Identify Niche Area for a New Business Venture

1. Introduction

1.1 Background

Every year, hundreds of new businesses are started in every city. Some businesses see higher successes than others. Some of the factors that decide whether a business is successful is the demand for such a business, competitors in the market, availability of patrons, and profit margin. The goal of this project is to offer some insights into how to make some of the decisions pertaining to starting a new venture.

1.2 Problem

Suppose a person is interested in starting a business venture. The person must first decide what type of business to start and where to start it. These two key factors decide all other parameters. Therefore, this project focuses on solving these two questions.

Let us look more closely at these two questions. We first select the potential list of cities where this person is interested in starting the business. This selection can be based on his domicile, i.e., where he/she is currently residing or where they want to live long term. In general, people have specific preference on which broad area they want to live in and therefore, we will limit our search space within that area.

In this project, we look at two potential cities – New York City and Toronto. The goal is to find 1) what business is viable in this area and 2) a neighborhood to start this business.

2. Data Acquisition and Preprocessing

Details of neighborhoods and boroughs in New York City was obtained from the dataset provided by Coursera at https://cocl.us/new_york_dataset. The corresponding data for Toronto was scraped from the Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. The BeautifulSoup (bs4) package was used to get the contents of the webpage. The contents were then extracted into a table, preprocessed (stripped of unnecessary values/entries) and structured to obtain a pandas dataframe. The geospatial coordinates of all the neighborhoods were then added to the dataframe.

To validate the choice of location, the BikeShare data (https://open.toronto.ca/dataset/bike-share-toronto-ridership-data/) from the Open Data repository from the City of Toronto is used (https://open.toronto.ca/). The most recent data available is from Q4 of 2018 and is used. The data was loaded into a dataframe. As a first step, the geocodes of each bike station is obtained and added to the dataframe. Geocodes were obtained from geocod.io by uploading a csv of the street addresses and batch processing. The locations with no geocodes was then removed. The number of rides from each location was then tallied and added to the dataframe as a new column. The idea behind using the BikeShare Data is to show that the location selected does have a lot of people visiting it. A good way to quantify people visiting that location is the number of bike rides originating from the location.

3. Methodology

Some exploratory data analysis was carried out in order to understand the cities – New York city and Toronto. As a first step, the different neighborhoods in each city was plotted using Folium to understand the spread of the city.

The next step was to understand and explore various venues in each neighborhood. For this process, the FourSquare location data was used. Near-by venues were obtained for each neighborhood in both cities using the FourSquare API. The number of venues per neighborhood was used as a measure of how "busy" or popular the neighborhood was. This means that neighborhood with more venues are likely to have more visitors than neighborhoods with fewer venues nearby.

The next step was to analyze the diversity of venues in these popular neighborhoods. Once this is done, we can identify potential areas to start a business, and the type of business to start based on what is lacking in the area.

The last step is to validate the choice of location using the BikeShare data. The BikeShare data provides a list of trips along with the start and end locations of the trip. This was used to figure out the number of trips that started from each bike station. The popularity of the bike station is considered to be a function of the number of trips originating from the station.

4. Results

Figures 1 and 2 show the distribution of neighborhoods in New York City and Toronto respectively. Based on the plots, it can be seen that **the neighborhoods in NYC are found to be spread with uniform density, however, the neighborhoods in Toronto become sparser as we move further from the Downtown.**

