**Predicting the areas to start an Indian Restaurant in Toronto**

Ankush Lakkanna

1st April, 2019

**1. Introduction**

**1.1 Background**

Toronto is one of the most developed cities in the world. It is easier to find a job there than to start a business of your own. Firstly, you will need to find an area where the business will flourish and then you will have to search for a vacant real estate for your business to start up. Now finding the right neighborhood is a tough task because each neighborhood has almost everything and you will need to find a place where the idea of your business is lacking so that you can develop a proper customer base. It is a tedious task looking around each neighborhood to search for the right neighborhood and that is where data science comes in. Using Data Science, we can find the best neighborhood to start your business in a very short span of time.

**1.2 Goal**

The goal of this project is to find the best place to start an Indian Restaurant. As an Indian myself, I would love to start a business in Toronto and it is a dream for the business to flourish.

**1.3 Problem**

The problem here is that it is very hard to find the right place to start an Indian Restaurant because Toronto has a very diverse set of restaurants everywhere, most of which seems established. So, for someone who wants to enter the restaurant business, it seems very hard to find the place to set up the business. It is very crucial to find the right spot because, constructing a restaurant is very costly and it is a onetime investment. It is even harder to relocate, once the business starts failing, so we must make the right choice in choosing the location.

**1.4 Plan of Action**

We can solve this problem by first listing out all the neighborhoods and then using the Foursquare API to find all the types of facilities in each neighborhood. We then find out the list of neighborhoods containing Indian Restaurants and then we use this to filter out the neighborhoods that do not contain Indian Restaurants. These neighborhoods are our targets. We can check the ethnicity of these Neighborhoods and find out the ones which contain the most Indian Population because Indians are our prime customer base. Finally, we can choose the best area to set up and Indian Restaurant.

**1.5 Interest**

This project will be useful for the people of India who wish to move to Toronto to start a restaurant of their own. We can even tweak this project to cater to people who want to start up Indian Dessert Shops. Also, we can tweak this project such that it can be used to analyze another city altogether, so that Indians moving to those cities can find the best area to set up a restaurant of their own.

**2. Data Acquisition and cleaning**

**2.1 Data Requirements**

Since, we are looking at finding the best neighborhood for setting up our Indian Restaurant, we will need the Toronto Neighborhood data. Like our previous projects, we need the borough of the neighborhood and the latitude and longitude of the neighborhood. Once we have this, we will have a clean dataset to work with. Since we are also looking at the ethnicity of our customer base, we will need the Toronto Ethnicity dataset. We should have the neighborhood and the borough and the ethnicity groups in that neighborhood.

**2.2 Data Sources**

We can collect the Toronto Data from online websites, and we can create a data frame out of it. We obtain the list of neighborhoods using <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>, for geospatial coordinates we use <http://cocl.us/Geospatial_data> and ethnicity data from <https://en.wikipedia.org/wiki/Demographics_of_Toronto>

**2.3 Extraction and Cleaning**

Data was scraped from the Wikipedia link by using the pandas.read\_html() function. It was one of the easier methods that I found, and it does a good job of scraping the website and automatically lists out all the wikitables in that Wikipedia website as an array. The wikitables are already in the form of a dataframe and we can assign the table we want to use to a variable as well.

One problem of using the read\_html() method is that the actual column names will be at row 0 and the column names of the dataframe will be numerical values. We can drop row 0 and using the rename function we can rename the columns.

