IBM Data Science Capstone Project Report

Exploring the Best Places to Establish a Chinese Restaurant in Toronto

June 2020

Table of Contents

Sr#.	Chapters	Page No.
1.	Introduction	3
2.	Data Section	4
	2.1. Data Pre-Processing	4
3.	Methodology	6
	3.1. Data Analysis	6
	3.2 Data Aggregation	7
4.	Results Section	11
5.	Discussion Section	17
6.	Conclusion	19

1. Introduction

Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario's northwestern shore. Toronto also has many green spaces, from the orderly oval of Queen's Park to 400-acre High Park and its trails, sports facilities and zoo.

Toronto is a city with a high population and population density, as per Wikipedia Page about **631,050** Chinese lives in Toronto. In order to open a Chinese restaurant in Toronto we first need to identify places or venues which will be more suitable for us to open this restaurant to make it a profitable business.

If we think of above problem we can start from finding the best possible neighborhood for establishing a Chinese restaurant in Toronto city based on the number of Chinese restaurants in the vicinity of the chosen spot, i.e. choosing a neighborhood with minimum competition and coming up with a few suggestive neighbourhoods that have business potential in terms of opening a new Chinese restaurant.

When an invester dream about doing investments into opening Restaurants, one of the important question is try to find best possible place or area with one of the least competition.

In this project we will locate such places which are ideal for opening a Chinese Restaurant.

This project will help two types of target audiences, first Individuals/Investers who are trying to establish new Restaurant business by finding areas which has low number of Chinese restaurants and secondly tourists e.g. Asian tourists or people who like Chinese food to help them choosing neighborhoods with easy accessibility.

2. Data Section

The required data set can be acquired from different data sources. The three data sources are listed below:

- Wikipedia to fetch boroughs and neighborhoods of Toronto.
- A .csv file https://cocl.us/Geospatial_data to fetch latitudes and longitudes corresponding to each postal code.
- The Foursquare API to fetch different public venues in the vicinity of the neighborhood.

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 with the **FourSquare 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.

2.1 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. Methodology

In this section we will explore methodologies we will use in this project, i.e. data analysis and statistical and machine learning approaches we will use on our above data.

3.1. 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 Chinese restaurant in Toronto city based on the number of Chinese 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.

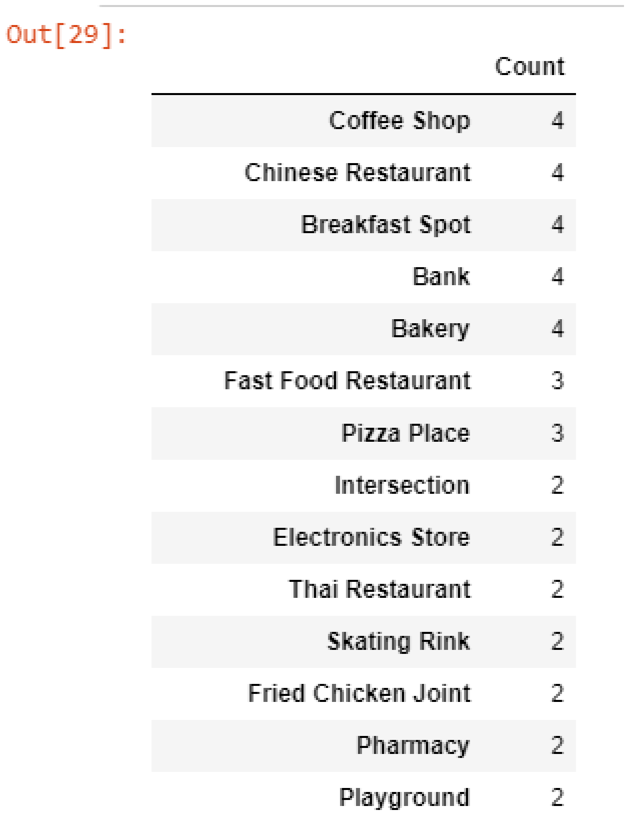
Since the project is based on borough-wise analysis. I split the final, clean data into separate data sets where each table will contain data pertaining to only one borough. This was done by retaining the rows that had the borough of our interest associated with it.

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.

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.

3.2. Data Aggregation

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



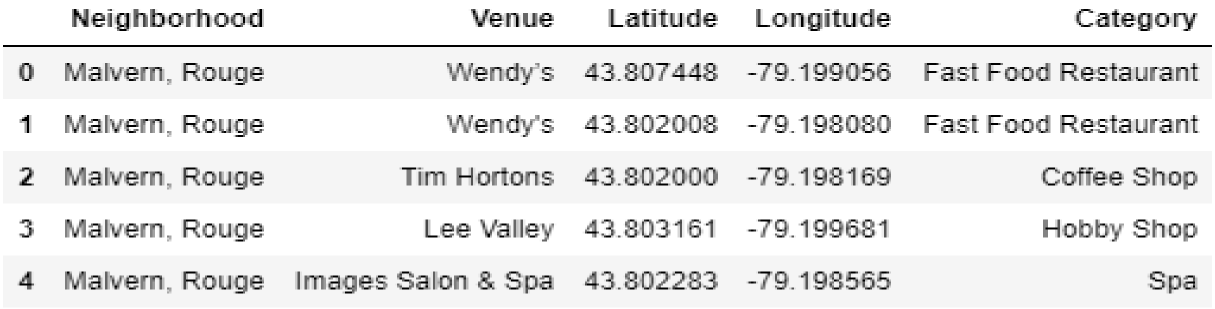
Out[29]:

