Restaurant Location Selection in New York City

IBM Data Science Professional Certificate Capstone Project

Charlie Kelley

12/31/2019

1.0 Introduction

1.1 Background

The business problem I am trying to solve is to find the best location for opening a new restaurant in New York City. In order to succeed financially, restaurants need the right combination of concept and customers which leads to the need for site selection analysis.

Research shows that fast casual restaurants are currently gaining more market share faster than any other type of food business as visits to fast causal restaurant have increased 6% over the last five years (QSR Magazine). Because of this, I have decided to open a fast casual restaurant concept that specializes in tacos with the target audience of younger consumers who live in the neighborhoods or will take public transportation to get there.

1.2 Objective

The objective of this analysis is to find a location that will maximize potential revenue for my new restaurant. 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 public transportation

To perform this analysis, I combined multiple publicly available datasets and used KNN cluster analysis to group all the neighborhoods into clusters and find the cluster that best matches my hypotheses above giving me a list of neighborhoods to explore further.

1.3 Interest

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. This

process can be used for businesses across industries from retail to professional services, and also refined and improved.

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 and could be used by others.

2.0 Data

2.1 Data Sources

To complete my analysis, I needed a variety of data from population to restaurant counts based on location. The most granular population data I could find is at the zip code level, where I could find total population and median age, but not breakdown of age categories so I could zero in on the 25-34 age group I am desiring. I decided that I will have to satisfice with median age. For additional context, I will have a neighborhood associated with each zip code. Latitude and longitude will also be added to each zip code row.

I used five unique datasets to achieve this:

1. **Foursquare API**
   From this API I will find out how many taco restaurants are located within a certain radius of each zip code. I will use a query to narrow down on establishments that sell tacos.

   Source: https://developer.foursquare.com

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

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

3. **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.

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

4. **US Zip Code Latitude and Longitude**
This file contains the latitude and longitude of each zip code in US.

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

5. **NYC Subway Locations**
This file contains the latitude and longitude of each subway entrance in NYC.

Source: https://datamine.mta.info/

2.2 Data Cleaning

To complete my KNN cluster analysis I need a single dataset that contains the following attributes:

- Zip Code
- Neighborhood
- Latitude
- Longitude
- Population
- Median Age
- Restaurant Count

2.2.1 Neighborhood Data

To start I scraped a NYC Department of Public Health website to get zip codes grouped by neighborhood and break out each zip code into their own row.

	Neighborhood	ZipCodes
0	Central Bronx	10453, 10457, 10460
1	Bronx Park and Fordham	10458, 10467, 10468
2	High Bridge and Morrisania	10451, 10452, 10456
3	Hunts Point and Mott Haven	10454, 10455, 10459, 10474
4	Kingsbridge and Riverdale	10463, 10471

	ZipCodes	Neighborhood
0	10453	Central Bronx
1	10457	Central Bronx
2	10460	Central Bronx
3	10458	Bronx Park and Fordham

Next, I create a new dataframe with Latitude and Longitude for each zip code and append to the original dataframe.

I do the same for Population and Median Age.

	ZIP	Population	Median_age
0	601	18088	37.1
1	602	40859	39.0
2	603	53162	39.2
3	606	6415	39.2
4	610	28805	39.7

	ZipCodes	Neighborhood	LAT	LNG	Population	Median_age
0	10453	Central Bronx	40.852779	-73.912332	79002	31.0
1	10457	Central Bronx	40.847150	-73.898680	71015	29.4
2	10460	Central Bronx	40.841758	-73.879571	58676	30.0
3	10458	Bronx Park and Fordham	40.862543	-73.888143	76276	29.4
4	10467	Bronx Park and Fordham	40.869953	-73.865746	99251	33.1

2.2.2 Restaurant Data

The next data component is restaurant data from Foursquare which I will use to find the count of taco restaurants in the zip codes I am interested in. I created an API that searches for the query “taco” and returns the ‘Category’ as well as latitude and longitude.

