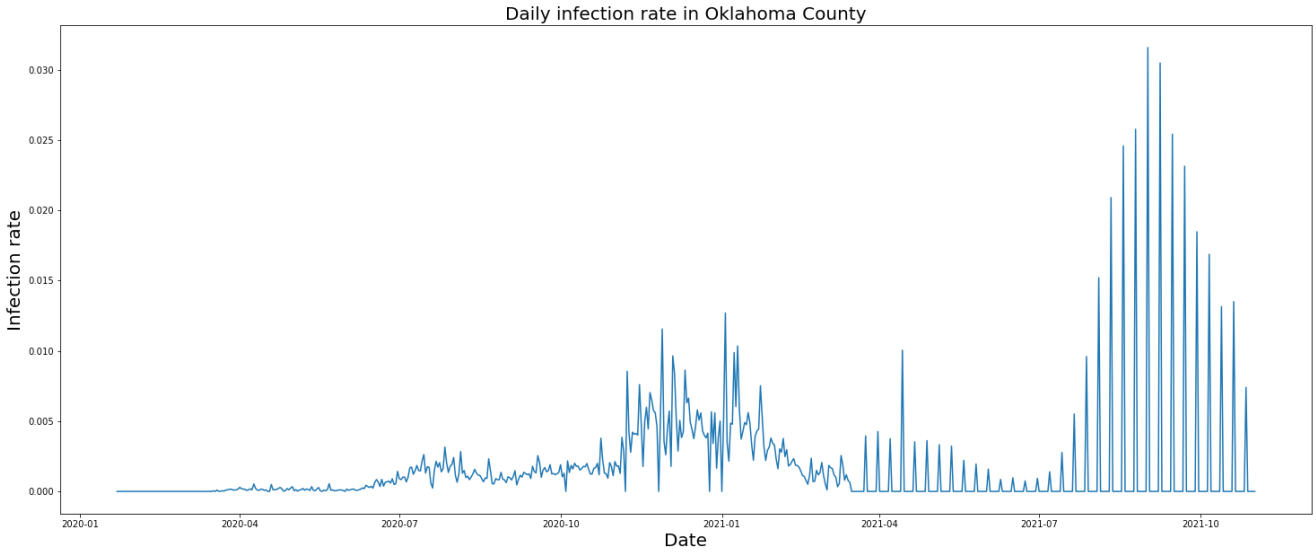
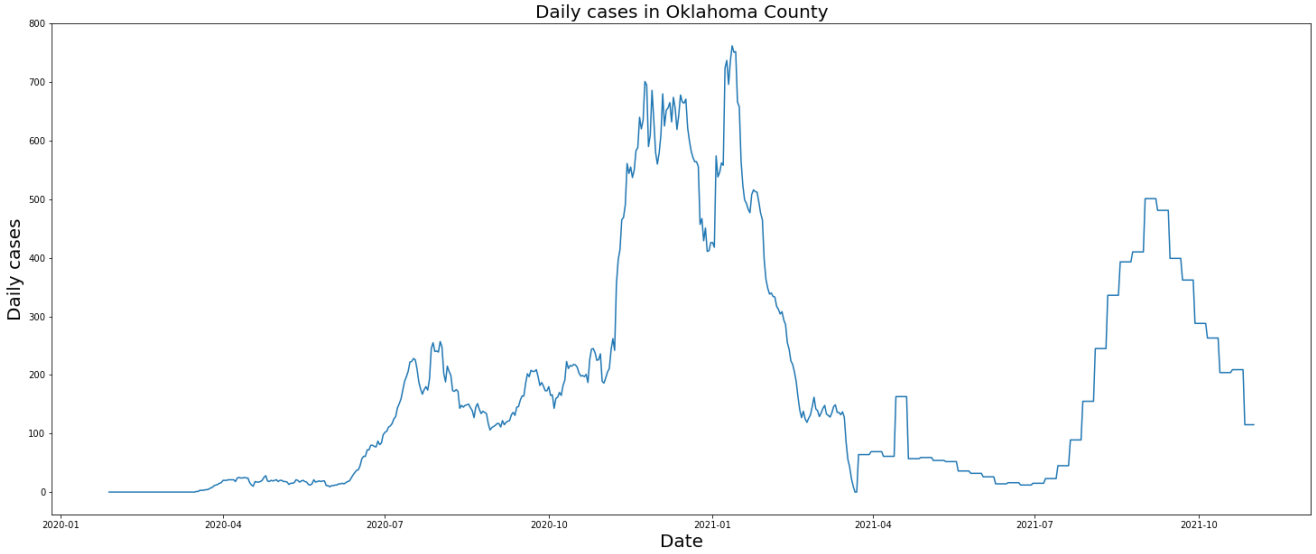
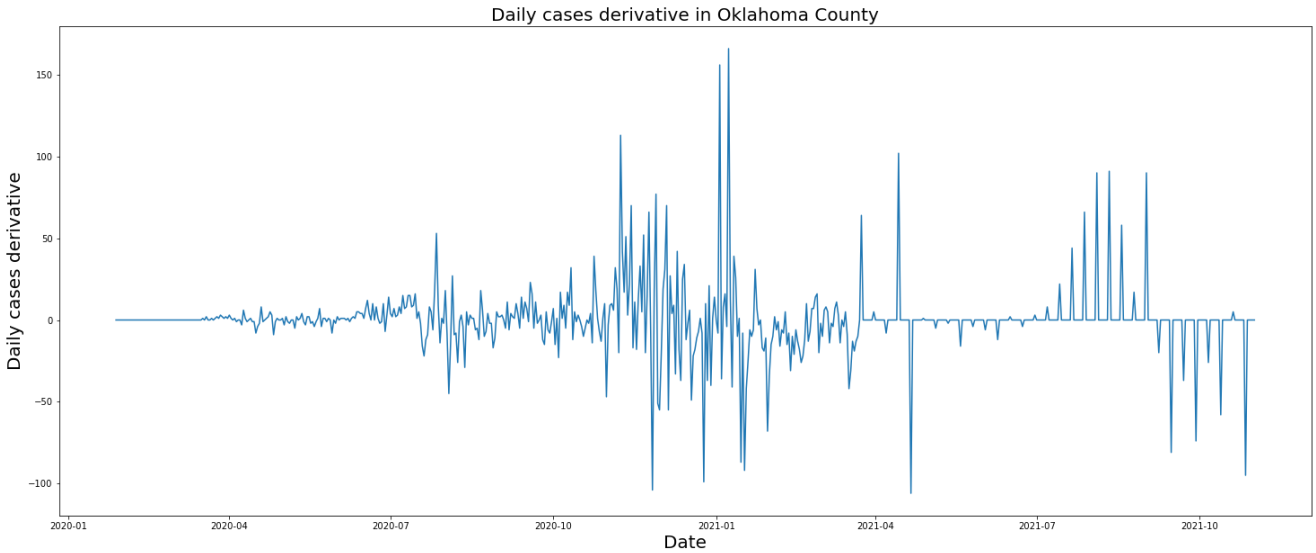

