OPENING A PIZZA ROBOT SALES OFFICE IN NYC

Capstone Project: Battle of the Neighborhoods¶

ABSTRACT

The following report was done for a fictional client in the food robotics industry in April-May 2020. It locates a series of neighborhoods in which to open a sales office using data analytic techniques. Report performed for Coursera Applied Data Science Capstone.

Analysis by Josh S

Coursera Data Science Capstone – IBM Data Science Professional Certificate

Contents

Introduction

Background:

In the wake of the post-lockdown phase of the Covid-19 pandemic, the restaurant industry requires significant changes to it's core business model that will enable better sanitation and implementation of recommended social distancing for employees and customers. New York City was the hardest hit by the pandemic and as the city with the highest population density is an ideal test case for technological solutions that will aid the achievement of implementing sanitation and social distancing guidelines while still achieving the goal of quality product and high-volume production of food for takeout, delivery and commercial clients. The help the restaurant industry needs may come in the form of robotics (made by companies such as Picnic (https://www.hellopicnic.com/press), which can preserve the input quality of ingredients while automating the busywork and reducing the number of physical employees in the kitchen space to enable social distancing and greater sanitation.

Client and Business Case/Problem:

This report was performed for a vendor in the food robotics industry in April-May 2020. The goal of this report is to identify neighborhoods ideal for placing a sales office for the client. _However, more generally, the analysis would be useful for any entrepreneurs in the food-service equipment industry seeking to locate their office in an ideal space. As New York City, specifically Manhattan, is famous for it's pizza, it is an ideal test market for the rollout of the pizza vending robot our client is making. Specifically: the client has asked "Where in Manhattan should we locate our office to maximize density of potential sales venue?"

Deliverable:

Specifically, our client wants a list of target neighborhoods with the a large number of pizza vendors*, so as to have more potential sales targets.

We can use our data analysis abilities and deploy a range of techniques to generate a list of ideal neighborhoods and then outline the advantages of each neighborhood (and flag potential downsides) which will enable our client to make an informed decision about the placement of their office.

Data

Our business case and client question should shape the data we are going to use. In this case, the number of client-relevant restaurants in each neighborhood (i.e. Italian and Pizza restaurants, but also fast food places serving pizza. Note that Foursquare already flags any restaurant that serves pizza, regardless of the other cuisines.)

Several data sources will be used in this analysis. For neighborhood data, NYU's Spatial Data Repository contains an excellent open-source map of New York's neighborhoods with geospatial point data (https://geo.nyu.edu/catalog/nyu_2451_34572) which we will use in this analysis [2]. The .json file has coordinates of all the neighborhoods in NYC. However, since our client is only interested in Manhattan, I cleaned the data and reduced it to city of Manhattan.

For locations within each neighborhood, We will be using the Foursquare API [3] to find data on concentrations of restaurants, including pizza places and Italian restaurants , as well as other venues and neighborhood characteristics. Foursquare is a technology company that offers software that enables users to upload real-time location and venue data. As a result, we can find up-to-date lists of venues and target specific types of venues using the API.

Finally, In terms of python libraries, there are several used for data gathering in this project:

We will be using the geopy library to get the latitude and longitude values of New York City.

The Folium library [4] will be used for data visualization. Several other packages will be used for our analysis, for more information, see the main text in the Methodology section.

