**IBM Data Science Specialization**

**Applied Data Science Capstone Project**

**The Battle of Neighbourhoods: New York vs Toronto**

***Author:* Musah Abdul-Mumuni**

**September, 2019**

**TABLE OF CONTENTS**

1. **Executive Summary………………………………………………………………………………………………..3**
2. **Contents…………………………………………………………………………………………………………………4**
3. **Business Understanding……………………………………………………………………………….…..4**
4. **Analytical Approach…………………………………………………………………………..……………..5**
5. **Data Analysis and Visualization………………………………………………………………………...6**
6. **Data Requirements……………………………………………………………………………….6**
7. **Data Collection & Understanding: Neighborhoods Segmentation …...6**
8. **Toronto Data**
9. **New York Data**
10. **Data Preparation………………………………………………………………………………...14**
11. **Toronto**
12. **New York**
13. **Modelling and Evaluation ……………………………………………………………..…….16**
14. **Toronto**
15. **New York**
16. **Deployment and Feedback…………………………………………………………………..21**
17. **Results……………………………………………………………………………………………………………………21**
18. **Conclusion…………………………………………………………………………………………………………...…21**
19. **Future Directions…………………………………………………………………………………………………….21**
20. **References ……………………………………………………………………………………………………………..22**
21. **Executive Summary**

**The Battle of Neighbourhoods: New York vs Toronto:** One of the biggest challenges in life is making a decision to setup a business in an unknown region. i.e diving into unchartered waters. The risk is not only that the business can collapse and the capital lost, but also ones reputation to take important decisions would be tarnished by investors and family members alike. But when properly done the reward is immense. A family that operates a Doner shop in Europe, precisely Berlin, the capital of Germany is hoping to achieve a similar success in a North American financial capital. The decision on which city to choose to launch the business in order to avoid failure, survive and grow is not an easy task.

For such an investment to succeed, it requires very good understanding of the business location. The decision of choosing which city in North America to launch the business can be time consuming, tedious and costly because it requires good knowledge of the popular venues, foot traffic and certain establishments with which one can gauge the level of patronage of a similar business.

An important method that will immensely assist the decision making process is the application of Data Science methods which includes the use of Python Jupyter Notebook and Foursquare API for collecting available data online for analysis.

A presentation of the process of collecting, preprocessing and analyzing Neighbourhood data of North America´s two richest financial capitals, New York and Toronto in order to make a decision on where to start an exotic fast food business is outlined in this report.

1. **Contents**
2. **BUSINESS UNDERSTANDING**

A family that operates a Doner (a very popular dish in Germany consisting of spiced lamb cooked on a spit and served in slices, typically with pitta bread ) shop in Berlin the capital of Germany in Europe is considering introducing its business to the North American market. The family is facing the choice of either establishing its business in New York or Toronto. The reason for choosing one of the two cities is that, Toronto being the Commercial capital of Canada and New York the commercial capital of the United States of America will offer the much needed foot traffic and also the exotic (Italian, Chinese, Indian, Jamaican Restaurants) eating habits that is associated with cosmopolitan neighbourhoods, a necessary ingredient for the survival and growth of such a Business venture.

On the basis of the fact that The US and Canada fall in the top 3 countries with the highest consumption of fast foods in the world, with the U.S coming first and Canada placing 3rd behind Japan, makes North America a reasonable location for establishing a Doner business. It is also important to consider that, more than 1 out of every 3 American adult eats fast food in a given day and a similar thing exist in Canada.

The presence of other similar exotic businesses and Beach, touristic, sports, leisure and recreational facilities are positive indication of a future for the business.

Data Science methodology is one of the modern world´s effective and efficient method of making decisions such as making an investment decision far away on another continent. To choose between New York and Toronto on the basis of where is more suitable to establish a Doner Business can be done using Python Jupyter Notebook and Foursquare API for collecting location data and required libraries are imported for preprocessing, clustering and analyzing the data.

Clustering, a machine learning technique is used to segment the neighborhoods with similar objects on the basis of each neighborhood data. These objects will be given priority on the basis of exotic restaurants and foot traffic in their respective neighborhoods.

From the work done, any exotic fast food entrepreneur outside North America wishing to invest in one of the top Financial Capitals in North America can find the model useful in deciding either to choose New York or Toronto.

