Covid Data Analysis Report: Spring 2022

Priyanka Budavi
University of North Carolina Greenboro
Email: p_budavi@uncg.edu

1

Abstract. *Coronavirus is a continuing worldwide pandemic, which has affected a lot of people. Our goal of the project is to develop an analytical framework to study the data coming from United States to understand patterns of COVID-19 effect and spread. The project was divided in five stages and each stage goal is as follows:*

- *Understanding the data*
- *Data Modelling*
- *Distributions and Hypothesis Testing*
- *Basic Machine Learning*
- *Creating the Dashboard*

1 Stage 1 : Downloading the data-sets and analyzing the effect of covid on various verticals of the nation

In Stage-1 we are trying to understand the different Enrichment stage and relate with the COVID-19 data-set. Using Python libraries we perform data wrangling with the COVID and enrichment data-set so that these files could be merged to analyze the spread of COVID19 pandemic.

The various verticals like demography , hospitals , politics and employment had serious effect because of the COVID19 and in this stage we are trying to analyze how each sector was affected due to Covid19 so that this information could help government agencies to take actions to prevent the spread of COVID19. For example, the hospital data-set gives information of how many beds are utilized due to covid and if a new patient wants to get admitted at a hospital the availability of beds information can be derived from the analysis of the project.

1.1 Preliminary Intuitions:

- **Confirmed covid cases:** When analyzing the confirmed cases of COVID-19 by county, we see that the confirmed cases start off slow in January and February 2020. By mid-March 2020, we can see that cases begin to exponentially increase in more populous counties in states on the east and west coasts, such as Nassau County in New York and Orange County in California. By early-to-mid April 2020, this trend has spread throughout the country. The number of new cases daily does not appear to start subsiding until early 2021, which coincides with the public administration of COVID-19 vaccines in the United States. This trend generally continues throughout Spring 2021, though we start to upticks in July and August 2021, which coincides with the arrival of the Delta variant of COVID-19 in the United States.
- **Covid Deaths:** For COVID-19 deaths by county, the trends we saw above can be found here as well. However, the trends occur a few weeks later than with COVID-19 cases. This seems logical, as people who end up dying from COVID do so a few weeks after contracting it. As such, we would expect daily COVID cases deaths to increase or decrease a few weeks after increases and decreases in daily COVID-19 cases.
- **County Population:** The need for the county population becomes apparent as we analyze confirmed cases and deaths. While the trends above are generally true for all counties, the raw numbers themselves vary greatly. Counties like Orange County in California have hundreds of thousands of confirmed cases, while others like Graham County in Arizona have less than 6,000. At first glance, this could lead someone to believe that certain counties may be under-reporting COVID-19 numbers. However, by bringing in the county population, we can see that Graham County only has a total population of less than 40,000 people. With that additional context in mind, 6,000 confirmed cases seem much more plausible. In order to account for these population discrepancies, we will be standardizing the COVID-19 confirmed cases and deaths. This will minimize the impact of population size and will put all counties on the same scale for analysis purposes.

1.2 Dataset used for analysis

In this task we merged the three datasets to perform the analysis :

- **confirmed_usafacts.csv** Gives info about no. of covid confirmed cases
- **covid_deaths_usafacts.csv** Gives info about no. of covid death cases
- **covid_county_population_usafacts.csv** Gives info about population in county and state

1.3 Pre-processing the dataset

The above CSV file were merged to perform the analysis of COVID19. Firstly, all the CSV files were converted from wide to long format and then each of the columns in the dataset were cleaned and converted to a particular datatype to perform opeartions. Below is the screenshot of the merged dataset. The code is available on the github.

	countyFIPS	County_Name	State	StateFIPS	Date	Confirmed	Deaths	population
0	0	statewide unallocated	AL	1	2020-01-22	0	0	0
1	0	statewide unallocated	AL	1	2020-01-23	0	0	0
2	0	statewide unallocated	AL	1	2020-01-24	0	0	0
3	0	statewide unallocated	AL	1	2020-01-25	0	0	0
4	0	statewide unallocated	AL	1	2020-01-26	0	0	0
...	...	...	...	...	...	...	...	...
2346471	56045	weston county	WY	56	2022-02-03	1491	17	6927
2346472	56045	weston county	WY	56	2022-02-04	1496	17	6927
2346473	56045	weston county	WY	56	2022-02-05	1496	17	6927
2346474	56045	weston county	WY	56	2022-02-06	1496	17	6927
2346475	56045	weston county	WY	56	2022-02-07	1508	17	6927

2346476 rows × 8 columns

Exporting the merged file to csv

```
merge_final.to_csv('../../../data/stage_1/superdataset.csv',index = False)
```

Figure 1: Merged Data-set

1.4 Enrichment Data-set Analysis with the COVID data-set

Description: The hospital bed data-set contains information about the hospital resources and the bed utilization values of both confirmed and suspected cases of COVID-19. It is a high dimensional data-set of 54 observations and 117 columns. Out of 117 columns I chose the below variables to define a new data frame which I find relevant to the covid data-set.

Column Name	Datatype	Description
State	String	Name of the state
Inpatient beds	Int	Number of beds
Inpatient beds used	Int	The number of beds used
Inpatient beds used covid	Int	Beds used for covid
critical_staffing_shortage_today_yes	Int	Staff Information
critical_staffing_shortage_today_no	Int	Staff Information
staffed_icu_adult_patients_confirmed_and_suspected_covid	Int	Suspected covid and confirmed cases
staffed_icu_adult_patients_confirmed_covid	Int	confirmed covid cases
total_adult_patients_hospitalized_confirmed_and_suspected_covid	Int	Total confirmed and suspected cases
total_adult_patients_hospitalized_confirmed_covid	Int	Confirmed cases
total_staffed_adult_icu_beds	Int	Number of ICU beds
inpatient_beds_utilization	Float	Beds used
adult_icu_bed_covid_utilization	Float	Beds used for Covid
inpatient_bed_covid_utilization	Float	Beds used for Covid
percent_of_inpatients_with_covid	Float	% Of covid patients
adult_icu_bed_utilization	Float	Beds utilized by adults
reporting_cutoff_start	String	Day wise information
deaths_covid	Float	Number of deaths

Figure 2: Hospital Dataset

Variables that I can relate to the Covid19 Data-set:

- The data types for the entire file are integers, float, or strings.
- The variable that could merge the two data sets is using the column 'State'.

