Using clustering technique to recommend Halal restaurant location in Toronto

100% HALAL CERTIFIED * 100% HALAL CERTIFIED * 100% HALAL CERTIFIED *
حلال
HALAL

Peiqing Lian (https://github.com/jasperlpq)

1. Introduction

1.1 Background

The purpose of this capstone project is to help people who is looking for a place to open a Halal restaurant in Toronto area. Although, there might have been quite amount Asian food places opening in Toronto, there may not be enough authentic Halal food around and it would be a great opportunity for those who want to start their own restaurant business. Ideas behind this project are divided into two parts, firstly, Halal food, as one of the Asian food branches, is not widely spread around the community. Secondly, there are not a lot of Halal immigration migrating to Toronto and thus helping Halal people to find a great location to start their business would enrich the diversity of the whole community in terms of food and culture.

1.2 Business problem

The objective of this capstone project is to find a trending but suitable location for people who are going to open a Halal restaurant in Toronto area using data science methods such as clustering approach.

1.3 Target audience

The people who want to open up a Halal restaurant in Toronto area

2. Data sources

- We will scrap Toronto neighborhoods data from Wikipedia and corresponding geo information such as latitude and longitude and corresponding borough.
- We then need data of different venues in different neighborhoods of that specific borough.
- In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API. After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meters.

The information obtained per venue as follows:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue
5. Name of the venue e.g. the name of a store or restaurant
6. Venue Latitude
7. Venue Longitude
8. Venue Category

3. Methodology

3.1 Exploratory Data Analysis

After getting the desired dataset, I extracted borough name that only contains Toronto as my targeted area source without focusing on other boroughs. Then Geocode has been used to get the geographical coordinates of Toronto and visualization of map of Toronto was performed to verify whether corresponding borough has right coordinates (see figure 3.1.1).