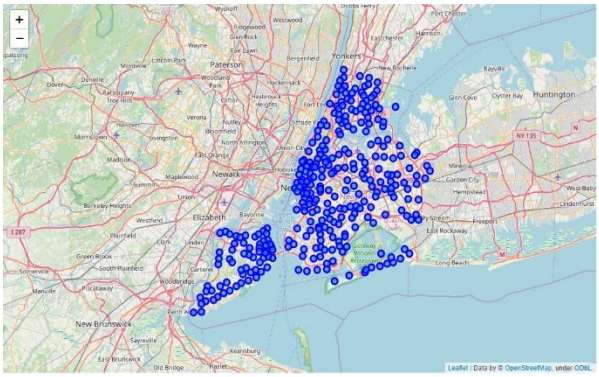


Figure 1: Neighborhoods in New York City

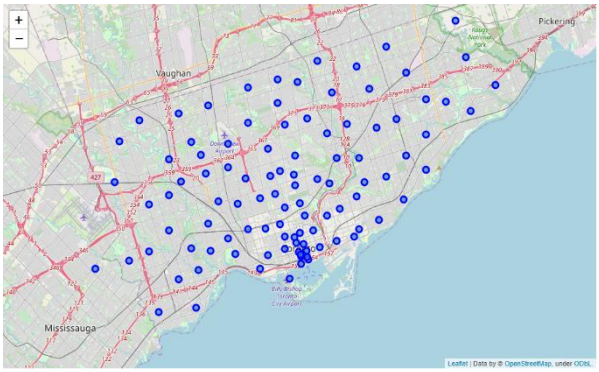


Figure 2: Neighborhoods in Toronto

From the above figures, it can also be seen that New York City is a much bigger City than Toronto as well.

The popular neighborhoods in both the cities were then extracted. This was done by selecting 100 venues near each neighborhood. Neighborhoods with more than 80 venues in close proximity were defined as popular or busy neighborhoods.

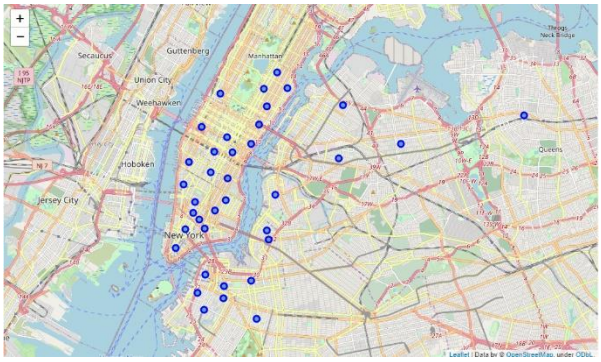


Figure 3: Popular neighborhoods in NYC

Figure 4: Popular neighborhoods in Toronto

The key here observation is that all the popular locations in Toronto are in the Downtown Toronto area, whereas its spread over 4 of the 5 boroughs in NYC, while it is localized to one borough in Toronto.

The next step was to analyze the diversity of venues in these popular neighborhoods. New York City has 423 different types of venues whereas Toronto has 263 different types.

NY has more venues offering the same facilities/experience compared to Toronto. Ex: NY has 77 yoga studios as opposed to 13 in Toronto. Similar trend in a lot of facilities

NY offers more multicultural experiences than Toronto. Examples of multicultural venues in NY not in Toronto: Argentinian Restaurant, Australian Restaurant, Austrian Restaurant, Arepa restaurant, etc.

This shows that NYC has a more diverse multicultural experience compared to Toronto. From this observation, we can infer that there is an opportunity to open a niche business in Toronto which it currently lacks. **Based on the above observations, it is advantageous to open a multicultural restaurant in Downtown Toronto. There seems to be lack of such restaurants in Toronto. For example, Downtown Toronto lacks a Kebab restaurant. In addition, the Downtown is a very popular place with many neighborhoods with more than 80 venues. This makes it an ideal place to start a business as the neighborhood is frequently visited by people in general.**

The next step is to identify a specific location within Downtown Toronto to start the restaurant. We will use k-means clustering to find one big cluster and thus its centroid. The key idea would be to cluster all the busy neighborhoods and locate the restaurant close to the center of the largest cluster. This is based on the assumption that we want the restaurant to be located at the center of the most popular neighborhood. Given that we only have a handful of popular neighborhoods, it does not make sense to cluster it into multiple groups. We thus set the cluster number as one and find the centroid of all the popular locations in Toronto. Based on this analysis, we get **the location (43.64878529 -79.37989929), which corresponds to Bay Street /King Street W.**

