Subway Surfing:
Where Should I Start My House Hunt?

IBM Data Science Capstone Project
Prepared By: Ethan Goldstein

Introduction

I am currently living in the outer suburbs of Toronto, and my commute to work is driving me insane. I would love to find a new place to live, ideally right on a subway line for an easy commute. Transfers between subway lines are okay, but I don't want to have to take a bus or any other form of transit besides subways.

So, which subway stop should I centre my house hunting on?

My ideal place to live would be within walking distance of a variety of grocery stores, pubs, parks, and gluten free restaurants. I love to read, and would prefer if at least one public library was within walking distance. The total count of these venues will be the primary metric I use in deciding which subway station should be at the center of my property search.

Data

Walking Distance

I am a moderately fast walker, and can walk ~1500 meters in about 15 minutes. I consider a 15 minute walk short enough to do complete without a second thought (or an urge to call an Uber/taxi), so will be using a distance of 1500 m as the 'walking distance' throughout this excercise.

Subway Stations

Data attribution*: Contains information licensed under the Open Government Licence – Toronto https://open.toronto.ca/dataset/ttc-routes-and-schedules/*

I will begin by exploring the TTC's (Toronto Transit Commission) stop locations, which is provided through the City of Toronto's Open Data Portal. The City publishes a GTFS (General Transit Feed Specification) file, which contains route definitions, stop patterns, stop locations, and schedules. More information about this file format can be found in the Google Transit API Documentation: https://developers.google.com/transit/gtfs/reference.

My primary interest is in the subway stop data, which is stored in the *stops.txt* file. The TTC's file contains the stop ID, code, name, latitude, longitude, and whether or not the stop supports wheelchair boarding for all modes of transit. Bus and streetcar stop data will have to be removed before we proceed. As well, the file contains separate coordinates for each subway station platform. Most stations will have two sets of latitude and longitude (northbound and southbound or eastbound and

westbound), while interchange stations may have four sets of coordinates. I have averaged each platform's coordinates to provide one overall latitude/longitude for each subway station, saved in the *TTC_Subway_Stops_cleaned.csv* file.

Venues

I will use Foursquare's location data to find the venues I am interested in that are within 1.5 km of each subway station. Specifically, I will use the Foursquare Places API *venue search* function. Details of thes API calls may be found in the Foursquare Developer documentation: https://developer.foursquare.com/docs/api/venues/search

Methodology

Cleaning The TTC Data

I began by cleaning the *stops.txt* file and transforming it into *TTC_Subway_Stops_cleaned.csv,* as described in the previous section. This exploratory analysis was done in Microsoft Excel, using the following process.

1. During my first look at the data, I noticed that the stop descripion for all subway platforms was in the format "*name* – *direction* PLATFORM ", where *name* is the name of the subway station and *direction* is the direction of travel on that platform. With this in mind, I removed all results that did not have the word "PLATFORM" in the description.
    a. Streetcar platforms that remained were removed from the data.
    b. One interchange station was shown under two names (BLOOR and YONGE). These were merged into BLOOR-YONGE station.
2. Using a combination of the IFERROR, TRIM, LEFT, and FIND formulas, the station name was copied into a new column. This will serve as the index for calculating the final latitude and longitude.
3. This station names column was copied to a new sheet, and duplicates were removed.
    a. The station count was checked against the numbers provided on the TTC's website (http://ttc.ca/Subway/index.jsp) and found to match.
4. An AVERAGEIF formula was used to calculate the average latitude and longitude for each station. The resulting worksheet has been saved as *TTC_Subway_Stops_cleaned.**csv***

Please see *TTC_Subway_Stops_cleaned.**xlsx*** file for the exact formulas used. This file is in the same Github folder as this report.

