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 will use 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 will only use the data available for U.S. entirely and/or state level. The U.S. datasets contains daily cases having different field such as: locations, new 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 two parts: Statistical analysis of the data and Using a predictive model for time-series forecasting.

3.1. Statistical Analysis of COVID-19 Data

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 50 states. 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 we plan to expend 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. In progress, we will try to analyzed the

data in all states and county level and will try to find the hotspot of certain region for different pandemic aspects.

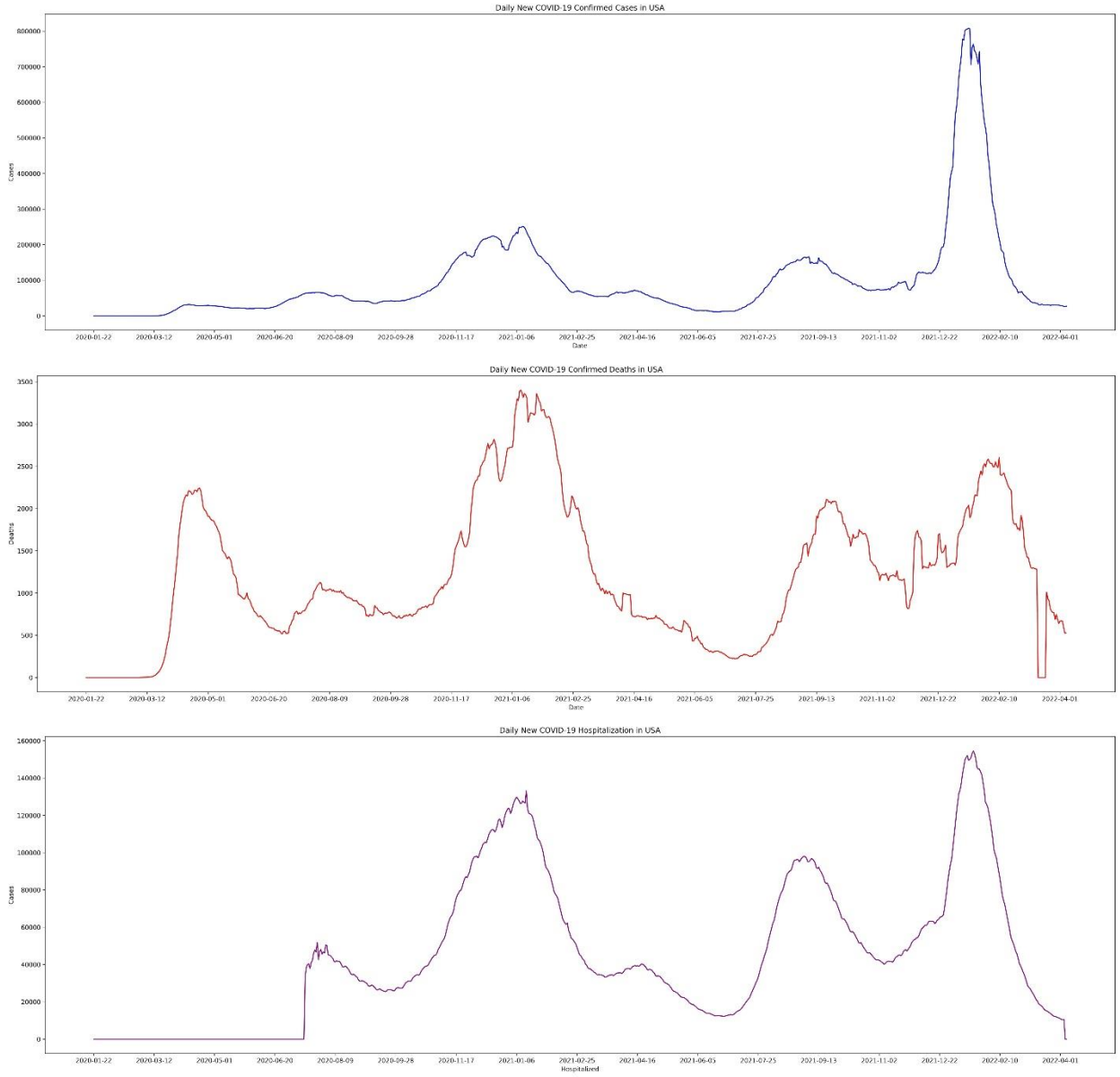

Figure 1: Covid-19 daily cases (top), deaths (middle) and hospitalizations (bottom) in the U.S. from January 2020 to April 2022

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. Figure 3 and Figure 4 also show the same characteristics for the state of North Carolina as well as South Carolina.

