Title: A Statistical Analysis of Different COVID-19 Aspects in the U.S. and Forecasting the U.S. Mortality-rate Time-series

Rezaur Rashid

1. Introduction

Coronavirus Disease 2019 (COVID-19) has been one of the biggest catastrophic events since 1921 Influenza Flu and World War II. It was declared as a global pandemic in March 2020, affecting almost every country in the world. As of February 22, 2022, globally over 400 million people have been infected and almost 6 million people have died [1]. According to CDC, over 78 million people have been affected by COVID-19 in the U.S. and almost 1 million people have died [2].

Since this pandemic is still growing, it is important to understand how to combat this disease effectively. A myriad amount of COVID-19 data exists out there and therefore, we can examine and analysis different aspect of the pandemic and prepare for the future crisis.

In this study, we provide a statistical analysis of the COVID-19 pandemic in the U.S. and predict a forecasting time-series on the mortality rate, namely we will use the data available only for U.S.

2. Dataset

For our project, we have used the COVID-19 data compiled by Our World in Data [3] to do our statistical analysis and the data from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [4] to do a time-series forecasting on covid-19 mortality rate. Although, these repositories contain COVID-19 data about most of the countries in the world, we have only use the data available for U.S. entirely and/or state levels. The U.S. datasets contains daily cases having different field such as: locations, new confirmed cases, deaths, recoveries, testing and hospitalizations, vaccination etc. for all 50 states in their county level as well as time-series data about confirmed cases and deaths.

3. Methodology

The project is divided into 3 parts: (1) Visualization of different aspect of COVID-19 situation, (2) Statistical hypothesis test for different covid cases between states using Oneway-ANOVA and Wilcoxon Rank Sum Test and (3) Using a predictive model i.e. ARIMA model for time-series forecasting.

3.1. Visualization of different aspect of COVID-19 situation

Doing a thorough statistical analysis of COVID-19 data in the U.S. gives us a visualized report and this will provide us some insights about the spread of the disease in the U.S., factors that are causing fatality-rate, how hospitals are coping with new cases, and how people are reacting and recovering after taking vaccination etc. for all the 51 states including DC. Namely, we have tried to recreate the few of the analysis presented in the paper (Dayaratna, K., & Badger, 2022) [5] related to U.S. for the pandemic, but expended this analysis broken into state level in this project. We have analyzed several distributions for different aspects of the COVID-19 situations in country level and some example states.

