Capstone Project - Battle of the Neighborhoods

Introduction

My client is an Italian restaurant entrepreneur. My client's goal is to open a new Italian restaurant in either Toronto or New York City. The area within the 2 cities has been narrowed down to Manhattan, and the downtown vicinity of Toronto. My client hires me as the data scientist and is interested in analyzing by mapping and clustering all restaurant venues in both cities to visualize the overall competition, similarities, and types of restaurant in each area. My client also wants to look at frequencies and most common of restaurant type categories as well and take this all into account for making a final decision. Also, we need to visualize just Italian restaurants on a map to analyze the competitive environment. Manhattan is world renowned for little Italy and many low to high price Italian restaurants. Toronto also has its own Little Italy area with many fabulous low- and high-priced Italian restaurants. In the past, the entrepreneur has had much success in going to areas that are like hidden gems. These hidden gems areas of town will have a lower volume of competition in major cities.

Background

NYC is the most populated city in the United States. NYC is also the most densely populated city in the United States. New York city is comprised of 5 major boroughs Brooklyn, Queens, Manhattan, Bronx, and Staten Island. New York does have a Little Italy section in Manhattan as well which my client wants to avoid due to heavy competition. My client chose Manhattan since it has the population with the most income per capita in the city, and also has the most tourists.

Toronto is the most populated city in Canada. Toronto is also the most densely populated city in Canada. Toronto is comprised of 6 boroughs Etobicoke, North York, Scarborough, York, East York, and Toronto. Toronto also has a little Italy section in the Toronto borough as well which my client wants to avoid due to heavy competition. My client chose the Toronto downtown area since this area has the most tourists and foot traffic.

Either of these cities can present a prime opportunity, however selection of the area using my client's critical decision factors of lower competition for Italian food, but a solid number of overall restaurant venues is key to my client's approach. This approach has worked before in other major cities.

My client has opened up a handful of successful Italian restaurants in major cities. My client believes in targeting areas of low penetration for Italian cuisine.

Data

Toronto

The project will focus a large portion of the analysis on Foursquare, Wikipedia, and the NYC dataset from the prior labs. My IBM free cloud hours are actually used up or additional data would have been brought in. That will be listed as possible next steps in the analytical approach below.
The specific link to the data sets are https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Brought in Geospatial Data for Toronto from https://cocl.us/Geospatial.data

Pulled in any neighborhood that contained Toronto in the string.

The wiki Toronto kept changing titles to Post Code/Postal Code, and Neighborhood/Neighbourhood so I adjusted for this each time I re-ran seem like something changed.

Used Foursquare to analyze the category type restaurants only took a while to figure out how to do an API call for a specific category.

Used Foursquare to also analyze and narrow the results for the API call down to just Italian restaurants.

NYC

For NYC data used https://cocl.us/new_york_dataset that already had longitude and latitude in there so no secondary data set was needed.

Then used Foursquare API to analyze just the specific restaurant category.

Used Foursquare to also analyze and narrow the results for the API call down to just Italian restaurants.

The Foursquare API will be used to run a similar analysis to our labs to analyze the least frequent distribution of neighborhoods of Italian restaurants, and that are above the overall mean in count of overall venues.
We can use the Explore or Search end points for the analysis in the Foursquare API.

Methodology

Process Flow/Machine Learnings Usage

Toronto

Imported all Libraries and primarily used pandas, json, geopy, requests, matplotlib, sklearn, and folium.

Pulled in Toronto data set from wiki and read into pandas DF using read_html.

Each time I did this something changed like the spelling of postal code to post code, or neighborhood to neighborhood. I adjusted by renaming these each time.

I analyzed the shapes and counts and removed any rows that had NA as borough.

I brought in the geospatial website to bring in the longitude and latitude.

I then merged this file with the wiki data frame and validated the shape and data types.

I then looked at boroughs that only had Toronto contained in the string since my client narrowed his decision down to Downtown Toronto.

Then a key problem I was running into was that when I narrowed the Foursquare category ID down to only restaurants venue types at the end when I ran the KMeans cluster I would error out with an error message saying the length did not match when I joined up my final cluster table. So, after a few hours of analysis I figured out I had to drop a few columns from this early step to be able to match up in the KMeans cluster and run without an error so the cluster labels matched the earlier data set in rows when merged. Each time I ran this I had to change which rows I dropped. This last time I dropped the last row that had a post office box, and Roselawn which had no outputs. So, I dropped these 2 rows which did not produce any restaurants to be able to run the KMeans cluster at the end.

