IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Frank Zheng>
<August 28, 2022>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- My methodology was mostly referred from Coursera provided examples.
- 1) New York City neighborhood data was imported and transformed.
- 2) Appropriate Venue Category IDs were obtained from Foursquare following some experimentation for proof of concept.
- 3) Needed functions were constructed (some of which were wholly repurposed from the Coursera examples).
- 4) The Foursquare Venue API was utilized to pull venue data by category type.
- 5) Category data was transformed into counts by neighborhood.
- 6) Counts related to the two pre-selected category count types were extracted, feature-scaled, transformed, and ranked by neighborhood.
- Murray Hill has the highest density and variety of restaurants.

Introduction

- We all know NYC stands out each year in annual GDP which causes population to increase each year. However, starting a business could be extremely difficult in NYC, so I wanted to find out the competition among restaurants in NYC if one day I could open a restaurant.
- Are the restaurants diverse in each region of NYC?
  - How much money does it take to open a restaurant?
  - Where could be a good place to open a restaurant?
  - Who are my consumers?

Methodology

Executive Summary

- Data collection methodology:
  - NYC neighborhoods and coordinates: https://geo.nyu.edu/catalog/nyu_2451_34572
  - Foursquare Venue Category Hierarchy: https://developer.foursquare.com/docs/resources/categories
  - Foursquare venue data API: https://api.foursquare.com/v2/venues/search
- Perform data wrangling
  - When the data is completely gathered, we will perform processing on that raw data to find our desirable features for each venue. Our main feature is the category of that venue. After this stage, the column "Venue's Category" will be One-hot encoded and different venues will have different feature-columns. After On-hot encoding we will integrate all restaurant columns to one column "Total Restaurants" and all food joint columns to "Total Joints" column. We assumed that different restaurants use the Same raw groceries. This assumption is made for simplicity and due to not having a very detailed dataset about different venues.

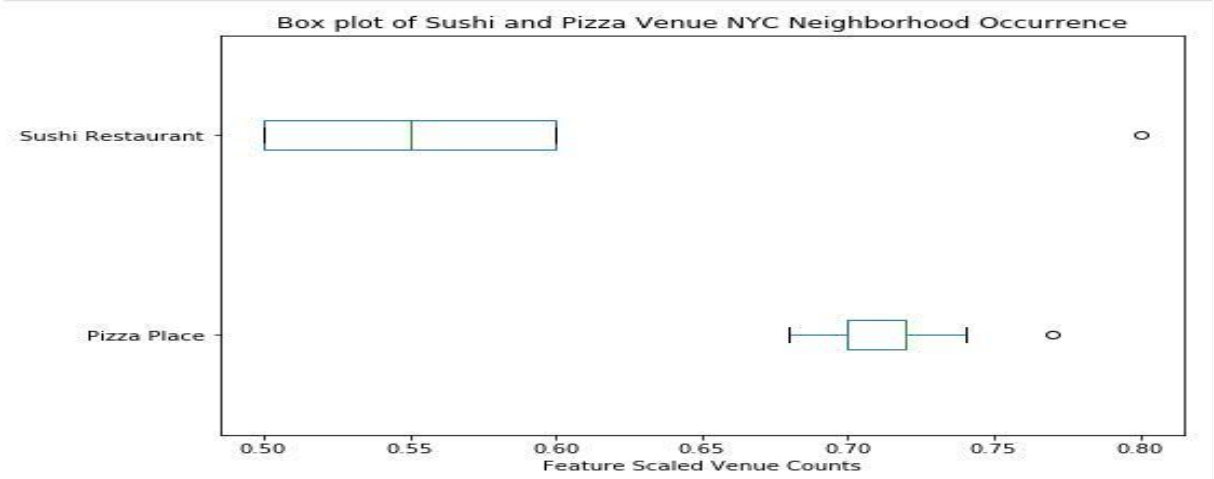
Data Collection

- We will collect data from the links provided before.
  - We will find all venues for each region using Foursquare API
  - Next using Foursquare API, we will find the Ratings, Tips, and Number of Likes for all the Restaurants.
  - Then we will consider all the neighborhoods with average rating greater or equal 9.0 to visualize on map.
  - Finally, we will visualize the Neighborhoods and Borough based on average Rating using python's Folium library.

