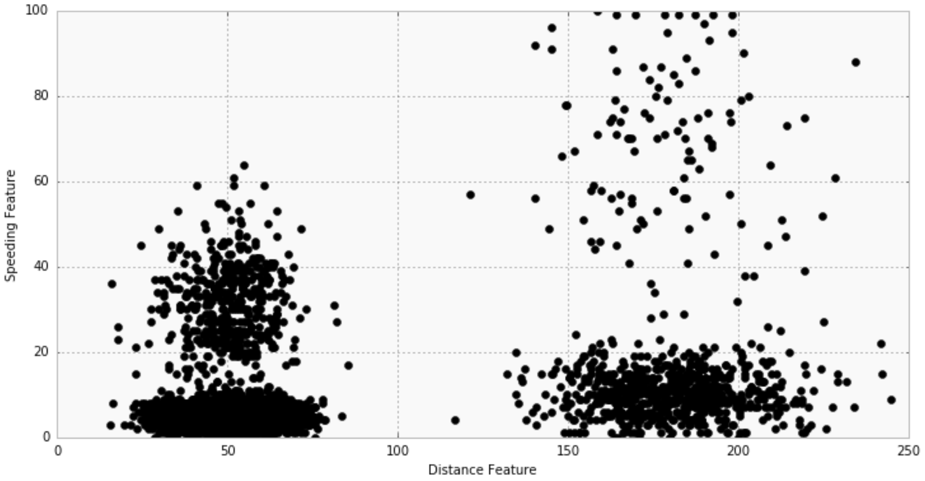
Ozan Gokdemir
Rain Kwan
Thomas Kane

Clustering Delivery Fleet Data Using K-Means, K-Medoids and Hierarchical Clustering

We decided to perform cluster analysis on a delivery fleet dataset. This public dataset is comprised of four thousand rows and two columns, thus, our data is two-dimensional. Features in the data are the distance(in miles) of the delivery location and the average speed(in mph) during the delivery. We used the pandas data analysis library to clean and process the data into a pandas dataframe that we can work with in Python. All three algorithms are implemented in the Python programming language.

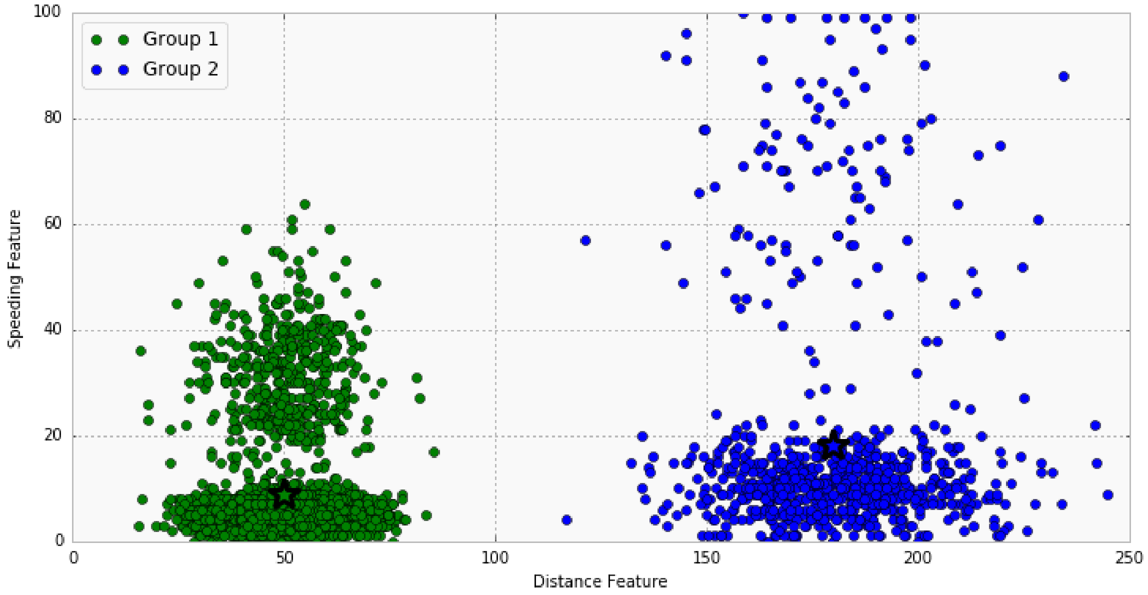


A plot of our raw data.