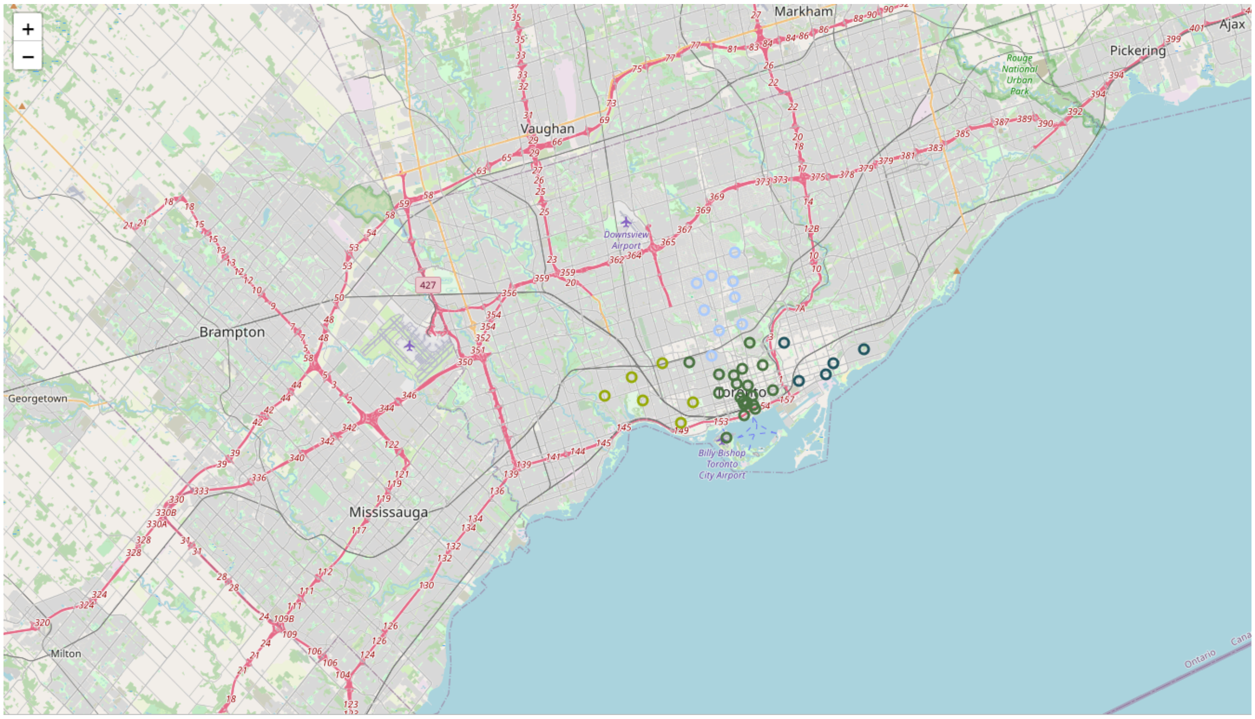
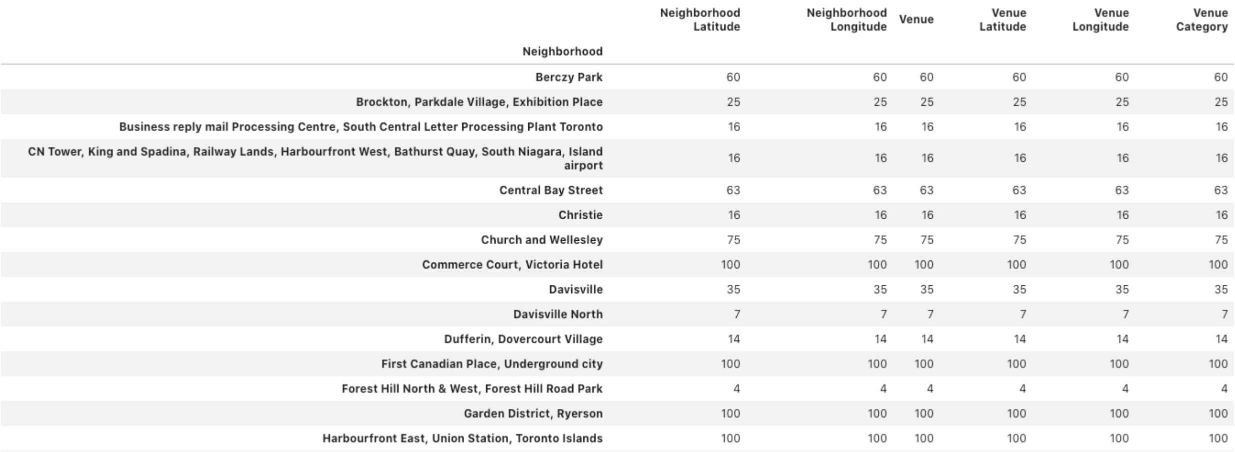


Figure 3.1.1 Map of Toronto with four main boroughs respectively are Central Toronto, Downtown Toronto, East Toronto, and West Toronto.

