**IBM Applied D.S Capstone Project**

Week 5 – Report

Camilo Echeverry

**Introduction:**

This is a capstone project for IBM Data Science Professional Certificate. In this project, I am going to explore a hypothetical concept to investigate how enough are the Indian Restaurants in Toronto, Canada. Therefore, it might be a great opportunity for an entrepreneur who is based in Canada. As Indian food is popular among the Asian community, so this entrepreneur might think of opening its business in areas where the Asian community resides. With the purpose in mind, finding the location to open such a restaurant is one of the most important decisions for this entrepreneur and I am designing this project to help him find the most suitable location.

Lots of people are migrating to various states of Canada and needed lots of research for good housing prices and repeated schools for their children. This project is for those people who are looking for better neighborhoods. For ease of accessing to Cafe, School, Supermarket, medical shops, grocery shops, mall, theatre, hospital, likeminded people, etc.

This Project aim to create an analysis of features for a people migrating to Scarborough to search a best neighborhood as a comparative analysis between neighborhoods. The features include median housing price and better school according to ratings, crime rates of that area, road connectivity, weather conditions, good management for emergency, water resources both fresh and wastewater and excrement conveyed in sewers and recreational facilities.

**Business Problem:**

To find the most suitable location for the entrepreneur to open a new Indian Restaurant in Toronto, Canada. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if an entrepreneur wants to open an Indian Restaurant, where should they consider opening it?

**Data:**

The data for this exploratory analysis was scrapped by the following [link](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), this dataset was scrapped from wikipedia on Week 3 and has information of latitude, longitude and zip codes. On the other hand, it’s necessary to used Foursquare API Data, basically to explore the need data about different venues in different neighborhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API. After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each neighborhood. For each neighborhood, we have chosen the radius to be 100 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighborhood.

2. Neighborhood Latitude.

3. Neighborhood Longitude.

4. Venue.

5. Name of the venue.

6. Venue Latitude.

7. Venue Longitude.

8. Venue Category.

**Methodology:**

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia: https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M I did the web scraping 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. However, it is only a list of neighborhood names and postal codes. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder Package, but it was not working so I used the CSV file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering these coordinates, I visualize the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius.

I have created a Foursquare developer account to obtain account ID and API key to pull the data. From Foursquare, I can pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Here, I made a justification to specifically look for “Indian restaurants”. Lastly, I performed the clustering method 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 and it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Indian food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

**DATA EXPLORATION** - Firstly, we need to get the list of neighborhoods in Toronto. Fortunately, the list is available on the web page (https://toronto7.com/postal-codes-in-toronto/). We must do web scraping using Python requests to extract the list of neighborhood data. However, this is just a list of pin codes, postal office names, and cities.

**DATA GEOCODING** - We need to get the geographical coordinates in the form of latitude and longitude to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert the address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a panda Data Frame.

**DATA VISUALIZATION** - Visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinate’s data returned by Geocoder are correctly plotted in the city of Toronto.

**INFRASTRUCTURE EXPLORATION** - Next, we make use of Foursquare API to get the top 225 venues that are within a radius of 625 meters. We need to register a Foursquare Developer Account to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude, and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category.

**DATA WRANGLING** - We are also preparing the data for use in selection. Based on the occurrence of infrastructures in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new infrastructures and which neighborhoods are most suitable to visitors to stay.

**DATA CLUSTERING** - Finally, we will perform clustering on the data 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 and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “no. of existing infrastructures”. The results will allow us to identify which neighborhoods have higher, medium, and lower concentration of infrastructures. Based on the occurrence of infrastructures in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new infrastructures.

**Results:**

Let’s analyze the results for the clusters, as we can see in the image; The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Indian restaurants are in each neighborhood, so:

* ****Cluster º0: Neighborhoods with the less number of Indian restaurants.
* Cluster º1: Neighborhoods with no Indian restaurants.
* Cluster º2: Neighborhoods with a more number of Indian restaurants.

**So it’s ton ice the result that we can see in the image, because it show us that there are many** neighborhood that have not yet an indian restaurant and if we analyze the poblation density of Indian people in Toronto maybe it coul give us an idea of business oportunity to start a new restaurant.

**Recommendations:**

Most of the Indian restaurants are in cluster 2 which is around Central Bay Street, Church and Wellesley, Berczy Park, Union Station, Richmond, lowest in Cluster 1 areas which are in North Toronto West and Parkade areas. Also, there are good opportunities to open near St James Town,Cabbagetown Looking at nearby venues it seems cluster 0 might be a good location as there are not a lot of Indian restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic Indian restaurant in these locations.

In this project, using k-means cluster algorithm I separated the neighborhood into 10(Ten) different clusters and for 103 different lattitude and logitude from dataset, which have very-similar neighborhoods around them. Using the charts above results presented to a particular neighborhood based on average house prices and school rating have been made.

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appreciation. This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.

**Future Works:**

This project can be continued for making it more precise in terms to find best house in Scarborough. Best means on the basis of all required things (daily needs or things we need to live a better life) around and also in terms of cost effective.

**Libraries Which are Used to Develope the Project:**

Pandas: For creating and manipulating dataframes.

Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.

Scikit Learn: For importing k-means clustering.

JSON: Library to handle JSON files.

XML: To separate data from presentation and XML stores data in plain text format.

Geocoder: To retrieve Location Data.

Beautiful Soup and Requests: To scrap and library to handle http requests.

Matplotlib: Python Plotting Module.