**Battle of the neighbourhoods**

Manisha Rawat

8th August 2020

**1. Introduction**

**1.1 Background**

People love to travel one place another experiencing things they would have never experienced if they were home. Travelling gives us an opportunity to explore unknown destinations, have different cuisines, try new fun activities.

Travelling makes us confident and independent. It helps us to deal with unexpected situations. We can learn a lot from our exposure from a new recipe, to a new, more effective solution to an ordinary problem or a new way of creating something.

## **1.2 Problem**

In this capstone project we'll be comparing different places on the basis of the facilities, population, diversity, etc. they have. This project will take the data about a particular destination and explore the places around it. You'll be able to compare two or more than two places on the basis of your interest.

**1.3 Interest**

The capstone project targets the people who loves travelling, are planning to shift to other location, are opening a new restaurant or any kind of business. Also it will be helpful for the residents to compare their key facilities with other parts of the country.

**2. Data**

**2.1 Data sources**

**•** Foursquare API: The Foursquare API has a database of more than 105 million places. This Capstone Project will be using Foursquare API as its prime data gathering source. This API will help us to perform location search, location sharing and details about a business. Foursquare users can also use photos, tips and reviews in many productive ways to add value to the results.

• URL: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> : for Postal codes, borough and names 0f the neighborhoods in Canada

• K-mean clustering algorithm will be applied to form the clusters of different categories of places residing in and around the neighborhoods. These clusters from each of those two random neighborhoods will be analysed individually collectively and comparatively to derive the conclusions.

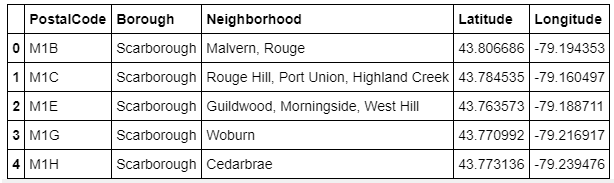
• Work Flow: HTTP requests would be made to the Foursquare API server using zip codes of the random neighborhoods to pull the location information (Latitudes and Longitudes). Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 700.

**2.2 Libraries used**

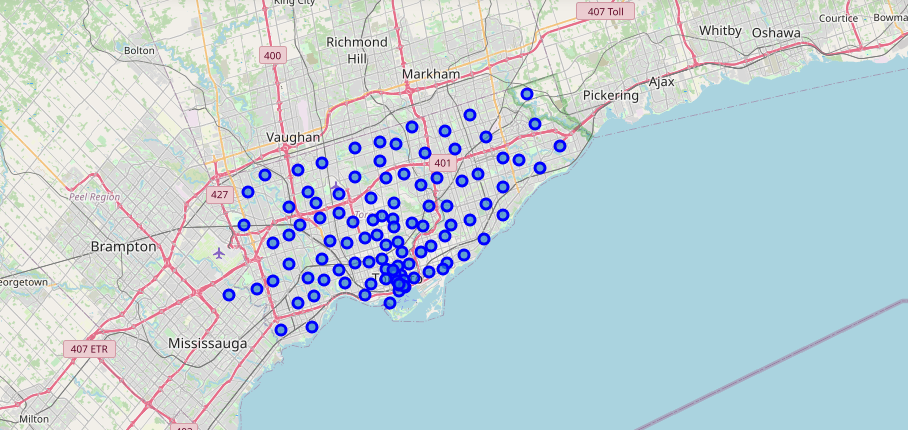
* Pandas - For data analysis and manipulation
* NumPy – To compute data in vectorized manner
* Geopy – To locate coordinates of locations
* JSON – To handle JSON files
* Requests – To handle http requests
* Sklearn – For machine learning algorithms
* Matplotlib – For interactive visualization
* Folium – For map rendering

**3. Methodology**

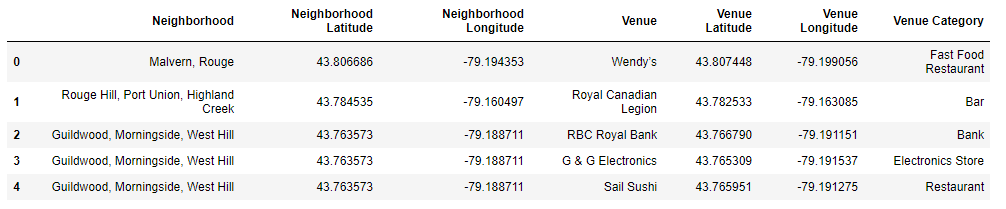
After importing the libraries successfully we proceed further. The source url is then scrapped and extracted in the form of dataframe. Our dataframe that will be used in this entire project consists of Postalcode, borough and neighborhood. The dataframe is further extended by adding latitudes and longitudes with respect to its location**.**

