases rate:



- we can see that the covid cases in Miami-Dade County was less from March 2020 to May 2021 all of sudden the cases increased in July 2021.There are around 50 to 60 cases for every 1000 people
- In September 2020 the covid cases in Lafayette County has sky rocketed. There are around 350 to 360 cases per 1000 people
- In May 2020 there are many cases in Liberty County

Top 5 highly infected county- normalized death rate:



- Similar to the covid -19 cases the deaths are also more in Lafayette County in the month of September 2020
- Initially in May 2020 the deaths in Liberty County are more
- we can see the deaths in Miami county are less compared with other counties
- we can also see there are a greater number of deaths in Dec 2020, March 2021 and April, May and June 2021, there are no deaths in Gulf County from September 2020 to December 2020.

SAIPAVAN TADIKONDA, Graduate Student

Generate weekly statistics for a specific state:

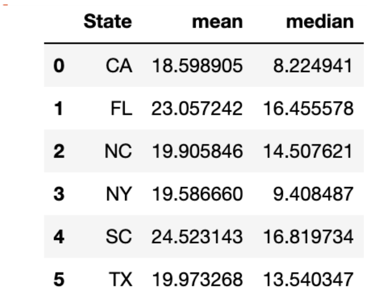
I have chosen North Carolina State. I was able to find the individual new cases and new deaths by using the diff function and converted the daily new cases and new deaths into the weekly data. And on this weekly data I determined mean median and mode for weekly new cases and weekly new deaths.

Compare the chosen state data with other five states:

The States which I have chosen for comparing are as follows:

1) California (CA)
2) Florida (FL)
3) New York (NY)
4) South Carolina (SC)
5) Texas (TX)

I obtained the individual new cases and new deaths across the selected state and normalized the obtained new cases and new deaths by population per 100000 as every state differs in the population. Then I converted the daily data into weekly data for further comparisons. The mean and median for weekly normalized new cases is as follows:



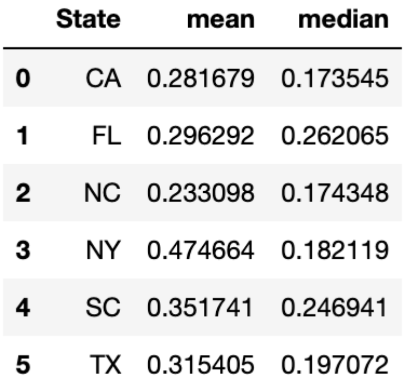
	State	mean	median
0	CA	18.598905	8.224941
1	FL	23.057242	16.455578
2	NC	19.905846	14.507621
3	NY	19.586660	9.408487
4	SC	24.523143	16.819734
5	TX	19.973268	13.540347

Fig. 1

From the above Fig1. We can observe that the mean is more for the South Carolina state followed by the Florida state. The mean of the California state is least among the six states.

We can also observe that the median for South Carolina is more followed by Florida.

The mean and median for weekly normalized new deaths is as follows:



	State	mean	median
0	CA	0.281679	0.173545
1	FL	0.296292	0.262065
2	NC	0.233098	0.174348
3	NY	0.474664	0.182119
4	SC	0.351741	0.246941
5	TX	0.315405	0.197072

Fig. 2

From the above Fig. 2 we can observe that the mean is more for the New York than other states. The North Carolina has least mean. Coming to the median the florida has the highest median.

Identify the peaks:

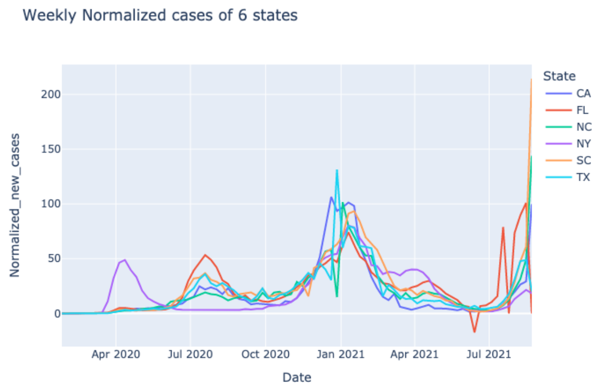


Fig. 3

From Apirl 2020- July2020 there is peak for the Normalized new cases for the state New York. From July 2020- October 2020 Florida State registered its peak. But from Nov 2020- March 2021 all the six states have registered peaks in between these months. After April 2021 there is a downfall in the cases across all the states. Again, after July 2021 we can observe the increase in cases. So, we can understand up to an extent that the spread of covid cases across these states is more in winter months compared to remaining months.

