Graph Clustering with Applications on Covid-19 Growth Data Across the United States

Joseph T. Fernandez Mentor: Dr. Enyue Lu
Salisbury University Department of Computer Science

Abstract

Covid-19 has killed hundreds of thousands and has infected more than 30 million people in the U.S. alone. The approach used in this project is to group counties with similar growth trends together. Grouping similar counties can be done using graph clustering to analyze Covid-19 growth data across all counties in the U.S. K-means clustering was used for the clustering. The K-means clustering algorithm works by finding nodes in a graph that are close to each other. It puts a centroid in the data and assigns the nodes to the nearest centroid, forming clusters. The centroids are moved to be in the center of data. Clustering helps take large amounts of data that would otherwise be challenging to analyze and make it simpler. Being able to analyze this data and find similarities and possible relations between them could help predict future trends across counties.

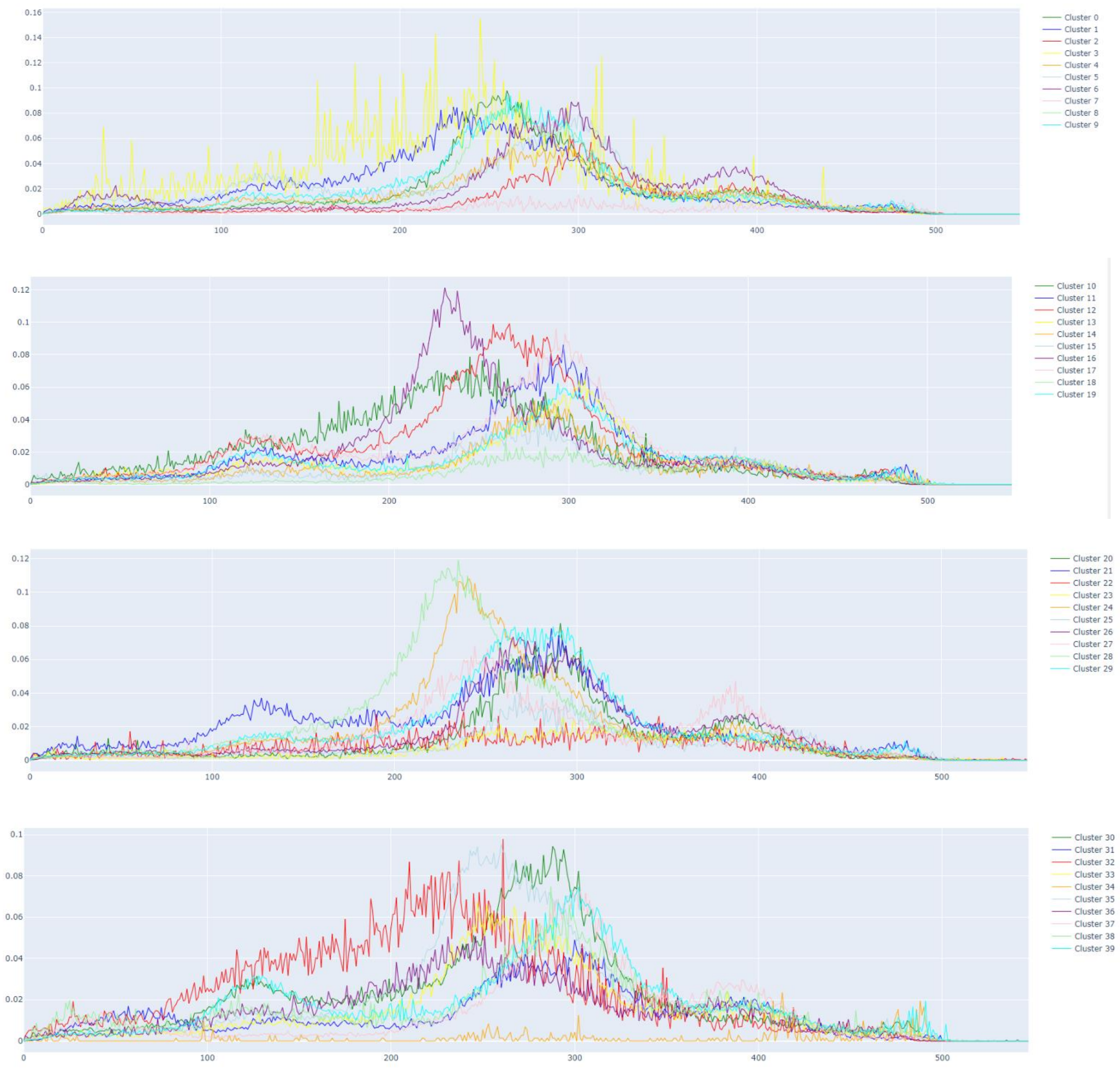
Methods

- Data from the New York Times covid-19 dataset is taken into a java program and sorted into counties by date.
- This data is then matched with the population data from the U.S. Census.
- The similarity measures between counties are calculated using a combination of the Euclidean distance formula and the Manhattan distance formula.

$$\sqrt{\sum C_i/P_i - C_j/P_j} * \sum |C_i/P_i - C_j/P_j|$$

K-Means Algorithm

- Creating a given numb.er of centroids with random coordinates
- The distance is taken between nodes (counties) cluster by cluster and the centroids.
- The centroids are moved in comparison to the other nodes on the graph
- Nodes are then assigned to their closest centroid.
- This process repeats until the centroids position in the graph does not change.



