For open ended problems like these, it's interesting to collaborate and understand how other people approach the problem.

I learnt a lot through the process, specifically:

* There are a number of ways to approach a given problem and talking to others, gives you a holistic idea of the problem
* I now know how to use rolling averages in python

2 specific experiences:

1) I had an amazing discussion with a fellow student in trying to look at the big picture of the problem.

Below is a summary of our discussion:

As per the prompt, we are interested in understanding how masking mandates have affected the progression of confirmed COVID cases.

Taking a step back, we agreed that there were a number of factors that determine Confirmed COVID cases numbers

* No of people vaccinated
* No. of people wearing masks
* No. of people at any given time that have COVID
* Climate (?)
* ...

However, we do know for certain (based on research papers) that wearing masks reduces the spread of covid.

We want to try and see if we can maybe quantify the relationship between:

Issuing Mask Mandates → More people wearing masks → slower progression of COVID

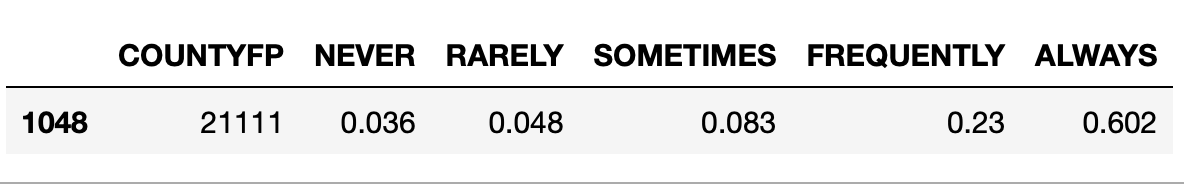
Looking at the visualization that was created (fig1), after the mask mandate was removed, there was a sudden spike in the cases.

This provides strong evidence in support of a modified version of the above claim:

Removing Mask Mandates → Faster progression of COVID

While this is a good indication of the claim, we came to the conclusion that we need more evidence to claim causality.

Based on a survey that was conducted from July 2nd - July 14th, in Jefferson County (KY), around 85% wear masks for the majority of the time.



This snapshot, corresponds to the just before/during the mask mandate was first issued (July 10th 2020)

Are these numbers due to the mask mandate, or would they be even higher if they were recorded after the mandate came into effect?

To be able to answer these questions, it would be nice to have a time series of mask survey data rather than a single snapshot that was provided. If we had this data for time periods where mask mandates were in place as well as removed, it would give us more insight.

2) I also followed the slack discussions and I learnt alot from the thread where people were discussing how they dealt with missing data.

Plotting the data for Jefferson County, Kentucky, there were a lot of days where no data was logged. Digging a little deeper, uptill July 2021, these days, mostly corresponded to holidays. However, since then, Saturdays and Sundays were also days with no data logged.

To smoothen and address this seasonality, the consensus seemed to be to take a 7 day average

