Final Report

1. Introduction:

The objective of this Project is to identify better facilities around different neighbourhoods which will help people exploring these facilities smartly to take effective decision by selecting great neighbourhood out of many other neighbourhoods in Toronto.

A lot of people are migrating to a various state of Canada who need to know many important factors like housing prices, reputed schools, available entertainment options, shopping location, food shop and hospital etc. This project aims to enable people looking for better neighbourhoods considering the above factors.

Business Proposal :

Here, I will determine the features of the neighbourhood and perform a comparative analysis between them. The features include median housing price and better school according to ratings, crime rates of that particular area, road connectivity, weather conditions, good management for emergency, water resources both fresh and wastewater and excrement conveyed in sewers and recreational facilities.

2. Target Audience :

This project will facilitate people by generating awareness of the surrounding area and the neighbourhood before moving to the new place. While planning for suggesting a better neighbourhood in a new city for the people who are moving, I found some challenges:

- Identifying similarity of people's social presence and culture
- Determining Connectivity to major location or landmark like airport, bus stand, city centre, shopping mall
- Analysing the facilities.

For this,

- List of housing in terms of prices
- List of educational institutions in terms of location, fees, rating, and reviews.

3. Data Description :

The data comes from dataset from the following link. It contains 227,428 check-ins in New York city. The data contains two files in csv format.
https://sites.google.com/site/yangdingqi/home/foursquare-dataset
Each file contains 8 columns, which are:

- User ID (anonymized)
- Venue ID (Foursquare)
- Venue category ID (Foursquare)
- Venue category name (Foursquare)
- Latitude
- Longitude
- Time zone offset in minutes (The offset in minutes between when this check-in occurred and the same time in UTC)
- UTC time

1. **Foursquare API:** In this project, I will use Four-square API as a prime data source. Because Foursquare has a database of millions of places, especially the API provides the ability to perform location search, location sharing and details about a business type and business category.
2. **Workflow:** I will use the credentials of Foursquare API features of near-by places of the neighbourhoods and mine data. Due to http request limitations, I will set the number of places per neighbourhood parameter to 100 and the radius parameter to 500.
3. **Libraries:**
    - Pandas: To create and manipulate data frame
    - Matplotlib: Python Plotting Module.
    - Beautiful Soup & Requests: To scrap and library to handle http requests.
    - Folium: Python visualization library would be used to visualize the neighbourhoods cluster distribution of using interactive leaflet map.
    - JSON: Library to handle JSON files.
    - XML: To separate data from presentation & XML stores data in plain text format.
    - Geocoder: To retrieve Location Data.
    - Scikit Learn: To import k-means clustering.

4. Methodology :

After extracting and reading the data, we will translate above data into a Pandas data frame for processing which would look like this. These are data elements that are needed when we call Foursquare web service call in order to get venues available in that neighbourhood (Neighbourhoods are not included here).

	VenueID	CategoryName	Visitor Count	Latitude	Longitude
0	49bbd6c0f964a520f4531fe3	Arts & Crafts Store	7	40.719810375488535	-74.00258103213994
1	4a43c0aef964a520c6a61fe3	Bridge	37	40.60679958140643	-74.04416981025437
2	4c5cc7b485a1e21e00d35711	Home (private)	1	40.716161684843215	-73.88307005845945
3	4bc7086715a7ef3bef9878da	Medical Center	1	40.7451638	-73.982518775
4	4cf2c5321d18a143951b5cec	Food Truck	4	40.74010382743943	-73.98965835571289

Then we will create a dictionary in order to decide which category is the most popular (commercial type)

```
[('Train Station', 943), ('Park', 778), ('Airport', 769), ('Bar', 756), ('Subway', 587), ('Coffee Shop', 447), ('Gym / Fitness
Center', 447), ('Food & Drink Shop', 426), ('Neighborhood', 362), ('Plaza', 342), ('Stadium', 339), ('Bridge', 272), ('Office',
264), ('Department Store', 240), ('Mall', 238), ('Burger Joint', 206), ('American Restaurant', 202), ('Road', 201), ('Bus Stati
```

```
'Bar' is the most visited commercial category according to given data.
```

After all this, we will check the coordinates within given n number of kilometres and count how many ‘Bars’ are there (venues selected as 2000 as a trial)

```
Coordinates with number of Bar shops within 4 kilometers according to 2000 venues.

('40.60613336268842', '-74.17904376983643') : 2
('40.719810375488535', '-74.00258103213994') : 0
('40.60679958140643', '-74.04416981025437') : 0
('40.716161684843215', '-73.88307005845945') : 0
```

Find the two neighbourhoods that are closest to the coordinate which has the greatest number of the specific shop type but lacking that within 4 kilometres.

5. Results and Conclusion :

In our sample of 2000 venues, we did find more than 10 coordinates that has no Bar (the most visited shop type according to sample) within four-kilometre sphere. And we did manage to get the neighbourhoods' names from foursquare database and pin down the two closest neighbourhoods, 'Bedford-Stuyvesant', and 'Turtle Bay', into the map. Of course, it should not be forgotten that the data used above is almost 6-year old so further research might be needed. Anyways, the results according to the data in hand can be checked from the map and analysis above can be of use for future entrepreneurs.