CS 349

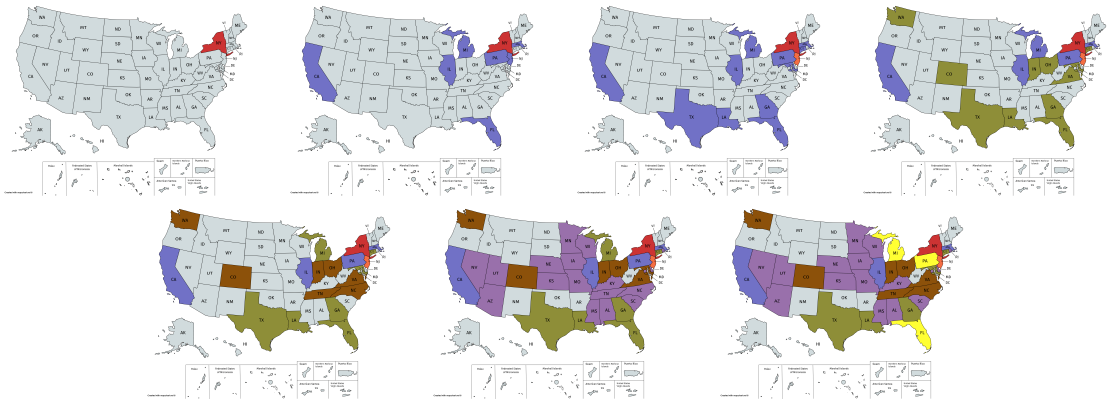
Take-home Final

Henry Raeder

Due on Wednesday 11:59pm, June 10th, 2020.

For my take-home final, I chose to do Option 1, which was working with the Covid-19 dataset from Johns Hopkins. In general, my plan was to work with the US data, as opposed to the global data, and try to find a pattern in how states' responses to the pandemic changed the overall spread of the disease in each of their populations. I decided to use K-Means clustering, because I wanted to find naturally occurring clusters in the data without applying any labels to them. Additionally, I didn't have any ground truth labels for training data, so K-Means seemed to be a reliable option to find clusters In general, what I expected to see was one of two things. The most obvious option would be that states were separated by geographic area, which would make sense based on the general idea of how a virus spreads. The other option I expected to see was based off of how long it took the virus to spread to a given place. It makes sense that the longer it took for the first case to be reported, the fewer overall cases there should be.

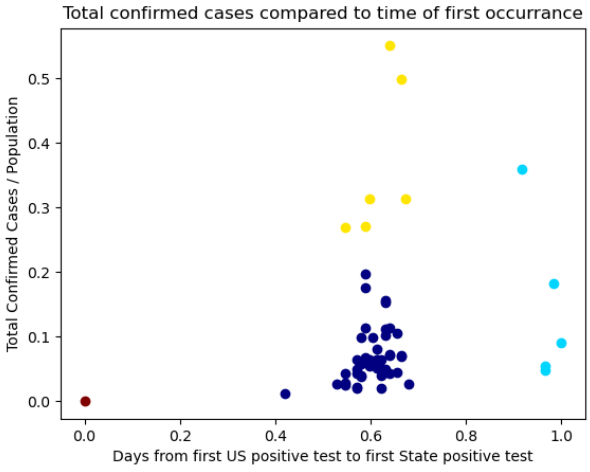
The first strategy I used was to take the overall confirmed case data for each state, and simply plug that aggregated data into a K-Means algorithm given a number of clusters in the range from 2 to 8. Essentially, each row in the features was every day of confirmed cases for a state/territory starting on January 22nd, 2020. From these clusters, I generated 7 maps, with each state colored according to its assigned cluster. These maps can be seen below. Note that, in these maps, gray counts as a cluster.

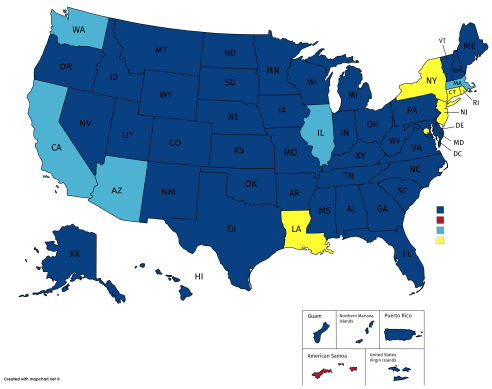


Now, it can clearly be seen from these graphs that there is no apparent pattern to the geography of the clusters. Generally, what would be considered harder-hit areas are grouped together, but which states are contained in which clusters change enough between clusters

that I do not believe anything of value could be gleaned from this. So, I decided to improve my approach.

I realized that there were some glaring errors in the way I looked at the data. For one, I did not explicitly account for how long it took Covid to reach each state. Additionally, I did not take account of the population of each state, which can obviously play a huge role in how a virus spreads. So, I made some quick changes to my feature set. Rather than just passing in the raw data, I broke each state down into two values. One was a relative percentage of the time in which Covid had been spreading in the state, and the other was the number of confirmed cases in that time as a percentage of the population of the state, as given by the deaths spreadsheet. For example, if the first case arrived in the US in Washington and 10% of its population had been confirmed positive, its feature set would be [1.0, 0.1]. I used the same set of clusters, but in this case I could explicitly graph the results because of the two-dimensional nature of the features for each state. I will not put all of my graphs below, but simply the n=4 graph, along with its map representation.





This data presents a much more obvious set of clusters, and the graph itself is very telling. It appears that there were 4 distinct clusters, although there was 1 cluster with only a single territory in it. There appear to be 3 main categories. These are states that got hit early and relatively hard, states that got hit after a while and saw a relatively low rate of infection, and states that got hit at the same time but saw much worse results.

In general, it appears that some of my hypotheses were true. There is some geographic causality, as the west and east coast were where the virus started its spread in the US, and that is where the worst cases are. One interesting thing we can also see is how well the clusters track with population density. Many of the non-dark-blue states have regions of

high population density (for example, Chicago in the lone light-blue state of Illinois, or New Orleans in Louisiana). However, the trend I was expecting of a relationship between time-of-infection and overall prevalence did not occur. Regardless of when a state became infected, there were some states with very little exposure and some states with a lot of exposure. Overall, the exposure of a state appears to be influenced more by population density and distance from the epicenters of the pandemic rather than the amount of time since exposure occurred.

Given more data, I would also like to expand on this study. The one major variable I could not control for was the prevalence of testing in any given state. For example, in some of the southern states where the testing rate is very low, there may be huge underreporting of confirmed cases, which would lead to a lower y-axis value. Additionally, I would like to control for known events that could lead to large amounts of infection. For example, Louisiana is strongly in the yellow cluster, despite being surrounded by dark blue states. This could be because of Mardi Gras, which happened right as the pandemic began to spread in the US. If there was more information regarding the specific geographic hotspots within states, it would be easier to assess why these differences exist, rather than just finding that they do.

A link to the Github repository containing my code can be found here: Link to Github repository. The code I wrote is contained in the exp folder, the file is named Henry_Experiment