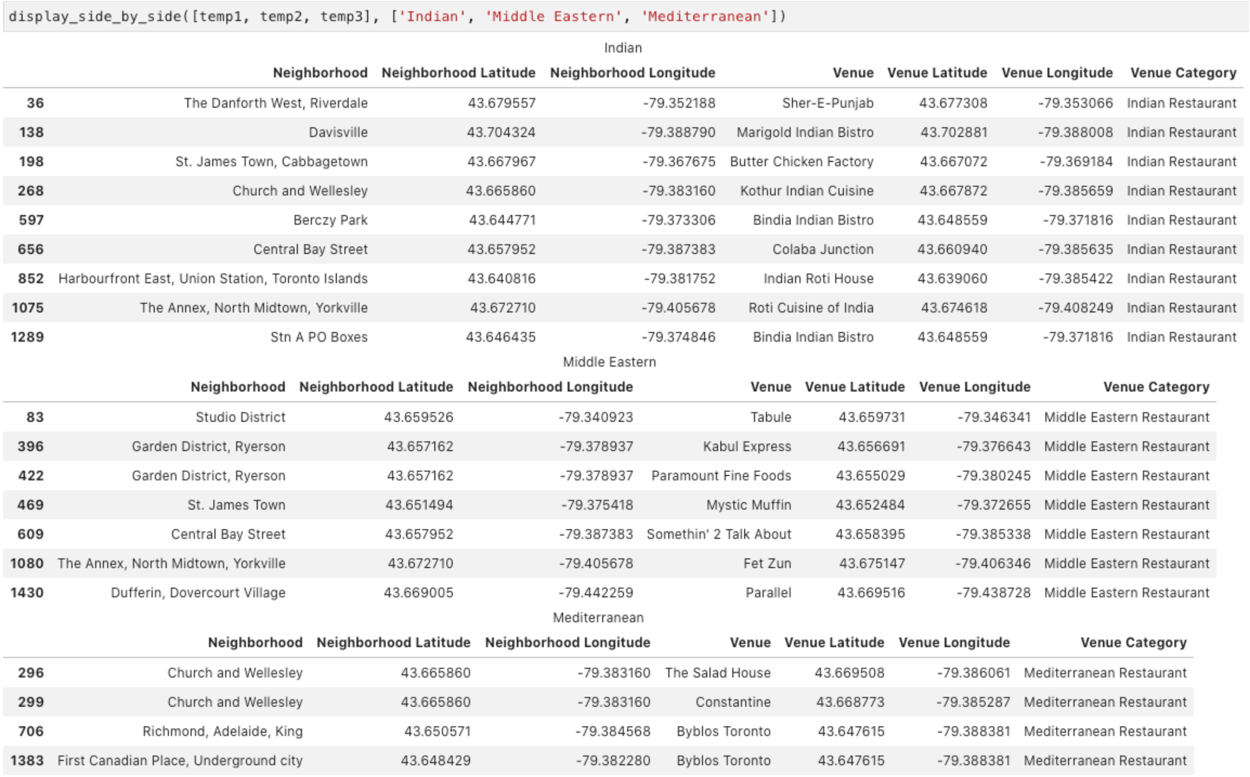
Next, I used Foursquare developer account to call API and got 100 venues within 500 meters of defined radius range (due to http request limitations the number of places per neighborhood parameter). We can see how many venues out there in all neighborhoods to get a glance of the data (**See Figure 3.1.2**).



	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Berczy Park	60	60	60	60	60	60
Brockton, Parkdale Village, Exhibition Place	25	25	25	25	25	25
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	16	16	16	16	16	16
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	16	16	16	16	16	16
Central Bay Street	63	63	63	63	63	63
Christie	16	16	16	16	16	16
Church and Wellesley	75	75	75	75	75	75
Commerce Court, Victoria Hotel	100	100	100	100	100	100
Davisville	35	35	35	35	35	35
Davisville North	7	7	7	7	7	7
Dufferin, Dovercourt Village	14	14	14	14	14	14
First Canadian Place, Underground city	100	100	100	100	100	100
Forest Hill North & West, Forest Hill Road Park	4	4	4	4	4	4
Garden District, Ryerson	100	100	100	100	100	100
Harbourfront East, Union Station, Toronto Islands	100	100	100	100	100	100

Figure 3.1.2 Number of venues returned for each neighborhood

Since our interested target food is Halal, I initially explored all venue categories, 235 in total, and did not find any within the data retrieved, I made an assumption that, based on food taste (note: After googling Halal food, it returns Indian, ME, and Mediterranean food as related information, so I assume people would like to choose those if Halal food is not found.), I decided to include Indian, Middle Eastern, and Mediterranean food as Halal-kind food (**See Figure 3.1.3**). Another reason I combined them all together is that individual part of data contains only a few and is not enough to use.



```
display_side_by_side([temp1, temp2, temp3], ['Indian', 'Middle Eastern', 'Mediterranean'])
```