Mapping the Subway Stations

We can now plot the stations on a map:

Figure 1 - TTC Subway Stations

Furthermore, we can demonstrate walking distance by plotting a circle with a radius of 1.5 km around each station:

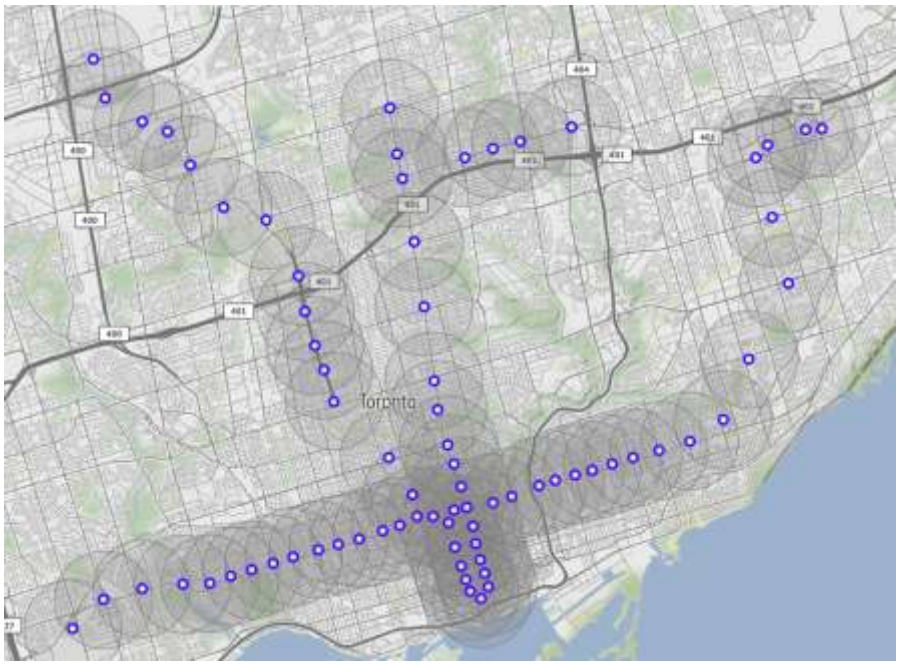

Figure 2- 1500 m radius of TTC Subway Stations

Venues from Foursquare

Next, I queried Foursquare for a list of interesting venues surrounding each subway station. This was easily completed using the /venues/search API call. You simply need to include your credentials, the starting location, radius of search, and category ID in the call to return a list of matching venues. The query was repeated for each venue type and each subway station, for a total of 304 requests.

For this analysis, we do not need venue specific data like their name or address; only the number of venues, by category, that are within walking distance of each subway station is required. As such, only the record count of each query was saved.

At this point, subway stations that have no libraries within the defined walking distance were removed from the result table.

Expanding the Scope by Clustering Neighbourhoods

While the number of venues surrounding each subway station should be enough information to meet our needs, there may be no available housing near that station. it could be beneficial to examine similar stations, grouped together in clusters.

The subway station / venue counts scraped from Foursquare were normalized using the Standard Scalar process, and stations were clustered using the K-Means algorithm. The optimal value of k = 3 was determined using the elbow method so the subway stations were clustered into three distinct groups.

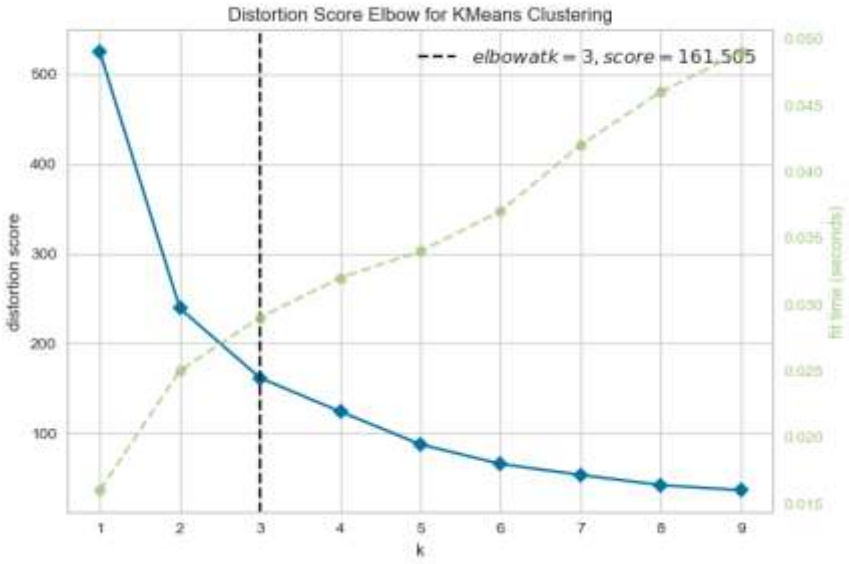

Figure 3 - Elbow Method Clustering of TTC Subway Stations

Figure 4 - TTC Subway Stations clustered to three groups

Results and Discussion

Cluster 0 – Inner Suburbs

This cluster contains 28 subway stops in the 'Inner Suburbs' of Toronto, that is, the lower-rise, more suburban areas of the city. There are likely to be less amenities around these stations, and the number will decrease significantly the further away from the station you are. These stations are shown in blue on the map above.