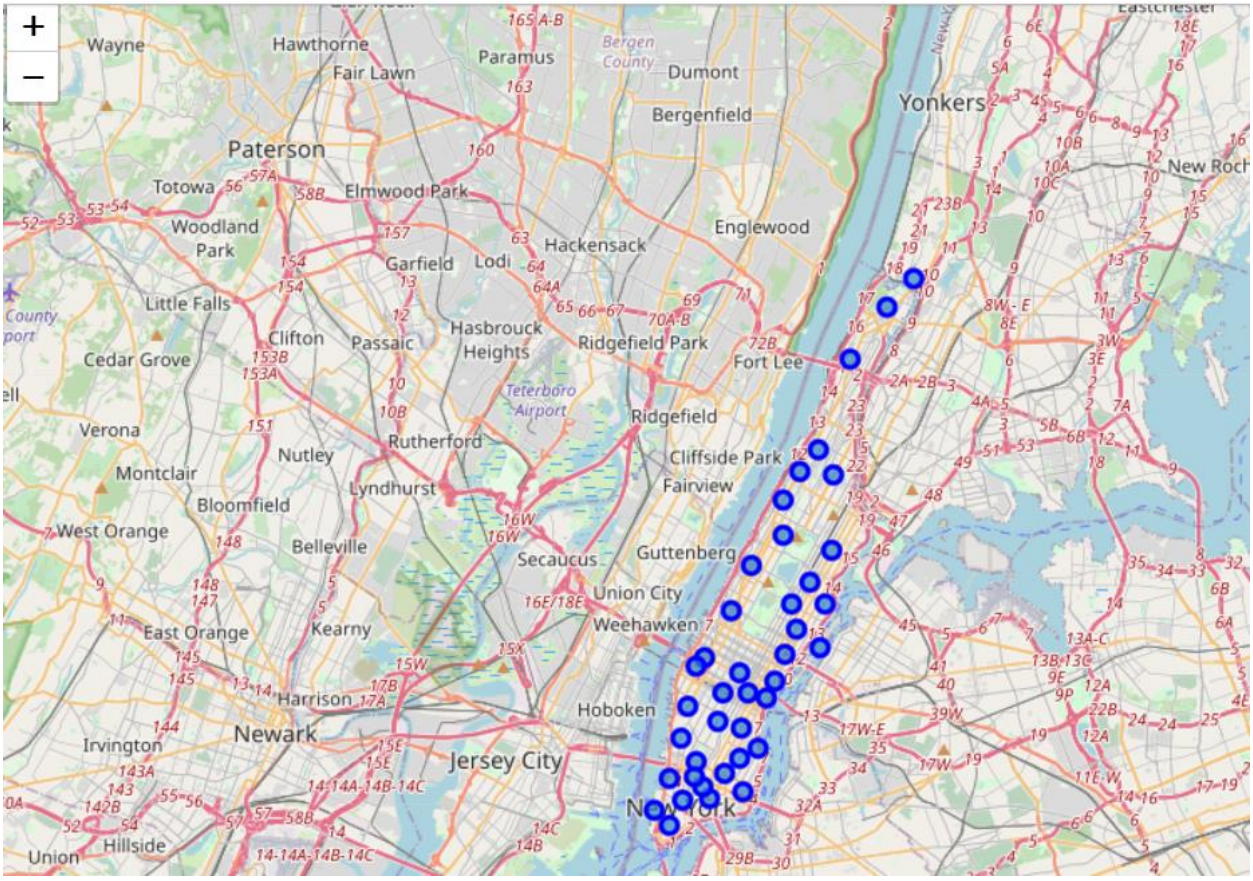
Utilizing data analysis and analytical techniques discussed in the labs from Coursera's IBM Data Science Professional Certificate, the data will be analyzed using the methods outlined below.

Methodology

Utilizing data analysis and analytical techniques discussed in the labs from Coursera's IBM Data Science Professional Certificate, the data will be analyzed using the methods outlined below.

As a database, I used a notebook written GitHub repository in my study.

I used Python’s **Folium** library to visualize Manhattan neighborhoods, creating the map of Manhattan with the neighborhoods superimposed on top as dots, by getting latitude and longitude using the Geopy package to create the map seen below.



