Restaurant Location Selection in New York City

IBM Data Science Professional Certificate Capstone

Charlie Kelley

1/23/2018

Table of Contents

Introduction

PROBLEM

I am opening a restaurant in New York City and want to find the most profitable neighborhood location. The restaurant will be fast casual tacos and the target audience is younger consumers who live in the neighborhood or will take subways to get there.

My initial hypotheses are that I can maximize revenue by selecting a location that has:

- High population density
- High concentration of 25-34 year olds
- Low concentration of existing taco restaurants
- Close proximity to subway stops

OBJECTIVE

My objective is to find out what neighborhoods in NYC best meet the above criteria to help pick the optimal neighborhood location.

Why is this interesting?

People would be interested in this project if they want to see a scalable and repeatable process for using publicly available data for selecting business locations based on customer profiles.

With so many possible locations to choose from, KNN clustering will help me identify similar locations as a starting place to pick a final one.

Data Understanding

- I will use 5 sources of data to complete my analysis.

FOURSQUARE API – EXPLORE BASED ON QUERY

From this API I will find out how many taco restaurants are located within a certain radius of each zip code.

https://developer.foursquare.com

NYC ZIP CODES AND NEIGHBORHOODS LIST

Scraped from NYC Department of Public Health – this table breaks down NYC Zip Codes by neighborhood for additional context.

https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm

US ZIP CODE POPULATION AND MEDIAN AGE

This file contains the population of each zip code in US as well as median age. It was pulled from the 2014 5-year American Community Survey.

https://www.census.gov/programs-surveys/acs/

US ZIP CODE LATITUDE AND LONGITUDE

This file contains the latitude and longitude of each zip code in US.

https://gist.github.com/erichurst/7882666

NYC SUBWAY LOCATIONS

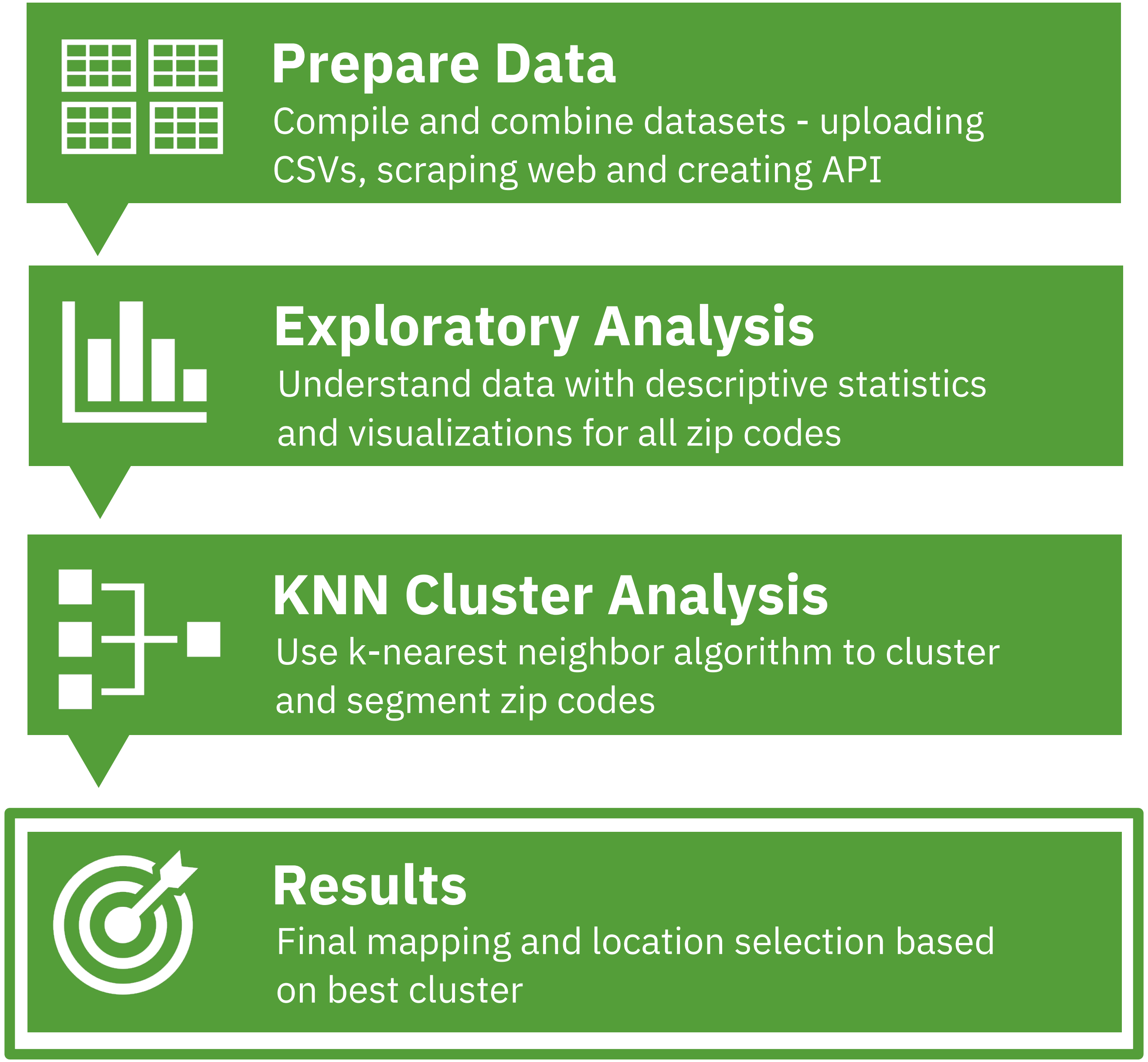
This file contains the latitude and longitude of each subway entrance in NYC.