Then I created a map of the Toronto neighborhoods I was going to analyze in Foursquare.

Then I pulled in the Foursquare venues using the explore functions with the category only restaurants.

I setup a for loop to get the information into a data frame.

Then I analyzed these restaurants only venues for my string that contained Toronto neighborhoods which I consider downtown Toronto.

This is the area my client wanted to map, analyze competition, most common venues, and clustering for similarities.

Then I reviewed the shapes and counts of the venue data frame I created from the Foursquare venues.

Then I used one hot encoding since my client wanted to analyze all of the restaurants at first by frequency and commonality to see competition in the area.

I analyzed the shape, and performed a group by neighborhood to pull in the data my client wanted to review on competition.

Then I pulled in a top 10 restaurant venue frequencies.

Then I created a data frame to analyze the competition with the 10 most common restaurant venues.

This will be one of the statistics that are used in the ultimate decision.

Then as another data point for a final decision we performed the machine learning KMeans algorithm. I used 5 cluster labels. This was used since my client wanted to see the clusters of the restaurants to analyze their similarities.

This was the part I spent a few hours trying to fix the error on which was driven by the merge of data having one or 2 different rows in data difference. I figured out eventually I had to drop these rows so the 2 data frames being merged were the same length for the KMeans.

Then I graphed the KMeans clusters of restaurants for my client to visualize similarities.

Next, I only pulled in a category for Italian restaurants only from Foursquare.

I used the same process from earlier using a for loop and API call to turn this into a data frame.

This was completed since my client wanted to visualize the Italian restaurants in the downtown area. Also, to look at volumes of Italian restaurants only by neighborhood for more concrete view on competition and to visualize on a map

NYC

For NYC I pulled in the data which already had all relevant information so no web scraping was needed from Wiki.

I pulled in the features and analyzed the format.

I assigned column names and created a data frame using a for loop.

Then I grabbed only the Manhattan borough since this why the second choice my client narrowed his choices down to.

Then I created a map of Manhattan with the neighborhoods for visualization.

Then I grabbed the Foursquare data just for restaurants category just like the Toronto analysis.

I used the API call and created that into a pandas data frame.

Then I looked at the shape and counts.

Followed the same process as Toronto and used One Hot encoding to get an idea of restaurant frequency and competition.

Then I created a data frame to analyze the competition with the 10 most common restaurant venues.

This will be one of the statistics that are used in the ultimate decision

Then as another data point for a final decision we performed the machine learning KMeans algorithm. I used 5 cluster labels. This was used since my client wanted to see the clusters of the restaurants to analyze their similarities.

Then I graphed the KMeans clusters of restaurants for my client to visualize similarities.

Next, I only pulled in a category for Italian restaurants only from Foursquare.

I used the same process from earlier using a for loop and API call to turn this into a data frame.

This was completed since my client wanted to visualize the Italian restaurants in the downtown area. Also, to look at volumes of Italian restaurants only by neighborhood for more concrete view on competition and to visualize on a map

Methodology calculations used algorithms used and why

Since my client was very concerned with competition in neighborhoods, overall restaurant venues, and similarities within each neighborhood I chose one hot encoding to get an idea of what the most frequent or top 10 results were for each neighborhood. Then I used the machine learning KMeans clustering to get an idea of similarities within each neighborhood. These were the key results my client needed to make his final decision on where to locate his next restaurant venture.

Results

Toronto

Once I cleansed my data set I had 39 rows and 5 columns that including the longitude and latitude.

I had to drop 2 additional rows that did not have restaurant results to be able to cluster.

There were 1365 restaurant venues in downtown Toronto area.