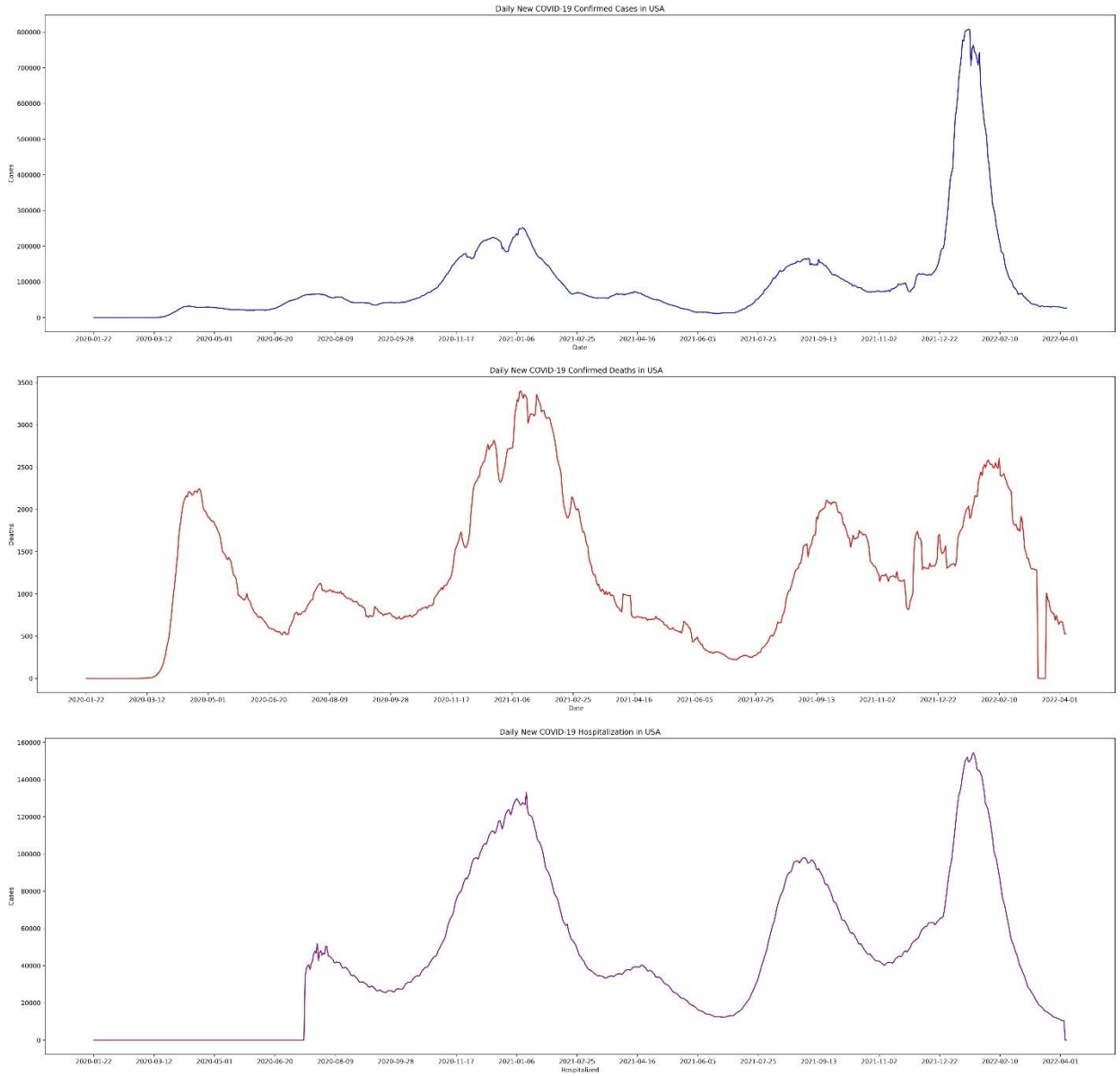

Figure 1: Covid-19 daily cases (top), deaths (middle) and hospitalizations (bottom) in the U.S. from January 2020 to April 2022

Figure 1 illustrates that after the initial break out of the pandemic in January 2020, the infections, deaths, and the hospitalizations went to pick in January 2021 and two other pick cases in September 2021 and January 2022 due the Delta and Omicron variant respectively all over the U.S.

Similarly, Figure 3 show the same characteristics for the state of North Carolina and as well as Figure 4 with a comparison between North Carolina vs South Carolina as an example of two states.

In Figure 2, we can see the total confirmed cases vs total deaths according to different age groups in the U.S.