	Count
Coffee Shop	4
Chinese Restaurant	4
Breakfast Spot	4
Bank	4
Bakery	4
Fast Food Restaurant	3
Pizza Place	3
Intersection	2
Electronics Store	2
Thai Restaurant	2
Skating Rink	2
Fried Chicken Joint	2
Pharmacy	2
Playground	2

Figure 1: Count of each venue

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 a Chinese restaurant. Since Canada is a country having a considerable number of Chinese. Opening a Chinese 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 Chinese Restaurants along with its frequency. The main idea behind this solution is to avoid places that already have a high density of Chinese 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 Chinese restaurants will help the new business, flourish unopposed.

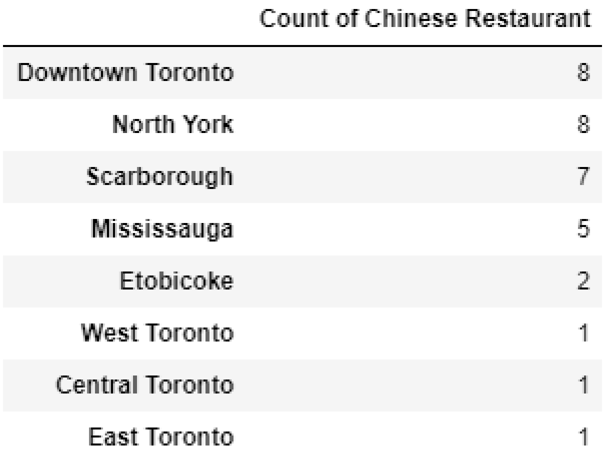
The process flow followed for solving this problem statement is similar to the one followed for borough analysis. We again used the same function that was used for borough analysis to call the foursquare api. But instead of passing the data frame which was segregated according to individual boroughs, 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.



	Neighborhood	Venue	Latitude	Longitude	Category
0	Malvern, Rouge	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Malvern, Rouge	Wendy's	43.802008	-79.198080	Fast Food Restaurant
2	Malvern, Rouge	Tim Hortons	43.802000	-79.198169	Coffee Shop
3	Malvern, Rouge	Lee Valley	43.803161	-79.199681	Hobby Shop
4	Malvern, Rouge	Images Salon & Spa	43.802283	-79.198565	Spa

Figure 2: Data set containing list of venues in Toronto city.

We then extract rows having information about Chinese 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 Chinese Restaurants in each borough. The data frame is depicted in the image below.



	Count of Chinese Restaurant
Downtown Toronto	8
North York	8
Scarborough	7
Mississauga	5
Etobicoke	2
West Toronto	1
Central Toronto	1
East Toronto	1

Figure 3: Count of Chinese 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 Chinese Restaurants on the map of Toronto. The map will have a marker at the position of the restaurant, this depicts that a Chinese restaurant is present in that neighborhood. Additionally, when a person clicks on the marker, a label pops up depicting the number of Chinese restaurants in that neighborhood. This will give a very good idea to viewers about the location of Chinese restaurants by looking on the map, at the same time it will depict the count of Chinese restaurants in a particular neighborhood. A person who plans to open a Chinese restaurant will avoid places that have a large number of Chinese restaurants and will look for places that pose minimum competition. This can be done by viewing the map and studying the location of Chinese restaurants carefully.

