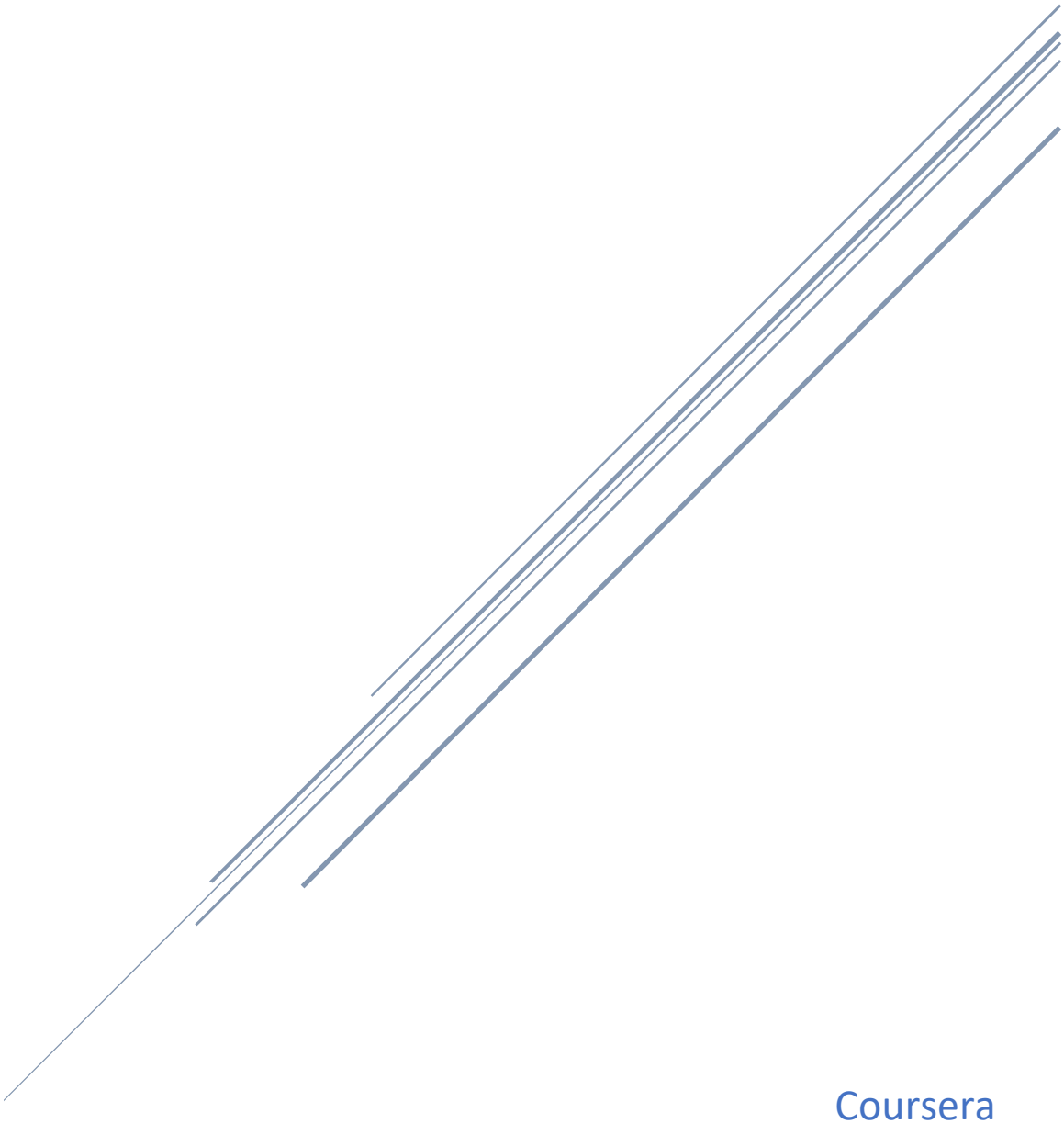
Identifying optimum locations for starting a Restaurant in Toronto, Ontario in Canada

Mohab O. Hassan



Coursera

Applied Data Science Capstone

Table of Contents

Table of Figures

Introduction

Background

Although the global food service market, defined as sale of food and beverages for immediate consumption, was worth US$ 3.4 Trillion in 2018 and is expected to reach a value of US $4.2 Trillion by 2024 [1], starting a new business in the restaurant industry nowadays is challenging. That's due to the existence of various competitive players in the market. In addition, there are large fast-food chains that expand easily from one location to another. According to the Foodservice Industry Forecast report in Canada, the commercial food service sales are expected to grow to US $77.5 billion in 2020 [2].