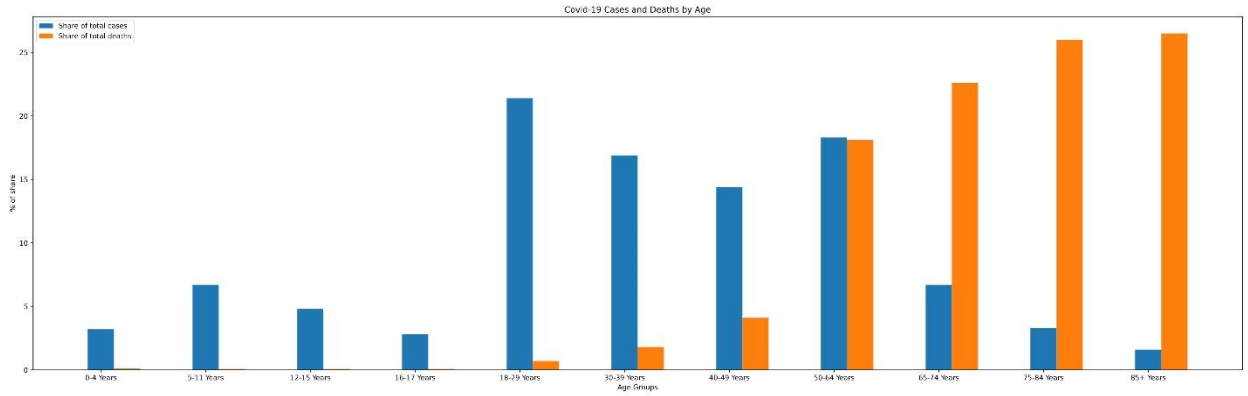

Figure 2: Covid-19 cases and deaths of U.S. We can see that the elderly population (aged >65) shares a higher mortality rate although they are not the large group of Covid-19 patients.

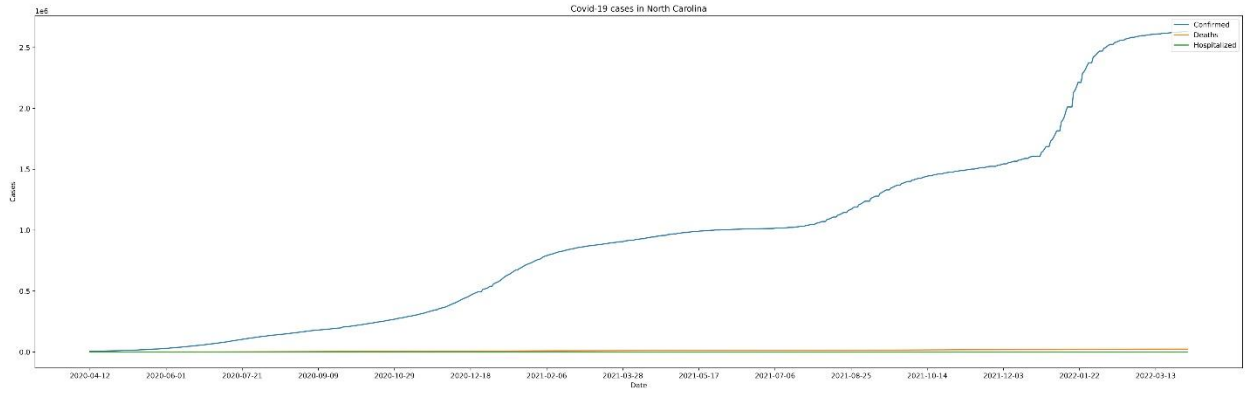

Figure 3: Covid-19 total confirmed cases, deaths and hospitalizations in North Carolina

