IBM Data Science Capstone Project Report

Exploring Toronto and Searching the Best Place to Establish an Indian Restaurant

Table of Contents

Chapter-1

1.1-Introduction

Hello esteemed readers, as part of the IBM capstone project course, I was assigned the task of formulating a hypothetical problem and coming up with a data science-oriented solution to unearth meaningful insights and conclusive results.

As the title suggests, the task here is to explore the beautiful city of Toronto, to traverse through its neighbourhood areas and finally come up with a few suggestive neighbourhoods that have business potential in terms of opening a new Indian restaurant.

When an individual dream about venturing into the hospitality sector (i.e. Hotels & Restaurants). The prime aspect of such a business is to locate a place where the restaurant can be set-up. There are a lot of questions that arise when someone is hunting for a place to set-up a business. The main aspect during this hunt is to locate a place which has the least competition. This enables the new restaurant to thrive without hinderance. We will be locating such places in this project.

1.2-Target Audience

While formulating the problem statement, I took into consideration the prospect of coming up with a predicament that could be faced by individuals in the real world. Hence, I came up with a problem statement that aims at finding the best possible location for a proposed Indian restaurant in the city of Toronto.

This will serve two purposes for two different audience sets. Firstly, this will help individuals who are aiming to start a new business in the hospitality sector (i.e. Restaurants), find a place that has the least concentration of Indian restaurants. Secondly, this could help tourists, choose places (neighborhoods) based on their personal preferences. For example, a neighborhood with a good bunch of chinese restaurants or a neighborhood which is home to a considerable number of parks that tourists can possibly be interested in.

Chapter-2

2.1- Data Set Description

The data set required for the following project was acquired from three different data sources. The three data sources are listed below,

- A Wikipedia Page to fetch boroughs and neighbourhoods of Toronto city.
- A .csv file to fetch latitudes and longitudes corresponding to each postal code.
- The foursquare api to fetch different public venues in the vicinity of the neighbourhood.

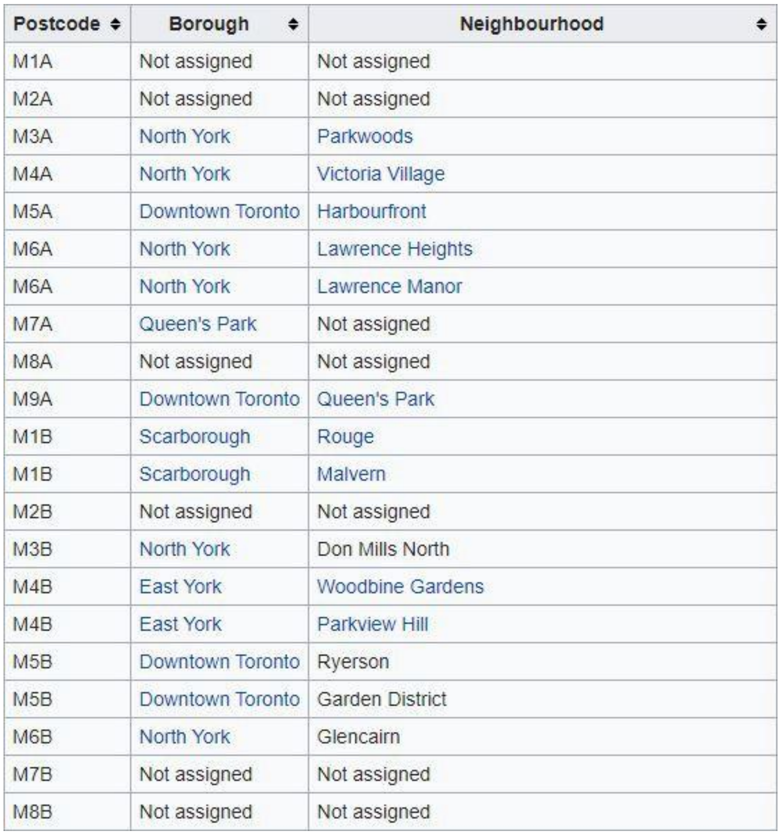
The Wikipedia page contains a table of postal codes followed in Toronto, along with the boroughs and neighbourhoods in Toronto city. The .csv file provides us with the latitude and longitude co-ordinates of each postal code followed in the region of Toronto. This data is beneficial since these co-ordinates are then used in tandem with the four square api to give out a list of popular venues in each neighbourhood.

The data is comprehensive, and yields valuable insights related to Toronto city that eventually helped us in unearthing conclusive results and observations. The data source, as it is perceived at the start of the project is unclean and required intensive pre-processing in order to convert it to a working set, capable of handling machine learning algorithms and visualization operations that were implemented on it.

2.2-Data Set Before/After Comparison

The data set was firstly acquired from the Wikipedia page. As I studied the data set, I noticed that there were a lot of missing values which were stated as not assigned in their respective spaces in the table and working with such data is not advised. It results in the machine learning or visualization algorithms to misbehave and provide incorrect outputs/results.

The data set before pre-processing algorithms were implemented can be seen in the image given below.



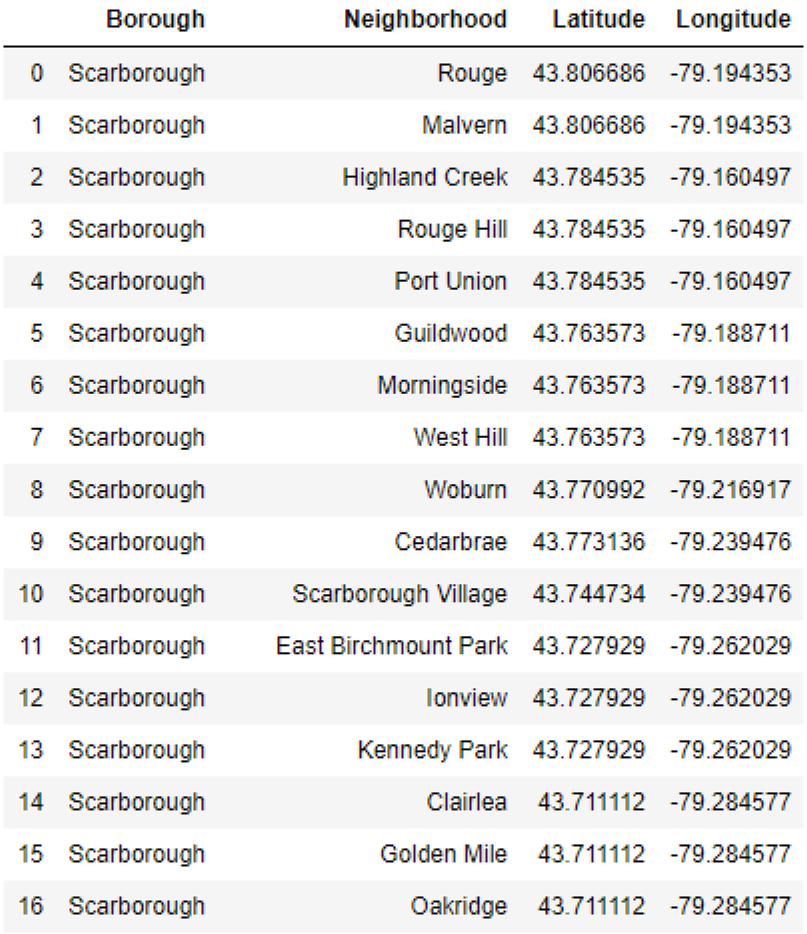
Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M6A	North York	Lawrence Heights
M6A	North York	Lawrence Manor
M7A	Queen's Park	Not assigned
M8A	Not assigned	Not assigned
M9A	Downtown Toronto	Queen's Park
M1B	Scarborough	Rouge
M1B	Scarborough	Malvern
M2B	Not assigned	Not assigned
M3B	North York	Don Mills North
M4B	East York	Woodbine Gardens
M4B	East York	Parkview Hill
M5B	Downtown Toronto	Ryerson
M5B	Downtown Toronto	Garden District
M6B	North York	Glencairn
M7B	Not assigned	Not assigned
M8B	Not assigned	Not assigned

Fig 1: Unclean Data Set

As a reader you can clearly see that the above data set is hard to read since it is not uniform at several places and likewise it is very difficult for machine learning and visualization algorithms to work on such data.

