Clustering US States on COVID-19 Restrictions

DATA607 Final Project

Rick Powell

December 2024

1 Introduction

In March 2020, the world suddenly changed. Around the start of 2020, news was going around about a new sickness that was rapidly spreading, but it wasn't until March 2020 that it truly landed in the United States. This was the start of the COVID-19 pandemic in the United States.

Through this paper, we will examine data covering the COVID-19 data and focus on how certain government regulations impacted population health in the community, and can be used to cluster our data.

Our data was sourced from Emanuele Guidotti and David Ardia's COVID-19 data hub published in the Journal of Open Source Software in 2020 (Guidotti and Ardia [2020]). The data was collected from a multiple different sources and aggregated and stored on GitHub. For the United States data, the list of sources can be found here: COVID-19 Data Hub. Below, I have included a table which goes over the variables that were used in this project, along with a description of each of the variables. For more information on the exact details for each variable, this link, will take you to the documentation which goes over all additional variables (inlcuding those not included in this paper.

For this paper, we will use the data for the 50 US States, and the data for the United States as a whole. For the United States, we will just focus on some of the large scale metrics in order to see the timeline of the pandemic.

Variable	Description
confirmed	Cumulative number of confirmed cases
deaths	Cumulative number of deaths
recovered	Cumulative number of patients released from hospitals or reported recovered
tests	Cumulative number of tests.
vaccines	Cumulative number of total doses administered.
population	Total population.
school_closing	Policy for whether schools should be closed
workplace_closing	Policy for whether workplaces should be closed
facial_coverings	Policy for whether facial coverings must be worn
stay_home_restrictions	Policy for whether individuals should stay at their homes

Table 1: Description of Variables

For the individual 50 states and Washington DC, we will look at the policies set in place by the state governments and how they changed over time. Having the timeline of the pandemic will give us some insight as to what was happening when as the states' policies changed.

2 Data Cleaning

The COVID-19 dataset is both large and messy, so our first step when working with the data was to clean it up. Our first cleaning step was to pare down our list of variables in order to make sorting through it easier. We did this by taking another subset of our dataframe. This time we excluded all of the columns that we were not planning on using for this project.

Next, we focused on cleaning up our categorical variables. In the initial data set, certain variables were listed as negative values. These negative values were used to identify policies that represent a best guess of the policy in force, but may not represent the real status of the given area. For this paper, we have decided to treat those negative values as representative of their areas, so we converted those values to positive using absolute value.

Finally, I removed the following territories from my data set: Northern Mariana Islands, Virgin Islands, Guam, American Samoa, and Puerto Rico. I chose to remove these territories because they frequently had missing values

and did not have the same levels of reporting as the rest of the United States.

3 Data Analysis

For our data analysis, I wanted to begin by looking at the overall statistics for the entire United States population. The first graph [1] is an overview of the cumulative rates of confirmed COVID cases, deaths caused by COVID-19, and the vaccine doses administered. It is important to see the overall impact of COVID on the entire United States as it allows us to see and understand the policy decisions that were made by each state over the following months and years.

