Part 1 : Step 3

We aim to visualize through a time series graph, what are the changes in the infection rate (or the change in number of cases on a day level) for Harris County in Texas.

The 3 datasets provided were cleaned and standardized so that we can merge them to create one dataset with meaningful information. We extract the FIPS of the county from one dataset and use it to merge the second dataset. The mase use by county dataset did not have time series information because it is based on survey responses at a given point in time, so that is not of much use in this particular analysis.

The final raw dataset used for this analysis has the date and the number of covid cases on that day, and the masking mandate as of that day. We then engineer some features to create effective visualizations to understand the effect of the change in masking mandates.

There are many assumptions we make for this analysis. Due to limited data available in the 3 datasets given, we assume that the factors external to the data have no effect on the infection rate. The deaths due to covid, the effect of vaccinations etc has not been considered due to the simplicity of the analysis expected.

We first plot and see what the masking mandates look like over time. And then we try to map the effect of those changes in the change in the number of covid cases.

The line graph shows the number of COVID cases daily, and the color of the line depicts the masking mandate at that point in time. For the times when we do not know the mandate, we color the line blue, orange for when there was a masking mandate in place, and green when there was no mandate in place. When the mandate was put in place, there was a dip in the number of cases, and we also see a rupture point at the place near the dip in the graph. This validates our assumption that there are decreases in the number of covid cases when people start wearing masks.In addition to this one, there are more rupture points, in the graph showing changes in the infection rate. The population of the county is assumed to be static over the time period considered for analysis, so the trend in the visualization is not affected whether we take into account the infection rate or the number of COVID cases.