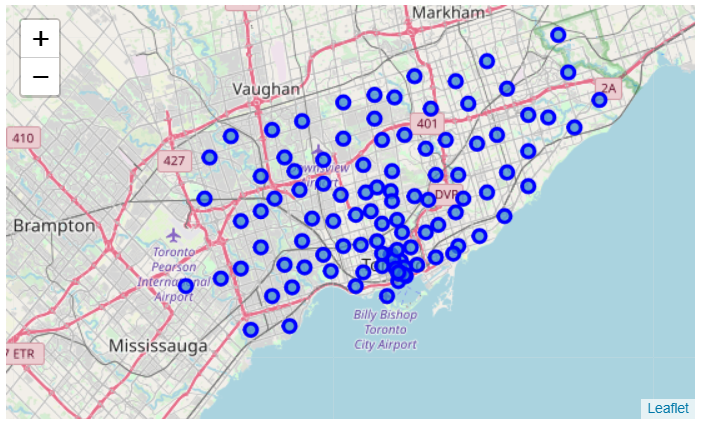
The dataset did contain a lot of missing values. The rows with missing neighborhood names were assign their respective borough names and the rows with the missing borough names were removed. What resulted from this was a dataframe containing the postcode, borough and the neighborhood.

I merged all the neighborhoods having the same postal code into one row to eliminate redundancy. Using the postcodes in the geospatial coordinates dataset I was able to merge the latitude and longitude of the neighborhoods accordingly.

**3. Exploratory Data Analysis**

**3.1 Plotting the Map of Toronto**

Now that we have all the names of the neighborhoods, boroughs, postcodes, latitudes and longitudes, lets plot the neighborhoods using the folium package. Folium does provide interactive maps for the plots we create. We can zoom in and out and we can move the map around. The picture below shows the map of Toronto with all the neighborhoods highlighted with blue circular dots.



**3.2 The Foursquare API**

The Foursquare API is one of the best ways to get venues of each area. We use the client ID, client secret, the version, latitude and longitude to obtain the list of venues in an area. The results are in the form of a json file.



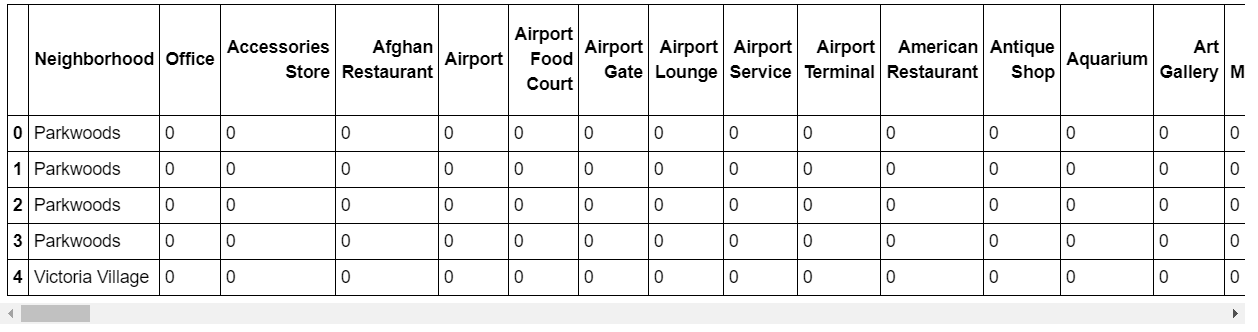
**3.3 Venues and categories**

We use the json\_normalize() function to create a dataframe of the columns we need. We are mostly interested in the categories of the venues.