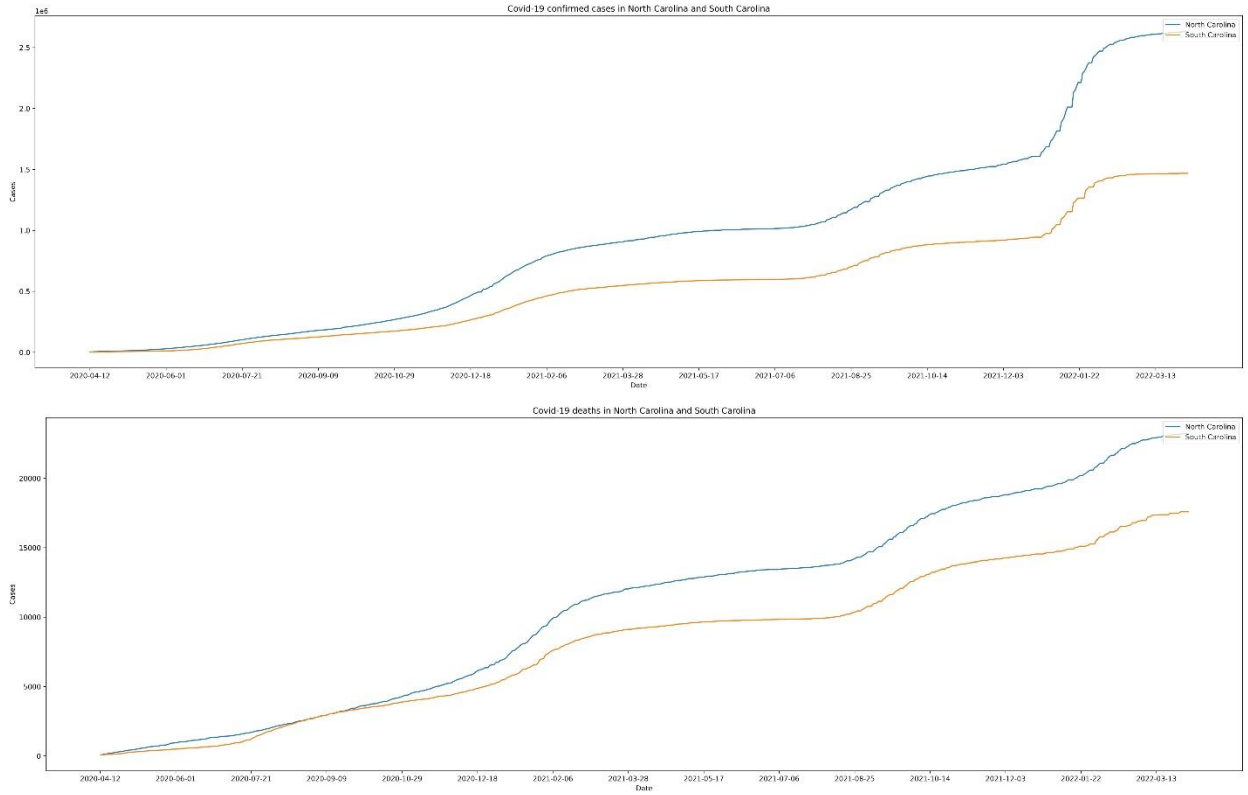

Figure 4: Covid-19 total cases (top), total deaths (bottom) comparison between North Carolina and South Carolina

In the Figure 5 (top), we can see the different aspect of vaccinations in all over the U.S from December 2020 to March 2022. From the observation we find that people that have taken at least one dose of vaccination, they have taken the second does as well (mostly). But the number of booster doses is low and not all the fully vaccinated people have taken the booster dose. The same scenarios are observed in Figure 5 (bottom) for the state of North Carolina.

Figure 6 illustrates the histogram of daily vaccinations for North Carolina and South Carolina in comparison.

