**Capstone Project - Battle of Neighborhoods**

1. **Introduction:**

- A real estate company in the United States is looking to expand its market, and they decide to go with Toronto. Their plan is to develop a system that recommends the next perfect house for their customers based on their preference.

1. **Data:**

We will be using:

• Wikipedia ( https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Cana da:\_M) to get the data about the real estate in Toronto

• Geospatial data for Toronto (http://cocl.us/Geospatial\_data) to get the geographical coordinates of each postal code

• Foursquare API to obtain more information about venues

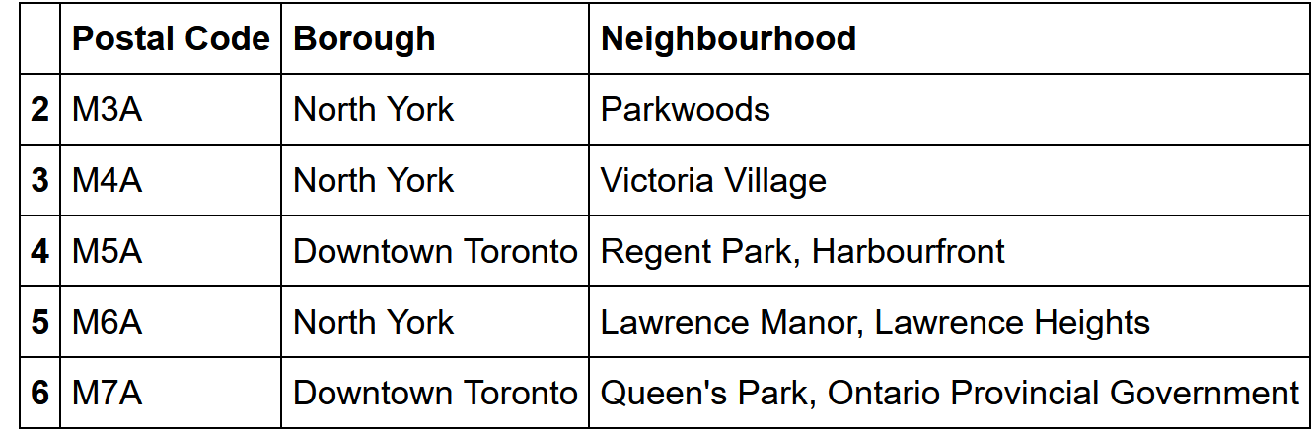
• Random user data, with a random number of preferences to check how our system works and to improve it.

1. **Methodology:**

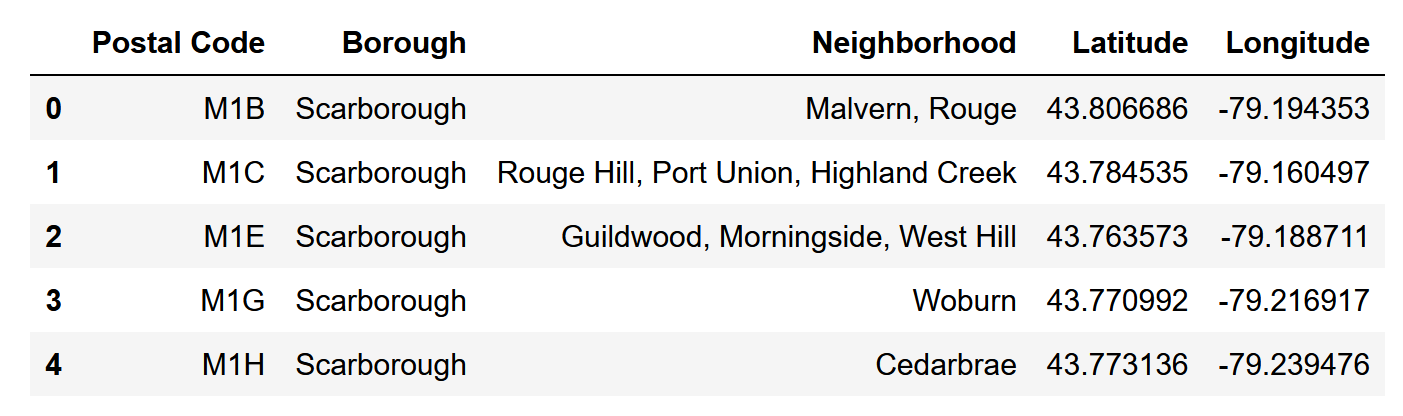
In this project, we will create a recommender system to help users find their perfect home.