	name	categories	lat	lng
0	Estrellita Poblana	Mexican Restaurant	40.854010	-73.888923
1	Grito Mexican Grill	Mexican Restaurant	40.853643	-73.930886
2	Guacamole	Mexican Restaurant	40.869405	-73.916408
3	Tacos El Paisa	Mexican Restaurant	40.853318	-73.930760
4	Guadalupe Bar and Grill	Mexican Restaurant	40.867334	-73.920863

Next I create a function that assigns the restaurants to a zip code from my New York City neighborhood dataframe based on its latitude and longitude as seen below. Because of Foursquare API call limits I only pull 10 restaurants per zipcode.

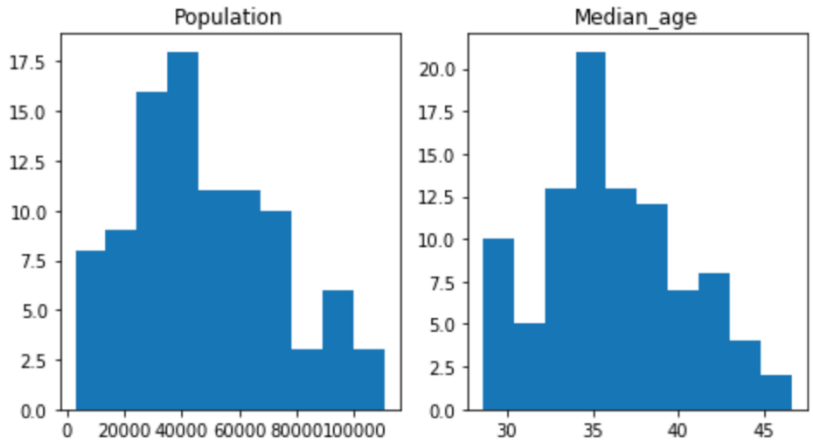
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	10458	40.862543	-73.888143	Chipotle Mexican Grill	40.860335	-73.890507	Taco Place
1	10467	40.869953	-73.865746	La Estrellita Poblana	40.867077	-73.867595	Mexican Restaurant
2	10451	40.820479	-73.925084	La Perla Restaurant	40.817281	-73.922249	Mexican Restaurant
3	10451	40.820479	-73.925084	Taco Bell	40.817215	-73.923504	Taco Place
4	10454	40.805489	-73.916585	Jalisco Tacos	40.806481	-73.915559	Taco Place

With these two dataframes I can get into the exploratory analysis and KNN machine learning.

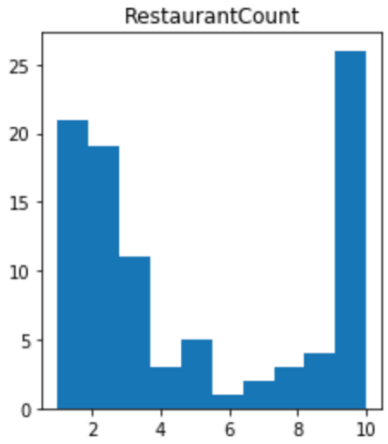
3.0 Methodology

3.1 Exploratory Analysis

Population and Median Age are fairly normally distributed by zip code, although they do skew to be younger and more populous with a longer tail of older and smaller zip codes.

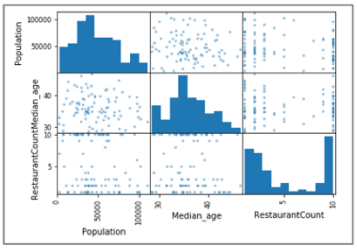


Restaurant count per zipcode is the opposite. Zip codes tend to have few or many taco restaurants.



