Clustering New York City Neighborhoods

IBM Data Science Capstone Project

Stephen Ingraham

Introduction

- Resident of Manhattan wants to relocate to another borough of New York City
- He wants to find a neighborhood with a similar mix of amenities to his current neighborhood
- We can apply data analysis and machine learning to solve this problem

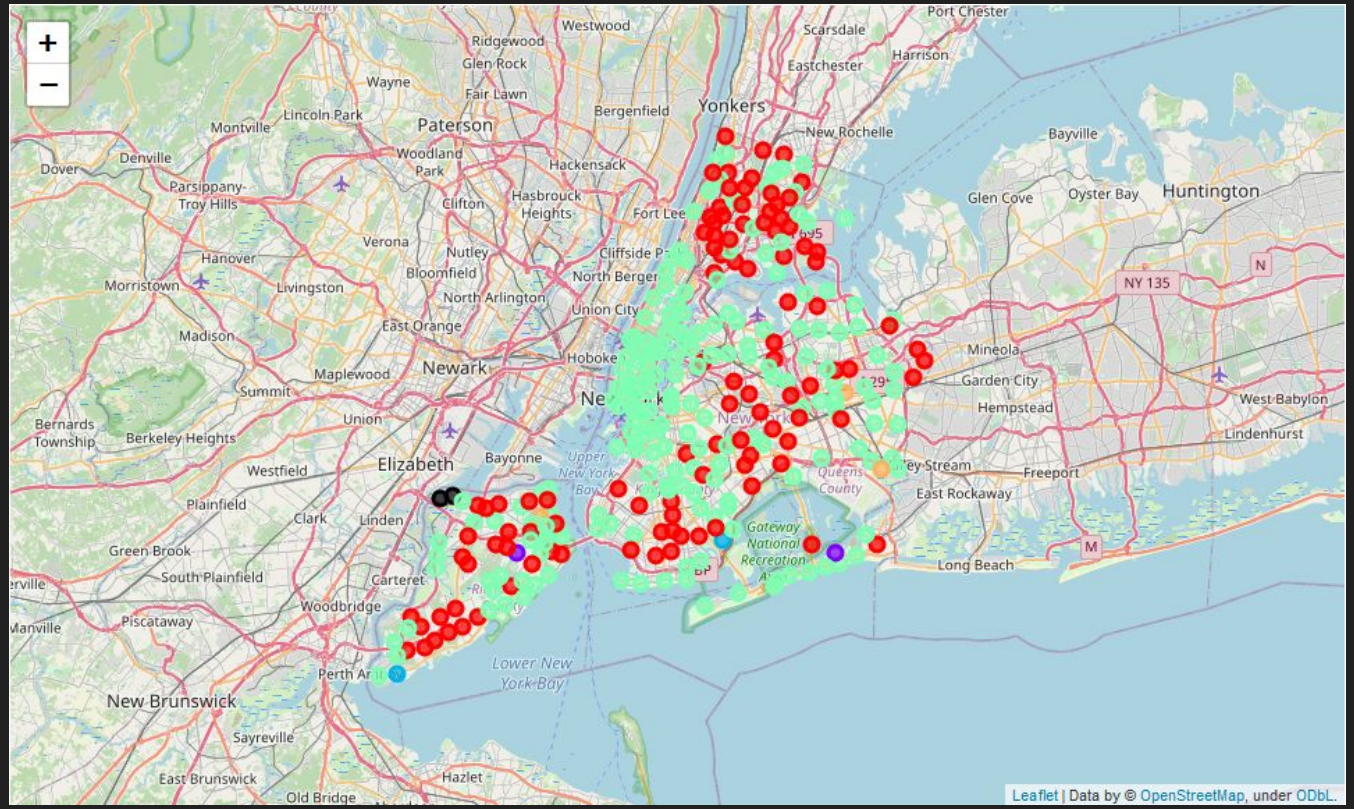
Data

- Two data sources:
  1. NYU neighborhood data: contains names and latitude/longitude coordinates for each neighborhood in New York City
  2. FourSquare location data: use the FourSquare API ‘explore’ endpoint to access data about venues in each NYC neighborhood
- Process this data and organize it into Pandas dataframes with Python scripts

Methodology

- Set up a Pandas dataframe where each row is a neighborhood, and the columns are each of the venue categories
- Dummy variables are used to indicate the presence or absence of venues from each category within a 500 meter radius of the neighborhood's latitude/longitude coordinates
- Perform k-means clustering on this data
- Determine what the top venue categories are in each neighborhood, and analyze how this relates to the clusters
- Generate a map of New York City, with each neighborhood labeled with a dot which corresponds to its cluster assignment

Results



Results

- **Cluster 0** (red): primarily have restaurants, food stores (pizza places, bakeries, etc.)
- **Cluster 1** (purple): two neighborhoods with an identical top venue profile. Parks, yoga studios, and flower shops are among the top venues.
- **Cluster 2** (blue): top venues such as pools, farms, and fields
- **Cluster 3** (cyan): primarily have restaurants and food stores. Also have a more diverse set of venues than Cluster 0, including parks and other non-food-related amenities
- **Cluster 4** (orange): primarily feature dog runs, farms, fields, and restaurants

Discussion

- Clusters 1, 2, and 4 contain only two or three neighborhoods each
- Clusters 0 and 3 are have quite similar venue profiles, but cluster 3 neighborhoods tend to have a more diverse range of amenities, such as parks and gyms
- Further analysis options:
  - k-means can be applied repeatedly to find optimal k-value, and with different initial conditions for the cluster centroids
  - Venue categories can be aggregated into supercategories (e.g. combine all restaurant types into a single restaurant category) to compare the relative density of amenity types in each neighborhood. This may yield quite different clusters.

Conclusion

- A Manhattan resident who wishes to relocate to another NYC borough with a similar mix of amenities has many options
- All Manhattan neighborhoods are in Cluster 3. These neighborhoods are also abundant in Staten Island, Brooklyn, and Queens.
- He may want to avoid the Bronx, where these kinds of neighborhoods are less common.