**Choosing the best neighbourhood to move to in Toronto**

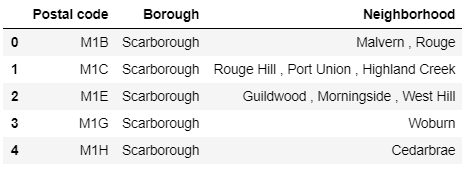
Description Miguel lives in the North-West side of Toronto in Downsview where he enjoys what his neighbourhood has to offer, such as bars, restaurants, cafes and parks. He recently received a job offer in South-East Toronto at the Studio District, the career prospects at this new company represent what he was looking for and he is interested in taking this job. The problem is Miguel now has to move closer to his new job as commuting if not feasible from the place he lives as the salary is high enough he is not concerned about the rent prices, he is just focused on his social life thus venues and facilities nearby is what is the most important for him. There are plenty of options in the surrounding areas but he would like to choose the perfect location with venues similar to the ones of his neighbourhood.

To solve this problem Miguel is thinking about getting a list of all the possible neighborhoods in the city where he can move to, data regarding to the venues and facilities. This data will be taken from Foursquare where he will pull all the venues for each neighbourhood on Miguel’s options. Based on this data, he will build a classification model to help him determine the most suitable location for him. Venues data: <https://foursquare.com/>

#### Neighborhood Grouping

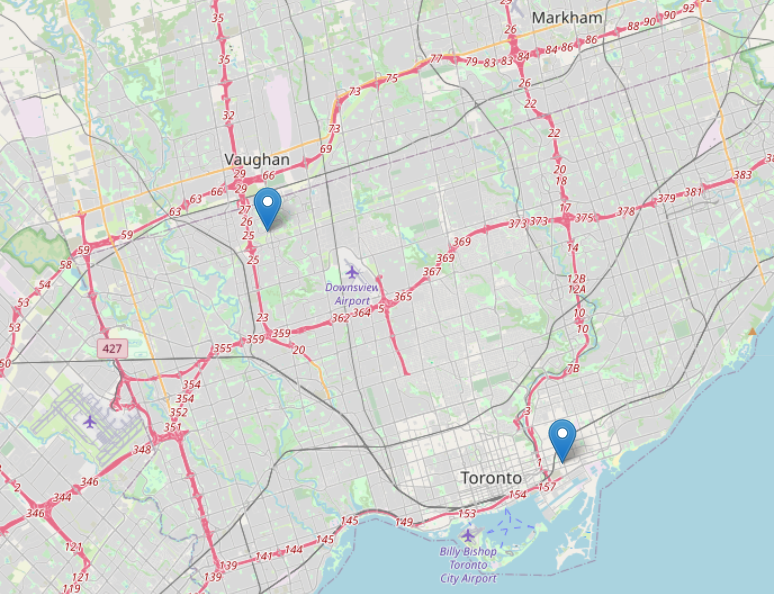
The data corresponding to the neighborhoods in Toronto is taken from Wikipedia. The data extracted is then cleaned and each neighborhood is grouped according to its respective Postal Code

Data Source: wikipedia

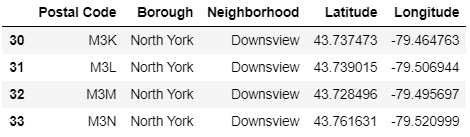


#### Exploring Miguel's Neighborhood

Downsview is located in the north west part of Toronto and Studio District in the south east part of the city



Four postal codes correspond to Downsview



### **Extracting data from Foursquare**

Location data is data describing places and venues, such as their geographical location, their category, working hours, full address, and so on, such that for a given location given in the form of its geographical coordinates (or latitude and longitude values) one is able to determine what types of venues exist within a defined radius from that location. So, for a given location you will be able to tell if restaurants for example, exist nearby, or if schools, or parks, or gyms, or community centres exist nearby. Also, how many of each category exist and how each surrounding venue is reviewed by other people. So this is what's referred to as location data for this project Foursquare will be chosen fetch the Data, but what is Foursquare?

Foursquare is a technology company that built a massive dataset of location data. What is interesting about Foursquare is that they were very smart about building their dataset. They actually crowd-sourced their data and had people use their app to build their dataset and add venues and complete any missing information they had in their dataset. Currently its location data is the most comprehensive out there, and quite accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, and is currently being used by over 100,000 developers, and this number is only growing.