Hence pre-processing algorithms were implemented on the above data set to clean it, make it uniform and remove inconsistencies at places where they occurred. This enabled a seamless implementation of various crucial algorithms that later assisted me in unearthing crucial insights and helped me in solving the problem that we are tackling in this project.

The data set after the preprocessing look clean, neat and consistent and is a blessing to work with for a data scientist as well the algorithms that were later implemented on them. A cleaned sample of the data set is given below in figure 2.



	Borough	Neighborhood	Latitude	Longitude
0	Scarborough	Rouge	43.806686	-79.194353
1	Scarborough	Malvern	43.806686	-79.194353
2	Scarborough	Highland Creek	43.784535	-79.160497
3	Scarborough	Rouge Hill	43.784535	-79.160497
4	Scarborough	Port Union	43.784535	-79.160497
5	Scarborough	Guildwood	43.763573	-79.188711
6	Scarborough	Morningside	43.763573	-79.188711
7	Scarborough	West Hill	43.763573	-79.188711
8	Scarborough	Woburn	43.770992	-79.216917
9	Scarborough	Cedarbrae	43.773136	-79.239476
10	Scarborough	Scarborough Village	43.744734	-79.239476
11	Scarborough	East Birchmount Park	43.727929	-79.262029
12	Scarborough	Ionview	43.727929	-79.262029
13	Scarborough	Kennedy Park	43.727929	-79.262029
14	Scarborough	Clairlea	43.711112	-79.284577
15	Scarborough	Golden Mile	43.711112	-79.284577
16	Scarborough	Oakridge	43.711112	-79.284577

Fig 2: Clean Data Set

Chapter-3

3.1-Data Pre-processing

The data that we need for this project is available at varied places and it is very difficult for a data scientist to perform meaningful analysis, if he/she does not have the right data to work with.

Hence, I first started with cleaning my data. The first step I performed was to scrape data from the Wikipedia page that consisted of all the boroughs and neighborhood along with their postal codes. I converted it into a data frame since they are the best data structure to work with when it comes to analysis using visualization techniques. The data frame, still consisted of many values that can be treated as missing values, since the postcodes were not assigned to any borough or neighborhood. Missing values can cause a discrepancy in results when we approach later stages of the project. Hence, I got rid of all the rows that had missing values present in them. Mind you, getting rid of missing values does not mean that I have lost crucial information. On the contrary, I have made sure that the useful data we have in hand is not hindered by the missing data, that can usually work as outliers, and disrupt results.

The second step included importing data from a .csv file. The .csv file consisted of latitude and longitude co-ordinates of each postal code. This .csv file was imported into a data frame for ease of analysis in the later stage. Followed by which, I merged the data frame consisting of borough and neighborhood information and the data frame consisting of the co-ordinate values. The merge was implemented on the postal code column which was later dropped from the final table since it was not of any use for further analysis.

Data pre-processing in my opinion is one of the most time-consuming aspect of any data-science project. It takes a lot of patience and mental thinking to mold the data into a form that you want. If we get the data pre-processing step wrong, we are sure to deviate from our final results, that will lead to a person drawing incorrect conclusions from the results.

3.2-Data Analysis

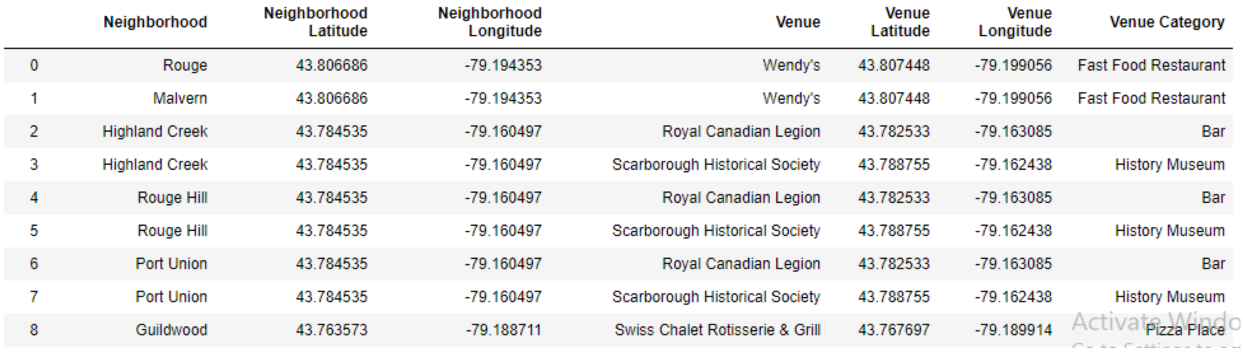
The data analysis phase included two significant tasks that had to be done in order to get answers to our problem statement. The two aspects of our problem statement included.

- Borough Analysis.
- Finding the best possible neighborhood for establishing and Indian restaurant in Toronto city based on the number of Indian restaurants in the vicinity of the chosen spot, i.e. (Choosing a neighborhood with minimum competition).

Firstly, I started with borough analysis. In order to get the data required for the different venues in a particular borough, I used the foursquare api. Foursquare api was linked to my code when the client id, client secret and the version of foursquare api was passed. This meant that I had a connection with foursquare api, and that now I can just call the foursquare api for any venue information required, pertaining to any borough in Toronto city.

Since the project is based on borough-wise analysis. I split the final, clean data into separate data sets where each table will contain data pertaining to only one borough. This was done by retaining the rows that had the borough of our interest associated with it.

After completing the above step, I wrote a function that would call the four square api and access data such as venue name, venue category, venue latitude, venue longitude and later, combine it with the borough table that we extracted in the earlier step. I also dropped the borough column since it is not necessary for our analysis.

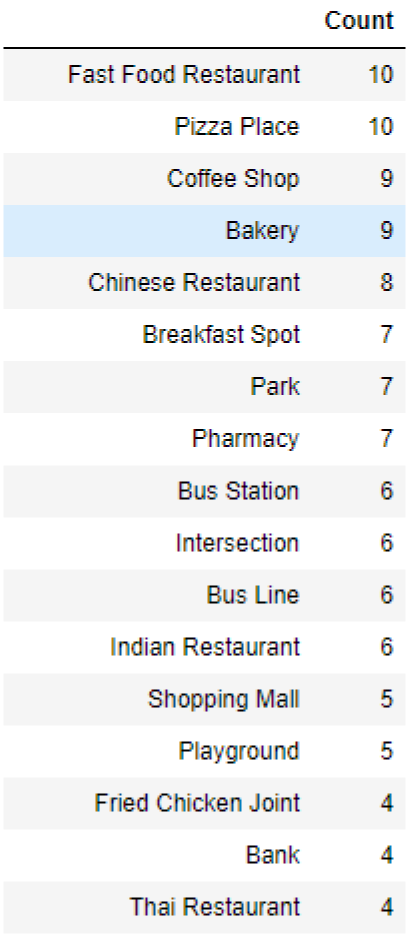


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
2	Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
3	Highland Creek	43.784535	-79.160497	Scarborough Historical Society	43.788755	-79.162438	History Museum
4	Rouge Hill	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
5	Rouge Hill	43.784535	-79.160497	Scarborough Historical Society	43.788755	-79.162438	History Museum
6	Port Union	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
7	Port Union	43.784535	-79.160497	Scarborough Historical Society	43.788755	-79.162438	History Museum
8	Guildwood	43.763573	-79.188711	Swiss Chalet Rotisserie & Grill	43.767697	-79.189914	Pizza Place

