2020

IBM Data Science
Professional Certificate
Capstone Project

EZZAKRAOUI Meriem

Opening new Park in Toronto

[THE BATTLE OF NEIGHBOURHOODS]

Abstract

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of 2,731,571 in 2016. Current to 2016, the Toronto census metropolitan area (CMA), of which the majority is within the Greater Toronto Area (GTA), held a population of 5,928,040, making it Canada's most populous CMA.

Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. So a group of investors is looking to open a new Park in Toronto, and asking for recommendation about the best Location where implementing the new Park?

This project seeks to explore data insights using Data Science Methodology 101 to identify the best Districts in Toronto that may be under-served by the number of existing Parks, and what are the factors that lead to such insights like Population Density. The audience for this project is investors having a vested interest in responding to market demand by making investment and operations decisions based on data insights.

I- Introduction of Data Science Methodology 101

I-1 Data Science Methodology

Data Science Methodology is a system of methodology used in a particular area of study or activity to address a question in hands. This methodology aimed at increasing the use of Data Mining over a wide variety of business applications and industries.

I-2 Data Science Methodology 101

DS Methodology 101 begins with spending the time to seek clarification supporting the goal. Before even starting collecting data, objectives and goals need to be defined.

DS M101 is iterative and never ends. It starts from Business understanding to Analytic Approach. Since DS Methodology is like cooking, the Data Science should seek for all ingredients required, collect them, understand them and prepare them to meet the desired outcome.

II- From Problem to Approach and From Approach to analytics (Methodology section)

The group of investors is looking to choose the better place to implement a new Park in Toronto which is an international center of business, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

So our task is building recommendations of some districts in Toronto where this group can open a new Park using Clustering since we don't have any labeled data on which we can learn our machine.

We need to segment our District using Population and Number of existing Parks here which is an unsupervised Approach to create segments with groups of objects that are similar to other objects in a cluster, and dissimilar to objects in other clusters.

Doing that, we will be able to interpret Clusters and deduct the one with a Max Population and Minimum existing Parks.

III- From Requirement to Collection (Data Section)

As we should use Open Source Data to get Districts in Toronto and use **population** and the **number of existing Parks around**, are the main criteria for this first iteration, we will start by collecting this data using:

- Scrapping from Wikipedia to get Demographic Data and Focus on Population Only in this iteration:

URL: https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

- Geocoder to geo-locate Districts of Toronto to be used in Foursquare API
- Foursquare API to get all geographic information about Neighbourhoods and explore Toronto Districts to get all venues and specially the existing Parks.

IV- From Understanding to Preparation

After collecting Demographic Data from Wikipedia (Fig-1), geo-locate them using the Name of Districts (Fig-2), and get all Neighbors around these Districts in a radius of 500 (Fig-3), we got these 5-Top Categories of Neighbors (Fig-4).

	Name	Population
1	Agincourt	44577
2	Alderwood	11656
3	Alexandra Park	4355
4	Allenby	2513
5	Amesbury	17318

Fig1 - Demographic Data from Wikipedia

	Name	Population	Latitude	Longitude
1	Agincourt	44577	43.601717	-79.545232
2	Alderwood	11656	43.650787	-79.404318
3	Alexandra Park	4355	43.711351	-79.553424
4	Allenby	2513	43.706162	-79.483492
5	Amesbury	17318	43.743944	-79.430851