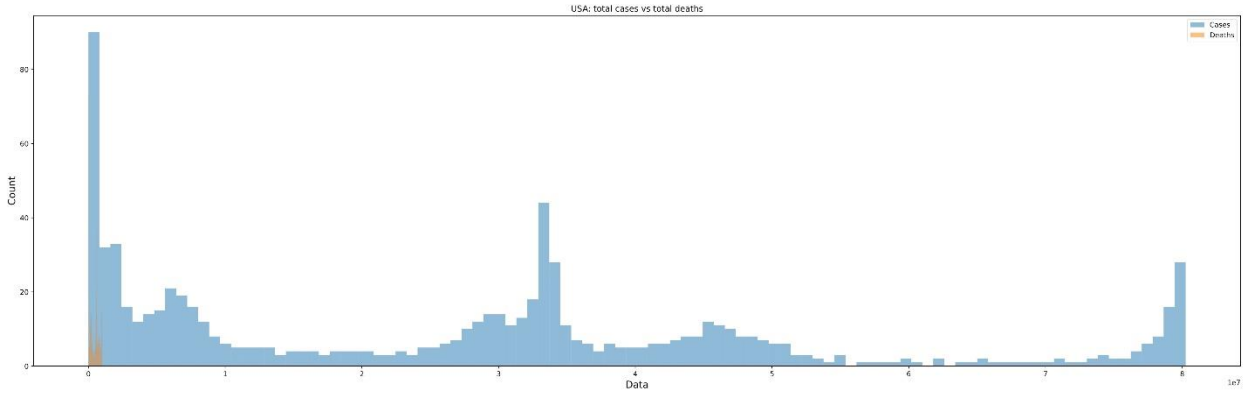

Figure 2: Covid-19 total cases, deaths distribution in the U.S. from January 2020 to April 2022

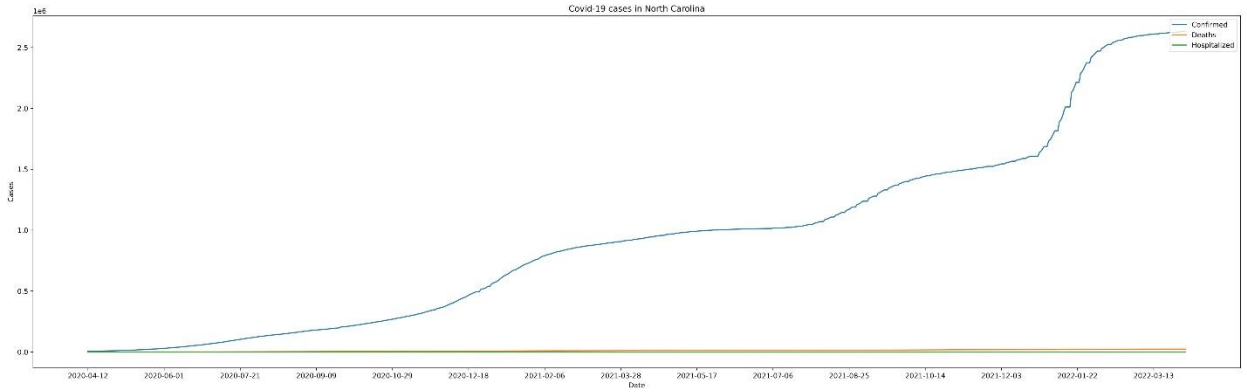

Figure 3: Covid-19 total confirmed cases, deaths and hospitalizations in North Carolina