Hypothesis : The hospital bed data-set will help us understand how the pandemic has challenged the critical care capacity and put a strain on the healthcare system. If we compare the number of confirmed COVID-19 cases with the available hospital beds in the same area, we can conclude that patients who need extreme supportive treatment and could not be admitted to the hospital are left with no choice but to self-isolate and treat the virus at home. This further would increase the spread of the virus as members of the family could be infected. Such infected people can spread the virus if they meet others and the chain continues which cannot be controlled and increases the confirmed covid cases.

In this task the three data-sets i.e., confirmed, deaths, population were merged to form a super data-set. Prior to that I performed analysis by selecting ALABAMA as my state of interest to check the trends in COVID cases and deaths. After plotting the graph with the data-set, I was able to understand that initial days of the week the cases were consistent but in the mid-week that is on 3rd Feb 2022 to 4th Feb 2022 there was a sudden spike in the number of covid cases and the deaths. I also noticed that there were no cases reported on the weekends and again there was a rise in covid cases. Thus, I conclude that there is a weekly increase in the covid cases and deaths. Below graphs show the trends of total cases and deaths of last 7 days in ALABAMA state.

```
sum = last_week_coviddata.sum()
print(sum)
plt.plot(sum)
```

```
2022-02-01    1229300
2022-02-02    1229300
2022-02-03    1229300
2022-02-04    1240496
2022-02-05    1240496
2022-02-06    1240496
2022-02-07    1245876
dtype: int64
[<matplotlib.lines.Line2D at 0x1ef329f1f40>]
```

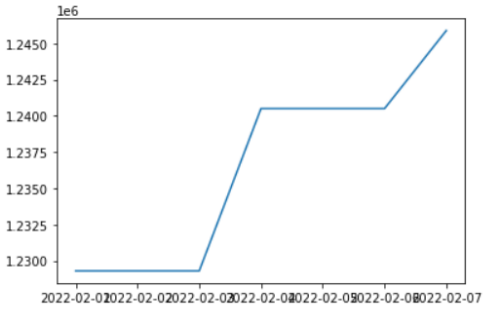

Figure 3: Covid Cases Alabama State

```
sum = last_week_coviddeaths.sum()
print(sum)
plt.plot(sum)
```

```
2022-02-01    17215
2022-02-02    17215
2022-02-03    17215
2022-02-04    17371
2022-02-05    17371
2022-02-06    17371
2022-02-07    17387
dtype: int64
[<matplotlib.lines.Line2D at 0x1ef3324da30>]
```

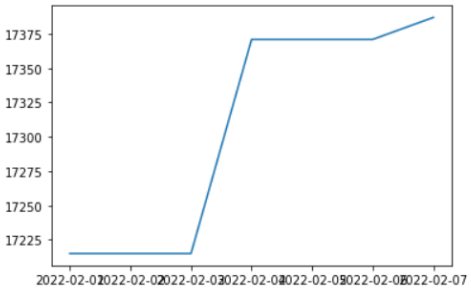

Figure 4: Covid Deaths Alabama State

2 Stage 2 : The goal of Stage II is to develop the data for modeling and comparative analysis.

In Stage -2 we develop the data for modeling and comparative analysis. We compare the cases and deaths of the US with other countries and also analyze the data at state level and county level. We perform statistical analysis and compare how COVID has affected at country , State and county level and visualize the graphs using plotly. For stage-2 world data-set was used https://ourworldindata.org/coronavirus-source-data and here we are trying to compare different countries and trying to understand the trends of cases and deaths.From the worldwide COVID-19 data-set, retain data for countries with population densities similar to that of the US. The countries chosen for comparison are Brazil, Democratic Republic of Congo, Iran, Mexico, Myanmar, South Africa and Spain.Some of the countries chosen have no records of deaths before certain days in 2020. With this in view, data for all the eight countries chosen has been filtered by dates accordingly.We will therefore, be analyzing weekly new cases and deaths for weeks starting from 03/22/2020 to 02/05/2022. In order to compute weekly averages for the world data, first group the data by iso_code to group countries, then compute the differences amongst total_cases and total deaths column values for dates from March 21, 2020 to February 5, 2022. Once the differences have been computed, delete the row corresponding to the date 03/21/2020. Save the daily data with new cases and deaths generated. (NOTE: We use values for 03/21/2020 to compute new cases and deaths for 03/22/2020 which is when we start our analysis). Below are the screenshots of USA.

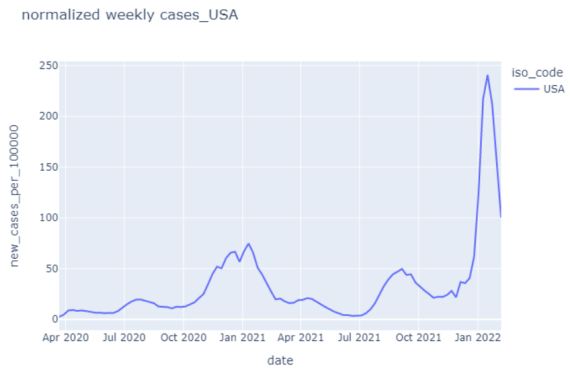

Figure 5: Covid Cases Normalised - USA

Inference : There are three peaks in COVID-19 cases in December 2020-January 2021, August 2021 and January 2022. These seem to coincide with the peaks before the roll-out of vaccines started in the United States in January 2021, the resumption of in-person classes in educational institutions and the third peak (which is the highest) with the rapid rise in cases due to the spread of the infectious Omicron variant. COVID-19 deaths have had multiple peaks as seen in April-May 2020 which saw a huge spike in certain areas of the country as the virus first started spreading. There are subsequent peaks, some of which are not very high. The deaths rose to the maximum in January 2021 and saw a decline after the roll-out of the vaccines. Deaths peaked again around October 2021 in the US (which follows the spike in cases in September 2021 and is as high as the first spike noted in COVID-19 deaths) probably because of a general disregard for the safety protocols. Post this spike, it started to decline before reaching a peak again in January 2022, which can be attributed to the Omicron variant.

