KICK START OWN BUSINESS @ TORONTO

IBM CAPSTONE PROJECT

INTRODUCTION

As a part of Coursera IBM Data Science professional certificate capstone project we are asked to cluster the neighbourhood of city Toronto. This project ‘Kick-start own business at Toronto ‘ is focused towards finding a best solution to customer/investor who wants to start a new business in city Toronto.

The customer wants to know the locations where he can probably start a new business and what business he should be taking so that it will get him maximum business plus profit.The Customer also wants five options of business for him to choose from for a particular neighbour city in Toronto where he wishes to start a business.

To address the above problem we use Data science technique which can provide more explanatory solution to the customer/investor . In this project using different algorithms available in Data Science and with support to various API and libraries we will provide more accurate and data driven solution to the problem.

DATA (ACQUIRING DATA)

To solve the problem we use mainly three types of data sets.

1. Data set that consist of neighbouring cities of Toronto with their postal codes.
2. Population and Dwellings data with the in the neighbouring cities of Toronto with their postal codes.
3. Geo spacial data that contains the longitude and latitude vales of all the neighbouring cities of Toronto with their postal codes.

Using the Foursquare API we will collect all the venue details in the neighbouring cities of Toronto with their venue category. By combining the venue data with the population data and dwelling data we will be able to recommend a model to solve the problem and identify the best list of business and the place to start these business can be obtained.

METHOLOGY

Data Cleaning (Preparing data for Analysis)

The data which we was collected from various source won't be not ready for processing directly. Before doing any operation with the data we need to clean the data initially.

The data cleaning consist of selecting only those data that make an impact on our result and discarding all the other data from the collected data set.
So here in our case data that we collected from wikipedia page has many not assigned values in it which we will remove since they won't make any impact on our result.It also contains borough names where Toronto is listed as second name . We change all the borough names into Toronto which contains Toronto word in it.Since we are only interested in Toronto data we remove all the rows except the one with Borough as Toronto.

The data containing population information about Toronto is also cleaned by removing all unwanted rows and merged with the main Toronto data along with geo spatial data that contains longitude and latitude values of all neighbours in Toronto.

We use Foursquare API for getting all the venues within the specified radius of Toronto. We pass the Longitude and latitude values of all the neighbours in Toronto to the API which returns us the venues located in that neighbouring place within specified radius. Along with the venue names it also contains the venue category.

Data Analysis

Since our business problem is to find the most suitable business that an investor can choose to start in Toronto. For addressing this problem we will fetch the Least common venues from the list of venues that we have obtained from Foursqaure API.

Least Common Venues are the ones that an investor can look into and start his new business. The least common venue will have more chances of generating business since people in that area have not been to such services earlier or they may have travelled few distance to get that services.

Since this business also depends on the population of that particular area where the investor wishes to start business we have included the population data to our data set which makes our solution more convincing.

Clustering

Here we use k-means clustering . For doing this I have chosen value of k=8. This random number 8 I have selected after multiple checking using different values of k from 1 to 10. With k =8 I could obtain more convincing solution than any other.

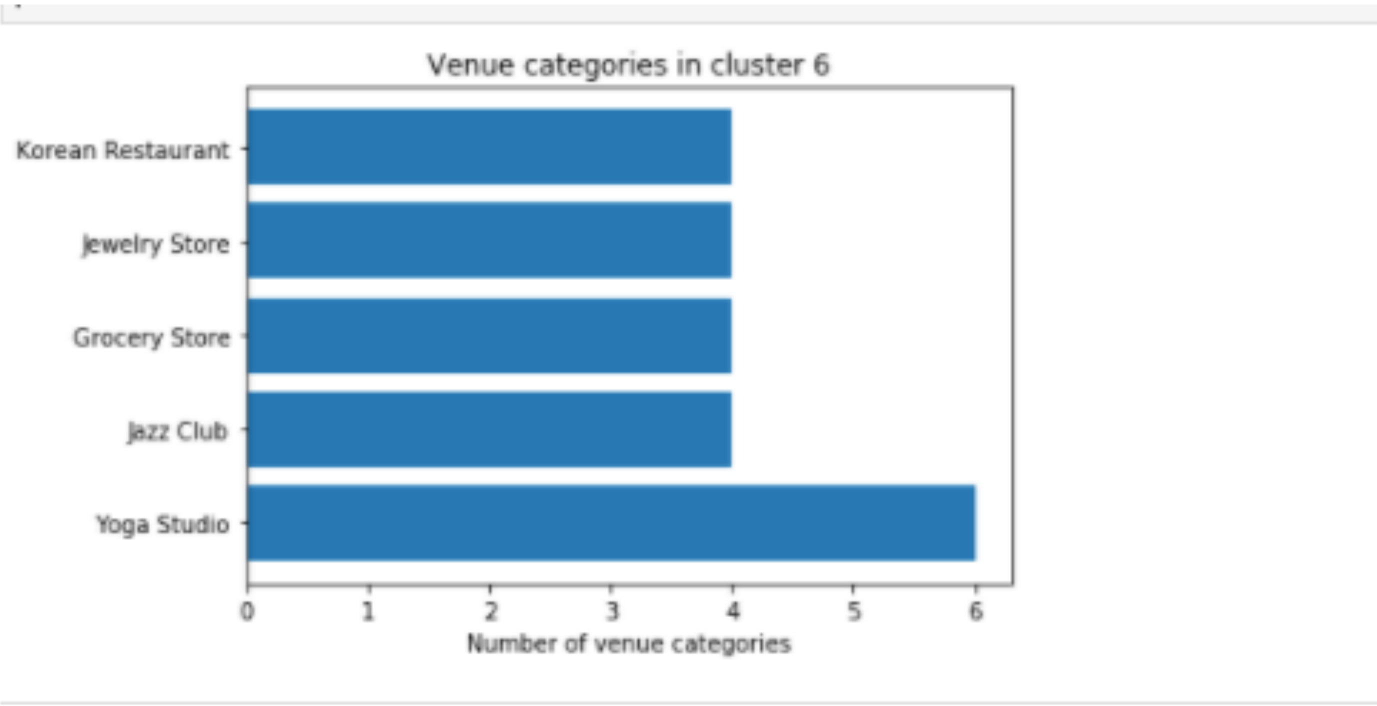
Once the clustering od data is done the same is represented using map and each row is given a cluster label based on the cluster they fall in.