I then utilized Foursquare's API to get the data on pizza restaurants, limiting to the first 100 venues and defining a search radius of 500 (meters). I did this by creating a function that uses category id from Foursquare to get ONLY the pizza places (or client-relevant places that serve pizza, i.e. fast food restaurants, etc.) called getPizzavenues, and then calling that function to create a pandas dataframe called nyc_venues that uses the function we just define to find all the pizza places in Manhattan and put them into a dataframe—see below for the head(first 5 results) of the dataframe)

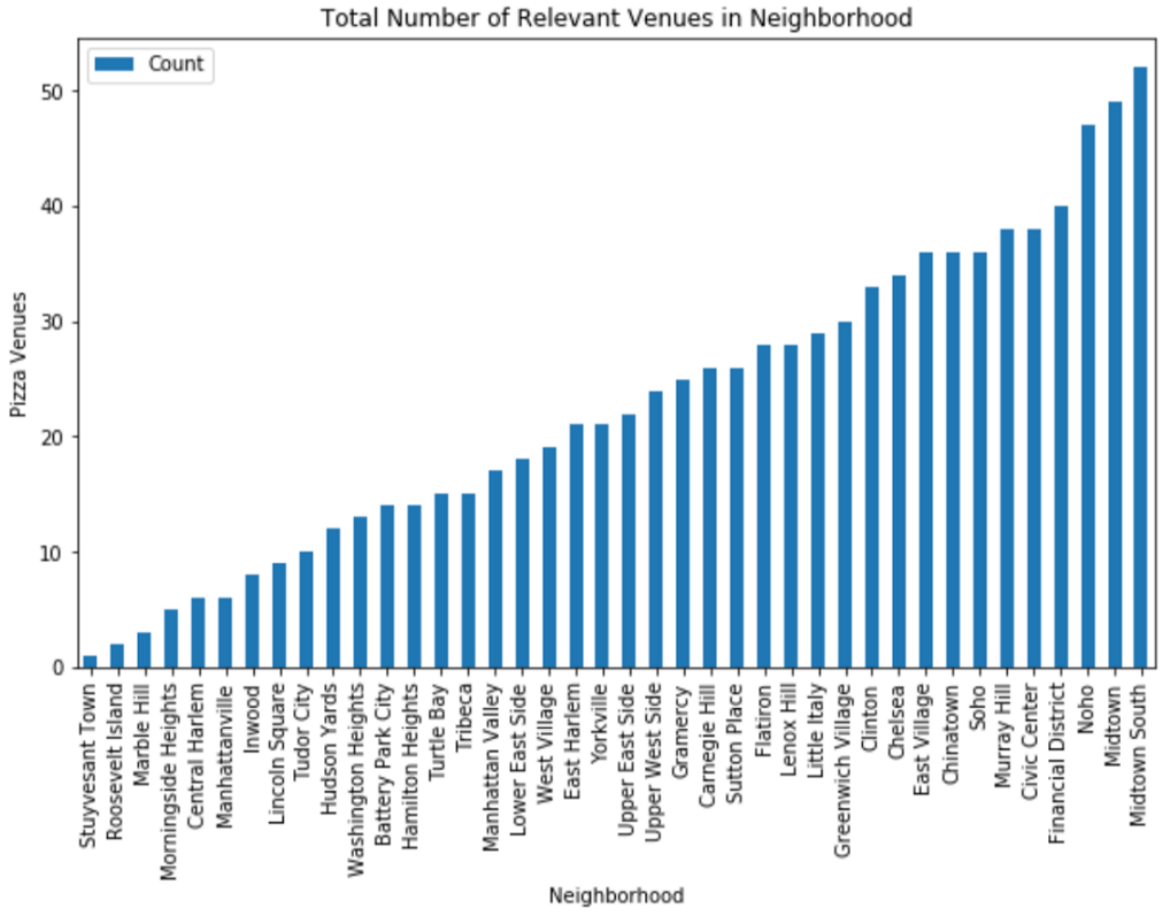
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.910660	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.910660	Pizza Hut Express	40.873201	-73.907861	Pizza Place
2	Marble Hill	40.876551	-73.910660	Kennedy Chicken & Pizza	40.874986	-73.909184	Fast Food Restaurant
3	Chinatown	40.715618	-73.994279	Scarr's Pizza	40.715335	-73.991649	Pizza Place
4	Chinatown	40.715618	-73.994279	Williamsburg Pizza	40.718303	-73.991046	Pizza Place

Next, I created a new dataframe summarizing the counts of the above dataframe by neighborhood, briefly checking our table to make sure it makes sense, putting into a table like so

:

	Neighborhood	Count
0	Stuyvesant Town	1
1	Roosevelt Island	2
2	Marble Hill	3
3	Morningside Heights	5
4	Central Harlem	6

I then proceeded to graphically analyze the dataframe using matplotlib, which generated the below bar chart depicting the count of relevant venues in each neighborhood.