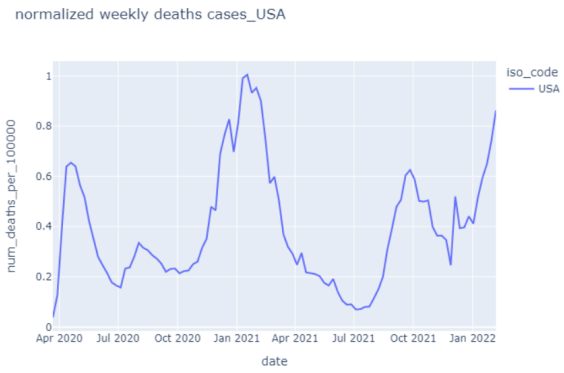

Figure 6: Covid Deaths Normalised - USA

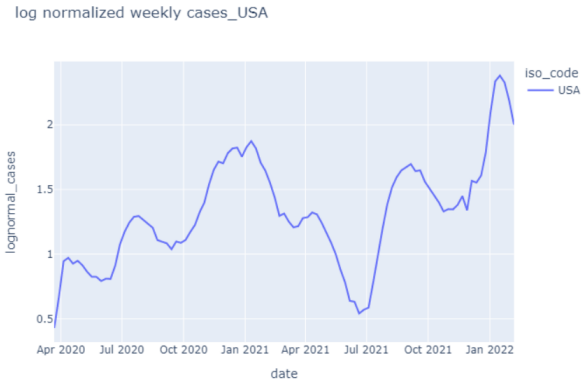

Figure 7: Covid Cases log Normalised - USA

2.1 Comparing USA trends with other countries :

- **Mexico:** The first wave of Covid cases in Mexico was reported in August 2020. The second wave which came in February 2021 was the most impactful in terms of confirmed Covid cases and reported deaths. Mexico city and the State of Mexico most impacted and were assigned red zone which resulted in strict lockdown regulations. The third wave of Covid cases was reported in September 2021. The government had lifted restrictions on social and Business activities which caused a high number of new cases in Covid. January 2022 had an impact due to the Omicron variant which spread throughout

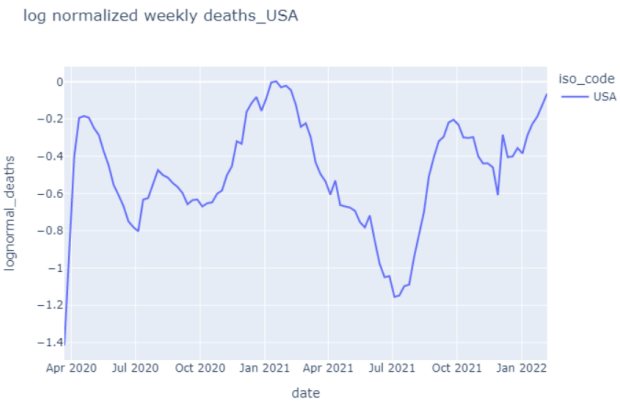

Figure 8: Covid deaths log Normalised - USA

Mexico and caused a high number of Covid cases. The death count decreased since April 2021 as most of the citizens were vaccinated.

- **Iran:** There was a large spike in COVID cases in late April/early May of 2021 that differs from other countries. It's unclear what caused the spike, but it appears that Iranians as a whole were struggling to get vaccinated. The leader of Iran banned imports of American and British vaccines, meaning they had to get Russia's or China's vaccines instead. They had their own vaccines, but struggled to mass produce them. By the time the Delta variant hit, only 2% of citizens were vaccinated. This can also be seen in the COVID deaths, where spikes occur roughly around the same time as each COVID case spike, with an equal amount of severity. This is different from other countries with higher vaccination rates. Lastly, it appears that the Omicron variant is currently hitting the country hard, with a current major spike in cases. Deaths are only now starting to uptick, but we'd anticipate seeing a similar spike soon.

- **Brazil :** In Brazil, the first noticeable peak in covid cases occurred in the month of August 2020 which starts declining by October 2020. The peak occurred because of the covid outspread in winter months(June to September) in the southern hemisphere and a high number of covid testing. So the number of cases and deaths increased. From September 2020 cases again started increasing and reached their second peak in the month of April 2021. The cases are much higher than the first wave because of the Delta variant. Also, a huge peak can be been in the number of deaths because of this. This second wave effect went on until July 2021 after which cases started decreasing. In January 2022, cases started increasing and reached their peak because of the new omicron variant. But as a large population is vaccinated in Brazil so we see no high peak in the number of deaths. Compared to the USA, Brazil has fewer cases and deaths and the trend is different than US. Just the peak in the month of January 2022 is the same. Moreover, Brazil has cases peaks going throughout the plot.

- **South Africa:** Peaks were observed in the month of August 2020 , Jan 2021 , August 2021 and again in the month of Jan 2022. In Aug 2020 , It was the first wave of covid and thus there was an increase in the number of cases. In Jan 2021, there was sudden increase in the number of cases and reason is social gathering. It was that time of the year where people want to meet their families and celebrate Thanksgiving , Christmas and New Year together. This is one reason for increase in COVID cases. Also, 501.V2 a new covid variant was discovered and until then no vaccination had started for the local people. By the end of January, the vaccine was imported from India to prevent the spread of the virus. But the research says that the main reason for the outbreak is related to the new variant or a lack of compliance with health guidelines during the holiday period. In August 2021, the third wave of covid broke and this led to multiple variants of covid in the country which were highly transmissible(Delta variant). The President declared the system was beyond its breaking point, with insufficient beds and barely enough oxygen which led to more deaths. In Jan 2022, there was an outbreak for OMICRON virus which led to the increase in the number of cases.

Note: The inference only for few countries are shown in the report. For other countries mentioned above refer the note book on GitHub.

2.2 Trends of a particular state and the highly affected counties

In this task a particular state was chosen and the trends of cases and deaths were compared with normalized values and log normalized values of covid cases and deaths. I chose ALABAMA state to perform the analysis and five counties which were sorted by their cases and deaths column to find the counties that were highly affected due to covid. The counties are Winston county , hale county , Franklin county , clay county ,Clarke county.

