Analyzing Neighborhoods to Identify Expansion Opportunities

An Upscale Italian Restaurant Chain Explores International Expansion into North America

Background

- Who

  An Upscale Italian Restaurant Chain based in Milan, Italy

- What

  The chain is exploring international expansion into the North American market. New York City, USA and/or Toronto, Canada have been identified as the best fit for the initial expansion locations.

- How

  The expansion strategy for the chain focuses on identifying neighborhoods that are similar to the Milan neighborhoods that the chain and other Italian restaurants currently operate in.

Background

- Problem

  An intelligent methodology for identifying New York City and Toronto neighborhoods and analyzing them in relative to relevant Milan neighborhoods is needed.

- Solution

  A methodology that utilizes Python to gather neighborhood information from sources on the internet and then utilizes machine learning techniques to identify the New York City and Toronto neighborhoods that are best fits for future expansion based on the chain's expansion strategy.

- Stakeholders

  The stakeholders who will benefit the most from this project are the members of the chain's management team tasked with driving the North America expansion program.

Data

- Data Sources

  The raw data for the project was scaped from various sources on the internet. Wikipedia and the Coursera website provided neighborhood information. Geopy was utilized to obtain geospatial coordinates for the neighborhoods. The Foursqaure API provided information on the commercial venues in each neighborhood.

Out[197]:

	City	Borough	Neighborhood	Latitude	Longitude
0	Milan	Centro storico	Brera	45.473479	9.188408
1	Milan	Centro storico	Centro Storico	45.447112	9.094054
2	Milan	Centro storico	Conca del Naviglio	45.458560	9.177745
3	Milan	Centro storico	Guastalla	45.458252	9.200023
4	Milan	Centro storico	Porta Sempione	45.477128	9.170598
5	Milan	Centro storico	Porta Tenaglia	45.477821	9.181593
6	Milan	Stazione Centrale, Gorla, Turro, Greco, Cresce...	Adriano	45.513572	9.251202
7	Milan	Stazione Centrale, Gorla, Turro, Greco, Cresce...	Crescenzago	45.509219	9.247484
8	Milan	Stazione Centrale, Gorla, Turro, Greco, Cresce...	Gorla	45.504945	9.224539
9	Milan	Stazione Centrale, Gorla, Turro, Greco, Cresce...	Greco	45.502184	9.211233
10	Milan	Stazione Centrale, Gorla, Turro, Greco, Cresce...	Loreto	45.484535	9.215276

(103, 5)

Out[204]:

	City	Borough	Neighborhood	Latitude	Longitude
0	Toronto	North York	Parkwoods	43.753259	-79.329656
1	Toronto	North York	Victoria Village	43.725882	-79.315572
2	Toronto	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	Toronto	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	Toronto	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
5	Toronto	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242
6	Toronto	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	Toronto	North York	Don Mills	43.745906	-79.352188
8	Toronto	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	Toronto	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937

(10478, 7)

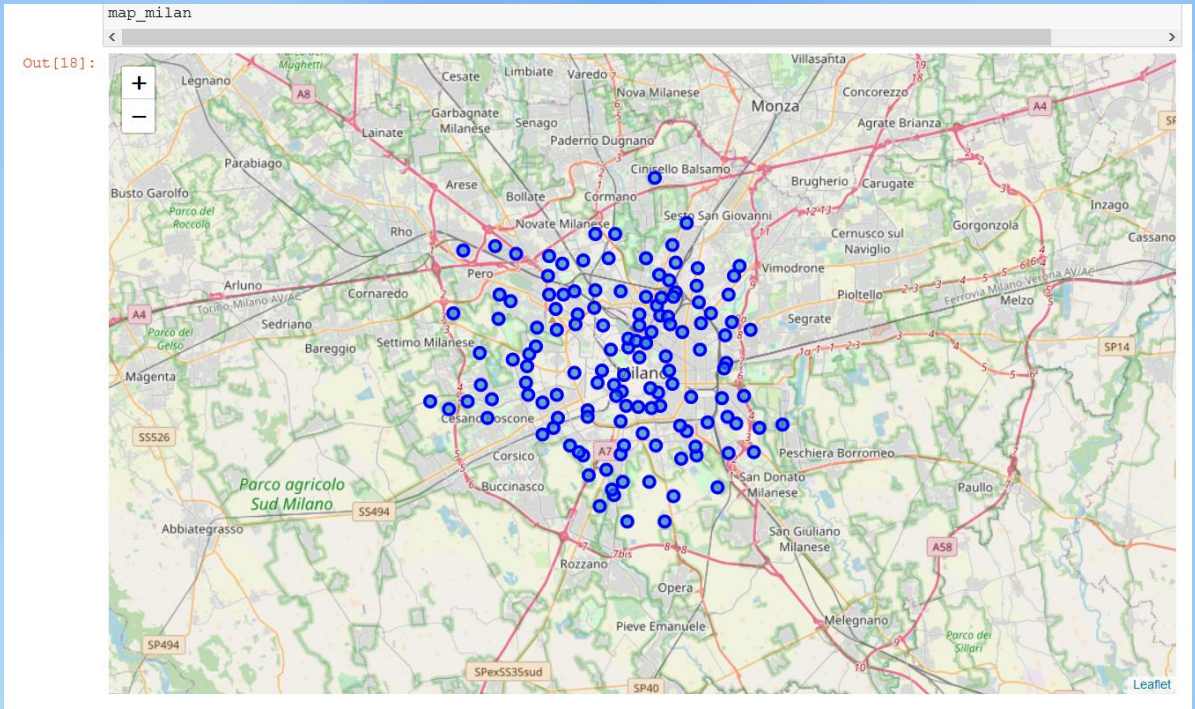
Out[43]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop
4	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop

