Predicting Places to open Café in Toronto

Youssif khairy

October 31, 2019

1. Introduction

1.1 Background

Toronto City is one of the most famous cities in the world and one of the most visited cities in the world. Opening new café is one of the beneficial things any investor or stakeholder will be interested in. Cafés in some areas win a lot of money and others don't. So, where to open new café Is our target area. Therefore, it is advantageous for to predict where to open their new café.

1.2 Problem

Data that might contribute to determine where to open new café might contain Neighbourhood, Borough, Postcode, longitude and latitude of the Neighbourhood. This Project aim to predict where to open new café in Toronto, Ontario.

1.3 Interest

Obviously, Stakeholders that are interested to open new café in Toronto, for competitive advantage and business values. Others who benefit from this investment as people of Toronto and tourists.

2.Business Problem

The objective of this project is to find the best location for stakeholders to open new coffee shop in Toronto, Canada. Using data science methods and machine learning methods as clustering, this project aims to answer business question: For stakeholders in Toronto, Canada where to open new coffee shop?

3.Target Audience

Stakeholders who want to open new cafe` in best locations in Toronto, Canada.

4.Data Acquisition and cleaning

4.1 Data Sources

Most of data of Toronto As Neighbourhoods ,boroughs and postcode Can be found here, Also Longitude and latitude of each Neighbourhood can be found here.

4.2 Data Cleaning

Data is scraped from Wikipedia as mentioned above. There was a lot of boroughs that doesn't contain values, so it was wise to drop them. There was problem with dataset that there was a duplication of neighbourhoods, so grouping same neighbourhoods with to same postcode value.

5. Methodology

First, I need to get all the neighbourhoods in Toronto so I scraped the Wikipedia page here. I scraped this page using BeautifulSoup library in python. This only contains Postcodes and neighbourhoods only, but I also need the coordinates of each neighbourhood. So, I tried to use geocoder but there was a problem with it due to the charge, so I used CSV file here to get each coordinates(latitude and longitude) of each neighbourhood. After that I used folium package to draw map to see Toronto and the positions of its neighbourhoods to be sure that we are walking on the right way. It is now the time for Foursquare API. I used it to get 100 venues in the radius of 500. And extracted the name and category and location of each venue. Then I calculated the mean value for every venue for each neighbourhood. Only what I need is the venue is café. So, I extracted the mean values only for it. Lastly, I used clustering method by using K-Means clustering it is one of the popular unsupervised machine learning algorithms and it highly fits this project as well. I used 3 clusters for this algorithm to obtain the locations with high and medium and low number of cafés.

6. Results
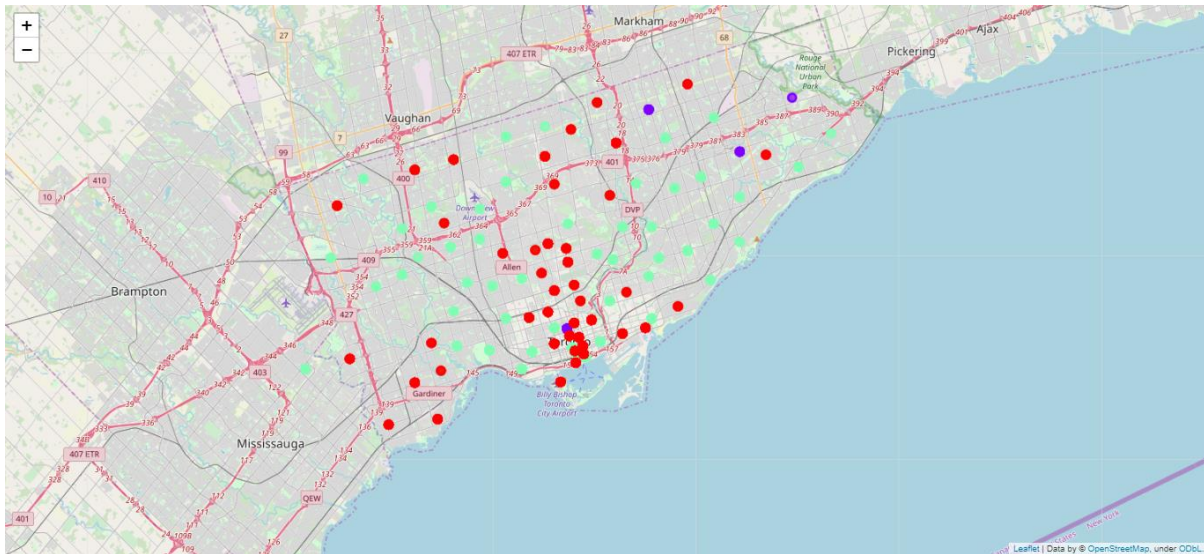
The results of K-Means clustering algorithm is 3 clusters:

Cluster 0: with too much cafés in this area (Red coloured)

Cluster 1: with fewer number of cafés in this area (Blue coloured)

Cluster 2: with no number of cafés in this area (Cyan coloured)

7. Recommendation

It seems that if you get nearer to the centre of the city you will get much cafe's but the areas far from the centre are much fewer ones. SO, it's smart to open in places as Victoria Village or Etobicoke marked as cluster 2 as it will be less competition with other cafés.

8. Limitations

I only take in consideration the occurrence and the existence of cafés. There can be other factors to take in consideration such as population density, income of residents and the rent that could influence the decision of opening new café. However, it's impossible to implement all these features in this short time.

9. Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.

10. References

List of neighbourhoods for Toronto is:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

CSV file for extracting location for each neighbourhood:

https://cocl.us/Geospatial_data

Foursquare API to extract information about neighbourhoods:

http://foursquare.com/