https://datamine.mta.info/

Methodology

PROCESS OVERVIEW

- I am going to use Python to perform my analysis.
- Data is going to be brought together from multiple sources including Census CSVs, scraping from websites and a Foursquare API.
- I will normalize the data and use a k-nearest neighbor algorithm to determine neighborhood segments, from which I will have a list to choose a location from.
- As a final step, I will overlay a map of recommended locations with subway stops to find best locations.



Prepare Data

Compile and combine datasets - uploading CSVs, scraping web and creating API

Exploratory Analysis

Understand data with descriptive statistics and visualizations for all zip codes

KNN Cluster Analysis

Use k-nearest neighbor algorithm to cluster and segment zip codes

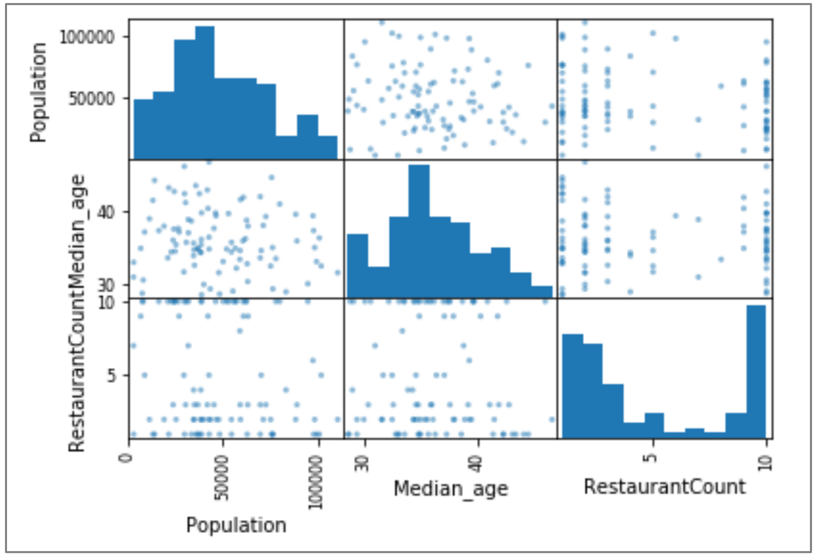
Results

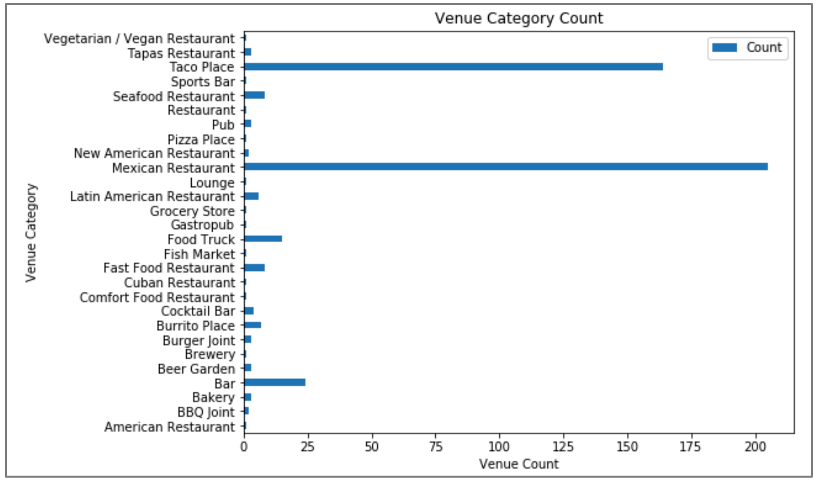
Final mapping and location selection based on best cluster

