Comparing Neighborhoods of Toronto And NYC to Predict Restaurant Success

Keegan McCleary-Sharpe

December 3, 2020

Introduction -

This project showcases the ability to use data science toolkits on real-life problems. Geographical data can be scraped from the web, segmented, and subsequently applied to a relevant business problem. The analysis for this particular problem will be completed within Python, using primarily the Pandas, Geocoder, and Folium libraries, as well as Scikit-learn for K Means clustering.

Business Problem:

A potential client has established the following question:

"We have a series of restaurants that perform very well throughout several neighborhoods of New York City. We are looking to expand our operations into Canada, particularly the Toronto area, however, we are unsure which neighborhood we should target in order to access a similar clientele to our New York chain."

Problem Analysis:

Based on the question, it stands to reason that we can address this business problem using geospatial data and clustering the neighborhoods of New York City and Toronto. We can look first at the neighborhoods of New York City that contain existing locations of the client's restaurant, and we can characterize those neighborhoods based on surrounding venues.

Once characterized, we can do the same for the neighborhoods of Toronto, and then perform unsupervised clustering on the neighborhoods of both cities. If two neighborhoods cluster together, they will have similar venues throughout the neighborhoods, and for this we can infer similar clientele. We can then highlight the neighborhoods of Toronto that cluster with New York City neighborhoods with the client's successful locations.

Data -

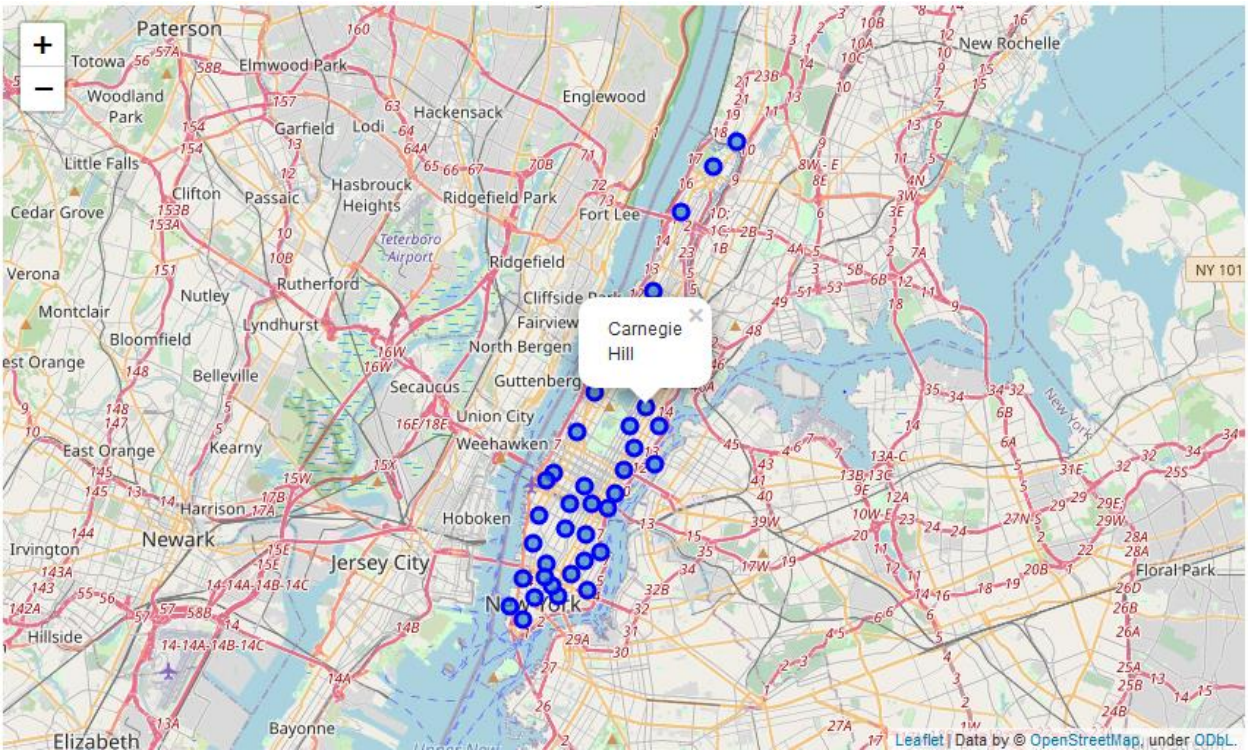
To perform the above analysis, we will need a dataset that outlines the five boroughs of New York City, as well as the neighborhoods within them. We have obtained that dataset in the form of a JSON file from IBM's server. We will also need a similar dataset for the Toronto area, which we can scrape from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

After we have loaded, cleaned, and began exploratory analysis on the neighborhood data for each city of interest, we will also need to access Foursquare's API. This API will allow us to identify the most frequent type of venue within a neighborhood and identify the characteristics of a neighborhood that allow the client's locations to be successful.