Out[11]:

Neighborhood	American Restaurant	Arcade	Arepa Restaurant	Asian Restaurant	Bagel Shop	Bakery	Bar	Boat or Ferry	Brazilian Restaurant	Breakfast Spot	...	Sandwich Place	Seafood Restaurant	Snack Place	Spanish Restaurant	Sports Bar	Sushi Restaurant	R
Allerton	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
Annadale	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	3	
Arden Heights	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	2	
Arlington	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
Arrochar	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	

5 rows × 54 columns



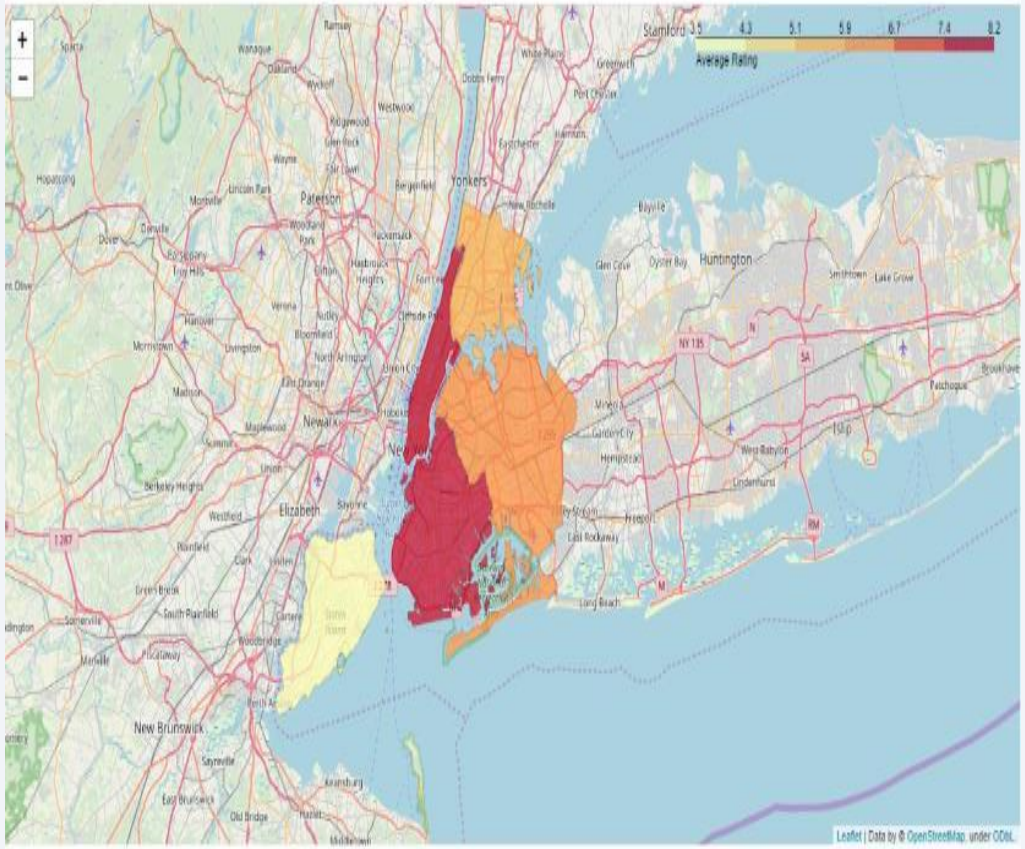
Data Collection – Cont.

Neighborhood	Pizza Place	Sushi Restaurant
Murray Hill	0.77	0.80
Gramercy	0.70	0.60
East Village	0.68	0.60
Bensonhurst	0.72	0.55
Lefrak City	0.72	0.55
Ocean Parkway	0.72	0.55
Downtown	0.72	0.50
Brooklyn Heights	0.70	0.50

