From this assignment, I learned that data analysis relies heavily on assumptions. Unlike what I have experienced before in my academic career where all data is provided in a nice and usable way, real data suffers from many problems and the information provided may not capture the whole picture, such as many edge cases and side factors. For example, the data contains information at a county level. Analyzing and visualizing for all counties will be overwhelming to our viewers, but analyzing and visualizing for only one county comes with the problem of small data size. Our visualization only captures around 600 days of data, and many things could have happened in these 600 days, such as government elections and new variants that heavily affect the trend of our data. However, since COVID has been around for only so long, we do not have more data and are forced to make the assumption that the mask mandate is the factor affecting confirmed cases. Such an assumption is realistically false, but practically necessary because we cannot scientifically account for all other factors and model them into our data. We can only make our best effort to reduce the effects of these other factors. For example, to account for the incubation period, we simply delay the mask mandate time by the average incubation 5.6 days (rounded to 6 days), which is obviously rough but the best estimate we can have. We also make the assumption that the entire population is at risk of COVID-19 when calculating the infection rate because estimating the count of population at risk requires much more data, such as daily activities, which is unavailable to us. We have also assumed that the vaccination status is largely constant, which is untrue since more population gets vaccinated as time goes on. However, I still make this assumption because my data of Fulton indicates it always has a mandate in place, so I compare it with another county, and here I am instead assuming that the two counties have similar vaccination status, which is less dependent on time.

Another thing that it is important for data scientists to discuss on acceptable and scientifically reasonable approaches to address problems. Sometimes problems are in the data patterns. For example, the data indicates that more people are confirmed of COVID on Monday. Such periodicity is likely due to medical facilities operating more on weekdays as opposed to weekends, and it interferes with our analysis. Discussion from Slack suggests a solution of averaging over 7 days to avoid this periodicity, which I also think is reasonable. Other times, problems are due to lack of data. For example, for my county, Fulton, mask mandate is issued throughout the date range of available data. There are not any assumptions I can make to solve this problem of not having mandate vs. without mandate data to compare. A discussion from Slack suggests that we may choose to compare with other counties and measure the mask coverage using the New York Times data of identifying masks wearing habits. I chose Valley county to compare because it has the lowest proportion of survey participants “always” wearing a mask. Doing so carries other risks. For example, the demographics of the two counties might be different. The geographical locations of the two counties might cause COVID to spread to the two counties at different times. The population density may cause COVID to spread more easily in one county over another. Nevertheless, with the current available information, this is the best we can think of to perform our analysis, and I would definitely not think of this approach without

the discussion in Slack. Therefore, discussing with peers is important to both come up with new approaches and discuss validity of different approaches when facing problems.

Lastly, for the result of analysis, although Fulton seems to have more population wearing masks than Valley, it appears from the graphs that the two counties' infection rates do not differ significantly. This is a surprising result to me, but due to so many restrictions of data and assumptions made, as stated above, this conclusion may not be valid. It also depends on the rest of class for an overall conclusion to see if this is also true for other counties.