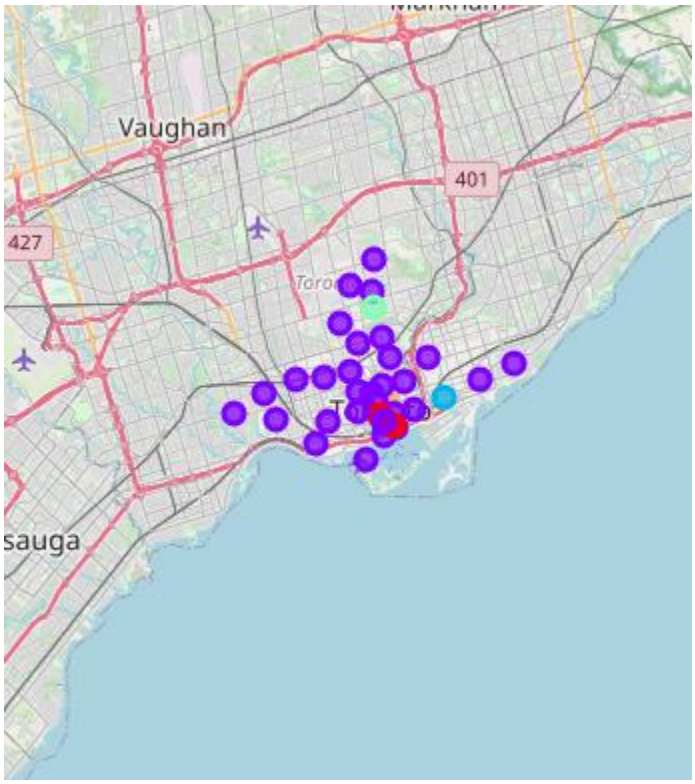
I maxed these at 100 restaurant venues per neighborhood.

Commerce Court, First Canadian Place, Toronto Dominion Center hit the max amounts of 100.

There were 89 unique categories of restaurants

When looking at the most common venues Italian restaurants placed in the top 5 in 13 of the 37 neighborhoods, and in the top 10 in 20 of 37 neighborhoods.

In the KMeans cluster 0 has 2 neighborhoods, cluster 1 has 32 neighborhoods, Clusters 2 through 4 each have only 1 neighborhood. So, this analysis illustrates that 86% of the clusters are together in one cluster.



For Italian restaurants only there are 202 venues. The top 5 neighborhoods are Commerce Court at 26, Toronto Dominion Center at 24, First Canadian Place at 20, St James Town at 20, and Central Bay Street as 16.

These amounts were in line with the top volumes of overall restaurant venues.

NYC

Once I cleansed my data set, I had 40 rows and 4 columns that including the longitude and latitude.

There were 2787 restaurant venues in the Manhattan area.

I maxed these at 100 restaurant venues per neighborhood.

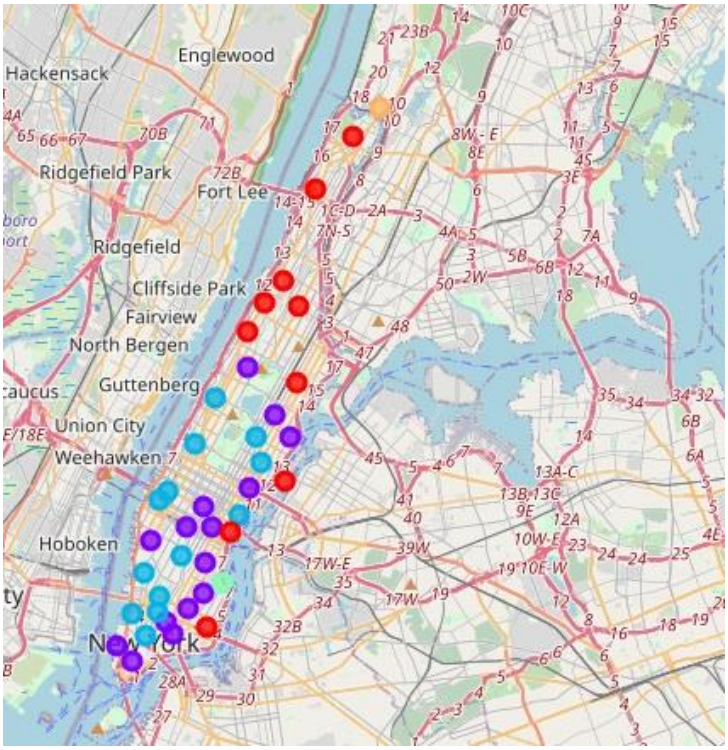
Chinatown, Clinton, East Village, Financial District, Greenwich Village, Lenox Hill, Little Italy, Midtown, Midtown South, Noho, Soho, and West Village all maxed out at 100.

There were 126 unique categories of restaurants.

When looking at the most common venues Italian restaurants placed in the top 5 in 22 of the 40 neighborhoods, and in the top 10 in 23 of 40 neighborhoods.

NYC appears to have many more Italian restaurants ranked higher and 15 neighborhoods even had the number 1 most common ranking.

In the KMeans cluster 0 has 10 neighborhoods, cluster 1 has 15 neighborhoods, Clusters 2 has 13 neighborhoods, clusters 3 and 4 each have only 1 neighborhood. These clusters are much more spread out than Toronto primarily between the 3 clusters.