Stations include: BAYVIEW, BESSARION, DON MILLS, DOWNSVIEW PARK, ELLESMERE, FINCH, FINCH WEST, GLENCAIRN, HIGHWAY 407, KENNEDY, LAWRENCE, LAWRENCE EAST, LAWRENCE WEST, LESLIE, MCCOWAN, MIDLAND, NORTH YORK CENTRE, PIONEER VILLAGE, SCARBOROUGH CENTRE, SHEPPARD WEST, SHEPPARD-YONGE, VAUGHAN METROPOLITAN CENTRE, VICTORIA PARK, WARDEN, WILSON, YORK MILLS, YORK UNIVERSITY, YORKDALE

Cluster 1 – The Core

Includes the downtown core, these 19 subway stations are in the heart of the city and will have the greatest concentration of amenities within walking distance. These stations are shown in red. The largest differentiating factors between these stations are the count of Gluten Free Restaurants and

Pubs; the stations that form a 'U' through the downtown core are within walking distance of the most venues. I will begin my house hunt by examining this cluster of stations.

Venue Count of Cluster 1

Stop Name	Libraries	Gluten Free Restaurants	Grocery Stores	Pubs	Parks	Total
DUNDAS	32	13	50	50	49	194
QUEEN'S PARK	30	10	49	50	48	187
COLLEGE	29	11	48	50	47	185
OSGOODE	22	14	49	50	48	183
ST PATRICK	22	13	49	50	48	182
QUEEN	19	13	48	50	48	178
ST ANDREW	19	13	49	50	46	177
MUSEUM	24	5	50	47	48	174
WELLESLEY	24	4	49	47	47	171
UNION	15	11	45	50	46	167
KING	18	11	44	47	46	166
BAY	20	4	50	41	49	164
ST GEORGE	19	4	50	39	49	161
BLOOR-YONGE	15	3	50	39	49	156
SPADINA	18	2	50	28	49	147
BATHURST	14	3	50	28	50	145
DUPONT	15	2	45	26	49	137
ROSEDALE	14	2	41	29	48	134
SHERBOURNE	12	2	42	27	48	131

Stations include: BATHURST, BAY, BLOOR-YONGE, COLLEGE, DUNDAS, DUPONT, KING, MUSEUM, OSGOODE, QUEEN, QUEEN'S PARK, ROSEDALE, SHERBOURNE, SPADINA, ST ANDREW, ST GEORGE, ST PATRICK, UNION, and WELLESLEY.

Cluster 2 – Midtown / Bloor Danforth

Includes 28 subway stops in Midtown Toronto and the stops on Line 2 – Bloor Danforth that are outside the downtown core. While primarily serving low-rise neighbourhoods, the street around these stations is very walk-able. Overall, this cluster would be my secondary choice for house hunting.

Stations include: BROADVIEW, CASTLE FRANK, CHESTER, CHRISTIE, COXWELL, DAVISVILLE, DONLANDS, DUFFERIN, DUNDAS WEST, EGLINTON, EGLINTON WEST, GREENWOOD, HIGH PARK, ISLINGTON, JANE, KEELE, KIPLING, LANSDOWNE, MAIN STREET, OLD MILL, OSSINGTON, PAPE, ROYAL YORK, RUNNYMEDE, ST CLAIR, ST CLAIR WEST, SUMMERHILL, and WOODBINE.

Conclusion

Of the seventy five subway stations in Toronto Subway system, three were not within 1.5 kilometers of a library and were excluded from further analysis. The remaining stations were clustered into three distinct groups, based on the number of Pubs, Gluten Free Restaurants, Grocery Stores, and Parks within walking distance of the stations. The cluster containing subway stations in and around the city's downtown core has the greatest number of these venues.

As Dundas Station is surrounded by the most venues, this is the subway station to focus my house hunt around. To expand my search, I can look for housing around any of the subway stations in Toronto's downtown core, as these all have a wide variety of venues that interest me nearby.