Clustering, Segmenting, and All That Jazz: Opening a Jazz Club in Toronto

Joseph Barden

Introduction

Toronto is a vibrant city geographically situated between Detroit and Montreal, both of which are renowned for their respective jazz scenes. Despite being a major metropolitan center, Toronto has not achieved the same distinction as a hub for jazz artists and aficionados. This presents the enterprising restaurateur/club owner with the opportunity to fill a niche by opening a new jazz club – a bar/restaurant specializing in soul food cuisine, with an intimate aesthetic and a performance space for jazz and blues artists.

Business Problem

The objective of this project was to determine the ideal location(s) for a jazz club that serves a blend of southern-style barbecue, soul food, and Cajun cuisine while providing local and touring musicians a venue for live performances. Using data science and machine learning methods, an analysis was conducted in order to provide information pertinent to the matter of where an entrepreneur ought to consider opening such a business within the city of Toronto.

Data

To solve this problem, data were obtained from three sources: 1) a list of Toronto neighborhoods scraped from Wikipedia, 2) geographic coordinates of Toronto neighborhoods acquired using the Geocoder library for Python, and 3) data on popular venues in Toronto, accessed using the Foursquare API.

Methodology

All methods were carried out in Python 3.6. A list of Toronto neighborhoods was acquired by scraping Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) using the Pandas data analysis library. Tabular data were pulled from the web page and read into a Pandas

dataframe. The data were cleaned to remove null values and grouped according to postal code and borough (Fig. 1)

	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Fig. 1: Cleaned Toronto neighborhood data scraped from Wikipedia.

Geographic coordinates for each postal code were read from a .csv file provided by Coursera (http://cocl.us/Geospatial_data). Latitude and Longitude values were merged with the existing neighborhood data. From the resulting dataframe, only boroughs within Toronto were selected for further analysis. This subset of data was read into a new dataframe (Fig. 2).

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
2	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

Fig. 2: Cleaned data for Toronto boroughs with geographic coordinates.

Using these data, a map of Toronto was created. Neighborhoods were plotted using the Folium spatial visualization library and segmented according to their respective boroughs (Fig. 3).

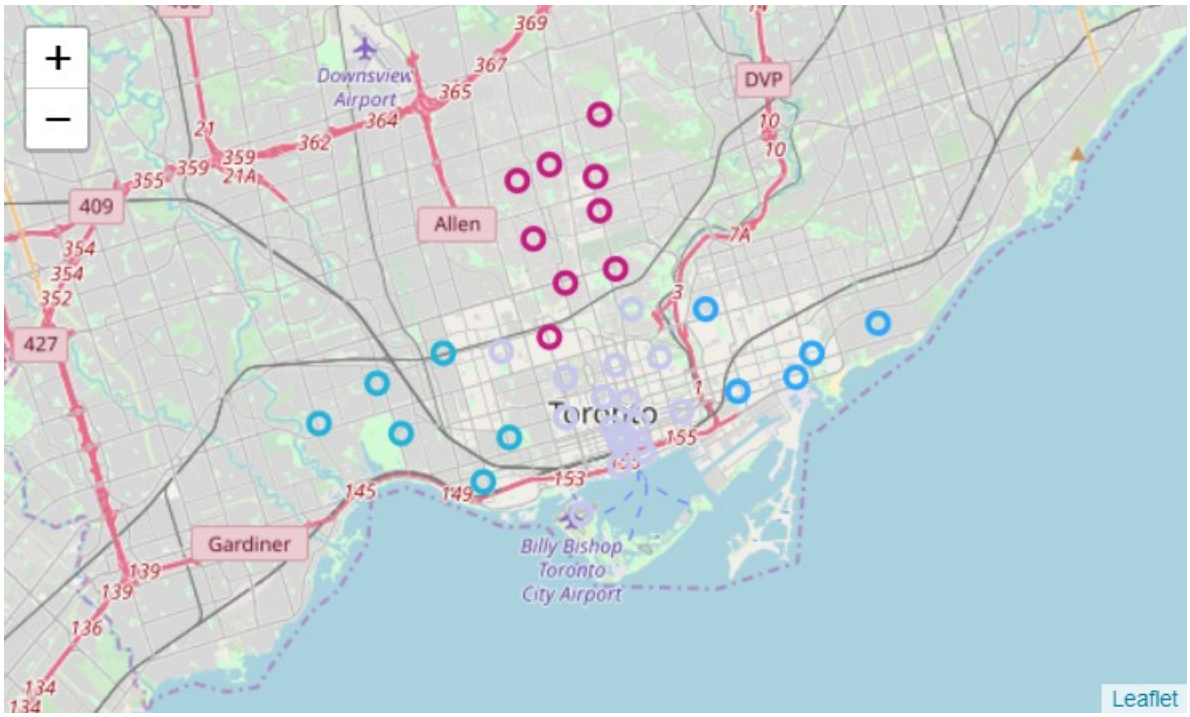

Fig. 3: Plot of Toronto neighborhood coordinates, segmented into four boroughs.

