Predicting Toronto Subway Stations for Coffee Shop Expansion

Bradley Jones

Introduction/Business Problem:

Expansion to new locations is something that many businesses regularly must consider and often execute. This can be problematic when there are many possible locations of expansion, with little way of knowing which locations will perform best. For my Capstone, I focussed on determining the best expansion locations for coffee shops such as the Canadian native Tim Hortons or Starbucks. My audience would consist of stakeholders within a coffee shop business and this is of importance to them because expanding to a sub-optimal location can be extremely costly to the business and is not a mistake that can be mended swiftly. To narrow my initial search, I targeted locations near Toronto subway stations, and I will be using the Foursquare API to retrieve the venue data in close proximity to each station.

Data:

To solve the problem previously stated, I will be using data from two different providers. The first will provide the location data of all subway stations in Toronto via four CSV files representing each subway line and will come in the format: Latitude, Longitude, Station Name. The data can be found at the following address: https://scruss.com/blog/2005/12/14/toronto-subway-station-gps-locations/. The second data provider will be Foursquare using their venue-search API. In order to find the popular food-venue types in close proximity to each subway station, I will need to make a unique API call using the individual station's latitude and longitude obtained from the previous data provider. The documentation for the API call I will be making can be found at the following address: https://developer.foursquare.com/docs/api-reference/venues/explore/.

Methodology:

The first step of my methodology was to load the Toronto subway station location data to dataframes via the Pandas library. This data can be accessed here. After this step I had four dataframes, one for each subway line in Toronto with them being Bloor, Yonge, Sheppard and Scarborough. The data cleaning process began with changing the column names as the initial names contained what should have been the first entry. To fix this, I concatenated the first entry with the existing dataframe and renamed the columns to 'Latitude', 'Longitude' and 'Station'. Next, I mapped the stations using the mapping library Folium and colour coded the markers by subway line which can be seen in **Fig 1**.

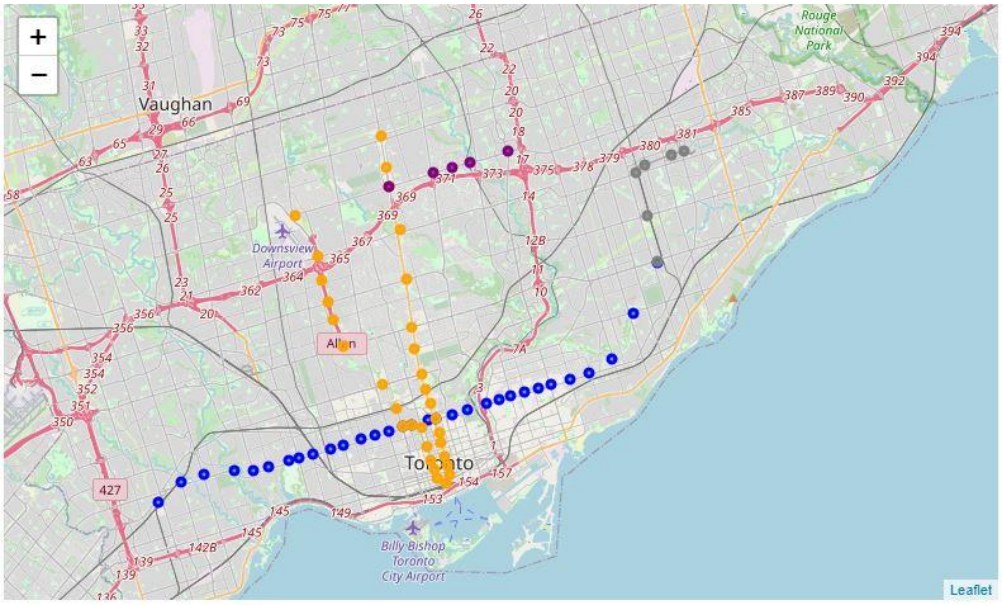

Fig 1

Through this data visualization I noticed that some subway stations overlapped and would skew the venue data I need for each station as I would be retrieving the same venue data twice. To fix this, one of the overlapping entries must dropped from the dataframe. These overlapping stations were ‘Spadina’, ‘St George’, ‘Bloor-Yonge’, ‘Sheppard-Yonge’ and ‘Kennedy’. After the overlapping entries were dropped, I updated the Folium map which can be seen in **Fig 2**.