Many small businesses start and fail everyday due to the lack of prior market surveys and studies to understand the possible needs for the targeted business locations. For example, in the food service market, critical factors can affect the success or failure of a business depending on the geospatial location of such a business as well as the existing rivals. Thus, a proper market study is required to understand the success factors of starting a business in the restaurant industry.

Problem

It is required to identify the optimum location(s) for starting a new restaurant in Toronto, Ontario in Canada. In this context, the optimum location can have several meanings. For example, it could mean a location where there is minimum competition, or a location where there is a weak competition or. These two options will be addressed in this report.

Proposed solution

To address the problem of identifying the optimum location for starting a new restaurant, the two suggested definitions of the optimum locations need to be defined as in the following:

1) A location with a minimum competition means a location where the number of restaurants is low with respect to the number of people or the current population living there.
2) A location with a weak competition means a location where there are some restaurants, but their quality or rating is low.

Data Acquisition and Cleaning

Data Sources

The type of data as well as the sources are listed in Table (1) below.

Table 1 - Data types and sources

Data type	Source
Postal Codes, Boroughs and Neighbourhoods in Canada	Wikipedia [3]
Postal Codes, and geographical locations in Toronto, Ontario	Coursera (CognitiveClass.ai)

The population in each Neighbourhood in Canada in 2016	Canada Statistics [4]
The geographical boundaries for each Neighbourhood in Toronto, Ontario	Canada Statistics [4]
Venues' IDs, names, locations, distances from the corresponding Neighbourhood centre location, postal codes and categories type	Foursquare API
Restaurants' ratings, likes, checkins, price tier and tips	Foursquare API

The first two data sources are used to identify the geographical location and postal code of each neighbourhood in Toronto. The 2016 population data from Canada Statistics is then linked to these two data sources using the postal code in both data sets. The geographical boundaries data area used to facilitate the plotting of data in choropleth maps. Using the geographical location of each neighbourhood, Foursquare API is used to get all the restaurants with their details such as ratings, like count, check-ins, tip count and price tier.

It should be noted that the population data available online is that of 2016. So, this data has been used current data from Foursquare API due to unavailability of population data in 2020.

In the Foursquare API, all the venues in each neighbourhood were searched using the intent keyword “browse” and with a limit of “1000” and a radius of “1000” m. This resulted in all the venues within a 1 km circle radius in each neighbourhood. The number of restaurants were more than 500 which is the maximum number of premium calls that can be made with Foursquare API. So, it took more than one day to get the details for each restaurant.

Data Cleaning

From the Wikipedia table of postal codes that starts with “M”, all the unassigned boroughs have been removed, and all the unassigned neighbourhoods have been assigned to their corresponding boroughs. The total number of neighbourhoods was 103.

In the 2016 population data, the postal codes that starts with “M” were only chosen. However, one postal code (namely “M7R”) didn’t have a population data so it was removed. The total neighbourhoods with given population data became 102.

The geographical boundaries were obtained from Canada Stats website in the form of a .shp file. The procedure of converting this file into .geojson file using QGis[5] is well explained here [6].

Using the Foursquare API, all the venues for each neighbourhood was searched. Then, only the venues that had the keyword “Restaurant” in the venue category column were chosen. Since some neighbourhoods were close to each other (or to be specific, they were closer than 1000 m which was the radius indicated in the search query for venues using the Foursquare API), there

were duplicates in the restaurants. They were not many though. All the duplicates were removed. Moreover, not all the neighbourhoods had venues. At the end, there were 881 restaurants in 99 neighbourhoods in 9 boroughs.

