Location Identifier – First outlet in a new Geography

- ## Introduction

  Technology and internet has reduced the world boundaries for companies. Today companies aim to deliver their products or services globally. No matter if the company is formed 100 years ago or couple of months back. Things are changing and moving very fast in this dynamic world.

  But a wrong start in a country could drastically affect the future in that company. Also while entering a new country or location is challenging as we are not aware of the dynamics of the location. Company has to evaluate many factors for finalizing each step. One important step is finding the location for first outlet in a new geography. A good product but placed in wrong location will not produce the sales as desired and we might waste energy in redeveloping the product. So identifying ideal locations for starting a new outlet in a new geography is very crucial.

- ## Problem

  Data can be used to identify an ideal location for starting a new outlet in a new geography. It should reach the maximum population looking out for the products in the same category.

  For example: A multinational company is looking to enter in Canada through Toronto. Toronto is a large city of about 630 sq km consisting of large number of neighbourhoods. So finding the suitable neighbourhoods for opening a sport equipment and accessories shop is very crucial. Since company does not have any resources in Toronto, it is manually not possible to explore every neighbourhood and determine the same.

  This project aims to find neighbourhoods which are suitable to start a new sports equipment and accessories shop

- ## Data Acquisition and Cleaning
  - ### Data Sources
    - Data about the location of existing stores is to be found out for the project. Same is obtained from the Foursquare API using the search query
    - Details of the venues nearby each of the identified sport shops are collected using the Foursquare API
    - The details of the neighbourhoods (list and names) are obtained from the url -
      'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'

- The location details of each of the neighbourhoods were collected from csv file provided in the Coursera course.
- List of all the venues in each of the neighbourhoods was acquired from the Foursquare API

  o **Data Cleaning**

  The data retrieved from the Foursquare API is in the .json form. Thus it is very crucial to understand the part of the results and transform it into a dataframe.

  It is also essential to understand the results received from the API. E.g. searching sport will also result in sports bar which is not a result we are looking for. Such categories are to be identified and dropped from the data.

- ## Methodology

  1. Data analysis

  All the sports shops in Toronto city along with their location and category are retrieved from the Foursquare API. The categories other than sport shop are deleted e.g. Record of sports bar or restaurants. These locations of existing sport equipments and accessories shops is plotted on map to check if they are clustered in a location or are spread across the city.
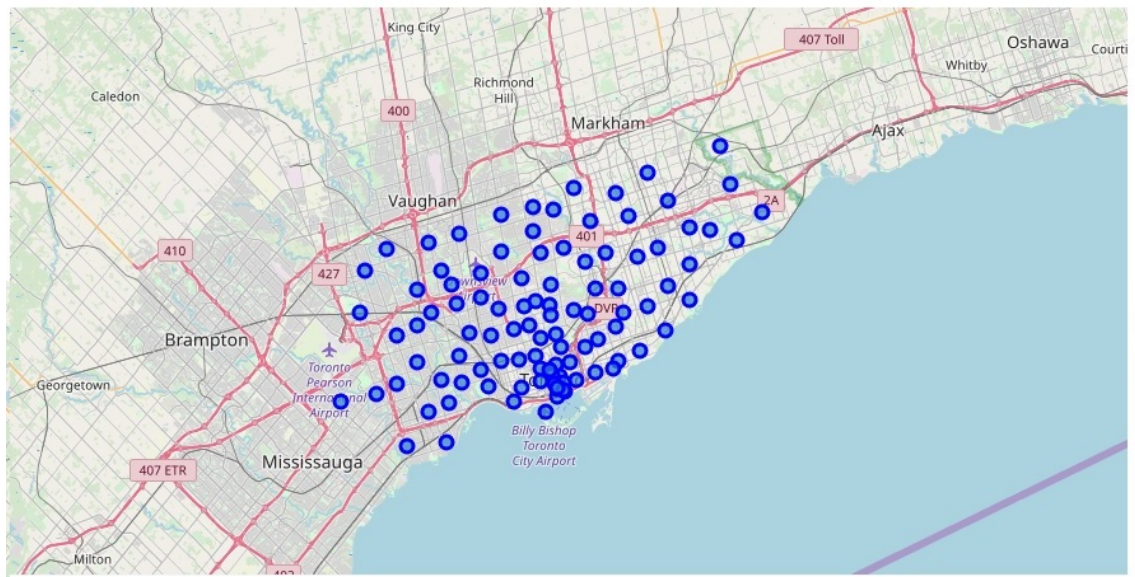
  Map1 – Existing sport shops in Toronto city

  Then details of the venues nearby each of these sport shops are obtained from Foursquare API. Categories of these venues are also captured. It was evident that the venues nearby each of the existing sport shops were similar – Restaurant, shopping area etc.