Inferences for cases of counties:

- The number of cases for the Franklin county is seen increasing in Jan 2022. This was the time when the new covid variant OMICRON spread had infected people. The lowest was in Hale county.
- The peaks were also seen in August 2021 to December 2021 and then again in Jan 2021 to April 2021. These peaks were observed due to social gathering and international travel to the united states. In July 2021 it was due to delta variant

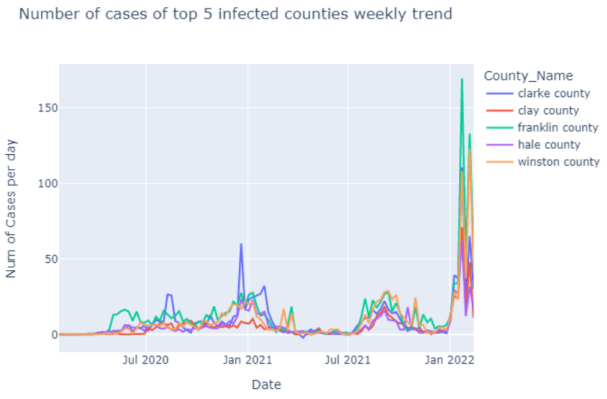

Figure 9: County cases trends top 5

and Jan 2021 was was due to the New years eve and Christmas.

- August 2021 - December 2021 , Winston county had the highest deaths and hale county had the lowest deaths.
- Jan 2021 to April 2021 , the highest peak was for Clarke county and lowest was for clay county.

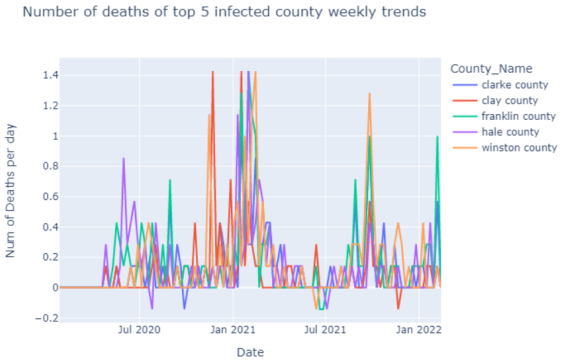

Figure 10: County deaths trends top 5

Inferences for deaths of counties:

- In the above graph we observe many deaths over the year from July 2020 to Jan 2022. The graph is not has sudden death rate increase in few months and less in other days of the year.
- In Jan 2022 , Franklin had the most deaths and Winston had lowest deaths. In mid of September 2021 Winston had the highest deaths and hale had less death rate.
- In Feb - March 2021 Clay and Clarke county had more number of deaths and clay county has less death rate. In July 2020 during the first covid wave hale county had the highest death rate and lowest in Winston county.
- The number of cases for the Franklin county is seen increasing in Jan 2022. This was the time when the new covid variant OMICRON spread had infected people. The lowest was in Hale county.

Note: Similarly the log and normalized values were calculated and the trends can be seen on GitHub.

3 Stage 3 : The goal of Stage II is to develop distributions and formal hypothesis tests for the intuitions you had in Stage I and II and utilize statistical modeling to prove/disprove them.

In this task the normalized cases and deaths which were generated in Stage 2 were used to plot the distribution. Here the data was discrete so PMF was calculated using the stats library.

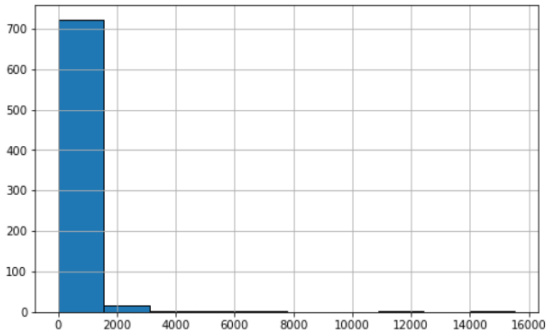

Figure 11: Histogram of cases ALABAMA Sate

Inference from the graph :

- The graphs shows the count of cases per day.
- The data is skewed at the right which means its a positive skew.
- The data has a peak initially and then falls. Thus the data is discrete
- Since the data is discrete, poison distribution is the best fit for this data.

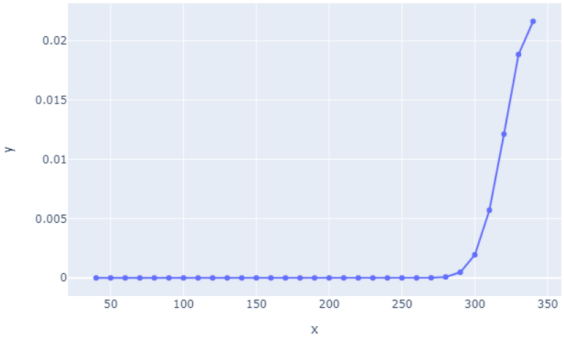

Figure 12: Scatter Plot using Probability Mass Function

Inference :

- I set the range from 40 to 350 and these are my interval where am calculating the probability mass distribution.
- Using the mean obtained from the above dataset I found the mean and plotted a graph with x axis being my interval and y axis being probability.

- From the graph we observe that initially the cases were low hence the probability in that Alabama state is zero but after few intervals the cases rise and hence the probability of a person contracting covid is high.

The moments of distribution for 5 states were calculated and compared with the ALABAMA state to see the trends of each state and know which state was highly affected.

Inference of Moments of Distribution Alabama State :

- The mean of the Alabama state for new cases per day is 340.1542787860785
- The skewness is positive hence the tail of the distribution is longer towards the right hand side of the curve.
- The kurtosis is positive and tells distribution is peaked and possesses thick tails.

Comparision of Moments of Distribution with other states :

- In kurtosis, Alabama is highly peaked and Minnesota is low.
- The data is positively skewed for all the states and the skewness is observed towards right which is same for all states.

