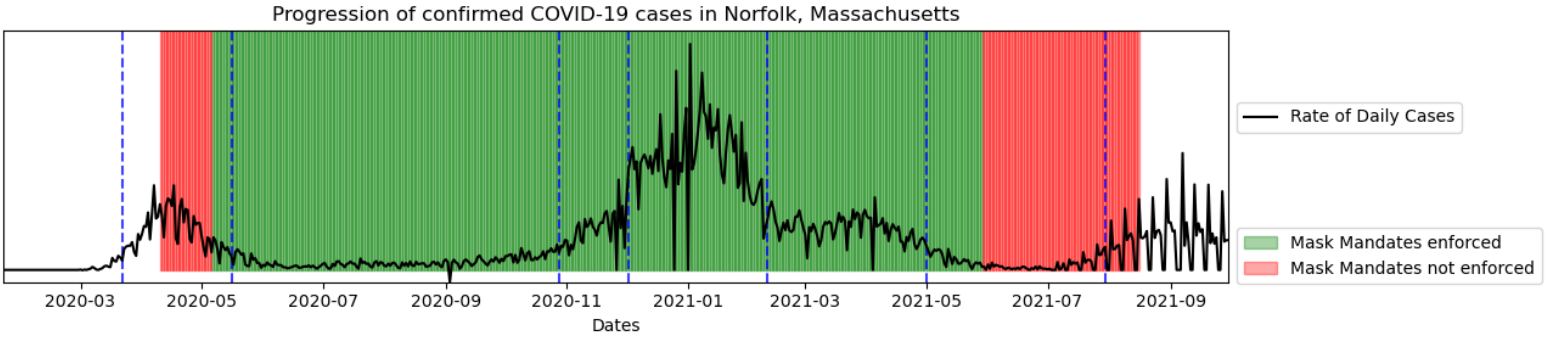
DATA 512 PROJECT PART 1

COLLECTIVE ANALYSIS

Github Link: https://github.com/anuhyabs/data512_project

County assigned to me: Norfolk, Massachusetts

Visualization plot:



Explanation of visualization:

The figure is a representation of the progression of confirmed daily cases of COVID 19 based on the Masking policies from February 2020 to October 2021 in Norfolk, Massachusetts. The red region in the visualization is the time period when the mask mandates were not enforced and the green region is the time period where the mask mandates were enforced. The black line shows the rate of daily confirmed COVID cases across the February 2020 to October 2021 time period. The dashed blue lines are indicative of changes in the direction of the cases automatically detected by the Pelt Algorithm. The datasets provided were originally filtered for Norfolk county. The Confirmed cases dataset consisted of a cumulative number of cases. The daily cases were calculated as the difference in the number of cases each day. This is plotted on the graph as the rate of daily cases. As it can be seen in the beginning 2020 between the time period of February-May 2020 the mask mandate was not enforced and there is a peak in the number of daily cases. As a result of the increase in the number of cases in the county as well as country wide, I believe the masking mandate policy was enforced in May 2020. Following the enforcement, there is a steady decline in the number of daily cases and there are no major increases in the number of cases until November 2020. From November 2020, there is a steady increase in the number of cases with it peaking in January 2021. It

must be noted that there has been no change in the masking policy but still there is a sharp increase in cases. This could be due to multiple reasons. Around the November-January time the second wave of COVID started and most hospitals were finding it difficult to admit patients due to a lack of resources. The first vaccinations had just come out for the public and there was an air of skepticism around the administering of vaccinations as well as a short supply of vaccinations. Thus, more people seemed to have become infected due to multiple reasons which may have had nothing to do with the masking mandate policy being still in place. From February 2021, a sharp decline of cases can be observed and as the number of daily cases become significantly less the mask mandate is removed. At the end of the graph, after the removal of the mask mandate, the daily cases seem to be rising again. This implies that masking policies did have some extent of impact on the progression of COVID cases.

As a side note, certain column in the dataset were unclear with no description provided (ex: order_code) and I was not able to understand how to use the mask_use_by_county dataset in my analysis as there wasn't sufficient information provided such as the dates when the data was captured.

Reflection Statement:

Before working collaboratively, I imported the data and explored it to understand what each dataset was, how it could be used to answer the research question and what features I would need to focus on. While exploring the data, I realized that the datasets were for all counties and I had to filter it for my county. Which meant that there had to be some sort of foreign key connection between each of the files. I found the FIPS column in the confirmed cases dataset and used that to find the county information in the other datasets. While collaborating I learned something new that there was another way of doing it by combining the FIPS_State and FIPS_County in the Mask Mandates dataset. The dataset also provided a cumulative count of COVID cases. When measuring the rate of infection growth it made a lot of sense to use the daily confirmed cases instead of cumulative numbers. If I were doing the project individually, I would have gone ahead with this assumption but after talking to my classmates and seeing the work it was reassuring to see that most of the class was also thinking in the same line. Some of them preferred using weekly change in cases as the daily change in cases was a little too noisy. Finally, when it came to visualizing the data and answering the research question, I was momentarily stuck as I did not know how to show the change in masking policies with a change in the rate of infection. I found the collaborative assignment to be very constructive as a lot of students had different ideas, some were trying models using Time Series, others were using Facebook's

Prophet algorithm and some others were using the ruptures library and Pelt Algorithm. I chose to use the ruptures library based on a blog I read online. The discussions however helped me understand what the library actually did. It detects the points in the data when there is a change in the trend. It did not specifically help me with regard to understanding the masking policies influence but did help me understand how to read the daily change in covid cases plot. I used this as a reference to plot the rate of infections with the period of masking mandates and was able to come to answer the research question.

The fact that some of my classmates were trying out Time Series modeling like ARIMS models confused me a little as I did not think the datasets has enough information to be able to apply such models and come to distinct conclusions but it was a great way to see the thought process behind the analysis of my classmates and compare it with my own to understand where I could improve. Another thing that I found interesting was that each classmate had different issues with the data like some of them had counties where no masking mandates were enforced at all for which different brainstorming was applied to be able answer the research question.

Attribution:

- The code to plot a barh for mask mandates was referenced from Charles
- I found the changepoint detection library Ruptures online but used a little bit of Tharun Reddy's code to understand what was going on with the technique.