Thus, this would be a **good location to start a Kebab restaurant in Toronto.**

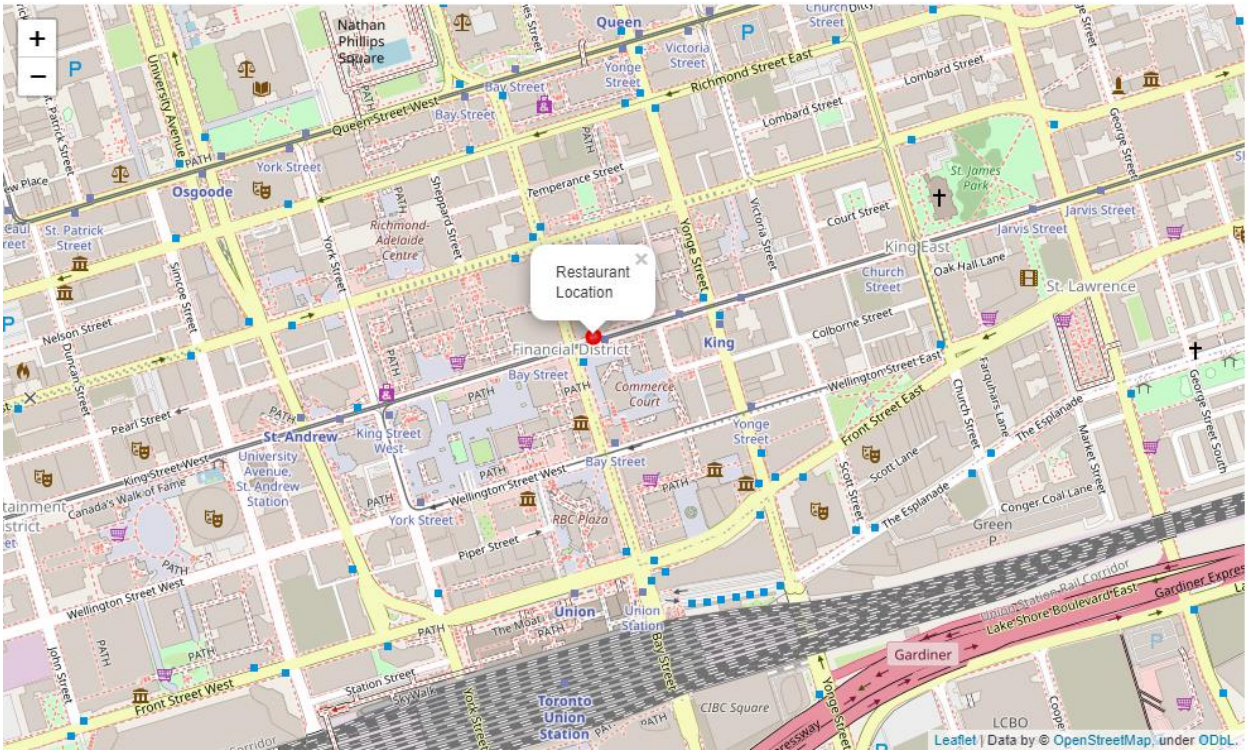


Figure 5: Chosen location to start a Kebab Restaurant

We narrowed down a location based on the FourSquare data. In practice, it is often advantageous to corroborate the findings by comparing results against another dataset.