I then used onehot encoding to create dummy variables for each venue type. While it might seem a bit off to use onehot, as the majority of these variables will be zeros, this is important because other venue types are included beside pizza places that might be relevant and we don't want to exclude them.

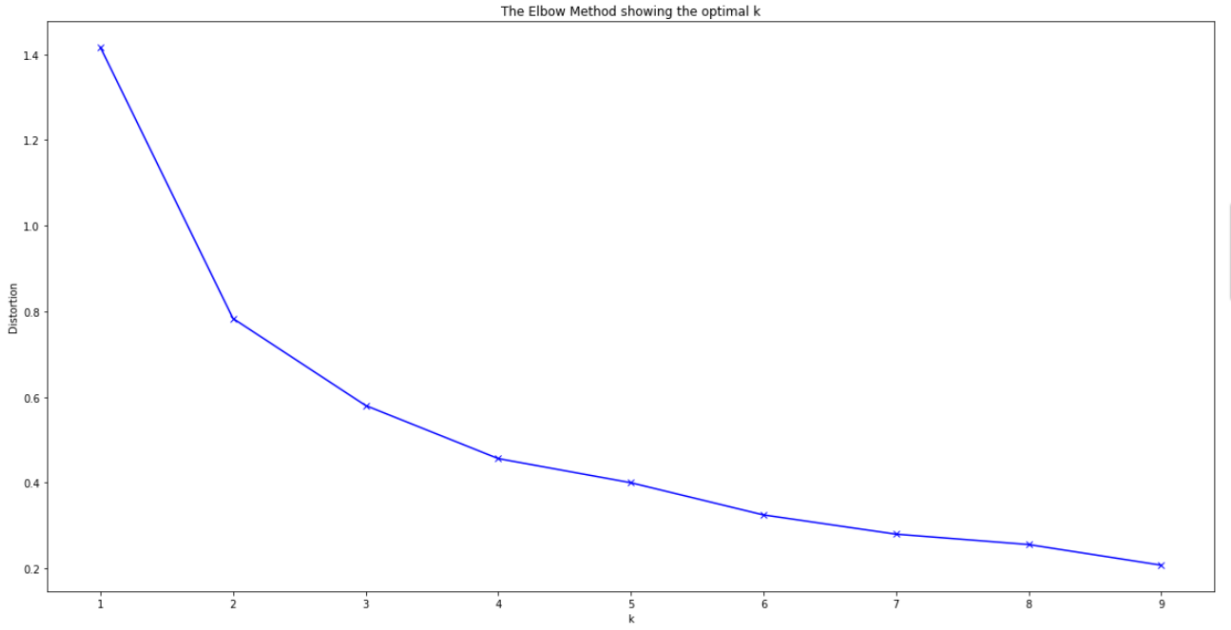
Next, I wrote a function to check the top 5 types of venues for each neighborhood—as expected, pizza place was the top result in almost all neighborhoods, but the top 5 also included Italian restaurants, gourmet shops, burger joints, pubs, American food, and grocery stores, among others. Our targets were neighborhoods with a large number of business that could be potential sales targets, especially where

italian restaurants were second to pizza places. To find out, I wrote another function to return the most common venues in each neighborhood and put them into a table.