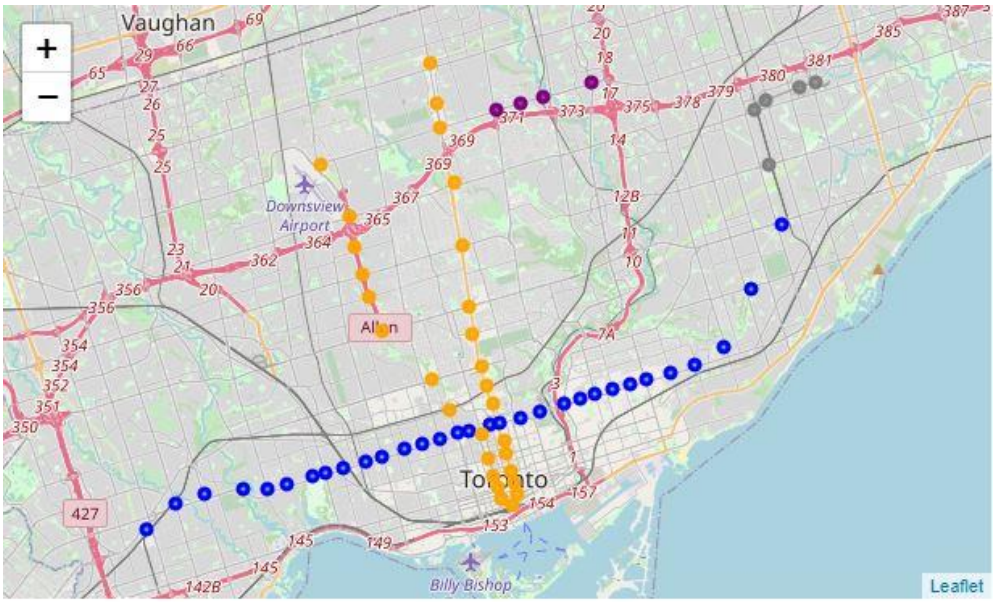

Fig 2

Since the data was now clean, I concatenated these dataframes together to create a dataframe containing data for every station. Now that I had all the station data cleaned and in a single dataframe, I began to make calls to the Foursquare API. The documentation for the type of call I made can be found here. As I wanted to keep the venues in relatively close proximity to the station, I set the radius to 250 which is measured in meters. Additionally, I set the sortByPopularity Boolean flag to 1 to ensure I will be receiving the most popular venues. The other parameters I had to set were the categoryId which was set to the 'Food' category id as this will only return food related venues and set the limit parameter to 10 as this will return the ten most popular food venues. Lastly, I set the latitude and longitude parameter to the latitude and longitude of the station from the dataframe containing all stations. From the venue data returned, the only features required were 'name', 'categories', 'latitude' and 'longitude'. Once the venue data was loaded into a dataframe, I was able to map the venues and their station using Folium. To start, I mapped the first station in the stations dataframe along with it's venues which can be seen in **Fig 3**.

Fig 3

Now that venue markers were being mapped correctly, I updated the map to colour code venue markers by their venue category. The updated map can be seen in **Fig 4**.

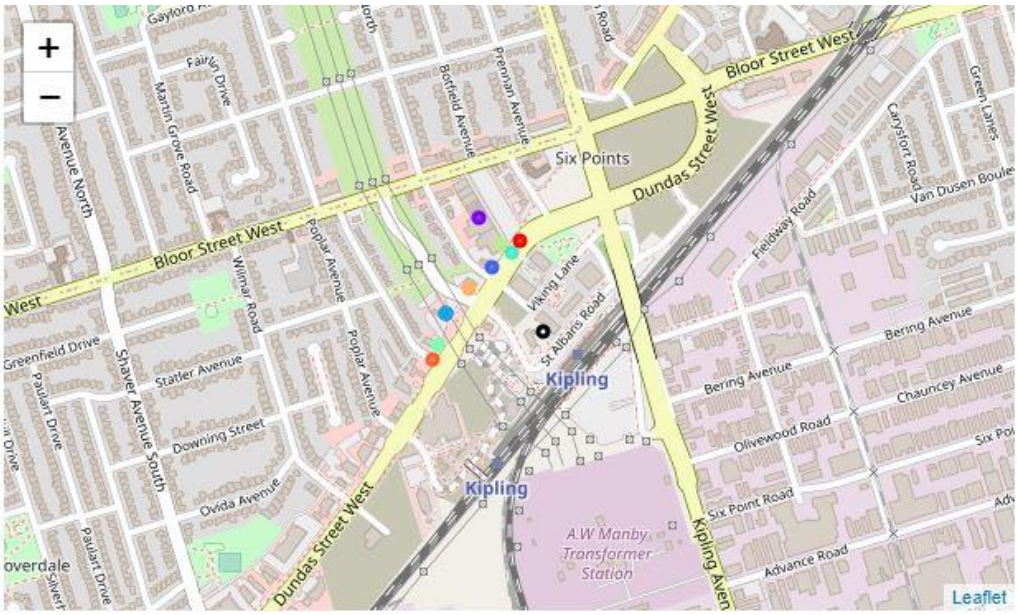

Fig 4

As the mapping appeared correct, I repeated the Foursquare API call for all other stations and loaded the required features to a new dataframe. In addition to the required features from the API, the new dataframe also contained the features of 'Station' indicating which station the venue belongs to, 'StationLat' indicating the station's latitude and 'StationLong' indicating the station's longitude. The first five entries of this dataframe can be seen in **Fig 5**.