Indian

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
36	The Danforth West, Riverdale	43.679557	-79.352188	Sher-E-Punjab	43.677308	-79.353066	Indian Restaurant
138	Davisville	43.704324	-79.388790	Marigold Indian Bistro	43.702881	-79.388008	Indian Restaurant
198	St. James Town, Cabbagetown	43.667967	-79.367675	Butter Chicken Factory	43.667072	-79.369184	Indian Restaurant
268	Church and Wellesley	43.665860	-79.383160	Kothur Indian Cuisine	43.667872	-79.385659	Indian Restaurant
597	Berczy Park	43.644771	-79.373306	Bindia Indian Bistro	43.648559	-79.371816	Indian Restaurant
656	Central Bay Street	43.657952	-79.387383	Colaba Junction	43.660940	-79.385635	Indian Restaurant
852	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	Indian Roti House	43.639060	-79.385422	Indian Restaurant
1075	The Annex, North Midtown, Yorkville	43.672710	-79.405678	Roti Cuisine of India	43.674618	-79.408249	Indian Restaurant
1289	Stn A PO Boxes	43.646435	-79.374846	Bindia Indian Bistro	43.648559	-79.371816	Indian Restaurant

Middle Eastern

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
83	Studio District	43.659526	-79.340923	Tabule	43.659731	-79.346341	Middle Eastern Restaurant
396	Garden District, Ryerson	43.657162	-79.378937	Kabul Express	43.656691	-79.376643	Middle Eastern Restaurant
422	Garden District, Ryerson	43.657162	-79.378937	Paramount Fine Foods	43.655029	-79.380245	Middle Eastern Restaurant
469	St. James Town	43.651494	-79.375418	Mystic Muffin	43.652484	-79.372655	Middle Eastern Restaurant
609	Central Bay Street	43.657952	-79.387383	Somethin' 2 Talk About	43.658395	-79.385338	Middle Eastern Restaurant
1080	The Annex, North Midtown, Yorkville	43.672710	-79.405678	Fet Zun	43.675147	-79.406346	Middle Eastern Restaurant
1430	Dufferin, Dovercourt Village	43.669005	-79.442259	Parallel	43.669516	-79.438728	Middle Eastern Restaurant

Mediterranean

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
296	Church and Wellesley	43.665860	-79.383160	The Salad House	43.669508	-79.386061	Mediterranean Restaurant
299	Church and Wellesley	43.665860	-79.383160	Constantine	43.668773	-79.385287	Mediterranean Restaurant
706	Richmond, Adelaide, King	43.650571	-79.384568	Byblos Toronto	43.647615	-79.388381	Mediterranean Restaurant
1383	First Canadian Place, Underground city	43.648429	-79.382280	Byblos Toronto	43.647615	-79.388381	Mediterranean Restaurant