]:

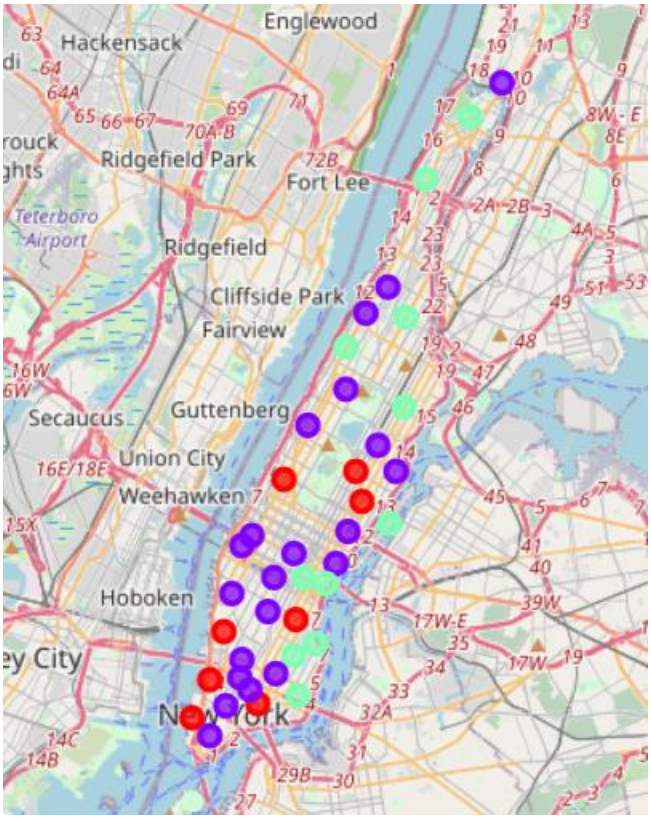
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Battery Park City	Pizza Place	Italian Restaurant	Gourmet Shop	Burger Joint	Deli / Bodega
1	Carnegie Hill	Pizza Place	Italian Restaurant	Café	Wine Bar	Bar
2	Central Harlem	Pizza Place	Grocery Store	Gluten-free Restaurant	Gay Bar	Food & Drink Shop
3	Chelsea	Pizza Place	Italian Restaurant	Bakery	American Restaurant	Bar
4	Chinatown	Pizza Place	Italian Restaurant	Bakery	Seafood Restaurant	Gourmet Shop

Finally, For collecting a set of target neighborhoods, rather than just picking one, we can use a clustering analysis. Using the data above, we can create a graph to determine the ideal k (i.e. number of clusters) at the 'elbow' of the graph and use that to cluster our neighborhoods. The elbow appeared to be around 3, so I chose 3 as my k value and ran the k-means clustering analysis accordingly.



Results

Now, we had completed our methodological analysis and the results were ready to visualize and analyze. The visualization of our cluster analysis is below:



In the cluster visualization above, cluster 1 is represented by red dots, cluster 2 by purple dots, and cluster 3 by green dots.

Based on these results, we can see that cluster 2 (see following screenshot for a table of the neighborhoods in cluster 2 and their accompanying most common venues) is likely to be the most ideal list from which to draw our potential locations. The list includes the top 3 results from the graph (NoHo, Midtown, Midtown South)--the only neighborhoods which have more than 40 pizza joints. Additionally,

the second most common category is Italian restaurants, which are likely to also be sales targets

```
In [ ]: #### Cluster 2 Results

In [44]: nyc_merged.loc[nyc_merged['Cluster Labels'] == 1, nyc_merged.columns[[1] + list(range(5, nyc_merged.shape[1]))]]
```