****

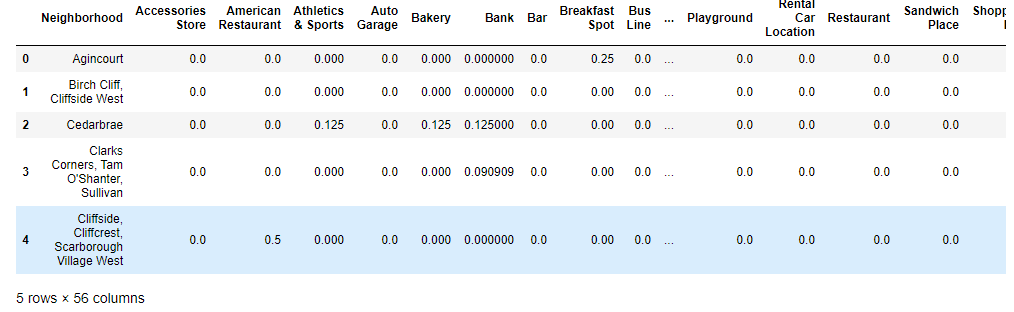
For this project I’ll be taking the Scarborough location of Canada. Presenting the neighborhood in the map:



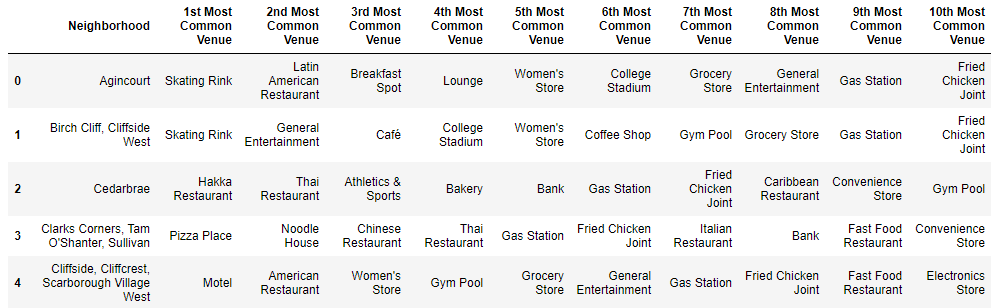
Now we have the neighborhood with its coordinates. We need to find and compare the facilities around the neighborhood. Thus we integrate with the Foursquare API to get the venues around the location. After integrating the notebook with the Foursquare API we extend our dataframe and add venues, venues’ category and its coordinates with respect to the locations.



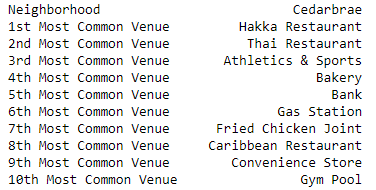
Analysing the neighborhoods on the basis of facilities around it. We calculate the mean frequency of occurrence of each venue category.



As we can see the neighborhoods have various facilities around them and it would be a bit difficult to analyse all of them. Therefore, we will be taking 10 most common venues of each neighborhood.



We can also search top 10 most common venue of any location in the neighborhood taking Cedarbrae as an example:



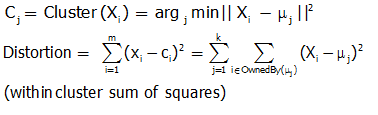
**K-means clustering algorithm**

K-means clustering is one of the simplest unsupervised learning algorithm that solve the well known clustering problem. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. So here we will be using k-means clustering algorithm as our data set is distinct and well separated in a linear fashion.