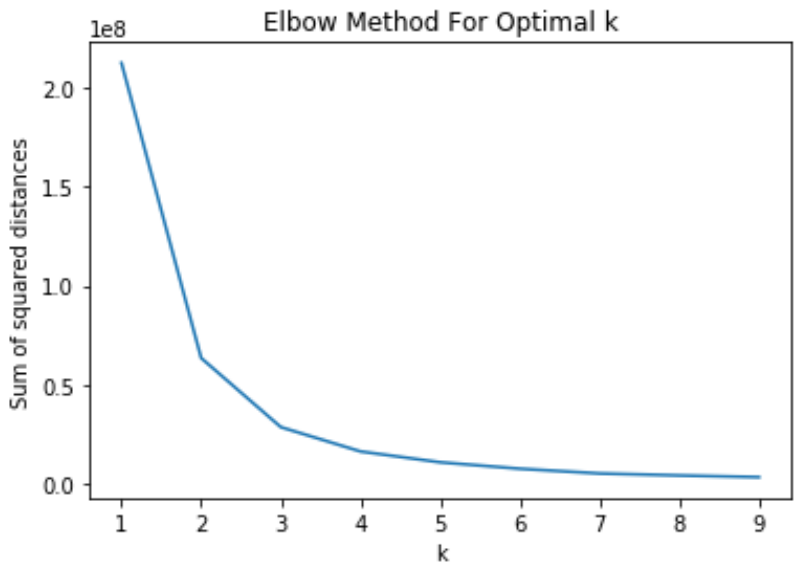
We will now use the BikeShare data from the Open Data provided by the City of Toronto to verify that the location we narrowed down for starting the business is indeed a popular location. Bikeridership data for 2018 Q4 is selected for this analysis (This is the latest dataset found).

	trip_id	trip_duration_seconds	from_station_id	trip_start_time	from_station_name	trip_stop_time	to_station_id	to_station_name	user_type
0	4158592	749	7061	10/1/2018 0:01	Dalton Rd / Bloor St W	10/1/2018 0:14	7042	Sherbourne St / Wellesley St E	Annual Member
1	4158593	433	7003	10/1/2018 0:06	Madison Ave / Bloor St W	10/1/2018 0:13	7280	Charles St E / Jarvis St - SMART	Annual Member
2	4158594	285	7024	10/1/2018 0:14	Dundonald St / Church St	10/1/2018 0:19	7028	Gould St / Mutual St	Annual Member
3	4158595	150	7190	10/1/2018 0:16	St. George St / Hoskin Ave	10/1/2018 0:18	7161	Beverley St / College St	Annual Member
4	4158596	744	7265	10/1/2018 0:21	Wallace Ave / Symington Ave - SMART	10/1/2018 0:33	7136	Queen St W / Close Ave	Annual Member

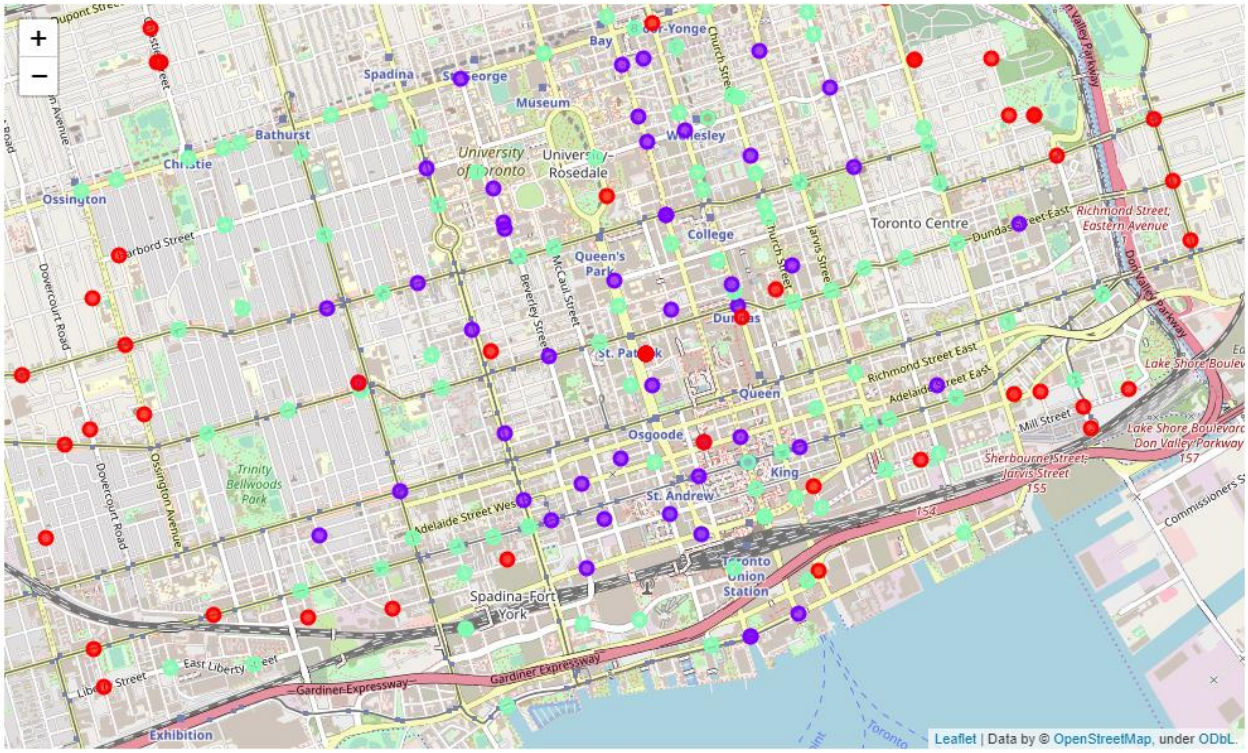
The raw data is converted to a form useful for this project by aggregating the trips based on start location. The number of trips from a rideshare location is a good indicator of how popular that location is. Here is a snapshot of the processed data (first few rows out of 357 total rows) used for clustering.

