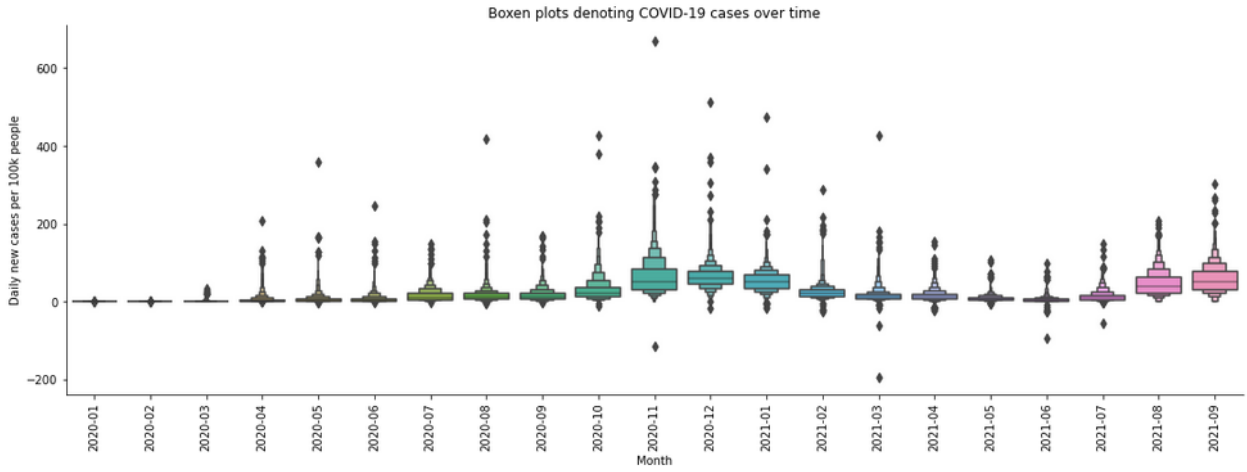
Design Document, COVID-19 Set

Consuelo Ugarte, Jake Hefty, Noel Ngui

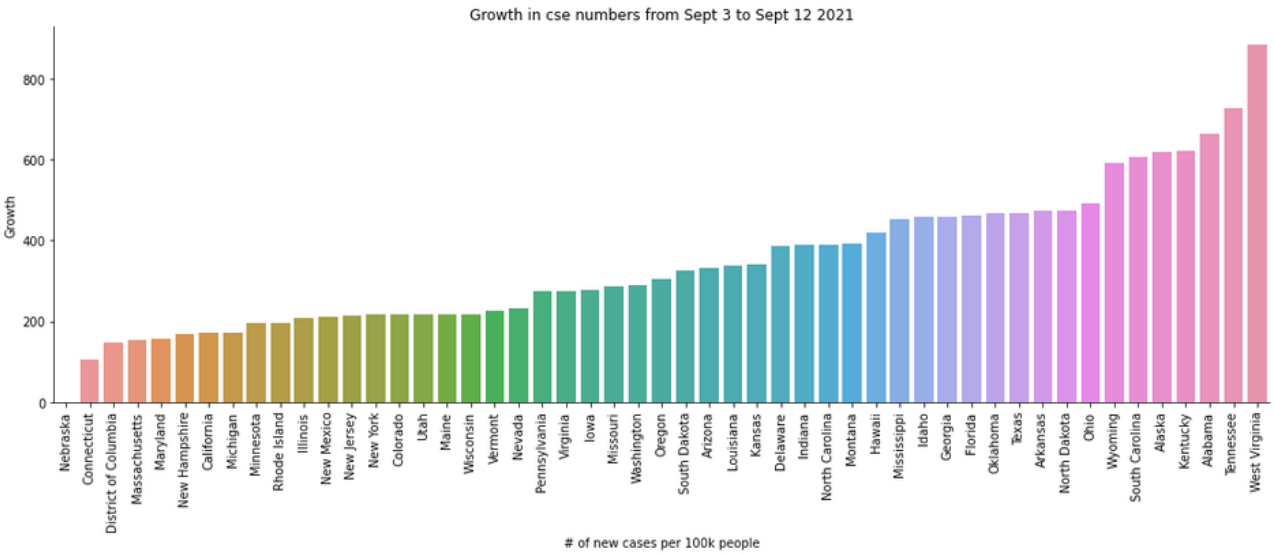
EDA

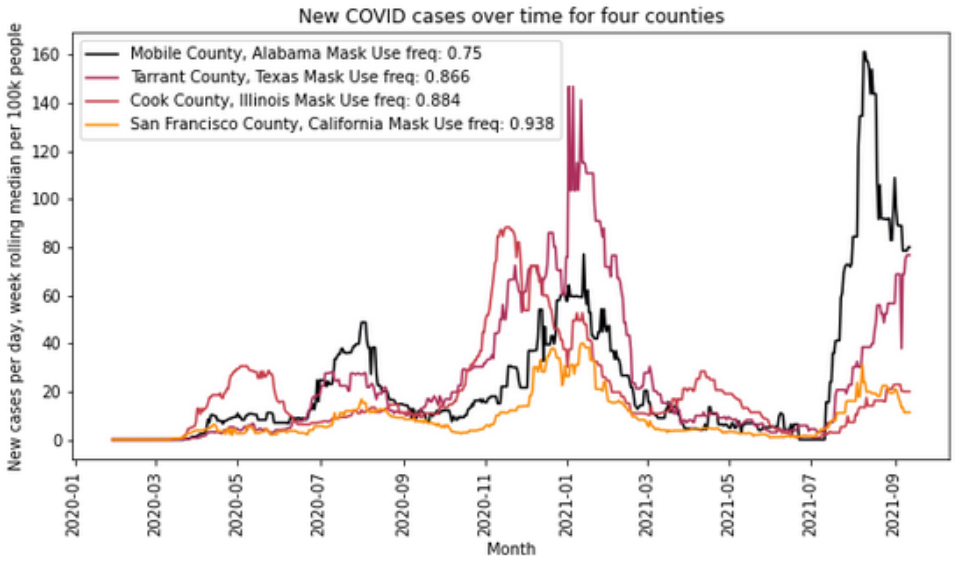
We created a visualization that shows the overall county-wise distribution of the number of COVID-19 cases per capita across the United States as a function of time through a series of boxen plots for every month, starting from January 2020 to September 2021.



We noted that the changes in concavity of the graph was interesting, and that they might form some sort of pattern.

We also graphed states' increases in number of COVID-19 increases for a week, and it almost seems as if states with low populations have the greatest increase in cases—but that might be attributed to intrinsic properties of the data that are not meaningful, since density is probably a important factor than population.





We graphed a few counties along with their mask use, and it was interesting to see a difference between cases—lower mask usage did not necessarily mean less new cases per day. However, we saw that a county had an inordinately high mask usage—which led us to hypothesize that maybe there exists a mask-usage threshold that makes mask wearing effective.

Hypothesis

A two-month cycle is an effective tool for predicting the spread of COVID-19. Effective means, in this case, that a model created using a two month cycle will predict COVID surges better than random chance.

What led us to this hypothesis