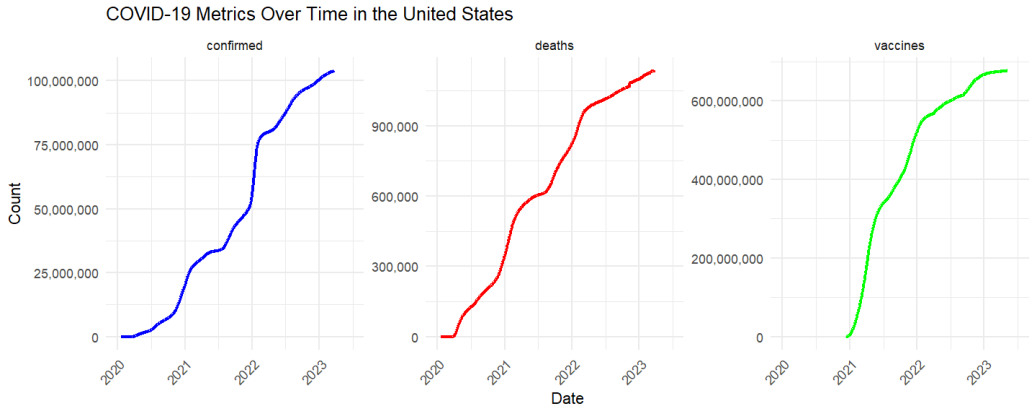

Figure 1: COVID-19 Com firmed Cases, Deaths, and Vaccines

The next group of graphs focus on the COVID-19 policies for each state and then how they evolved as the pandemic continued. For each graph, I have taken five snapshots, the first in July 2020, three months into the start of the pandemic and the most strict lockdowns. The next snapshot takes place six months later at the beginning of the new year in January 2021. Following that are July 2021, January 2022, and finally July 2022. I chose these snapshot because they allow a couple months to see the response to the initial outbreak, and then check-ins every six months until the data set used stopped showing results.

Our first graph in this set [2] looks at the Facial Covering requirements during the pandemic. COVID-19 is spread through airborne particle being breathed in and out and therefore masks were implemented as a cautious

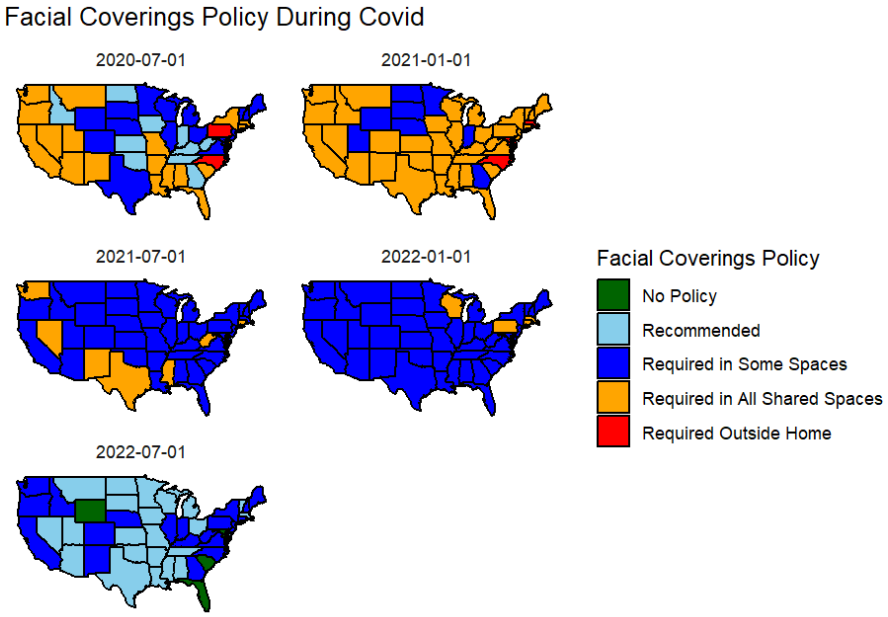

Figure 2: Face Covering Restrictions over COVID-19

method to reduce the spread. Although there was a lot of push back on the requirement to wear masks (mostly due to political grandstanding), we can see that many states very quickly adopted mask policies at the beginning of the pandemic. In the first 2 dates listed in this graph, we see a lot of states requiring wearing a mask, at least in some spaced. By the beginning of 2021, all states had some mask requirement. As the pandemic continued, we started to see the Orange requirements give way to more Blue (Required is Some spaces) and after 2 years of the pandemic, many states returned to a Recommended to No Mask policy.

