The Battle of Neighborhoods Report

2. Data

2.1 Data sources

Dataset containing geographical coordinates of New York neighborhoods would be retrieved from scrapping web https://geo.nyu.edu/catalog/nyu_2451_34572. This data would be transformed into Pandas dataframe and sliced it to contain only Manhattan borough of New York.

Foursquare API would be leveraged to get the venues using the search criteria 'Italian" for the Manhattan latitude and longitude and transformed it into Pandas dataframe. The data returned by API will be examined to check if it contains venues other than Italian restaurant and only cleaned dataset after removing those venues would be used for exploring venues

Afterwards, Foursquare API would be used again to retrieve venue details relating to 'likes' 'count' for each venue and added this data point as 'likes count' column to venues dataset. The 'likes count' would be used for determining the top 3 restaurant choices and offering preferred choice to solve the problem

Foursquare API would be used again to retrieve venue details relating to 'rating', though rating would not be available for all venues to verify the venues ranked high based on likes count would also be rated high.

Recommendation of top three venues choices would be based on twin criteria: high likes count and high rating to solve the business problem

2.2 Methodology

Use pandas libray to hold data, clean, explore and analyze dataset

Use numpy libray to handle data in a vectorized manner

Used geopy, folium libraries to create maps of New York and Manhattan with neighborhoods and restaurant venues superimposed on top

Use Foursquare API to search for venues categorized as Italian Restaurant and explore venue details relating to 'likes' count

Run unsupervised learning K-means algorithm to cluster the venues
Use Foursquare API to get venue rating, though rating is not available for all the venues