Battle of Neighbourhoods: Report

Prospects of Opening an Indian Restaurant in the City of Toronto, Canada.

A. Introduction

1. Description of the Problem and Discussion of the Background

According to a study, Canada contains the worlds's eighth largest Indian diaspora. There has been a 74% rise in the Indian Immigrants since 2001.Not only that, Indians account for one-fourth of Immigrants Population in Canada. Canada is also the third most popular country to pursue Higher studies for Indian students. Toronto is the financial capital of Canada and you can find Indians working in every sector making it one of the best places to set up an Indian Restaurant.

We will work trough this project step by step. For this week I shall describe the initial data preparation and future steps for the project.

2. Initial Data Preparation (Week 1)

2.1. Get The Names of Boroughs, Neighbourhoods and Land Market Price from Wikipedia.

The wikipedia page of Toronto contains the table of Postal code, Borughs and Neighbourhoods. I have used ax excel file of the table since the data instances and attributes were limited. Ideally, it iscreate the inital frame. Another dataframe was created using the file Geospatial Coordinates.csv which was provided by Coursera. The Average Market price of land per sq ft was obtained from a report. We create a final dataframe after merging the two dataframes mentioned above. It gives an overall picture of the corresponding Borough , as later on I have considered top most venues within 500 meters radius of the selected Boroughs. After this inital preparation, I moved on to the next step to obtain coordinates using Geopy library.

2.2. Processing the Information From Wiki to select areas of interest.

Our main focus area was the areas of Main Toronto city. So the dataframe was first cleaned by dropping the NaN values in the data. Later on, the Borough names with key word 'Toronto' in them were selected and a new data frame was formed which was complete with attributes: Postal Code, Borough, Neighbourhood, Latitudes, Longitudes and Land Market Price per sq. ft.

2.3. Obtain the Average Land Market Price Data from Web-Scrapping

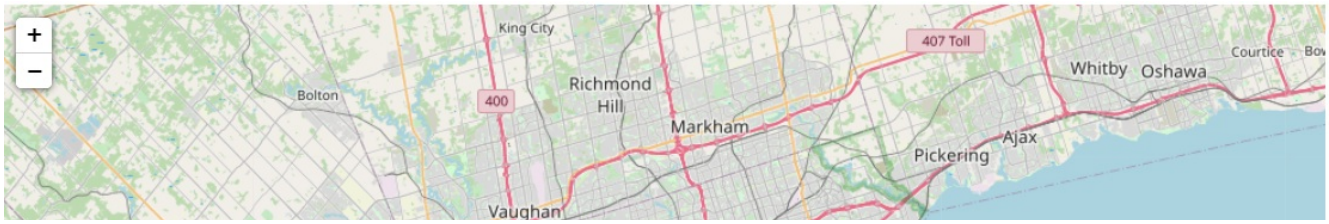
The average Land Market Price per sq.ft was extracted from a report named Toronto Fall report 2019 from a real estate website. The entries were manually added to the Toronto city data. The units for the price is in Dollars ($).

2.2. Foursquare Data (Week 2)

Here we make use of FOURSQUARE API to explore 100 venues within a radius of 500 m in and around Toronto.