Restaurant Details

The premium call (venue details) in Foursquare API allowed for getting details/features such as rating, likes and number of tips for each restaurant. However, there were many missing values in some of these features such as the ratings and price tier for instance. In the price tier, only 37 restaurants weren't assigned a price tier. By examining the other restaurants, it was found that the majority falls in the price tier "2" as shown in Figure (1). So the unassigned 37 restaurants were assigned to a price tier "2".

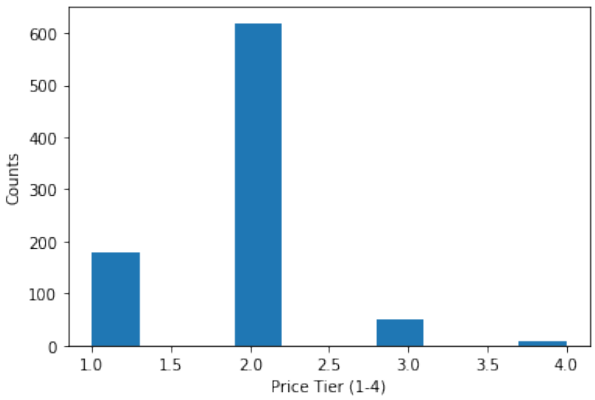


Figure 1 - Distribution of the Price Tiers in all the restaurants

In addition, more than 50% of the restaurants didn't have a rating. By looking at the distribution of those restaurants with given rating, it can be shown that the ratings are between 4.5 and 9.1 as shown in Figure (2). In order to get the missing values in the ratings, the other features of the restaurants such as likes count or tips count needed to be explored. A machine learning technique (namely single linear regression) was used on the known rating values with their corresponding likes count to investigate any correlation between them. As shown in Figure (3), the correlation was weak (~0.5) between the rating and likes. The R squared value of the trained data was very low as well (~0.25). So, it wasn't possible to estimate ratings values from the likes. The other features were also investigated but no good correlation was found. So, it was decided to fill the missing values of ratings with a random uniform distribution of values between the minimum and maximum given ratings as shown in Figure (4). This assumption is not accurate but it is done for the sake of this project only.

