Best Neighbourhood for Restaurant

Introduction

1.1 Background

Toronto is the capital city of Canada and is the most populous city with more than 2.7 million people in 2016. Toronto is divided into 6 major districts and more than 100 neighbours with their own characteristics.

1.2 Problem

Data around Toronto neighbourhood is widely available from multiple sources. The aim of this study is to find the best neighbourhood to open a restaurant in Toronto with consideration of the geolocation data and demographics data of each neighbourhood.

1.3 Target Audience

- New immigrants to Toronto
- People who plan to move to Toronto

Data

2.1 Data Sources

This study is based on publicly available data in Internet with three key data sources as below.

The first one is neighbourhood list in Toronto scrapped from Wikipedia using beautiful soup package. Data is available in table format. The following analysis is using this list as the basis for further information retrieval and analysis.

The second one is geolocational data from 'foursquare.com' using API calls. Data are collected for each neighbourhood based on the list obtained as above. The aim of this search is to gather nearby places with a radius of 500. This data supports the analysis by identifying the most popular places in each neighbourhood.

The last one is neighbourhood demographic data is downloaded from 'open.toronto.ca'. The Census of Population is held across Canada every 5 years and collects data about age and sex, families and households, language, immigration and internal migration, ethnocultural diversity, Aboriginal peoples, housing, education, income, and labour. City of Toronto Neighbourhood Profiles use this Census data to provide a portrait of the demographic, social and economic characteristics of the people and households in each City of Toronto neighbourhood. The profiles present selected highlights from the data, but these accompanying data files provide the full data set assembled for each

neighbourhood. In this study, we use population density information to complement the analysis on popular places.

2.2 Data Features

To open a restaurant in Toronto, neighbourhood geolocational data is valuable to be explored. This includes the nearby places around an area and people stays within the target area. 2nd data source listed above helps to identify the nearby places and 3rd data sources add further details to complement the analysis.

Methodology

3.1 Data Cleaning

(a) Download data from Wikipedia as first data source

Data shall be scrapped into a table format to facilitates data analysis in the following stages. This part is handled using beautiful soup package. Data is then transformed into pandas data frame.

(b) Request for geolocational data as 2nd data source

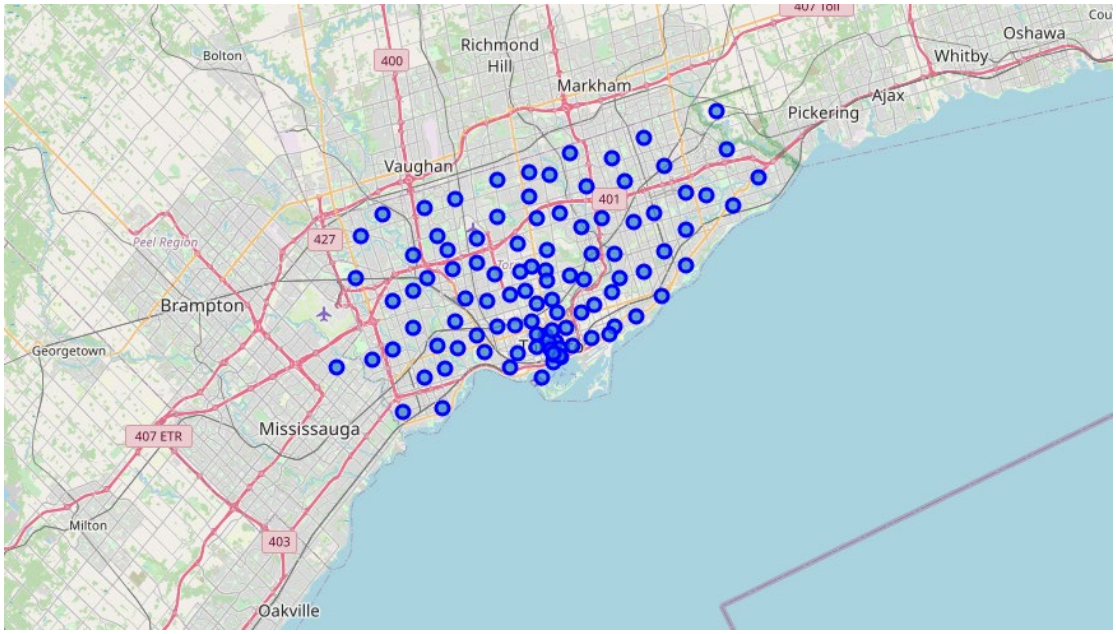
With help of four-square package, we can download nearby places of identified neighbourhood coordinates. Results are stored in a pandas data frame for further data processing. One-hot encoding is used to transform the geolocational data into binary format for each category.

(c) Download demographic data from Census of Population

Data are downloaded from the government website. These are additional data points to provide insights for population density for each neighbourhood. Given there are different name variations of data, names have to be cleaned and combined to ensure linkage across pandas data frames.

3.2 Data Visualization of potential candidates

With a pandas data frame of neighbourhood with coordinates information, we shall be able to visualise the data using folium package.



