

2020

Battle of Neighborhoods

Poobathi Subramani

FEB 2020



1. Abstract

This report discusses about comparing the business opportunity provided by two major cities – New York, USA and Toronto Canada. This is to support a decision making process of establishing a restaurant in one of these cities. Appropriate statistical methods and models have been adopted to analyze the data and visualize the data. A conclusion was arrived based on the results found in the analysis and visualization.



2. Introduction

One of my client reached out to me with an intention of opening a restaurant in either NY City or Toronto. He asked me to come up with suggestions with supporting artefacts.

Just to give a brief introduction about my client - he is food lover! He would like to travel across the globe, meet with people, and the community, understand their neighbourhood, their cultures etc. Out of everything, the one thing he loves the most is the food. From local roadside food to exotic international cuisines, he have lot of love and appreciation for their look, feel and taste. Being enjoying the variety of food for several years, it is always his wish to open a restaurant to serve delicious food for the craving taste buds.

Within the two great cities, New York and Toronto, he would like to know which city provide more opportunities to realise his wish and to be successful.

Provided there are several publicly available data sources like wiki, Foursquare etc., I would like to source the datasets, clean and transform as required, analyse using visual tools and techniques and summarise my findings to help my client to make a decision.

3. Data

a. Data Sources

For the analysis, the data acquired from publically available data sources. The following are the sources:

1. Foursquare is the major source of my analysis as it provides all the venues around the neighbourhoods. This source is capable of providing lot of additional information like ratings etc. but for this analysis, I will be focusing at the high level and using the appropriate level of data that can be obtained with a regular membership
2. Wiki - This is another lake of source. I will be using this to get the initial list such as zip/postal codes, boroughs etc.
 - i. Toronto Neighbourhood
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
3. Other sites - to get information like demographics etc.
 - i. NY City Zip <https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/>
 - ii. NY zip code Mapping
<https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.html>

b. Data cleaning and formatting

The data obtained from the sources is cleaned before the further use for analysis. Below are the steps involved in the cleaning process:

1. Toronto Neighbourhood data – Fixing 'Not assigned' Neighbourhoods - If the 'Neighbourhood' contains value 'Not assigned' then use the value from 'Borough'
2. Consolidating Neighbourhoods - There are more than one neighbourhoods for Postcode and Borough combination. in such cases, the neighbourhood values to be concatenated into a single row.
3. NY data contained an unwanted column 'ZIP Codes' so that was removed
4. The 'Zip' column contains zip in a string concatenated format and that was normalized
5. Unwanted columns ('State', 'Timezone', 'Daylight savings time flag', 'geopoint') were dropped from the NY Geo file
6. The zip code column name and the data type are updated
7. There are several NaN rows which are removed

4. Methodology

We have used K-Means Clustering technique in this analysis. Since we do not have any already known answer to our business problem, we need to use unsupervised algorithm. K-Means clustering is one of the simplest and popular unsupervised machine-learning algorithm. Based on the business problem, we would like to make a recommendation on which city would be appropriate for starting a business.

To draw recommendations, we will analyse the data, especially the food related establishments and its location to understand how well the two cities are doing and what are the opportunities both places might have. Therefore, we have grouped the collected data based on the categories so that we can see what kind of establishments are exist in the neighbourhoods. This will help to make decision as to what kind of restaurant would work better in the given place and of course, the ultimate question – which city would be the best.

We initially looked at all the venues within 500 meters distance around the neighbourhood location specified by latitude and longitude. The venue data was obtained from Foursquare. Each neighbourhood has more than one venues around it. Then we have gone through the list to identify the food/restaurant related venues so that we have filter and analyse only those venues. After filtered, we have grouped the venues based on the venue category. Then we have transformed the data using the 'one hot encoding' technic. This will allow us to find the mean of each cuisine type with respect to each neighbourhood. We have input this result into the KMeans model.

For the model, we have changed the K value to check the distribution of clusters across the population and we have finalized $k=8$. We have not used the complex model evaluation techniques like inter-cluster and intra-cluster variances.