Using the Foursquare API, venues within each neighborhood, and their respective geographic coordinates, were obtained. Foursquare also provides the categories of venues, and this feature of the data was used for analysis. Using one hot encoding, venues were separated into distinct categories, and the frequency of each category (i.e., the proportion of venues of that category) within each neighborhood was calculated. These data were then used to list the five most common types of venue in each neighborhood (Fig. 4). Furthermore, a new dataframe was populated with the top ten most common venue types for each neighborhood.

Using k-means clustering, Toronto was clustered based on the most popular venue categories within its respective neighborhoods. An optimal *k*-value of 4 was established using the elbow method. The clusters obtained in this way, however, proved less useful than anticipated. So, more relevant information was gleaned from the data by re-running the clustering algorithm using only information on venues similar to the hypothetical new business. Specifically, a new dataframe was created showing the frequencies per neighborhood of the following venue categories: jazz clubs, Cajun/creole restaurants, southern/soul food restaurants, BBQ joints, and performing arts venues (Fig. 5).

```
----Cabbagetown, St. James Town----
          venue  freq
0  Coffee Shop  0.07
1           Pub  0.05
2         Café  0.05
3       Bakery  0.05
4   Restaurant  0.05


----Central Bay Street----
                        venue  freq
0                 Coffee Shop  0.15
1             Ice Cream Shop  0.05
2                       Café  0.05
3  Middle Eastern Restaurant  0.05
4          Italian Restaurant  0.05


----Chinatown, Grange Park, Kensington Market----
                           venue  freq
0                           Café  0.08
1  Vegetarian / Vegan Restaurant  0.06
2                            Bar  0.05
3             Mexican Restaurant  0.04
4             Chinese Restaurant  0.04
```

Fig. 4: Five most common venue categories in three Toronto neighborhoods.

	Neighborhood	Jazz Club	Cajun / Creole Restaurant	Southern / Soul Food Restaurant	BBQ Joint	Performing Arts Venue
0	Adelaide, King, Richmond	0.010000	0.0	0.0	0.000000	0.000000
1	Berczy Park	0.017857	0.0	0.0	0.017857	0.000000
2	Brockton, Exhibition Place, Parkdale Village	0.000000	0.0	0.0	0.000000	0.045455
3	Business Reply Mail Processing Centre 969 Eastern	0.000000	0.0	0.0	0.000000	0.000000
4	CN Tower, Bathurst Quay, Island airport, Harbo...	0.000000	0.0	0.0	0.000000	0.000000

Fig. 5: Dataframe containing frequencies of relevant venue types, by neighborhood.

Again, the optimal *k*-value was determined by plotting the sum of squared error and using the elbow method. The k-nearest neighbors algorithm was used to assign each row of the dataframe to a

cluster. The cluster labels and geographic coordinates of each neighborhood were added to the dataframe (Fig. 6) in order to visualize the new clusters. The neighborhoods were plotted in Folium, with clusters denoted by color-coding (Fig. 7).

Results

	Neighborhood	Jazz Club	Cajun / Creole Restaurant	Southern / Soul Food Restaurant	BBQ Joint	Performing Arts Venue	Latitude	Longitude	Cluster
2	Brockton, Exhibition Place, Parkdale Village	0.00	0.0	0.0	0.0	0.045455	43.763573	-79.188711	0
20	Harbourfront, Regent Park	0.00	0.0	0.0	0.0	0.019608	43.757490	-79.374714	0
0	Adelaide, King, Richmond	0.01	0.0	0.0	0.0	0.000000	43.806686	-79.194353	1
19	Harbourfront East, Toronto Islands, Union Station	0.00	0.0	0.0	0.0	0.010000	43.786947	-79.385975	1
22	Lawrence Park	0.00	0.0	0.0	0.0	0.000000	43.770120	-79.408493	1

Fig. 6: Final cluster data.

