Business in the Neighbourhood

Karishma

October 9, 2019

1. Introduction

1.1 Background

There are more than 28 million small businesses in the United States, making up a whopping 99.7 percent of all U.S. businesses, according to the Small Business Administration. When you consider some of the most popular reasons to start a business, including having a unique idea designing a career that has the flexibility to grow with you, working toward financial independence, and investing in yourself — it's no wonder that small businesses are everywhere. But it takes a lot more than just capital money and experience to start a small business. Location plays a very important role in describing the business success. For a person not familiar with the city it takes a lot of research to select an appropriate location for business. Therefore, through this project the tedious task for selecting a location made easy.

1.2 Problem

The problem in establishing a successful business is to select a proper location for start-up. For example, the person may select a location entailing high competition or a location which might not have a single venue of the same business perspective. i.e. the idea of doing business in the location might not be feasible. Considering such challenges there is a requirement for a model which can provide with the safest options for business location.

1.3 Interest

The problem is very common and therefore, shares interests with a number of stake holders. The model would interest people looking to open a small business, a start-up in a city and need safest options for the target location.

2. Data

The data used in the project is of different format and origins. The types of data worked with in the project are as follows:

1. Web scraped data from the Wikipedia website. The data encompasses mainly the table of Postal code, Borough and Neighbourhoods associated with each. In the project as an example

the data of city of Toronto is web scraped from the Wikipedia site. The data is in tabular form but not normalized.

2. The user input given . The first input is the selective neighbourhood in the city. The second input is the type of business the user is willing to start. This is mainly the search query which will be used in the search process of venues using the FourSquare API which provides accurate location data.

3. Along with the locations in the city the coordinates of those locations are obtained via the geopy library.

4. The list of venues , their category , their latitude and longitude with respect to search query are provided via the FourSquare API.

3. Exploratory Data Analysis

In this part the proper analysis of the type and origin of data is performed. The relationship between different data is explored. The relationship between the attributes neighbourhood and venues is examined via the frequency of the number of venues present at a particular neighbourhood of the same post code. The same type of neighbourhoods are grouped together in the data set. The neighbourhoods containing high number of search query venues are least likely to be included in the result.

4. Analytic Approach

It is important because it helps identify what type of patterns will be needed to address the question most effectively. In this project, the problem is to search for appropriate locations in the neighbourhood to start a business. As the problem deals with exploring relationships between different factors therefore, a descriptive approach where clusters of similar data based on the number of venues in the neighbourhood and preferences are examined, would be the right analytic approach.

5. Data Understanding

In this part the proper understanding of the type and origin of data is performed. The data from different resources and different types is cleaned and transformed and then combined into dataframe. For example, the web scraping is done via the Wikipedia library in python. The data is then transformed and converted into a Pandas Data frame.

The location coordinates retrieved for each postal code are then added to the original data frame. The final data frame would look like this.

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

6. Data Preparation

The FourSquare API is used in the retrieval of venues located in every neighbourhood. The search query is passed in the URL providing the result. The result which is a json file is converted to dataframe and then the resultant dataframe is the combination of the neighbourhood dataframe and the venues and venue categories associated to each.

The dataframe might contain missing values which can be dropped . The resulting dataframe would be like this :

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Hillcrest Village	43.803762	-79.363452	New York Fries	43.803664	-79.363905	Fast Food Restaurant
1	Hillcrest Village	43.803762	-79.363452	Villa Madina	43.801685	-79.363938	Mediterranean Restaurant
2	Fairview, Henry Farm, Oriole	43.778517	-79.346556	Hero Certified Burgers	43.777295	-79.344584	Burger Joint
3	Fairview, Henry Farm, Oriole	43.778517	-79.346556	Michel's Baguette	43.777082	-79.344557	Bakery
4	Fairview, Henry Farm, Oriole	43.778517	-79.346556	New York Fries	43.778298	-79.343267	Fast Food Restaurant