Methodology –

To begin the analysis, it was necessary to clean the data received. Firstly, it was necessary to trim the New York City data to specifically contain only the neighborhoods of Manhattan, as the client is based out of Manhattan.

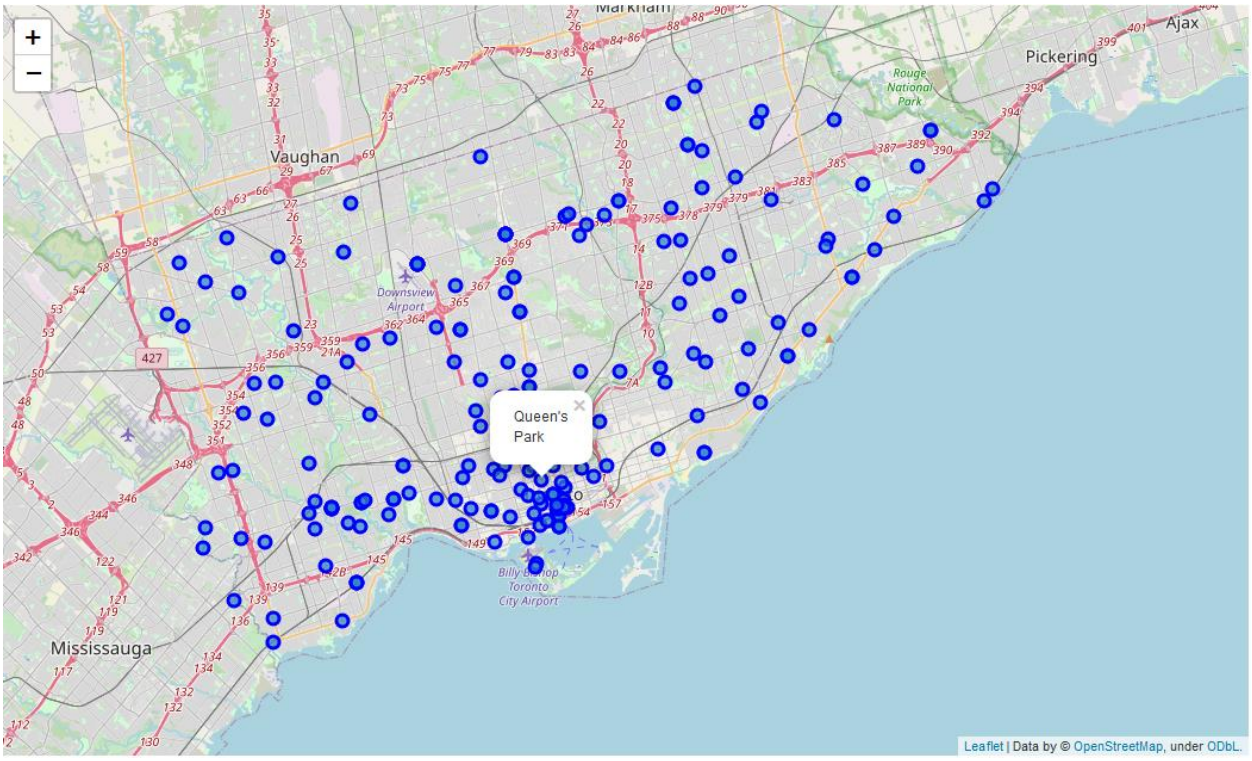
With the data cleaned, exploratory analysis could be performed. A visualization of the geospatial data was performed to become familiar with the location provided by the client. The neighborhoods were plotted along an interactive map to represent the geographical organization of the data. The client's most successful location is contained in the Carnegie Hill neighborhood, central to the area analyzed.



With the regions of interest visualized, it was then necessary to characterize those regions. This was done using Foursquare's API; the neighborhoods of Manhattan were passed through this API and the 100 nearest venues to each neighborhood's latitude and longitude were returned. The most common venue types were calculated based on relative frequency.