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

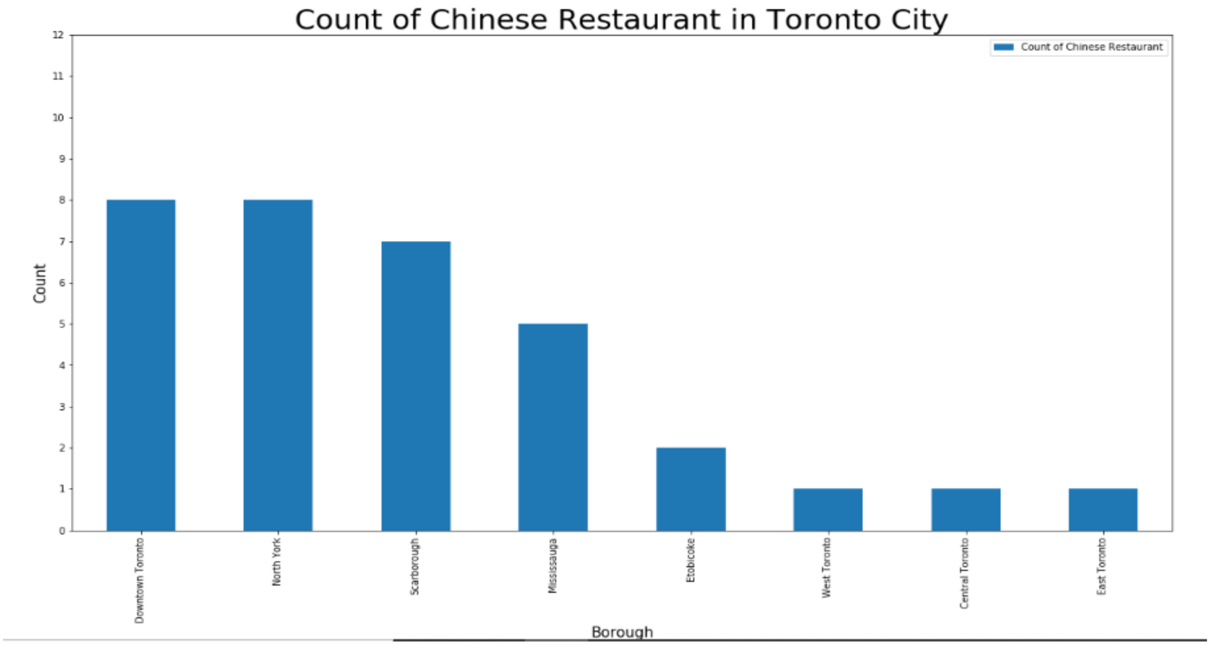


Figure 4: A bar chart depicting count of Chinese restaurants in each borough

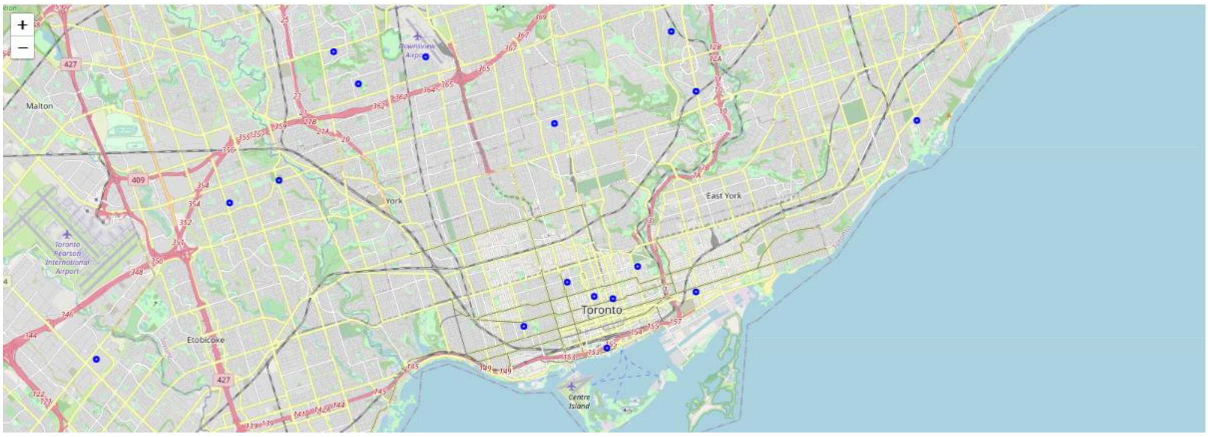


