Clustering New York City Neighborhoods

IBM Data Science Capstone Project

Stephen Ingraham

Introduction

A resident of a Manhattan neighborhood enjoys his local amenities, but he feels that apartment rents in the area are getting too high. He decides to relocate to a more affordable neighborhood in New York City. However, he wants to make sure that his new neighborhood has a similar mix of amenities. How can he identify neighborhoods in other boroughs of NYC which have similar amenities to his neighborhood in Manhattan? I will apply data analysis and machine learning techniques to answer this question.

Data

In order to solve this problem, I need a dataset which contains information about the various venues and amenities in each neighborhood of New York City.

The first step is to use a dataset from NYU which contains information about each neighborhood in New York City. I can use this data to identify the names and locations (latitude and longitude coordinates) of each neighborhood. Here is a link to the data:
https://geo.nyu.edu/catalog/nyu_2451_34572

The second step is to use FourSquare location data to access venue information for each neighborhood. I can use the FourSquare API functions to explore the venues around the latitude/longitude coordiantes of each NYC neighborhood in the NYU dataset.

Methodology

Processing the Data

I converted the NYU neighborhood data from JSON format to a Pandas dataframe with the relevant columns, including the name of the neighborhood, its borough, and its latitude/longitude coordinates.

Next, I used the FourSquare API ‘explore’ endpoint to access data about venues in New York City. I passed the latitude/longitude coordinates of each neighborhood and a radius (500 m) to the API in order to generate a JSON file with a list of the top 100 venues within this radius. I compiled these lists of venues into a Pandas dataframe. Then I used dummy variables to encode the ‘venue category’ feature of each venue as a number.

I then defined a function to find the most common venue categories in each neighborhood, and I put this data into a new dataframe.