## **Mathematical Formulation for K-means Algorithm:**

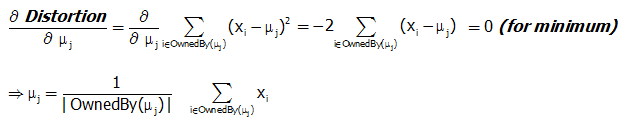
D= {**x1,x2,…,xi,…,xm**} à data set of m records

**xi**= (xi1,xi2,…,xin) à each record is an n-dimensional vector



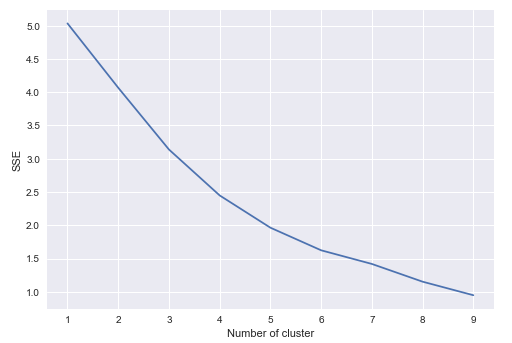
## **Finding Cluster Centres that Minimize Distortion:**

Solution can be found by setting the partial derivative of Distortion with respect to each cluster centre to zero.



For any k clusters, the value of k should be such that even if we increase the value of k from after several levels of clustering the distortion remains constant. The achieved point is called the “Elbow”.

This is the ideal value of k, for the clusters created.

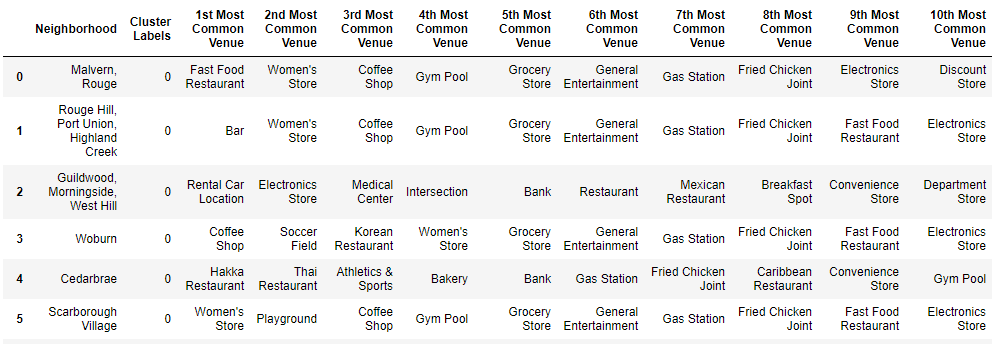


After performing the k-means clustering algorithm we got the value of k as 3. We extend our dataset and add cluster column to it. This will help us track which neighborhood belongs to which cluster.

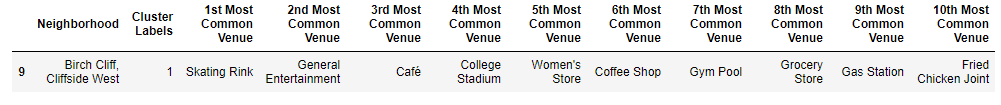


Distributing the neighborhood on the basis of the clusters:

**Cluster 0:**

****

**Cluster 1:**

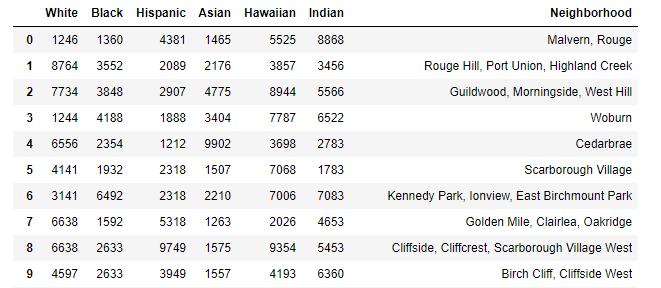
****

**Cluster 2:**

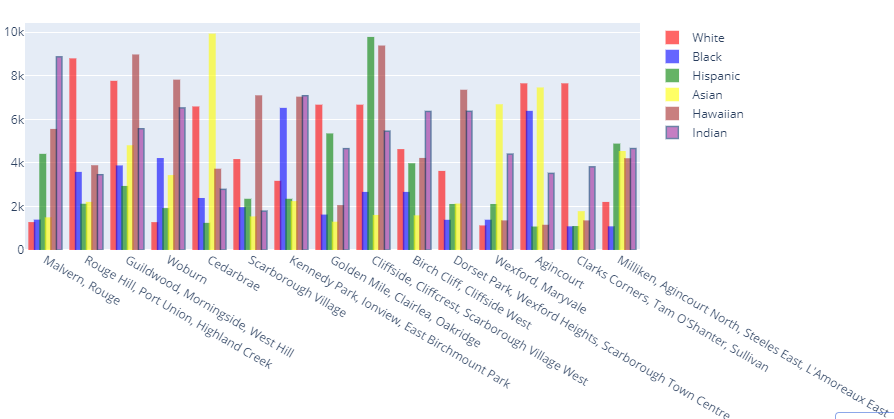


The neighborhood in the clusters have similarities. Now let us add more sources to our dataset such as:

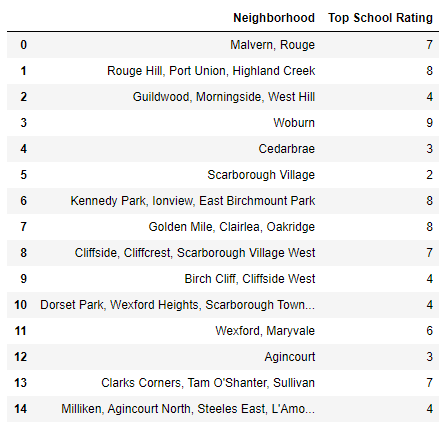
**\* Population distribution by ethnicity**

****

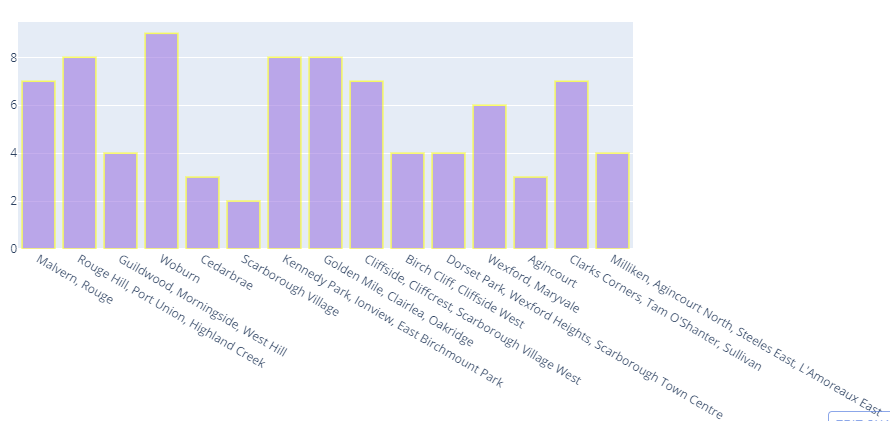
**Visualising the population distribution of the neighborhood**

