# Toronto Battle of Neighbourhoods

## Francisco Tarantuviez

### January 2, 2021

**1. Introduction**

**1.1 Background**

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America.

Toronto encompasses a geographical area formerly administered by many separate municipalities. These municipalities have each developed a distinct history and identity over the years. Former municipalities include East York, Etobicoke, Forest Hill, Mimico, North York, Parkdalse, Scarborough, Swansea, Weston and York. Throughout the city there exist hundreds of small neighbourhoods and some larger neighbourhoods voering a few square kilometers.

**1.2 Problem**

I am currently living in a small neighborhood in the borough of West Toronto, called Dufferin, Dovercourt Village. I love where I live because of all the amenities it has to offer me, such as restaurants, gyms, parks and so on.

I work as an operations manager in a successful company located in the same neighborhood that I am living in. The problem comes when they offer me a “dream” promotion which the job location of it is not the same as the actual. The job neighborhood is in Scarborough, and it’s called Cedarbrae.

This neighborhood is approximately 19.58km from Dufferin, my actual neighborhood. What it means is that I should move if I want to take this job opportunity. Therefore I want to make a data driven decision and find the optimal neighborhood between two parameters: nearness with the job location and similarity in relation to my actual neighborhood.

**1.3 Interest**

The personal interest of getting a promotion where I work and also can move to a neighborhood which is near the job location and is similar to my actual neighborhood, which I love.

**2. Data acquisition and cleaning**

**2.1 Data Sources**

Different data sources compose the final dataset. Starting from Wikipedia’s article: [List of postal codes of Canada: M](https://en.wikipedia.org/wiki/Postal_codes_in_Canada), getting the data via parsing the HTML code, and where the features name of neighborhood, postal code, borough were extracted.

From there, using the geolocator in the package *pygeo*, and the postal code of each neighborhood, we could obtain the latitude and longitude belonging to each respective neighborhood. Also were deleted the wrong values with NaN type assigned. The neighborhood distributions following the geolocation looks like the next image:

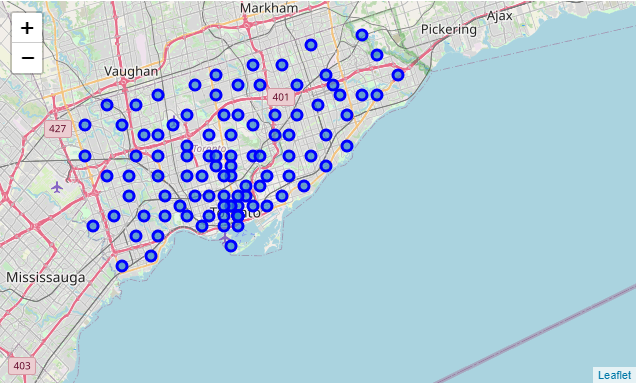


Figure 1.1: Neighbourhoods distribution in Toronto City

Applying the jaccard distance provided by *geopy*, and passing as parameters the longitude and latitude of the neighbourhoods, we got the distance between the job neighborhood and each single neighborhood in kilometers. In example, the neighborhood Parkwoods, located in North York, is 7.58 kilometers from Cedarbrae.

Already with the neighbourhood's location, the next step was to get the first 100 venues in a radius of 500 meters with respect to each neighborhood. This was achieved with the [Foursquare API](https://foursquare.com/) which provides a different useful information in relation with venues. From these data we extracted the following features: venue category, venue name, venue address, venue id (API id), category id (API id), venue latitude and venue longitude.

**2.2 Data cleaning**

All the data was merged into a single dataset. The data extracted from Wikipedia HTML table had a lot of insignificant places with values “not assigned”, so all of these values were removed.

Then, when extracting the location of the neighborhoods, some of the postal code query returns NaN values, so they also were removed after analysing the impact of these in the data.

Finally, visualizing the neighborhoods segmenting by borough, we could see the wrong value corresponding to Business reply mail Processing Centre, South Central Letter Processing Plant Toronto. As we can see in the map (look at figure 2.1), between the light green neighborhoods (Scarborough borough) there is a cercle with other colors.

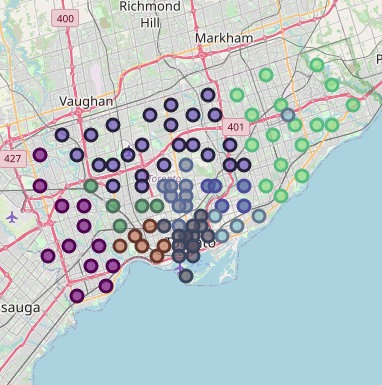


Figure 2.1: Wrong neighborhood assignation.

Business reply mail Processing Centre, South Central Letter Processing Plant is not even a neighborhood, but a mailing service, so also was removed from the dataset.

**2.3 Feature selection**

Because it was out of the scope of this research, the features venue id, category id, venue latitude, venue longitude and venue address were removed from the feature selection.