Figures 1-4. These graphs are the average growth trends across the counties in each cluster. Each graph has the growth trends from 10 of the clusters

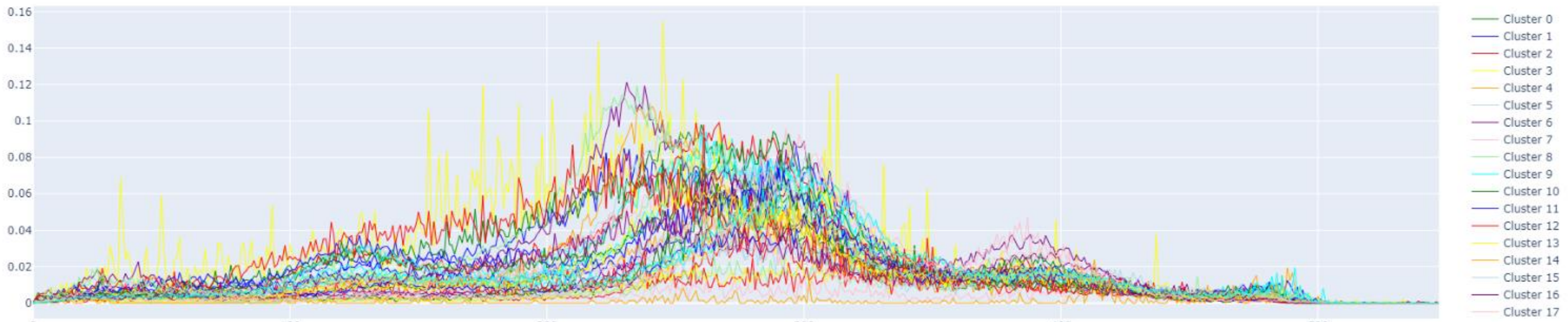


Figure 5. All average trends across all clusters

Figures 1-5. In the above graphs, the X-axis is the number of days starting from the first case in each county. The Y-axis is the number of new cases for that day over the population of the county. Each curve is the average of all of the counties in the cluster.

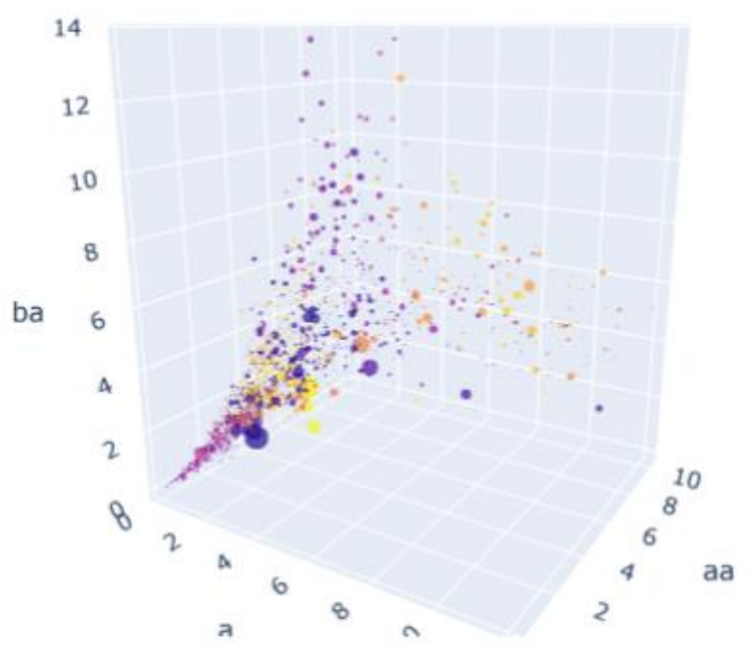


Figure 6. The clusters visualized in three dimensions. The colors signify the cluster and the size of the node signifies the population of the county. This graph only shows three dimensions(shown by the letters on each of the axes of the 3135 that were used in the K-Means algorithm.

Results

- Through testing, it was found that 40 clusters was the optimal number to show the difference in trends.
- In many of the clusters, it was found that geographic location was the largest determinant in the similarity of trends, with many clusters having many counties in them, but only two or three states
- Population was also a large factor in the similarity in growth trends between counties, however it was not nearly as large of a signifier as geographical location was.

Future Work

- Test different clustering algorithms, like Markov Clustering or K-Truss to compare results.
- Explore different testing algorithms that could compare differently sized counties better.

Acknowledgments

The work is funded by NSF CCF-1757017 under Research Experiences for Undergraduates Program.