Fig2 - District Coordinates

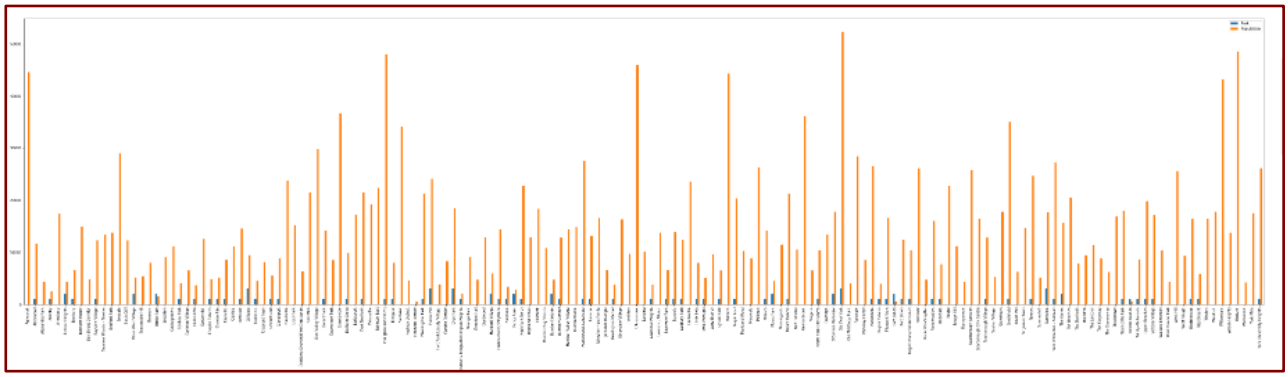
Name	Latitude	Longitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
Agincourt	43.601717	-79.545232	Il Paesano Pizzeria & Restaurant	43.601280	-79.545028	Pizza Place
Agincourt	43.601717	-79.545232	Timothy's Pub	43.600165	-79.544699	Pub
Agincourt	43.601717	-79.545232	Toronto Gymnastics International	43.599832	-79.542924	Gym
Agincourt	43.601717	-79.545232	Tim Hortons	43.602396	-79.545048	Coffee Shop
Agincourt	43.601717	-79.545232	Pizza Pizza	43.605340	-79.547252	Pizza Place

Fig3 - District Neighbors

VenueCategory	Count
Park	90
Café	81
Coffee Shop	57
Bakery	56
Italian Restaurant	56

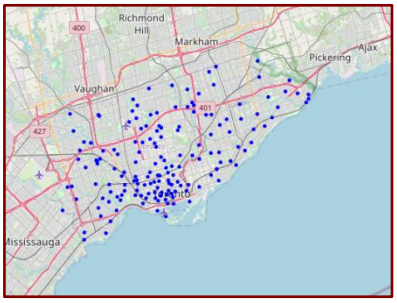
Fig4 - 5-Top Categories

We only keep **Park** Category as Neighbors to answer what demanded by Investors, and then we got cleaned DataSet of 163 District ready for making some Features like: *Number of existing Parks*.



The orange color represents Population in a District while Blue Color represent Number of existing Parks around this District in a radius of 500 (this number range from 0 Park to 3 Parks around).

We can get here an idea about some potential in some Districts with an important Population density, but no Park around.



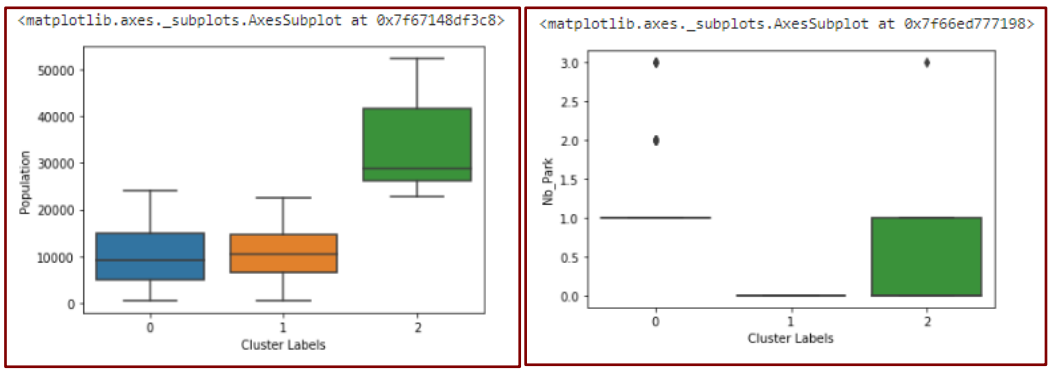
We can also see the repartition of these Districts in this Map using Folium:

V- From Modeling to Evaluation (Results section)

After getting our data ready, we can start our modeling.