Data

- Wrangling

  The neighborhood data was brought from Wikipedia into the notebook through Beautiful Soup which worked without any issues. The geospatial data from Geopy was generally very good however inaccurate coordinates for a few neighborhoods required those neighborhoods to be eliminated from the analysis. The commercial venue information from the Foursquare API was great and easily fit into the analysis.



Methodology

With the data collected and wrangled, it is time for the analysis which will follow an ordered methodology:

1. Build a dataframe for each city populated with the commercial venue information for each neighborhood.
2. Concatenate these dataframes into one master neighborhood venue dataframe and to similarly concatenate individual city data to build a master neighborhood location dataframe.
3. Work with the master dataframes and k-Means Clustering to build a new combined master dataframe that lists the ranked top ten commercial venue types for each neighborhood and subdivides the neighborhoods into clusters based on their commercial venue profiles. Silhouette analysis will be applied to the data in order to identify the best k to utilize during the final k-Means Clustering. in order to provide the most meaningful analysis.
4. Analyze this combined master dataframe to identify the subset of Toronto and New York City neighborhoods that are most promising for future expansion.
5. Final analysis of the subset of neighborhoods to identify the top neighborhoods most appropriate for supporting the upscale-style of Italian restaurant featured by the chain. This will be accomplished by utilizing the Foursquare API to find all Italian restaurants with a price tier of 4 (highest average price for an entree) in the subset neighborhoods.

Analysis

- Master Neighborhood Venue Dataframe

Contains information on the commercial venues in every neighborhood of all three cities.

The concatenate operation for joining the data from the three cities was necessary to create the master dataframe however it took almost 7 hours of processing time and exploded the size of the notebook file to over 52 GB. It will be beneficial to find a more efficient method for performing this operation to reduce the file to a more manageable size.

```
print(total_grouped.shape)
total_grouped.head(10)
```

```
(532, 503)
```

Out[75]:

	Neighborhood	Abruzzo Restaurant	Accessories Store	Adult Education Center	African Restaurant	Airport	Airport Service	American Restaurant	Amphitheater	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum
0	Acquabella	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Adriano	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Affori	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Arzaga	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Assiano	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	Baggio	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	Barona	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	Bicocca	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	Boffalora	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	Boldinasco	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Analysis

- Silhouette Analysis

  Used to identify the optimum k value to use in k-Means Clustering. Can range from -1 to 1 with higher numbers being better.

  The analysis was run on k values between 3 – 10:

K	Silhouette Value	K	Silhouette Value	K	Silhouette Value	K	Silhouette Value
3	0.031	5	0.048	7	0.050	9	0.063
4	0.043	6	0.045	8	0.057	10	0.042

  k = 9 has the highest silhouette value thus it was used in the k-Means Clustering

- K-Mean Clustering

  Machine Learning method used to segment the neighborhoods from the three cities into meaningful clusters of similar neighborhoods.

```
In [85]: # Performing the final k-Means Clustering

         # set number of clusters
         k = 9

         # run k-means clustering
         kmeans_total = KMeans(n_clusters = k, random_state = 0).fit(total_grouped_clustering)

         kmeans_total.labels_[0:10]

Out[85]: array([3, 1, 1, 0, 3, 1, 1, 1, 1, 1])
```

Analysis

- Combined Master Dataframe

Created from the k-Means Cluster analysis.

```
print(total_merged.shape)
total_merged.head(10)
```