Next, I wanted to look at a similar graph for School Closing requirements [3]. Similar to what we saw when we looked at the Facial Covering Requirements, Schools Closing Requirements were most strict immediately after the start of the pandemic. Every states, except for Illinois, required schools to be closed in at least some capacity, with most states having schools closed at all levels. What I found strange was, at the start of 2021, as most states were easing their restrictions from 6 months prior, Illinois actually moved to require closing at some level. By the end of the 2021 school year, many states were beginning to return to normal procedures with either no closing

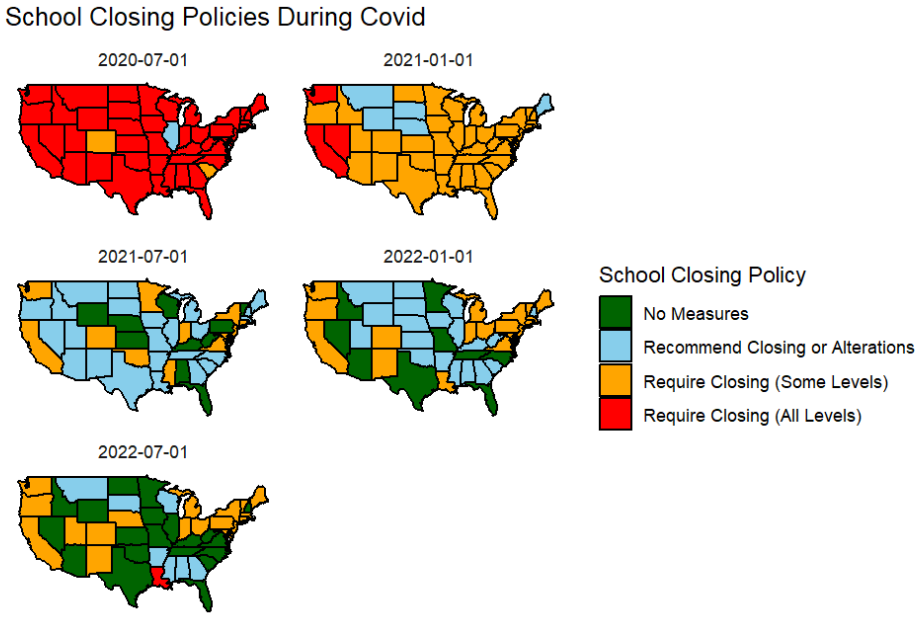

Figure 3: School Closing Restrictions over COVID-19

requirements or recommended closings or alterations. Another interesting point was seeing Louisiana turn to requiring all levels of school closing when most of the other states were beginning to relax their restrictions. This might have been caused by an additional COVID wave, which was common during the pandemic.

The final graph I wanted to look at in this set was concerning Workplace closing policies over the same time frame as the other two [4]. During the COVID pandemic, policies ranged from No Measures to Requiring All but the Essential Businesses close. My expectation for this section was to see graphs similar to the school closing graphs and for the most part that seemed to be the case. Both started with the most strict positions near the start of the pandemic. The largest differences between the two was that the most strict positions for workplace closings was that there were much fewer required closings at the beginning of the pandemic. Most of the initial closings were recommended or required some sectors to close. Very few required all non-essential businesses close right at the beginning of the pandemic. A second difference between the two graphs was that workplaces returned to normal policy (No measures) much faster than schools.