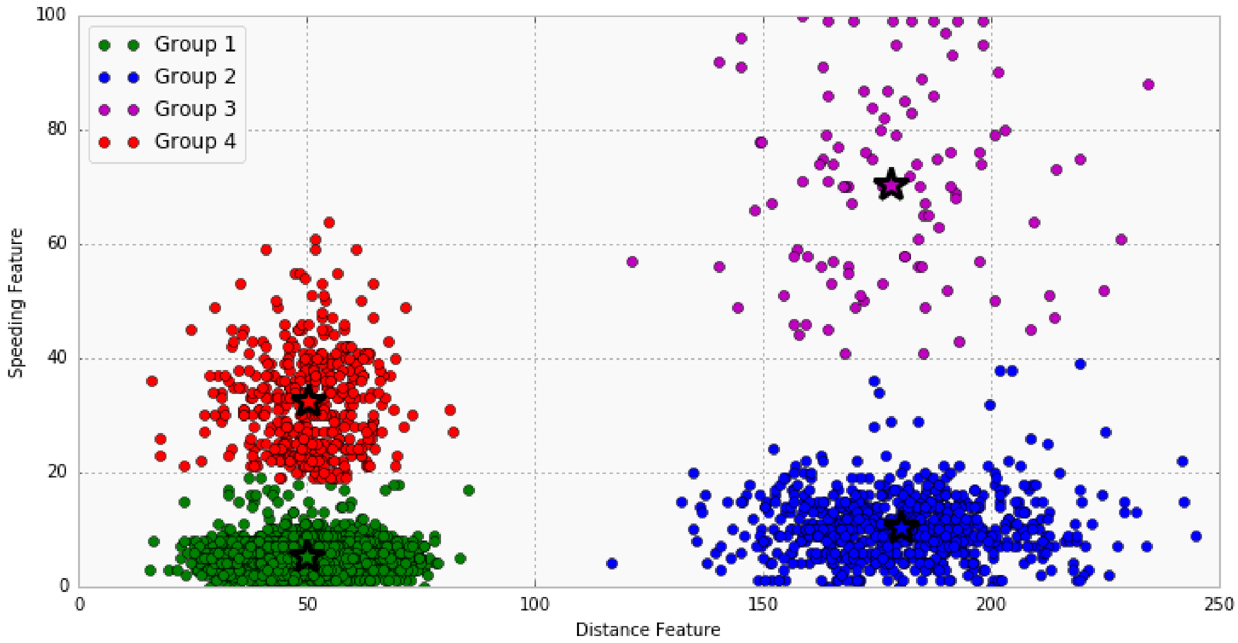
In order to reach the optimum cluster configuration in our K-Means implementation, we used inertia. Inertia evaluates the amount of in-cluster spread of data from the centroid point. We wrote a function that loops through all the currently computed clusters and calculates their inertia. Then, the sum of these inertia values is stored as a variable. Our intuition was that, if this sum-of-cluster-inertias value remains the same in two consecutive steps, then no more data points have shifted between clusters or moved and we reached the final state of the clusters.

In our implementation of the K-Medoids algorithm, our measure of distance between two data points was the Euclidean Distance. This decision was justified by the fact that our data is two-dimensional (e.g: contains two columns). We ran a do-while loop in which we computed the clusters starting with an initial set of medoids. We then repetitively computed new medoids and clusters based off of the previous clusters made. We broke out of the loop at the point where none of the medoids changed. This implied that we had finally reached to a stabilized set of clusters. Since we did not have to use the inertia criteria in K-Medoids, we observed that the runtime for our K-Medoids implementation was shorter than that of our K-Median implementation. Furthermore, we noticed the difference between the methods of finding the centroid in K-Means and K-Medoids. We believe that implementing this step facilitated our permanent understanding of the difference between these two algorithms.

The Hierarchical Clustering algorithm was the hardest one to implement and visualize. Upon realizing the necessity of utilizing a tree data structure in order to keep track of the hierarchy of clusters, we imported the implementation of a tree data structure in Python from an online source. We clearly indicated this source and where it was used within our code documentation. We realized that, by nature, Hierarchical Clustering would merge clusters until there is only one cluster containing all the data. Therefore, merging must be stopped at a certain point based on criterium like diameter, radius or density of the clusters. Unfortunately, due to type-related syntactical issues, we were not able to accomplish this step. Since we already knew that there were four apparent clusters in our data, we stopped the merging process when had reached to four clusters.



A plot of K-Means clustering of the data when k = 2



A plot of K-Medoids clustering of the data when k = 4

Our analysis of the final clusters in K-Means indicates that the data can be grouped into two clusters based on the feature of distance. This implies that we can identify these two groups as urban(on the left) and rural(on the right) deliveries(see plot above). Based on the density of the urban deliveries cluster, we can claim that most of the delivery orders are for urban areas which makes sense intuitively.

When we further investigated the data by creating four clusters with K-Medoids we found that urban and rural drivers are also separated into speeding drivers(on the top) and the drivers who follow the speed limit(in the bottom). Based on the plot above, we inferred that urban drivers have a lower speed threshold and they tend to follow the speed limit more than the rural drivers. We believe this results from more abundant camera control, police presence, traffic jams, traffic signs and intersections in the urban driveways than those in the highways and rural roads.