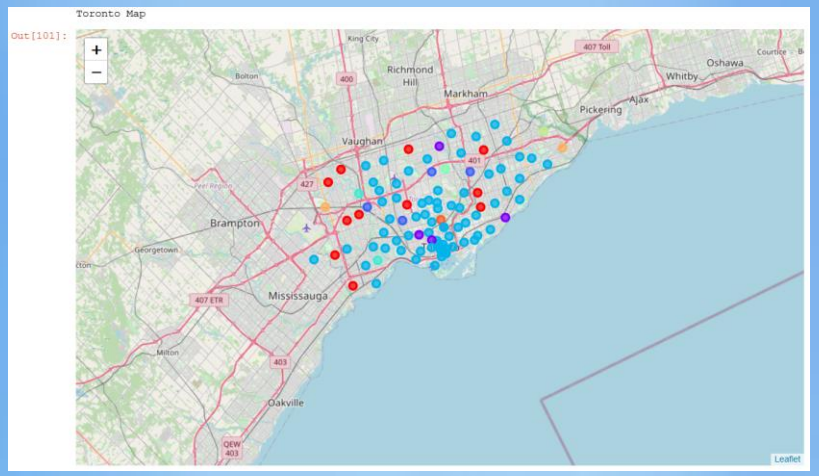
```
(538, 16)
```

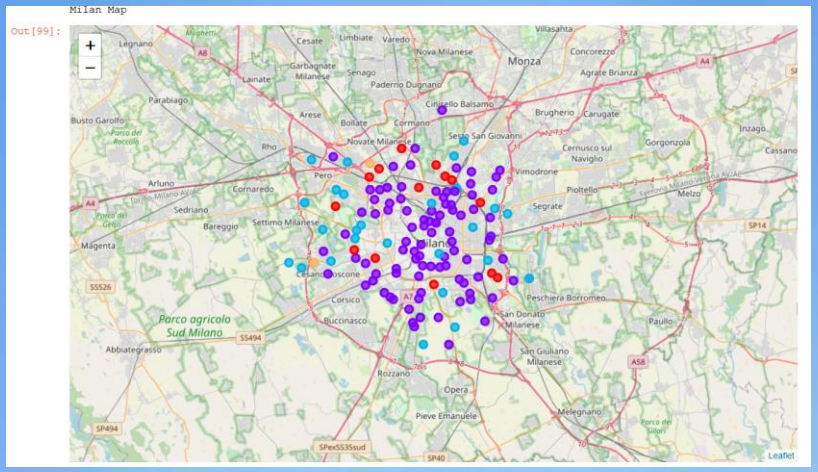
Out[93]:

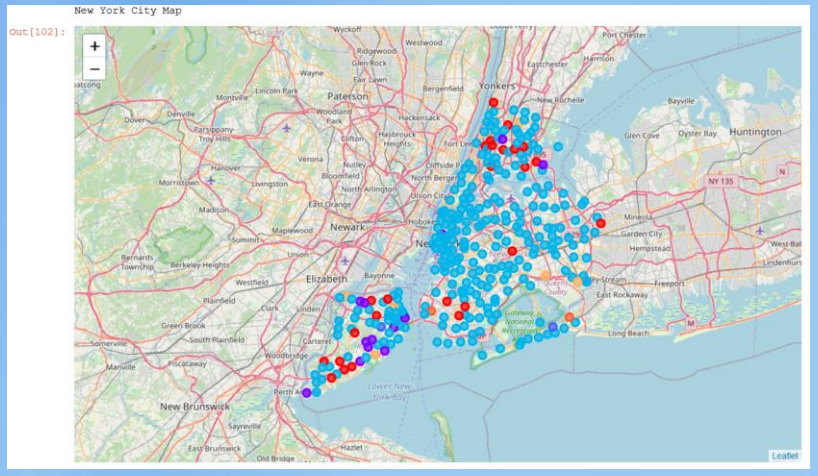
	City	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	
0	Milan	Centro storico	Brera	45.473479	9.188408	1	Italian Restaurant	Ice Cream Shop	Hotel	Japanese Restaurant	Café	Pizza Place	Restaurant	C
1	Milan	Centro storico	Centro Storico	45.447112	9.094054	1	Bar	Pizza Place	Italian Restaurant	Park	Shopping Mall	Cupcake Shop	Fried Chicken Joint	Mov
2	Milan	Centro storico	Conca del Naviglio	45.458560	9.177745	1	Italian Restaurant	Café	Ice Cream Shop	Pizza Place	Cocktail Bar	Seafood Restaurant	Bistro	
3	Milan	Centro storico	Guastalla	45.458252	9.200023	1	Italian Restaurant	Tram Station	Pizza Place	Ice Cream Shop	Restaurant	Kebab Restaurant	Bar	
4	Milan	Centro storico	Porta Sempione	45.477128	9.170598	1	Italian Restaurant	Pizza Place	Cocktail Bar	Japanese Restaurant	Ice Cream Shop	Bakery	Gastropub	Med
5	Milan	Centro storico	Porta Tenaglia	45.477821	9.181593	1	Italian Restaurant	Japanese Restaurant	Café	Chinese Restaurant	Ice Cream Shop	Pizza Place	Wine Bar	
6	Milan	Stazione Centrale, Gorla, Turro, Greco, Cresce...	Adriano	45.513572	9.251202	1	Italian Restaurant	Ice Cream Shop	Soccer Field	Performing Arts Venue	Trattoria/Osteria	Toy / Game Store	Cultural Center	Fis
7	Milan	Stazione Centrale, Gorla, Turro, Greco, Cresce...	Crescenzago	45.509219	9.247484	1	Italian Restaurant	Café	Ice Cream Shop	Bus Stop	Metro Station	Supermarket	Pharmacy	H
8	Milan	Stazione Centrale, Gorla, Turro, Greco, Cresce...	Gorla	45.504945	9.224539	1	Italian Restaurant	Pizza Place	Hotel	Plaza	Park	Spa	Brewery	C
9	Milan	Stazione Centrale, Gorla, Turro, Greco, Cresce...	Greco	45.502184	9.211233	3	Supermarket	Seafood Restaurant	Chinese Restaurant	Café	Hotel	Plaza	Pet Store	C

