IBM DATA SCIENCE CAPSTONE PROJECT

THE BATTLE OF NEIBOURHOOD BEST NEIGHBORHOOD WITH MORE CHINESE RESTAURANT IN MANHATTAN

BEST NEIGHBORHOOD WITH MORE CHINESE RESTAURANT IN MANHATTAN

- The purpose of this Project is to help people in exploring better facilities around their neighborhood. It will help people making smart and efficient decision on exploring great neighborhood out of numbers of other neighborhoods in Mahattan, New Tork.

TARGET AUDIENCE

- Lots of Chinese people are migrating to various states of US and Chinese food are becoming now more and more populars nowadays. This project is for those people who is explorting Mahattan and are interedted in finding more Chinese restaurants.

DATA

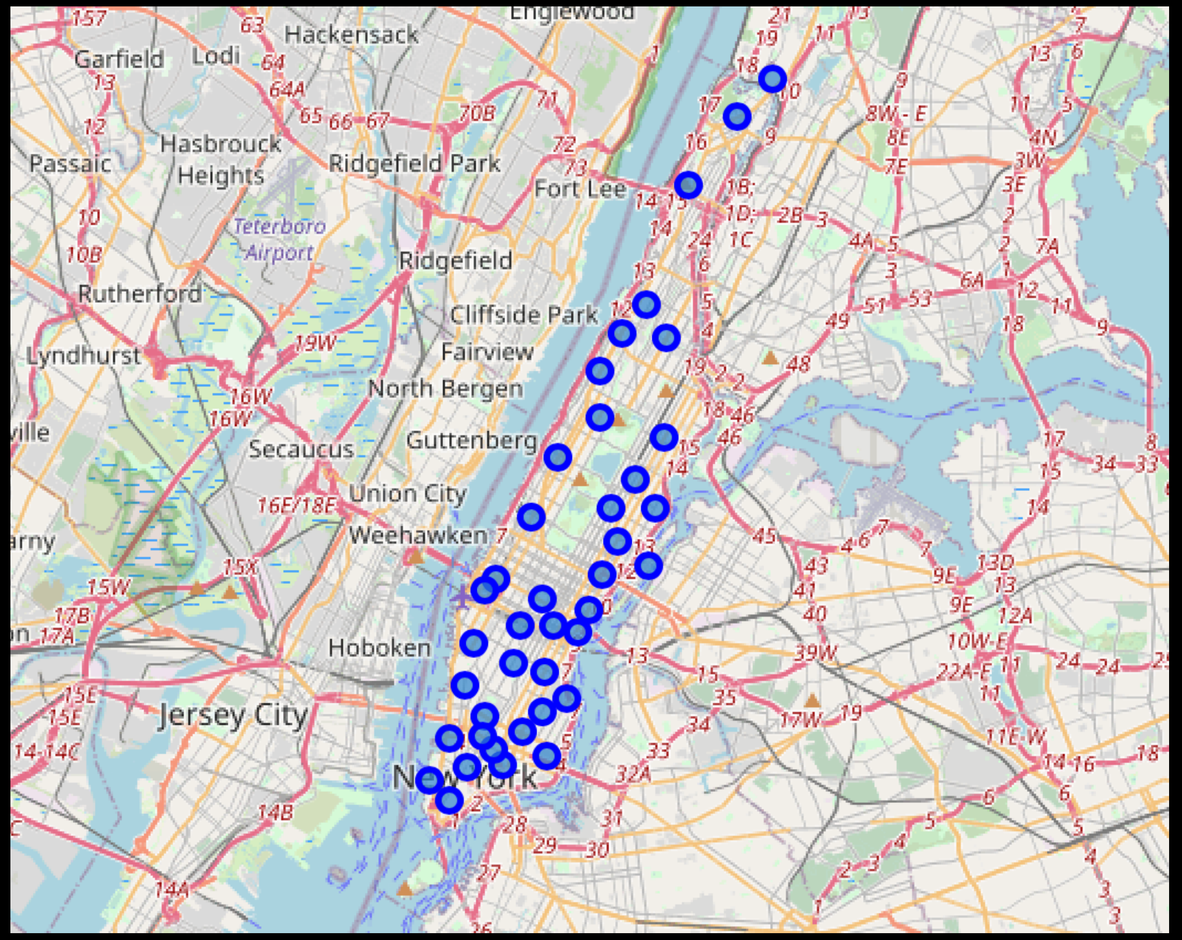
- We will use New York dataset which we scrapped on Week 3. Dataset consisting of latitude and longitude, zip codes.

- Data Link: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

EXTRACTING THE DATA

- Read data through JSON file and create a datafram
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

MAP OF NEIGHBORHOODS OF MANHATTAN



METHODOLOGY

- First, I need to get the list of neighborhoods in Manhattan, New York. I extracted relative information from the JSON data and did standard data preprocessing and cleaning. Then I put them into a Pandas dataframe, and use Geocoder to retrieve coordinates.