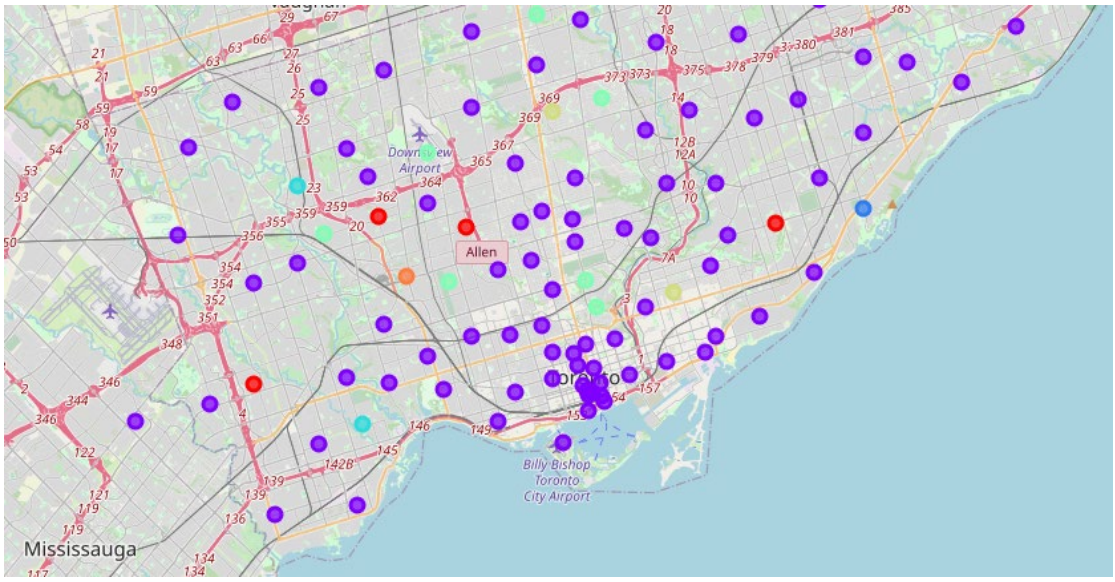
The neighbourhood coordinates are evenly spread within Toronto. This servers as the basis of this study. The next step is to filter away noise so that we can locate the best neighbourhood to open a restaurant.

3.3 Clustering

As we are dealing with unlabelled data, we can use unsupervised machine learning algorithm to build clusters among these neighbourhoods.

With the one hot encoded data frame prepared in earlier steps, we able to use K-means to cluster data points with consideration of nearby places.

The clustered neighbourhood can be visualized as below map.



3.4 Neighbourhood Selection

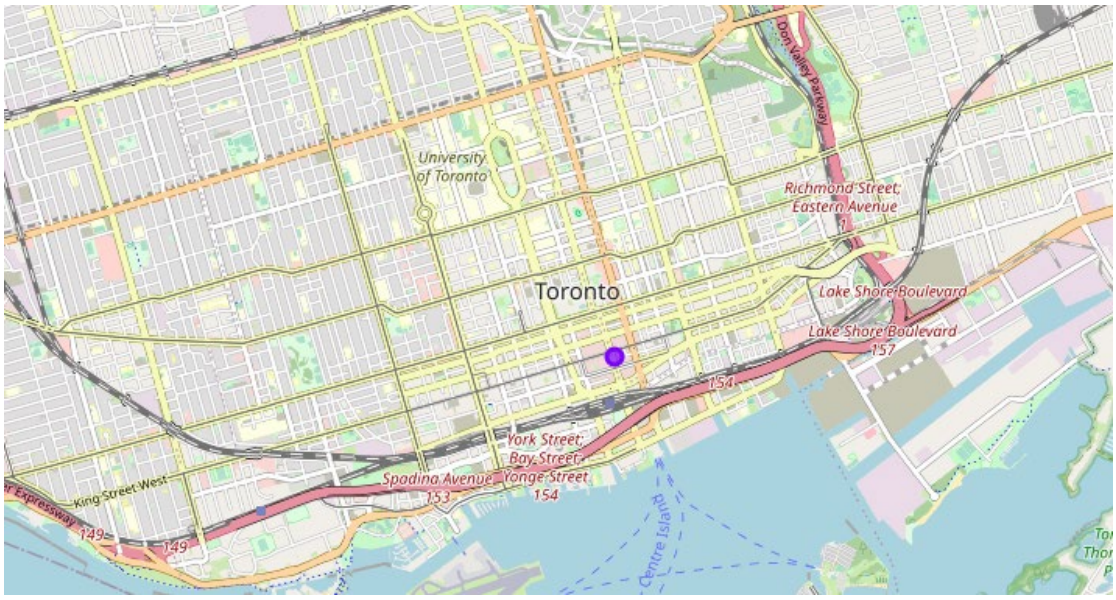
With clustering information as below, we will be able to tell the places with 'restaurants' in the top 3 popular venues. As we have 3rd data source with populate density, we are able to filter further and select the neighbourhood that linked to 'restaurant' keyword with highest population density.

Results

Below is the result of the analysis. The best place selected is Commerce Court with population density of 3700.

	Neighborhood	density_int
48	Commerce Court, Victoria Hotel	3710

It can also be visualised as below –



Discussion, observations, and future directions

The place turned out to be the CBD area of Toronto with higher population density which is not uncommon. This analysis can be further enhanced using more demographic data points. This will also give better clustering power of K-means to produce a more accurate result.

Conclusion

In this study, I gathered publicly available information from website and construct a mechanism to cluster neighbourhood information based on popular nearby places. The result set are further filtered based on demographic information of neighbourhood. The final recommendation is to open a restaurant in CBD area of Toronto, specifically area around Commerce Court, Victoria Hotel.