Final Capstone Project Report – Week-5

IBM Applied Data Science Capstone

Introduction :

This is a capstone project for Coursera's IBM Data Science Professional Certificate. In this project, a hypothetical scenario has been portraited for a concept that there might be better opportunity to open a new Yoga Studio in Toronto. Therefore it might be a great opportunity for an entrepreneur who is from Canada. As the Yoga is very popular among Western as well as Asian communities. Identifying the best possible location to open such a Yoga Studio is one of the most important decisions for this entrepreneur and this project is to design to help and find the most suitable location.

Business Understanding :

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Yoga Studio in the city of Toronto, Canada. By using data science methods and tools along with machine learning unsupervised algorithm i.e. Clustering, this project aims to provide solutions to answer the business requirement / question : ***If an entrepreneur wants to open a Yoga Studio in Toronto city in that case which location in Toronto city should he/she optimally consider focusing business competition and future growth.***

Target Audience :

The entrepreneur who wants to find the location to open a Yoga Studio in Toronto city.

Data Requirements

To analyse and solve this problem, there will be need for below data requirements :

- ✓ List of neighbourhoods of Toronto, Canada
- ✓ Latitude and Longitude of the neighbourhoods

- ✓ Venue data related to existing Yoga Studios in Toronto for final model input and comparison study.

Data Collection :

- ✓ Scrapping of Toronto neighbourhoods from Wikipedia link i.e. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- ✓ Getting Latitude and Longitude data of these neighbourhoods via Geocoder package
- ✓ Using Foursquare API to get venue data related to these neighbourhoods

Methodology :

At first the list of neighbourhoods in Toronto, Canada needs to be extracted. This is possible by extracting the list of neighbourhoods from Wikipedia:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The web scraping has been done by utilizing pandas HTML table scraping method as it is easier and more convenient to pull tabular data directly from a web page into the data frame.

This is only a list of neighbourhood names and postal codes. Then extract the coordinates to utilize Foursquare to pull the list of venues near these neighbourhoods. To get the coordinates, Geocoder Package has been used. Unfortunately the process was not working so the CSV file provided by IBM team has been utilized to obtain the coordinates of Toronto neighbourhoods.

After that data visualization has been performed on the map of Toronto using Folium package to verify whether these are correct coordinates. Then used the Foursquare API to pull the list of top 100 venues within 500 meters radius.

To achieve this a Foursquare developer account has been created in order to obtain account ID and API key to pull the data.

From Foursquare, pulled the names, categories, latitude, and longitude of the venues. With this data, it can be found that how many unique categories from these venues. With this, analysed each neighbourhood by grouping the rows by neighbourhood and taking the mean on the frequency of occurrence of each venue category. This is required for the further clustering process.

Then searched for the venue category as “Yoga Studio”. After that, clustering method has been applied by using k-means clustering.

K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms. As this algorithm is ideally matched for this given business problem and has been identified as the Analytics Approach for this project.

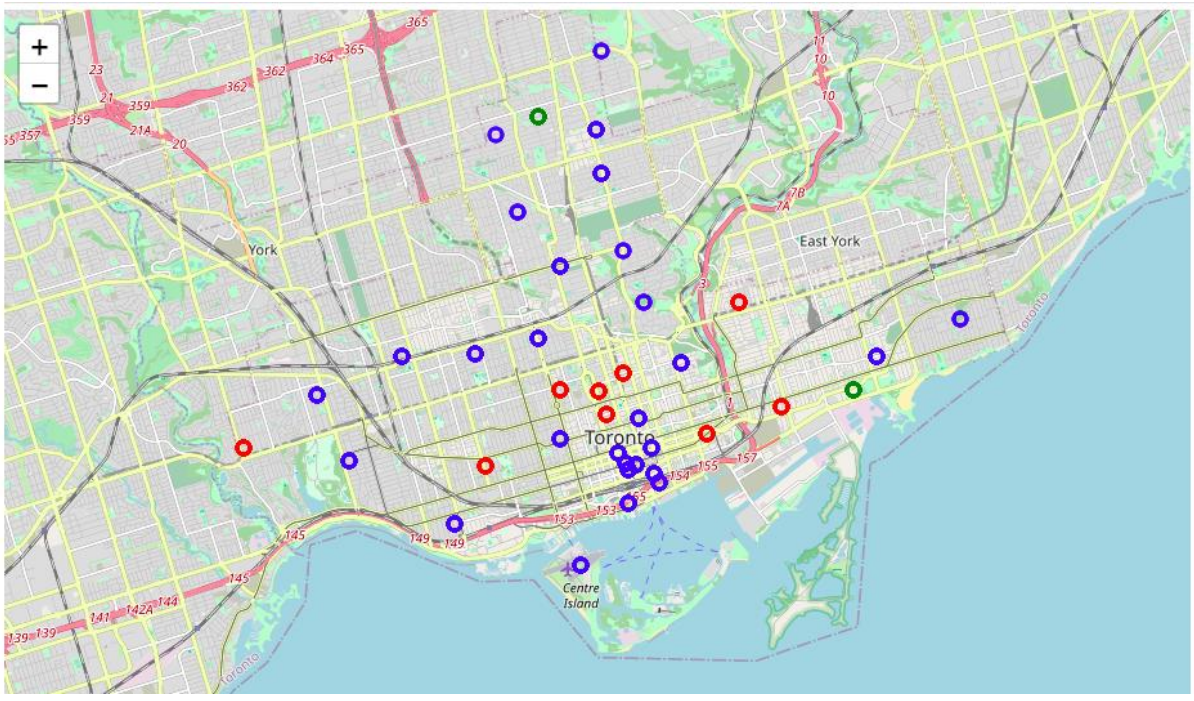
The neighbourhoods in Toronto area has been clustered into 3 clusters based on their frequency of occurrence for “Yoga Studio”. As per the result of the algorithm the ideal place to open a new Yoga Studio can be recommended.

Result :

The results from k-means clustering show that we can categorize Toronto neighbourhoods into 3 clusters based on how many Yoga Studios are situated at each neighbourhood:

- Cluster 0: Neighbourhoods having most numbers of Yoga Studios i.e. ten studios
- Cluster 1: Neighbourhoods having only one Yoga Studio
- Cluster 2: Neighbourhoods having only two Yoga Studios

The results has been plotted on the map as part of the data visualization with colour coding for each clusters such as Cluster 0, Cluster 1 and Cluster 2.



Recommendations :

Among three clusters, cluster 0 having most of the Yoga Studios i.e. around 10. Only 2 studios are present under cluster 2 and only 1 studio is present under cluster 1.

Near North Toronto West, Lawrence Park and South Toronto, having only 2 studios. There are good opportunities to open new one around this area i.e. cluster 2.

Also near Stn A PO neighbourhood , only one studio is available. It looks cluster 1 might also be a good location to setup a new Yoga Studio.

We can have the conclusion for this project and recommends to the entrepreneur to open a new Yoga Studio in the above mentioned locations.