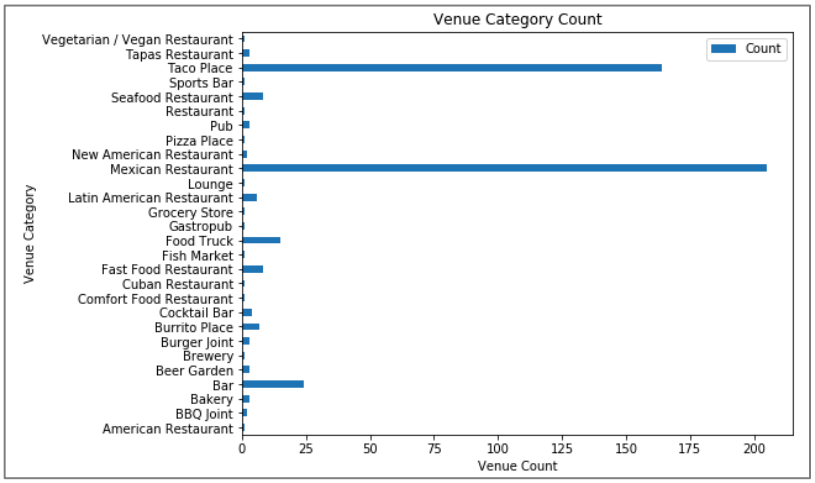
There also seemed to be little to infer from the relationships between of zipcode Population, Median Age and Restaurant Count as they were not correlated to eachother at all, which also means I can use them all as variables in my machine learning analysis.

Correlation



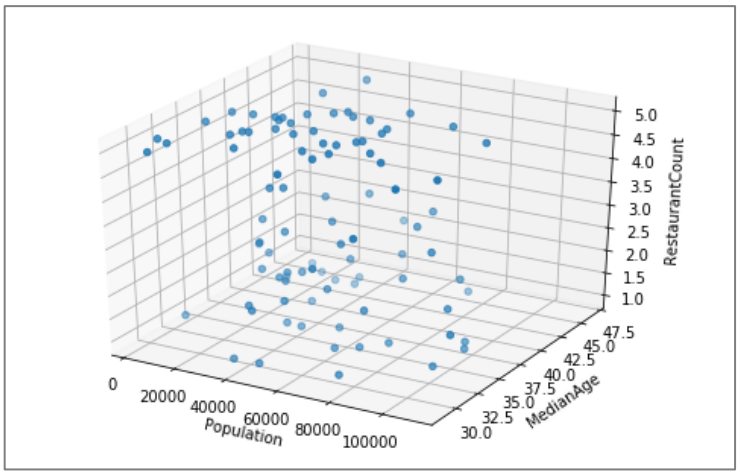
	Population	Median_age	RestaurantCount
Population	1.000000	-0.118227	-0.160044
Median_age	-0.118227	1.000000	-0.034871
RestaurantCount	-0.160044	-0.034871	1.000000

Getting into the restaurants, there are 469 within the zipcodes included in this analysis. The categories of ‘Taco Place’ and ‘Mexican Restaurant’ were the largest by far of the group.



3.2 KNN Cluster Analysis

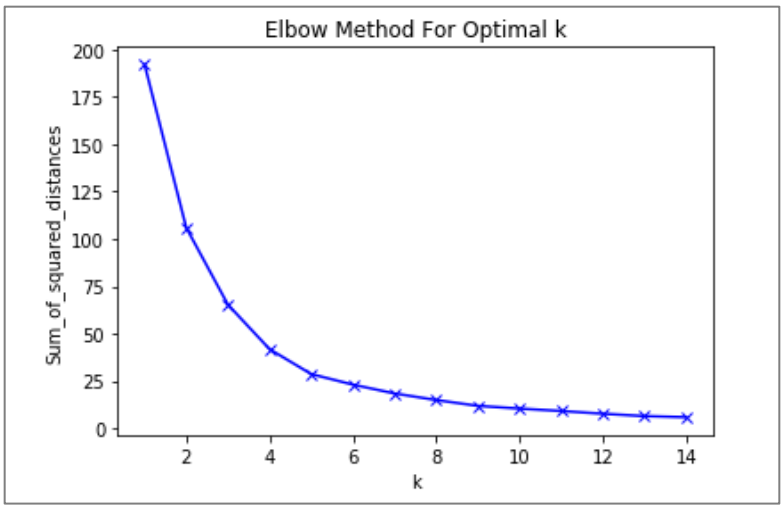
My research goal is to find a list of similar zipcodes to further explore putting a restaurant in, and when these three variables were plotted together, visually it is hard for me as a human to see how to group them into clusters. I believe that using the K nearest neighbor (KNN) algorithm will help me break these zipcodes into more manageable and correct clusters for further analysis.



