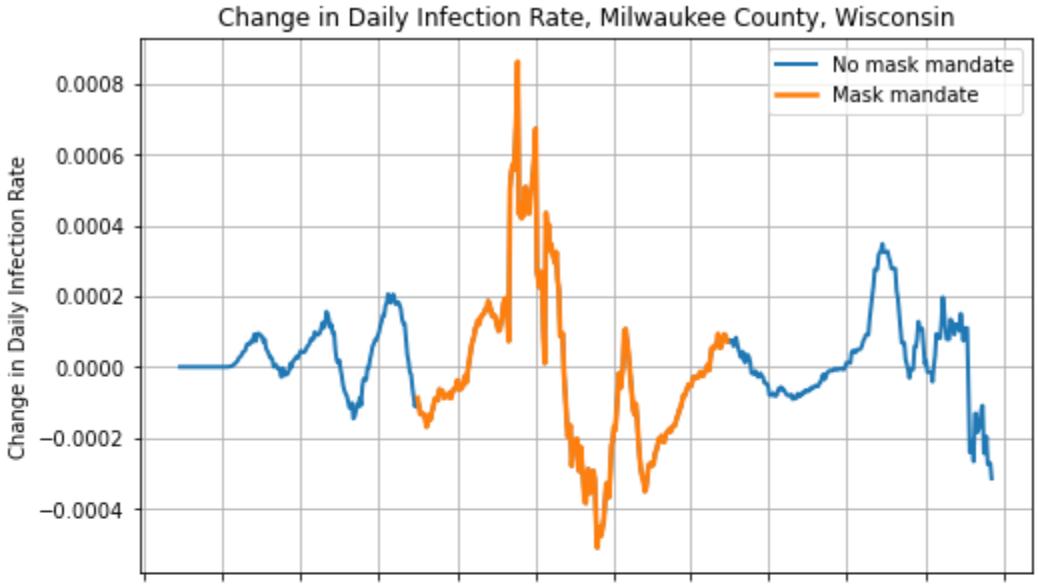
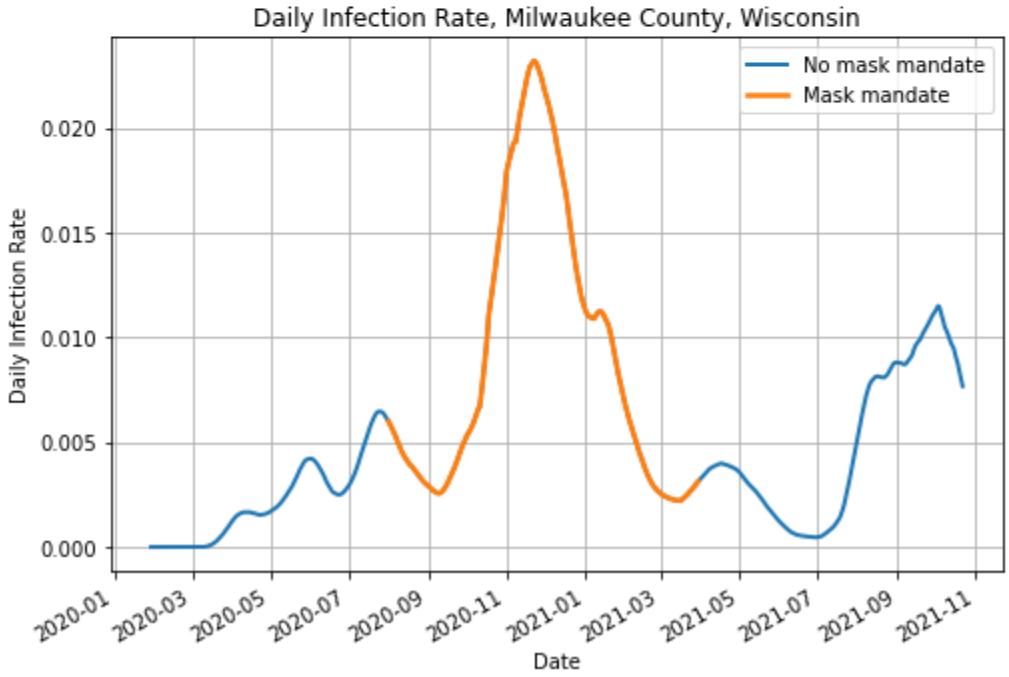
A4: Common Analysis

Patrick Peng (ID: 2029888)
DATA 512, Autumn 2021

Visualization





Explanation of Visualization

This visualization shows the daily COVID-19 infection rate in Milwaukee County, Wisconsin on the bottom plot and the daily rate of change in the infection rate on the top plot. Different plot colors indicate whether a mask requirement was in effect on a specific date. Infection rate is a parameter that describes the frequency of new infections in a population during a specific time period and is calculated as:

$$\frac{number\ of\ infections}{number\ of\ people\ at\ risk\ of\ infection}$$

This visualization uses data taken from two sources:

- The raw confirmed COVID-19 case counts, downloaded from the Kaggle repository of John Hopkins University COVID-19 data.
- The CDC dataset of masking mandates by county.