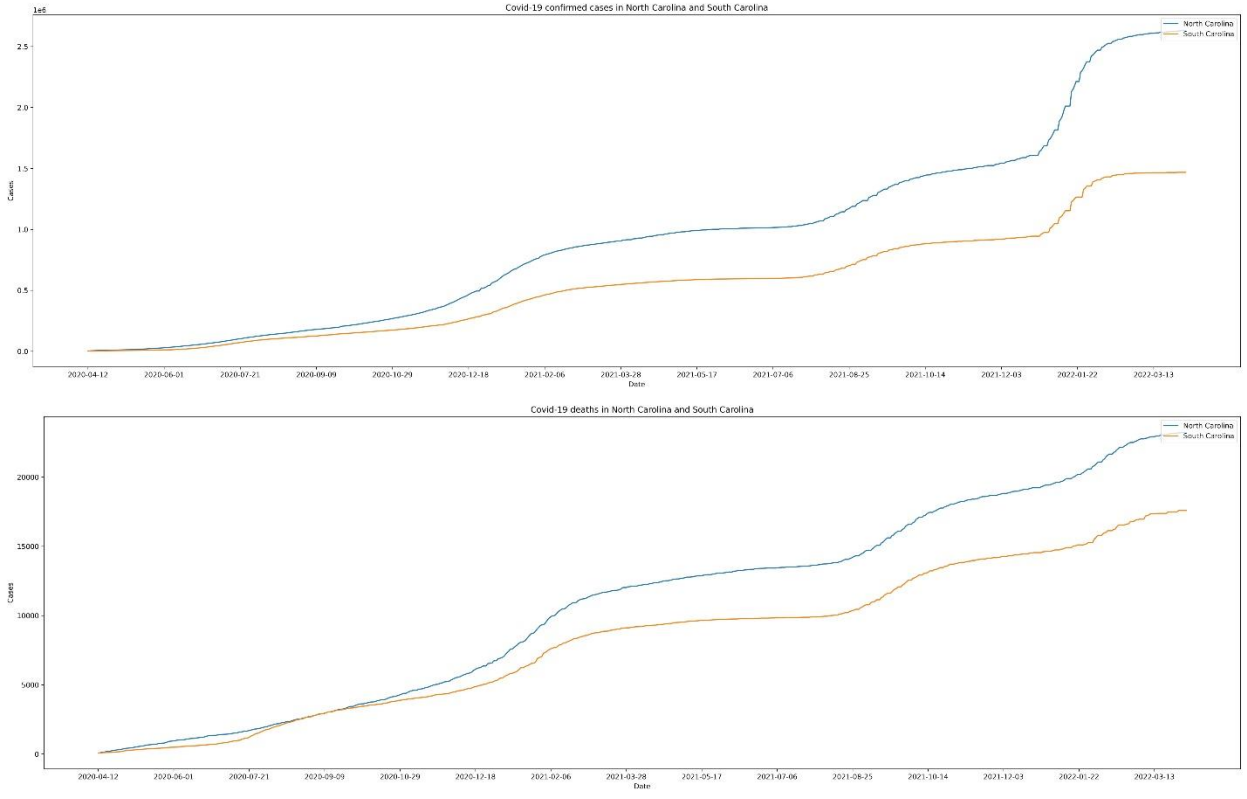

Figure 4: Covid-19 total cases (top), total deaths (bottom) comparison between North Carolina and South Carolina

