Capstone Project - The Battle of Neighborhoods - Coursera - Applied data science - IBM

Analyzing Subway Stations in Toronto

1.Introduction

Toronto's main public transportation system is operated by the Toronto transit Commission (TTC). The backbone of its public transport network is the Toronto subway system, which includes three heavy-rail rapid transit lines spanning the city, including the U-shaped Line 1 and east–west Line 2 [1]. Additional two smaller lines (3 and 4) are also part of the subway network in Toronto. Note that the city (municipality) of Toronto includes, beside the original city of Toronto, also North York, Etobicoke and Scarborough.

The stations are characteristically different one from the others. In particular, the purpose (or characteristic) of the subway station may be related to the nearby central venues. For example, subway stations near the university have different features than stations in rural areas. These lead us to the first stage of the analysis, which is to examine the present subway network.

The transit score in general, and the vicinity to subway stations in particular may be one of the important features in the wellbeing of the citizens of the Great Toronto Area (GTA). Findings suggest that low-income households, on average, see greater improvement in travel time savings to destinations in Toronto than other groups and the overall population [2].

Ideally, the subway network, together with other TTC services (e.g. buses, streetcar), is expected to cover the following areas. The departure stations are located in dense areas, and in low-income households neighborhoods (where one cannot afford a car for example), and destination stations are in the vicinity of business areas, shopping malls, tourist popular venues, or universities.

Analyzing the present state of the GTA's subway network, together with spatial and demographic features of the neighborhoods themselves will help the leader to decide

(where assume limited sources) whether this current situation is satisfactory or that urgent new subway stations or lines must be added.

2. Questions and Data to be Used

(i) Goal:

The main questions surface in this analysis are:

1) What are the main features of the subway stations? (e.g. the group might be end-stations, transfer to other lines stations, stations near unique features such as airpost, university, malls)
2) How are the different groups of stations spatially distributed?
3) Is there a relation between the demographic features of the neighborhoods the subway crosses them?
4) Can we suggest improvements or additional subway stations that need to be built using the data from the above questions?

(ii) Data and Methods:

A: Clustering Subway Stations Using Nearby Venues

To analyze the current state of the subway network, together with the different features of each station, I will use the FourSquares database. First, I will extract from FourSquares the subway stations in Toronto. Next, I will explore the nearby venues for every subway station found in the previous step. In the last step of the examination the current state, I will cluster the stations using similarities on their nearby values.

B: Clustering Subway Stations Using Demographic Features of the Neighborhood

To explore the current state of the subway network, in relation to the demographic features of the population nearby, I will use data published in Wikipedia [3], where I will concentrate on two features of each neighborhood; yearly income, and its density (i.e population/areal unit). Here, I want to examine whether the current state if satisfying. Here,

since we expect that low income households communities and dense neighborhoods will be more beneficial from having a subway station nearby, we hope to find these neighborhood transit scores to be high.

C: Clustering Toronto Neighborhoods Using Demographic Features

Here I assume that the current state of the subway network and the number of stations are not sufficient and some improvement must be done. By exploring the most dense neighborhoods or neighborhoods with low-income households, we can suggest improvement to the current routes, and to detect what neighborhoods that need (relatively urgently) new stations. Since [3] includes demographic data by neighborhoods' names, I also use [4] to find the location of each neighborhood.

3. Methodology

(i) Data

I find the subway station location using FourSquere . I use only the relevant data for us; the name of the subway station, and some location data (address, postal code, latitude and longitude) .

	name	location.address	location.crossStreet	location.lat	location.lng	location.postalCode
0	Osgoode Subway Station	250 University Ave.	at Queen St. W.	43.650877	-79.386824	M5H 3E5
1	Queen Subway Station	171 Yonge St	at Queen St	43.652373	-79.379191	M5C 2L7
2	Dundas Subway Station	300 Yonge St	under Dundas St	43.656096	-79.380785	M5G 2B3
3	St Patrick Subway Station	449 University Ave.	at Dundas St. W.	43.654818	-79.388331	M5G 1W8
4	St Andrew Subway Station	147 University Ave	at King St W	43.647773	-79.384939	M5H 1J9

Table 1: Example of the data of subway locations given from FourSquares.