	Station	StationLat	StationLong	Venue	VenueLat	VenueLong	VenueCategory
0	Kipling	43.63802	-79.536388	Starbucks	43.640187	-79.538053	Coffee Shop
1	Kipling	43.63802	-79.536388	Apache Burger	43.639257	-79.537725	Burger Joint
2	Kipling	43.63802	-79.536388	Tim Hortons	43.638374	-79.538893	Coffee Shop
3	Kipling	43.63802	-79.536388	Wendy's	43.638372	-79.538910	Fast Food Restaurant
4	Kipling	43.63802	-79.536388	Pho House	43.639528	-79.537217	Vietnamese Restaurant

Fig 5

With all the venue data now saved to a dataframe, I used Folium again to map all stations and all their venues with the venues still being colour coded based on their venue category. This map can be seen in **Fig 6**.

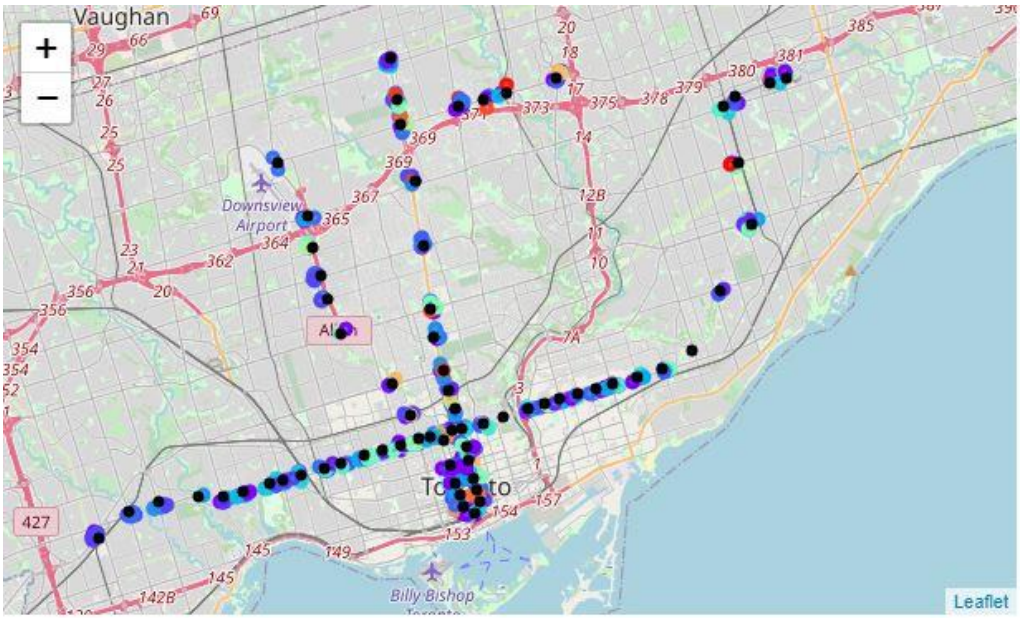

Fig 6

As I now had all the necessary data, I began my exploratory data analysis. I started by one-hot encoding the dataframe containing all venues with the columns being the different venue categories. From this, I could take the frequency of each venue category and rank them by popularity in close proximity to the subway stations. A list of the twenty-five most popular food venue categories can be seen in **Fig 7**.

```
----Most Popular Food Venues at Subway Station----
                       venue  freq
0               Coffee Shop  0.18
1                      Café  0.06
2       Fast Food Restaurant  0.05
3                Pizza Place  0.05
4                 Restaurant  0.04
5                    Bakery  0.04
6         Italian Restaurant  0.03
7             Sandwich Place  0.03
8          Chinese Restaurant  0.02
9                Grocery Store  0.02
10       Japanese Restaurant  0.02
11           Bubble Tea Shop  0.02
12          Sushi Restaurant  0.02
13           Thai Restaurant  0.02
14       Fried Chicken Joint  0.02
15                Food Court  0.02
16          Asian Restaurant  0.01
17                 Gastropub  0.01
18             Ice Cream Shop  0.01
19         Indian Restaurant  0.01
20       American Restaurant  0.01
21        Mexican Restaurant  0.01
22                 Juice Bar  0.01
23         Korean Restaurant  0.01
24  Mediterranean Restaurant  0.01
```