Fig 3: Data set with venue details

I then moved on to grouping our venues based on the venue categories. The occurrence of a venue category will be calculated from the result set that we extracted in the previous step and a new data frame will be created that has only the venue category and their respective counts. This gives us a good idea about the different category of venues present in that borough along with its frequency. Moreover, this will also help tourists, choose places to visit in Toronto city based on the sole factor that whether their place of interest is present in a particular borough and what is the frequency with which their venue of interest appears in that borough. This process is repeated until we have the results for every borough in Toronto city

A sample data set for a particular borough containing the venue category and their respective counts is shown below.

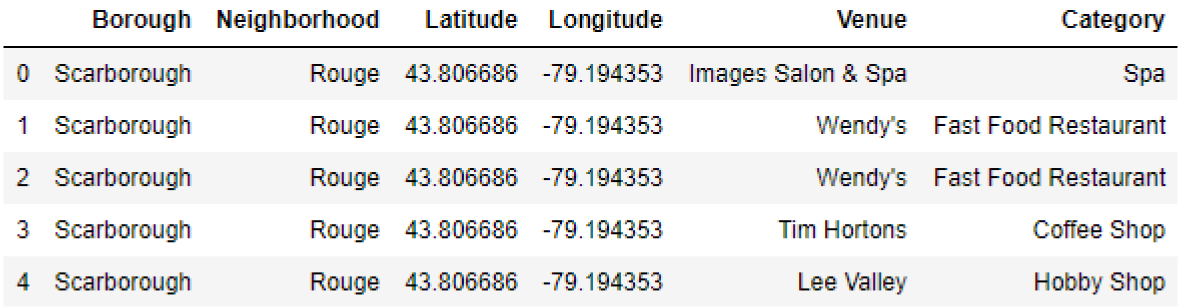


	Count
Fast Food Restaurant	10
Pizza Place	10
Coffee Shop	9
Bakery	9
Chinese Restaurant	8
Breakfast Spot	7
Park	7
Pharmacy	7
Bus Station	6
Intersection	6
Bus Line	6
Indian Restaurant	6
Shopping Mall	5
Playground	5
Fried Chicken Joint	4
Bank	4
Thai Restaurant	4

Fig 4: Data set with venue count.

As the next step of my capstone project. I started with the analysis of the whole Toronto city with the aim of finding neighborhoods and boroughs that could be best suited for establishing an Indian restaurant. Since Canada is a country having a considerable number of Indians. Opening an Indian restaurant is not a bad idea. But picking the right location is an important factor if a person expects a sustained income from his/her business. Hence in this part of the project, we will display a map showing markers that will depict neighborhoods that already have Indian Restaurants along with its frequency. The main idea behind this solution is to avoid places that already have a high density of Indian restaurants present. It is advisable to establish a business at a place where it will face the least competition. Hence selecting a neighborhood that has no Indian restaurants will help the new business, flourish unopposed.

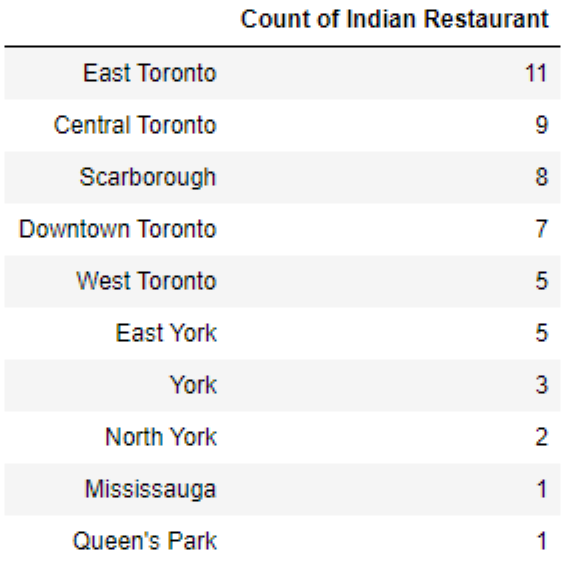
The process flow followed for solving this problem statement is similar to the one followed for borough analysis. We again used the same function that was used for borough analysis to call the foursquare api. But instead of passing the data frame which was segregated according to individual boroughs, we passed the data frame containing information about all boroughs and neighborhoods in Toronto city. This gave us a large data frame containing information about almost all venues in Toronto city along with the categories of those venues.



	Borough	Neighborhood	Latitude	Longitude	Venue	Category
0	Scarborough	Rouge	43.806686	-79.194353	Images Salon & Spa	Spa
1	Scarborough	Rouge	43.806686	-79.194353	Wendy's	Fast Food Restaurant
2	Scarborough	Rouge	43.806686	-79.194353	Wendy's	Fast Food Restaurant
3	Scarborough	Rouge	43.806686	-79.194353	Tim Hortons	Coffee Shop
4	Scarborough	Rouge	43.806686	-79.194353	Lee Valley	Hobby Shop

Fig 5: Data set containing list of venues in Toronto city.

We then extract rows having information about Indian Restaurants and discard the rest of the entries in the data frame since they are of no use to us. The data set that we have in hand now is a concise data set with information important to our project. We then create a new data frame that consists a count of Indian Restaurants in each borough. The data frame is depicted in the image below.



	Count of Indian Restaurant
East Toronto	11
Central Toronto	9
Scarborough	8
Downtown Toronto	7
West Toronto	5
East York	5
York	3
North York	2
Mississauga	1
Queen's Park	1

Fig 6: Count of Indian restaurants in each borough

This data is then plotted on a bar chart for ease of understanding and to also grab the attention of viewers who are reading this document.

Moreover, I also planned on depicting the location of Indian Restaurants on the map of Toronto. The map will have a marker at the position of the restaurant, this depicts that an Indian restaurant is present in that neighborhood. Additionally, when a person clicks on the marker, a label pops up depicting the number of Indian restaurants in that neighborhood. This will give a very good idea to viewers about the location of Indian restaurants by looking on the map, at the same time it will depict the count of Indian restaurants in a particular neighborhood. A person who plans to open an Indian restaurant will avoid places that have a large number of Indian restaurants and will look for places that pose minimum competition. This can be done by viewing the map and studying the location of Indian restaurants carefully.

Given below are two images, the first depicting the number of Indian restaurants in each borough, whereas the second image depicting the location of Indian restaurants with the help of a map of Toronto city.

