Data 512: Human Centered Data Science

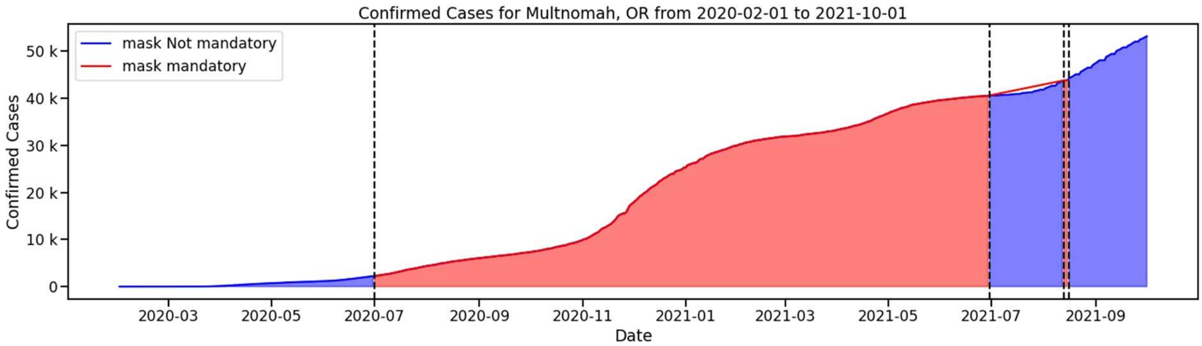
Project Part 1: Collaborative Analysis

Juhi Choubey

Analysis Explanation:

We will cover the analysis in the top-to-bottom order to understand the process and all the caveats surrounding the calculation methodology of certain metrics.

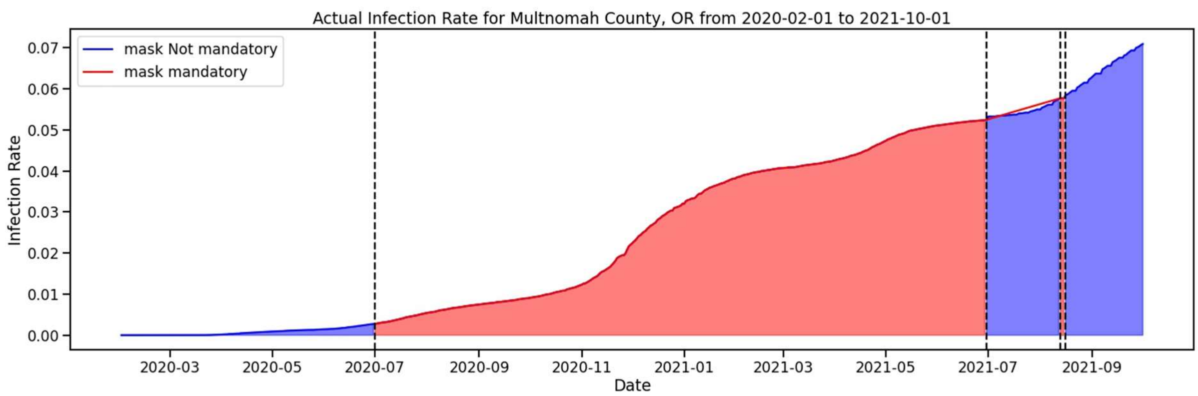
Viz 1 *(optional)*: Time series showing the trend of confirmed cases



This figure shows the rise in daily confirmed covid-19 cases in Multnomah County, Oregon between Feb 2020, and Oct 2021. The x-axis holds the daily days in this time, while the y-axis represents the number of confirmed covid-19 cases. The vertical dashed lines indicate the change in masking policy for this county – with the red shaded region representing days where masking mandate policy was put in action while the blue shaded region representing the days where masking mandate policy was not active.

Note: The data regarding masking mandate was unavailable for dates before Jul 2020 and dates after Aug 2021. Before Jul 2020, covid was beginning to spread and there was no masking policy implemented. Therefore, instead of dropping these records, we have updated the masking policy for these dates to 'Not Mandatory'. Same for dates after Aug 2021.