Next to each subway station we use FourSquare to find its nearby venues. We cluster all subway stations into 5 clusters using the common nearby venues. This is given in Table 2.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	0	Bathurst Subway Station	Korean Restaurant	Grocery Store	Bakery	Ice Cream Shop	Mexican Restaurant	Coffee Shop	Eastern European Restaurant	Japanese Restaurant	Bubble Tea Shop	Burrito Place
1	2	Bay Subway Station	Boutique	Italian Restaurant	Coffee Shop	French Restaurant	Hotel	Restaurant	Japanese Restaurant	Spa	Café	Sushi Restaurant
2	2	Bloor-Yonge Subway Station	Coffee Shop	Italian Restaurant	Café	Spa	Sushi Restaurant	Hotel	Japanese Restaurant	Yoga Studio	Boutique	Restaurant
3	2	Broadview Subway Station	Pub	Coffee Shop	Burger Joint	Restaurant	Pizza Place	Convenience Store	Sandwich Place	Café	Rental Car Location	Ramen Restaurant
4	4	Castle Frank Subway Station	Convenience Store	Park	Nightclub	Outdoors & Recreation	Metro Station	Deli / Bodega	Dance Studio	Fish & Chips Shop	Filipino Restaurant	Fast Food Restaurant

Table 2: Example of the data of nearby venues of subway stations, and their cluster using this nearby common venues.

To characterize the neighborhoods of Toronto, I use the data from Wikipedia [3]. This gives us the population, density, Average income and Transit Commuting percentage for each neighborhood.

Since the demographics of neighborhoods do not include their location, I merge the DataFrame given from [3] with the data from [4], thus my new table will include the locations.

	Name	Population	Density (people/km2)	Average Income	Transit Commuting %	Latitude	Longitude
1	Agincourt	44577.0	3580.0	25750.0	11.1	43.794200	-79.262029
7	Bathurst Manor	14945.0	3187.0	34169.0	13.4	43.754328	-79.442259
9	Bayview Village	12280.0	2966.0	46752.0	14.4	43.786947	-79.385975
11	Bedford Park	13749.0	6057.0	80827.0	15.2	43.733283	-79.419750
13	Birch Cliff	12266.0	3525.0	48965.0	11.4	43.692657	-79.264848

Table 3: The location and demographic data for each neighborhood.

(II) Methods

(A) Clustering

We use k-clustering (K=5) to characterize the subway station in Toronto. To do so, I use the frequency of each venue as a location in the multi-dimensional space (each dimension refers to a different category of venues, e.g. Cafe, bank etc.).

(B) Binning

I binned the average outcome for a given neighborhood for 10 equal sized bins. Therefore, every neighborhood is mapped into 10 sets. Then the data is presented with a color scale between red (low average income) and blue (high average income). In Statistics it is called deciles.