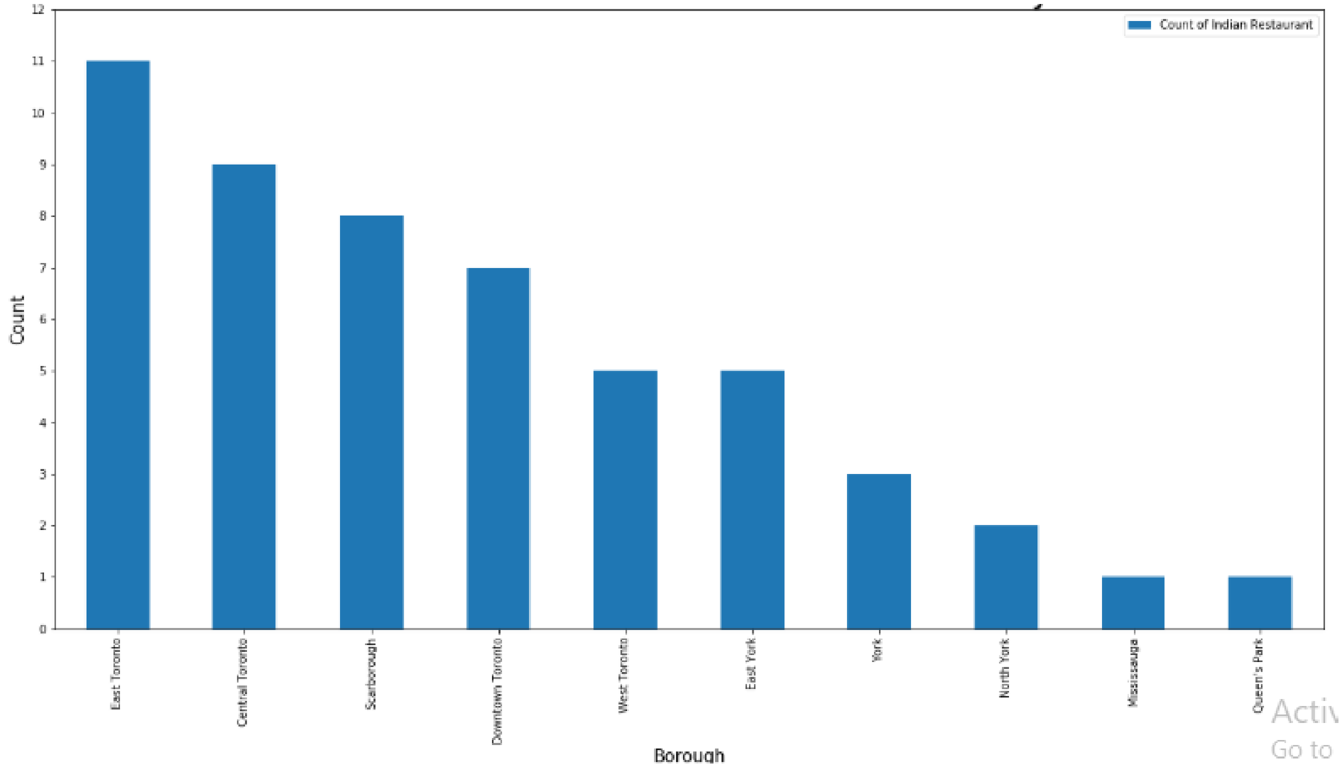


Fig 7: A bar chart depicting count of Indian restaurants in each borough

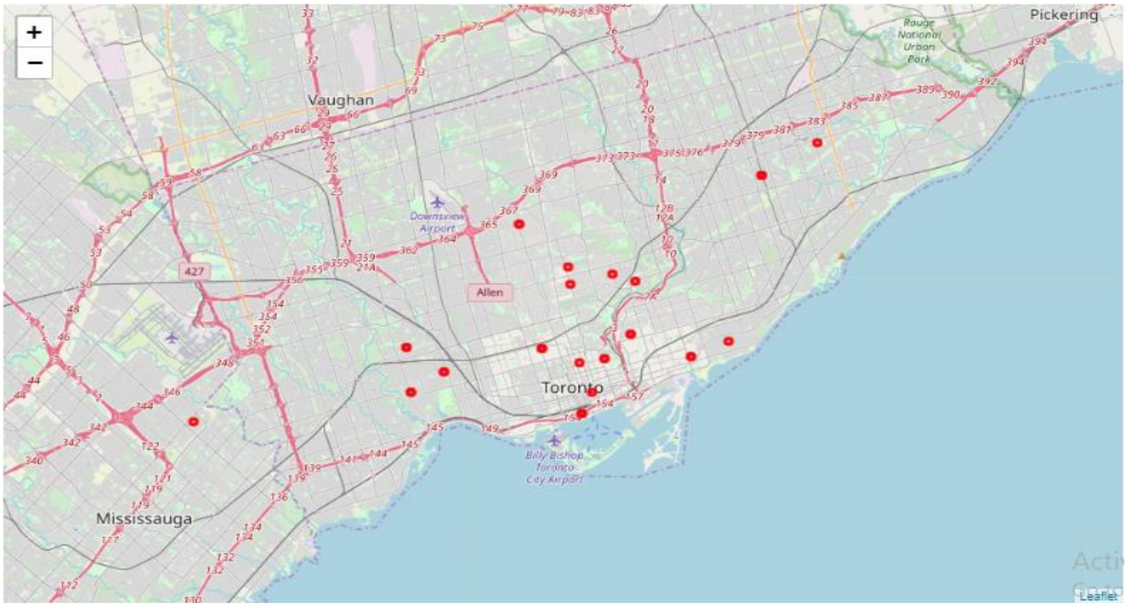


Fig 8: Map of Toronto portraying the location of Indian restaurants in Toronto city

Chapter-4

In this section we will discuss the results that we acquired, after implementing the various data science methodologies. We will primarily discuss about the analysis we have done pertaining to each borough, since the analysis that we did on the Toronto city data set was discussed in the previous section.

The results that we acquired after a thorough data analysis stage have been depicted in the form of a bar-graph for the ease of understanding. Given below are the bar charts containing borough wise data analysis for different venues that each borough has along with its frequency.

The list of different boroughs that make up Toronto city are given below:

- Scarborough
- East York
- North York
- York
- Downtown Toronto
- West Toronto
- East Toronto
- Central Toronto
- Etobicoke
- Mississauga
- Queen’s Park