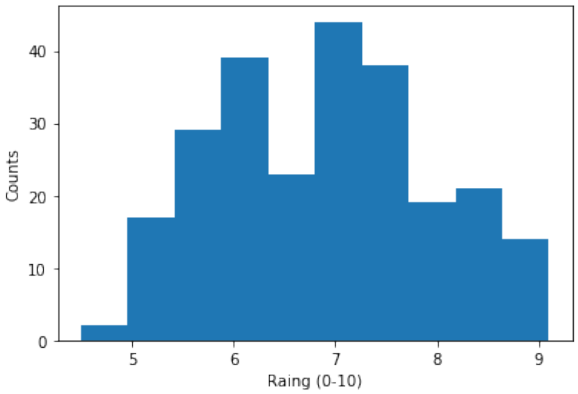


Figure 2 - Distribution of the available Ratings in the restaurants

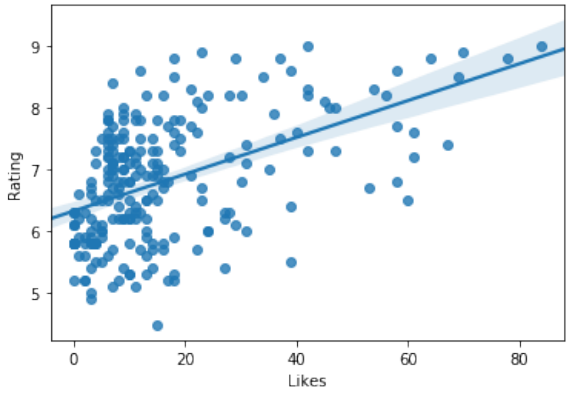


Figure 3 - Single Linear Regression between the Like Count and Rating

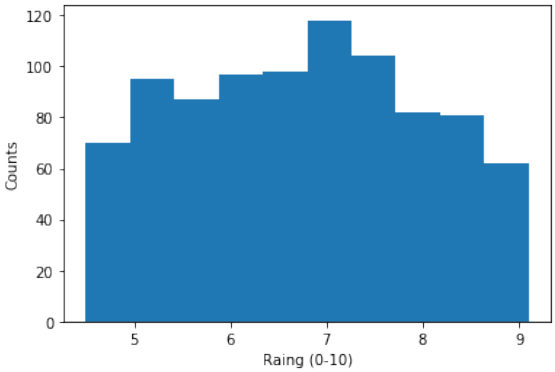


Figure 4 - Distribution of Rating in all restaurants after using a uniform random distribution for the missing values

Exploratory Data Analysis

Types of Restaurants

Looking at the 881-restaurant data gathered, it was found that there are 65 different type of restaurants (or in other words, different cuisines) as shown in the bar char in Figure (5). Notice, however, that the type of restaurant with the largest count is actually named just “Restaurant” which means it is unclassified or it doesn’t belong to a specific cuisine. So, any data analysis that would be done based on the cuisine will be biased. For this reason, all the restaurants will just be considered as generic restaurant without considering their cuisine.