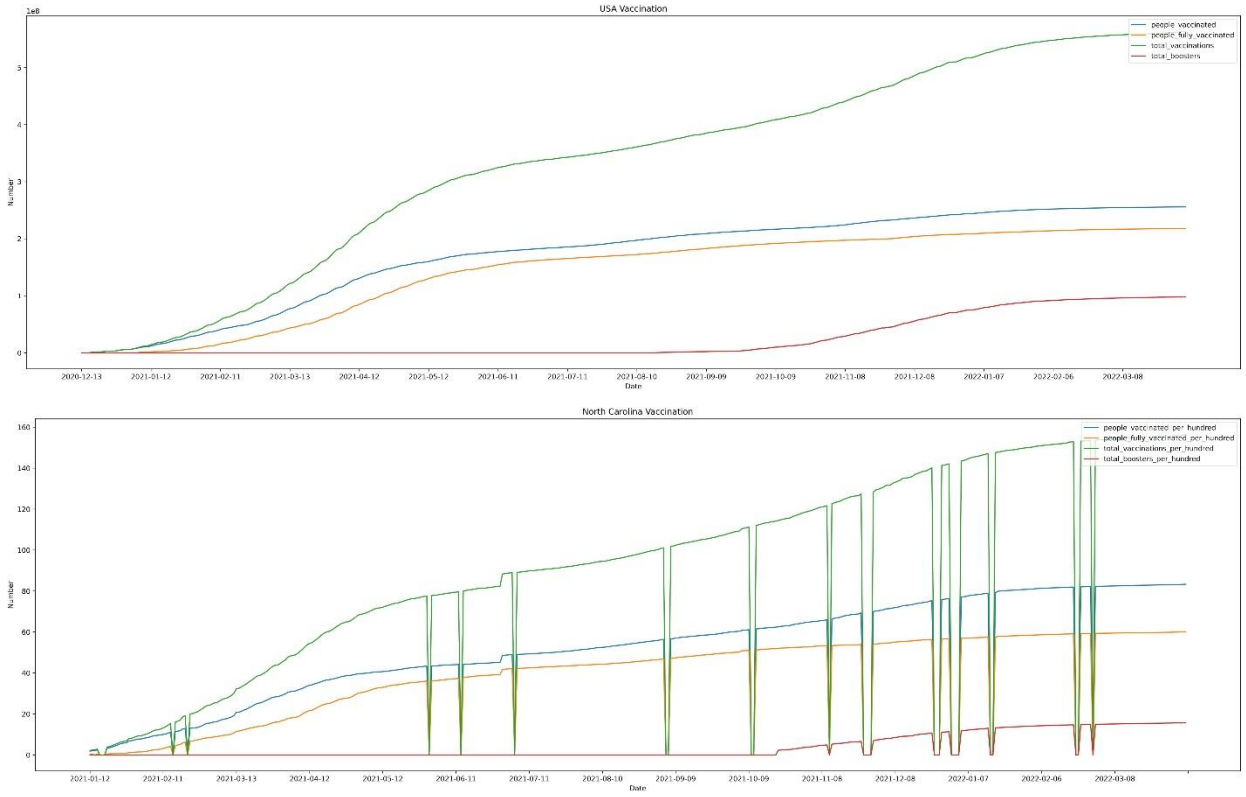

Figure 5: Covid-19 Vaccination updates: U.S. (top), North Carolina (bottom)

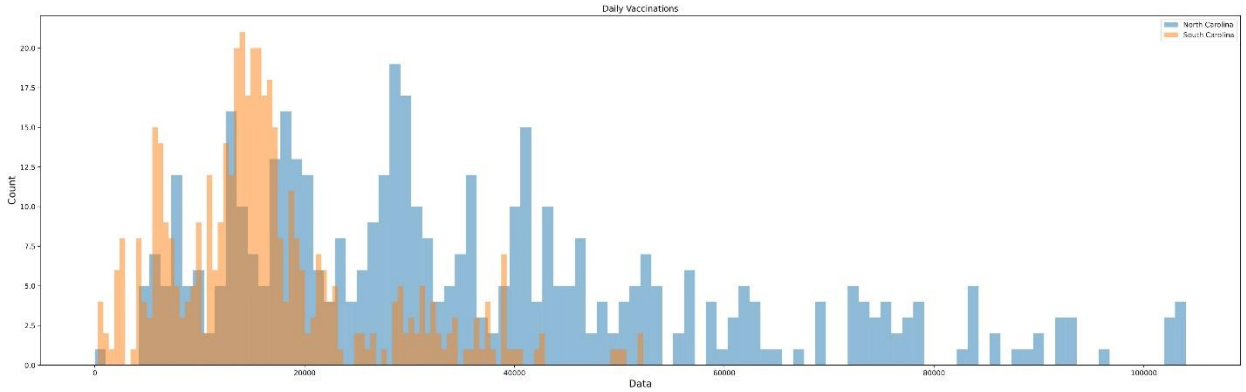

Figure 6: Covid-19 daily Vaccination distribution comparison between North Carolina and South Carolina

3.2. Statistical Hypothesis Test for Different Covid-19 Cases Between States:

3.2.1. Oneway-ANOVA Model

Since we have different case data in different categories like daily confirmed cased, deaths, recovered cases and mortality for all 51 states including DC, we have tried to do a null hypothesis test weather there is differences in group mean (group as in states).