Analysis

- Cluster Maps for Each City







Analysis

- **Subset of Toronto and New York City Neighborhoods Most Promising for Expansion**

The k-Means Clustering produced 9 clusters of similar neighborhoods. Interestingly, one cluster (Cluster 1) was dominated by neighborhoods that had Italian restaurants as a highly common type of commercial venue in the neighborhood. Furthermore, this cluster contained 68% of the Milan neighborhoods thus indicating that the 17 New York City and Toronto neighborhoods contained in the cluster were similar to a majority of Milan. After analyzing all nine clusters, it was determined that this cluster contained the subset of Toronto and New York City neighborhoods that are most promising for expansion based on the chain's expansion strategy.

City	Neighborhood	City	Neighborhood	City	Neighborhood
Toronto	Christie	New York City	Greenwich Village	New York City	Richmond Town
Toronto	Bayview Village	New York City	Mariner's Harbor	New York City	Shore Acres
Toronto	Birch Cliff/Cliffside West	New York City	Great Kills	New York City	Elm Park
Toronto	Univ. of Tornto/Harbord	New York City	Tottenville	New York City	Egbertville
New York City	Belmont	New York City	Old Town	New York City	Lighthouse Hill
New York City	Edgewater Park	New York City	New Dorp Beach		

Analysis

- **Identifying the Top Neighborhoods for Upscale Italian Restaurant Expansion**

The foursquare API was utilized to further analyze the Cluster 1 subset by researching which of the member neighborhoods currently support upscale (price tier = 4, i.e. highest average entrée price) Italian restaurants.

City	Neighborhood	Price Tier 4 Italian Restaurants	City	Neighborhood	Price Tier 4 Italian Restaurants	City	Neighborhood	Price Tier 4 Italian Restaurants
Toronto	Christie	1	New York City	Greenwich Village	11	New York City	Richmond Town	1
Toronto	Bayview Village	1	New York City	Mariner's Harbor	2	New York City	Shore Acres	3
Toronto	Birch Cliff/Cliffside West	1	New York City	Great Kills	2	New York City	Elm Park	2
Toronto	Univ. of Tornto/Harbord	1	New York City	Tottenville	1	New York City	Egbertville	1
New York City	Belmont	18	New York City	Old Town	4	New York City	Lighthouse Hill	1
New York City	Edgewater Park	4	New York City	New Dorp Beach	2			

This analysis determined that the two New York City neighborhoods of Belmont and Greenwich Village contained the vast majority of the upscale (price tier 4) Italian restaurants in the subset. Thus, Belmont and Greenwich Village are the most promising candidates by following the chain's strategy for future expansion.

City	Neighborhood
New York City	Belmont
New York City	Greenwich Village

Conclusion

The goal of this project was to identify neighborhoods in Toronto, Canada and New York City, USA that fit the international expansion strategy for an upscale Italian restaurant chain based in Milan, Italy.

By utilizing neighborhood information from the internet and venue information from the Foursquare API, we were able to apply k-Means Clustering to intelligently cluster the neighborhoods of the three cities. This clustering clearly identified a subset of four Toronto neighborhoods and thirteen New York City neighborhoods that are similar to Milan neighborhoods where Italian restaurants thrive. The foursquare API was utilized to further analyze this subset by researching which of the member neighborhoods currently support upscale (price tier = 4) Italian restaurants.

This analysis determined that the two New York City neighborhoods of **Belmont** and **Greenwich Village** are the best fits for future expansion based on the company's international expansion strategy.