CAPSTONE PROJECT: THE BATTLE OF NEIGHBORHOODS (WEEK 1)

By Ajay Mutreja (April 2020)

1. BACKGROUND:

Ever since the US President Donald Trump tightened the visa policies for existing and new immigrants to the USA, the US companies have started to look at Canada as an alternative playground for their businesses, specially the IT companies which depend on skilled workers from all over the world.

Canada Government took the opportunity to offer new visa schemes for skilled workers to immigrate to Canada, Toronto, Ontario as an example has offered many nouvelle benefits for such companies and its workers.

A global relocation company has huge demand from their customers to provide Canada Destination Information and best match of the neighborhoods for their clients.

They have asked me to build a program on their app, in which the clients just need to enter the city from where they are moving in to Toronto and the system matches them to the best neighborhood, so that people can satisfactorily settle in quickly and can plug and play into their jobs without worrying.

"Toronto Neighborhood" is a test case, if the system is successful the company may implement it for all Canada and scale up later to the global application.

2. BUSINESS PROBLEM:

Recently, a family approached the relocation company. They wish to move from Mumbai, India to Toronto and they have the following requirements.

1. They want to move to neighborhood which matches their social needs, be around Indian Community, grocery store, restaurants and around city center.
2. They want to know the housing prices in that neighborhood, and
3. They want to know about the Schools in the neighborhood for their two children.

3. TARGET AUDIENCE:

Anyone who wishes to move to Toronto from anywhere in the world.

This Project aim to create an analysis of features for a people migrating to Toronto to search a best neighborhood as a comparative analysis between neighborhoods.

The features include matching the neighborhood with the lifestyle as close as possible to the life style of the city people are migrating in from, expected housing prices and various school options based on their ratings, it may help people to get awareness of the area and neighborhood before moving to a new city to start a new fresh life.

4. PROBLEM ADDRESSED:

Build a program to give a dependable recommendation, based on real time data analysis.

5. THE DATA SCIENCE WORKFLOW & DATA DESCRIPTION:

This project will rely on public data from Wikipedia and Foursquare.

Canada Neighborhood Data - (Source Identified - Scraped from Wikipedia - Canada,. Ontario, Toronto, Postal Code)

Data Link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Data wrangling and Cleaning - Data is found but is not in a useable form, so data wrangling and cleaning is performed to create appropriate data frames.

The cleansed data will then be used alongside Foursquare by invoking Foursquare API credentialed location information

Foursquare location data will be leveraged to explore or compare districts around Central New Delhi, India with neighborhoods in Toronto, where consumers go for shopping, dining and entertainment.

Get data about different venues in different neighborhoods of the specific borough.

For each neighborhood, we chose the radius of 100 meter.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes, top get following information;

Neighborhood, Neighborhood Latitude, Neighborhood Longitude, Venue, Name of the venue e.g. the name of a store or restaurant, Venue Latitude, Venue Longitude, Venue Category.

Data manipulation and analysis to derive subsets of the initial data, segment them, group them, apply K Means clustering algorithm to cluster the neighborhoods similar to Mumbai and use the cluster data frames to output the property prices and school information.

Visualization: Analysis and plotting visualizations, using various mapping libraries.

Libraries which are used to develop the Project:

PANDAS: For creating and manipulating data frames.

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: To scrap and library to handle http requests.

MATPLOTLIB: Python Plotting Module.

6. CONCLUSIONS:

Using k-means cluster algorithm I separated the neighborhood into 10(Ten) different clusters and for 103 different latitude and longitude 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.

Observe the difference between Map 1 and Map 2, Map 1 Plots all the postal codes in Downtown Toronto however in Map 2 plots only the recommended neighborhoods based on the Indian community density clusters. Also the average housing pricing index gives the average prices of various neighborhoods using the merged data frame and the school ratings of various schools in the neighborhood.

Based on these information the immigrant can easily choose where to stay, how much to budget and what to expect. This helps in confidence immigration leading to customer satisfaction.

7. DISCUSSION:

This program can be built for Canada, For Canada Postal Codes Data is available,

But data cleaning and converting them into data frame may not be able to be automated, as the formats of data tables on Wikipedia are non-standard.

The Neighborhood Data may not be available for all the cities of the world, therefore establishing global program may be a huge challenge.

Average Housing Price index and School ratings for all the cities may not be available, therefore the data will have to collected through primary surveys

The data may be available but may be available in different language, therefore additional ML translation program may have to be invoked for data gathering.

8. ACKNOWLEDGEMENT:

My Work is in continuation to the inspiring work of many capstone project pursuers on the same subject.