(C) Geographical Mapping

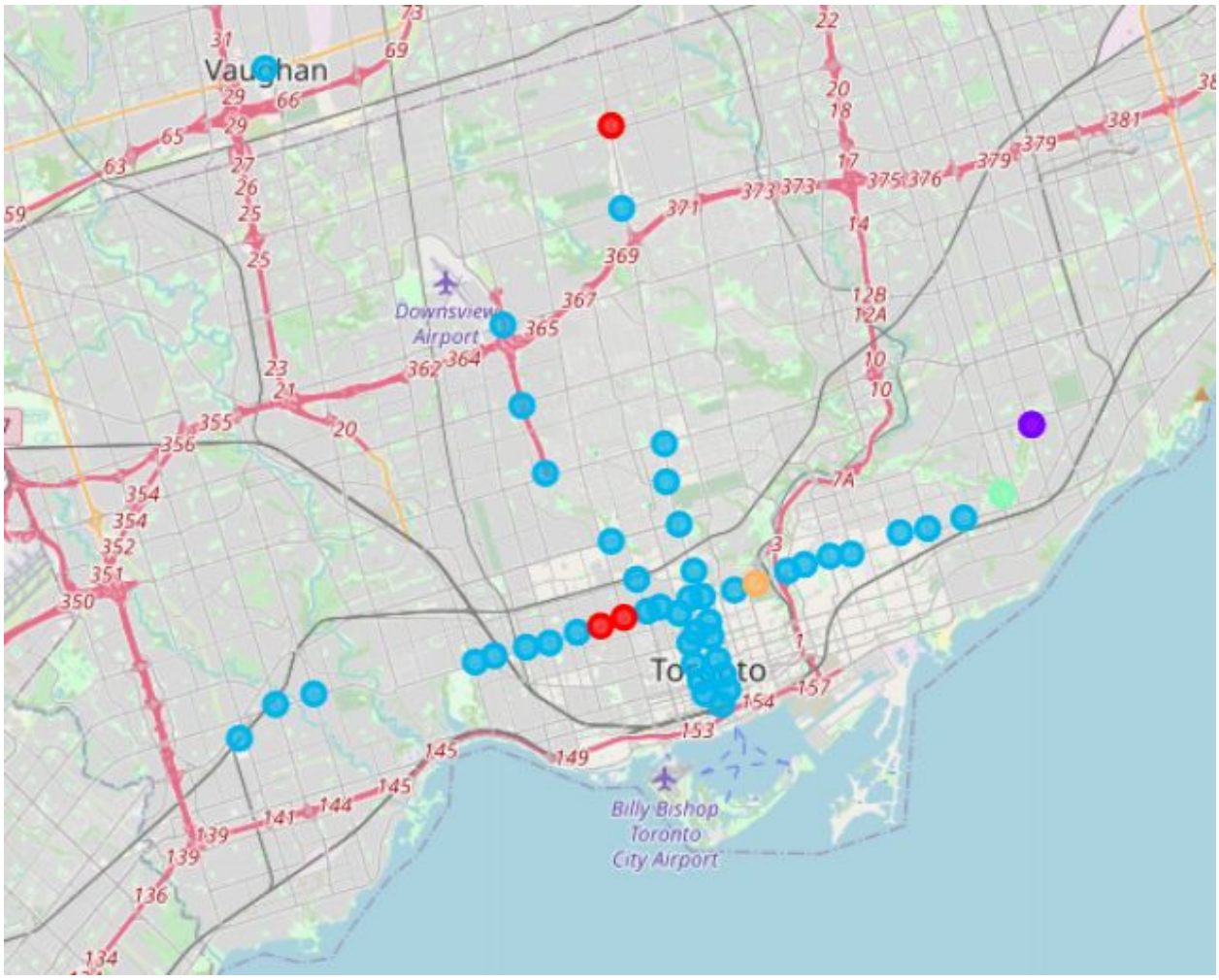
Using Folium maps to analyzing the geographical pattern

4. Results

(i) Subway stations locations and their clustering using nearby venues frequency

Map 1: The locations of the subway station in Toronto.

To see whether there are some patterns on the subway stations using the nearby venues. We cluster the stations into 5 clusters. The results for the 5 different clusters are given the following map.



map 2: The subway stations in Toronto after clustering them into 5 clusters using their nearby venues frequency. Each color represents a different cluster.