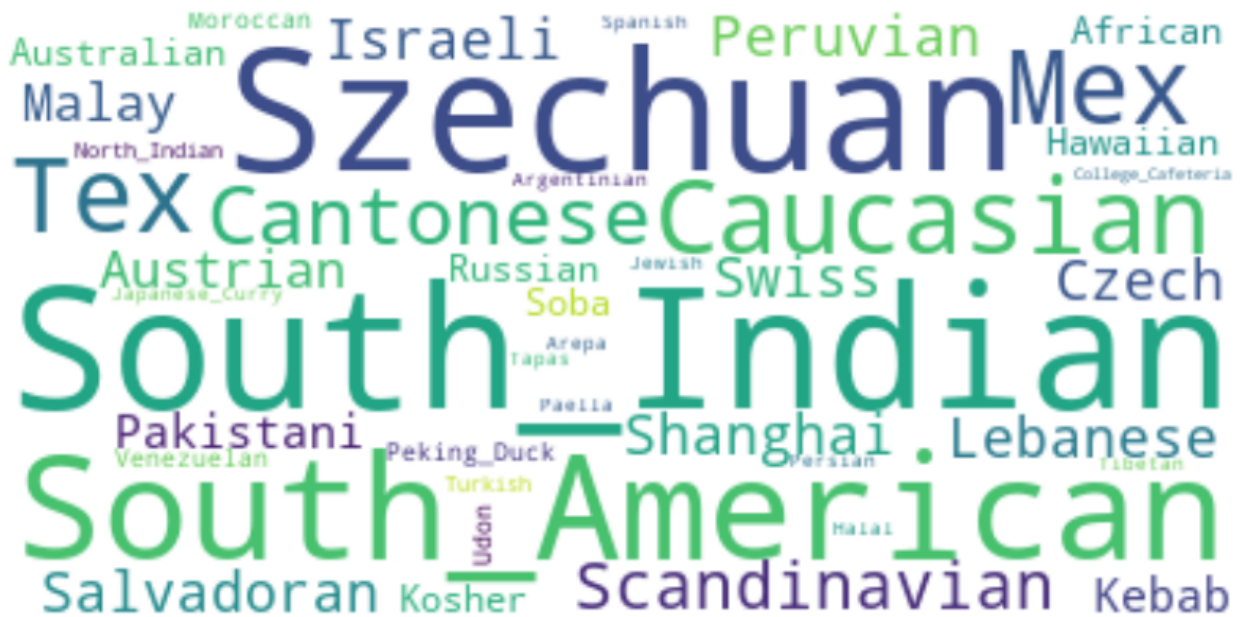
The model generated the cluster lables, which we combined with the original dataset and used that to plot the maps of Toronto and New York cities. We have used Folium to plot the map.

A visual inspection in the maps helped as to see how the neighbourhoods are clustered.

5. Results

From the analysis, we have uncovered interesting facts about the two cities as far as the food related establishments are concerned. We have noticed that, from the Foursquare data, the NY city has more venues than Toronto City. New York City has 7000+ venues whereas Toronto has 4000+ venues. In addition, for the subject of our interest, the count goes like 3000+ in New York and 1000+ in Toronto.

Though there are several commonalities exists between NY and TO, the distinct food industries that are existing in NY is larger than the establishments in TO. The below Word Cloud shows the cuisines that are exist in NY but not in TO.



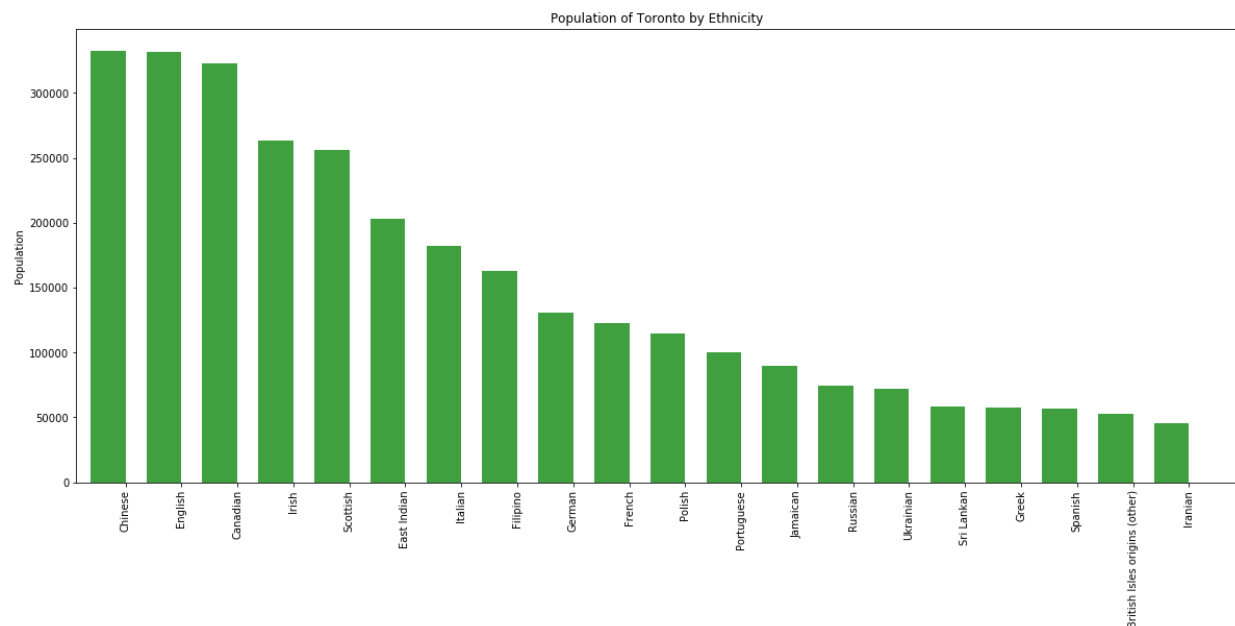
The Word Cloud highlights the more frequent cuisines in the bigger font size. That means there are more consumers in the NY City for those cuisines comparatively. For example, look at the following cuisines:

1. South Indian
2. South American
3. Szechuan
4. Caucasian
5. Scandinavian
6. Salvadoran

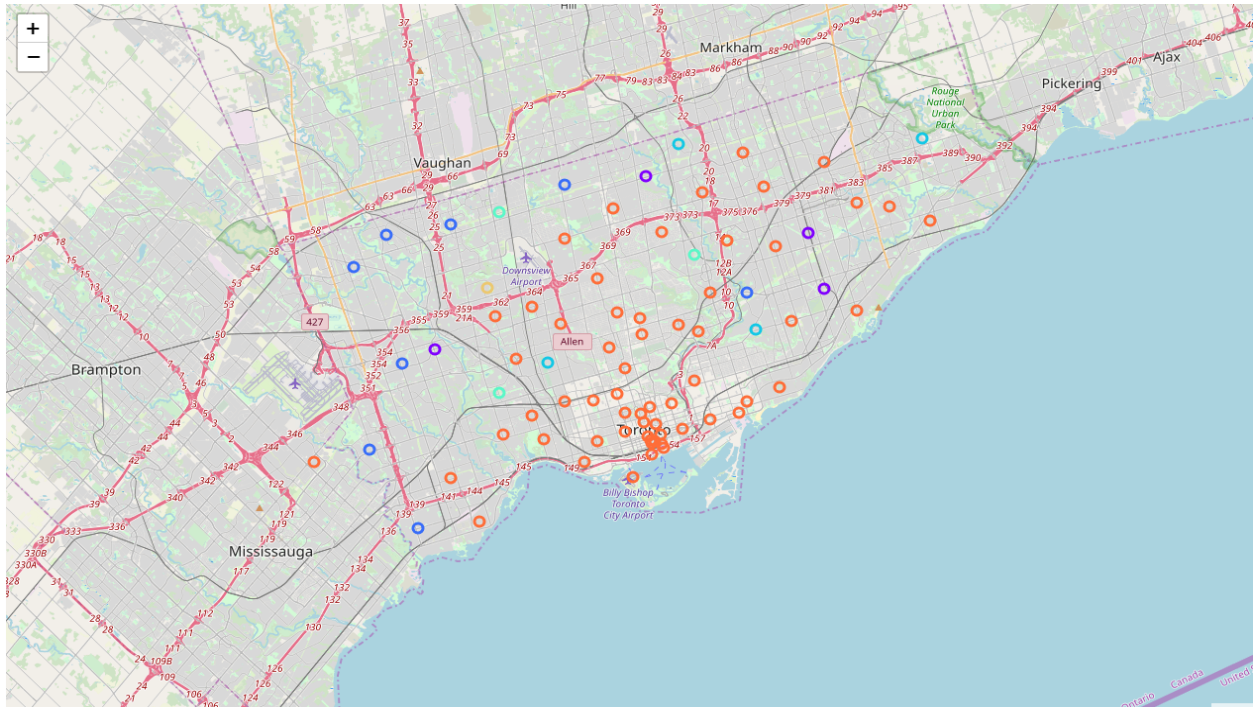
Now, let's look at the cuisines exist in TO that are not exists in NY.