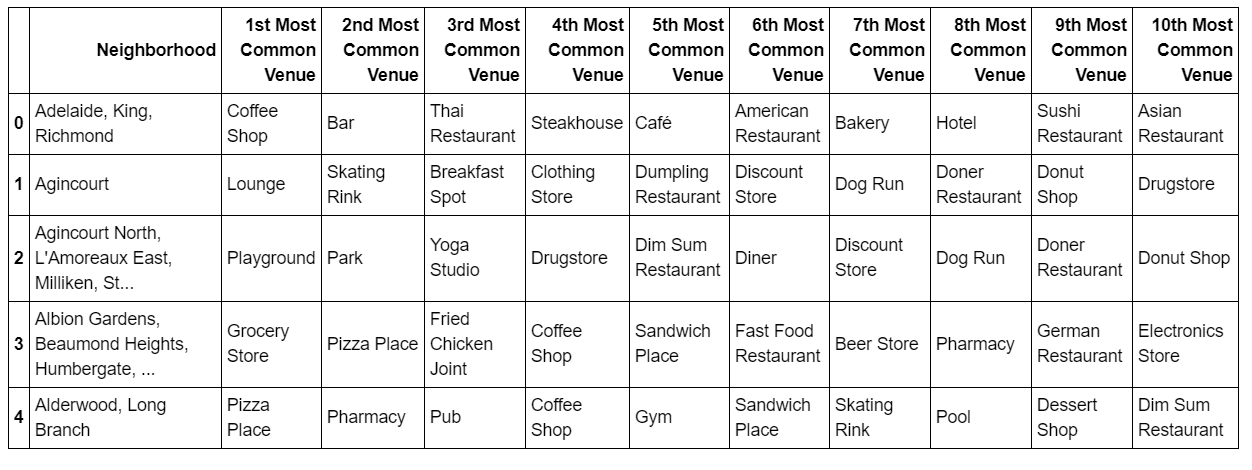
**3.4 One Hot Encoding**

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. We perform one hot encoding for each type of venue for each neighborhood in Toronto which gives us a dataset like the one shown below



**3.5 Top 10 venues**

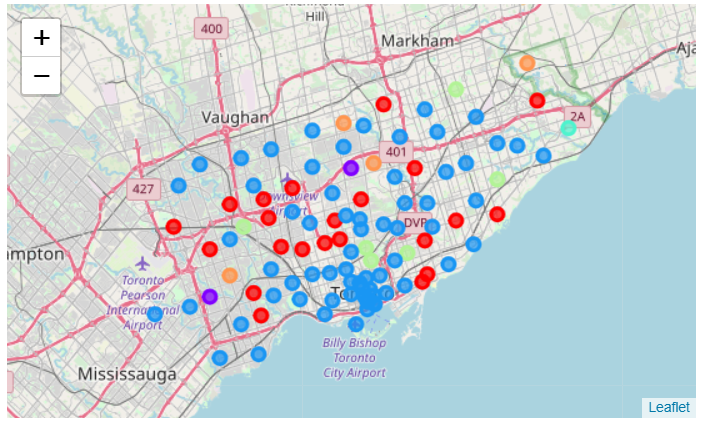
Now that we have all the venues for each neighborhood, we then create a dataset which contains the top 10 most common venues for each neighborhood. The common venues are based on the venue category and not by the name of the venue.



**4. Predictive Modelling**

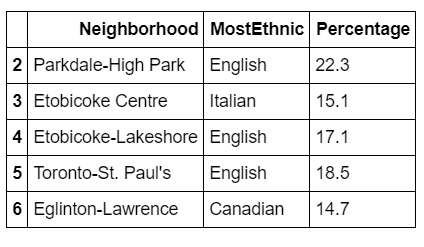
We want to find neighborhoods that are similar to each other. This is based on the similarity of the venue categories for the top 10 venues for each neighborhood. I found that K-Means was appropriate for performing clustering. It could provide proper cluster labels for each neighborhood.

On performing K-Means, we get a list of cluster labels. I merged the cluster labels with the top 10 venues dataset. The I sorted the dataset according to cluster labels. Then I plotted the clusters on a folium map with a different color for each cluster.

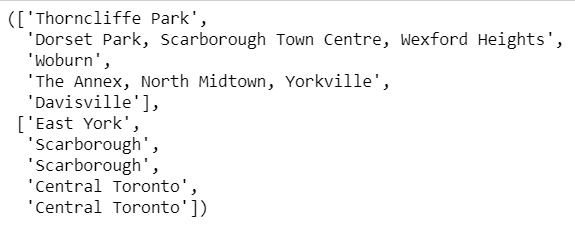


**5. Ethnicity Analysis**

Now that we have all the places and the venues, we need the ethnicity of people for each area in Toronto. The best dataset that I could find was the ethnicity of each electoral area. It also consisted of a percentage of the most ethnic group in each area.