4.1-Scarborough

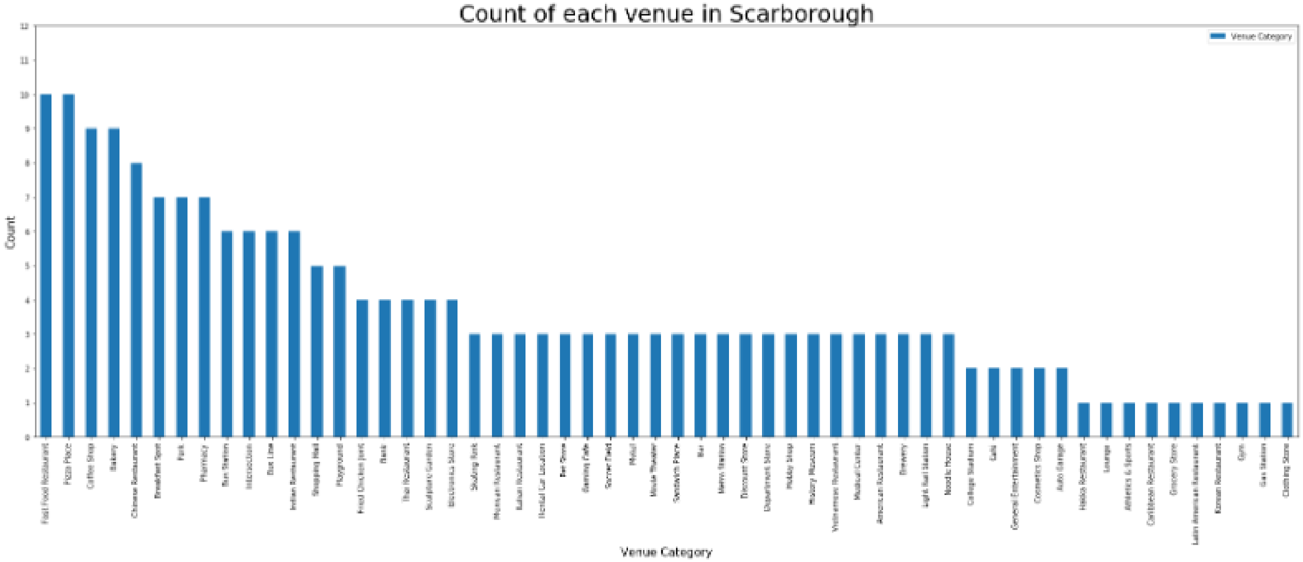


Fig 9: Bar chart for Scarborough

4.2- North York

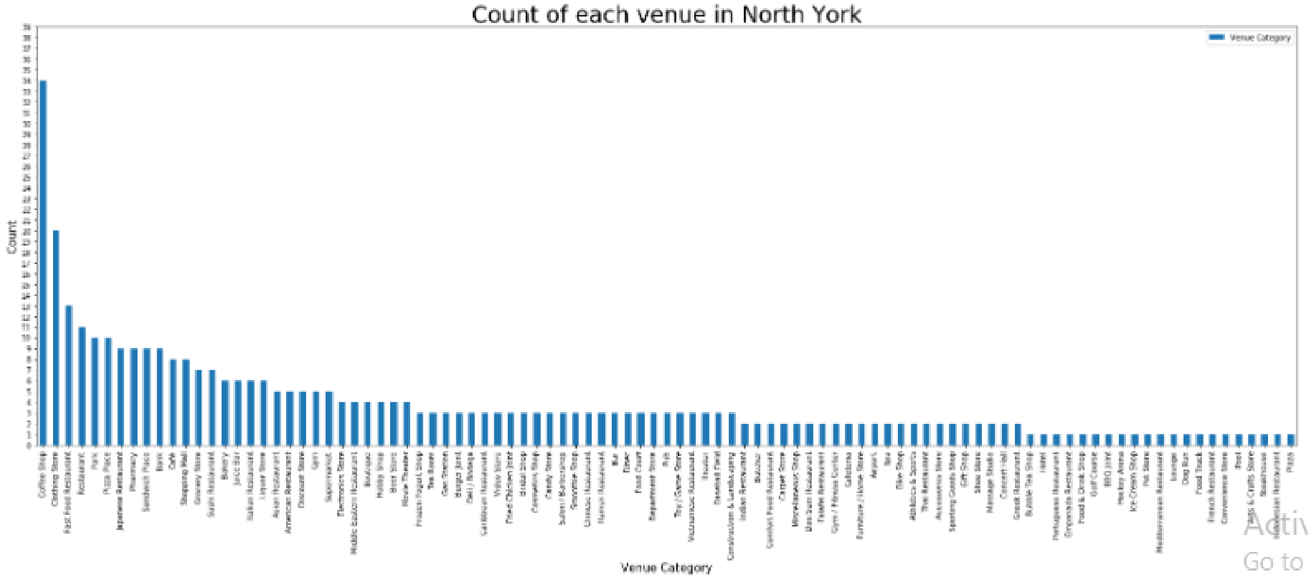


Fig 10: Bar chart for North York

4.3-East York

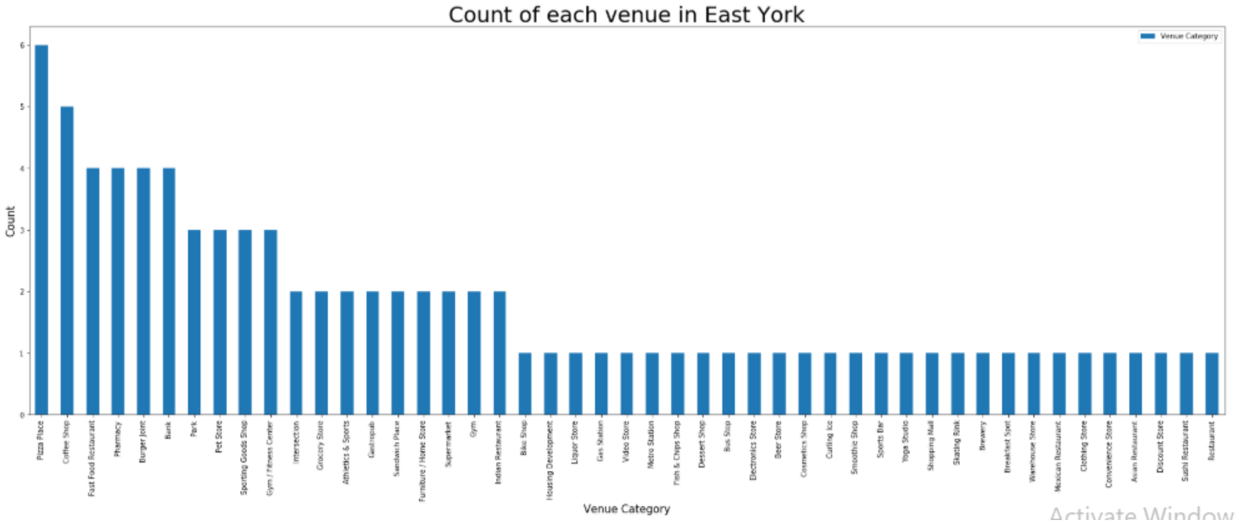


Fig 11: Bar chart for East York

4.4-York

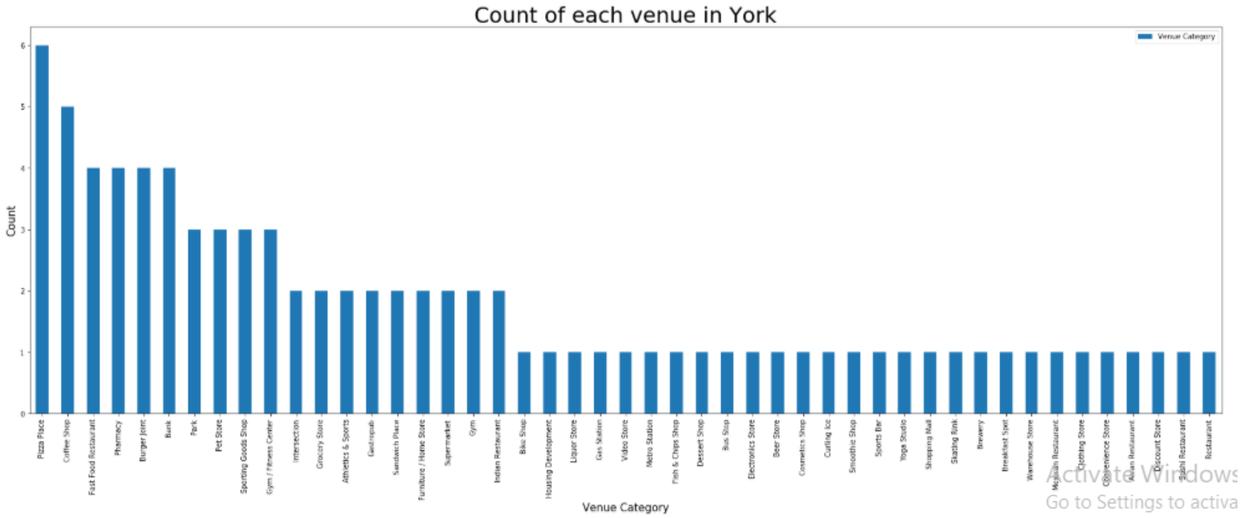