The Figure 7 illustrates the anova-summary/myLm-summary statistics done using Oneway-ANOVA full model in 'R' to get the null hypothesis and we see that the group means differs significantly from the overall mean of different variables in four cases mentioned above.

```
Analysis of Variance Table

Response: myData_confirm
             Df     Sum Sq    Mean Sq F value    Pr(>F)
states       50 1.6385e+16 3.2771e+14  729.75 < 2.2e-16 ***
Residuals 36769 1.6512e+16 4.4907e+11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Variance Table

Response: myData_death
             Df     Sum Sq    Mean Sq F value    Pr(>F)
states       50 4.6082e+12 9.2164e+10  1363.7 < 2.2e-16 ***
Residuals 36769 2.4851e+12 6.7586e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Variance Table

Response: myData_recover
             Df     Sum Sq    Mean Sq F value    Pr(>F)
states       50 1.1337e+14 2.2674e+12  161.12 < 2.2e-16 ***
Residuals 36769 5.1745e+14 1.4073e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Variance Table

Response: myData_mortality
             Df Sum Sq  Mean Sq F value    Pr(>F)
states       50 2.3107 0.046215  312.29 < 2.2e-16 ***
Residuals 36769 5.4413 0.000148
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.2.2. Wilcoxon Rank Sum Test for Pairwise comparison between states vaccinations

Here, we have tried a different statistical hypothesis test where we wanted to compare the daily vaccinations between North Carolina and other states with a pair-wise test. Since, we have Covid-19 datasets for different states which are non-linear in natures, instead of a parametric statistically significant test (e.g. t-test), we have used a 'Mann–Whitney U test' in python which usages the 'Wilcoxon signed-ranked test' for different sample size. This non-parametric statistical test is used to compare whether there is a difference in the dependent variable for two independent groups. It compares whether the distribution of the dependent variable is the same for the two.

Table 1 illustrates all the p-values from the pair-wise non-parametric Wilcoxon test between North Carolina and other states. We can see from the table that almost for 46 states we can reject the null hypothesis that the vaccinations distribution is not same for these groups whereas for the states of Michigan, New Jersey, Ohio and Virginia we can reject the null hypothesis that there is no difference with North Carolina.

	States	p-value			States	p-value
North Carolina	Alabama	3.85E-70		North Carolina	Missouri	1.23E-36
	Alaska	5.65E-146			Montana	9.18E-142
	Arizona	8.68E-12			Nebraska	3.67E-121
	Arkansas	4.70E-103			Nevada	1.85E-93
	California	1.84E-122			New Hampshire	7.84E-127
	Colorado	2.24E-21			New Jersey	9.19E-01
	Connecticut	1.41E-51			New Mexico	2.17E-105
	Delaware	1.04E-139			New York State	3.30E-59