Daily infection rate in Oklahoma County
Infection rate
0.030
0.025
0.020
0.015
0.010
0.005
0.000
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
Date


Daily cases in Oklahoma County
Daily cases
800
700
600
500
400
300
200
100
0
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
Date


Daily cases derivative in Oklahoma County
Daily cases derivative
150
100
50
0
−50
−100
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
Date

Second, we would like to understand what you got out of the collaborative activities in this assignment. You should write a reflection statement that highlights one or two specific things that you learned from answering the research question posed in this assignment. Your reflection statement should include specific attributions for any/all code, methods and techniques that you reused. Your reflection statement should be no more than 2 written pages.

My visualization displays the daily infection rate in Oklahoma County, Oklahoma between January 22, 2020 and November 1, 2021. According to Wikipedia, the infection rate can be defined as

$$\frac{the\ number\ of\ infections}{the\ number\ of\ those\ at\ risk\ of\ infection}$$

Now this is a fairly straightforward equation, however being at risk of infection is open to interpretation and must be clearly defined, accordingly. Using the mask use data, I specified those that were at risk of infection as the percentage of the population that rarely or never wore masks. According to the mask-use dataset, 9.6% of the people never wear masks and 6.4% of people rarely wear masks in Oklahoma County, for a grand total of 16%. I also made a few assumptions in order to make my day-to-day computation of the infection rate more consistent. I assumed that those who had been infected could not be infected again. I also assumed that the previous proportions stayed fairly constant as the pandemic progressed, and so 16% of those who had not been infected was the total number of those at risk of infection each day. Therefore the daily infection rate was defined as

$$\frac{number\ of\ infections\ on\ day\ i}{0.16\times(population\ of\ Oklahoma\ county - population\ that\ had\ already\ been\ infected\ as\ of\ day\ i)}$$

Overall this employs the data from the mask-use dataset and the confirmed cases dataset. Unfortunately Oklahoma County and Oklahoma, as a state, did not have any mask mandate data since the beginning of the pandemic, and so I was unable to incorporate that data into the visualization, as it would have required me to utilize outside data. But this is essentially a time series plot with the date on the *x*-axis and the daily infection rate on the *y*-axis.