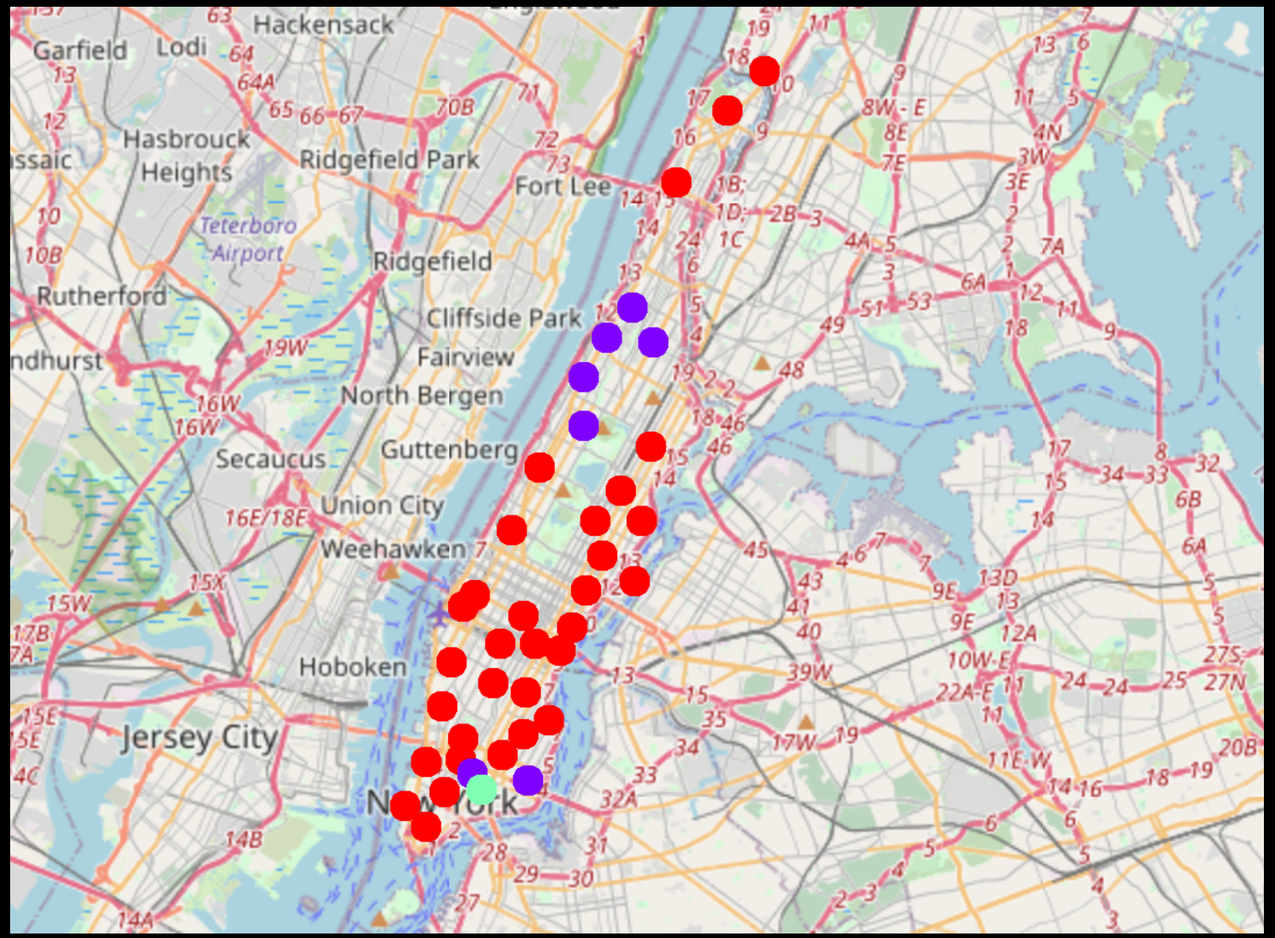
- Next, using Foursquare API, I pulled the names, categories, latitude, and longitude of the venues. With this data, I couldcheck how many unique categories that I can get from these venues. Then, I analyzed each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

METHODOLOGY

- Finally, I focused specifically on “Chinese restaurants”. I used the k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighborhoods in Manhattan into 3 clusters based on their frequency of occurrence for “Chinese food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to explore the restaurant.

RESULT

- cluster 0 : less Chinese restaurant
- cluster 1: medium
- cluster 2: more Chinese restaurant



RESULT

- The neighborhood with Chinese restaurants are in cluster 2 which is Chinatown. Two other neighborhoods close to Chinatown, little Italy and Lower East Side, are in cluster 1 which also have more Chinese restaurants.

- There is another center with more Chinese restaurants, which consists of 5 neighborhoods: Mahattan Valley, Morningside Heights, Manhattanville, Hamilton Heights, and Central Harlem.

- Hence I recommend there two centers for exploring Chinese restaurant, and Chinatown is the first place to check out.