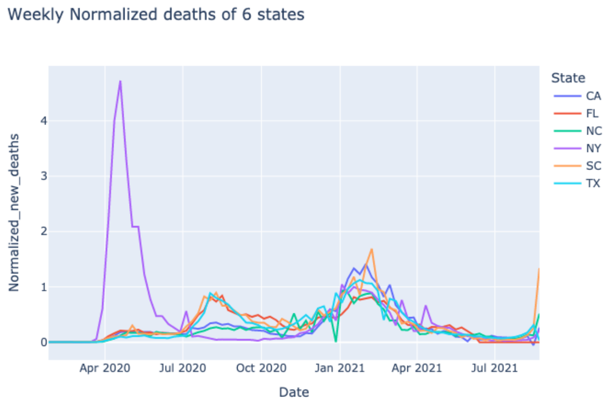


Fig. 4

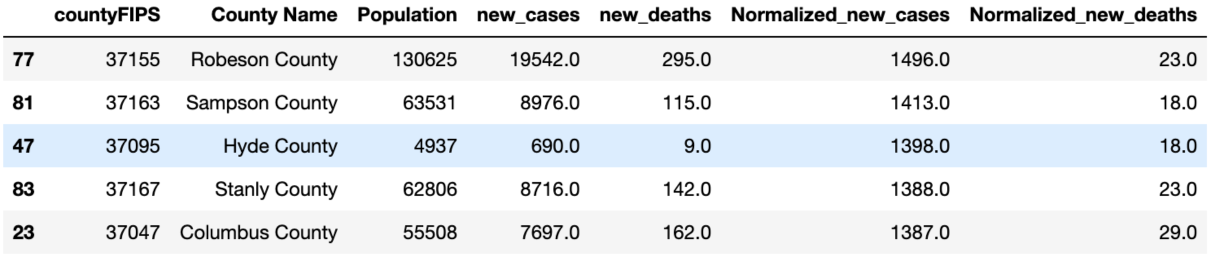
From the above Fig. 4 we can observe that for the State New York in between April 2020-June 2020 there is large peak in the death. It registered its highest peak in between these months. And the states like Florida, Texas, South Carolina showed some peaks in between August 2020- September 2020. All the six states registered their peaks in between the Jan2021-March 2021.

Do they compare with US patterns?

Yes, they can be compared with the US patterns. If we observe US case patterns there are a greater number of cases and deaths in between November 2020-March 2021. The above taken six states also showed their peaks in between these months. So, we can say that these patterns matched with US patterns.

Identify top five counties in NC having high cases and death rates:

The top counties having high cases is as follows:

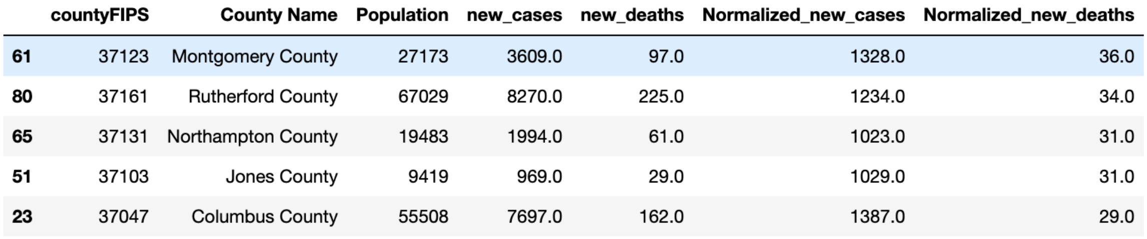


	countyFIPS	County Name	Population	new_cases	new_deaths	Normalized_new_cases	Normalized_new_deaths
77	37155	Robeson County	130625	19542.0	295.0	1496.0	23.0
81	37163	Sampson County	63531	8976.0	115.0	1413.0	18.0
47	37095	Hyde County	4937	690.0	9.0	1398.0	18.0
83	37167	Stanly County	62806	8716.0	142.0	1388.0	23.0
23	37047	Columbus County	55508	7697.0	162.0	1387.0	29.0

Fig. 5

Robeson County in NC had the highest cases.