Out[44]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Marble Hill	Pizza Place	Fast Food Restaurant	Grocery Store	Gluten-free Restaurant	Gay Bar
4	Hamilton Heights	Pizza Place	Italian Restaurant	Café	Pub	Deli / Bodega
5	Manhattanville	Pizza Place	Italian Restaurant	Gourmet Shop	Gay Bar	Food & Drink Shop
9	Yorkville	Pizza Place	Bar	Italian Restaurant	Pub	Deli / Bodega
12	Upper West Side	Pizza Place	Italian Restaurant	Vegetarian / Vegan Restaurant	Bar	Breakfast Spot
14	Clinton	Pizza Place	Italian Restaurant	Bar	Café	Diner
15	Midtown	Pizza Place	Italian Restaurant	Pub	Café	American Restaurant
17	Chelsea	Pizza Place	Italian Restaurant	Bakery	American Restaurant	Bar
18	Greenwich Village	Pizza Place	Italian Restaurant	American Restaurant	Rock Club	Gluten-free Restaurant
22	Little Italy	Pizza Place	Italian Restaurant	Bakery	Bar	Deli / Bodega
23	Soho	Pizza Place	Italian Restaurant	American Restaurant	Bar	Lounge
25	Manhattan Valley	Pizza Place	Italian Restaurant	Sandwich Place	Cocktail Bar	Gay Bar
29	Financial District	Pizza Place	Café	Italian Restaurant	Gourmet Shop	Bar
30	Carnegie Hill	Pizza Place	Italian Restaurant	Café	Wine Bar	Bar
31	Noho	Pizza Place	Bar	Italian Restaurant	Hookah Bar	Bakery
32	Civic Center	Pizza Place	Italian Restaurant	Bar	Burger Joint	Food & Drink Shop
33	Midtown South	Pizza Place	American Restaurant	Bar	Sandwich Place	Italian Restaurant
34	Sutton Place	Pizza Place	Italian Restaurant	Dessert Shop	New American Restaurant	Sports Bar

Discussion and limitations

As you can see from the above analysis, New York City is a big place, so we focused on Manhattan to have more clear results. There is a high population density, so any of our target neighborhoods are likely to have enough potential customers, and we need only focus on our potential sales venues. The pizza joints appear from the graph to be highly concentrated in several neighborhoods--only three have more than 50 pizza joints: NoHo, Midtown, and Midtown South. These would appear to be the three most ideal locations for a sales office.

I chose to use the K means algorithm as part of this clustering study. When I tested the data using the Elbow method, I set the optimum k value to 3. The visualization of the clusters illustrates that they are somewhat scattered. However, Cluster 2 is likely the best fit. This is particularly true given that Cluster

3 includes as it's second-most category grocery stores--which compete with pizza joints and wouldn't be suitable for targeting our client's products, whereas Cluster 2 contains as it's runner up primarily Italian restaurants, which do fit our client's criteria.

Note that this report has a number of limitations. Note that very different approaches can be tried in clustering and classification studies, K- Means may not be the only or best one. Moreover, results may differ depending on which method is used. Additionally, only a small subset of the total data from Foursquare was used, owing to the focus on pizza joints. This was in part owing to the fact I used the limited Sandbox tier, limiting the number of calls. A more premium/paid analysis could be possible with expanded access. For more detailed and accurate guidance, the data set can be expanded and include a number of other variables, including the relative cost of offices, from places such as Zillow [5] which owing to data compatibility issues with the NYC geospatial repository data were excluded.

I also performed data analysis through this information by adding only static data--dynamically updating from APIs would likely yield more accurate results on a more frequent basis.

Finally, I ended the study by visualizing the data and clusters on the Manhattan map. In future studies, adding additional data would likely yield more informed results.

Conclusion

In this project, we have gone through the process of identifying a business problem, providing background, and locating the relevant data required. We then cleaned, prepared and extracted the data, performed visualization of this data which gave us a graphical picture of pizza restaurant count in Manhattan, NY. Finally, we performed a machine learning analysis using k-means clustering. This gave us 3 clusters of neighborhoods for the client to choose from, which our client can use to make their decision of where to locate a sales office.

Now that we have explored our data, we can see that NYC has many pizza places. However, the most pizza places are located in three neighborhoods, NoHo, Midtown, and Midtown South. These neighborhoods are all excellent targets for the client. Of course, of these three, the client should pick the one with the lowest cost, or one from cluster two that is adjacent but lower price.

Thank you for reading this report.

References

1) [Picnic] (https://www.hellopicnic.com/press)
2) [NYU Geospatial Data Repository - NYC Neighborhoods]
(https://geo.nyu.edu/catalog/nyu_2451_34572)
3)[Foursquare] Foursquare API
4) [Folium] (https://python-visualization.github.io/folium/)
5) [Zillow Housing Data] (https://www.zillow.com/research/data/)