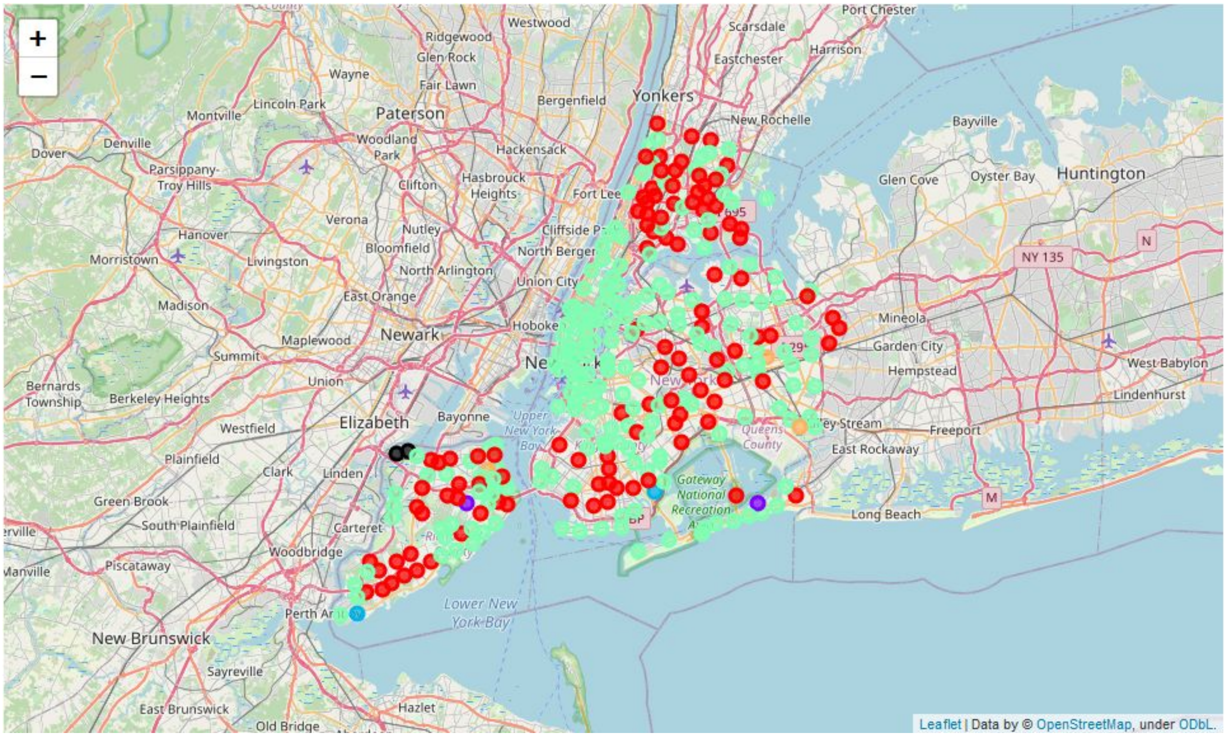
k-Means Cluster Analysis

Once I had the dataframes set up, I applied a well-known machine learning algorithm called k-means. For this algorithm, I defined a distance function between each data point. In this case, the distance function was defined by the number of top venue categories two neighborhoods have in common. Next, I selected the number of clusters (k=5) that I wanted to segment the neighborhoods into. I randomly selected initial points (centroids) for the k clusters. The algorithm then calculated which centroid each data point is closest to, and assigned the point to that centroid's cluster. Each centroid was then moved to the center of the points in its cluster. The algorithm applied this process iteratively until the cluster assignments reached a local optimum solution. At this stage, the intracluster distance should have been minimized, and the intercluster distance should have been maximized.

The Python library, scikit learn, has an implementation of the k-means algorithm. I used it to cluster all New York City neighborhoods according to the similarity in their top venue categories. Then I created an updated Pandas dataframe that included the cluster assignments (along with the top ten venue categories).

Results

I used the Folium library in Python to display a map of New York City where each neighborhood is labelled with a colored dot which corresponds to its cluster assignment. Cluster 0 is red, cluster 1 is purple, cluster 2 is blue, cluster 3 is cyan, and cluster 4 is orange. In addition, two neighborhoods in Staten Island did not have adequate data available (colored black). The map is shown on the next page. An interactive version of the map is available in the Jupyter notebook for this project in the same repository as this report.



Evidently, all of the neighborhoods in Manhattan fall into Cluster 3 (cyan). The Bronx appears to have a greater proportion of Cluster 0 (red) neighborhoods, while Staten Island, Brooklyn, and Queens have a mix of Cluster 0 (red) and Cluster 3 (cyan) neighborhoods. The other three clusters have only two or three neighborhoods each, suggesting that these are outliers in some way. Also, two of the neighborhoods on Staten Island were apparently not assigned to clusters due to an error in the data (they are colored black).

It would be interesting to know what these clusters are telling us about the feature of interest--the top venues in a neighborhood. When I analyzed the top venue information for the neighborhoods in each cluster, I identified the following patterns:

Cluster 0 (red): primarily have restaurants and food stores (e.g. pizza places, bakeries, etc.)

Cluster 1 (purple): each of the two neighborhoods has an identical top venue profile, with parks, yoga studios, and flower shops among the top venues.

Cluster 2 (blue): top venues such as pools, farms, and fields.

Cluster 3 (cyan): tend to have restaurants and food stores among the top venues. However, neighborhoods in this cluster appear to have a more diverse set of venues, including parks, public transit options, and other non-food-related amenities.

Cluster 4 (orange): primarily feature dog runs, farms, fields, and restaurants.

Discussion

The differences between Cluster 3 and Cluster 0 neighborhoods are not immediately clear, however, it does appear the Cluster 3 neighborhoods are more diverse. In particular , the Cluster 3 neighborhoods appear to have better access to amenities such as parks and gyms.

Clusters 1, 2, and 4 only contain two or three neighborhoods each. They appear to be distinguished from the primary clusters (0 and 3) by the relative prevalence of outdoor venues such as fields and parks.

Ideally, the k-means algorithm should be applied repeatedly with different k values in order to determine which k maximizes the intercluster distance and minimizes the intracluster distance. Also, we may find that even with the same k-value, other local optimum cluster groupings may result. In fact, this seems like a definite possibility because the top venue categories for New York City neighborhoods are not significantly different. It may be difficult to identify any number of clusters with meaningful differences within the same city.

One approach might be to group similar categories (e.g. Italian restaurants, Greek restaurants, etc.) into supercategories (e.g. restaurants) in order to capture information about the relative density of venues in these supercategories. For example, It might be more useful to compare the density of restaurants between various New York City neighborhoods than it would be to compare the prevalence of Italian restaurants.

Conclusion

A resident of Manhattan who wishes to relocate to another New York City neighborhood with a similar mix of venues should consider moving to one of the Cluster 3 neighborhoods. Such neighborhoods are abundant, especially in Staten Island, Brooklyn, and Queens (see map above).