We saw a pattern that was two-month like in dips in new COVID-19 surges.We did think of modeling COVID-19 cases alone, but we thought that that might be biting off more than we can chew. So instead, we thought that perhaps we can model the next change in positivity of the slope, instead of the magnitude of the number of cases. This theory of the two month pattern in COVID surges was corroborated by multiple news agencies so it might be revealing to compare our results and observe if the data was potentially misleading.

Why might this be useful

Our model, if successful, could be used as a type of tracker for when COVID cases are going to increase rapidly, or when they are going to fall. A successful model could be used to predict future COVID surges and allow the government to prepare/react in a way that could prevent further COVID deaths.

Proposition of type and methods of modeling

We propose to use non-linear regression, maybe using a fourier transform, or some cyclic function like sine. We intend to look for other data like vaccines, mask use, and maybe even immunity from prior COVID cases to feed into our model. We might have to collect more data, because we are modelling things over time, and the existing data has columns that focus on non-time sensitive data. Data on

vaccination and mask-use is unfortunately focused on a single point in time, so we hope to collect more data. Luckily, the CDC provides open source data on vaccinations and potentially mask use. If mask use data is hard to come by, we might look into using data about the implementation and lifting of mask mandates which is much easier information to find. Additionally, the political leaning of a state can be taken into consideration, as we observed that political leaning and new COVID cases were correlated, which would be handy if we decided to model a set of states rather than all of the U.S.