**Results and Discussion**

For our results we received a 60% accuracy rate through the means of using train and test data for our k-means.

In figure 1 we can see a silhouette chart of our test data after it has been processed with k-means. Based on the chart here, we can see harsh negative values for many of our clusters which indicates that k-means believes that the data points are so like one another that nine clusters may not have been the best choice for k despite there only being nine regions.

Chart

Description automatically generated

Figure : Silhouette Values of Clusters

Our work suffered from a few limitations from our data. The data had inconsistencies, with some data points having an increasing and decreasing infection rate even within the same regions and divisions. These limitations seem to be caused by the data being measured at different points of outbreak across the United States. A certain area may have already seen the brunt of the pandemic and is experiencing a decline in infections, while other areas in the same region may have a ramping up infection rate due to Covid just reaching them. This alone makes it very tricky to determine the region of an area based on infection rate alone

**Conclusion**

We discovered through this case study that correctly sorting data into clusters is not as simple as running a basic k-means algorithm. Most data sets, especially ones like this Covid set, require a significant pre-processing period before it can be properly sorted and assigned clusters. Through our toil we managed to receive a 60% accuracy rate. With this accuracy rate, our shortcomings become apparent, but this may reflect the overall difficulty for health officials to process the data in the real world to help combat the pandemic.

We believe we can expand upon our findings by processing additional data, and rather than trying to determine each region, we could use k-means to determine the most at-risk locations for disease breakouts.