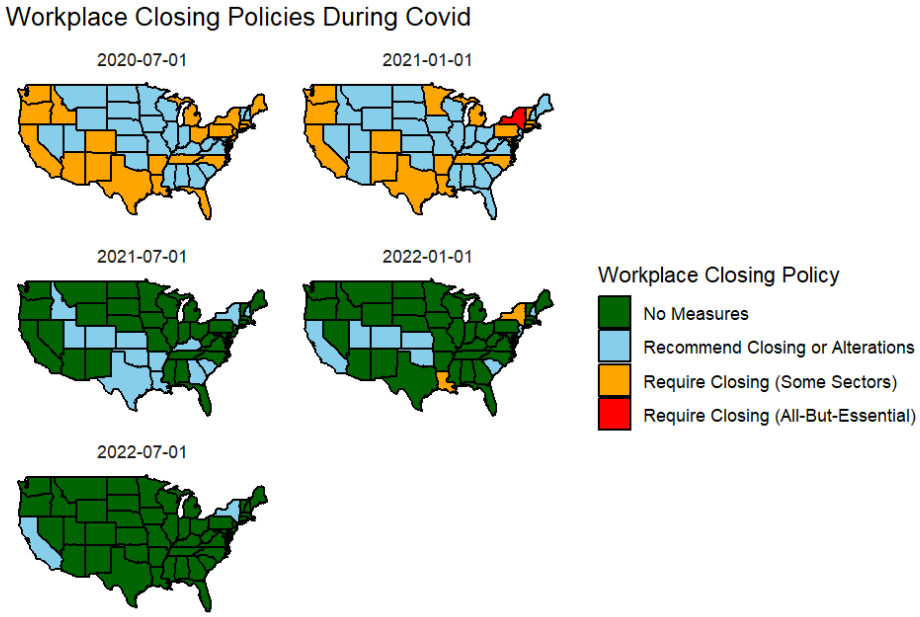

Figure 4: Work Closing Restrictions over COVID-19

4 Clustering

In order to cluster the US States, I wanted to look at each states' COVID-19 policies for the three topics we examined before (facial coverings, school closing, and workplace closing), and each state's Stay at Home policy. I took the combination of those four policies at each of the five timestamps above (at the beginning of July and the beginning of January for each year between March 2020, and December 2022) and used that combination to cluster our states. This clustering gives us an overview on how each state reacted to the initial wave of COVID, how their policies changed as the pandemic continued, and which states they were similar to.

Originally, when I was deciding on how I wanted to cluster my data, I originally wanted to use a heatmap. Having presented on heatmaps for my group presentation, I thought it would be a good way to use some of that knowledge for a different project. However, I did not like how my heatmaps were looking and I wanted to add more features. So I scrapped my idea for the heatmap, but I wanted to investigate the clustering further. Using hierarchical clustering was a good way to see how the states are clustered

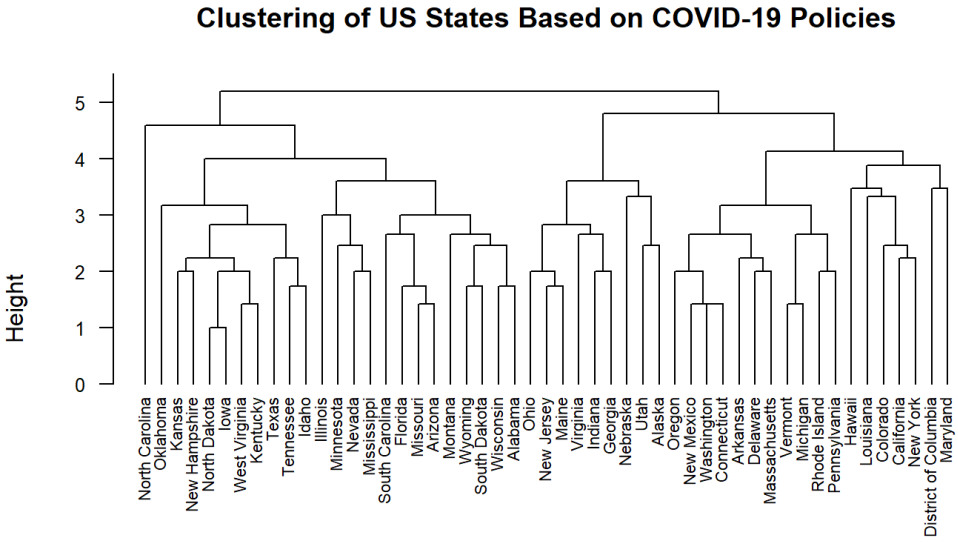

Figure 5: Clustering of US States

based on the similarities with the other states. Above is the hierarchical clustering of the US States.