3.1 Modelling of New Cases and Deaths

In this task the normalized cases and deaths were used to plot the distribution with 'x' axis being the random interval against the probability. I used a combined data-set to show the cases and deaths distribution for all the states in a single graph. below is the graph for all sates

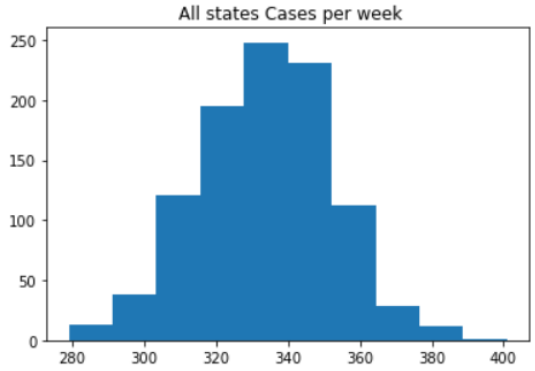

Figure 13: Model distribution of New Cases

Inference for Fig 15:

- We observe that all the states were having same deaths rates during a given interval
- The whole data-set we observe the probability of death rate in all states is in the range 0.20 - 0.25. This is a weekly data-set and hence the death rates seem to decrease at the start and end of a particular interval.
- The reason for the decrease could be reduced number of cases. Thus we compare that as the number of cases rose, the death rates also increased and vice versa.

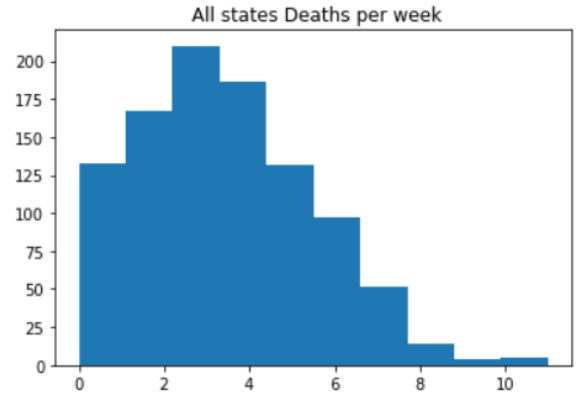

Figure 14: Model distribution of New Deaths

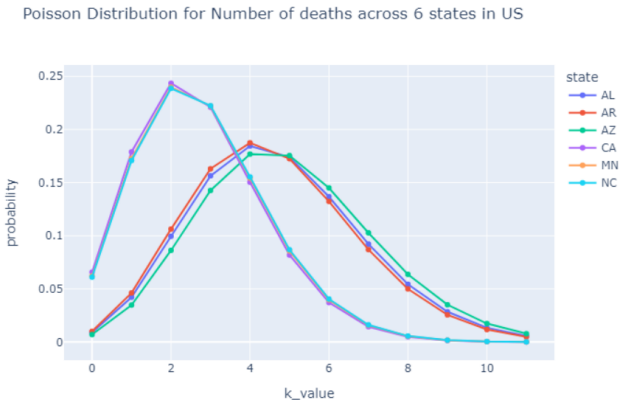

Figure 15: Poisson Model of all states

3.2 Perform correlation between Enrichment data variables and COVID-19 cases to observe any patterns.

For this task, used the merged hospital data-set with covid which was created in stage 1.

Inference for Figure 16 :

- Normalized cases and inpatient_beds_used_covid have a positive correlation which means when the cases increase the beds accommodated by patients affected due to covid also increase.
- Normalized cases and critical_staffing_shortage_today_yes there is a positive correlation as the staff may be affected due to covid which lead to the staff shortage. Hence these variables have a positive correlation.
- Also the adult_icu_bed_utilization variable also has a positive correlation.

4 Stage 4 : Machine Learning

The goal of Stage IV is to utilize machine learning and statistical models to predict the trend of COVID-19 cases and deaths.

```
hospital_dataset.corr()
```

	countyFIPS	Confirmed	Deaths	population	critical_staffing_shortage_today_yes	critical_staffing_shortage_today_no
countyFIPS	1.000000	-0.027332	-0.056518	-0.100925	-0.001407	0.135137
Confirmed	-0.027332	1.000000	0.438004	0.308859	0.071617	0.039671
Deaths	-0.056518	0.438004	1.000000	0.507920	0.108526	0.057129
population	-0.100925	0.308859	0.507920	1.000000	0.273683	0.152813
critical_staffing_shortage_today_yes	-0.001407	0.071617	0.108526	0.273683	1.000000	0.790629
critical_staffing_shortage_today_no	0.135137	0.039671	0.057129	0.152813	0.790629	1.000000
inpatient_beds	-0.141380	0.080617	0.118960	0.315873	0.773114	0.555159
inpatient_beds_used	-0.113336	0.080892	0.119311	0.316527	0.776637	0.549680
inpatient_beds_used_covid	-0.020640	0.076709	0.111782	0.300430	0.786548	0.647618
staffed_icu_adult_patients_confirmed_and_suspected_covid	-0.131886	0.076678	0.112516	0.301173	0.772314	0.649240
staffed_icu_adult_patients_confirmed_covid	-0.140730	0.076918	0.113006	0.302071	0.772406	0.646108
total_adult_patients_hospitalized_confirmed_and_suspected_covid	-0.024751	0.076700	0.111836	0.300371	0.786260	0.649682
total_adult_patients_hospitalized_confirmed_covid	-0.044216	0.077396	0.112938	0.303324	0.792600	0.642355
total_staffed_adult_icu_beds	-0.175751	0.081022	0.120005	0.317424	0.770491	0.548318
inpatient_beds_utilization	0.649294	0.016122	0.021439	0.054912	0.056541	-0.267111
percent_of_inpatients_with_covid	0.245431	-0.012732	-0.023412	-0.048198	0.205983	0.761189
inpatient_bed_covid_utilization	0.564973	-0.005130	-0.013410	-0.022496	0.251067	0.725395
adult_icu_bed_covid_utilization	-0.146198	-0.029609	-0.045346	-0.110971	-0.018224	0.535432
adult_icu_bed_utilization	0.101814	-0.027887	-0.048876	-0.098799	-0.524726	-0.096351
deaths_covid	0.375719	0.022778	0.027466	0.090411	0.216365	0.537738
normalized cases	0.001708	0.210305	0.081090	-0.004619	-0.032154	-0.019908
normalized deaths	-0.035333	0.018998	0.154194	-0.011310	-0.037246	-0.019519

Figure 16: Correlation of enrichment data-set with the covid 19 data-set

For this task the zero values from both the columns deaths and cases were removed in-order to get the first occurrence of cases and deaths. The data-set derived after cleaning the data we perform linear and non linear regression to see the trends of these cases and deaths. Also, the confidence interval along with the new prediction line was determined. The regression was performed for a particular state and then compared with five highly affected counties. The linear and non linear regression for both the data-sets was performed and the trends were compared. Also, the regression model for country USA was compared with other countries to see the trends. Below are the screenshots of regression model for country , states and counties.