1. **ANALYTIC APPROACH**

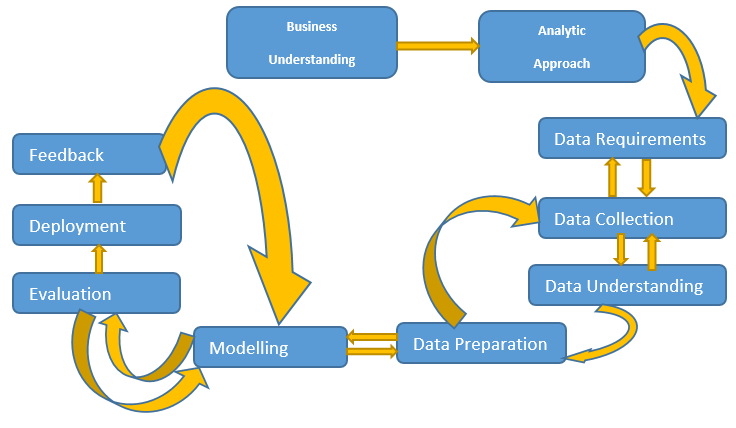


Fig. 1

The chart above summarizes the Data Science methodology to be applied in solving the problem of choosing between New York and Toronto for Doner Business establishment. It starts with Business Understanding, where a definition of the goal is establishing a Doner Business which will be successful in a North American financial capital is defined. It is followed by Analytical Approach where the guidelines or patterns for achieving the goal is defined. In this case, with the use of Python Jupyter notebook and Foursquare API, Neighbourhood data are collected online of the two cities and analyzed.

Data Requirement, collection, understanding and preparation stages are done iteratively. Questions are continuously asked in the process and more or similar and better data are required, collected and prepared for analysis to create a model from which solutions concerning establishing other related businesses in a similar setting can be found by manipulating the model. The model is then evaluated by checking other Neighbourhood venue within the model. Model evaluation goes hand-in-hand with model building as such, the modeling and evaluation stages are done iteratively. When satisfied of its performance it is deployed by establishing the business in the chosen city. Continuous feedback is fed into the model to achieve optimum functioning of the model.

1. **DATA ANALYSIS AND VISUALISATION**
2. **Data Requirement**

The data to be used are the Neighbourhood data of New York and that of Toronto. The data contains the postcodes, Boroughs and the corresponding neighbourhoods from which the venues where establishments like restaurants, café etc can be found for use in the analysis.

1. **Data Collection & Understanding**

By making use of Foursquare API, the opportunity to explore the data of two cities, in terms of their neighborhoods can be made. The data sets are imported from Wikipedia and IBM open data. The data also include the information about the places around each neighborhood like restaurants, hotels, coffee shops, parks, theaters, art galleries, museums and many more. A selection of one Borough from each city is done to analyze their neighborhoods. Queens from New York is chosen for the U.S. and Scarborough in Toronto is chosen for Canada.

The reason for choosing Queens is that it is the Borough with the highest Neighbourhoods and it has almost the largest population in New York. Scarborough is only second to North York in terms of highest neighbourhoods and population in Toronto and unlike North York it is located near a beach.

IBM Watson Jupyter Notebook was used for the exercise. First of all the raw data had to be imported. For the Toronto data it was scraped from Wikipedia. Later coordinates data for the city of Toronto are collected and added in order to support finding the venues. The original data of venues collected for the analysis was not uniform for both cities since Queens has a larger population than Scarborough and data returned for Scarborough within a radius of 500m was not satisfactory. Therefore the radius for collection of data was increased for both cities from 500m to 1000m.

**Neighborhoods Segmentation**

Neighborhoods Segmentation is done by importing the Boroughs and neighborhood list of Toronto from Wikipedia and converting it to data frame using pandas package in python. Then, another data set comprised of location data of neighborhood and boroughs was imported. It was in .csv format and then converted to data frame. After Cleaning the data set, two tables were merged to get the final Toronto neighborhood data set. The url of the Toronto data is as follows: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

A list of the libraries imported and downloaded for the project are as follows

|  |  |
| --- | --- |
| LIBRARY | USAGE |
| pandas | For data analysis. For data processing, CSV files |
| numpy | To handle data in a vectorized manner. For scientific computing, arrays, algebra, matrices |
| beautiful soup | It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping |
| json | library to handle JSON files. is a very common data format used for asynchronous browser–server communication, including as a replacement for XML in some AJAX-style systems. |
| request | to handle requests. It is to make HTTP requests simpler and more human-friendly. |
| matplotlib | It provides an object-oriented API for embedding plots into applications |
| scikit-learn | Simple and efficient tools for data mining and data analysis |
| cluster | allows to create several groups (clusters) of objects from a list. |
| folium | Helps create several types of Leaflet maps |
| geopy | makes it easy to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources |

Table. 1

1. ***Toronto Data***

Looking at the content and Structure of the data: The first 15 rows of the Toronto data making up of Postcode, Borough and Neighbourhood data as presented in a pandas dataframe. Putting together neighborhoods which have the same postcode in one row under the particular postcode

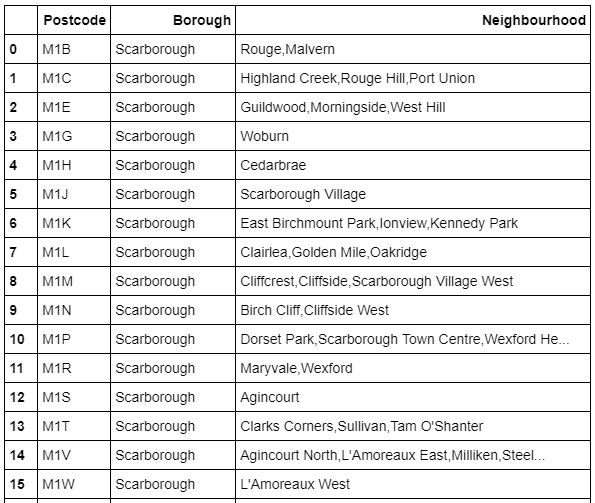


Fig. 2 initial dataframe of Toronto

Neighbourhood with missing names "Not assigned" are given the corresponding Borough names.

After importing csv file into a pandas dataframe containing the corresponding coordinates of the city of Toronto merging the coordinates together with the city names and postcode in Toronto



Fig. 3 dataframe of Toronto data without missing values

**Summary of Boroughs and their Neighbourhoods in Toronto**

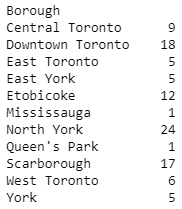


Fig 4 summary of Boroughs and their number of neighbourhoods in Toronto

**Collecting and analyzing data for North York.**

North York was first considered to be the Neighbourhood in Toronto for the establishment of the business. It has the highest number of Neighbourhoods, 24 with a population of 656 thousand compared to Scarborough having 17 Neighbourhoods and a population of 632 thousand.

Its data was collected and analysed in an attempt to pick it as the Main contender for Queens in New York but after looking closely at its location on the map as shown below, Scarborough was chosen instead.