Fig 12: Bar chart for York

4.5-East Toronto

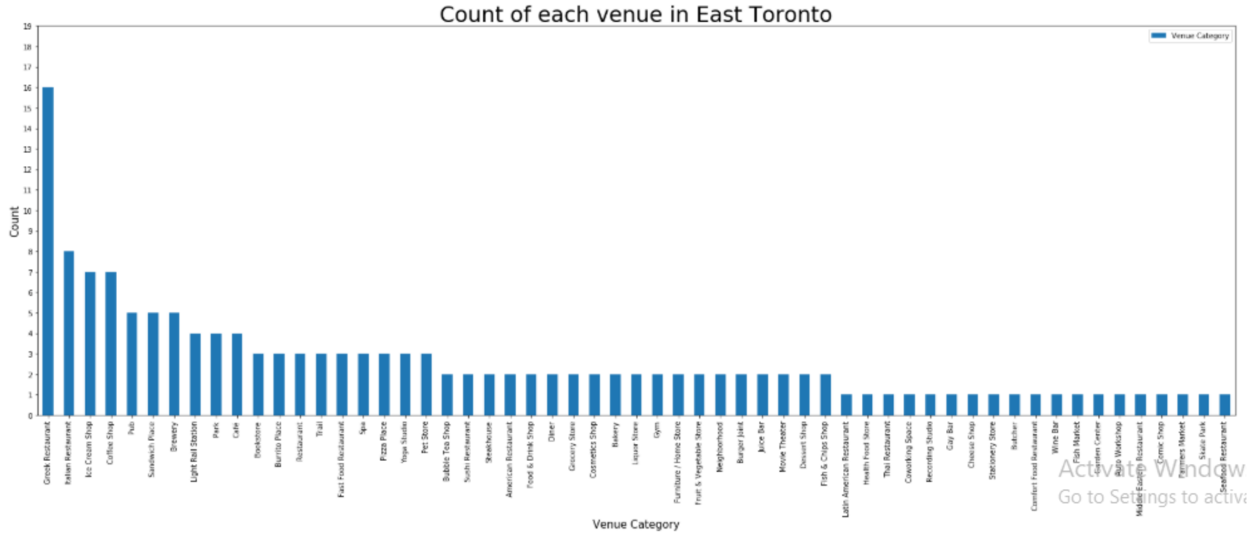


Fig 13: Bar chart for East Toronto

4.6-Central Toronto

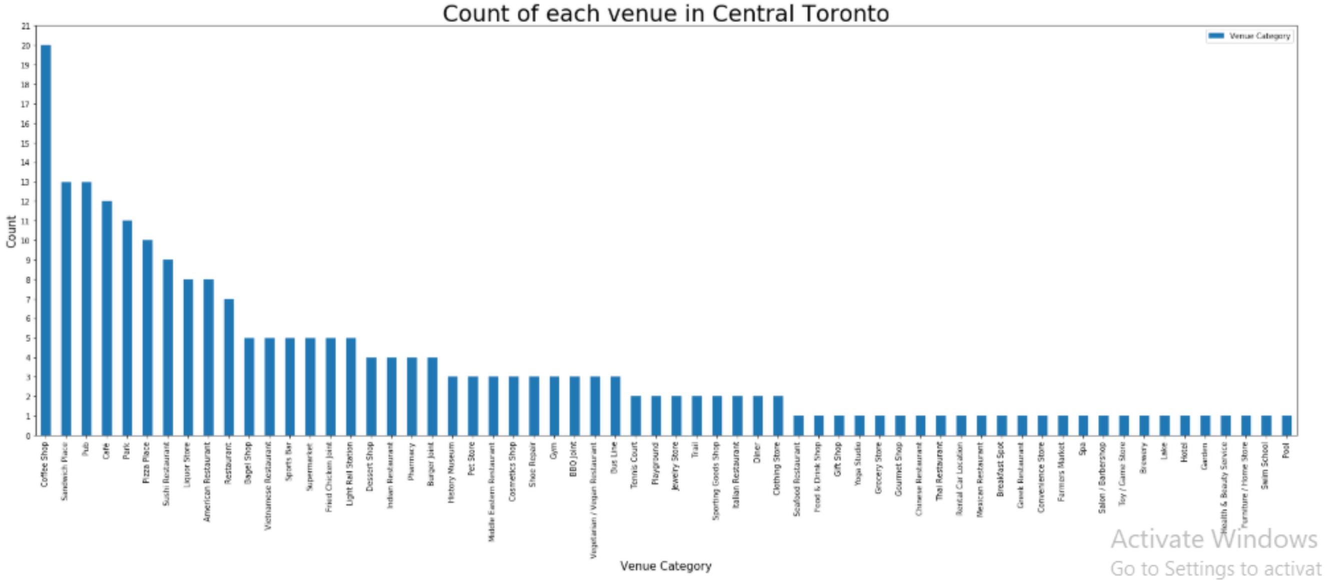


Fig 14: Bar chart for Central Toronto

4.7-Downtown Toronto

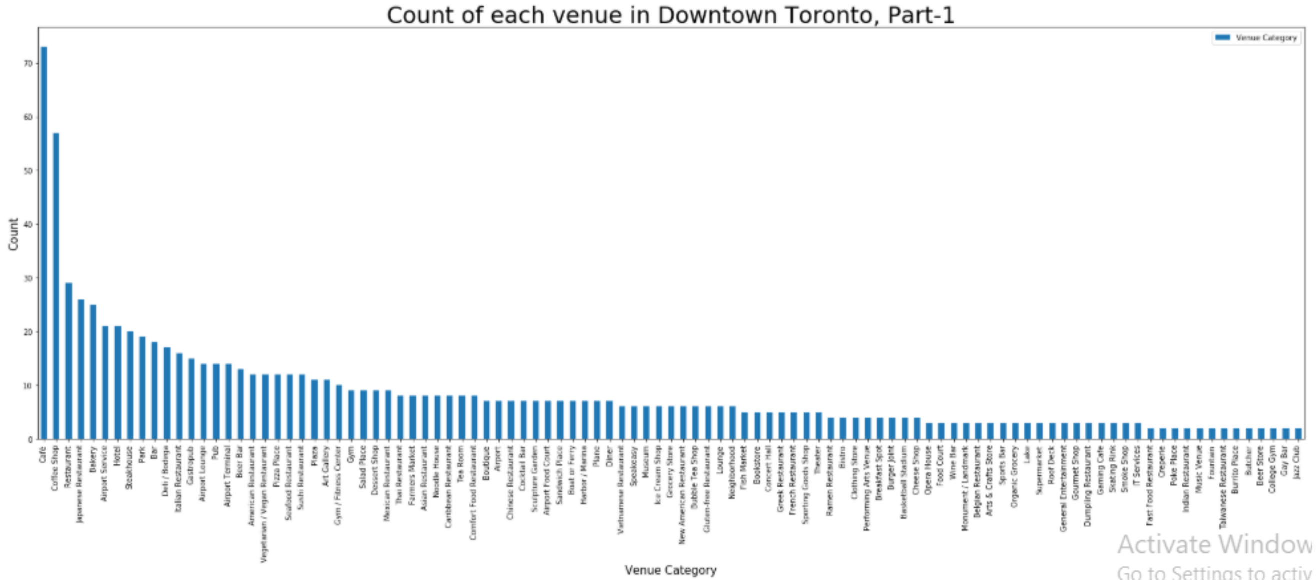


Fig 15: Bar chart for Downtown Toronto (Part-1)