  Thus venues in a neighbourhood can help determine the suitable locations for a sports shop. SO it was decided to segment the neighbourhoods of Toronto based on the venues in that neighbourhood

2. Modelling

Since the neighbourhoods were to be clustered and segmented, K Means clustering algorithm was decided to be used. For the same venues and their category in each of the neighbourhood in Toronto was obtained using Foursquare API. Each of the categories was then converted into categorical data. For each neighbourhood mean of the venue category was calculated and added in the dataframe. This dataframe was then used for K Means Clustering.

- **Results**

K Means segmented the neighbourhoods into 5 clusters (0-4)
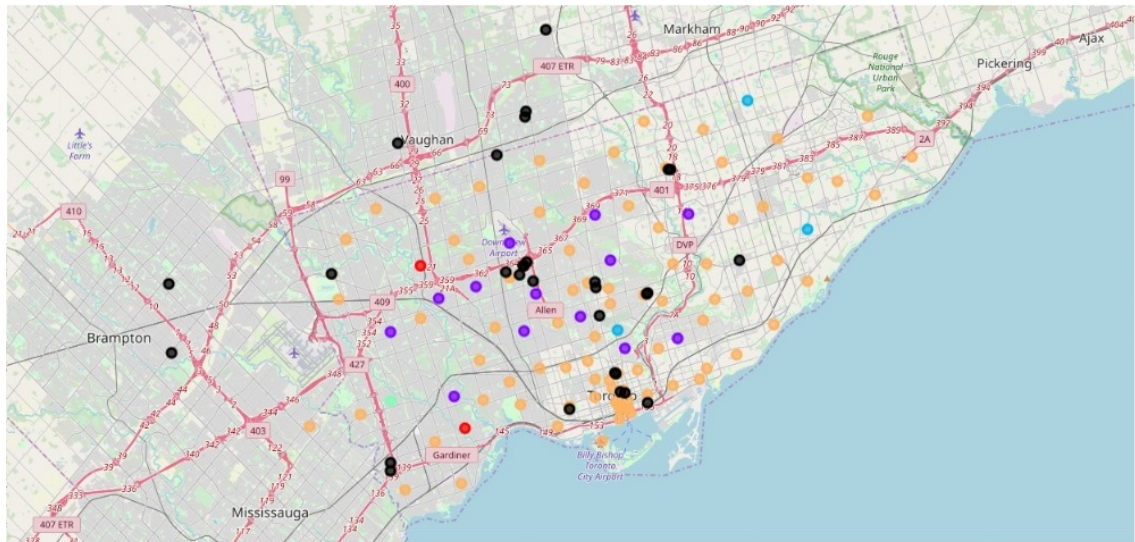
Map 2- *Maps showing distribution of clusters and existing sport shops (black) in Toronto*

Each cluster had similar venues than the other. The following Inferences were drawn

- **Cluster 1, 2 and 3 –** These clusters did not have many shopping areas or playgrounds. Since the company is planning to open only one outlet for start, it should be in the location either near to playgrounds or shopping centres and restaurants. These areas for this reason was not having existing sports outlet.
- **Cluster 0** (displayed in green in map 2) - This cluster has neighbourhoods with Baseball fields. However there are not many restaurants or other venues which attracts people. Baseball players do frequent this area. However given the climatic conditions this area has high footprint in summers only. Company decided to keep this these locations for future Baseball themed outlet.
- **Cluster 4 –** These are the neighbourhoods where the existing sports shops are spread and the neighbourhoods with similar venues like playgrounds, shopping centres, restaurants etc. Thus these neighbourhoods attract people who are out for SHOPPING. Hence neighbourhoods from this cluster are considered for next step.

Next Step – Further data from neighbourhood of cluster 4 like rental cost, space availability and footprint (using Foursquare Trending for 2 months everyday) the Location of the First Sport Outlet in Toronto will be finalized.