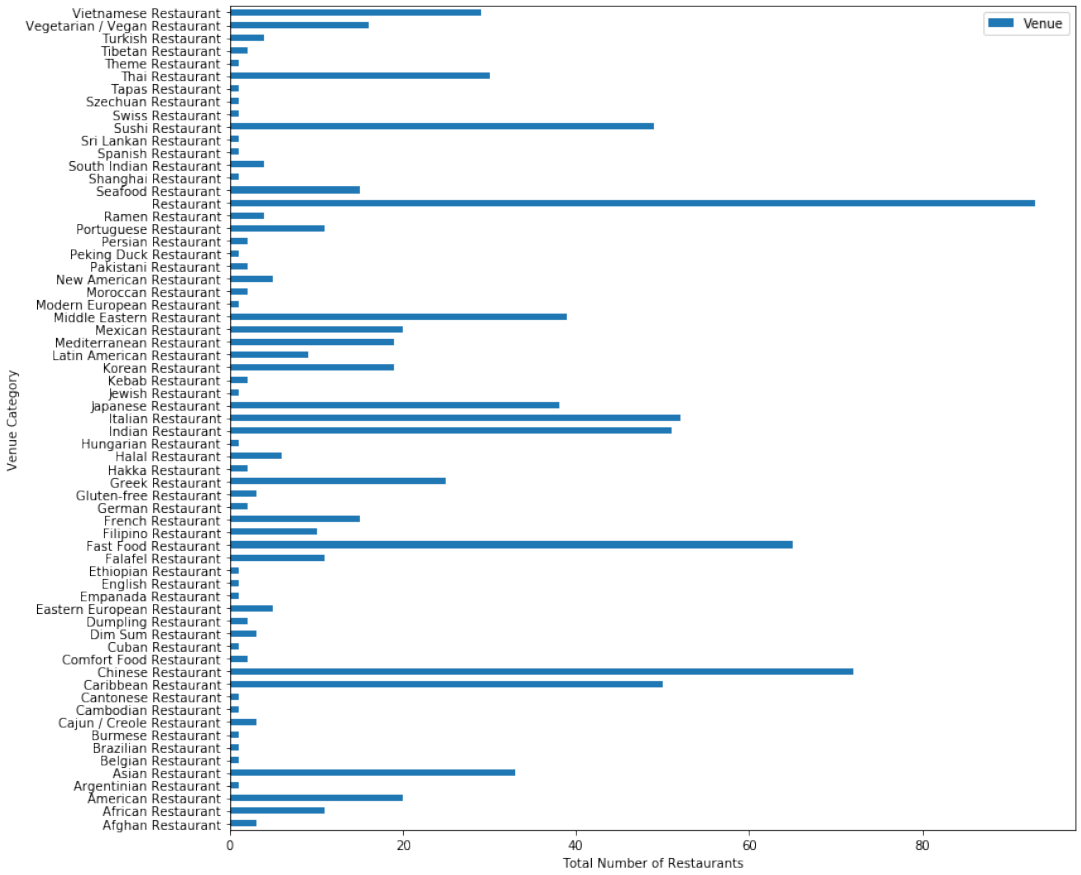


Figure 5 - Total number of restaurants in each restaurant category type

Number of Restaurants in each Neighbourhood

The number of restaurants in each neighbourhood was visualized on a choropleth map as shown in Figure (6). There are clearly distinct locations with many restaurants such as Downtown Toronto, Maryvale in Scarborough, Riverdale in East Toronto, and Bedford Park in North York.

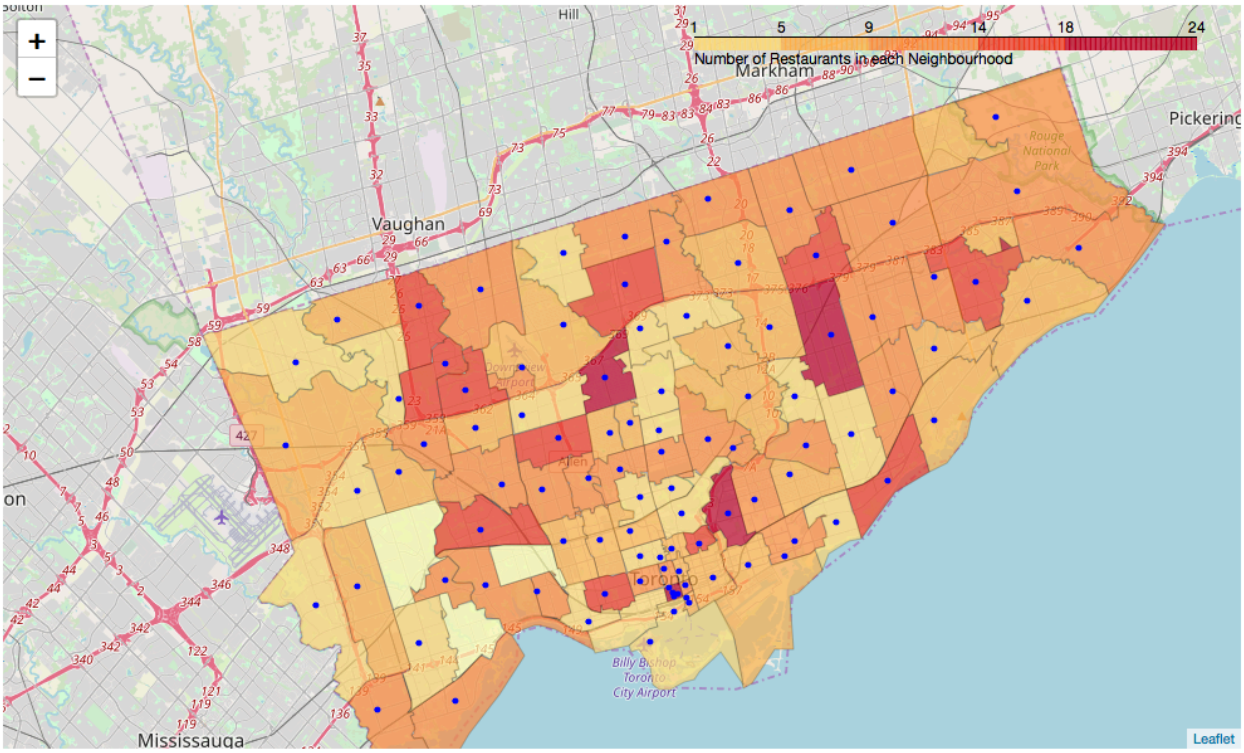


Figure 6 - Number of Restaurants in each Neighbourhood

Population of each Neighbourhood

The best way to show the population in each neighbourhood is by using again a choropleth map as shown in Figure (7) below. Surprisingly, the neighbourhoods near the Toronto Downtown area are less populated than those at the borders of Toronto such as Rouge, Malvern in Scarborough and Etobicoke. The reason is related to the fact that there are many neighbourhoods concentrated at the vicinity of the Toronto Downtown area. But there are only a few neighbourhoods in the boroughs at the border such as Rouge and Malvern.

