COURSERA CAPSTONE PROJECT

IBM DATA SCIENCE PROFESSIONAL SPECIALIZATION

OPENING A NEW HOTEL IN CANADA

BY:- NITIN VERMA

JUNE 2019

1) INTRODUCTION

Tourism is a great way to generate revenue. It provides a boost the country's GDP. Tourists try to visit the place where a great number of hotels are available, making it convenient for them to relax and enjoy. Therefore hotels play a great role in attracting the tourists and generating revenue. Hotels can also result in increase in employment rate. Various shops are also set near the hotels to attract the tourists. However if a great number of hotels exist in an area, this results in competition, which can result in increasing the fare for a room in the hotels. Due to which the number of tourists visiting that place decreases.

1.1) BUSINESS PROBLEM

The objective of this capstone project is to study the distribution of various hotels in the cities of Canada. After studying the distribution determining which places have high number of hotels already present in that area and determining which areas would result in greater revenue if a new hotel is setup in that area using machine learning and data science methodology.

1.2) TARGET AUDIENCE

This project is particularly useful for the investors and developers who are looking to develop a hotel or invest in a hotel development project. The project will help in guiding the investors as well as the developers in identifying the locations with higher probability of generating high revenues.

2) DATA

To solve the given problem the following data is required:

- List of all the major cities in Canada.
- Latitudes and longitudes of all those cities. This is required to plot the graph and also get the venue data.
- Venue data, in this case the data about hotels. This data will be required to perform clustering.

2.1) SOURCES OF DATA

Firstly we would require the list of cities. Fortunately, the list of cities is available at *https://simplemaps.com/data/world-cities*. The dataset available here is very beneficial for the study. The dataset contains the list of all the cities in the world. Also the latitudes and longitudes of the cities are also provided. However the dataset contains many other attributes also like *city_ascii, iso2, iso3, admin_name, capital, population, id.* These attributes have to be dropped. Secondly, since the dataset contains the list of all the cities in the world the, Canadian cities have to be identified and the data set has to be cleaned.

2.2) ATTRIBUTES TO BE USED

city, latitude, longitude attributes will be used for the study. However, in order to clean the data set and identify the cities of Canada , attribute- *country* will be used.

3) METHODOLOGY

Firstly we need to get the location of all the cities in Canada. Fortunately this data set can be found at *https://simplemaps.com/data/world-cities*. The dataset contains the list of all the cities of the world and many unnecessary attributes or columns. Therefore the dataset need to be sorted and only the cities present in Canada have to be selected. Many unnecessary attributes like- *iso2, iso3, population, id, city_ascii, capital* etc have to be discarded. Another requirement is the coordinates i.e. latitudes and longitudes of those cities. Again the data set proved to be helpful because the dataset already contains the coordinates of the cities.

After acquiring the necessary details about the cities they were superimposed on the map of Canada which was created using Folium package.

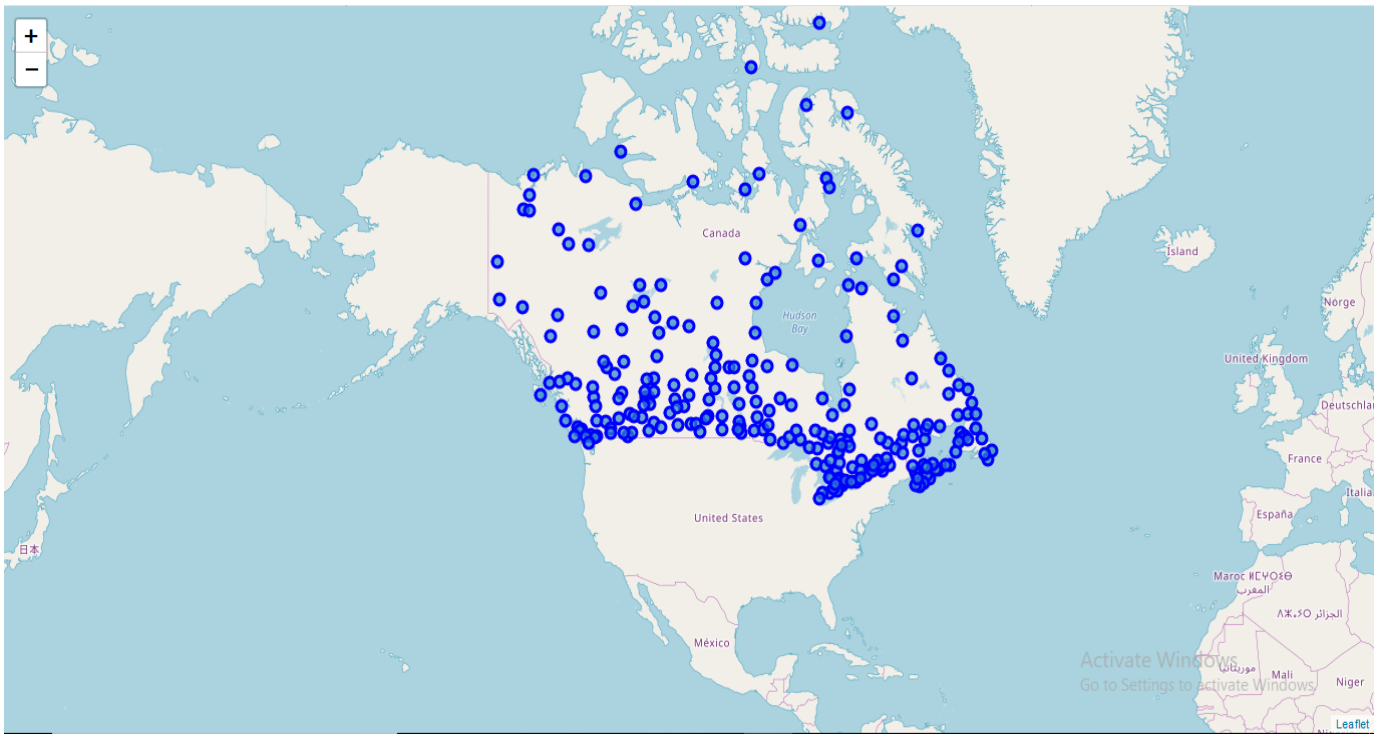
After plotting the map, *Foursquare API* was used to acquire the venue details within 500 meter radius. They were inserted in a new dataframe. Many values were returned for each city in the dataset. For example the city *Abbotsford* had 4 values in *VenueCategory* whereas *Baddeck* had 10. Then it was checked that whether the *VenueCategory* contained *hotels*. Fortunately hotel entry was present along with *gym, diner, park, bakery, grocery store* etc. After that each city was analyzed and the results were grouped together.