Figure 5: Map of Toronto portraying the location of Chinese restaurants in Toronto city

4. Results Section

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

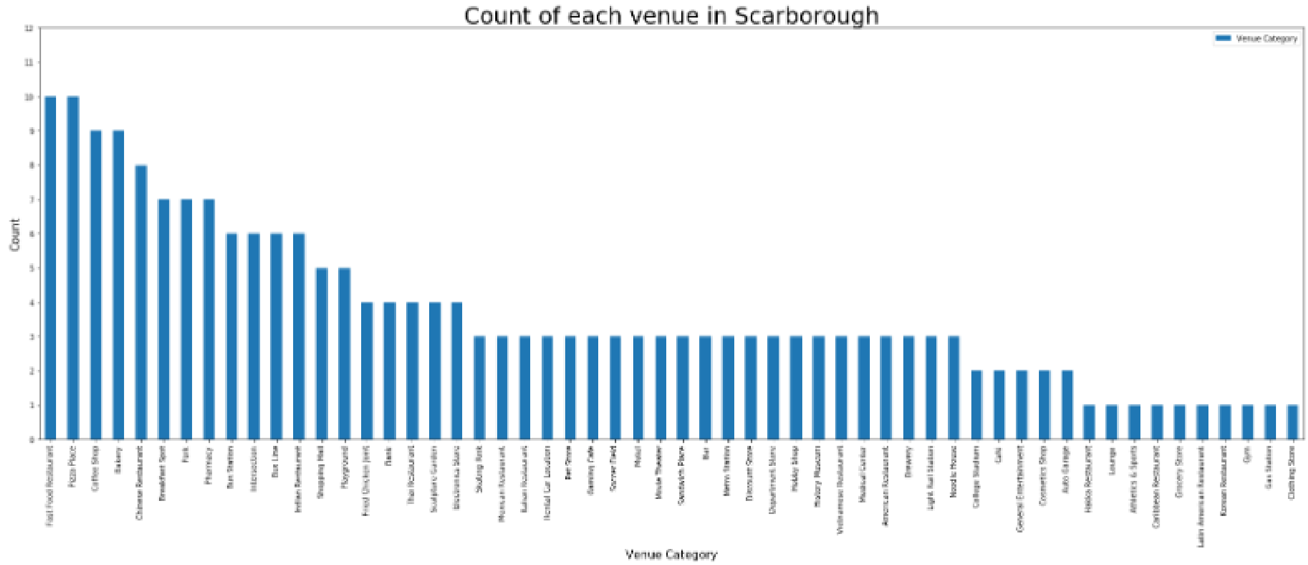
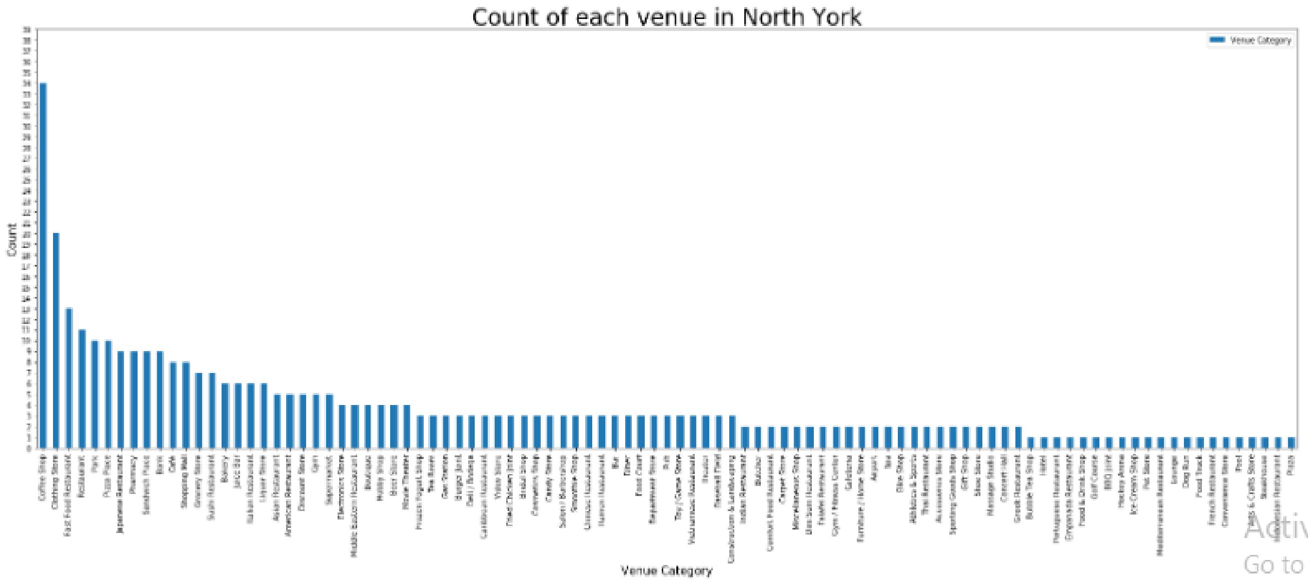