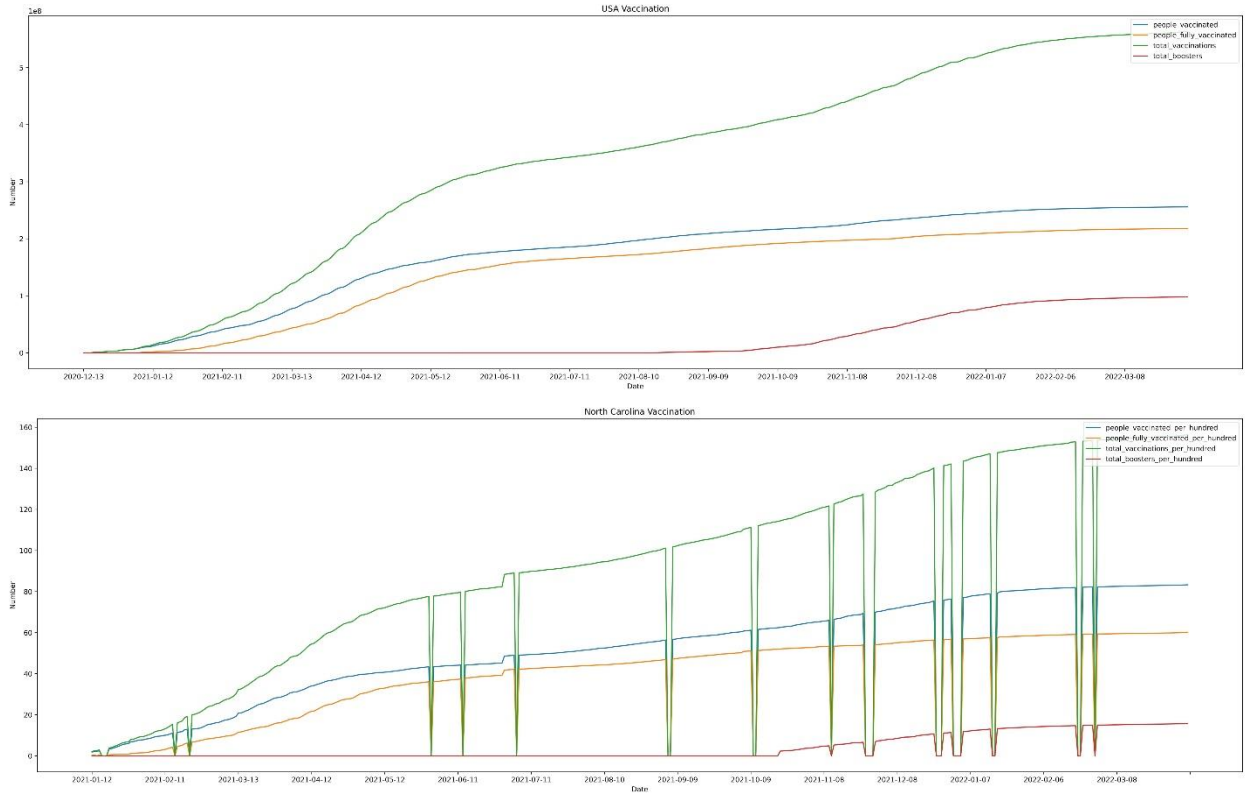

Figure 5: Covid-19 Vaccination updates: U.S. (top), North Carolina (bottom)

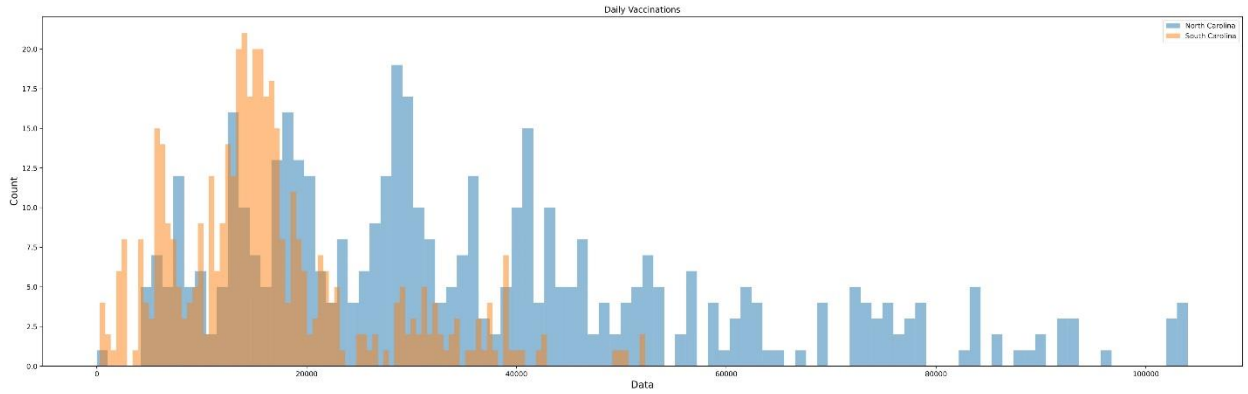

Figure 6: Covid-19 daily Vaccination distribution comparison between North Carolina and South Carolina

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

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 will use the 'Wilcoxon signed-ranked test' since the population data doesn't have a normal distribution as well. This non-parametric statistical test is used to compare two independent samples. Our plan for the next milestone is to compare North Carolina Covid-19 cases (deaths, hospitalizations, vaccination) against the US country level cases using Wilcoxon test and to observe how the distribution

very from each other. Secondly, we will use an 'One-way Anova test' to comapare the Covid-19 case scenarios across all the states.

