Dr. Trajanov

CS 777

Aaron T.

6/17/2025

Taxi Hub Pickups

For this project I will explore how K-means clustering could be beneficial to taxi drivers around New York City by creating hubs around the pickup points with the largest trip distance. I believe this could lead to maximizing their profit per ride. I will be using the taxi pickup longitudes and latitudes, along with total ride cost (including tip) as the features for my model from the New York City Taxi trip reports from 2013. This should produce areas around the city where taxi drivers on average have the highest probability of making the most money on a given ride.

I began with looking over the data from the *NYC Taxi trip reports from 2013* to determine criteria for a valid ride for modeling (Fare, Trip Distance, Total Amount, Pickup Longitude, Pickup Latitude, Tip). Making sure that the fare, trip distance, tip, and total amount where all positive numbers greater than 0 since a trip can't go backwards or a driver losing money after providing a service. Then after a quick google search about New York's coordinates I set up parameters for longitudes (74.0060° W) and latitudes (40.7128° N) ending up on (-74.5 <= longitude <= -73.5) and (40.4 <= latitude <= 41.0) around these coordinates.

After cleaning and filtering the data, it was now time to start a spark program to begin analyzing the dataset. I used the pyspark.ml library to create a K-means model using the features: Pickup Longitude and Pickup Latitude to train the model and then return the prediction cluster. After manually testing some K-values for the number of clusters to create I ended up using 19 since that was where the model performed the best in terms of the highest *Silhouette Score* of 0.5707 and the lowest *Within Set Sum of Squared Errors* of 118.08. A silhouette score of 0.5707 explains that the points in the clusters are reasonably placed. When setting the K-value to 4 the model achieves its highest silhouette score of 0.6839 but the *Within Set Sum of Squared Errors* rises to 683.87 explaining a greater error or distance between a point to the group (cluster) average.

The result of the model shows that Cluster 2 with Longitude=-73.785479, and Latitude=40.646630 is the prime location for Taxi drivers to converge to get the furthest rides (15.47 miles) which in turn leads to the highest fare price ($52.47). I would like to improve on this model by creating a function to calculate the optimal K-value. There is also room to continue explore by adding in drop-off coordinates to then see where the taxi driver should go next based on that drop-off location.