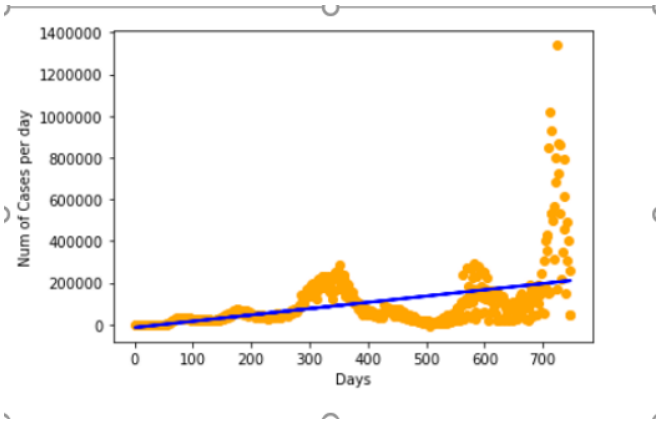

Figure 17: Linear Regression Model for Cases

Inference for linear regression cases:

- From the plot of linear regression for cases we can see that the model fit the data in a good manner.
- We can observe the trend line is increasing as Number of cases per day is increasing.
- We can infer from the graph, as number of days are between 700 to 747 there is peak / surge in new cases of Covid_19.
- We can infer That the highest peak of the cases is at the end of the graph for linear model.

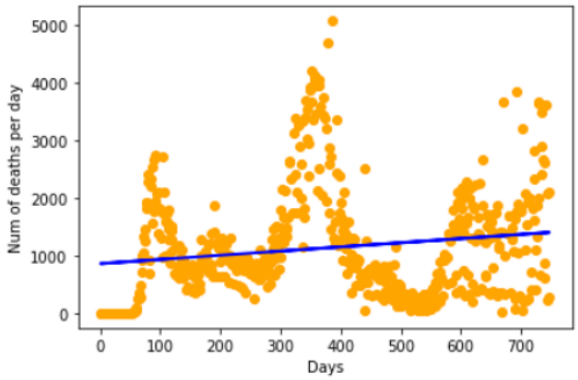

Figure 18: Linear Regression Model for Deaths

Inference for linear regression Deaths:

- From the plot of linear regression for deaths, we can see that the model fit the data in a nice manner.
- We can observe the trend line is increasing as the number of deaths per day is increasing.
- We can infer from the above diagram that the trend is positive from the start, and it is increasing.
- We can infer as number of death cases in linear model is increasing or at the peak it is not affecting the trend line.

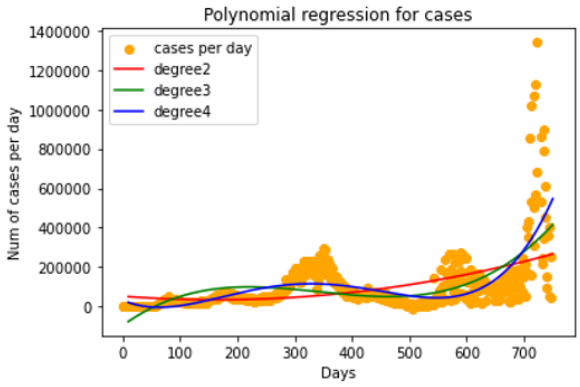

Figure 19: Non Linear Regression Model for Cases

Inference for Non linear regression Cases:

- For nonlinear cases, it can be seen that the regression lines are at the peak in the end.
- For nonlinear regression model of $degree = 2$, it fits the model in a linear manner.
- For nonlinear regression model of $degree = 3$, it fits the model in a curved manner.
- For nonlinear regression model of $degree = 4$, it fits the model in a multi curved manner.
- The model with $degree = 4$ fits the data in the best way.

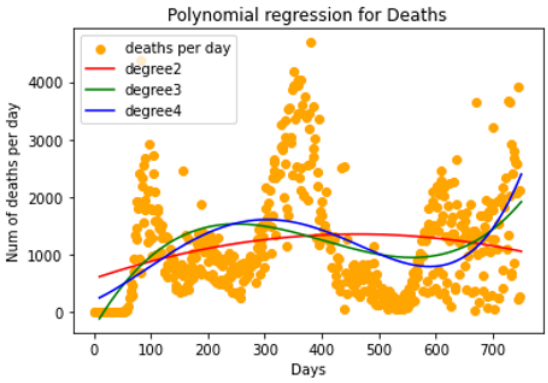

Figure 20: Non Linear Regression Model for Deaths

Inference for Non linear regression Deaths:

- For nonlinear deaths, it can be seen that the regression lines are at the peak in the end.
- For nonlinear regression model of degree = 2, it fits the model in a nonlinear curvature manner.
- For nonlinear regression model of degree = 3, it fits the model in a curved manner.
- For nonlinear regression model of degree = 4, it fits the model in a multi curved manner.
- The model with degree =2 has a declining curve at the end of the data points. While degree 3 and 4 have an increasing curve at the end of the data points.
- The model with degree = 4 fits the data in the best way.

