**Final Report**

**Best Neighbourhood Recommender system**

**Business Problem:**

This project is for the raw material contractors who are looking to start new business in the Downtown Toronto Borough of Toronto city. The contractors who want to deliver good quality, fresh vegetables and other raw materials to the different types of restaurants, cafes, breakfast spot and bakery in city, for that this recommender system will recommend best neighbourhood in which he/she can build a warehouse so that it will cover most of the restaurants nearby and provide quality service to them by providing on time and fresh delivery of raw materials.

Example for the above scenario is if the contractors starts his business in the area near by old famous restaurants, then the fresh vegetables and other raw materials can be provided to them on time and there would be no delay for restaurants cooks to start their job. Also there will be very less cost of transportation in this case. In this way the contractor will be appreciated for his work and will gain more income in that area.

Finding the best neighbourhood for the contractors to start his business is the problem here and our recommender system will suggest sorted list of the best neighbourhoods to the contractor**.**

**Data:**

For this project we need the following data:

1. Data is scrapped from Wikipedia table and stored in dataframe.

Data source: https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M

Description: Data contains three columns as below:

1. Postal Code : Postal code of the area
2. Borough : Name of the borough
3. Neighbourhood : Neighbourhood areas of the borough
4. Second dataset used is Geospatial\_Coordinates.csv which is provided in the course

This dataset contains latitude and longitudes of the different postal codes.

1. Third Data source used is foursquare API to get the venue details in neighbourhood.

Data source: Foursquare API- "https://developer.foursquare.com/"

Description: By using this API we will get all the venues and its details in each neighbourhood.

Example: We can get the venue details by passing latitude and longitude to the foursquare API.

**Final Dataset contains following columns:**

1. Postal Code : Postal code of the area
2. Neighbourhood : Neighbourhood areas of the borough
3. Neighbourhood latitude : latitude of neighbourhood
4. Neighbourhood longitude : longitude of neighbourhood
5. Venue : name of the venue
6. Venue summary : Summary of the venue
7. Venue category : category in which venue belongs
8. Distance : Distance of the venue from neighbourhood

**Methodology:**

1. **Collecting Dataset:**

Data is collected from Wikipedia table <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> . We will use different postal code from table and co-ordinates of each postal code to find venues in different neighbourhood.

1. **Using Foursquare API to retrieve venues:**

Once the data is collected and converted into dataframe, we are using Downtown Toronto Borough data here. To find venues in different neighbourhood of Toronto we used foursquare API which will provide us with the different venues and its details in neighbourhood. We kept radius limit of 1500 to limit the venues in area.

1. **Processing and cleaning Data Obtained:**

After all the data is obtained from Wikipedia and foursquare API, data needs to be cleaned to apply various techniques on it. Main useful feature from collected data is ‘venue category’ which helps us in identifying restaurants. First we will one hot encode venue category, then we make groups by neighbourhoods and then we will take count of total restaurants and add ‘Total Restaurants’ column in dataframe. Same is done for joints and ‘Total Joints’ columns is added in dataframe.

1. **Applying K-Means Clustering Machine Learning Technique:**

Now that our processed data is ready we will apply machine learning Model on the data. In this problem we used K- Means clustering machine learning model to cluster neighbourhoods. We used cluster value of 5 so that we can divide all the areas in 5 clusters. After Clustering is done we created new dataframe which will give us neighbourhood and its cluster number.

**Results:**

For results we focused on centres of each clusters and found that G5 is the best group. Total sum is calculated from addition of Total restaurants, total joints and other restaurants. According to our results we will recommend group 5 as the best neighbourhood to start the new business and group 4 is worst neighbourhood.

**Discussion:**

Based on results group 5 is the best neighbourhood and its details as follows:

Neighbourhood: 'Kensington Market, Chinatown, Grange Park'

Neighbourhood Latitude: 43.6532057

Neighbourhood Longitude: -79.4000493

Based on results group 3 is the second best neighbourhood and its details as follows:

Neighbourhood: 'Christie'

Neighbourhood Latitude: 43.669542

Neighbourhood Longitude: -79.4225637

**Conclusion:**

In this problem I analysed that 'Kensington Market, Chinatown, Grange Park' is the best neighbourhood in Downtown Toronto Borough. Also Venue category feature was very useful to cluster the data. K-Means clustering is very powerful machine learning technique to cluster our data in groups and gives very good output in real life scenarios as well.