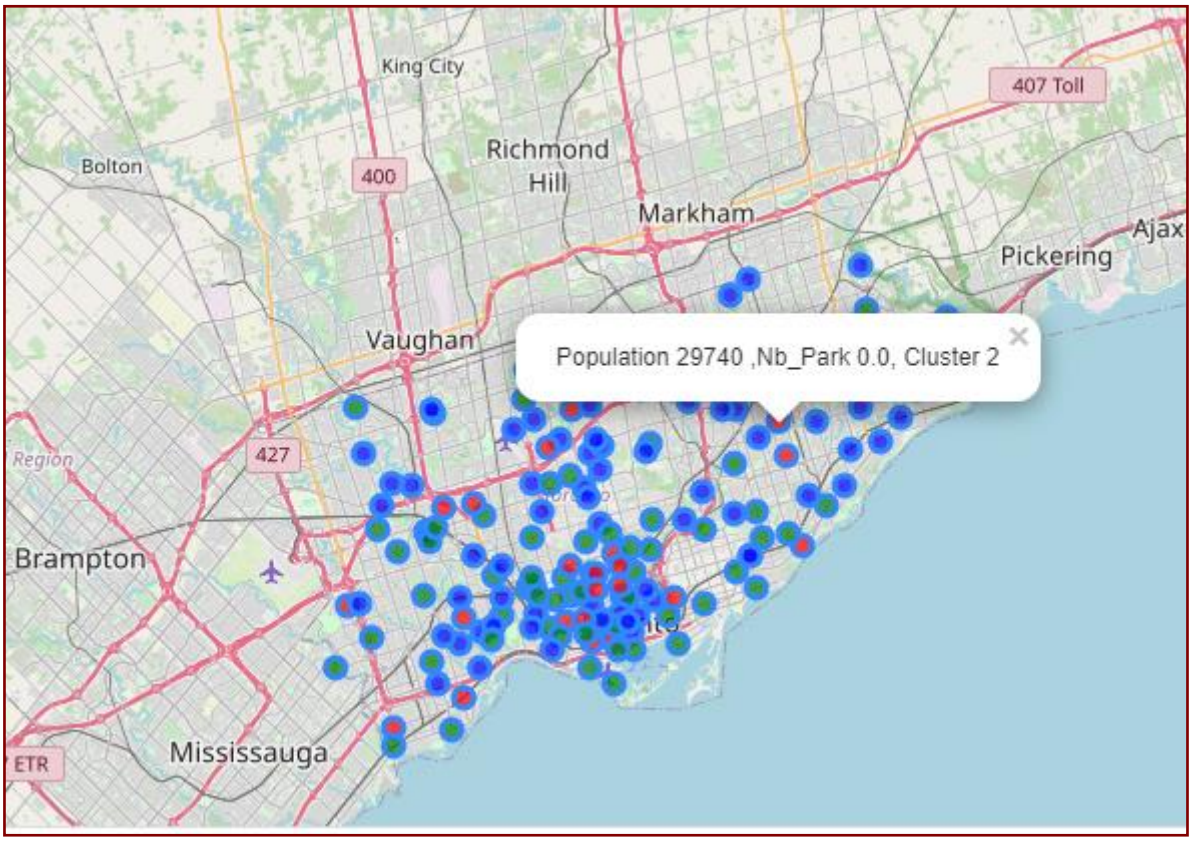
In this case we need to segment our District using K-Means Clustering with 2 Inputs: Population and Number of existing Parks.

After running the Model we get 3 Clusters in this iteration:



Interpretation: The Cluster-2 Have maximum values of Population and at most 1 Park.

We can also visual our Clusters in the Map using Folium:



Conclusion/Discussion section

Using some criteria such as *Population* and *Number of existing Parks Neighbors* to choose the right District where we can implement Our Park in this iteration we got these first results:

- Districts where the best Place should be implemented based on the Population are those in Cluster-2 with Maximum Population and Minimum of Parks.
- Districts where there is already a high competition, to be eliminated in a next iteration with more criteria.

By adding more criteria and more data we could filter these districts in Cluster-2 again in order to get the best District to implement the new Park.