Figure 6: Bar chart for Scarborough

4.2- North York



Activ
Go to

Figure 7: Bar chart for North York

4.3-East York

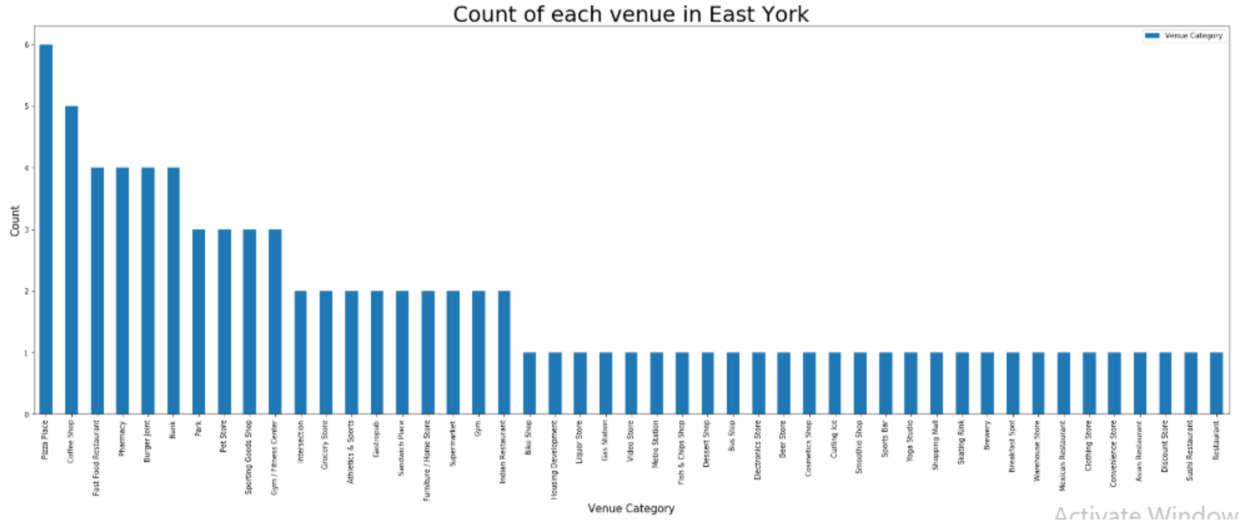


Figure 8: Bar chart for East York

4.4-York

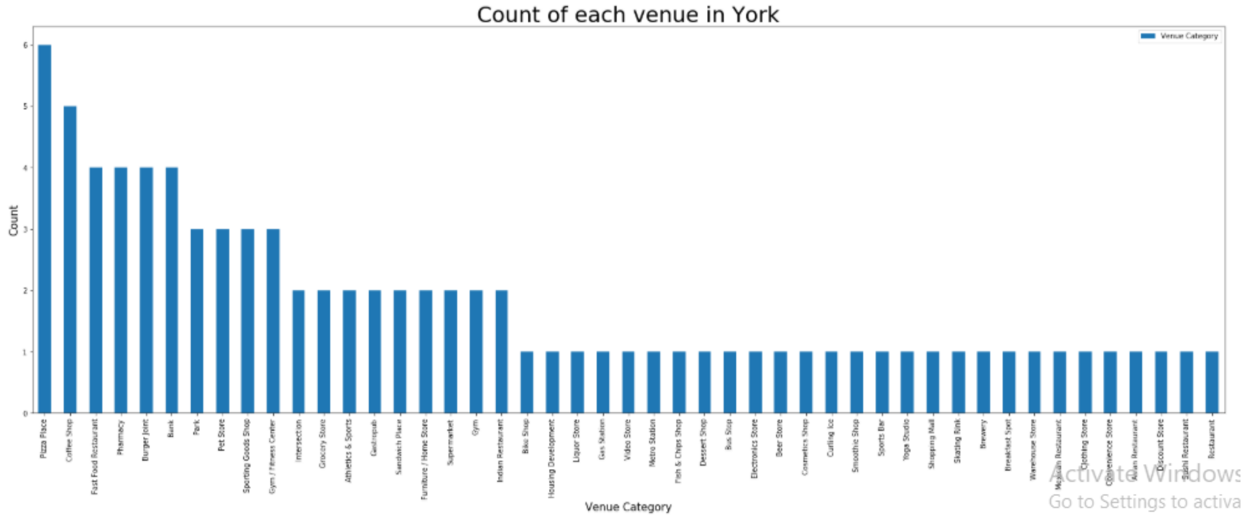
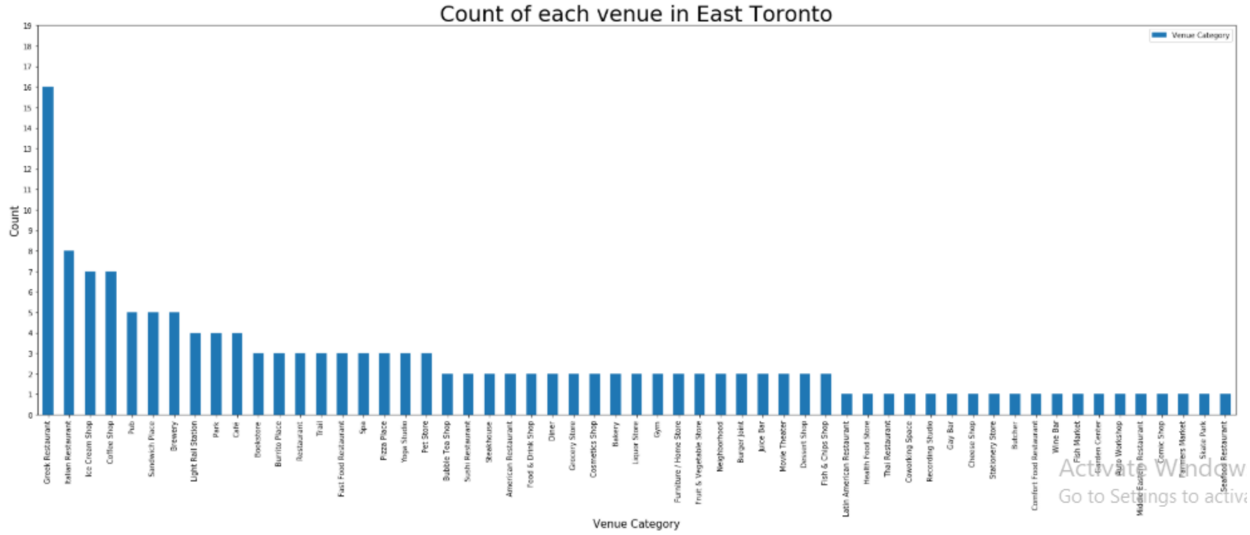