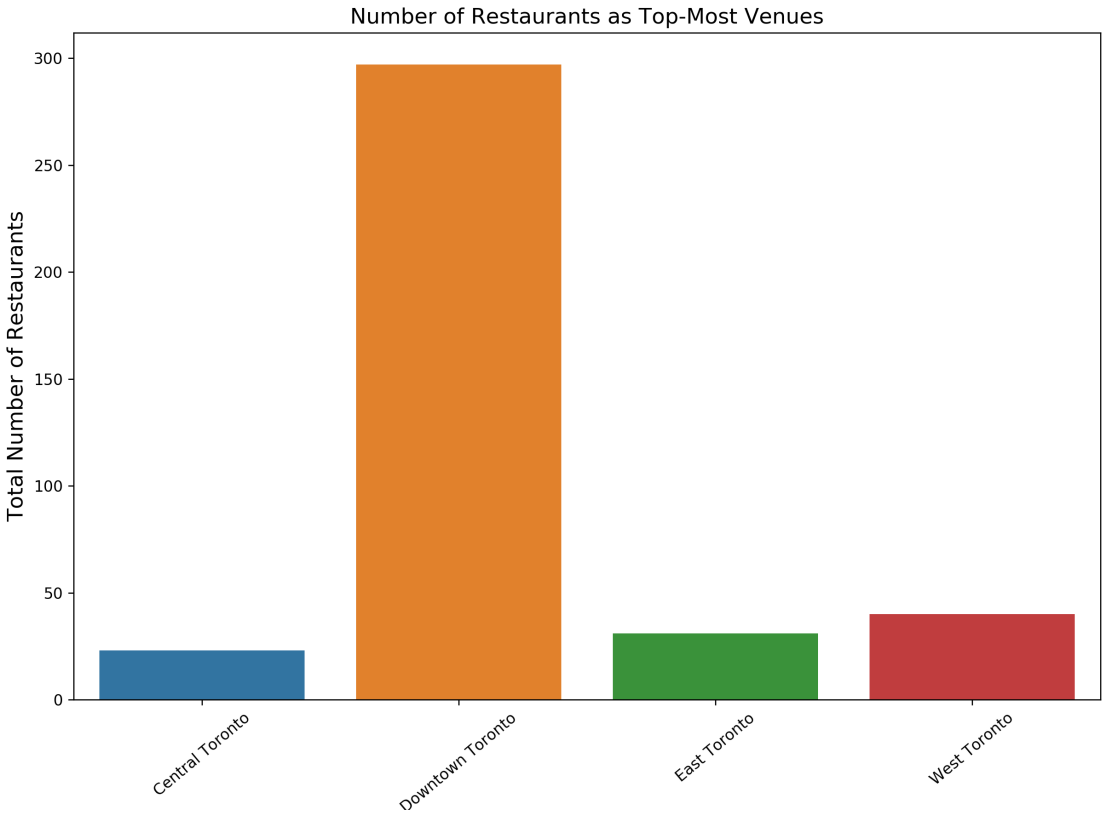
To decide which Borough is the best for opening a restaurant, we will explore the areas around the Main Toronto city which is divided into boroughs named East Toronto, West Toronto, Central Toronto and Downtown Toronto. Given below are the steps:

1.Create the get request url (Foursquare ID and Secret are necessary): \ 2.Assign Number of Venues as 100. \ 3.Radius of Search Would be 500 m. \ 4.Create a json from the request object (Need requests Module). \ 5.Create the lists Containing all the information. \ 6.From the lists create the dataframe.



3.1 Exploratory Data Analysis

From the Foursquare API, it was found that there are 1604 venues spread over 233 unique categories in Toronto within a radius of 500 m. Our main focus was restaurants so after running few lines of codes, we filtered our dataframe using the Venue Category as 'Restaurants'. This new data frame was merged with our main data frame using 'Outer Join' operator so that we can classify the restaurants according to the Boroughs.



Here, it was found that Downtown Toronto hosts the most number of restaurants followed by West Toronto, East Toronto and Central Toronto respectively.