	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	Park	Hotel	Gym	Coffee Shop	Memorial Site
1	Carnegie Hill	Coffee Shop	Café	Bookstore	Italian Restaurant	Gym / Fitness Center
2	Central Harlem	African Restaurant	Chinese Restaurant	Bar	Seafood Restaurant	American Restaurant
3	Chelsea	Coffee Shop	Bakery	American Restaurant	Art Gallery	Café
4	Chinatown	Chinese Restaurant	Cocktail Bar	Bakery	Dessert Shop	American Restaurant

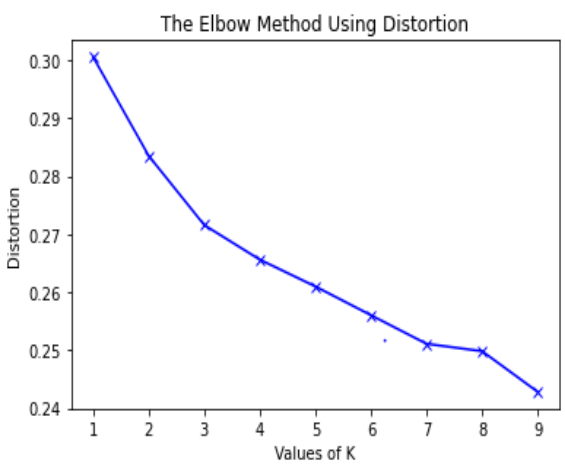
Finding the most common venue types of a neighborhood allowed the areas to be characterized by the venues contained, and for this, one could infer the type of clientele that a neighborhood contains and/or attracts. In order to avoid biasing the analysis in any way, all neighborhoods of Manhattan were kept, not just those containing the client's coffee shops. Similarly, all neighborhoods of Toronto were collected, cleaned, visualized, and characterized in the same manner.

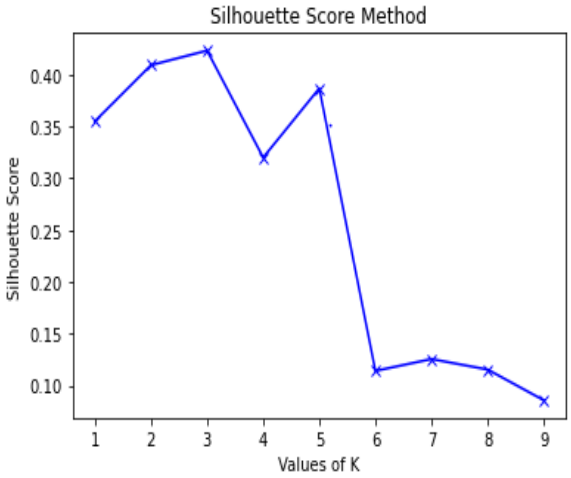


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adelaide	Coffee Shop	Restaurant	Japanese Restaurant	Café	Gastropub
1	Agincourt North	Bakery	Bank	Movie Theater	Beer Store	Liquor Store
2	Albion Gardens	Grocery Store	Fast Food Restaurant	Liquor Store	Sandwich Place	Hardware Store
3	Bathurst Quay	Coffee Shop	Café	Harbor / Marina	Park	Dance Studio
4	Bloordale Gardens	Donut Shop	Convenience Store	Intersection	Deli / Bodega	Print Shop

With the neighborhoods of both cities characterized, it became possible to analyze the similarities of these neighborhoods. In order to find these similarities the cleaned data from each city were combined into a single table containing the city, neighborhood, and relative frequency of each venue type returned by Foursquare's API. The format of this data lent itself well to K Means clustering, such that the neighborhoods of Manhattan and Toronto could be grouped based on the frequencies of similar venue types.

Before the clustering could be performed and analyzed however, it was necessary to optimize the k value (appropriate number of clusters) for the analysis. This optimization was carried out using a combination of the Elbow Method and the Silhouette Score. These are both statistical methods that demonstrate optimal k values (through error reduction calculation).





The Elbow Method and Silhouette Score were both in agreement that k = 3 was the optimal number of clusters for this analysis. The point of inflection of the Elbow Method occurs at k = 3, as does the global maximum of the Silhouette Score. For this, k = 3 was selected as the optimal number of clusters.

With the optimal number of clusters selected K Means clustering was performed and each neighborhood was labeled with the appropriate cluster label.

	City	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Manhattan	Battery Park City	2	Park	Hotel	Gym	Coffee Shop	Memorial Site
1	Manhattan	Carnegie Hill	2	Coffee Shop	Café	Bookstore	Italian Restaurant	Gym / Fitness Center
2	Manhattan	Central Harlem	2	African Restaurant	Chinese Restaurant	Bar	Seafood Restaurant	American Restaurant
3	Manhattan	Chelsea	2	Coffee Shop	Bakery	American Restaurant	Art Gallery	Café
4	Manhattan	Chinatown	2	Chinese Restaurant	Cocktail Bar	Bakery	Dessert Shop	American Restaurant

This clustering allowed the neighborhoods of both Toronto and Manhattan to be stratified into groups that were statistically similar according to the frequency of venue types.