Figure 9: Bar chart for York

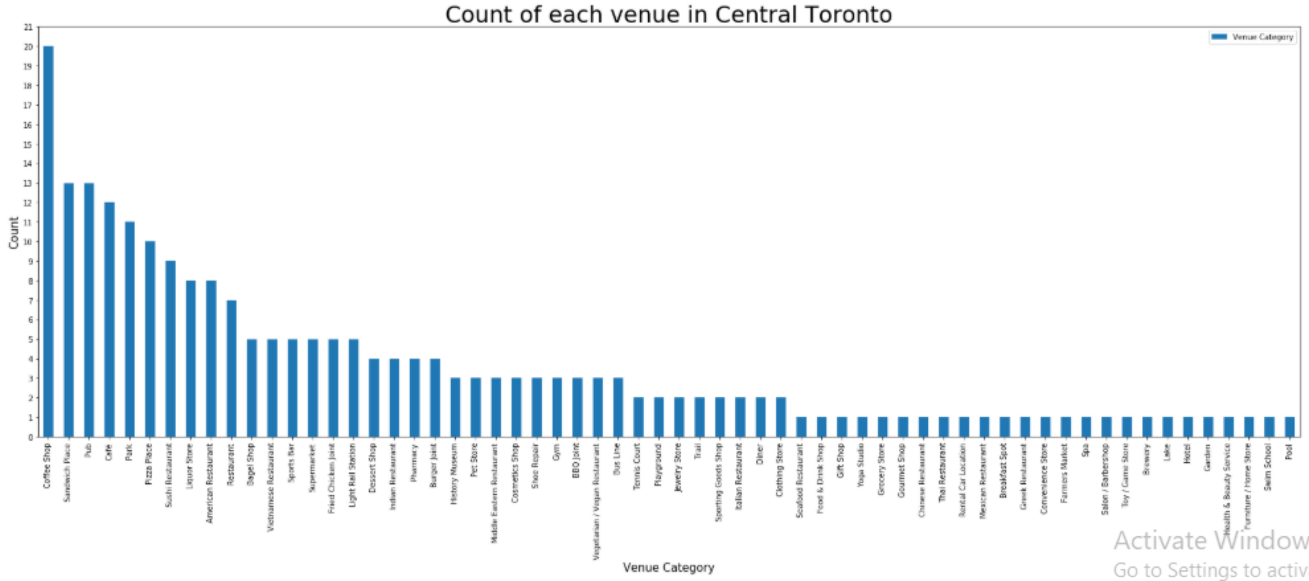
4.5-East Toronto



Activate Windows
Go to Settings to activa

Figure 10: Bar chart for East Toronto

4.6-Central Toronto



Activate Windows
Go to Settings to activat

Figure 11: Bar chart for Central Toronto

4.7-Downtown Toronto

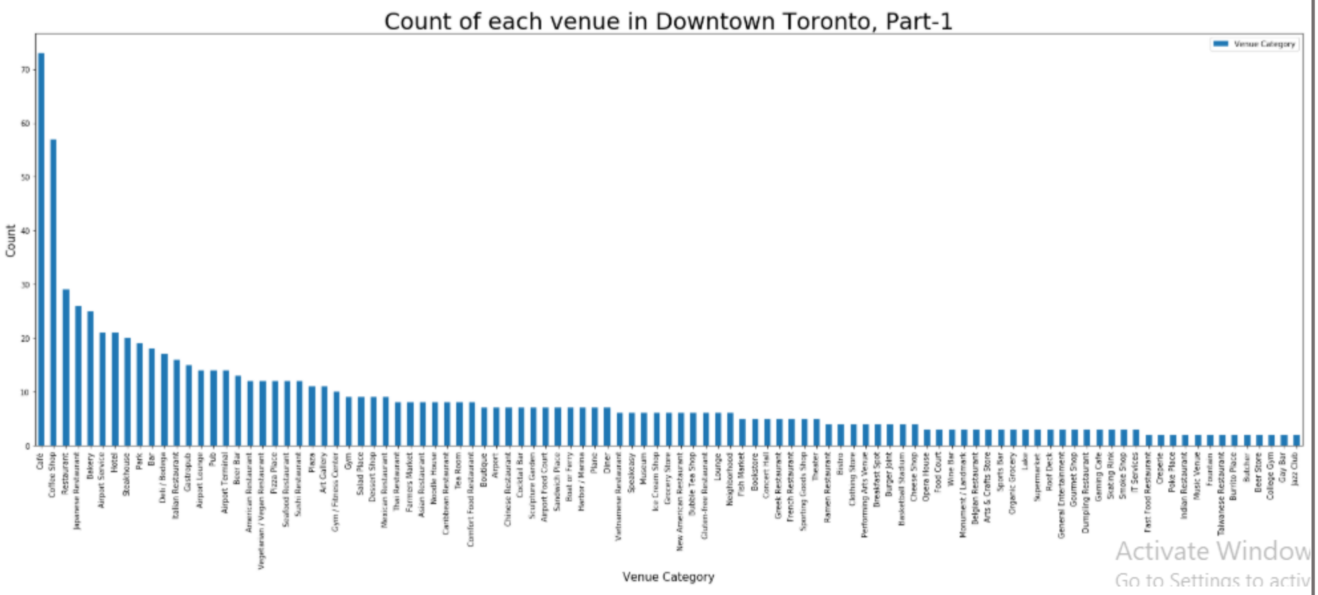


Figure 12: Bar chart for Downtown Toronto (Part-1)