3.2 Next step was to obtain information about the types of retaurants in each neighbourhood.

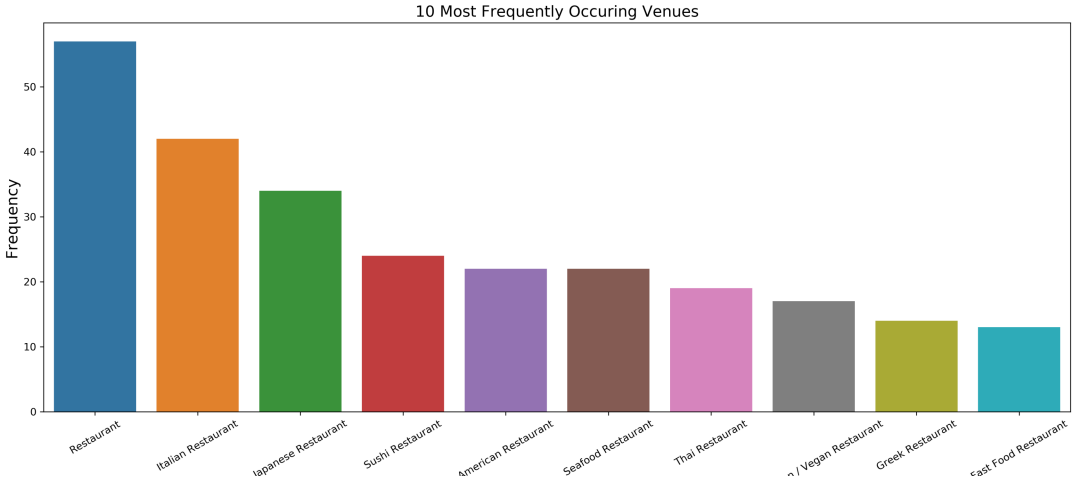
This can be done by following the steps given below:

1.Create a data-frame with pandas one hot encoding for the venue categories.

2.Use pandas groupby on Neighbourhood column and obtain the mean.

3.Transpose the data-frame at step 2 and arrange in descending order.

It is also important to find which types of restaurants are most popular around Toronto to get a clear idea of the preferences of the residents of the city. The study of the dataframe showed the following results stating clearly that the most popular cuisines were Uncategorised Restaurants followed by Italian and Japanese Restaurants as shown below.



4. Clustering the restaurants of Toronto

Finally, we try to cluster these Neighbourhoods based on the frequency of venue categories using K-Means clustering. So our expectation would be based on the similarities of venue categories, and the neighbourhoods will be clustered thereafter. Using K-Means algorithm from Scikit-learn library we obtain 5 clusters.

5. Results

The resutls of the exploratory data analysis and clustering are summarized below:

Restaurants, Italian restaurants and Japanese restuarnts top the charts of most common venues in the neighbourhoods of Toronto.

Downtown Toronto has the most number of restaurants and all categories of restaurants can be found here.

East Toronto is highly dominated by Fast Food Restaurants.

Japanese and Middle Eastern restaurants dominate the Central Toronto region.

Downtown Toronto hosts the most number of restaurants followed by West Toronto, East Toronto and Central Toronto respectively.

Central Toronto is the most expensive in terms of Average Land Market Price per sq. ft followed by Downtown Toronto, West Toronto and East Toronto.

6. Discussion

According to this analysis,neighbourhoods like Berczy Park,Brockton , Parkdale Village , Exhibition Place, Business reply mail Processing Centre will be suitable for an upcoming restaurant since there are hardly any proper eating joints in the most popular venue in these areas.

Also seen from the web-scrapped data, the average land price in and around West/East Toronto is much cheaper compared to the neighbourhoods close to central Toronto. So, definitely this region could potentially be a target for starting quality restaurants.

Some drawbacks of this analysis are: \ The clustering is completely based on the most common venues obtained from Foursquare data.There are many other factors that can be considered like the population, the age criteria,distnaces from various popular spots/offices etc.

Furthermore,from the kind of data distribution of restaurants as seen in the graph, this results also could potentially vary if we use some other clustering techniques like DBSCAN.

7. Conclusion

To conclude, this project gives an overview of how real life data-science projects look like.The analysis was done using of some frequently used python librairies to scrap web-data, use Foursquare API to explore the Toronto City Data and saw the results of segmentation of Neighbourhoods using Folium leaflet map. Also, some of the drawbacks and chance for improvements to represent even more realistic pictures

are mentioned.