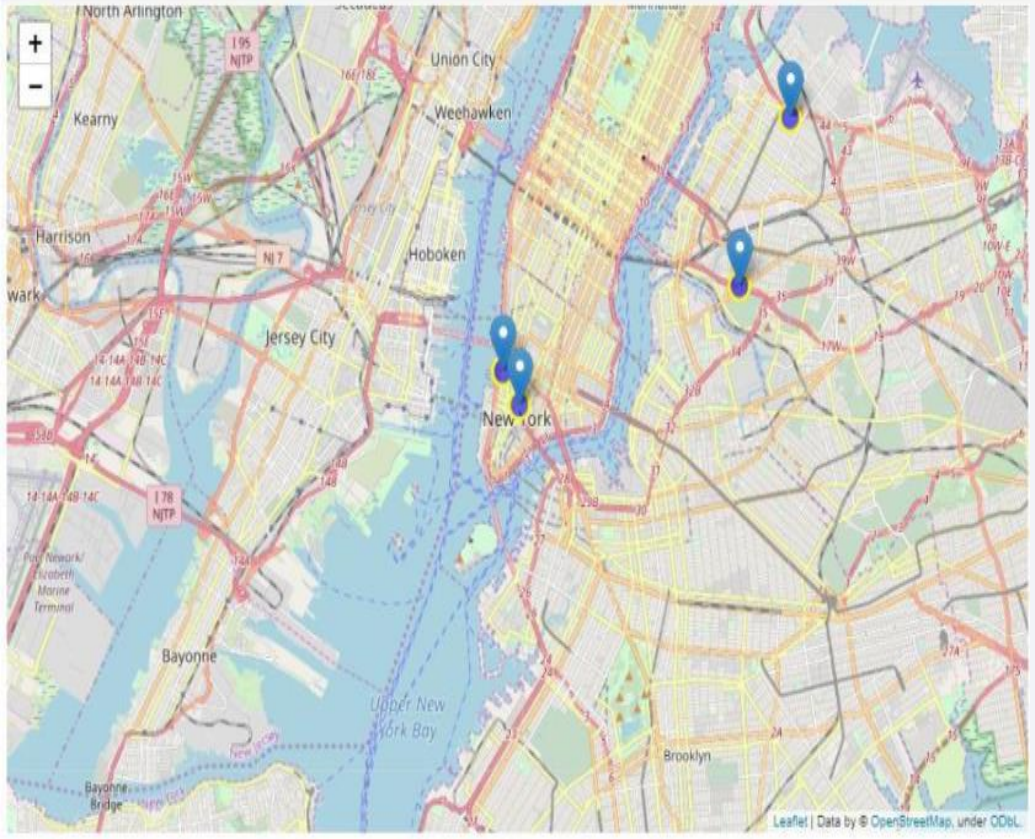
	Neighborhood	Average Rating
12	Civic Center	9.100000
69	Tribeca	9.100000
0	Astoria	9.000000
5	Blissville	9.000000
75	West Village	8.800000
44	Midtown South	8.800000
43	Midtown	8.800000
29	Gramercy	8.733333
25	Fort Greene	8.700000
11	Chelsea	8.700000

EDA with Data Visualization

Borough based on average rating:



Neighbourhoods based on average rating:



EDA with SQL

- We define a function to interact with Foursquare API and get top 100 venues within a radius of 1000 meters for a given latitude and longitude. Below function will return us the venue id , venue name and category.
- Now we will define a function to get venue details like like count , rating , tip counts for a given venue id. This will be used for ranking.
- The above result shows that there are 306 different Neighborhoods in New York.
- Now let create a BAR PLOT to show different Neighborhoods in New York.
- From the above Bar Plot, we can see that Queens has highest number of neighborhoods.
- Now we will get the ranking of each restaurant for further analysis.

Predictive Analysis (Classification)

- Now, we focus on the centers of clusters and compare them for their "Total Restaurants" and their "Total Joints". The group which its center has the highest "Total Sum" will be our best recommendation to the contractor. {Note: Total Sum = Total Restaurants + Total Joints + Other Venues.} This algorithm although is pretty straightforward yet is strongly powerful.

Results

- Based on this analysis, the best recommended neighborhood will be:
- {'Neighborhood': 'Agincourt',
- 'Postal Code': 'M1S',
- 'Neighborhood Latitude': 43.7942003,
- 'Neighborhood Longitude': -79.26202940000002}
- Manhattan has the potential of opening a restaurant.

Conclusions

- The selection of the two venue category types as a metric were not arbitrary. Part of my background includes urban planning, so I am used to looking at land use and business sites as correlate and indicator of other attributes. I sampled result sets from various combinations of Foursquare API pulls for geographic areas with which I was familiar.
- There is always room for improvement and hence the above solution I have provided can also be improved for best results depending upon the data we have.

Appendix

- https://github.com/frankzheng918/Capstone-Project

Thank you!