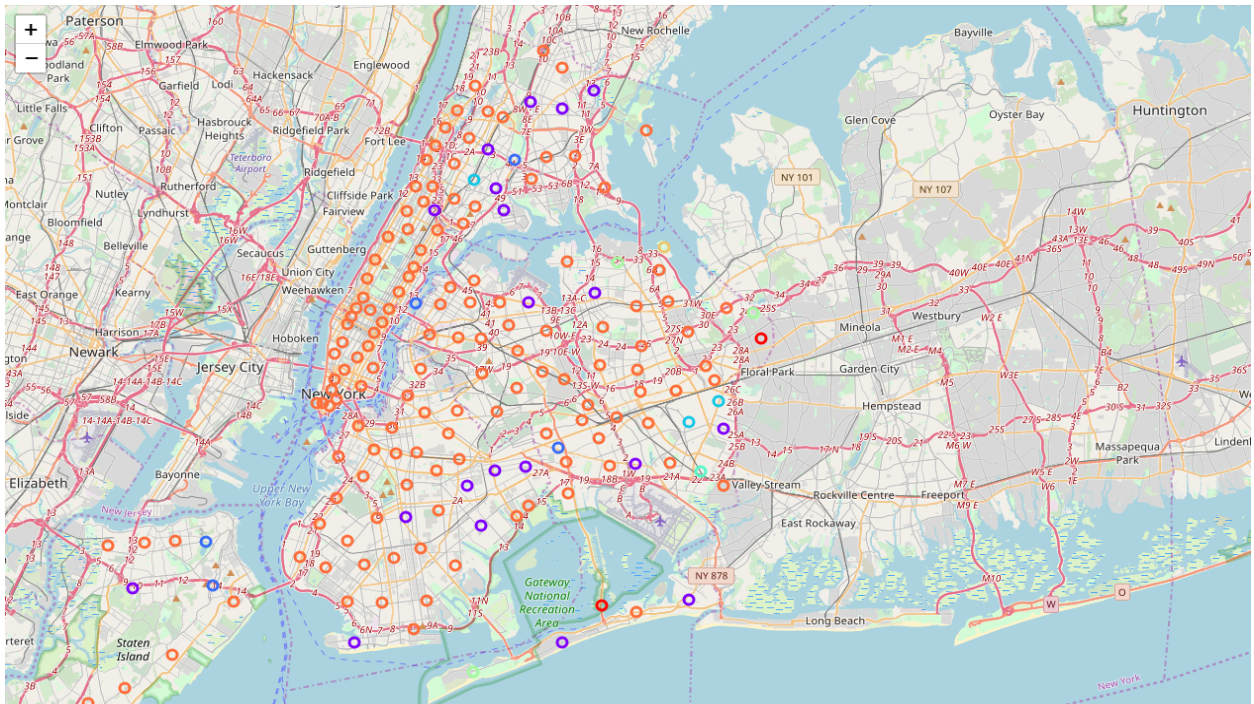
The Word Cloud shows that there are very less speciality cuisines in Toronto. Moreover, cuisines like 'Airport Food' are fairly considered to be a location-based, regular cuisine rather than an actual speciality cuisine (like Peruvian Cuisine)




i. Clustering of neighbourhoods:
Toronto:



New York:





The clustering shows how each neighbourhood is classified or clustered based on the type of restaurants available in and around the neighbourhood. We can see that the clustering look almost similar between both the cities.

6. Discussion

Based on the results above, it seems that, though Toronto and New York are similar in several aspects, especially, multi-cultural, multi-ethnic etc. the population living in Toronto do not have much options in multi-cuisine restaurants. Toronto have only half of the Cuisine types as compared to NY City.

Out of 7000 venues in New York, 3000 venues are restaurant/food related which is 57%. Whereas in Toronto, out of 4000 venues, 1000 venues are restaurant/food related which is 25%.

There are 105 distinct cuisines in New York, but only 75 in Toronto, which makes only 71% of New York cuisines.

7. Conclusion

There are 41 distinct cuisines in New York are not in Toronto. This definitely creates a potential business opportunity. Based on the above discussion, **we strongly recommend starting a restaurant in Toronto City**. And, the restaurant could be of any speciality type which is more popular in New York (depicted in the Word Cloud) as per the client's individual preference and skillset. The reason why we are suggesting for 'any' type is that because we have not done any deep dive as to see what is the top cuisine types are.