Fig 7

From this list it can be seen that the ‘Coffee Shop’ venue category performs exceptionally well when in close proximity to subway stations. For my next step, I wanted to cluster subway stations based on similarity of popular venue categories and in order to do this I must first prepare the data for clustering. To do this, I again one-hot encoded the venues dataframe but this time grouped entries by their Station and took the mean of each venue category for the station to retrieve that categories frequency. With this new information obtained, I was able to then sort the venue categories by way of popularity per each station. This sorted dataframe can be seen in **Fig 8**.

	Station	1st Most Popular Venue Type	2nd Most Popular Venue Type	3rd Most Popular Venue Type	4th Most Popular Venue Type	5th Most Popular Venue Type	6th Most Popular Venue Type	7th I Pop Ve
0	Bathurst	Coffee Shop	Sandwich Place	Pizza Place	Ramen Restaurant	Restaurant	Korean Restaurant	Fried Chicke Joint
1	Bay	Coffee Shop	Bakery	Café	Japanese Restaurant	Gourmet Shop	Burger Joint	Ethiop Restau
2	Bayview	Coffee Shop	Café	Burmese Restaurant	Mediterranean Restaurant	Restaurant	Burger Joint	BBQ J
3	Bessarion	Persian Restaurant	Sandwich Place	Pizza Place	Fish Market	Breakfast Spot	Buffet	Coffee Shop
4	Broadview	Coffee Shop	Pizza Place	American Restaurant	Mexican Restaurant	Ramen Restaurant	Fast Food Restaurant	Falafe Restau

Fig 8

With the data now in a more usable format, it was time to determine the optimal number of clusters for K-Means Clustering. To solve this, I used the Elbow Method which can be seen in **Fig 9**.

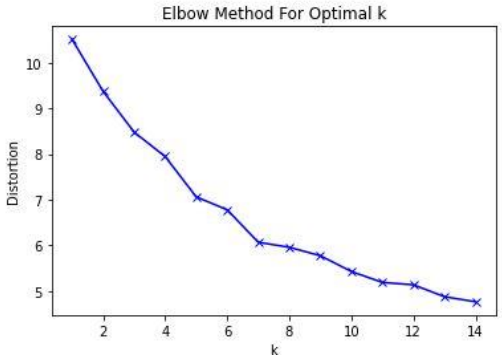

Fig 9

Upon completing this method, it appeared evident that 7 clusters would be optimal. I ran the KMeans algorithm provided by the Sklearn library on the venue frequency dataframe to retrieve the cluster labels and appended a new column called ‘ClusterLabels’ to the sorted dataframe. Stations in the same cluster are stations with similarly popular venue categories. With the stations now being clustered appropriately, I mapped them, and colour coded the stations based on their cluster. This map can be seen in **Fig 10**.