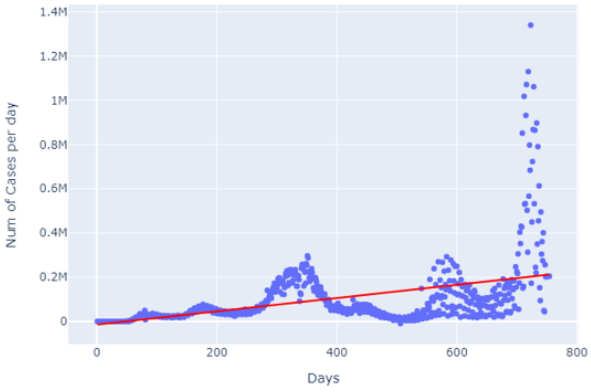

Figure 21: Prediction for for cases

Predicted trendline cases data

- For future prediction of upcoming week's cases, we have generated data of next seven days using the model.predict().
- Then we have concatenated it with the original data for cases and days.
- After creating a new updated data frame, we used plotly to plot the trendline same as above.
- We can see that the trendline is increasing since the first day of cases reported.
- And it can be predicted from the trendline that in the upcoming week the cases are going to increase.
- Also, we can see a slight difference of trendline between the original new cases trendline and the upcoming week's cases trendline.

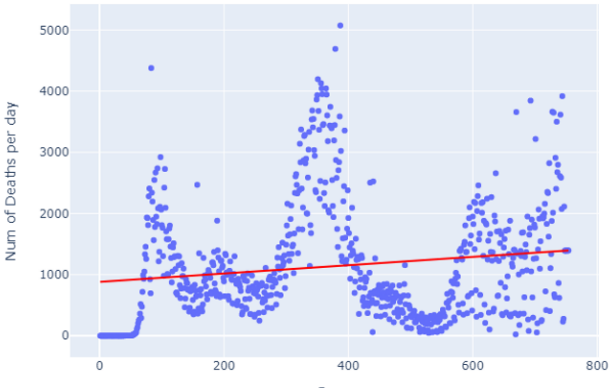

Figure 22: Prediction for for deaths

Predicted trendline for deaths data

- For future prediction of upcoming week's deaths, we have generated data of next seven days using the model.predict().
- Then we have concatenated it with the original data for deaths and days.
- After creating a new updated data frame, we used plotly to plot the trend-line same as above.
- We can see that the trend-line is increasing since the first day of deaths reported.
- And it can be predicted from the trend-line that in the upcoming week the deaths are going to increase.
- Also, we can see a slight difference of trend-line between the original new deaths trend-line and the upcoming week's deaths trend-line.

4.1 Comparison of USA with other countries

Brazil Regarding COVID cases, Brazil appears to have a flatter curve than the US, and its nonlinear model with 2 degrees is actually showing a downward curve to the data, as opposed to the US' upward curve for all models. Brazil differs even more with COVID deaths. All 3 models predict a downward trend in COVID deaths, while the US predicts an upward trend with all models.

Spain COVID cases for Spain have a nearly identical trend to the US for the 2 nonlinear models, though the trend is flatter in nature. Deaths show a different picture. Of all the countries we looked at, Spain has the most consistent downward curve as the days since first infection increase. It and Myanmar are the only two countries with an overall downward trend in the linear model. For more information about the individual member tasks, refer to the member task reports in the GitHub repository for Team 4. **Note: For more country comparison refer the notebook on GitHub**

4.2 County comparison and different types of testing

The county with highest number of cases and deaths was found by sorting the data-set and then linear and non-linear regression modelling was done. With this, confidence intervals for the regression line was calculated and also the trend-line of the cases and deaths were created. The notebook on GitHub has detailed description of each county and their trend-lines.

Comparison of State Alabama with other five counties - Trend of New Cases

- We observe that the new cases trend have been increasing toward the end of the graphs for both the state and the counties.
- Initial both at the state and the county level the cases were low
- The polynomial of degree 4 fits well along the regression line for the both the states and the counties

Comparison of State Alabama with other five counties - Trend of New Deaths

- We observe that the new deaths trend have been increasing towards the 300 to 400th days and again the rise was seen from 500-700th days
- Due to the increasing trend of new cases the rise in deaths cases were seen.
- The polynomial of degree 4 fits well along the regression line for the both the states and the counties

4.2.1 Point of no return , Hypothesis Testing and Chi-Square Testing

To calculate the point of no return I used columns from the merged data-set with beds and total beds used. So based on these values if the graph line crosses the total utilized bed, then there is no point of return. I tried to apply similar techniques by summing the values and plotting the graph against the new cases / deaths with the no. of days column. I selected two variables from the data frame with which I can compare. The reason covid and hospital is related because when the cases increase, and the beds are totally accommodated then there will be no point of return. So, I calculate the sum of the columns to perform this task and check if utilized bed cross the total beds then there is no point of return. Later I compare the data with another state, I chose North Carolina to check the trends of the two states.

Inference

- From the above plots we observe that the trend in both the states for cases and deaths is the same.
- The border line which explains that whether a parameter has a point of no return or not.

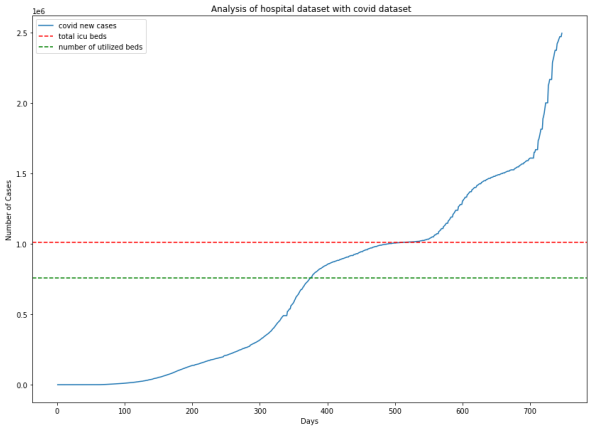

Figure 23: Point of no return for Cases

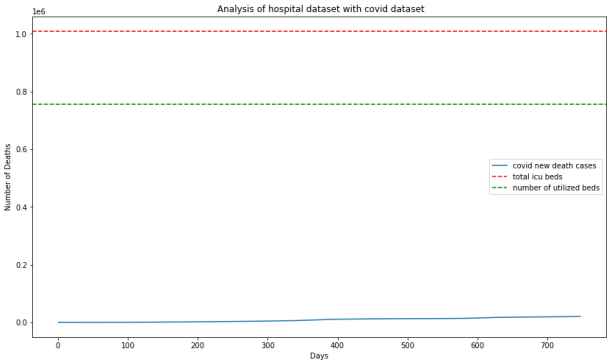

Figure 24: Point of no return for Deaths

- I have taken two parameters to compare against the covid data-set. So if the cases increase then there are no beds available.Also the Number of utilized beds cross above the total ICU beds then there is no point of return.
- In both the states the point of return has not yet arrived as the data-set used doesn't have such values for the utilized bed to cross the total beds.