North York latitude 43.7708175 & longitude -79.4132998

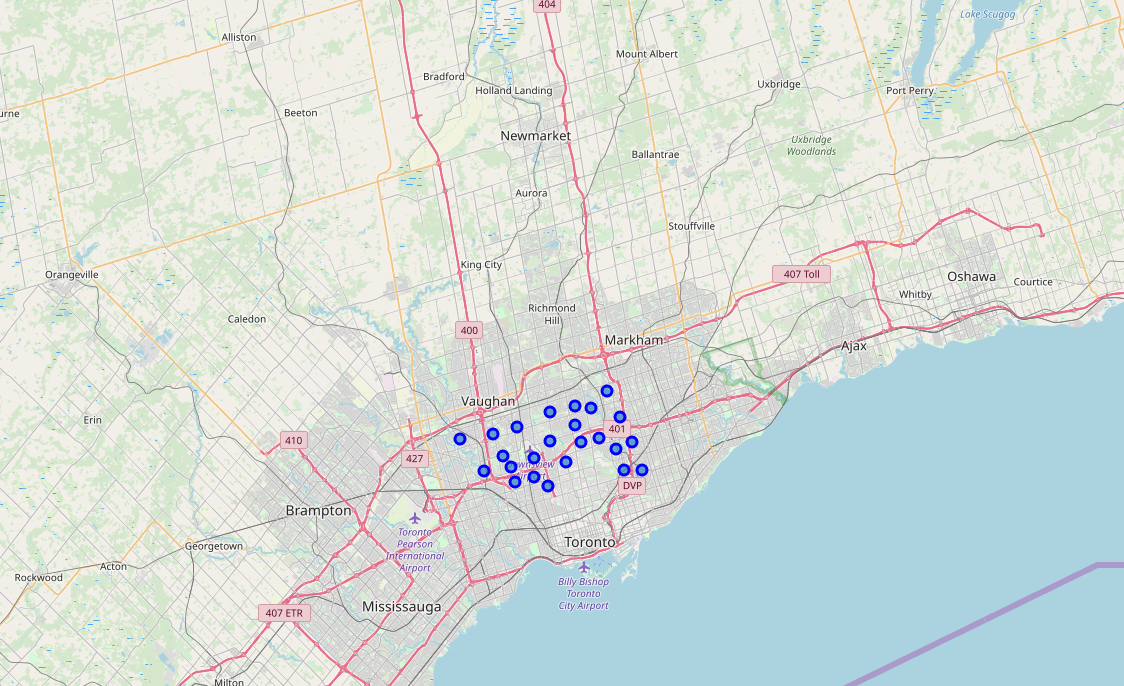


Fig. 5 map of North York

Map of North York above clearly shows it is away from the coast which was the reason Scarborough was chosen instead of North York.

Most tourists and foot traffic are heavily concentrated around beaches thereby increasing the buying of fast and exotic food especially in Summer.

Down Town Toronto was not chosen despite its higher Neighbourhoods of 18 because its population is just 199 thousand which is less than that of Scarborough

**Scarborough**

Scarborough which has almost the biggest Boroughs (17 Neighbourhoods) in Toronto with almost the highest population in Toronto of 632 thousand was chosen. Among the reasons why Scarborough was chosen ahead of North York in Toronto is that Scarborough just like Queens is closer to the coast than North York, despite North York having slightly higher population and more Neighbourhoods

The first five results of the Neighbourhoods in Scarborough with Rouge, Malvern being the first Neighbourhood data is given in the dataframe below