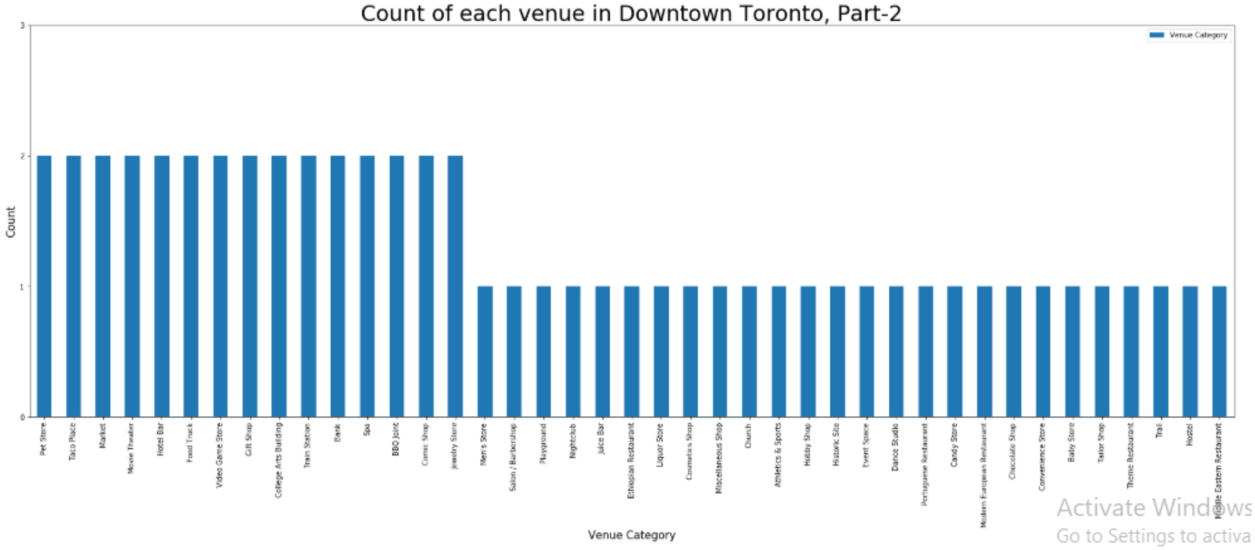


Fig 16: Bar chart for Downtown Toronto (Part-2)