Hypothesis testing and Chi-Square My Hypothesis is that with the increase in the beds utilization the number of new cases and new deaths are increasing. To state the Null hypothesis and Alternate hypothesis I would define it as ,

- Null Hypothesis (Ho): The beds utilization has no effect on increase in cases.
- Alternate Hypothesis (Ha): The bed utilization has an effect with the increase in cases.

I will try to to approve or reject a hypothesis by calculating the mean of the data-points. As shown in the above plots we can observe the mean of the data-points for both the variables and come to a conclusion that the proposed experiment is true. The mean for bed utilized is beyond the line of the mean of cases. We observe that the beds utilized is more than the mean of the cases which proves that the null hypothesis is correct. The two variables beds and cases are increasing and dependant on each other.Also, The data-set has a negative values which

means that there were no beds available and the people who wanted a bed were added to the wait-list. Thus in the graph the beds bin values are hidden as it goes negative.

5 Stage 5 : Dashboard

The final stage aims a developing a simple interactive dashboard based on the analysis we have done so far

5.1 Prediction of new cases and deaths

In this task, we have created a dash with selection of daily and logarithmic cases/deaths. In addition to this, we have given provision for selection of linear and nonlinear trend-line by a radio button. In addition, there is a date selection feature, which allows for selection of date, and graph is modified within the provided date range. Below are the screenshot for the same:

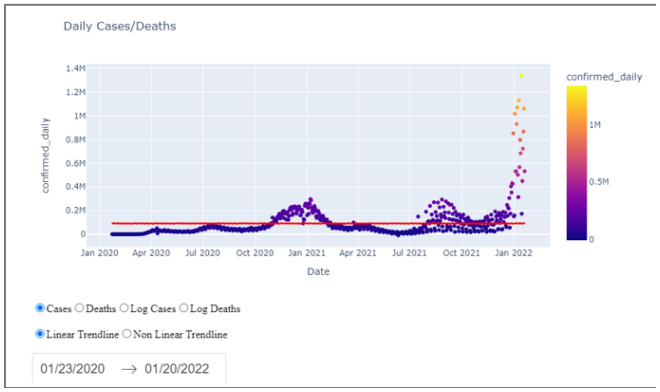

Figure 25: Linear Cases

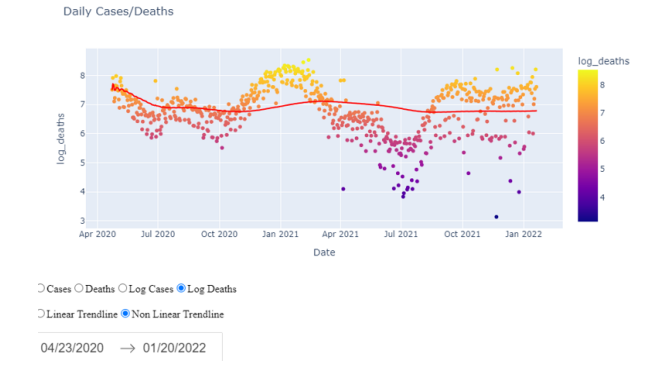

Figure 26: Non Linear Deaths

5.2 Moving average by states

In the dashboard we can see that the moving average can be viewed on a particular State. By default, Nevada is selected, and we can view the graph for different states. There are radio buttons to view the case type like, Cases, Deaths, Log Cases and Log Deaths.When the user selects a particular state and the choose one of the radio buttons, the trend-line along with the moving average can be seen. The ultimate purpose of moving averages is to identify long term trends. They are calculated by averaging a group of observations of a variable of interest over a specific period. Such averaged number becomes representative of that period in a trend line. Below is the screenshot for moving average for different states.

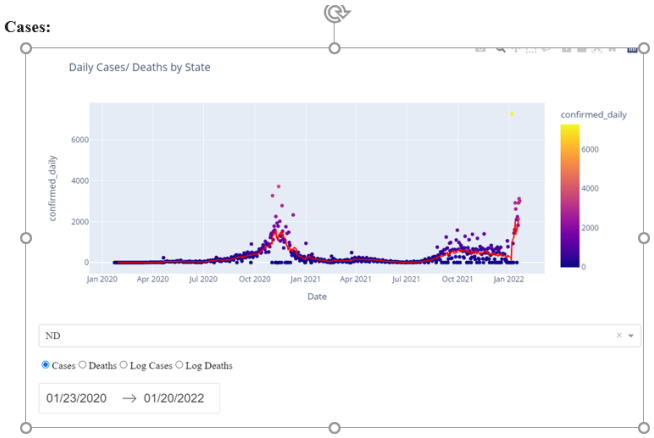

Figure 27: Moving average for states

5.3 Creating Map on Dashboard

For this task, cloropleth of plotly was used to generate the USA map for state and county. In the dashboard two radio buttons were provided to view cases and deaths and when the user hovers the data on the map then the county information can be viewed. Below is the screenshot of USA Map with hover of cases and deaths

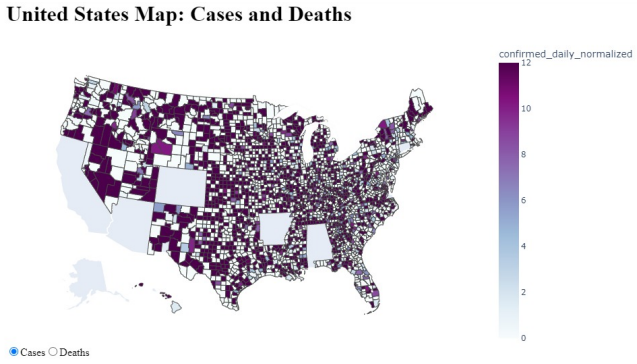

Figure 28: Map of USA

References

[1] https://github.com/UNCG-CSE/Spring-22_COVID-Team_4

[2] https://github.com/UNCG-CSE/Spring-22_COVID-Team_4/tree/main/doc