Starting a Wholesale Coffee Supplier Business in Delhi, India

Kanishk Kumar

February 28, 2021

1. Data

1.1 Data Requirements and Collection

Most of the neighborhood data can be found on Wikipedia. This Wikipedia page will be scraped for Borough and Neighborhood details of Delhi. Coordinates of the neighborhoods from the scraped data will be obtained using Geopy. The number of venues, their type and location in every neighborhood will be obtained using Foursquare API.

1.2 Data Cleaning and Feature Extraction

The first data is a Wikipedia page about the neighborhoods of Delhi. We will scrape the page and create a data frame consisting of two columns; Borough, and Neighborhood. We remove any rows that do not have borough assigned. Then, we will be using the Geocoder python package to retrieve the neighborhood's coordinates. It will return 108 rows and 4 columns.

The second data is stored inside Foursquare Location Data, and we will use Foursquare API to access it. We utilize the neighborhood names to retrieve popular venues around a specific radius. As a result, the same venue categories will be returned to different neighborhoods. We can use this idea to cluster the neighborhoods based on their venues representing services and amenities.

We will run the k-Means algorithm to perform this clustering with different number of clusters (k). The features will be the mean of the frequency of occurrence of each venue category. Finally, we can visualize the cluster model using the Folium module.

To sum up, we will use the 1st data to obtain the exact coordinates for each neighborhood based on their names, allowing us to explore and map the city. We will then use the coordinates and Foursquare credentials to access the 2nd data source through its API and retrieve the popular venues along with their details, especially for coffee shops.