Now that we had a hierarchical clustering, we needed to determine how many clusters was appropriate. Looking at our dendrogram, we could easily claim that there are multiple ways we could cluster the data and present a strong argument. In order to solve this issue, we will look at a silhouette graph. Normally, I am a fan of the elbow method to determine the number of clusters, but in this instance where k is equal to 2, I think the silhouette method looked cleaner. Above is our graph [6] showing the Silhouette method to determine the optimal number of clusters we should be using for our Clustering. In this graph, we see the largest value occurs when k is equal to 2, which is highlighted in red. Therefore, we plan to use 2 clusters to divide up the states.

To address possible uncertainty in our clustering, we can look at the silhouette score for our chosen k. We can see that for our optimal k equals 2, our silhouette score is .20, which is relatively low. This suggests that our clusters may not be well-separated or that some of our data points might be poorly assigned to clusters. We would have liked to see a higher silhouette

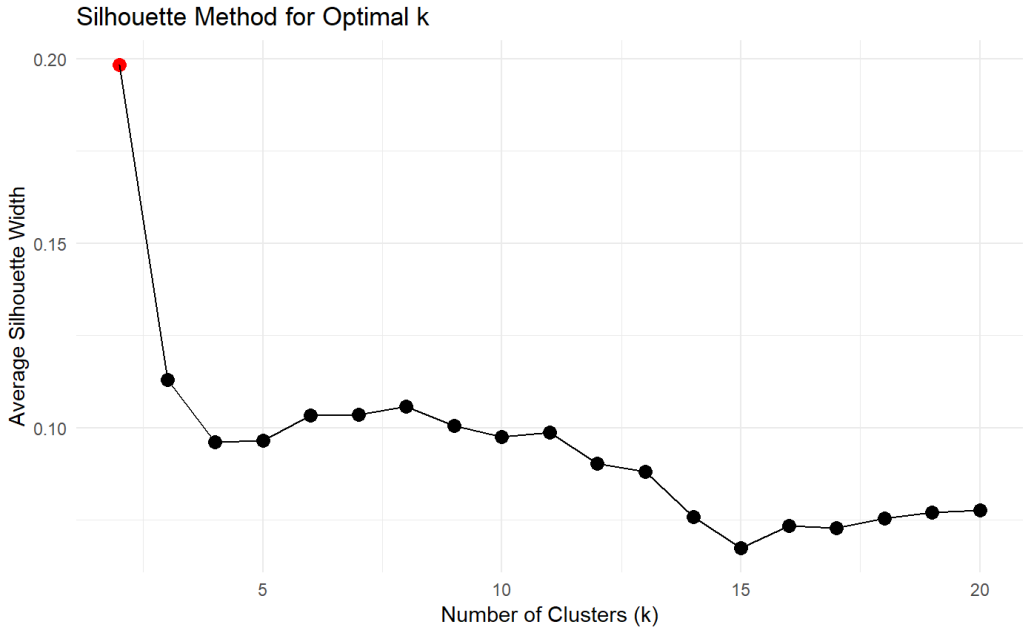

Figure 6: Silhouette Method showing Optimal k = 2

score (closer to 1), which would indicate more distinct and cohesive clusters.

Finally, we have our map [7] of the US divided into our twos clusters. Although they do not appear in the graph, both Alaska and Hawaii fell into Cluster 2. This clustering takes our dendrogram from above and splits it into the optimal two clusters that we found using our silhouette method. We can see that we have blue pockets along the New England and Northeast areas, along with the West Coast, while we have red pockets dominating the Southeast and Midwest areas.