Something we do see here is a fairly normal level of volatility through March 2021. However, this occurs at a very low infection rate (< 1%). It's difficult to understand why this is the case, as it could be due to low volume of testing, or people simply not getting infected. But if we also look at the accompanying graph that displays the daily cases, we see that the number of cases spikes in the Fall/Winter of 2020. For full transparency, this daily cases plot has been smoothed out over 7-day windows to account for the delay between contraction and testing positive for COVID. We also have the derivative of the daily case plot, which gives further insight into the volatility of the data, particularly because of its own volatility around the zero axis. We can see how the infection rate, cases, and derivatives change between the months of July 2020 and March 2021, as well as the seasonality being reflected after April 2021. After

doing some external research, I found that Oklahoma did not actually institute a mask mandate at all, so the general population did not have much defense against the virus, except for the approximately 74% of people that frequently or always wore their masks in Oklahoma County. Nonetheless, because of our lack of data, we cannot see what the effects of masks were because there were not tangible mandates put into place. Interestingly, after April 2021, we see much more volatility, but also much more seasonality. The infection rate spikes at seemingly consistent intervals, suggesting there is some systematic trend in the testing patterns or when people contract COVID. It's possible that people only get tested and test positive after social events on the weekends, resulting in this seasonal trend in these spikes of daily infection rates, but we cannot be sure. Furthermore, due to the lack of mask mandates, the infection rate is generally trending upward after around July 2021, which is also displayed in the second graph of daily cases.

Having this assignment be a more collaborative activity truly allowed me to work with my peers and gain different perspectives and approaches to solving similar problems and conducting analyses. For example, every student was assigned some subset of the same dataset, partitioned according to county in the United States. While in theory, we should all have been able to conduct the same analysis, more or less, producing varying results, not all of us had all the data necessary to conduct these analyses. One of the major discrepancies between datasets was the lack of mandate data for some counties. This data would have been extremely helpful for providing deeper insights on the effects of masks on the number of cases and infection rate from county to county. Oklahoma County did not have any mandate data and so I made a few attempts to improvise and impute this data. I tried looking for data from other counties in Oklahoma to see if there was a general pattern I could find that could replace the missing values that I had to work with. Unfortunately, the entire state of Oklahoma did not have any data on mask mandates, and furthermore, I found out (via my own research) that Oklahoma did not actually enforce any mask mandates to its population. However, this information was not in the data so I decided not to use it, and simply worked with the data that I had. One interesting suggestion on the Slack channel was to then look at surrounding states—these included Kansas, Arkansas, and Texas. Unfortunately, I was unable to find any relevant mask mandate data for these states either, and so it seemed as though I would not be able to leverage any sort of information regarding the enforcement of masks in Oklahoma County to conduct the analysis.

The questions also helped me think about the analysis in a way I might not have originally thought to consider. Looking at the rate of change of daily cases is also useful in looking at whether the number of cases being detected is slowing down or speeding up. This is another angle at which we can look at this problem to see how masks were affecting transmission. Typically, it's very easy to see this type of a real-world problem from a more pragmatic perspective, and simply look at what is happening. But because of the enormous amount of numerical information that we have at our disposal, we should be able to use a mathematical approach as well, as it provides different insights that we would otherwise not be able to produce.

Once again, my data did not have mandate data, and forced me to really go into a deep think about how I could work around this type of a problem, particularly because data is not always clean and readily available in the real world, and so it's very important to be able to know how to handle these scenarios. Furthermore, we also had to account for the fact that there was a 7-10 delay in contraction of COVID, and so this needed to be accounted for when plotting the daily number of infections. I used the code in one of the Slack threads to smooth out the plot over 7-day windows, by taking the mean across each 7 day window. This removed any seasonality that would have been in my plot, and made the plot more interpretable to how the cases were really changing over time. The removal of seasonality mitigates the pattern we had already discussed that occurs after April 2021, where we see spikes in cases and infection rate and consistent intervals through October. Obviously there is some system pattern in the way testing is being administered that does not actually have anything to do with how and when people are infected. So it is essential to remove this seasonality as it does not reflect real life.

Overall, this was an incredibly informative and useful exercise in manipulating and analyzing real world data, as well as employing ideas and approaches from many different minds in a collaborative setting. Many of these "soft skills" are also extremely useful in beginning a data science career, while also sharpening my data analysis skills.