Coursera Capstone

IBM Applied Data Science Capstone

Opening a New pub in Toronto, Canada.

By: Aniket Mitra

May 2020

Introduction

This is a capstone project for IBM Data Science Professional Certificate.

In this project I am creating a hypothetical scenario that there are not many enough pubs in Toronto. Therefore it might be a great opportunity for the entrepreneurs.

As liquor is very popular among western and cold countries, it might be very good opportunity for the investors to invest upon a lucrative business.

Business Problem

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new pub in Toronto, Canada. By using data science and tools along with machine learning algorithms, this project aims to provide a solution to the business problem.

Target audience of this project

This project is particularly useful to the investors, Liquor companies and businessmen.

Data

To solve the problem, we will need the following data:

• List of neighbourhoods in Toronto. This defines the scope of this project which is confined to the city of Toronto, the provincial capital city of Ontario.

• Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to pubs and restaurants. We will use this data to perform clustering on the neighbourhoods.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of TORONTO. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Toronto. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude.

Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Pub” data, we will filter the “Pub” as venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Pub”. The results will allow us to identify which neighbourhoods have higher concentration of Pubs while which neighbourhoods have fewer number of Pubs. Based on the occurrence of pubs in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Pub.

Results

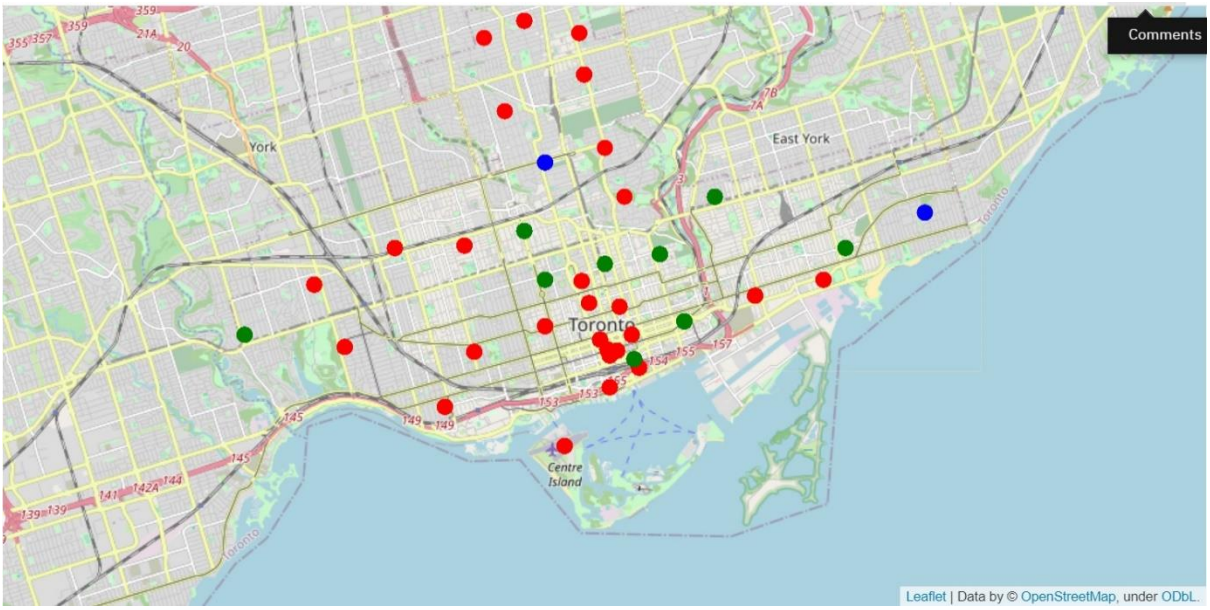
The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “pubs”:

• Cluster 0: Neighbourhoods with moderate number of pubs.

• Cluster 1: Neighbourhoods with low number to no existence of Pubs

• Cluster 2: Neighbourhoods with high concentration of pubs.

The results of the clustering are visualized in the map below.



Result/Recommendations

Most of the pubs are around cluster 2 that is in Stn A PO Boxes, The Annex, North Midtown, Yorkville, University of Toronto, Harbord, etc,. Cluster 1 has lesser number of pubs of pubs around whereas cluster 0 has very less or no pubs. Therefore it is suggested to invest on the following areas: First Canadian Place, Underground city, Berczy Park, Toronto Dominion Centre, Design Exchange.

Limitations and Suggestions for Future Research In this project, we only consider one factor i.e. frequency of occurrence of pubs, there are other factors such as population, choice, culture and income of residents that could influence the location decision of a new pubs. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new pubs.