Shubha Changappa Palachanda
Student ID: 2129848
DATA 512- Human-Centered Data Science

Reflection Statement

(Part 1 - Common Analysis)

Interactions with fellow peers are always beneficial in Data Science endeavors, especially when tackling complicated, ambiguous issues. It was critical for this exercise to gather and share feedback for people who were blocked or confused at any point in the study. Various approaches to the problem, acceptable assumptions, design considerations for the solution, and the use of diverse techniques were all extremely beneficial in developing this final visualization.

I started this exercise by first pulling in the 3 required data sets for my assigned county (Jefferson County, KY) and performing all the basic data cleaning EDA steps before creating the visualization. As I didn't have much experience with time series analysis, I took guidance from my class discord team discussions to get started with the analysis. I noticed that folks were utilizing the idea of identifying change points in the time series data to identify the time frames with abrupt variations in daily case rates. These discussions helped me a lot in understanding the concept of changepoints. Through a bit of self-study, I learned that change point detection can be great for use cases like Detecting anomalous sequences/states in a time series, detecting the average velocity of unique states in a time series, and even detecting a sudden change in a time series state in real-time. I tried fitting a model to identify the changepoints using the prophet module suggested by a classmate, however to my dismay, I realized that the fbprophet library doesn't work on certain versions of windows, conda, and even NumPy and pandas packages. I then looked up alternative methods to identify change points and came across the medium blog "Detecting the Change Points in a Time Series" where the Pelt Search method was utilized. I learned that the PELT method is an exact method, and generally produces quick and consistent results. It detects change points through the minimization of costs. It's also interesting to note that the algorithm has a computational cost of O(n)(where n is the number of data points) With this background, I then used the ruptures package to perform offline change point detection.

I then followed this up by incorporating the mask mandate policy into the changepoint detection plot to better understand the impact (if any) on the COVID-19 infections information for Kentucky County. I noticed in the group discussions that visualization was limited to only durations (time ranges) where the mask mandate policy data was available. I wanted to realize the change points in the infection rates outside of this time-bound, to see how much of an impact the mask mandates really make on the infection rates. Hence, I included all date points in the final visual which really helped in comparing the trends and change points during different time frames. When plotting the visual for all dates, I also noticed that there was a huge spike in daily cases and for most of the weeks in 2022 case counts were entered in only on the Saturday/Sunday of the week. This reminded me of the discussions we had during week 5 regarding why the counts are high/low over the weekends and how it corresponds to hospitalization and other factors which are currently not included in this common analysis exercise. Nonetheless, I utilized a 7-day moving/ rolling average approach to account for case counts not being updated for every single day in a week to help smooth out-trend information by creating a constantly updated average value in the plot.

Overall, I think collaborating and brainstorming with people helped a lot to understand the ask of the exercise and also to validate my assumptions and get feedback on the techniques I used. I would like to specifically thank Eli, Tharun, Nayantara, and Ariel for all discussions we took part in to gain unique perspectives for the analysis of the data in hand.