4.8-West Toronto

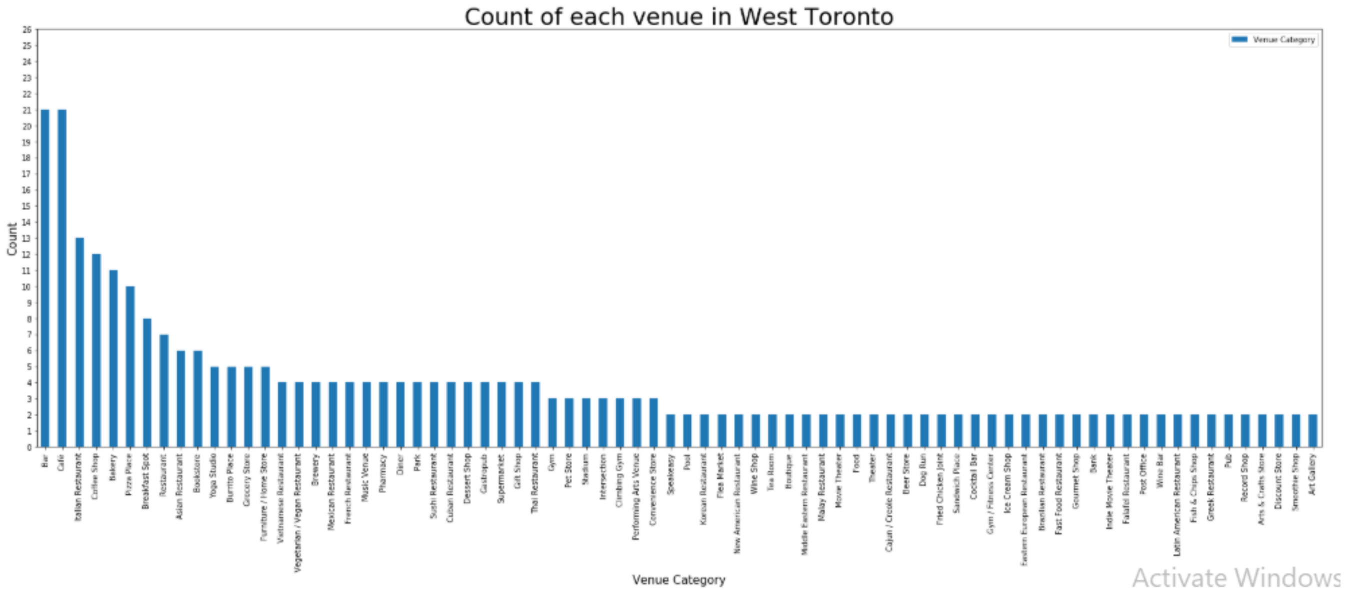


Fig 17: Bar chart for West Toronto

4.9-Etobicoke

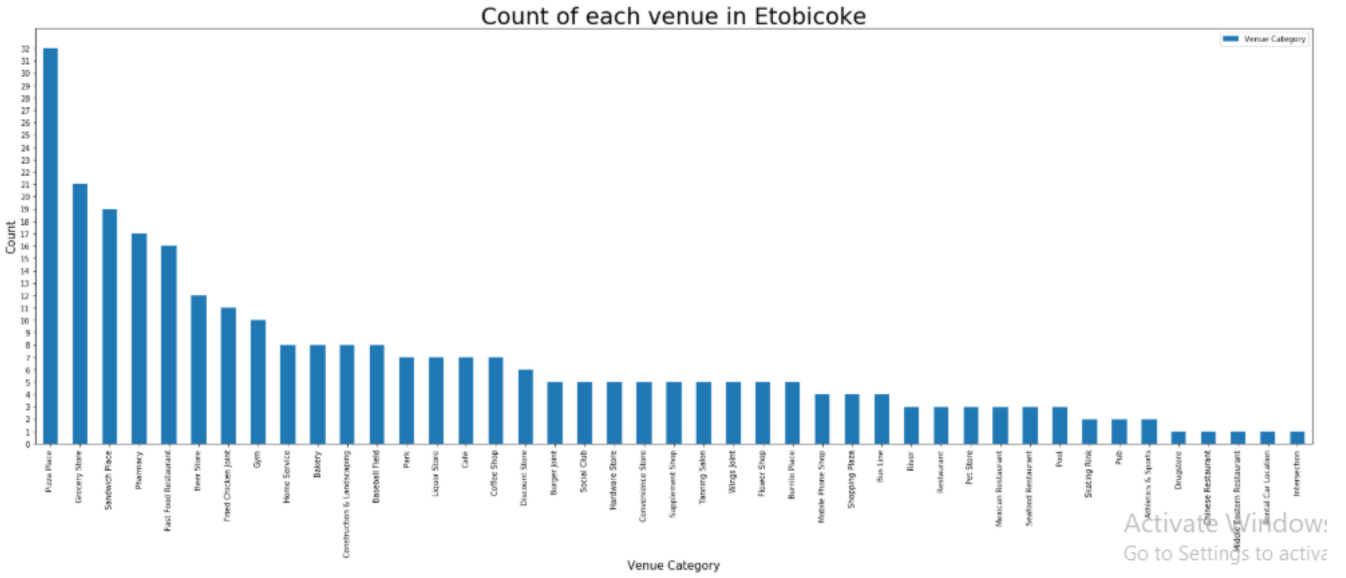
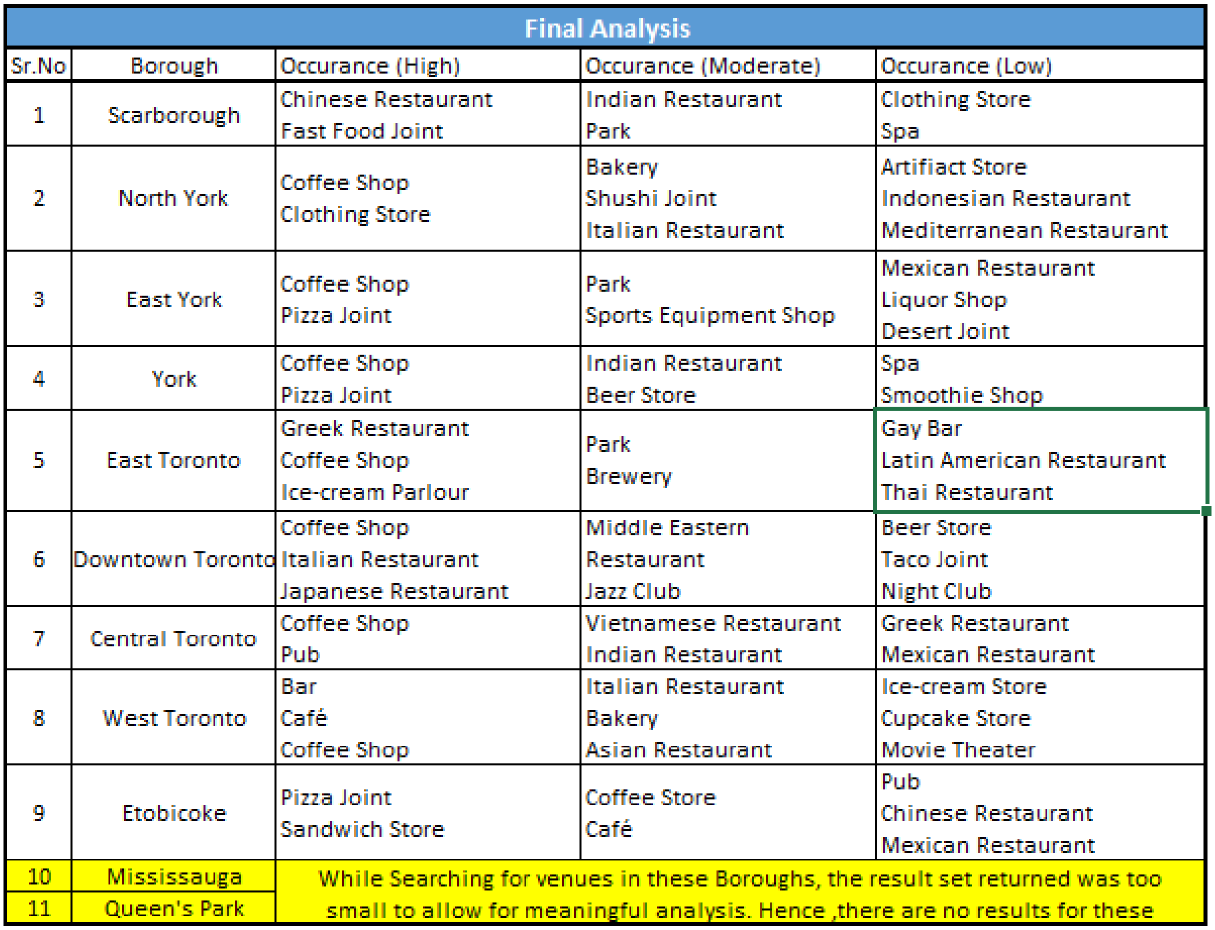


Fig 18: Bar chart for Etobicoke

The results that we acquired for Mississauga and Queen's Park were too small to perform meaningful analysis on. Hence, there are no bar charts for the aforementioned boroughs.

The table presented below provides readers with brief information about the different venue categories in a borough based on the frequency of their occurrence in data returned by four square api.



Final Analysis

Sr.No	Borough	Occurance (High)	Occurance (Moderate)	Occurance (Low)
1	Scarborough	Chinese Restaurant Fast Food Joint	Indian Restaurant Park	Clothing Store Spa
2	North York	Coffee Shop Clothing Store	Bakery Shushi Joint Italian Restaurant	Artifiact Store Indonesian Restaurant Mediterranean Restaurant
3	East York	Coffee Shop Pizza Joint	Park Sports Equipment Shop	Mexican Restaurant Liquor Shop Desert Joint
4	York	Coffee Shop Pizza Joint	Indian Restaurant Beer Store	Spa Smoothie Shop
5	East Toronto	Greek Restaurant Coffee Shop Ice-cream Parlour	Park Brewery	Gay Bar Latin American Restaurant Thai Restaurant
6	Downtown Toronto	Coffee Shop Italian Restaurant Japanese Restaurant	Middle Eastern Restaurant Jazz Club	Beer Store Taco Joint Night Club
7	Central Toronto	Coffee Shop Pub	Vietnamese Restaurant Indian Restaurant	Greek Restaurant Mexican Restaurant
8	West Toronto	Bar Café Coffee Shop	Italian Restaurant Bakery Asian Restaurant	Ice-cream Store Cupcake Store Movie Theater
9	Etobicoke	Pizza Joint Sandwich Store	Coffee Store Café	Pub Chinese Restaurant Mexican Restaurant
10	Mississauga	While Searching for venues in these Boroughs, the result set returned was too small to allow for meaningful analysis. Hence ,there are no results for these		
11	Queen's Park			

Fig 19: Overall Analysis of Boroughs in Toronto City

Chapter-5

During the course of this project, there have been a few assumptions and considerations that were taken into account so as to help in the successful completion of this project. While going through all of my implementation steps, the below given instances were taken into consideration and should be kept in mind while glancing through the process steps.

1. During the borough analysis phase, radius passed during the foursquare api call was considered to be 500 meters.
2. During the second phase of the project, which included finding an optimal place for setting up an Indian restaurant, radius passed during the foursquare api call was considered to be 700 meters. The change was necessary since taking a smaller radius resulted in missing out some neighborhoods.
3. The results returned by foursquare api can vary on a day to day basis, hence the number of venues returned by foursquare api can have slight changes.
4. While determining the best location for setting up an Indian restaurant, only one factor was taken into consideration, i.e. to prevent opening an Indian restaurant in a neighborhood that already houses considerable number of Indian restaurants (i.e. Minimum competition).
5. Two boroughs, Mississauga and Queen's Park were eliminated from our analysis since the search for venues in the aforementioned boroughs did not return considerable amount of results to work on.

Chapter-6

During the course of this capstone project, I was able to apply different data science techniques and tools that I learned in the IBM Data Science course. This helped me unearth meaningful insights from the data analysis that I did on the Toronto data set. The aspects I uncovered during the phase of data analysis are listed below.

- Borough analysis
    - Coffee shops are a venue that has a very high rate of occurrence in almost all the boroughs.
    - Mississauga and Queen's Park have very few venues to go to or choose from if you are a tourist.
    - Parks are the next venue that have the most occurrences amongst the different boroughs.
    - Downtown Toronto have the maximum and the most varied choices of venues to choose from for a tourist.
- Toronto city analysis (for establishing a new Indian restaurant)
    - East Toronto and Central Toronto are the two boroughs with the maximum number of Indian Restaurants.
    - India Bazaar, Thorncliffe Park and The Beaches West are the neighborhoods with the maximum number of Indian restaurants.
    - North York, Queen's Park and Mississauga are locations ideal for opening a new Indian restaurant based on our ideology that the new business will face minimum competition there.