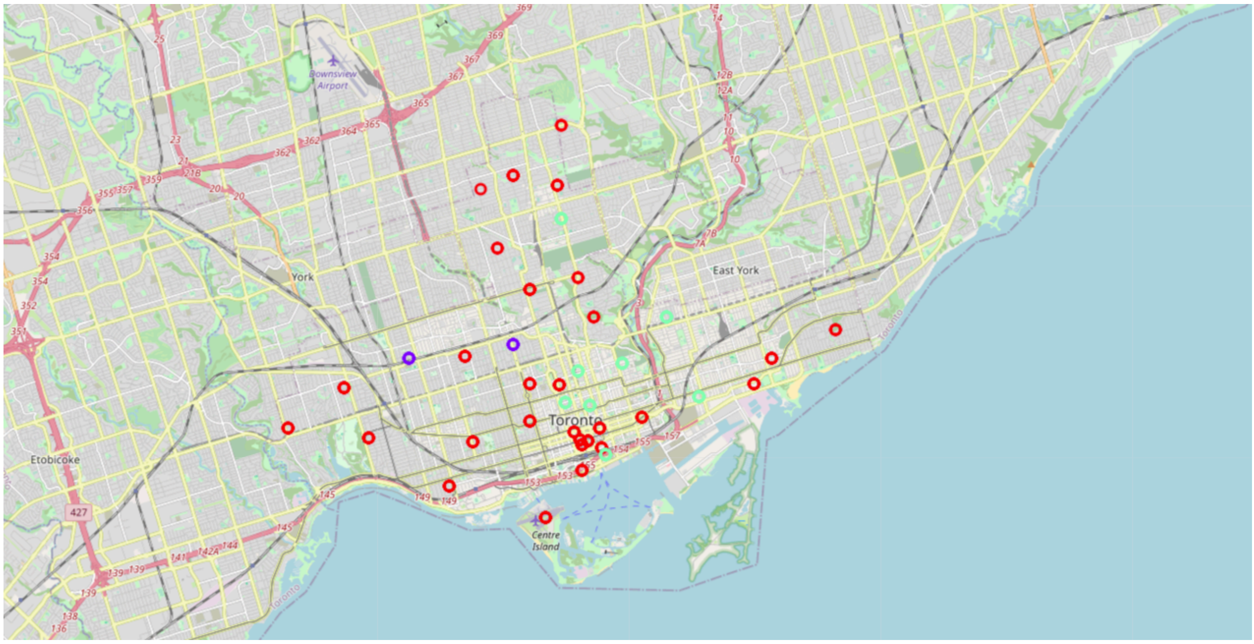
Figure 3.1.3 Indian, Middle Eastern, and Mediterranean

3.2 Clustering Approach

Before performing Kmeans clustering algorithm, I one-hot-encode all venue categories in each column and then group rows by neighborhood and take the mean of the frequency of occurrence of each category. Most importantly, I created a new data frame only that contains Halal restaurant frequency of occurrence of each corresponding neighborhood. With data prepared, I clustered the neighborhoods in Toronto into 3 clusters based on the frequency of occurrence for Halal food (combining 3 types of food categories, **referred to section 3.1**).

4. Results Section

Clusters visualized



Toronto neighborhoods are clustered into 3 different groups based on how many Halal restaurants in the neighborhoods.

Cluster 1 (Red): Almost there exists no target restaurants

Cluster 2 (Purple): Most target restaurants are in this cluster

Cluster 3 (Green): More Halal food than cluster 1 and less than cluster 2

5. Discussions and Recommendations

- There are more Halal food places in cluster 2 especial around North Midtown, Yorkville and Dover court Village.
- Cluster 1 has mostly close to zero frequency Halal food opening in areas such as Commerce Court, Adelaide, King, Richmond areas.
- Opportunity of opening up at St. James Town and Garden District shows some promising since it seems not quite competitive out there.
- Cluster 1 has most of diverse venue categories and opening a Halal restaurant in this cluster area might be a good place to thrive.
- It would be recommended that people can consider to start their business in cluster 1 and cluster 3 areas since there is little to no competition.

6. Defensibility and Future Work

In this capstone project, the underlying hypothesis is that the frequency of occurrence of Halal restaurant in each neighborhood is the only factor I took into consideration. Therefore, there are some future works and I would divide into two parts, namely data part and model part.

Data

More information can be included from various sources such as Google search and Yelp search. For instance, since we did not find any Halal food from the Foursquare nearby venue API, I can incorporate data from other data sources and analyze based on real labeled data. In addition, future data set should include not only the existence of restaurant but other factors such as population density, restaurant review, demographical data, housing price, violence rate, and etc.

Model

I only used 3 clusters as my initial clustering model. However, the process of determining the optimal k hasn’t been done. Methods like elbow method, silhouette score, and hypothesis testing can be performed to get the best k. In terms of model selection, other clustering method can be used such as hierarchical clustering as well.

7. Conclusion

In this capstone project, we have defined the business problem and target audiences, chose the dataset to analyze, did data exploration, and performed machine learning technique, clustering, to provide insights on finding the right location for opening up a Halal restaurant in Toronto area.

8. References

Postal Codes of Canada:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare developer API:

https://developer.foursquare.com/