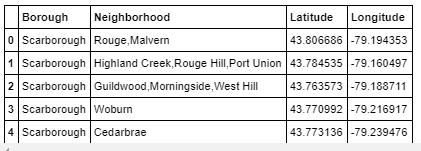


Fig. 6 dataframe of Scarborough data with coordinates

**Map of Scarborough**



Fig. 7 map of scarborough

**Cluster map of Scarborough**

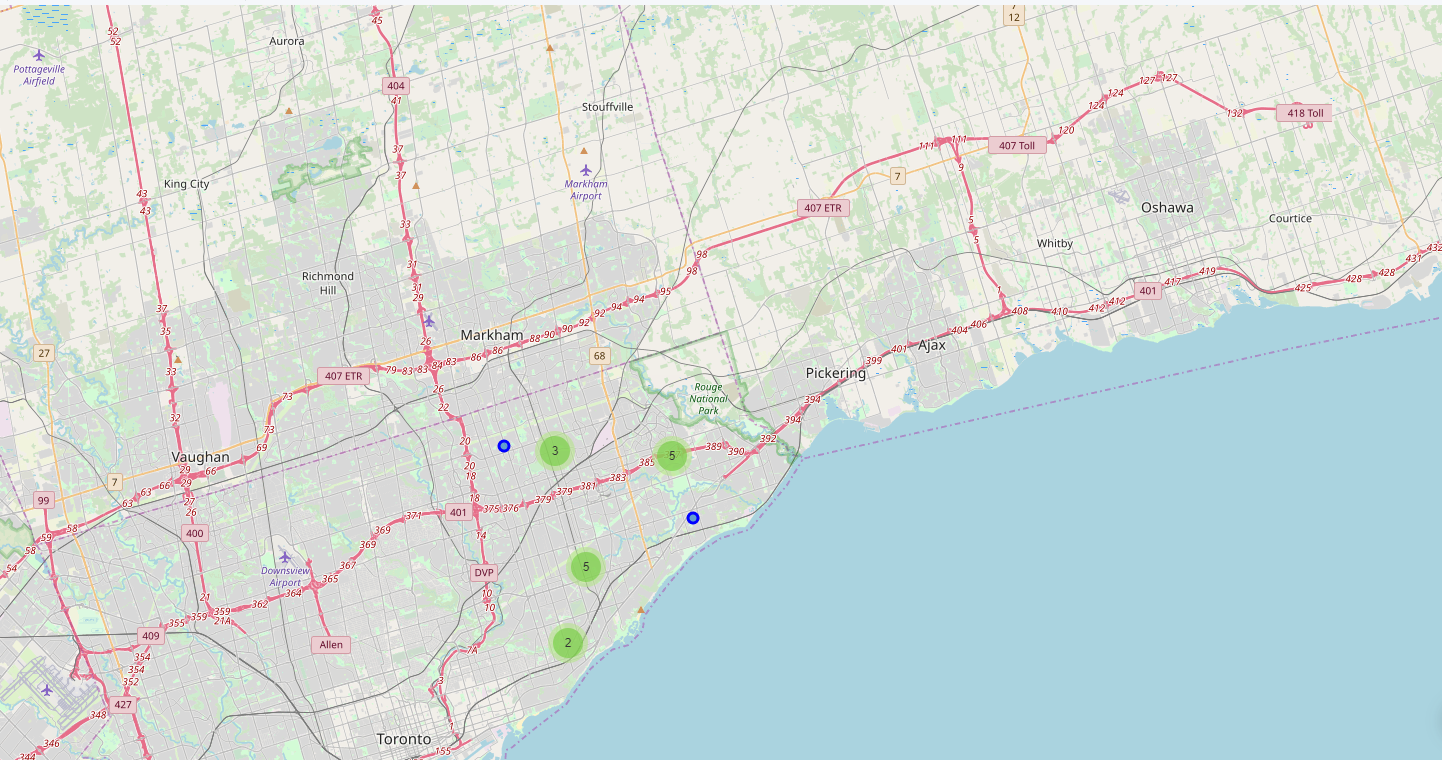


Fig. 8 cluster map of Scarborough

**1st Data Collection**

# Getting the top 100 venues within a radius of 500 meters of Rouge, Malvern

### After cleaning the json and structuring it into a pandas dataframe.



Fig. 9 Only one result was returned for the 500m radius

### **2nd Data Colleciton:**

Due to the very low number of venues returned in the first data collection, a second data collection is made where the radius is increased from 500 to 1000

# Getting the top 100 venues that are in Rouge, Malvern within a radius of 1000 meters.

# After cleaning the json and structuring it into a pandas dataframe

# And checking how many venues were returned by Foursquare



# Fig. 10 First five results of 19 venues returned for radius 1000

1. ***New York Data***

### importing data for preprocessing and understanding the New York City data and putting it into a dataframe. The following was used:

!wget -q -O 'newyork\_data.json' <https://cocl.us/new_york_dataset>



Fig. 11 imported data are placed in a dataframe with its coordinates

### Summary of Boroughs and their neighbourhoods in New York

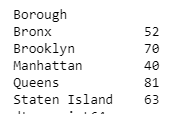


Fig. 12

**Queens**

### Queens Borough, New York has the most Neighbourhoods, 81 and a population of about 2.4 million (wikipedia) was chosen as the City in New York to contend for the Doner business establishment.

