1. **Introduction/Business Problem**

**New Business venture in a new city**

***Often times a new business start and then collapse within the first year, in fact statistics have shown that most unsuccessful business will close within the first 12 months. So it is imperative that before starting a new business one has to do a lot of homework about WHERE to open the business to maximize its probability of survival and also make it successful in the long run.***

***This project is about finding the best location to open a new business such as a restaurant in a new city like Toronto or New York with focus on ethnicity and population of each region of the city. Other considerations can be weather and transit access but those will not be detailed much. This would probably be future work.***

***Let’s call it the Recommendation Project.***

***Venue : New Indian restaurant in Toronto***

**2. Data Requirements**

**The following methods and techniques will be used to gather the required data to go about solving this problem:**

1. ***Get Toronto neighborhood data with their geographical information like latitude and longitude.***

***SOURCES :***

Postal Codes information from Wikipedia to provide neighborhood details such as Borough and population etc

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

This page will have to be scraped, cleansed and enriched. Then it will have to be transformed into STRUCTURED FORMAT using pandas dataframe.

Next we need to get Geographical coordinates of each postal code is available as CSV file at url '<http://cocl.us/Geospatial_data>'. These coordinates data have to be read and merged with data specified in Step 1 so that the new format would like below:

New Data Format with geographical coordinates: Postal Code, Borough, Neighborhood, Latitude, Longitude

***2. For each neighborhood, use Foursquare location data to gather all its different venues and categories***

Foursquare location data would be collected in the following format against each neighborhood: Neighborhood, Neighborhood Latitude, Neighborhood Longitude, Venue, Venue Latitude, Venue Longitude, Venue Category

Cluster the Neighborhoods and to evaluate results based on Top 5 neighborhoods picked randomly or manually.

# 3.0 Methodology

We will make use of various API within Python to clean the data, map the data and then use Foursquare API to get venues in each neighborhood. In this design we will choose the borough with the maximum number of neighborhoods and not go for famous locations such as Downtown. However as you will see form the code it is very easy to adapt the code to pick any other area. This is documented in the code.

There are 4 steps to go about this process (there may be more for other types of ventures)

1. Find out about where we can get all the data needed 🡪 See Data requirements above
2. Use the data to dissect the city and plot the neighborhoods accordingly (Matplotlib)
3. Find all the venues for the neighborhoods. We made use of the FOURSQUARE API.
4. Analyze the results and provide recommendation.

For 2: We need to build a data frame from the data that will have

* Postal Code
* Borough
* Neighborhood
* Latitude
* Longitude

Beautiful Soup will work its magic on the web data and create the above dataframe.

Using this data frame we will use MATPLOTLIB to map the data so we can see the concentration.

A count on the number of neighborhoods will also indicate to use the vicinity we are looking for.

At this point we can narrow our search to a specific borough and continue to work on this borough only for steps 3 and 4.

# 4.0 Results

For step 3, we will use the FOURSQUARE API to give us all the venues for each neighborhood in the borough that we have chosen in the previous step. We will receive a JSON object for each request to the API and we need to collect the data elements from the JSON object.

Finally the venue with the highest neighborhood count will probably be the optimal place for our new restaurant.

Again, we are looking at the place where there is enough neighborhoods and not necessarily famous places like DOWNTOWN or where there is a HIGH concentration of people who we think will be more likely to go to an Indian Restaurant. For this exercise the borough of NORTH YORK was the chosen one based on what was described in step 2.

Within North York the neighborhood with the most venues happens to be FAIRVIEW MALL area and this was our chosen location.

# 5.0 DISCUSSION POINTS

As I went through this exercise I realized that there can be many ways to “skin the cat”. Data is out there but there may be different ways to use the data and come to different conclusions. I think it is very important to set the CRITERIA of what will constitute the MOST LIKELY scenario you looking for before even attempting to use the data out there.

Another exercise would be to totally ignore boroughs and neighborhoods and only mine crowdsource data and come up with a location. Let’s say we do not want to go to a new location but rather share the market where it is already established that there are various Indian restaurants. We do not have to worry about whether people will go to that location because it is already established.

# 6.0 CONCLUSION

###### The Business Model has definitely changed over the past few years where DATA is once again at the centre and everything else revolves around data. Recommender systems are very widely used and if one uses any sort of online transaction to make purchases, one is sometimes puzzled with what else they want to sell to you.

Yes there seems to be TOO MUCH data out there and sometimes we are overwhelmed with the reviews, the advice and the general information about products or places but one thing for sure is that we do have CHOICES now. And we can do things with more confidence be it a simple purchase online or a big investment like opening a restaurant or any other business.