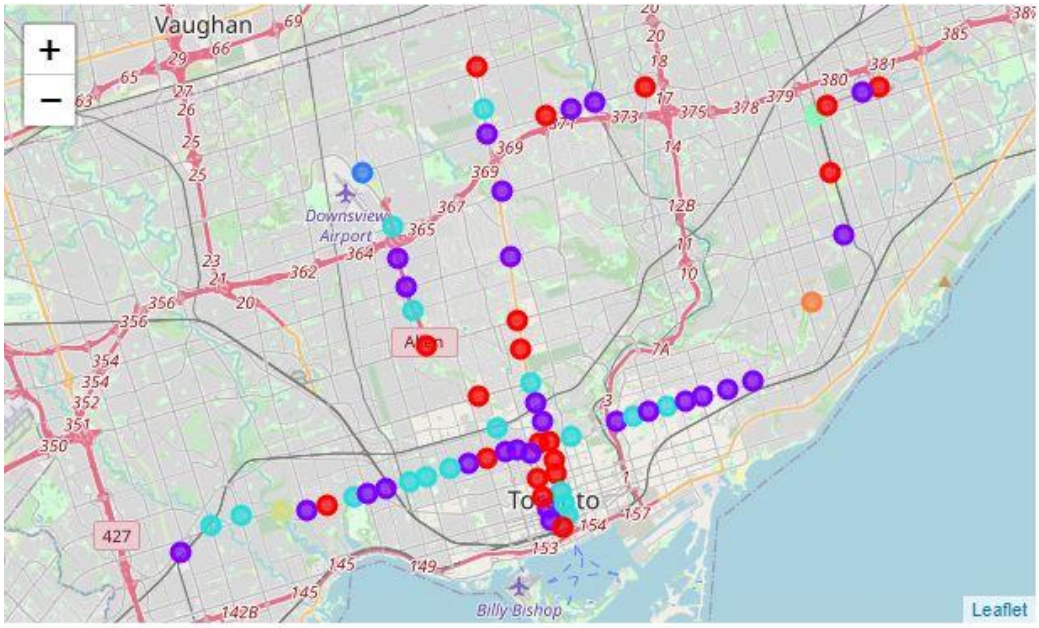

Fig 10

After analyzing the clusters, it was apparent that the first cluster was dominated by coffee shops so it could be ruled out of possible stations for expansion. It was also very clear that clusters 3, 5, 6 and 7 were made up entirely of outliers so these clusters were ruled out as well, leaving clusters 2 and 4. To determine which cluster of the two should be focussed, I took the average ranking of coffee shops for each cluster. The average rank of coffee shops in the second cluster was about 4.385 whereas the average rank in the fourth cluster was 6.333. This meant that on average, coffee shops in the second cluster outperform those in the fourth cluster and as a result the second cluster will be focussed. My final step was to drop the stations within the second cluster that already contained a coffee shop and map the resultant stations. This map can be seen in **Fig 11**.

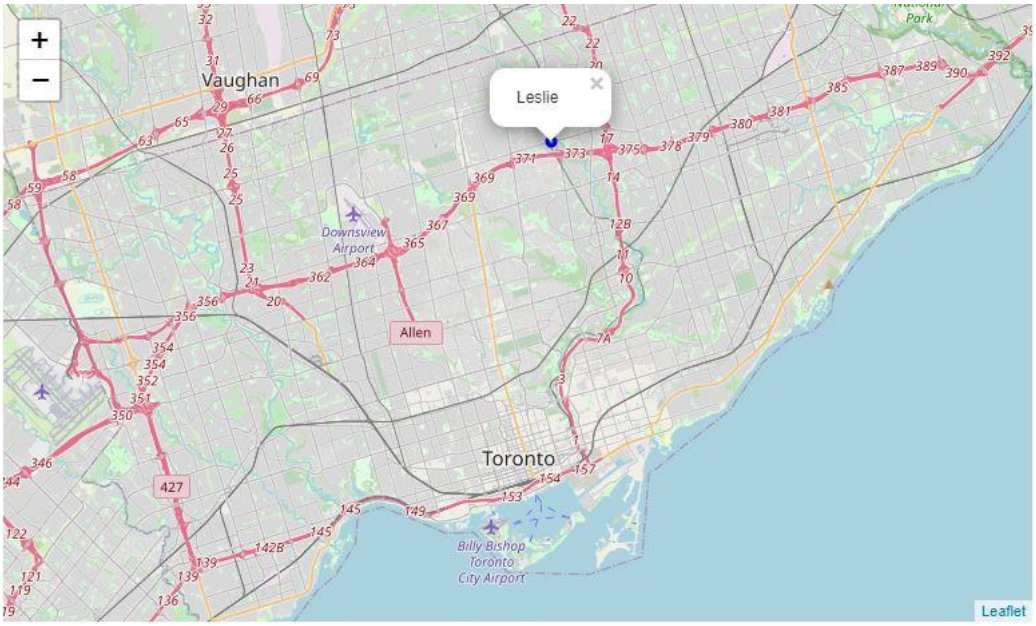

Fig 11

Results/Discussion:

Through the project, I was able to determine a variety of things. First is that, coffee shops are the top performing venue category in close proximity to subway stations therefore, stations should be focussed for possible expansion. The second thing I was able to determine was the rankings of venue category for each station and lastly, I determined that Leslie Station would be the optimal place for a coffee shop to open a new location at. These results could be affected by factors such as construction occurring in the nearby area, a lack of commercial properties within the given radius, etc. Based on these results, I highly recommend a coffee shop to expand within 250 meters of Leslie Station as coffee shops perform very well in other stations of the same cluster.

Conclusion:

For this project, I explored the relationship between subway station and their surrounding venues as well as analyzed what type of venue categories perform well in close proximity to subway stations. I then used the K-Means clustering algorithm to group stations with similarly popular venues in order to determine a location for a coffee shop to expand to in which it is likely that the coffee shop will become popular. This research can be of use to anyone who is a stakeholder of a coffee shop because choosing a good location for expansion will have great financial benefits.