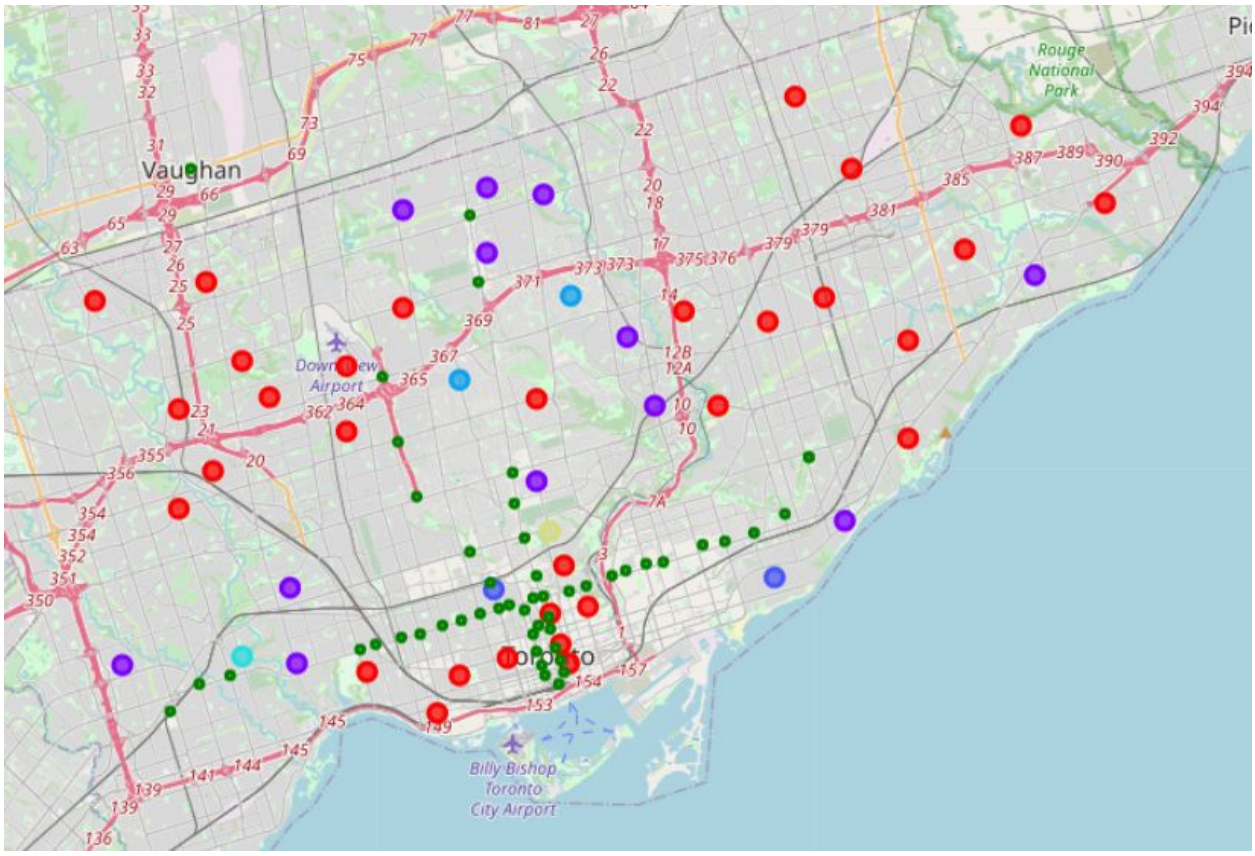
After clustering the subway stations using nearby venues, we conclude that most stations share similar nearby common venues.