The following map shows all the Postal Codes within Toronto that will be considered by Miguel



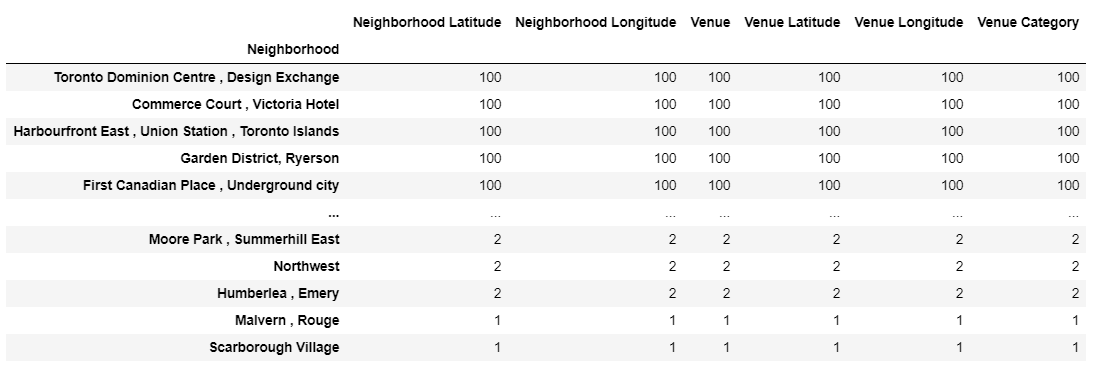
### Foursquare Data

After connecting to Foursquare's API the corresponding venues categories for each neighborhood were extracted as the sample below.

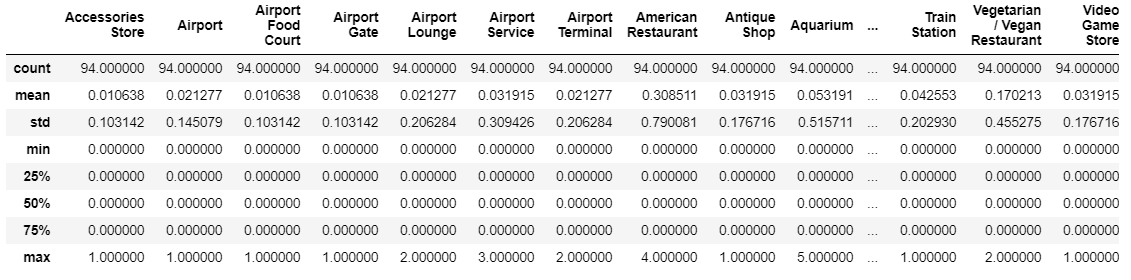


### Exploring the Neighborgoods

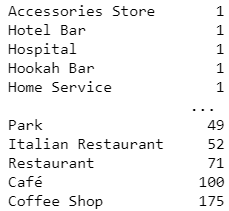
Based on the data, Miguel now wants to see what are the neigborhoods with the lowest and the largest amount of venues within the city.



Miguel Now decides to arrange the data in such way that he can visualize the total amount of venues by neighborhood. He performs certain statistical measures to gain some insights. He can now see that there is a high volatility in regards to the count of venues. This implies that there is significant difference within each neighborhood.

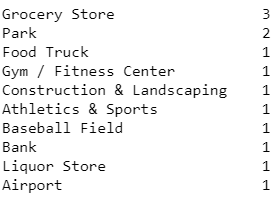


Miguel now wants to know what are the most popular and the least popular venues within the city. Based on the value counts he discovers that the Parks may not be as difficult to find as it is one of the most popular venues in the city.

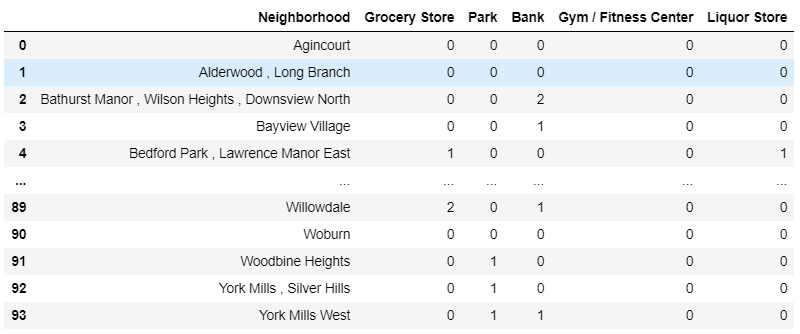


But what does Miguel's neighborhood has to offer?





Miguel considerts that what is really important for him to find in a Neighborhood is a Grocery Store, Park, Bank, Gym and a Liquor Store. Thus he clears the dataset to focus in these venues only.



### Problem solution hypothesis

To determine the possible neighborhoods that he would move into he decides to build a unsupervised learning model. This is given that there is no response variable Y and the aim is to identify the clusters with in the data based on similarity with in the cluster members. The algorithms he will to try to solve his problem are K-means and Hierarchical clustering.

Standardizing the dataset is essential , as the K-means and Hierarchical clustering depend on calculating distances between the observations. Due to different scales of measurement of variables, some variables may have higher influence on the clustering output. As such data is standarized as follows



In K-means, the number of clusters required has to be decided before the application, so some level of domain expertise would of help. Else we can use a scree plot to decide number of clusters based on reduction in variance