****

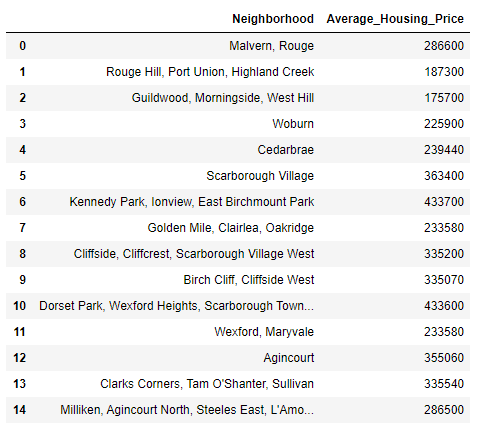
**\* Rating of schools**



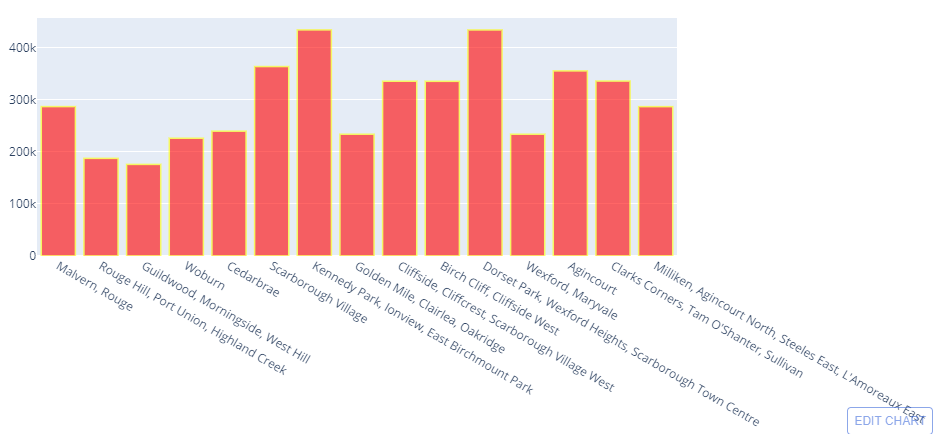
**Visualising rating of schools**

****

**\*Average housing price range**

****

**Visualising the average housing price range**

****

Let us try our project out by testing it by comparing two neighborhoods on the basis of our requirements and see which neighborhood wins.

For this report we will be taking:

Neighborhood 1: Woburn

Neighborhood 2: Cedarbrae

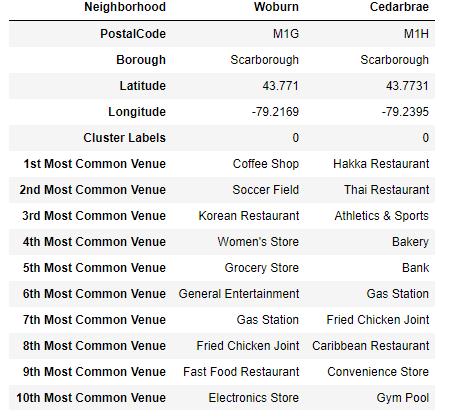
Problem: We want to shift to Canada in the neighborhood where there is:

\*More Indian population,

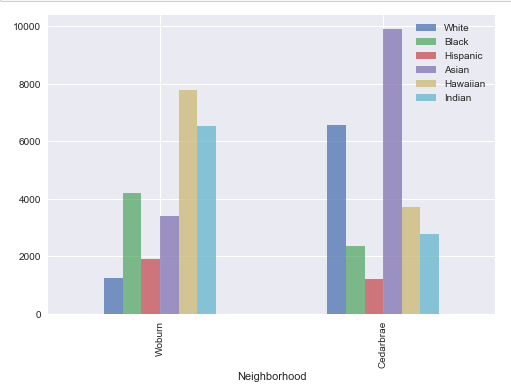
\*High school ratings,

\*Average house range is 200k-300k

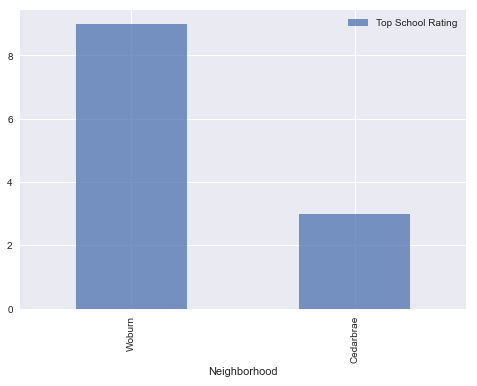
**Comparing on the basis of the top 10 most common venues:**



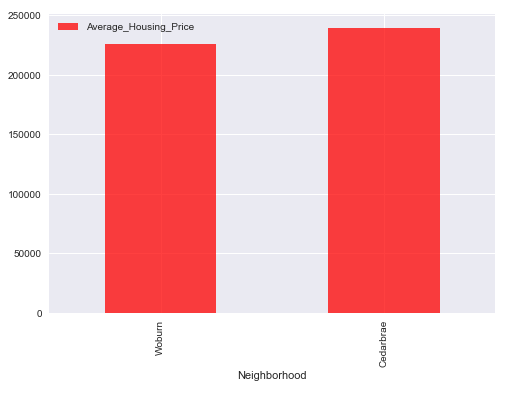
**Comparing the population distribution of both**



**Comparing on the basis of high school ratings**

****

**Comparing on the basis on average housing prices**

**Result**

On the basis of the above comparison we can conclude the following:

\*More Indian population:

Woburn: 6522

Cedarbrae: 2783

\*High school ratings:

Woburn: 9

Cedarbrae: 3

\*Average house range is 200k-300k:

Woburn: 225900

Cedarbrae: 239440

**Conclusion**

After completing the whole analysis we come to the conclusion that Woburn will be a better place to shift.