* 1. Preparing and obtaining data:

First, we scrape the data we need about Toronto housing situation through Wikipedia and then refined the data by removing data that is labeled “Not assigned”. Then, we substitute unnamed neighborhoods for the name of relevant borough. After that, we group the table by the postal code the result is shown in the table.



The next step is to incorporate this table with the geographical coordinates of each postal code to our data. The result is shown in the next table



* 1. Random user:

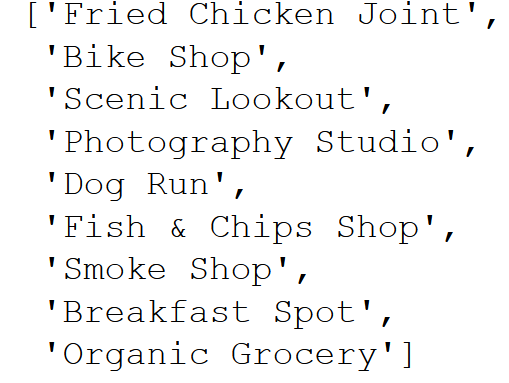
To generate a random user, we will get a list of available categories in the city and select a random number from 1 to 10. Then, from the list of categories we will sample the same amount obtaining the list of categories that our user will have interest in. Now we create a table with the categories as the columns and one row, where the values are 1 if the user has that category in his list and 0 for vice versa. This will result in a user profile that will be used in the recommendation system.

* 1. Recommendation System:

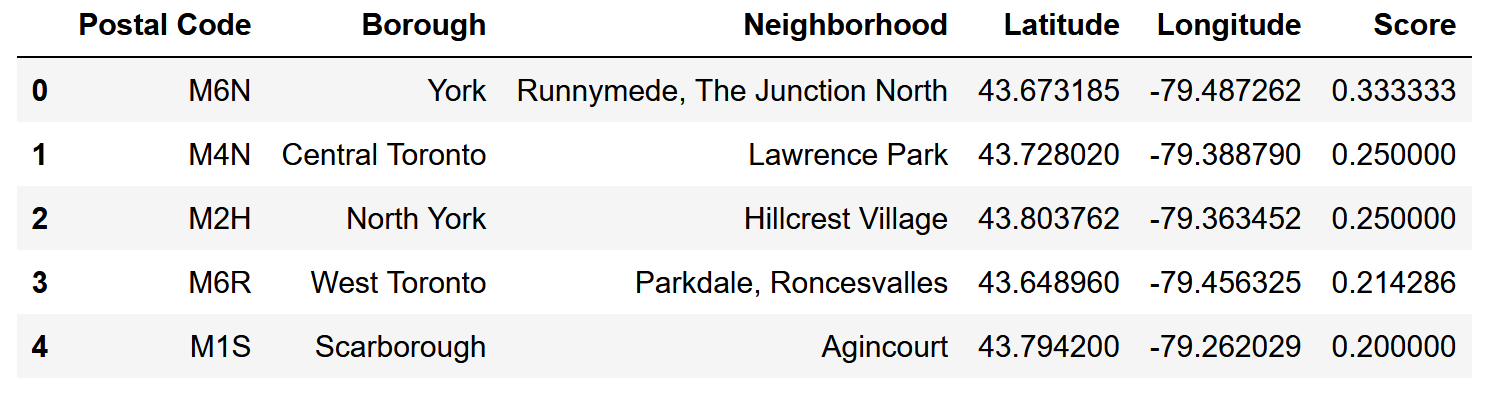
To make a recommendation system, we compare our user profile to the table with the neighborhoods and the mean of value for the amount of venues of each category in it. So we multiply both matrix and apply a sum for each row. As the result we get a new matrix with the neighborhoods and the score for each one of them. The higher the score the better the neighborhood matches the user’s interests. If we merge this table with the Table 2. We will be able to print in a map where are the better neighborhoods for our user.