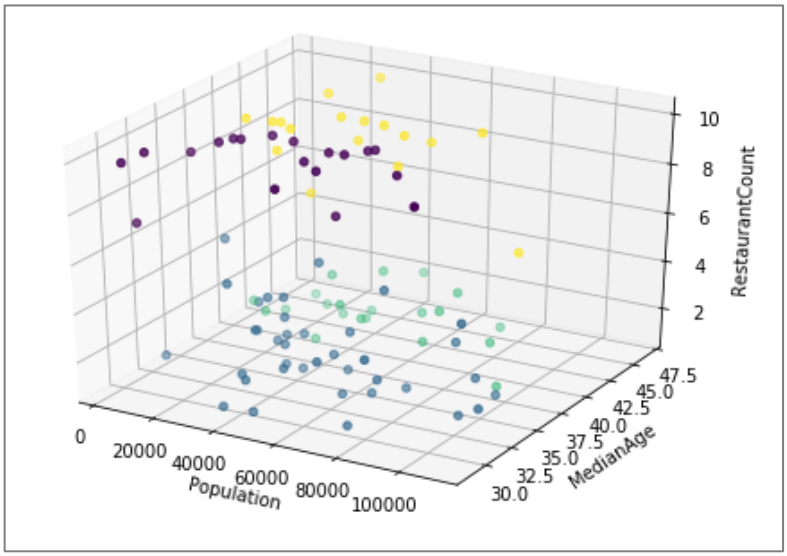
The KNN algorithm will partition all the zipcodes into k number of clusters based on their mean. This will give me clusters of zipcodes to compare which best meets the criteria I set out of younger population, higher population density and low number of taco restaurants.

To start the KNN, I create a dataframe for modeling and normalize the data in it.

Next, I need to find the optimal number of clusters. To do this I use the elbow which runs the KNN algorithm on the data and shows me the sum of squared distances for each number of clusters. I then find the ‘elbow’ of the chart where the sum of squared distances is not much different than the one before.



For this data I think 4 clusters is the right k, so I make that my input in my model. I then run the model against the normalized data and find my clusters.



4.0 Results

The clusters show different characteristics when the mean of the inputs are compared:

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

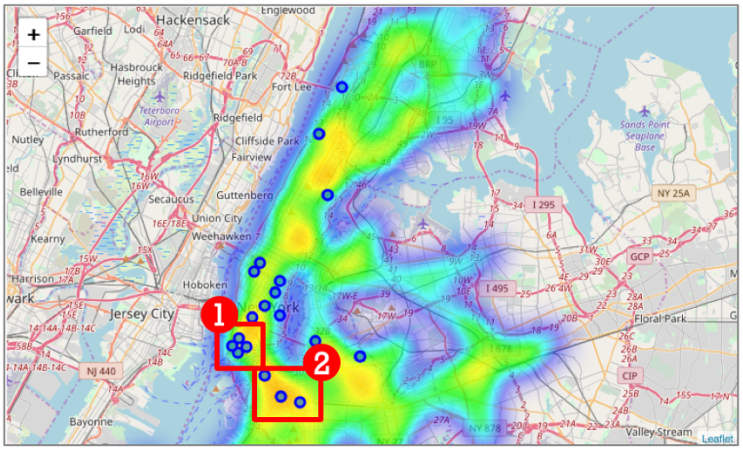
Looking at these clusters I can infer that:

- **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

Based on my initial assumptions, I think Cluster 0 is the best option for me to explore more. It includes 19 unique zipcodes and 11 unique neighborhoods.

