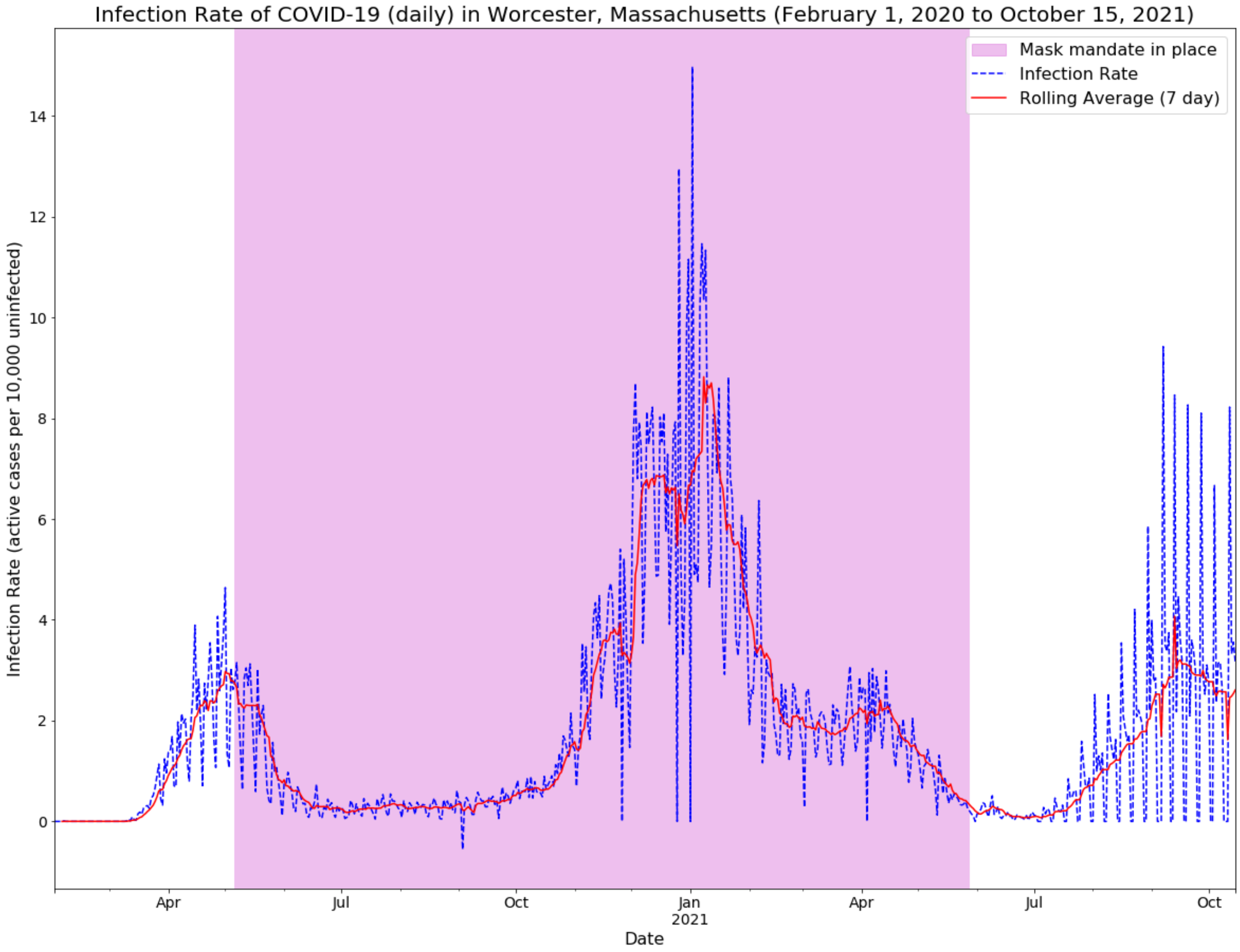
Castle Leonard
November 4th, 2021
Common Analysis
DATA 512: Human-centered Data Science

