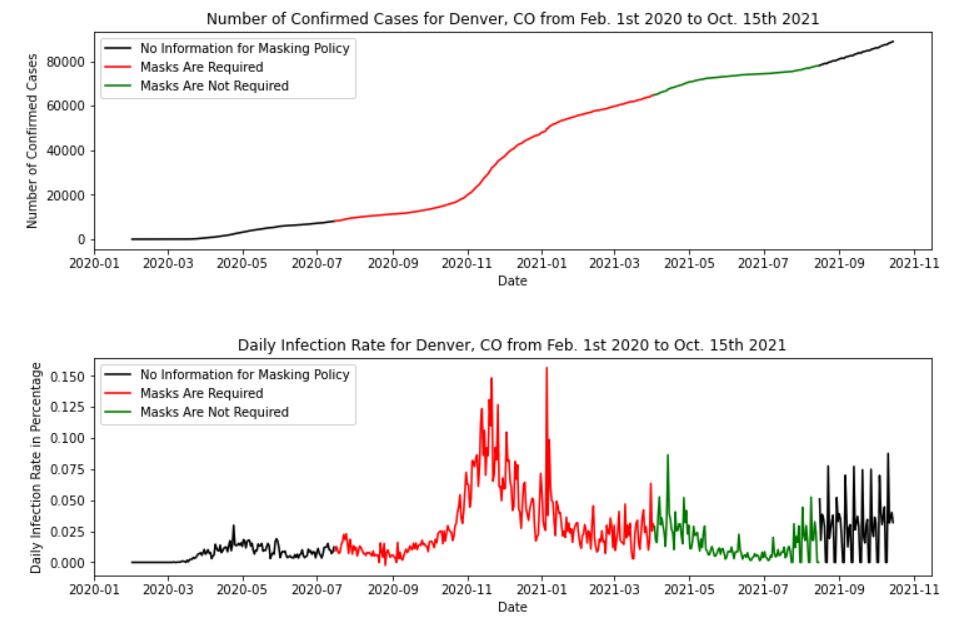
Visualization



Explanation of the Visualization

The two time-series graphs in the figure show the trend of number of confirmed cases and infection rate for Denver, Colorado from February 1st 2020 to October 15th 2021. Both graphs are encoded with three colors to represent the mask policy for a specific range of dates, with balck being unknown, red being required and green being not required. Viewers can first look at the top graph to get a sense of how the number of cases changes daily, and then look at the bottom one to observe the trend of infection rate.

The x-axis for both graphs indicate the month and year of the data. For the top graph, the y-axis is the number of cumulatively confirmed cases ranging from 0 to 8100. For the bottom graph, the y-axis is the daily infection rate, which is calculated as the daily number of confirmed cases divided by the size of population at risk. To get daily confirmed cases, I used cumulative cases from the current date to subtract that of the previous date. Because people who have been infected are rarely going to be infected again, second-infection cases are ignored and population at risk is defined by the total population of Denver minusing number of cumulative confirmed cases. Thus, the second plot can be interpreted as the derivative of cumulative cases in Denver, which allows viewers to quickly understand the trend of daily change.

The effect of masking policy on confirmed cases is very obvious from the two graphs. The most rapid increase happened around late November 2020 as the slope is very steep on the first graph and daily infection rate is very high in the second plot as well. Among 9 months of masking policy, the most rapid increase happened about halfway through, yet the derivative of confirmed cases dropped dramatically since January 2021. The slope of the first graph is a lot smoother and the infection rate is significantly smaller in the second graph. Such a trend even continues when masks are no longer required in April 2021, which means a relatively safe environment is guaranteed with a low infection rate in Denver.

Reflection

One lesson I learned from answering the research questions is to not make naive assumptions about data or any definitions. The original dataset ‘masking mandates by county’ provides information about whether people are required to wear a mask in public over time. Originally, when I saw there are ‘NA’ values in that column, I simply converted the values to ‘No’ because it’s more convenient to have binary values for the policy feature. However, after talking with one of my classmates, I think it is necessary to keep the ‘NA’ values separately since there is actually no information available for the policy. Although it may be understandable to combine ‘NA’ values with ‘No’ values, such a choice could potentially introduce external bias that changes the context of interpretation for the visualization.

Initially, I also naively assumed that the definition of people at risk is simply the total population of a county. After reading the discussion posts on Slack, I realized that there are many considerations that need to be taken into when calculating the size of population at risk. Using the total population, I’m essentially ignoring a large number of residents who have a very low chance of getting COVID-19 and as a result downplaying the infection rate by including too many people that shouldn’t have been included. In the end, I decided to take Patrick and Grant’s suggestions and define population at risk as the difference between total population and cumulative confirmed population.

There are of course other issues that are ignored by choosing such a definition. For example, it is possible for people to get COVID-19 twice, and some people may have extremely low chances of getting COVID-19 because they don’t go out very often. For this collaborative activity, different people may make different definitions based on their reasoning, but as long as a detailed explanation is provided, viewers should be able to interpret the visualization correctly.

In addition to not making naive assumptions, I’ve also learned not to over complicate the model in this assignment. There are many moments when I feel the research questions are too hard to answer and have the impulsion to take many complicated actions. There are many underlying factors that could substantially change the trend of the visualization. For example, more people may choose to test for COVID-19 during the holiday season due to travel restrictions and thus more cases are reported for asymptomatic people. However, it takes much effort and time if I take those factors into account because it’s questionable whether there will be external data available to test for such a hypothesis. As a result, I try to avoid making overly complicated assumptions as much as possible to make my visualizations straightforward but also reasonable.