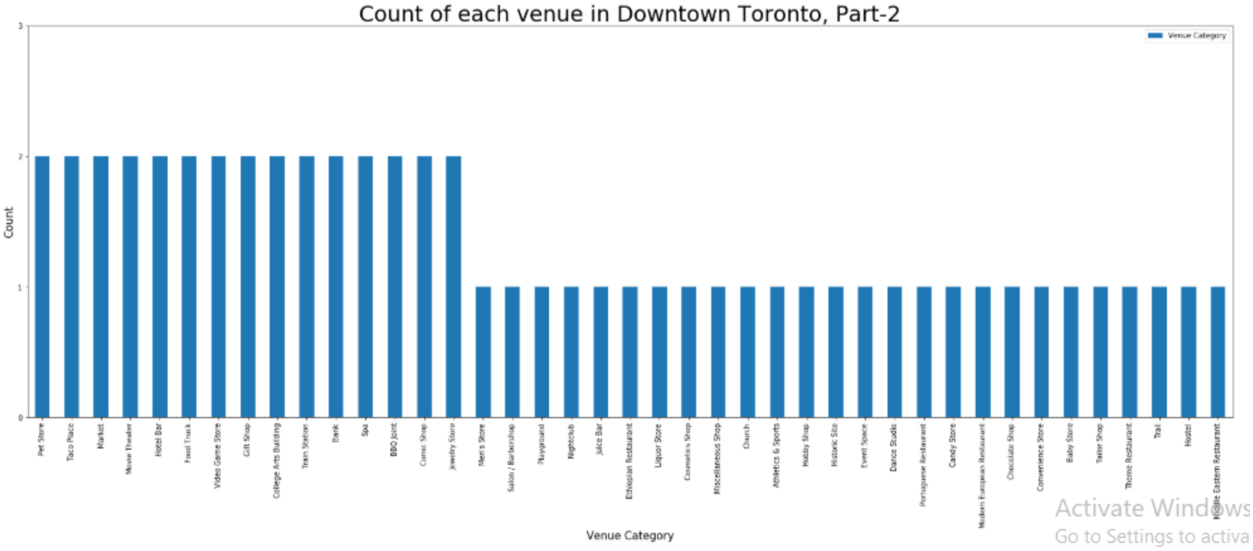


Figure 13: Bar chart for Downtown Toronto (Part-2)

4.8-West Toronto

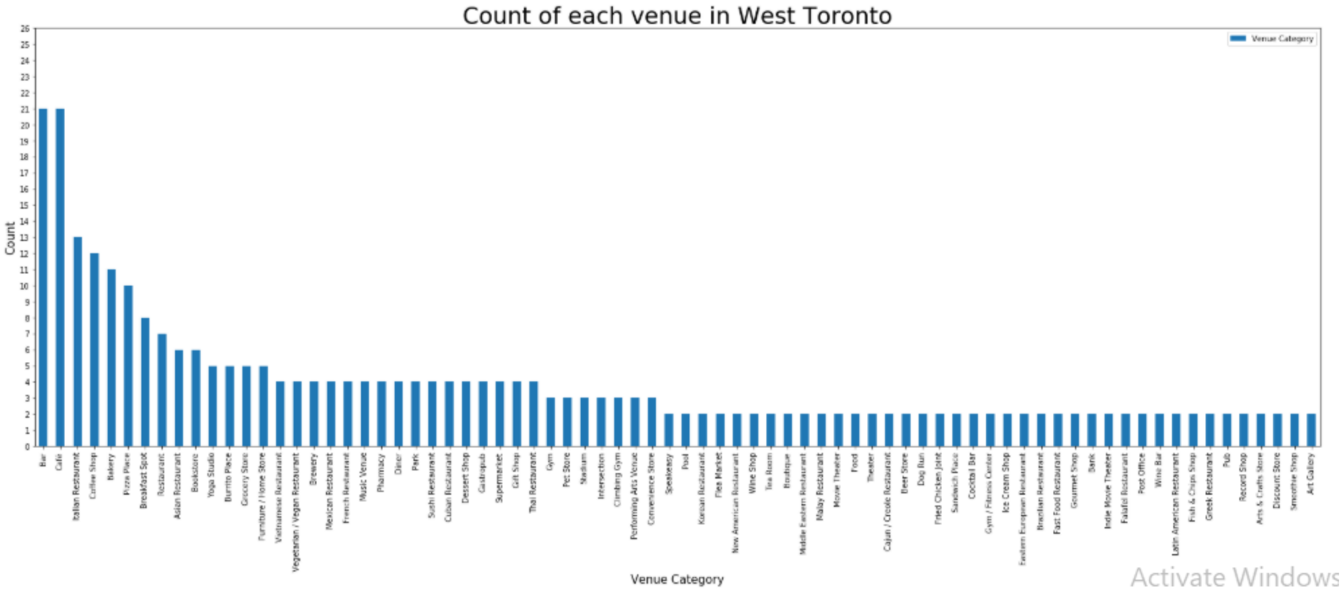


Figure 14: Bar chart for West Toronto

4.9-Etobicoke

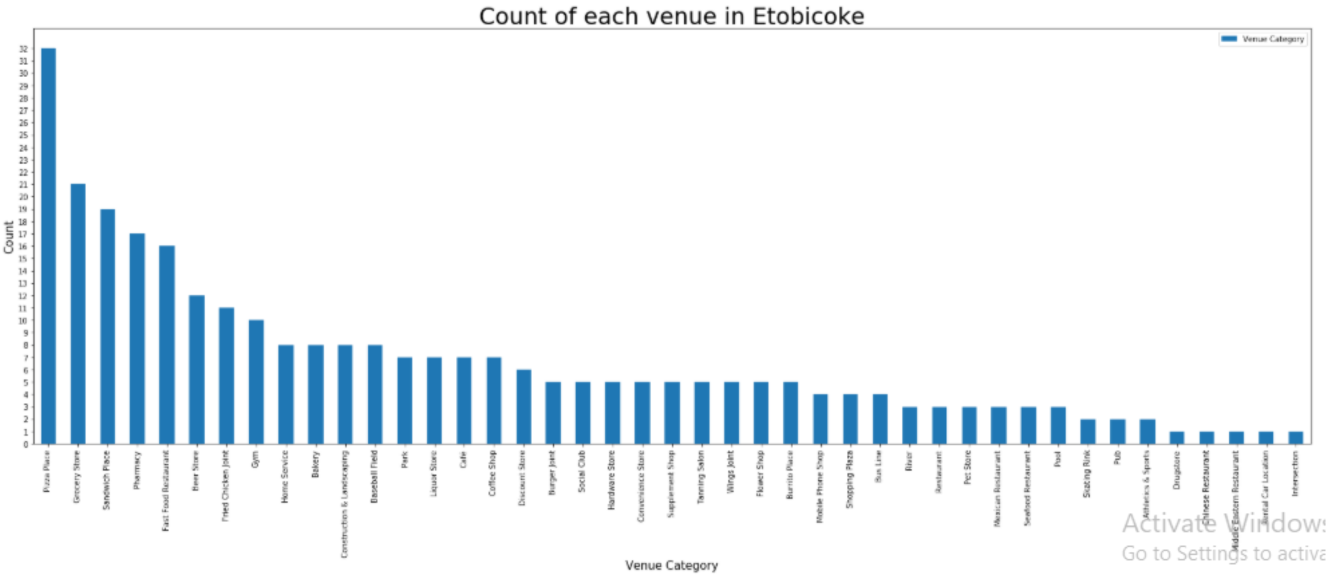
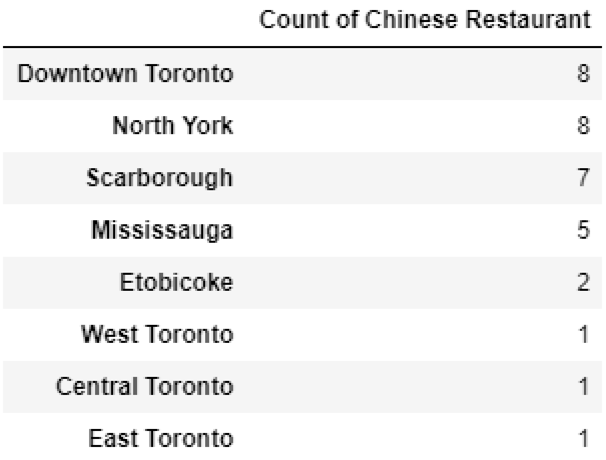