Visualization



Explanation of Visualization

This visualization includes data on Worcester, Massachusetts, a county about an hour west of Boston with a population of about 862,000 according to the 2020 census. The data comes from a Kaggle repository of John Hopkins University COVID-19 data (RAW_us_confirmed_cases.csv) containing daily reports of confirmed COVID-19 cases and a CDC dataset of mask mandates (masking mandates by county). The figure itself tracks the daily infection rate (y-axis) over time, from February 1st, 2020 to October 15th, 2021 (x-axis). Infection rate is defined as the number of new cases per 10,000 uninfected population. The uninfected population uses the fixed population from the 2020 census and subtracts the number of active cases on that day. Cases are considered active starting two days before the test is confirmed and for the following 8 days, under the assumption that people with positive tests have been infected for some time prior and the test results also contain some delay. The 10 day active infection assumption is the rough midpoint between the shortest and longest 7 and 14 days estimates of infectiousness. The pink shaded portion of the graph shows the time period in which masking mandates were in place for all indoor, public spaces and outdoor spacing when distancing could not be maintained. Lastly, to better see the slope of the infection rate a smoothed, rolling average is plotted in red over top of the raw data. The 7 day day rolling average was smoothest because it aligned with the weekly reporting patterns that had zero reported cases on some days.

The most interesting pattern is that the initiation of the masking mandate is a peak of infection, which could imply that the mandate stemmed the tide of infections, and that concludes at a trough before the infection rate again grows from July into October. It seems likely that the confounding factor would be public fear of COVID-19 which would move politicians to put measures in place and may coincide with greater caution and adherence with advice to isolate as much as possible. By the end of May, the lowered infection rate, an impatience with COVID safety measures, and a desensitization of the public to the dangers could explain why the mandate was lifted. Second most notable is that there was a dramatic peak of infection within the mask mandate at the beginning of January. This may be explained by the winter and the holiday season in the US. More time indoors and perhaps more exposure through social gatherings and lenience with masking around friends and family may explain the phenomenon.

There are too many potential confounders to say for certain, but It appears that masking policies may be effective at reducing the rate of infection when aligned with public beliefs, other associated precautions, and incentives.

Reflection Statement on Collaboration

From the discussions on Slack I was able to engage with my fellow students on the best metrics and visualization tactics to communicate the relationship between masking mandates and infection rate. It was particularly helpful to see the similarities and differences across counties with respect to masking mandates and the shape of the distribution.

From the standpoint of learnings from the data, I have only conjecture currently and would not say I have reached any conclusive learnings, however I learned about the strange formalities of Governor Orders as I read through each decree. In the Commonwealth of Massachusetts COVID-19 Order No. 31 (https://www.mass.gov/doc/may-1-2020-masks-and-face-coverings/download) there were a series of justifications structured as a series of statements in separate paragraphs, each beginning with an all caps, bolded "**WHEREAS**" that give the impressions of artifacts from a time of monarchy.

From a technical data manipulation perspective, I was pleased to learn about some pandas functions I had not previously encountered: rolling() and shift(), which allow windows of aggregation to be set and then shifted up or down to affect the value's alignment with the source data. This was particularly useful when I wanted to presume that confirmed cases had on average been active for 2 days prior to a confirmed positive test. I was also surprised when I varied the rolling average window size and found that it was very sensitive to the patterns in the data, becoming extremely noisy when the window was not aligned exactly with the 7 day cycle of the data.

I'd like to thank Grant Savage for his contribution of the rolling time window average idea and code snippet (df.column.rolling(window=7).mean().round()), which made the slope of the change more visually trackable and thus helped the visualization speak to the question at hand. Similarly, the discussion initiated by Emily Linebarger which discussed the interpretation of the prompt and Patrick Peng's contribution of discussing how to define infection rate and population at risk.