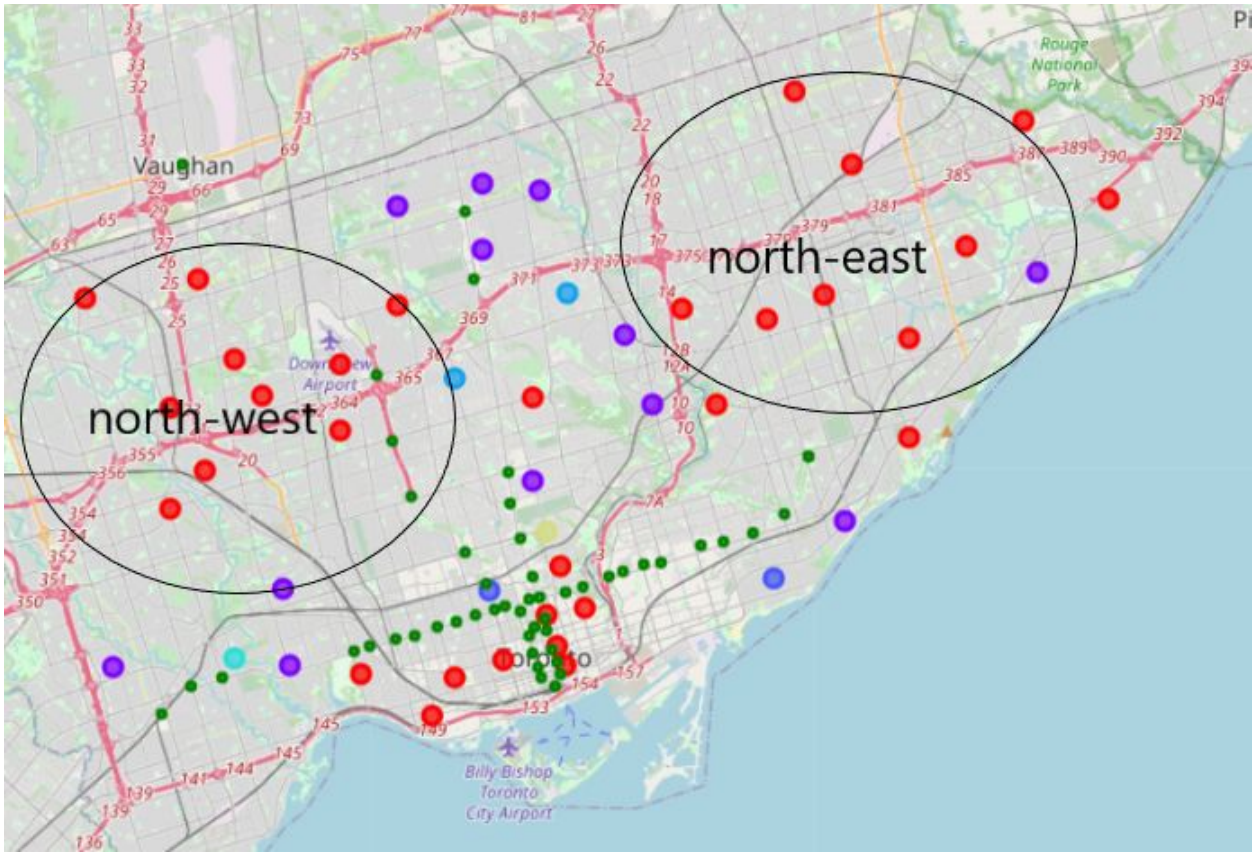
In the largest cluster (marked with blue), the most common venue near the subway stations is mostly cafes or coffee shops (where in some stations it is the 2nd most common). Then, in the purple cluster, which contains only one subway station, the most common venue is a

bakery. In the red cluster the most common venue is a Korean restaurant, which suggests a relatively high percentage of koreans in these areas. The other two clusters also contain one subway station each. The common venue in the green cluster is a gas station and the second most common is a park, and in the orange cluster we obtain that convenience store is the most common venue and the second most common is a bank.

The suggested conclusions from the above results are given in the following. First, the most common venue near the subway stations is a cafe (or coffee shops). Second, the red, purple, green and orange stations (see map above) have unique features. The areas near red subway stations may be populated with a larger percentage of Koreans (compared to other stations). Furthermore, the green (Victoria Park Subway Station), and purple (Warden Subway Station) do not seem to be located in highly populated areas, and might be used for transfer through the stations. The last statement relies on the fact that in the most common venues list there is a bus station, gas station, intersections etc. Last, In the orange cluster (Castle Frank Subway Station), the 3 most common venues are convenience store, park, and a night club, thus this station has some unique features.