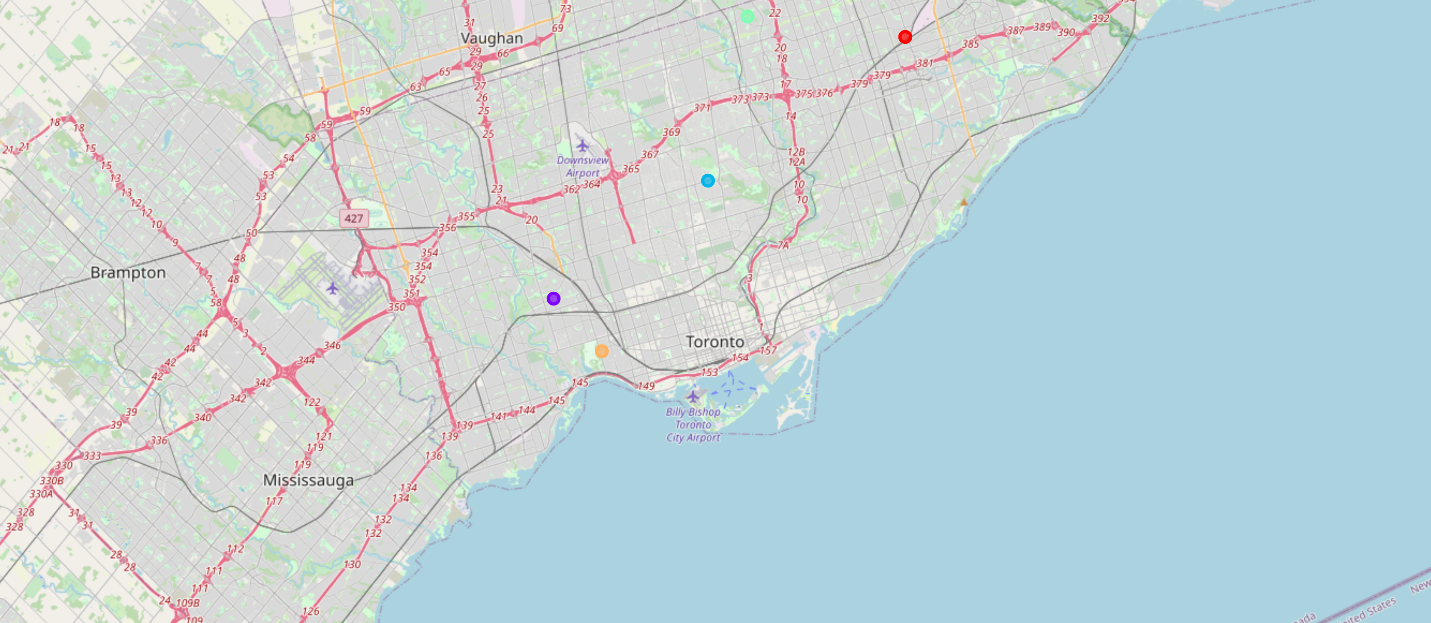
1. **Result:**

For the prototype, this user’s preference is



With this user’s preference, we got the below table, and a map, representing the top 5 suitable destination





1. **Discussion:**

From the table above, there are two good neighborhoods for our user are “Runnymede, The Junction North” and “Lawrence Park”. From the table, we can observe that the difference between the scores of each place is not significant and the reason behind that, maybe, because the user’s preference is common, and do not contain any outlier such as “Airport food court”

1. **Conclusion:**

This is a sample of a content-based recommender system but still need more development. We need more data in order to improve our accuracy, and also make it available with rural places. A lot of work still need to be done to improve this.