To calculate the infection rate from the case count data, I make the following simplifying assumptions:

- All infections are asymptomatic for the first 7 days (value chosen midway between the generally acknowledged range of 2-14 days).
- Upon the appearance of symptoms on day 7, every infection is reported and added to the confirmed case counts dataset. In other words, no infections go uncounted.
- Every infected patient recovers 14 days after symptoms appear (21 days after first infection). Nobody dies.
- Patients are contagious for the full 21 days of the infection and are not contagious after recovery.
- Recovered patients cannot be reinfected.

Using these assumptions, I can calculate the number of active infections on each day (the sum of that day's new infections and ongoing infections from prior days) and population at risk (the 2020 Census count for Milwaukee County less the cumulative number of recovered patients), then divide them to obtain the daily infection rate. The daily change in infection rate was obtained by calculating the day-to-day difference in infection rate.

The (bottom) infection rate chart describes the fraction of Milwaukee County's at-risk residents who were infected with COVID-19 on each day. For example, we can see that nearly 2.5% of county residents had active COVID-19 infections in December 2020, and that a mask mandate was in effect at the time.

The (top) change in infection rate chart describes the daily change in infection rate. Positive values indicate that the infection rate is increasing and negative values indicate a decrease in the infection rate. We can see that positive values on this chart tend to correspond to the build-up of the distinct infection "waves" seen on the bottom chart where the number of

infections is increasing very quickly, and negative values correspond to the suppression of each wave.

Reflection

I think the assignment document could have been improved to make the expected deliverables clearer. Based on the Slack discussion with classmates, there appeared to have been widespread confusion among the students as to whether the assignment was merely to create the time series visualization of infection rate (as specified in Step 2) or to actually perform time series modeling to predict infections in the counterfactual case (as is hinted at in Step 1).

While just plotting infection rates is trivial to perform, it does not yield any interesting insights into the effectiveness of mask mandates. In fact, at least in the case of Milwaukee County it may leave the viewer with the counterintuitive impression that mask mandates worsened the pandemic. We can see that the most rapid increase in infection rates occurs about 3 months after the imposition of the mask mandate, but it would be incorrect to infer that this was a result of the mandate rather than one of the numerous potential causes that are not included in the visualization.

We are more likely to answer the research question by comparing predicted no-mandate infection rates against the actual rates with the mandate in place, but such modeling is not covered in any UW MSDS core course and learning it is presumably outside the scope of a Human Centered Data Science course. There are other methods to draw such comparisons, such as fitting exponential growth curves to the build-up of each infection wave (something I attempted), but it is difficult to show the results of these fits on the specified visualizations. Furthermore, blindly fitting models to data without any underlying knowledge of epidemiology feels like an excellent opportunity to repeat the "cubic model" gaffe made by former Council of Economic Advisers chair Kevin Hassett, who early in the pandemic used a stock Microsoft Excel cubic polynomial fit to the COVID-19 infection data and infamously predicted that the pandemic would end by May 2020. While the stakes here are not as high (none of us are advising the President), it still seems like a practice to be avoided.

Lacking any prior knowledge or experience with epidemiological modeling and unsure whether the terms of this assignment required me to learn it, I opted to just create the time series plots specified in Step 2, fully aware that there were no meaningful conclusions to be drawn from them regarding mask mandate effectiveness. However, I also decided to perform some modeling of my own in my notebook (not submitted as part of this assignment) after reading Grant Savage's reply in Slack about trying to fit GLMs to the data at different time intervals before and after the imposition of mask mandates. Having heard that pandemics spread at an exponential rate, I tried to fit the function y=A*exp(B*x) to the infection rate curve at various points to see whether the coefficient B (which influences how rapidly the function increases) changes after the mask mandate is put in place. If mask mandates decrease the rate of disease transmission, we would expect the value of the B coefficient to be smaller after the mandate takes effect. I used the following code snippet that Grant posted, with some modifications, to

quickly and easily take intervals from the time series that corresponded to the exponential build-up of each infection wave:

```
mask = (df.index >= start_date) & (df.index < end_date)
Df.loc[mask]
```

As can be seen in the infection rate plot, Milwaukee County experienced two distinct “waves” of infection starting in April and June 2020, prior to the mask mandate, and two waves while the mask mandate was in effect, in September 2020 and March 2021. I ignore data after about May 2021 as the combined effect of vaccines and the Delta variant render any comparison to earlier data inappropriate. After fitting the exponential model to each wave’s build up (from trough to peak) using Numpy’s polyfit method, I found the following:

Wave	Mandate?	B coefficient value
1 (April 2020)	No	0.032
2 (June 2020)	No	0.038
3 (Sept 2020)	Yes	0.038
4 (March 2021)	Yes	0.013

We find that post-mandate Wave 3’s coefficient is practically the same as the pre-mandate Waves 1 and 2, but Wave 4’s coefficient is significantly lower than the rest. However, with only two waves before and after the mask mandate, and with one of the post-mandate waves (3) almost certainly driven by non-mask factors, it is hard to draw any conclusions about the effect of mask mandates. This method of analysis is also somewhat sensitive to the exact interval boundaries used for selecting points for fitting, as that can affect the resulting B coefficient value significantly.