	District of Columbia	2.22E-143			North Dakota	9.92E-145
	Florida	1.20E-59			Ohio	8.36E-01
	Georgia	1.94E-02			Oklahoma	5.17E-77
	Hawaii	8.95E-122			Oregon	7.65E-49
	Idaho	3.07E-131			Pennsylvania	8.55E-09
	Illinois	2.11E-08			Rhode Island	8.41E-135
	Indiana	1.13E-29			South Carolina	1.76E-52
	Iowa	1.96E-87			South Dakota	2.76E-143
	Kansas	2.90E-98			Tennessee	2.67E-26
	Kentucky	4.46E-66			Texas	9.71E-91
	Louisiana	5.98E-72			Utah	1.57E-86
	Maine	3.58E-125			Vermont	5.47E-143
	Maryland	6.60E-13			Virginia	8.46E-01
	Massachusetts	2.69E-04			Washington	6.52E-04
	Michigan	7.25E-02			West Virginia	5.63E-120
	Minnesota	1.62E-23			Wisconsin	4.92E-25
	Mississippi	3.40E-109			Wyoming	3.66E-147

Table 1: Pairwise Wilcoxon test of Covid-19 daily Vaccination distribution between North Carolina and other U.S. states.

3.3. Time-series Forecasting using Predictive Model

In this section, we have tried to predict the mortality rate for the U.S. time-series COVID-19 data for the state of North Carolina. The daily mortality-rate is the ratio between the daily deaths divided by daily confirmed cases. We used a forecasting model in python called ARIMA (Auto-Regressive Integrated Moving Average) that captures a suite of different

standard temporal structures in time series data. Figure 7 illustrates an example time-series forecast for the Covid-19 cases in North Carolina.
One observation from the predive model is that, although the model did a good job in training phase, the prediction doesn't have good accuracy due to the diverse nature of the covid-19 cases.

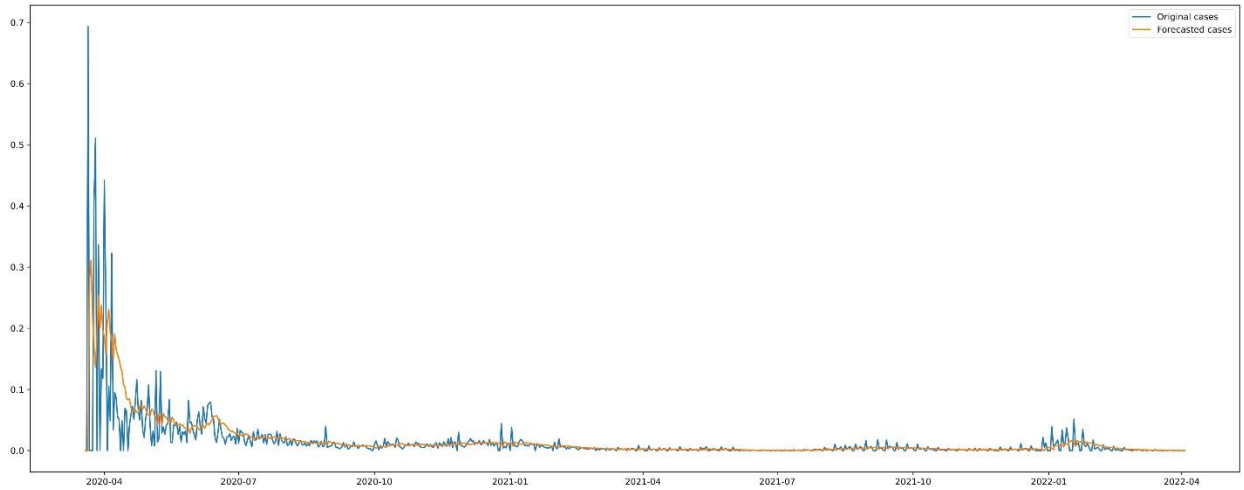

Figure 7: Covid-19 daily cases time-series forecast for North Carolina

4. Conclusion:

This project explores the effects of COVID-19 pandemic in the United States for different case scenarios using visualization and statistical analysis. Our finding is that the growth of pandemic has decreased in 2021 although there was a high peak in early 2022 due to Omicron variant. Another finding from the visualization is that, the vaccination helped reduced the number of hospitalization. We have also explored some statistical hypothesis test where we have found that the case scenarios within each states is significantly different than others.

Paper:

Dayaratna, K., & Badger, D. COVID-19: A Statistical Analysis of Data from Throughout the Pandemic and Recommendations for Moving On. 2022.

References

1. WorldoMeter: https://www.worldometers.info/coronavirus/

2. CDC: https://www.cdc.gov/coronavirus/2019-ncov/index.html

3. Our World in Data: https://github.com/owid/covid-19-data/tree/master/public/data

4. CSSE, JHU: https://github.com/CSSEGISandData/COVID-19

5. Dayaratna, K., & Badger, D. COVID-19: A Statistical Analysis of Data from Throughout the Pandemic and Recommendations for Moving On. 2022.