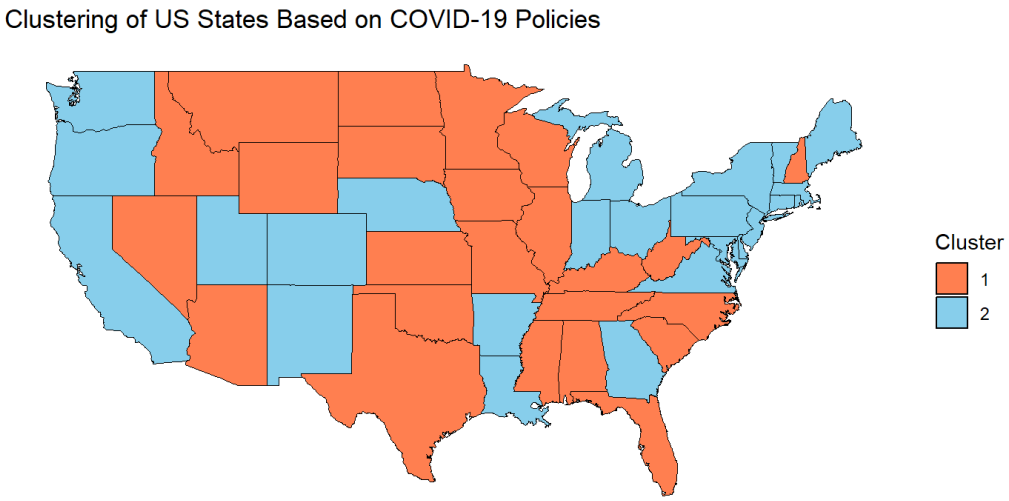

Figure 7: US States Split into 2 Clusters per COVID-19 Policies

5 Conclusion

Using the Policies implemented by their state governments, we were able to split up the United States into the two clusters shown in Figure 7. An interesting thing to note when looking at our clusters, they almost appear to be split on political lines, where a majority of the left leaning states are blue, and the right leaning states ended up being red. I have added Figure 8 above which shows the 2020 election results in order to compare the differences between our clustering and the election results. There are difference between the expected political lines and COVID-19 policy lines, but it was fascinating to see how looking at a couple of healthcare policies and how they changed over a two and a half year period was enough information to split the data into these clusters.

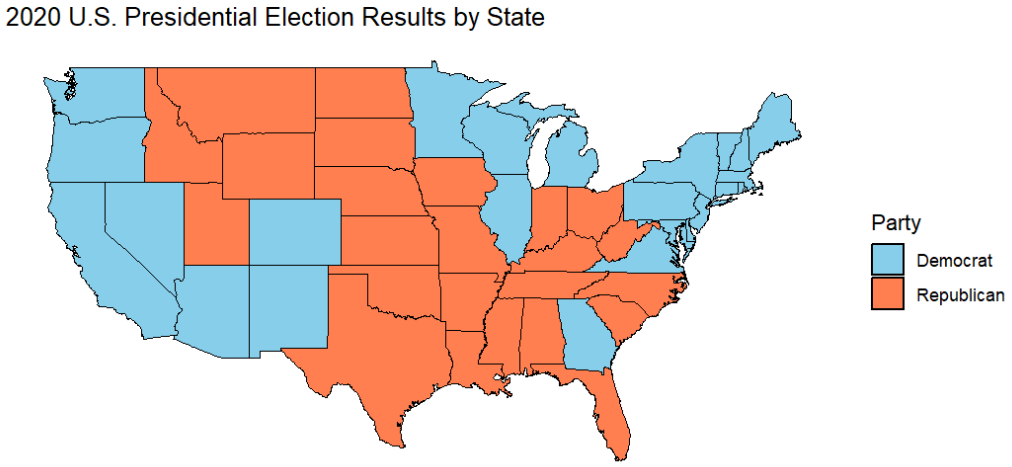

Figure 8: US States split up by 2020 Election Results

References

Emanuele Guidotti and David Ardia. Covid-19 data hub. *Journal of Open Source Software*, 5(51):2376, 2020. doi: 10.21105/joss.02376.