For Italian restaurants only there are 893 venues. The top 5 neighborhoods are Little Italy at 57, Greenwich at 54, Soho at 53, Financial District at 50, and Nohu at 47.

These amounts were in line with the top volumes of overall restaurant venues.

Discussion

Toronto and Manhattan had a similar number of neighborhoods, however the Manhattan area is 3 times the square miles as Toronto so this has to be taking into account as well. I would expect Manhattan to have more venues. They had double the amount of restaurant venues along with more unique restaurants but due to the size that is to be expected. Toronto's neighborhoods were much more similar and 86% of those neighborhoods were in one KMeans cluster in terms of restaurant venue similarity. Manhattan had more differentiation between the neighborhoods with the largest cluster only making up 38% of neighborhoods. So, the neighborhoods had much more differentiation in terms of restaurant venues when compared to downtown Toronto. When using the category type Italian restaurants Toronto had 35% in the top 5 restaurant venues, and 54% in the top 10. NYC had 55% in the top 5 along with 58% in top 10. So, the top 10 were similar in terms of proportional percentages. However, NYC had a whopping 15 neighborhoods with an Italian restaurant as the number most common 1 venue. Toronto had 0 neighborhoods where an Italian restaurant made up the number 1 spot. When analyzing total number of Italian restaurants both cities followed the similar trend of the venues with the highest overall restaurant venues also had the highest overall number of Italian restaurants for the most part. Manhattan had many more Italian restaurant venues, however the size is roughly 3 times larger in terms of square miles. I would still conclude that Italian restaurants are more popular in Manhattan since the discrepancy in volumes is even larger than the square mile difference.

With my client wanting an area with lower competition and a solid number of overall venues we can drill into the data and make some recommendations.

First, my client has had success in large cities in low competition areas, however my client always takes into consideration the overall number of venues as well so they do not set up shop in an area that doesn't have many people going out to eat. So it's a fine line between low competition and a decent number of overall venues of what fits the bill for my client.

So basically, we start by ruling out any top 5 ranked most common Italian restaurants venues in neighborhoods. For Toronto the clustering doesn't really show much differentiation, while NYC shows a lot of dis-similarity between neighborhoods. I would use the clusters that have Italian restaurants in them as being popular and choose the like neighborhoods with Italian restaurants that are lower than top 5 most common ranked or not on the list. Ruling out the top 5 removes almost 50% of the neighborhoods.

When with the remaining venues I look for a median to higher volume total venue area with a lack of Italian restaurant competition. For NYC I exclude clusters 0 which has no Italian restaurants, cluster 2 which is highly competitive for Italian restaurants, and cluster 3 and 4 which have limited information. That leaves me with cluster 1 as my selection. I rule out the higher ranked Italian restaurants in that cluster that leaves me with 6 possibilities that meet my clients criteria Chinatown, Midtown, Murray Hill, Manhattan Valley, Battery Park, and Midtown South.

Toronto has high homogeneity and a lack of Italian restaurants in top positions. For the Toronto clusters I exclude the 4 clusters that had 2 or less results, and choose the largest cluster. In the largest cluster I eliminate all of the neighborhoods that have an Italian restaurant rated in the top 10 most common venue per my client's requests. This leaves me with 14 Toronto neighborhoods Richmond, Queen's Park, The Beaches, Dufferin, Little Portugal, Lawrence Park, Davisville North, Forest Hill, The Annex, Kensington Market, Summerhill West, CN Tower, Rosedale, Church and Wellesley, and Studio District. This means in this growing diverse market there should be a large opportunity to open a new Italian restaurant.

There will need be next steps in this final analysis of the location will be a deep dive into income per capita, rents, crime, safety, public transportation on these final neighborhoods in NYC and Toronto to make a final decision. The KMeans clustering, mapping, and Foursquare research in Italian and all restaurants categories is not enough to deliver a final answer to my client. We have narrowed the list from 77 neighborhoods down to 20 possibilities which may meet my client's criteria.

Conclusion

New York and Toronto seem to be dis-similar in terms of restaurant venues. Toronto has high homogeneity within restaurant venues shown in clustering, while NYC does not. NYC has many neighborhoods with Italian restaurants as the number 1 most popular venue, Toronto does not. Following my client's instructions, and using KMeans clustering allows me to give my client a narrowed down for further research in income per capita, crime, transportation, tourists, foot traffic, ethnic make-up and place his restaurant in a popular growing area with a lack of good Italian restaurants.