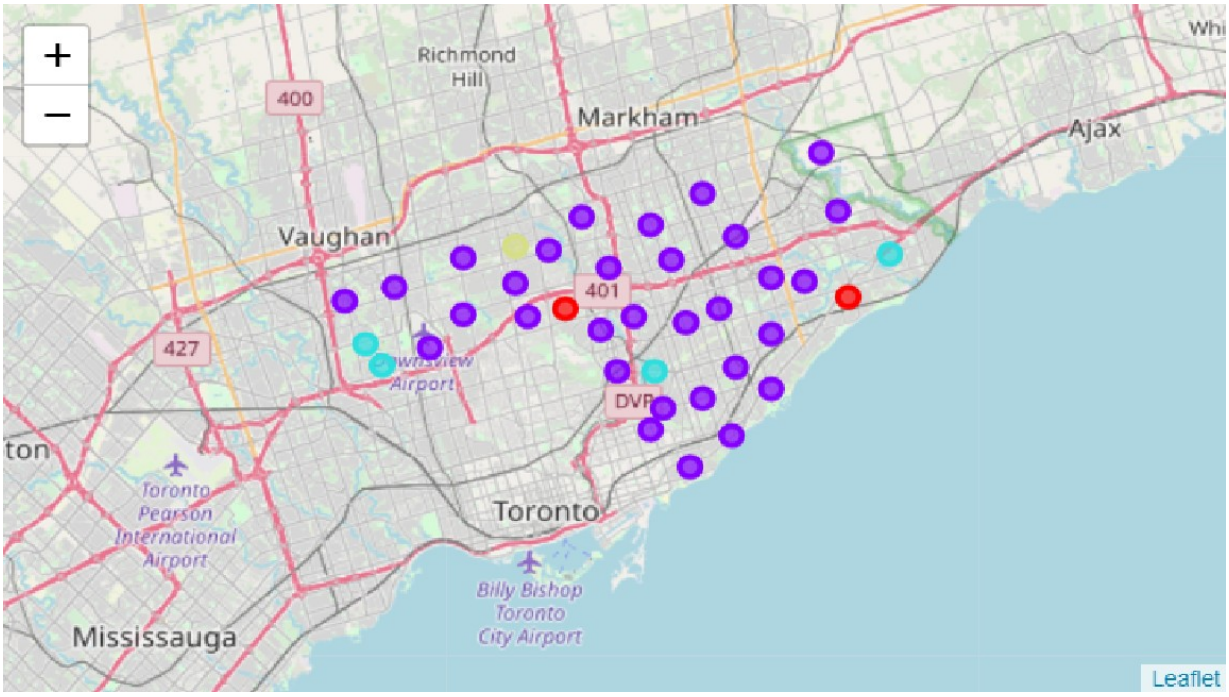

Fig. 7: Toronto venue clusters.

The k-means clustering results in four clusters of neighborhoods based on the number of particular venue types (e.g., jazz clubs, Cajun/creole restaurants) in each neighborhood. Cluster 1 contains no venues of interest except performing arts venues. Cluster 2, the largest cluster, is largely devoid of each of these venue types. Of the relevant venue categories, jazz clubs and BBQ restaurants are best represented in Cluster 3. Finally, Cluster 4 consists of a single neighborhood, High Park/The Junction South, in which Cajun-style cuisine can be found.

Discussion

A cursory look at the data confirms the initial assessment of Toronto as a city with a niche that is ready to be filled. Toronto has a dearth of jazz clubs and performing arts venues (five and three, respectively), with neither category counted among the ten most popular venue types in any neighborhood. Furthermore, the second, more focused segmentation reveals that the few jazz clubs that already exist within Toronto are not well distributed through the city, being largely relegated to outlying neighborhoods to the east west (e.g., Berczy Park, St. James Town).

Similarly, Toronto is home to few restaurants specializing in the renowned cuisines of the southern United States: barbecue, Cajun/creole, and soul food. For only a single neighborhood, High Park/The Junction South, does the Cajun/creole style find itself as the tenth most popular venue category. Toronto eateries are dominated by more generic restaurants and coffee shops, with ethnic establishments dominated by sushi and other Asian cuisines.

The relative rarity and uneven distribution of these types of venues establishes Toronto as an untapped market for a bar/restaurant that serves up southern-style cooking and live jazz. Furthermore, the clustering analysis reveals a handful of caveats – neighborhoods that the hypothetical restaurateur would do well to avoid when selecting a location for a new jazz club, in order to maximize market share.

Conclusion

For the entrepreneur seeking to open a jazz club in Toronto, Cluster 2 of the final analysis shows the most promise with respect to potential business locations. Not only is it the largest cluster with 31 neighborhoods, but it is virtually devoid of the types of venues identified as potential competitors. The only neighborhoods in this cluster to be avoided are those which already feature jazz clubs: Adelaide/King/Richmond and Commerce Court/Victoria Hotel.

Cluster 1 also presents viable options, with Harbourfront/Regent Park being perhaps the optimal choice. Apart from an utter lack of competition within the neighborhood, it is also situated near the city center, away from the slight concentrations of extant jazz venues. Neighborhoods in Clusters 3 and 4 should be avoided on the grounds that they contain slightly higher proportions of potential competitors.