### Getting the geographical coordinates of Queens, New York and part of its Neighbourhood data

Queens latitude 40.6504178 & longitude -73.7971341



Fig. 13 Queens Borough was chosen to represent New York

### creating map of Queens using latitude and longitude values

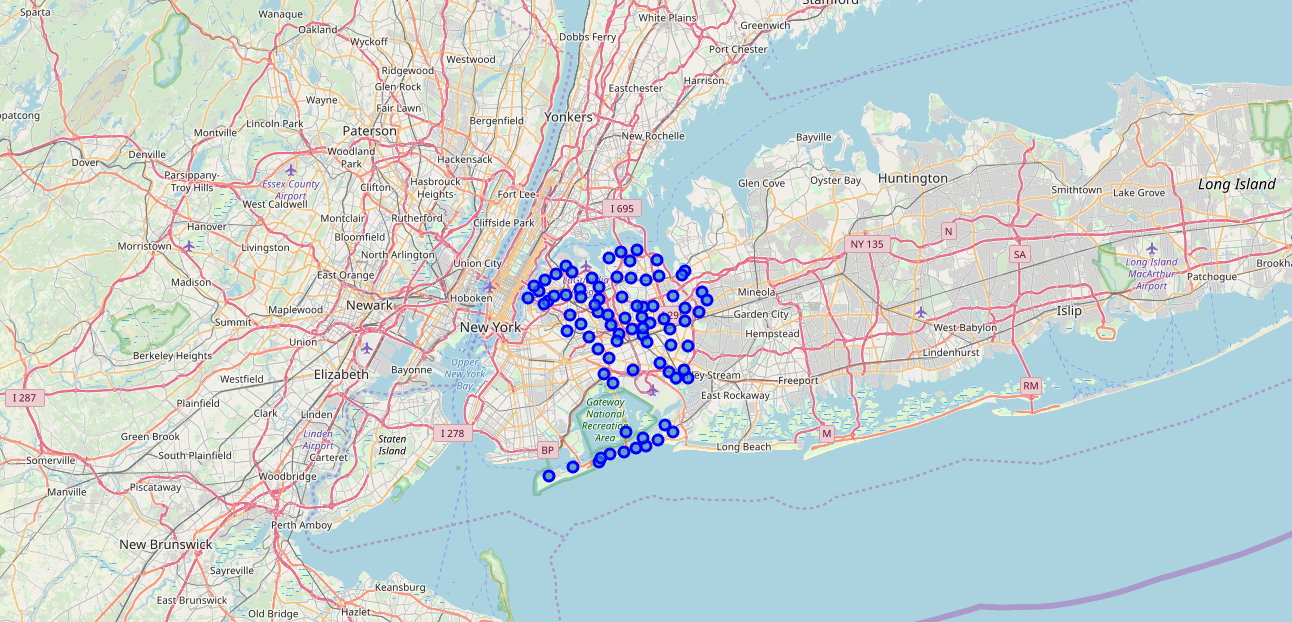


Fig. 14 map of queens with its Neighbourhoods

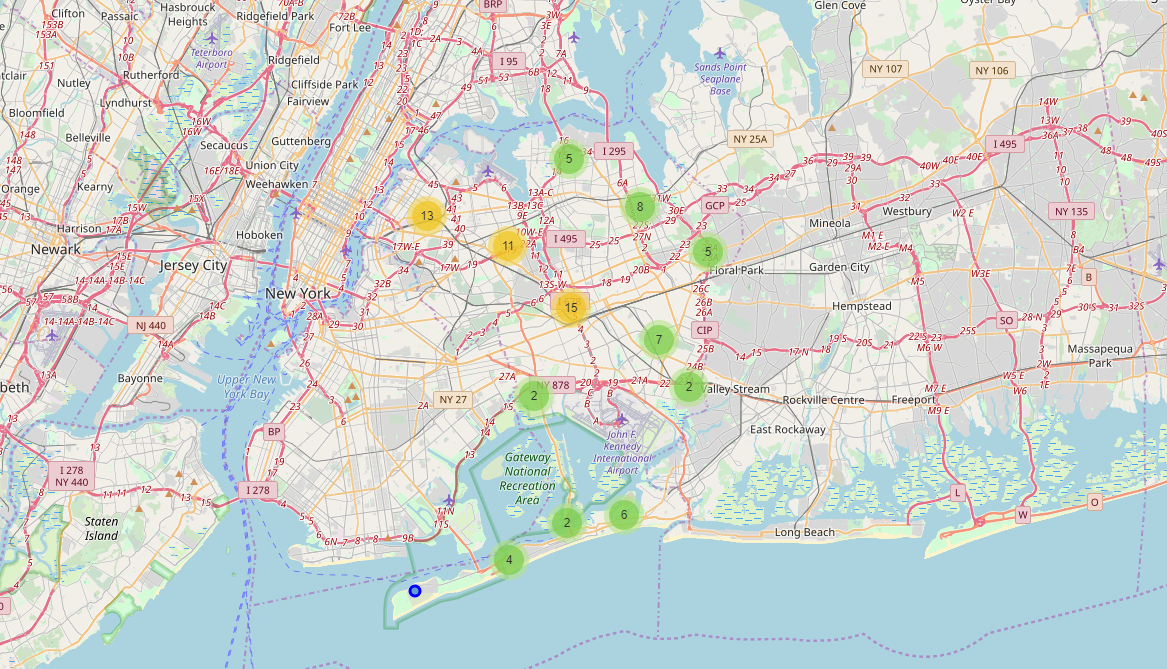


Fig. 15 visualising the cluster map of Queens

### Exploring the first Neighborhood in our Queens dataframe by getting the neighborhood's name which is “Astoria”

**Getting the top 100 venues that are in Astoria within a radius of 1000 meters.**

### Exploring the first Neighborhood in our dataframe by getting the neighborhood's name, Astoria. Getting the top 100 venues that are in Astoria within a radius of 1000 meters.