RESULT

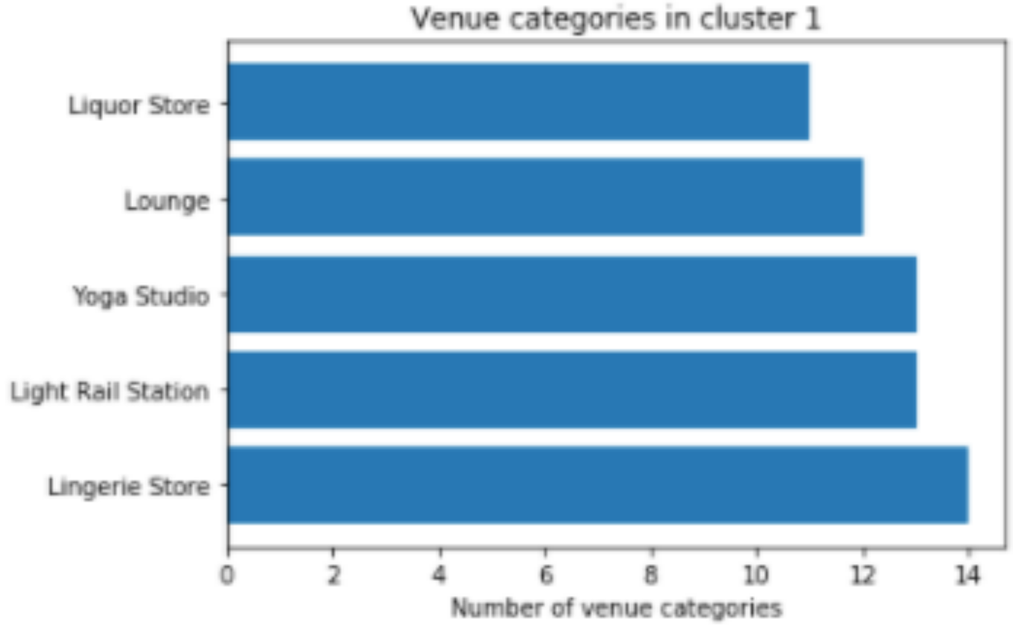
Our aim was to give the investor list of best business he can start in Toronto. For that purpose once the clustering is completed All different types of venue categories in the final data set is grouped according to their cluster label and their sum is calculated.

The one venue category that has more no of occurrence in the data set will be the least common venue appearing in almost every neighbours of that particular cluster. So for an investor this will be the best option of business we can advice him to start.

For representing the same in graphical way I used Bar graph which is shown below.The picture shows that in cluster 6 the investor can prefer Yoga Studio as his first preference and leaving that option he has got Jazz Club or Grocery store or Jewellery store or Korean Resto as his later options.



Same way we can give investor a different set of venue category in cluster 1 to choose which is shown in the bar graph below.



Here the investor can choose Lingerie store as his first preference. The Light rail station is not a business type so we can discard that and suggest Yoga studio or lounge or liquor stores as his other options he can choose from ho have maximum business.

DISCUSSION

In this project we have considered only venue and population data as the main factors that decides the business in an area. We could add more data such as income of people who stay in that area and their expenditures data and type of people stay in the area are the few more data we could add to make this model a better version.

Here in the venue categories it list all the venues which also include non business entities. So eliminating these non business entities from the venues will help in generating a more clear result.

The random value of K that we choose for clustering also make an impact since they are random values. We need to try the clustering using different values of k which I have done in this case. So choosing different algorithm for clustering will give us much better and vulnerable model.

CONCLUSION

This project has shown how to recommend an investor who is looking to start a business in Toronto. So by choosing the least common venues that is currently available in neighbouring of Toronto will generate him maximum business. The model which I have developed has got lot of scope for evolution by adding more informations that can affect the business in an area. In the current data that we have chosen I recommend Yoga studio as the best business that an investor can start in Toronto since it is least common venue in cluster 6 and second least common venue in cluster 1.