(ii) Subway Stations Locations Compared to Neighborhoods Averaged Income


Vaughan
Rouge National Park
Billy Bishop Toronto City Airport



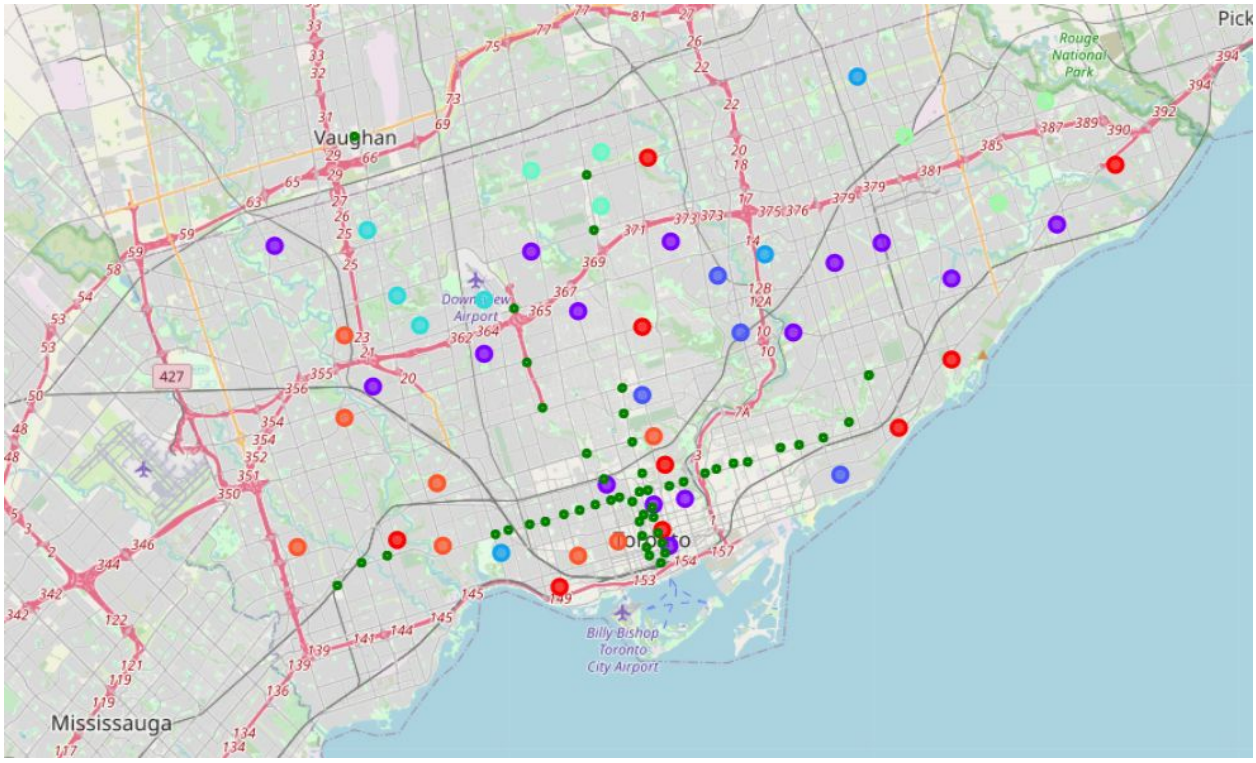
map 3: The neighborhood in Toronto presented in different colors. Eech color represents the decile of average income. The color scales between red (lowest decile) to blue (highest decile). The green dots represent the subway-stations locations. The above panel shows the raw map, In the lower panel, the two regions; the north-east and the north-west part of the city are emphasized.

Map 3 shows us that the downtown area and the north-east and north-west of the city has lower average income. (remind that the scale from red to blue represents low to high average income (respectively). The subway stations are given with small green dots.