	Street	Latitude	Longitude	Zip	Count
0	Fort York Blvd / Capreol Ct	43.640400	-79.399500	M5V	2354
1	Lower Jarvis St / The Esplanade	43.648220	-79.370922	M5E	1106
2	St. George St / Bloor St W	43.667510	-79.399821	M5R	2744
3	Madison Ave / Bloor St W	43.686100	-79.402500	M4V	1662
4	University Ave / Elm St	43.656322	-79.389114	M5G	1271
5	King St W / York St	43.647914	-79.383565	M5H	1918
6	Bay St / College St (East Side)	43.660809	-79.385849	M5S	3447
8	Wellesley St / Queen's Park Cres	43.663611	-79.390628	M5S	1363
9	King St E / Jarvis St	43.650458	-79.371903	M5C	1765
10	King St W / Spadina Ave	43.645446	-79.395150	M5V	1045
11	Wellington St W / Portland St	43.643177	-79.399538	M5V	1278

In order to use k-means clustering, we first need to find the number of clusters to be used. The elbow method was used to find the number of clusters in this data. Based on the curve, the number of clusters was chosen as 3.



The rideshare bike stations were classified into three clusters using k-means clustering. The following plot shows the clusters (each cluster is color coded in a different color).



We can see that the area we selected to start the new Kebab restaurant does indeed lie in a very popular area. The area has a high concentration of cluster 1 (purple points) and cluster 2 (green points) than Cluster 0 (red) points. The cluster 1 and cluster 2 points have a higher average number of rides than cluster 0.

5. Discussion

Two potential cities were considered to start a new business venture. Based on exploratory analysis of the cities, it was found that the two cities have very different distributions of neighborhoods. This was leveraged to decide between the cities and figure out a suitable location for the business. In this process, the lack of diverse multicultural restaurants in Toronto presented itself as a prefect opportunity to start a business. It is a new avenue to explore with a lot of potential for growth.

Deciding which location to open the restaurant within Toronto presents a challenge in that all the popular spots in Toronto are clustered in the Downtown area. So, it meant that clustering wasn't an option – so the centroid of the popular locations was chosen.

The last part of this project focused on using Toronto BikeShare data to validate out choice of location. The analysis showed that our location choice is indeed in a popular neighborhood. The use of an independent dataset to validate the results gives more confidence to out analysis. In this process, k-means clustering was used to cluster the bike stations in Toronto. We clustered the bike stations into three clusters using k-means clustering.

6. Conclusion

In this project, the cities of Toronto and New York were studied to identify a good location and potential ideas for a new business venture. Datasets on neighborhoods of New York City and Toronto was used in conjunction with FourSquare data to determine choice of city as Toronto. Furthermore, the location and business idea was decided based on these datasets. The choice of location in Toronto was then compared with the BikeShare data to get more confidence in the analysis.