Choosing My NYC Neighborhood

Peer-graded Assignment: Capstone Project
Coursera IBM Data Science Certification

David Rodwick

November 21, 2019

Contents

Introduction

Background

I am looking to move and have always wanted to live in New York City.

Business Problem to be resolved:

Money is a problem. So, there's that. But before I can even begin to tackle the money problem, I have to take on the neighborhood problem: NYC is very big. Even were I to limit my choice to just a particular borough, there would still be a lot of neighborhoods from which to choose. So, I need a simple metric as a starting point; the foremost goal being a means of simply narrowing my options by utilizing readily available data.

Interested Audience

Anyone looking for: a simple way to access and employ the FourSquare API; better understand the steps involved in using spatial data to inform decision making; or trying to come up with a logical way of developing a metric.

Data

Data and Sources

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

Other Tools

Jupyter Notebook:

https://labs.cognitiveclass.ai/tools/jupyterlab

Python Libraries:/Modules

Numpy, Pandas, json, Requests, Matplotlib, Sklearn, Folium, Pyplot

Methodology

My methodology was almost directly derived from the provided *Coursera* examples.

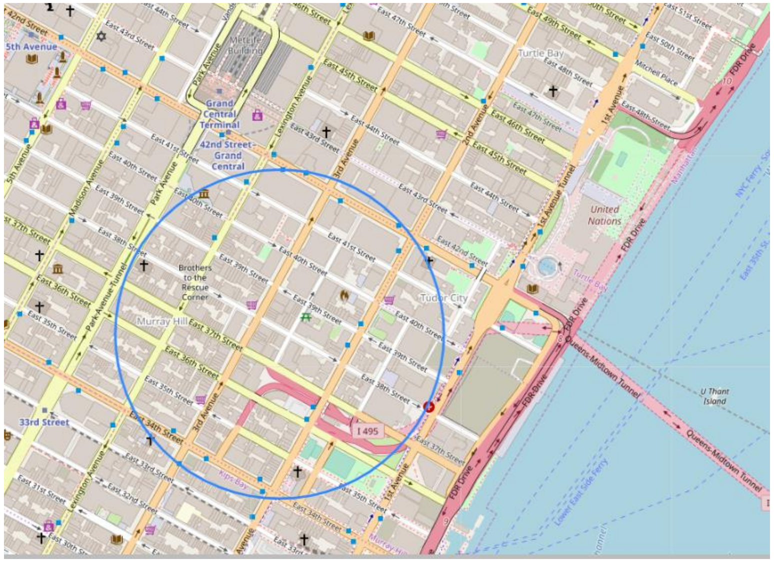
Within a Python Jupyter Notebook , needed libraries were imported then:

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.

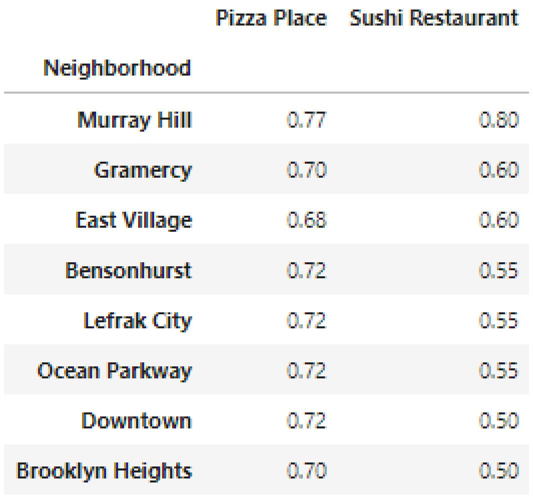
Results

Though above average occurrence of Pizza Places in Murray Hill is important and reason enough to sign a lease, it is still less so than the preternatural Sushi density that renders it an oasis of culinary extrema.

Murray Hill: lots of Sushi and Pizza!



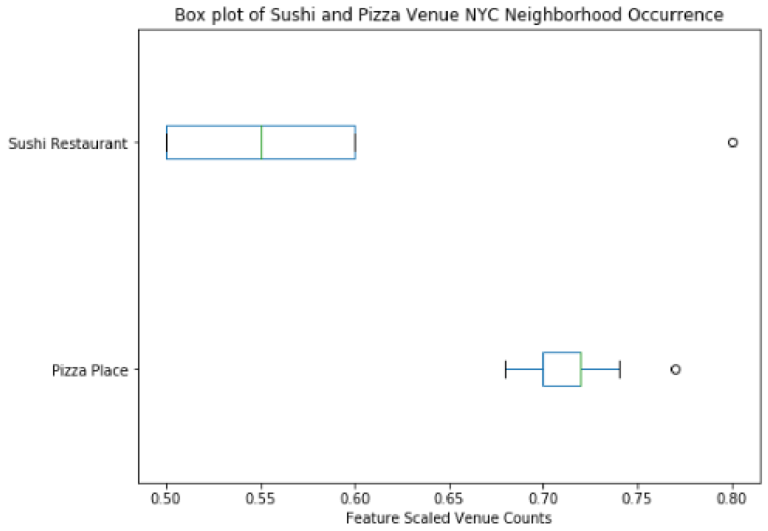
The Best of the Best: an index of culinary joy by NYC neighborhood.



	Pizza Place	Sushi Restaurant
Neighborhood		
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

<u>Discussion</u>: *Concerning Venue Category IDs as Proxy Indicators*

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 are many things I like in my daily life and environment. Without going into further detail, I ascertained that an areas supporting greater density of Pizza Places and Sushi Restaurants correlated well with a nice cross section of them.



I found Pizza Places far easier to come by than Sushi Restaurants (*see left*). So, Sushi was prioritized within the metric. I also looked at areas with below average density of the subject categories and ascertained other attributes I was interested were not present, thus, rendering (and confirming) them as poor residential candidates given my own tastes and lifestyle.

Conclusions

1. After examining a variety of combinations, I was surprised to see how well the Venue Category ID selection (Pizza and Sushi!) hypothesis held up for me. After looking at the area amenities, I really would like to live in Murray Hill.

2. Were I to do this analysis again, I would spend more time than I did examining the FourSquare and New York neighborhood data content and structure. Both contained anomalies (errors, nominalistic problems, and logical inconsistencies) that I think would impair serious analysis (- anything involving money, life or limb).

3. Though I found Folium produced attractive output, I thought it was clunky, not particularly user friendly, and not terribly versatile.