Viz 2 *(optional)*: Time series showing the actual rate of infection

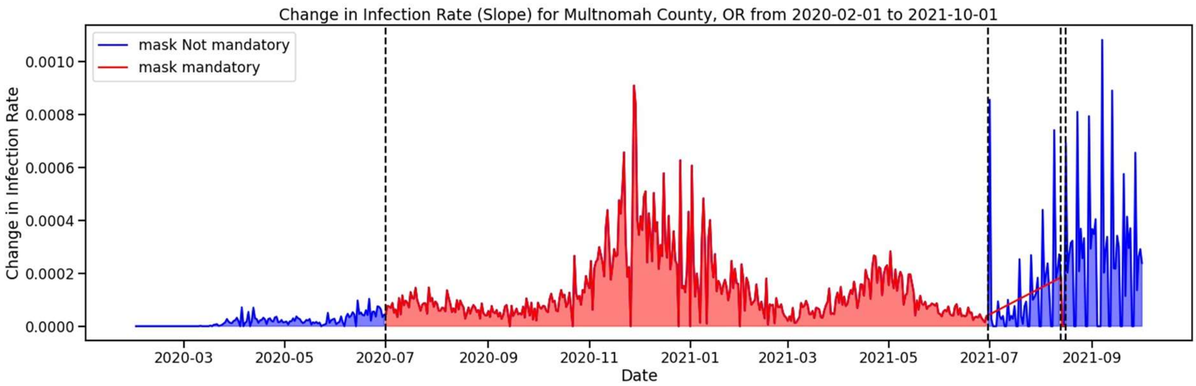


This figure shows the rise in daily rate of infection of covid-19 cases in Multnomah County, Oregon between Feb 2020, and Oct 2021. The rate of infection was calculated as:

Infection Rate = # Confirmed Cases / (Total Population - # Confirmed Cases)

Caveat: In using such methodology, we are aware that there are many considerations that are not put in place. For instance, the population at risk is a selective set in the total population that is currently healthy and not infected with covid. This includes all people who are alive, not currently infected, and already recovered from covid if they had it earlier. For this analysis, we have only removed the current infected folks and not considered the rest. Additionally, factors such as hospitalization and vaccination are bound to bring down the infection rate or even make some significant change to the infection rate calculations. Since the data for these is unavailable and not considered in the current analysis, it will instil some assumption errors into our understanding of rate of infection.

Viz 3 *(required)*: Time series showing the changes in the rate of infection



This figure shows the change in infection rate with time with both mask mandate enforced (red) and not enforced (blue). Change in infection rate is calculated by taking the difference in the infection rate between 2 consecutive dates (current-previous). We can see that highest peak in change in infection rate occurred around December 2020 – this is sort of indicative of the new variants of covid-19 that were introduced during that time. We can also see that the change in infection rate observed a steep fall twice: once around March 2021 and next around July 2021: same time around when mask mandate was removed. The viz does not really highlight the effect of mask mandate policy on change in infection rate.

Owing to this, I also proceeded to model the time series of number of confirmed cases with AR (1) process (lag of 1 was highly significant per autocorrelation plots), and conditioning on mask mandate. I observed a regression coefficient of -86 associated with mask mandate on forecasting # confirmed cases (with p value < 0.05). This suggests that there is a significant negative effect of mask mandate on # confirmed cases: when mask mandate is imposed (=1), there is a slight fall in # confirmed case. However, given the above caveats addressed, we will still take this model with a pinch of salt – it is not accurate.

Refection on Collaborative Analysis:

This analysis was one of the most nuanced ones we have been assigned in this course so far. Firstly, the lack of a good data landscape meant that we were left with making multiple assumptions concerning our data points. Secondly, to enable reasonable analysis, it was imperative that we consider and reflect through the epidemiology around covid-19 infections. These factors demanded a good discussion to incorporate sound reasoning in the analysis – as it is highly unlikely for one person to come up with every nuance, assumption, and though process behind the analysis.

Performing the cleaning and standardization of the available datasets to have 1 master file with all the relevant information was the first step in the analysis. There wasn't really any collaboration required in this step as the dataset provided were pretty standard without much to clean. There were certain null records that required attention: for some time at the beginning and the end of the dataset, the data pertaining to mask mandate was unknown. I followed up with our class fellows on discord and learned that while some folks decided to drop these null records, others proceeded to update the mask mandate policy for these dates to 0 or 'mask not required'. The reasoning was pretty logical – as the mask mandate was not fully in action in the very beginning of covid-19 spread and has become considerably lax at a later time frame. I decided to follow the latter approach.

I was able to chart out the 2 visualizations concerning the trend of number of daily confirmed cases and change in daily confirmed cases. With regards to rate of infection, I followed up on Kaggle to look at the concept of Rate of Infection and how to calculate this. I found a pretty relevant article that considered many factors (as highlighted in Analysis Explanation Viz 2); however, due to lack in data availability, I decided to proceed with a simple base methodology in calculating rate of infection. I discussed my approach with many people from class – and most of them agreed that the low data landscape hindered most of the analysis, and that diving deep into the epidemiology tended to mess up the analysis further. Hence, for Part 1: the analysis performed is benchmark.

There was also a good discussion around detecting 'changepoints' in the time series. While some people proceeded to implement and detect changepoints in the given series, my thoughts were different. I did not believe detecting changepoints was the problem we are trying to solve. The actual problem at hand was to understand if mask mandate was significant in affecting the number of confirmed covid-19 cases. So, instead of going the changepoint route, I moved on to build a base model considering mask mandate as a feature to forecast number of confirmed covid cases and check if this feature was statistically significant in estimating number of confirmed covid cases.

I would like to attribute this step to my class fellow: Syed Obaid Dawarki, as I was able to discuss my findings and align on whether this approach is relevant to our analysis. I find that it made perfect sense to make this part of the project a collaborative work since there were many areas wherein, we needed to understand different possible approaches so we could see the signal from the noise (*pun intended*).