K means clustering was used after the results were sorted. The numbers of clusters were selected as 3. Again Folium package was used to visualize the clusters.

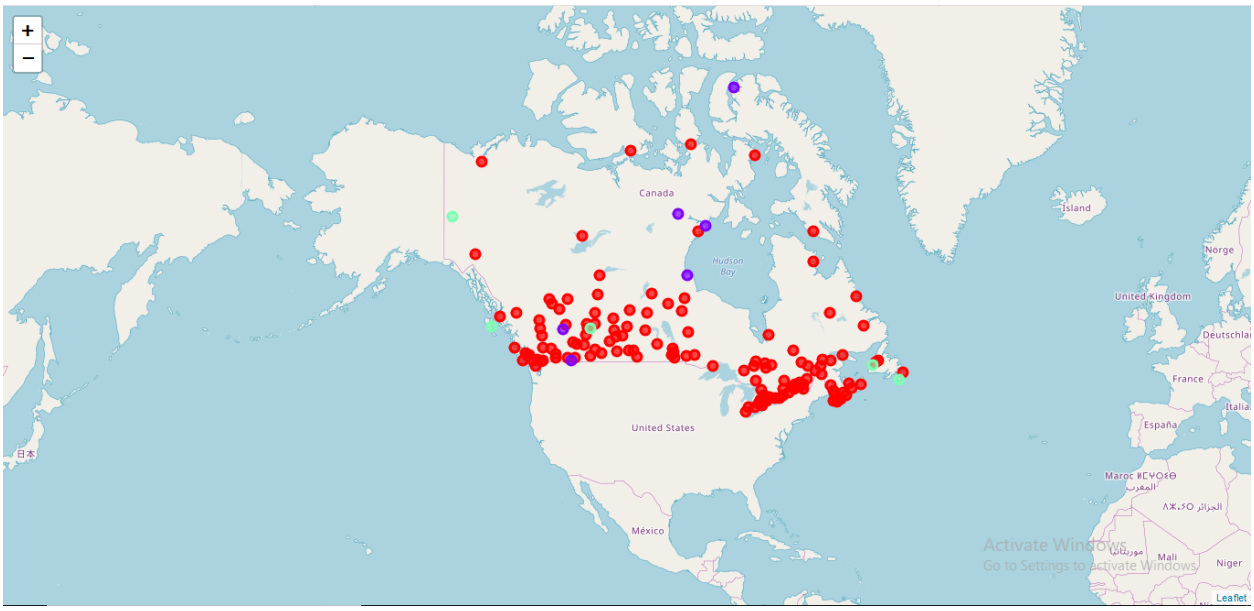
Then finally the results were analyzed.

4) RESULTS

The cities were plotted using the Folium package and the following result was obtained:



The following outcome was obtained which shows various clusters:



5) OBSERVATIONS

On the basis of clusters the following observations were made:

- High number of hotels exists in cluster 2.
- Moderate number of hotels exists in cluster 1.
- Low number of hotels exists in cluster 0.

Cluster 2 is represented red color. Cluster 1 by blue and cluster 0 by green.

6) DISCUSSION

Based upon the results that were obtained it is safe to say that setting up a new hotel at the location defined in cluster 2 will be very risky as very high concentration of hotels are already present. Setting up a new hotel there will result in competition. It is highly recommended that new hotel should be setup at either location mentioned in cluster 1 or cluster 0, but preferably cluster 1.

7) CONCLUSION

In this project we have gone through the process identifying the business problem, specifying the data required, extracting the data, preparing the data, cleaning the data and performing machine learning by clustering the data based on similarities and lastly providing recommendations to the stakeholders i.e. developers and investors. It is recommended that setting up a new hotel at locations mentioned in cluster 1 or cluster 0 can result in higher revenue generation.