Exploratory Analysis

INFERENCES

- Population and median age are distributed fairly normally by zip code, although slightly skewed to being younger and smaller.
- Zip codes tend to have either few or many restaurants that serve tacos (*Note: because of API limitation only 10 restaurants per zip could be counted*).
- There is no correlation between any of the three dimensions.
- While 28 venues serve tacos in the identified zip codes, the most likely restaurant category to serve tacos are ‘Mexican Restaurant’ and ‘Taco Place’.

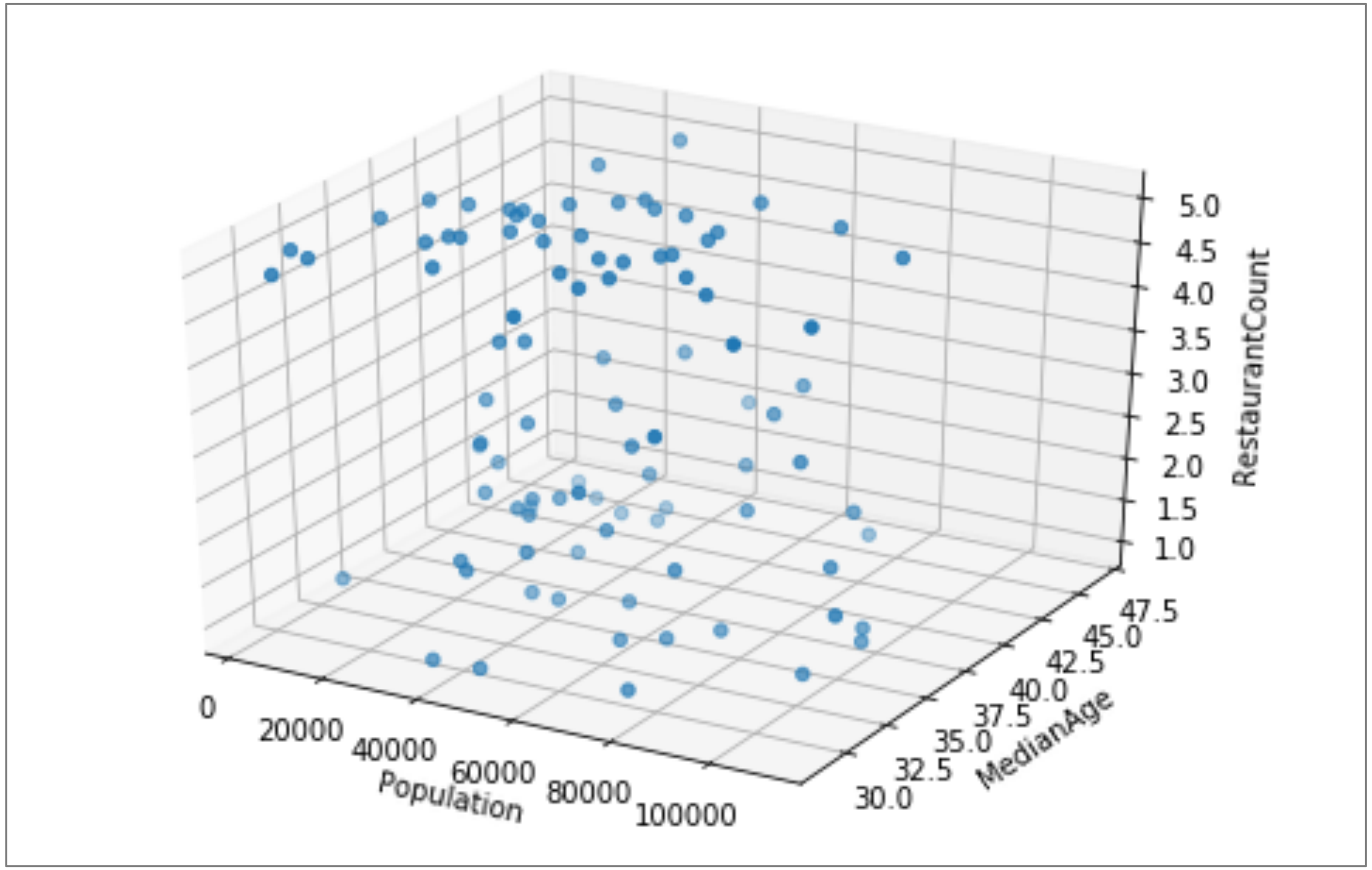




Exploratory Analysis