The dataset contains the venues that belong to a same category and in the same neighbourhood. Therefore, group the data of the same neighbourhood and frequency count the number of such rows. The resulting data frame would be like this:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bathurst Manor, Downsview North, Wilson Heights	9	9	9	9	9	9
1	Bayview Village	3	3	3	3	3	3
2	Bedford Park, Lawrence Manor East	21	21	21	21	21	21
3	CFB Toronto, Downsview East	1	1	1	1	1	1
4	Don Mills North	4	4	4	4	4	4

7. Descriptive Modelling

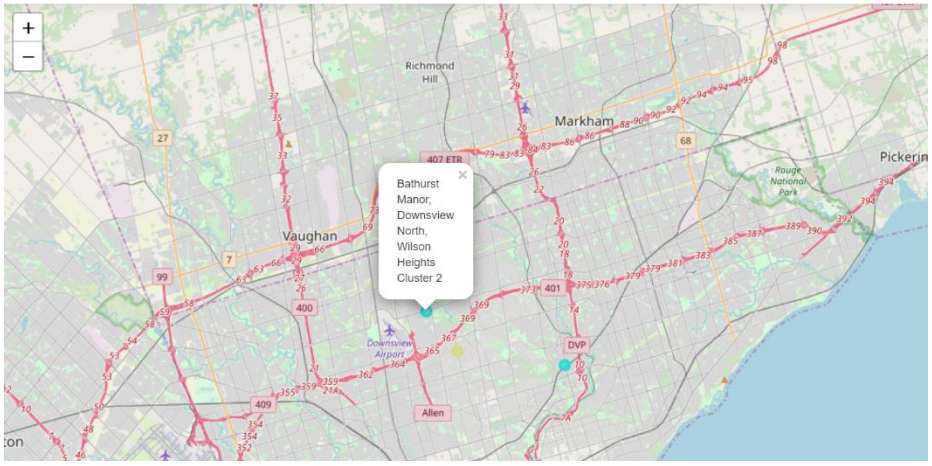
As discussed in the analytic approach section , the model used in the project would be descriptive model. The algorithm used in the procedure is the K-means clustering algorithm.

The number of clusters are 4 considering the suitable number of venues in the entire borough. The clusters are then plotted along the original map of the borough. In the evaluation phase, the clusters having the maximum number of venues and those having least number of venues are eliminated . This is because the objective of this project is to ensure the safest options regarding the location of business. And the neighborhoods containing large number of such venues would result in high competition and the neighborhoods having least number of venues considering (0,1,2,3) would be not feasible for opening the business due to various different reasons. Therefore, the most safe option is to select the neighborhoods containing an average number of such venues. And the result of successful business depends on so many different factors including domain expertise. Therefore, this is the best option.

7. Result

The final result would be a map of borough built using library folium which on which the clusters are superimposed. Apart from this the result would contain the dataframe with appropriate locations along the borough and their respective location coordinates.

For example :



Along with the mapped data the data in the tabular format is also output of the model.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Cluster Labels	Venue Latitude	Venue Longitude
0	Bathurst Manor, Downsview North, Wilson Heights	43.754328	-79.442259	2	43.755316	-79.440895
2	Bedford Park, Lawrence Manor East	43.733282	-79.419750	3	43.733725	-79.419436
9	Flemingdon Park, Don Mills South	43.725900	-79.340923	2	43.726201	-79.340690

8. Conclusion

The project is based on the problem which a business investor would face when searching for a proper location to start of their business. The model build is a descriptive model defining relationships between data i.e. neighbourhoods containing different number of venues. Where the same type of neighbourhoods form one cluster . The objective is to provide the safest possible options to the user for the location. The modelling is done via the K-means

clustering algorithm of ML. The user finally gets the valuable output in the form of map pointing to different clusters of neighbourhoods along with the tabular data of the list of neighbourhoods safe for establishing business.