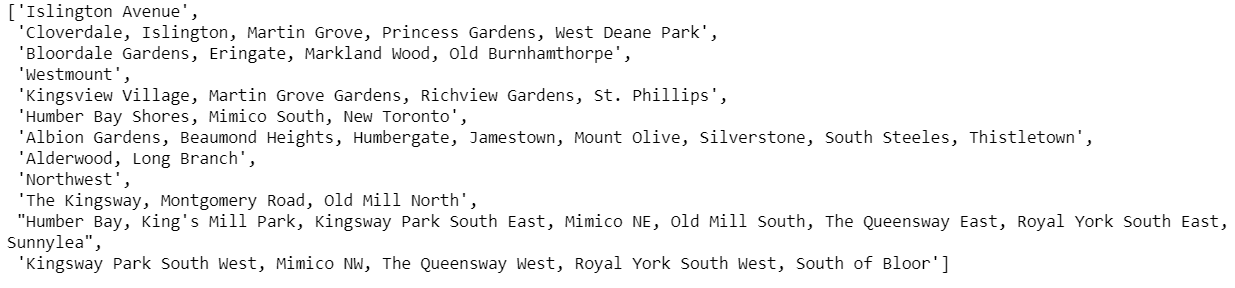
Since we have borough names, I decided to check the boroughs where the most common venues are Indian Restaurant. I printed out the neighborhood and boroughs to get an idea of the areas to avoid in order to get a loyal customer base.



I then searched for the places containing Indian population. Unfortunately 3 out of the 4 places containing the most Indian population was in Scarborough. The remining area was in Etobicoke, more specifically Etobicoke North.



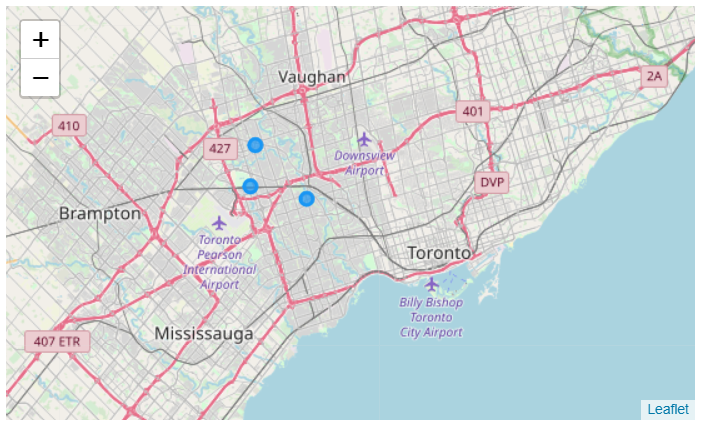
Etobicoke did not have a popular Indian Restaurant even though it had the most Indian population. Now, we know all the areas in Etobicoke but we need to find the neighborhoods most specifically the ones in Etobicoke north. These are the neighborhoods in Etobicoke



I then found the coordinates of the south latitude to get the neighborhoods in Etobicoke North.



These are the most apt neighborhoods to construct an Indian Restaurant in. I plotted the areas on the map of Toronto to get a visual idea of the neighborhoods.



**6. Conclusion**

In this study, I analyzed the venues of each neighborhood in Toronto, the places where Indian Restaurants were common and the ethnicity of each neighborhood so that we could get the best neighborhoods to target in setting up an Indian Restaurant. This model can be tweaked so that it can help other people, for example, Chinese people to identify areas with the most Chinese people, places that contain Chinese Restaurants and finally they will get their target areas to start up a Chinese Restaurant.

**7. Future directions**

I was only able to use the electoral areas to find out the ethnicity of the people in Toronto. I feel that if I find a more accurate dataset that contains the top 10 ethnic groups, with percentages, for each neighborhood of Toronto, then the accuracy of this model can be increased sevenfold. Having ethnicity data for each neighborhood can give us more insights as in giving us a rank for each neighborhood in Toronto and us displaying the top 5 neighborhoods to start a Restaurant in. It is indeed difficult to have a dataset like that, but if done so can be used for many other purposes as well.