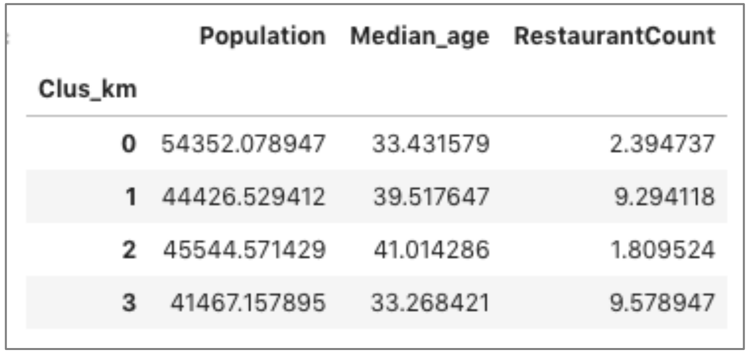
3D SCATTERPLOT

- Shows relationships between Population, Median Age and Restaurant Count by Zip Code



KNN Cluster Analysis

- Using the Elbow Method I determined that the optimal k is 4
- The four clusters are described by their means below:



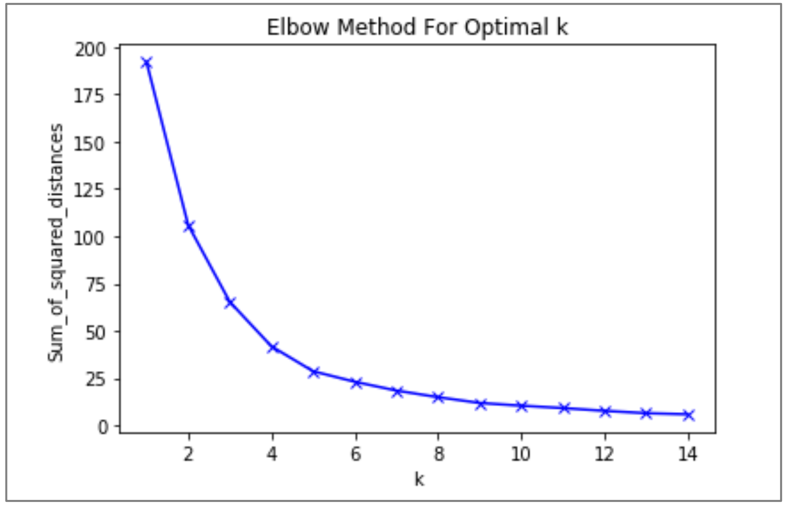
Clus_km	Population	Median_age	RestaurantCount
0	54352.078947	33.431579	2.394737
1	44426.529412	39.517647	9.294118
2	45544.571429	41.014286	1.809524
3	41467.157895	33.268421	9.578947

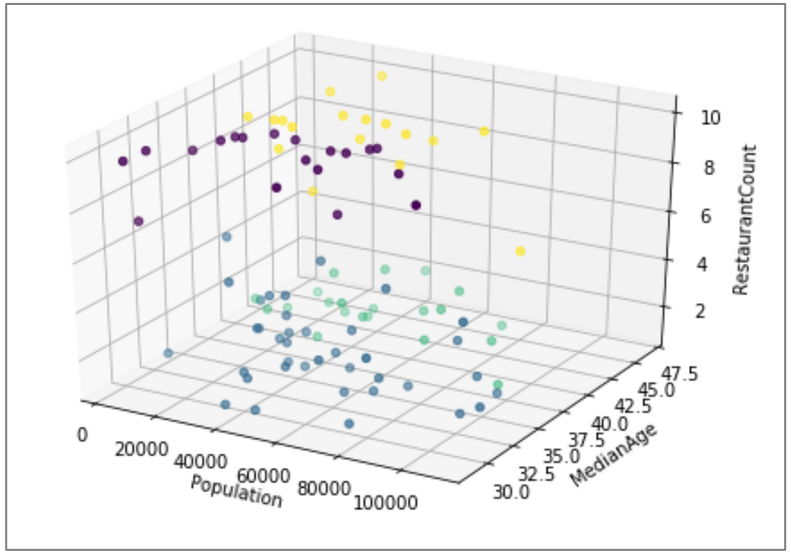
Cluster 0: Lots of customers and little competition

