Coursera Capstone Report

Italian Restaurants of New York City

By: Jagat Kohli

Date: 6/22/2020

Introduction:

You are a young entrepreneur with a knack for food. You have always wanted to start a niche Italian restaurant, but do not know where start the business. You have experience living in New York City, so you want to narrow it down to the neighborhoods in that city. NYC is a large city with a large market and opportunity. In this project, I explore the different factors that can affect the decision for where in New York City would be ideal to start a restaurant.

Business Problem:

The objective of this project is to gather and analyze restaurant data from New York City. By using data analysis and machine learning tools, we want to find out where in the city would provide the best opportunity to open an Italian restaurant.

Target Audience:

This project would be relevant to Italian restaurant owners who are looking to expand their franchise and entrepreneurs looking to start an Italian restaurant. With a city as big as New York City, there is a surplus of restaurants with numerous styles of cuisines, so knowing which neighborhood would lead to the most traffic for your restaurant is ideal.

Data:

For this business problem we will need to gather the following data:

- A data frame of the neighborhoods of New York City as well as their respective latitude and longitude coordinates. This data will be useful to find patterns and similarities between the neighborhoods.

- Data of restaurants within the New York City area to give a sense of where the competition is located.

Data Extraction Methods:

To obtain the data for the neighborhoods of New York City, I will use the URL of the dataset provided in Week 3 lab and clean the data frame to get the list of neighborhoods and their respective latitude and longitude coordinates.

To get the restaurant data, I will use the Foursquare API to properly view other Italian restaurants. I will use this by creating a Foursquare account and obtaining a Client ID and Client Secret number to call the relevant data.

For example:
https://foursquare.com/explore?mode=url&ne=40.713435%2C-73.89164&q=italian%20restaurants&sw=40.466801%2C-74.291267

Methodology:

To start, I need to retrieve the data from the URL (https://geo.nyu.edu/catalog/nyu_2451_34572) given from the Week 3 lab, and convert it from a json file to a pandas data frame. This data needed to be cleaned and merged with latitude and longitude coordinates by using the geopy package from Python.

Next, I need to call the Foursquare API by retrieving my personal Client ID and Client Secret. Once inserted into my code, I want to query all venues within 100 meters from each neighborhood's coordinates. For this I need to create a for loop to filter out the outputs for each venue, so I just have a data frame with the venue's name, coordinates, and category. This data will then be added to the data frame with each venue's respective neighborhood.

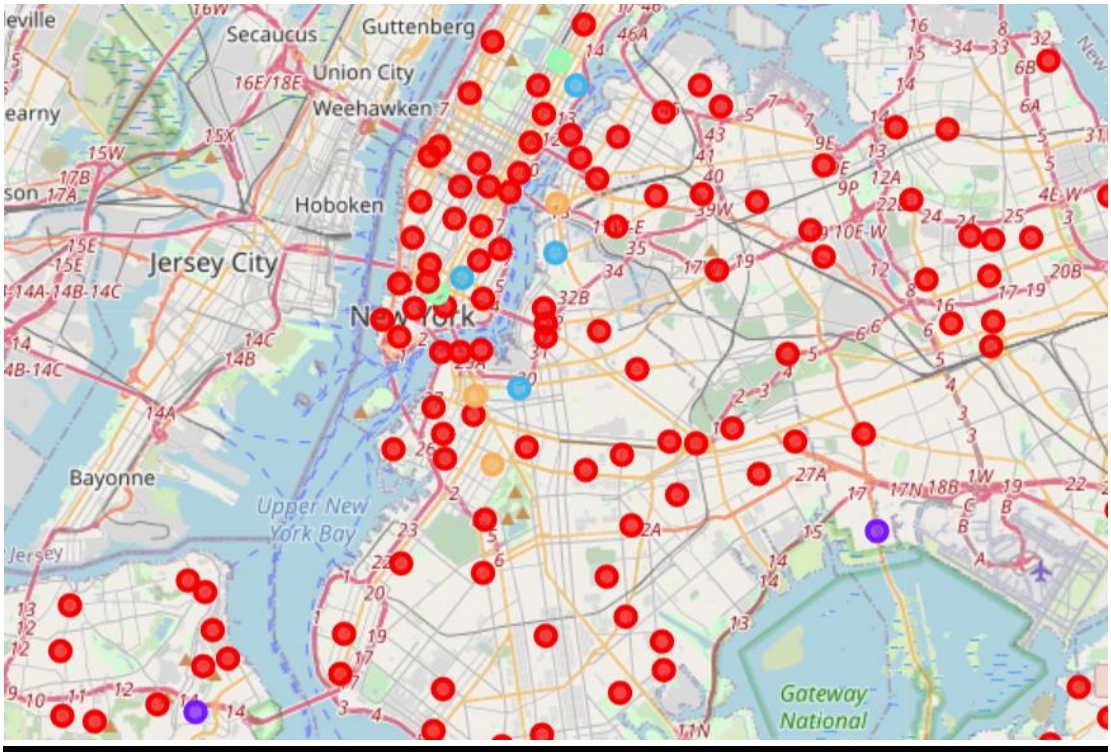
Next, I will use one hot encoding to analyze the venue categories in each neighborhood. Then will slice the data frame to eventually get the frequency of just the Italian restaurants.

Finally, I will use the k-means clustering method to find which neighborhoods are dense with Italian restaurants. I will use k value of 5 since there are more than 200 different neighborhoods in the city. After setting the k-value, I need to create the cluster labels for each neighborhood and join it with the primary data frame. After that I can separate the 5 clusters in their own data frame and analyze them individually based on the frequency row and by producing a New York City map to make sense of the data visually.

Results:

After implementing a k-means cluster analysis with k = 5, we receive the following data with the corresponding colors represented on the map:

- Cluster 1: few to no Italian restaurants
- Cluster 2: moderate high amount of Italian restaurants
- Cluster 3: few to moderate amount of Italian restaurants
- Cluster 4: high amount of Italian restaurants
- Cluster 5: few Italian restaurants



Discussion:

Based on the 5 clusters created, Clusters 2, 3, 4, and 5 seem to be heavily populated areas with an abundance of Italian restaurants. These neighborhoods are also belonging to some of the largest tourist attractions in the city, mostly located in southern Manhattan. Cluster 1 encompasses the most neighborhoods and represent a low frequency of Italian restaurants in those areas. By looking at the map, there is lot of opportunities for more Italian restaurants, especially in the Queens borough.

Conclusion:

Ideal places for an Italian restaurant are certainly in the neighborhoods that are in red, or Cluster 1. Even though there are many neighborhoods to choose from, you may want to consider the amount of people in these areas as well. Since there are more people and tourists in the Manhattan borough you may be better off placing your business in the neighborhoods adjacent to those near southern Manhattan that also are categorized in Cluster 1. Some great neighborhoods would be Greenwich Village, Soho, and the Lower East Side.