The dataframe for the first 5 results are as follows



Fig. 16 top 100 venues within a radius of 1000m from Astoria

### checking how many venues were returned by Foursquare.100 venues were returned by Foursquare.

1. **Data Preparation**
2. ***Toronto Data***

II Exploring Neighborhoods in Scarborough by creating a function to repeat the same process to all the neighborhoods in Scarborough

checking the size of the resulting dataframe (90, 7) i.e 90 rows and 7 columns



Fig. 17 exploring neighbourhoods in Scarborough

### checking how many venues were returned for each neighborhood

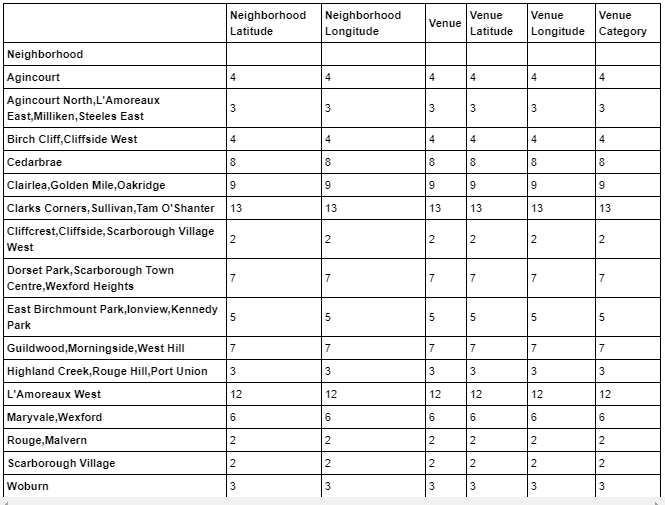


Fig. 18 Scarborough Neighbourhoods and corresponding venues

1. ***New York Data***

# Exploring Neighborhoods in Queens by creating a function to repeat the same process to all the neighborhoods in Queens, running a function on each neighborhood and create a new dataframe called Queens\_venues

### checking the size of the resulting dataframe (2154, 7) i.e 2154 rows and 7 columns



Fig. 19 first five results of venues in Queens

### checking how many venues were returned for each Neighbourhood

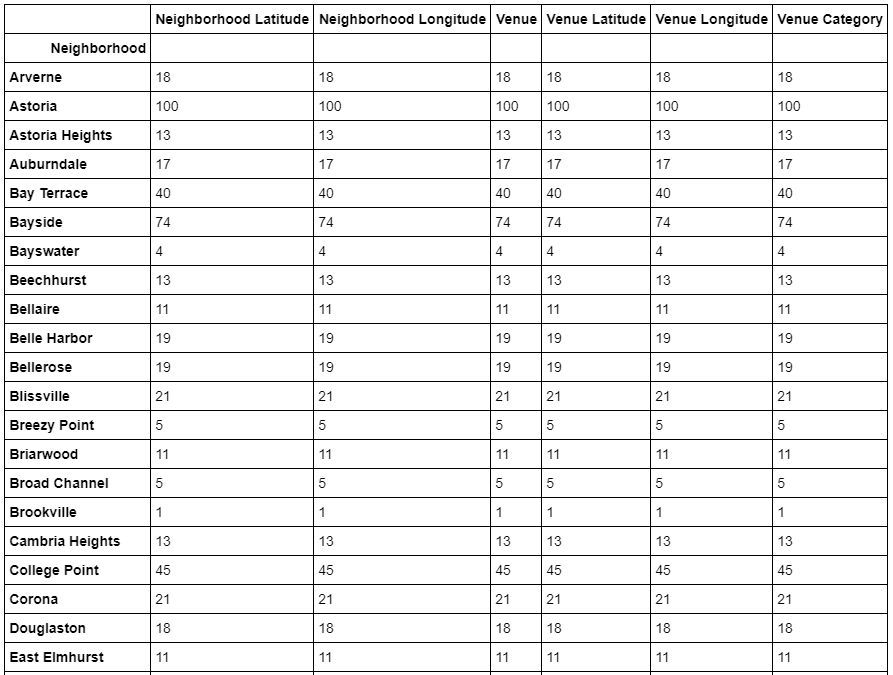


Fig. 20

### Finding out how many unique categories can be curated from all the returned venues. There are 270 unique categories.

1. **Modelling and Evaluation**

***a. Toronto Data***

Analyzing Each Neighborhood and checking the size of the new dataframe(90, 55) i.e 90 rows and 55 columns. Then grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category. checking the new size (16, 55)i.e 16 rows and 55 columns