Cluster 1: Fewer customer and lots of competition

Cluster 2: Fewer customers and little competition

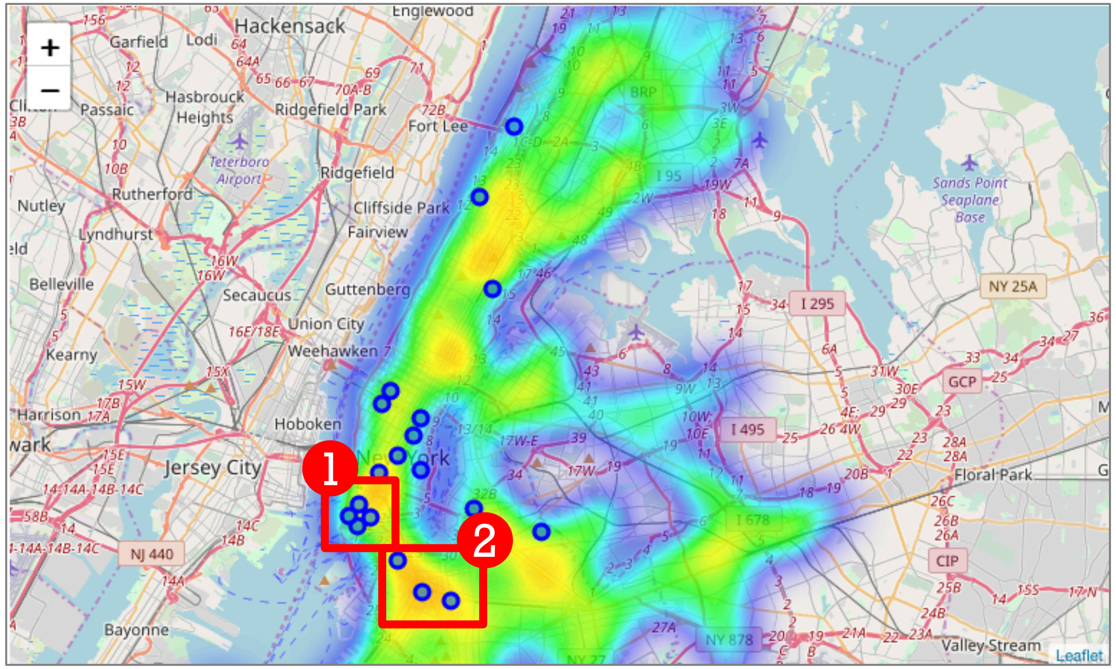
Cluster 3: Lots of customers and lots of competition





Results

- Cluster 0 appears to provide the best zip codes matching my initial hypotheses – this provides a list of 19 zip codes.
- To narrow down even further - final step in analysis is to overlay that cluster of zip codes with the subway stops.
- The pins in the map to the right show a zip code in Cluster 0, the heatmap shows number of subway stops.
- **Lower Manhattan** and **Northwest and Central Brooklyn** are two groups of zip codes prioritized for further exploration.



1 **Lower Manhattan** – 4 zip codes in cluster near density of subway stops

2 **Northwest and Central Brooklyn** – 3 zip codes in cluster near density of subway stops

Discussion

LIMITATIONS

I faced many limitations that could affect this analysis. Because of the Foursquare API I could only limit the number of locations per zip code to 10 before using all my allotted calls from the service. I also only used subway locations – it is possible there are better public transportation modes I should consider such as bus stops and bike share locations.

With those considerations, I am still happy with my analysis as a simple template for answering my initial business question.

FURTHER STUDY NEEDED

- Does competition help or hurt? My hypothesis is that nearby taco places will help my profitability, but further study might show that being the only nearby taco restaurant is better for business.
- How does concentration of complimentary venues – or venues likely to be visited next after eating at a taco restaurant – contribute to the success of a restaurant?
- My analysis only contained subway stops – there are also buses and bikeshares available. Does including these additional modes of transportation change my results?
- Once I select a location – can I use Foursquare API to mine further data about features and attributes of well liked restaurants in the location?

Conclusion

- Even with the limitations and further study needed described in the Discussion section – I believe this study is an interesting look at site selection and finding data to meet initial hypotheses. Using KNN classification helped narrow down zipcodes into more manageable clusters that were beyond what a human could easily infer, and overlaying subway stops further narrowed down to specific locations to look into.
- Overall I feel this study was an interesting application of both exploratory and machine learning analysis to a problem many businesses face.