Figure 15: Bar chart for Etobicoke

5. Discussion Section

As in introduction we established that it will be best choice to open a new Chinese Restaurant in area where competition will be less so hence below our result table shows that number of Chinese restaurant in each Borough-wise analysis.



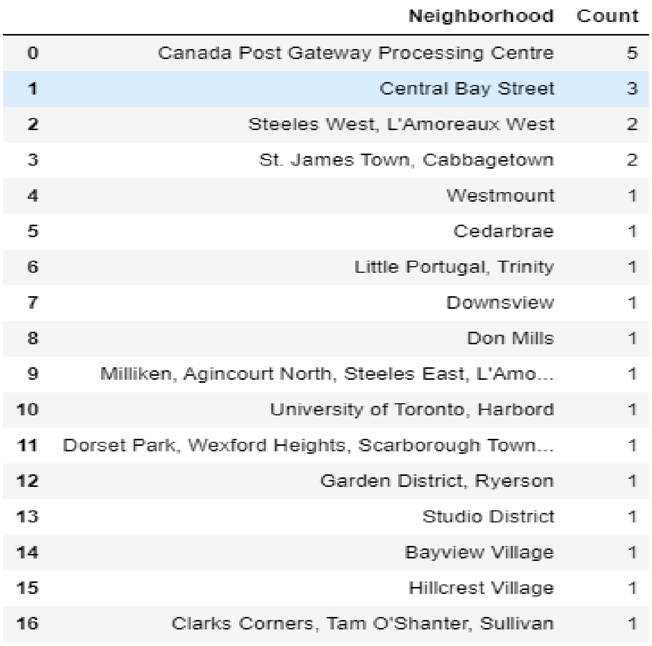
	Count of Chinese Restaurant
Downtown Toronto	8
North York	8
Scarborough	7
Mississauga	5
Etobicoke	2
West Toronto	1
Central Toronto	1
East Toronto	1

Figure 16: Table of Chinese Restaurants

Above table clearly shows that **Downtown Toronto**, **North York** and **Scarborough** have more number of Chinese Restaurants

Also we can clearly say that **York** and **Queen's Park** have no Chinese Restaurant, so these two may be best Boroughs for opening a new Chinese Restaurant.

If we apply Neighborhood-wise Analysis, we get following table:



	Neighborhood	Count
0	Canada Post Gateway Processing Centre	5
1	Central Bay Street	3
2	Steeles West, L'Amoreaux West	2
3	St. James Town, Cabbagetown	2
4	Westmount	1
5	Cedarbrae	1
6	Little Portugal, Trinity	1
7	Downsview	1
8	Don Mills	1
9	Milliken, Agincourt North, Steeles East, L'Amo...	1
10	University of Toronto, Harbord	1
11	Dorset Park, Wexford Heights, Scarborough Town...	1
12	Garden District, Ryerson	1
13	Studio District	1
14	Bayview Village	1
15	Hillcrest Village	1
16	Clarks Corners, Tam O'Shanter, Sullivan	1

Figure 17: Table for Neighborhoods

Above table show count of Chinese Restaurants in each Neighborhood we can now easily tell which Neighborhood will be best for opening a new Chinese Restaurant.

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 a Chinese 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.

6. Conclusion

During this capstone project, I was able to apply different data science techniques and tools that I learned. 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

- Through Borough-wise analysis we can say that **York** and **Queen's Park** are best Boroughs in Toronto for opening a new Chinese Restaurant
- Through Neighborhood-wise analysis we can say that **Canada Post Gateway Processing Centre** has more Chinese Restaurants and thus not suitable for opening a new Chinese Resturant, while those Neighbhoorhoods with frequency of 0 and 1 count can be considered instead.