### creating the new dataframe and display the top 10 venues for each neighborhood.



Fig. 21 first five results of top 10 venues

### running k-means to cluster the neighborhood into 5 clusters and merging data for each neighborhood

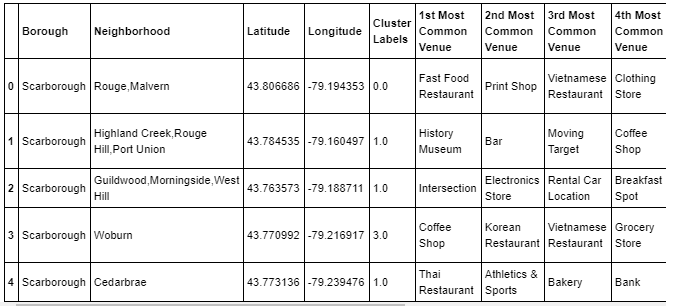


Fig. 22 initial clustering

# Examining the Clusters:

### each cluster is examined to determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, a name for each cluster can be assigned.

### first cluster (fast food, Vietnamese restaurant, fried chicken joint)

****

Fig. 23 1st cluster Fast food Restaurant is the 1st most common venue

### (Thai, Indian, Chinese, Caribbean restaurants, pizza place)

****

Fig. 24 2nd cluster has Indian, Thai and Pizza as the 1st most common venues

### (Vietnamese, Fast Food restaurant)

****

Fig. 25 3rd cluster

### (Korean, coffee shop, fast food restaurant)

****

Fig. 26 4th cluster

(Fried Chicken joint, Fast Food restaurant)



Fig. 27 5th cluster

# *b. New York Data*

# Clustering Neighborhoods

first results



Fig. 28

### a visualization of the resulting clusters

# 

# Fig. 29 resulting cluster map of Queens Borough

# Examining the Clusters

### each cluster is examined to determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, a name for each cluster can be assigned.

### (Korean, Filipino, Indian, Pizza, Sushi, Vietnamese, Chinese, Italian, Mexican, Greek, Caribbean, Thai restaurants)



Fig. 30 1st cluster has many exotic restaurants as part of the first most common venue

### (park, falafel, European, Eastern European, empanada restaurant)

### 

Fig. 31 2nd cluster even though park is the first most common venue there are other exotic restaurants making the top ten cut

### (Beach, Fast food, Falafel, Empanada restaurant)

### 

Fig. 32 3rd cluster even though Beach is the first most common venue in this cluster a couple of exotic food restaurants made the top ten

### (Falafel, empanada, Fast Food restaurant)



Fig. 33 4th cluster

#### (Empanada, Falafel, Filipino, fast food restaurant)



Fig. 34 5th cluster

# Deployment and Feedback

# When the decision on the city is made and the business is established, say New York is chosen, there should be continuous feedback fed into the model. Questions whose answers should be input into the model are “are the numerous exotic restaurants in the Neighbourhood acting as a factor for high patronage or acting as competition therefore less patronage”

# Results

### After clustering the data of the respective neighborhoods, both cities (Boroughs) have venues which can be explored for exotic restaurants. In the first cluster the first of Queens, the most common venues includes exotic restaurants like Latin American, Thai, Mexican, Chinese and Pizza restaurants. Toronto had its second cluster also made up of a number of exotic restaurants.

### The neighborhoods are much similar with regards to them having wide ranging exotic fast food restaurants. The main difference is that New York with its large population and many Neighbourhoods offers more exotic food restaurants than Toronto.

1. **Conclusion**

The descriptive model created shows the location of exotic restaurants in North America`s two financial capitals New York and Toronto and the top most common venues. From the clusters we can see that exotic restaurants, such as Chinese, Thai, and Greek make up the top most visited places. It is also clear that despite Toronto having its share of exotic restaurants, New York with its large population and many Neighbourhoods offers more exotic restaurants. The decision to choose to establish a Doner business in New York over Toronto can be established. However, Toronto has the advantage of offering less competition but a population that seems very interested in exotic restaurants.

1. **Future Directions**

A more thorough analysis could be made by comparing income, spending on food especially exotic, fast food, etc. Also compare North York (Toronto) with Queens (New York) or Manhattan in New York with North York. The distance between the venues can also be computed to find a place with the highest number of customers or visits to make the location of business establishment optimum.

References:

1. *Wikipedia*
2. *IBM Cognitive Class Labs*
3. *IBM Data Science Methodology lecture*