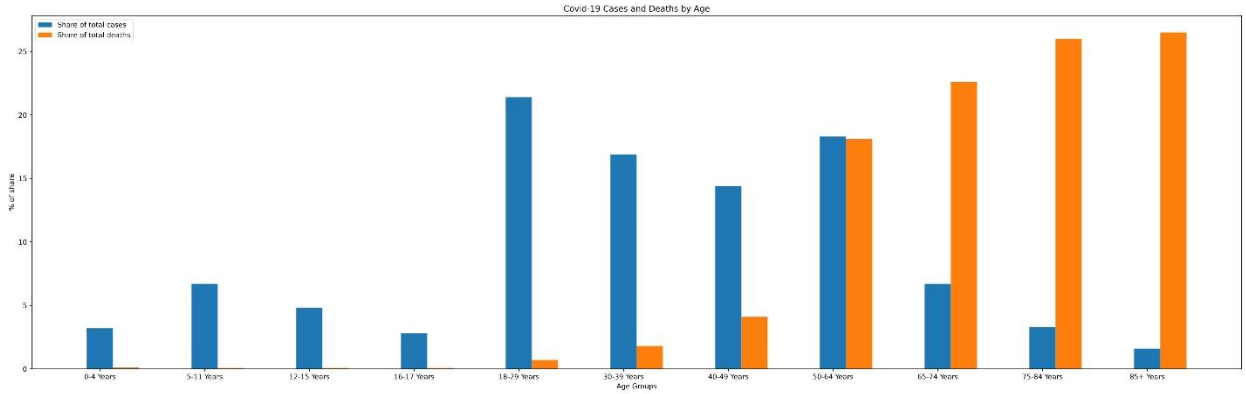

Figure 8: Covid-19 cases and deaths of U.S. We can see that the elderly population (aged >65) shares a higher mortality rate although they are not the large group of Covid-19 patients.

3.2. Time-series Forecasting using Predictive Model

In this section, we will try to predict the mortality rate for the U.S. time-series COVID-19 data for most of the states in their county level. The daily mortality-rate is the ratio between the daily deaths divided by daily confirmed cases. We will use 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. We will use different accuracy metrics such as: Mean Absolute Percentage Error (MAPE), Correlation between the Actual and the Forecast (corr), and Min-Max Error (minmax) to compare the models' performance series for different states as well as test their statistical significance. Figure 7 illustrates an example time-series forecast for the Covid-19 cases in North Carolina.

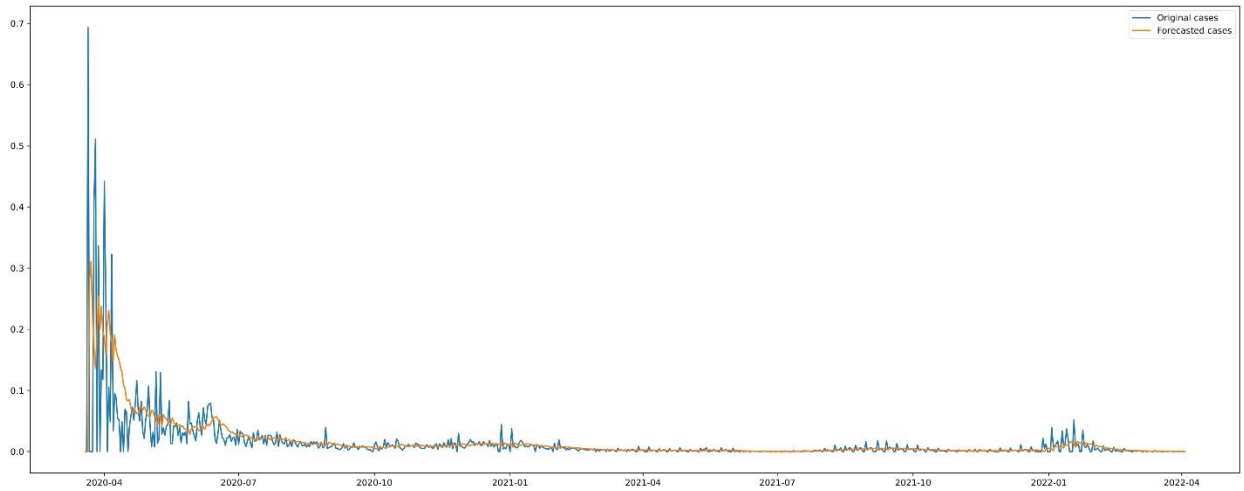

Figure 7: Covid-19 daily cases time-series forecast for North Carolina

Next Milestones:

1. More in-depth analysis of the Covid-19 data in states level.
2. Apply Wilcoxon test to check the null hypothesis between North Carolina and US case samples
3. Use One-way Anova to compare the sample for different Covid-19 case scenarios from multiple states
4. Provide a table for the comparative analysis between different state Covid-19 data of the statistical findings in a table.
5. Apply the Arima model to predict the time-series forecasting.

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.