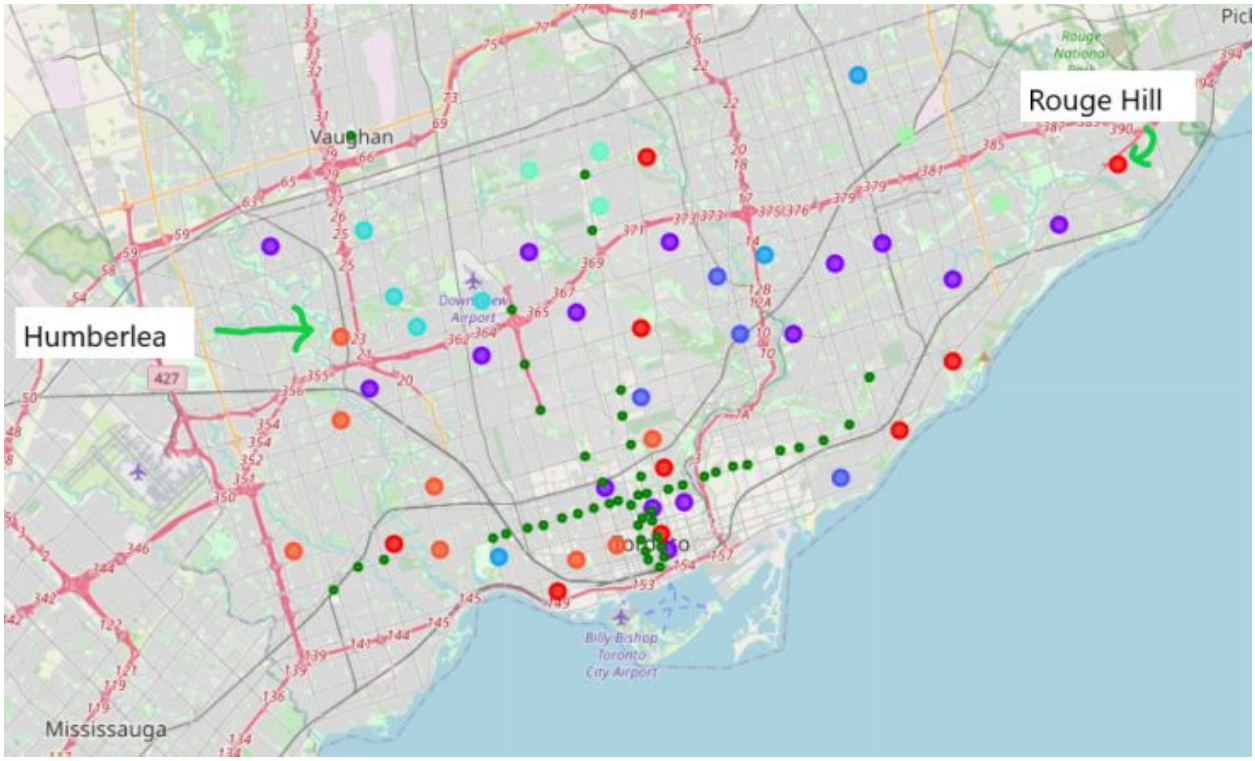
Clearly, two most-needed regions (north-east and north-west neighborhoods) do not have access to subway service, see Map 3 lower panel. In addition, the current subway lines cross rear mid- and high-income neighborhoods.

To clarify, I note that there might be some (temporal) correlation between the lower-income neighborhoods to lack of access to subway lines. This correlation might be accumulated during the years, e.g. a new subway line makes house prices increase such that poorer population will not be able to continue to live in this area. I do not examine such temporal effects.

In Map 4 I present high (red) and low (blue) population density neighborhoods. It shows that most dense neighborhoods have access to the subway. However, some high density neighborhoods, including for example Humberlea or Rouge Hill, have no (reasonable) access to subway stations, see for example Map 5.



map 4: The density of population (population/area) for a given neighborhood given in color scale between red (high density) to blue (low density).



Map 5: Two examples for neighborhoods with high density of population but with no access to the subway network.

5. Discussion, Summary and Conclusions

I have shown the geographical locations of the subway stations in Toronto, and found some stations with uniques features. In addition, I have found that low-income regions, north-east and north-west, have no access to subway stations even though they are in much need. This is, in contrast, with comparison to many neighborhoods next to the subway, with higher income on average. Thus, I suggest to connect these areas to the subway network. Moreover, I have found some neighborhoods with high density populations, where connecting these neighborhoods to the subway network will be highly beneficial.

As a general comment I note that all the data here is given from free sources, thus the accuracy of the data is not guaranteed and must be taken with caution.

References

[1] https://en.wikipedia.org/wiki/Public_transportation_in_Toronto

[2] http://www.metrolinx.com/en/docs/pdf/board_agenda/20190912/20190912_BoardMtg_Ontario_Line_Initial_Business_Case_EN.pdf

[3] https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

[4] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M