Results –

The neighborhoods of Manhattan and Toronto were separated into three groups; both neighborhoods containing the client's coffee shops were separated into Cluster Three, together, along with a large variety of Toronto neighborhoods.

Statistically, the Toronto neighborhoods contained in Cluster Three were similar to the client's neighborhoods.

Cluster 3

```
m_t_most_common.loc[m_t_most_common['Cluster Labels']==2]
```

	City	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Manhattan	Battery Park City	2	Park	Hotel	Gym	Coffee Shop	Memorial Site
1	Manhattan	Carnegie Hill	2	Coffee Shop	Café	Bookstore	Italian Restaurant	Gym / Fitness Center
2	Manhattan	Central Harlem	2	African Restaurant	Chinese Restaurant	Bar	Seafood Restaurant	American Restaurant
3	Manhattan	Chelsea	2	Coffee Shop	Bakery	American Restaurant	Art Gallery	Café
4	Manhattan	Chinatown	2	Chinese Restaurant	Cocktail Bar	Bakery	Dessert Shop	American Restaurant
5	Manhattan	Civic Center	2	Coffee Shop	Gym / Fitness Center	Spa	Cocktail Bar	Hotel
6	Manhattan	Clinton	2	Theater	American Restaurant	Gym / Fitness Center	Coffee Shop	Italian Restaurant
7	Manhattan	East Harlem	2	Mexican Restaurant	Bakery	Thai Restaurant	Deli / Bodega	Latin American Restaurant
8	Manhattan	East Village	2	Bar	Pizza Place	Mexican Restaurant	Ice Cream Shop	Wine Bar
9	Manhattan	Financial District	2	Coffee Shop	Pizza Place	Café	Bar	Cocktail Bar

Cluster Three was then analyzed further, specifically the characteristics of the neighborhoods containing the client's coffee shops. Both Carnegie Hill and Chelsea are characterized by having Coffee Shops as the most common venue in the neighborhoods. They also both contain cafés in the top five most common venues. For this, Toronto neighborhoods from Cluster Three with Coffee Shops as the most common venue were extracted and analyzed.

	City	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Manhattan	Carnegie Hill	2	Coffee Shop	Café	Bookstore	Italian Restaurant	Gym / Fitness Center
43	Toronto	Bathurst Quay	2	Coffee Shop	Café	Harbor / Marina	Park	Dance Studio
61	Toronto	Grange Park	2	Coffee Shop	Café	Sandwich Place	Art Gallery	Arts & Crafts Store
63	Toronto	Harbourfront	2	Coffee Shop	Café	Restaurant	Hotel	Italian Restaurant
64	Toronto	Harbourfront West	2	Coffee Shop	Café	Restaurant	Hotel	Italian Restaurant
177	Toronto	Harbourfront East	2	Coffee Shop	Café	Restaurant	Hotel	Italian Restaurant
205	Toronto	Queen's Park	2	Coffee Shop	Café	Sandwich Place	Italian Restaurant	Bubble Tea Shop
211	Toronto	Runnymede	2	Coffee Shop	Café	Bakery	Bank	Pizza Place
214	Toronto	St. James Town	2	Coffee Shop	Café	Pizza Place	Grocery Store	Indian Restaurant
216	Toronto	Studio District	2	Coffee Shop	Café	Cosmetics Shop	Vegetarian / Vegan Restaurant	Japanese Restaurant

Any of these Toronto neighborhoods should be able to provide reasonable regions of interest for the client. Not only did they feature coffee shops as the number one most common venue, but also cafés as the second most common.

Discussion –

Due to the cluster analysis, the neighborhoods of Toronto featuring Coffee Shops and Cafés as the first and second most common venue types respectively are bring recommended to the client as likely neighborhoods in which successful coffee shops could be opened.

Queen's Park and Harbourfront were specifically highlighted to the client, as not only do they meet the above requirements, but they also contain Italian Restaurants, which is yet another

characterizing feature of the neighborhood in which the client had already found success, Carnegie Hill.

Conclusion –

In conclusion, using the geospatial data of the Manhattan, New York City area, an area containing the client's successful coffee shops, and the geospatial data of Toronto, Ontario, Canada, an area to which the client is interested expanding, along with Foursquare's API and K Means Unsupervised Clustering, we are able to provide locations of Toronto that are statistically similar to the locations in which the client has already found success.

Actionable recommendations have been provided to the client based on the statistical analysis of the geospatial data, and the data provided by the venue endpoint of Foursquare's API. The client's target expansion will allow them to be more successful by reaching a clientele that has already proven to fuel their coffee shop business.