The top counties having high death rates is as follows:



	countyFIPS	County Name	Population	new_cases	new_deaths	Normalized_new_cases	Normalized_new_deaths
61	37123	Montgomery County	27173	3609.0	97.0	1328.0	36.0
80	37161	Rutherford County	67029	8270.0	225.0	1234.0	34.0
65	37131	Northampton County	19483	1994.0	61.0	1023.0	31.0
51	37103	Jones County	9419	969.0	29.0	1029.0	31.0
23	37047	Columbus County	55508	7697.0	162.0	1387.0	29.0

Fig. 6

Montgomery County in NC had the highest death rate.

Plot the weekly trends for the top 5 infected counties:

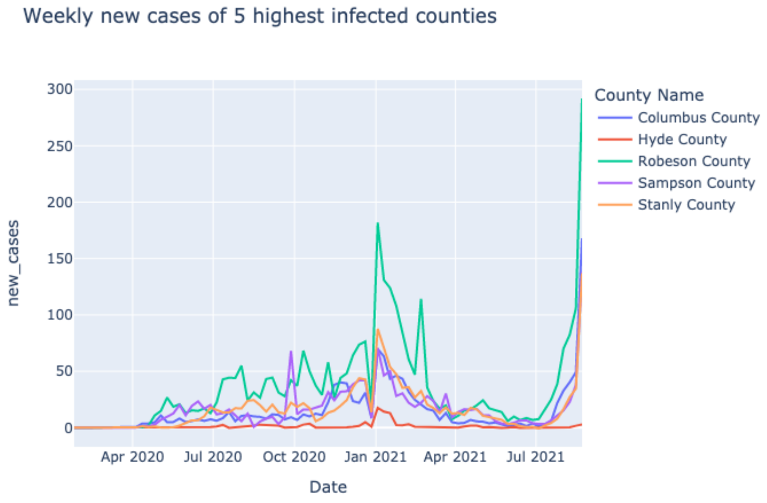


Fig.6

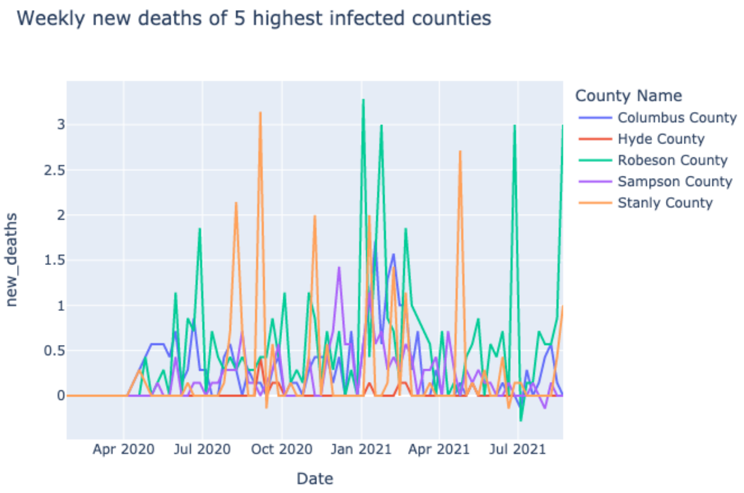


Fig. 7

From the above figures 6 & 7 we observe that all the top 5 infected counties registered their highest peaks in between Jan 2021 and March 2021. The reason might be because of cold climatic conditions the spread might be huge among these counties. After April 2021 there was complete downfall in the cases. And again, after July 2021 there is rapid increase in the cases. This might be because of a new variant among these counties.

Coming to the deaths among these counties different county's registered their peaks in different months. For example, Stanly county had registered its peak in September 2020. This might be because of lack of Medical facilities in that county. Robeson County had registered its highest peak in January 2021. This might be because of cold climatic conditions.

Do the counties follow State pattern?

Yes, all these five counties follow State patterns. Even if we look into the NC state patterns highest peaks of cases and deaths are registered in between Jan 2021- March 2021. Even in the top5 counties the highest peaks are in between Jan2021-March 2021.