	ZipCodes	Neighborhood	LAT	LNG	Population	Median_age	RestaurantCount
12	11238	Central Brooklyn	40.679171	-73.963804	51959	34.8	10
19	11201	Northwest Brooklyn	40.693682	-73.989693	56488	34.9	10
22	11217	Northwest Brooklyn	40.682306	-73.978099	38567	35.4	8
25	11211	Greenpoint	40.712597	-73.953098	94681	30.1	10
30	11237	Bushwick and Williamsburg	40.704160	-73.921139	55258	29.1	10
33	10001	Chelsea and Clinton	40.750633	-73.997177	22767	34.4	10
35	10018	Chelsea and Clinton	40.755319	-73.993114	8062	30.6	10
38	10029	East Harlem	40.791763	-73.943970	77857	33.5	10
40	10010	Gramercy Park and Murray Hill	40.739065	-73.982255	30708	35.7	10
41	10016	Gramercy Park and Murray Hill	40.745224	-73.978297	50112	33.1	10
44	10012	Greenwich Village and Soho	40.725581	-73.998078	24846	34.5	10
48	10005	Lower Manhattan	40.706027	-74.008835	7570	28.8	10
49	10006	Lower Manhattan	40.709614	-74.012954	2950	31.0	7
50	10007	Lower Manhattan	40.713848	-74.007755	6748	34.9	9
51	10038	Lower Manhattan	40.709278	-74.002562	20997	33.6	10
54	10003	Lower East Side	40.731829	-73.989181	57068	32.2	10
55	10009	Lower East Side	40.726399	-73.978631	61806	36.2	10
64	10031	Inwood and Washington Heights	40.825288	-73.950045	58912	33.4	8
66	10033	Inwood and Washington Heights	40.850545	-73.933983	60520	35.9	10

To add another layer of inferential analysis beyond the KNN clustering, I want to look at the density of subway stops near these zipcodes to narrow my choices even further. To do this I create a heatmap of all the subway stops and overlay the zipcodes from Cluster 0 to see if there are any zipcodes that stand out.



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

Two neighborhoods and corresponding zipcodes stand out: Lower Manhattan which has 4 zipcodes around a density of subway stops and Northwest and Central Brooklyn which has 3 zipcodes around a density of subway stops. These are the seven zipcodes I believe my restaurant will be most profitable in.

5.0 Discussion

I faced many limitations that impacted the outcomes of this research.

First of all – I could only call a maximum of 10 restaurants from Foursquare's API because of limitations on my account. It is possible that not having a cap on the number of restaurants per zipcode could impact my analysis and change of the clusters were partitioned.

I also was only able to find subway stop latitude and longitudes and there are many more modes of public transportation that shoud be considered. Perhaps adding bus stops and bike shares would impact my analysis.

Another limitation was that I based my analysis on population count and median age as this was the only way to tie this demographic data to zipcodes and neighborhoods. In a perfect world – I would have been able to directly tie the number of residents in the exact age group I was looking for to each zipcode to match my initial assumption.

While working on this report – I had some questions that would require future analysis:

- Does competition help or hurt? My hypothesis is that nearby taco places will hurt my profitability, but further study might show that being the only nearby taco restaurant is bad 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?
- How would overlaying real estate rental data help me decide where to locate my restaurant as I could be priced out of some locations.

6.0 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.

Sources

https://www.qsrmagazine.com/news/study-fast-casual-traffic-only-segment-grow-last-5-years