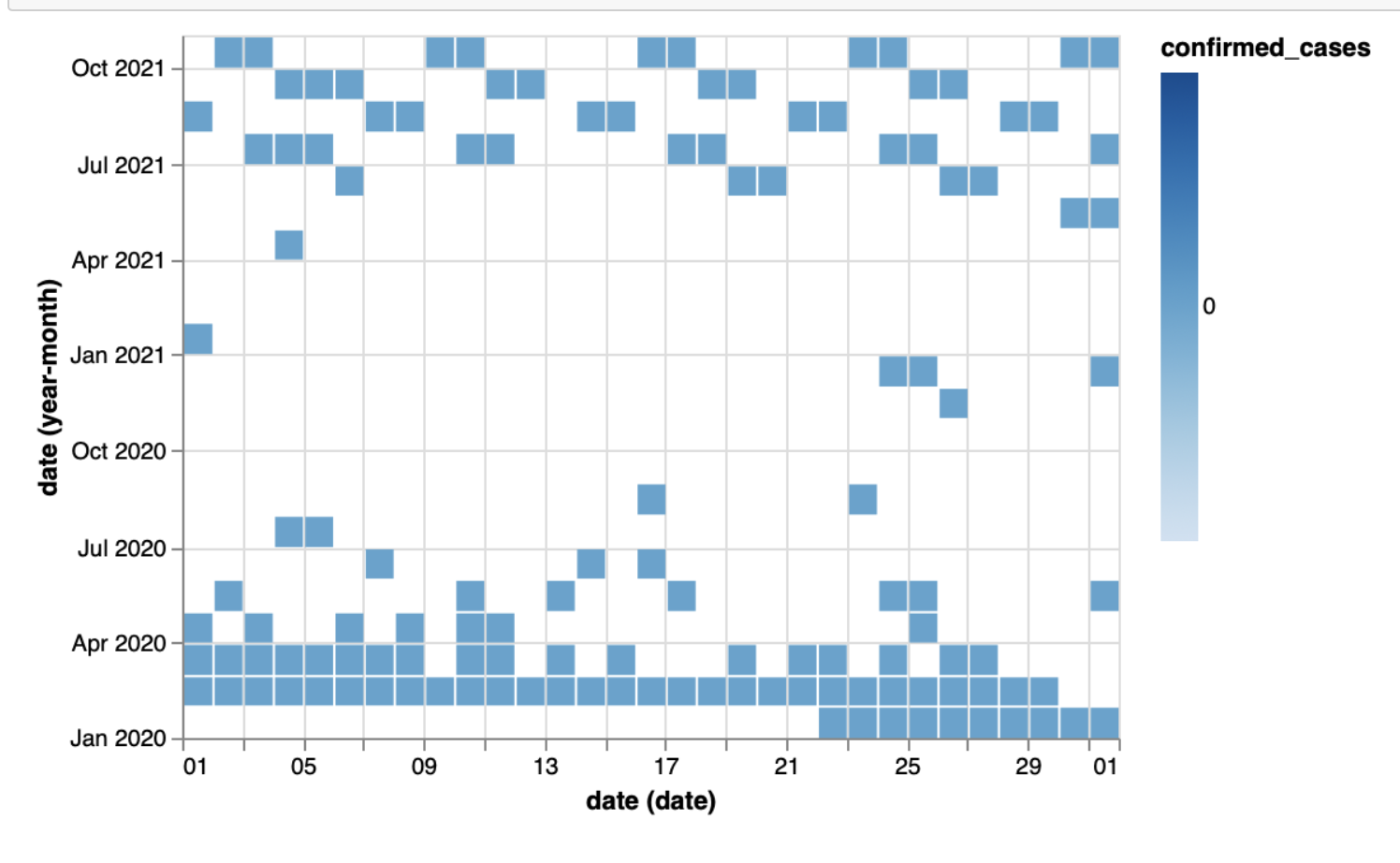
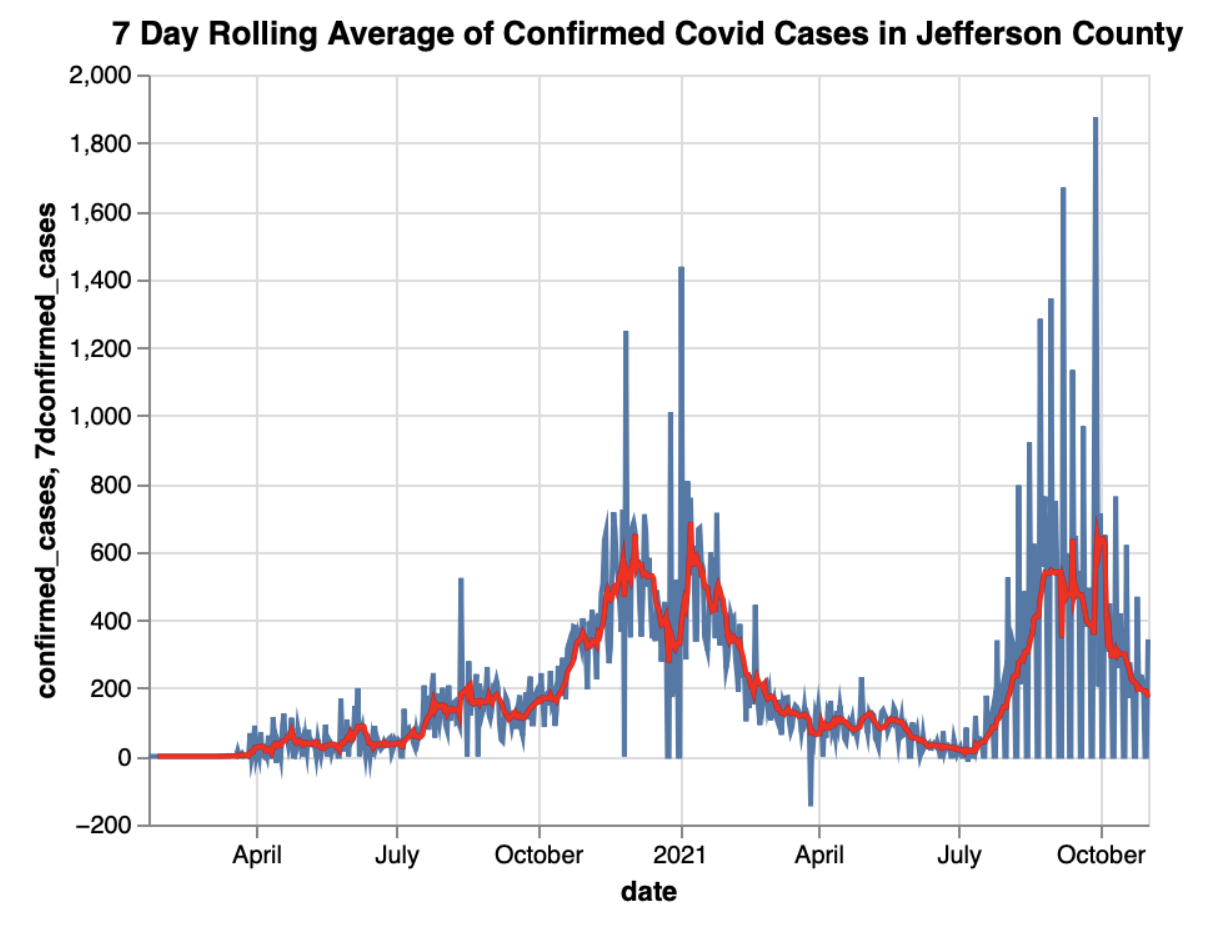


Fig:(left): 7d rolling average of confirmed covid cases in Jefferson County, KY; (right) days for which case count was 0 in Jefferson County, KY.

The issue with the 7 day average is that it creates a lag in the noticeability of a change in behaviour. This added to the fact that cases have an inherent lag (from point of getting covid to the point of it being reported), causes the graph to be shifted slightly to the right. I decided to not correct for that at the moment, but rather be cognizant of it and mention it in the analysis.