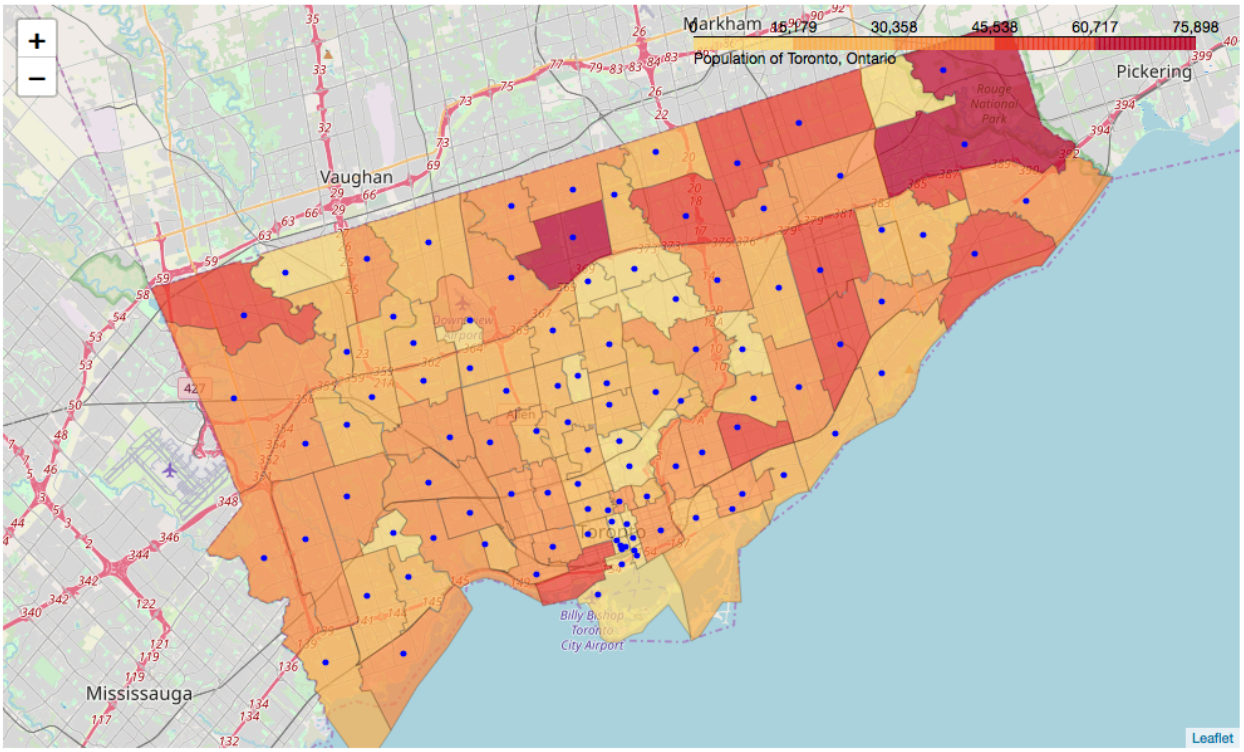


Figure 7 - Population in each neighbourhood in Toronto, Ontario

Results and Discussion

Lack of Restaurants in Neighbourhoods

A simple way to explore whether there are sufficient restaurants in each neighbourhood to serve all the people living there is by dividing the population number by the number of restaurants in each neighbourhood. This will give us a percentage that shows the need for more restaurants in each neighbourhood. This can be seen in the choropleth map in Figure (8). With a population of ~40,000 and only one restaurant, it looks like Willowdale West is very low on the number of restaurants needed to serve its large population. This could potentially be a very good place to start a new restaurant. It should be noted that there could be many other restaurants that actually exist in Willowdale West but are not in the Foursquare database.

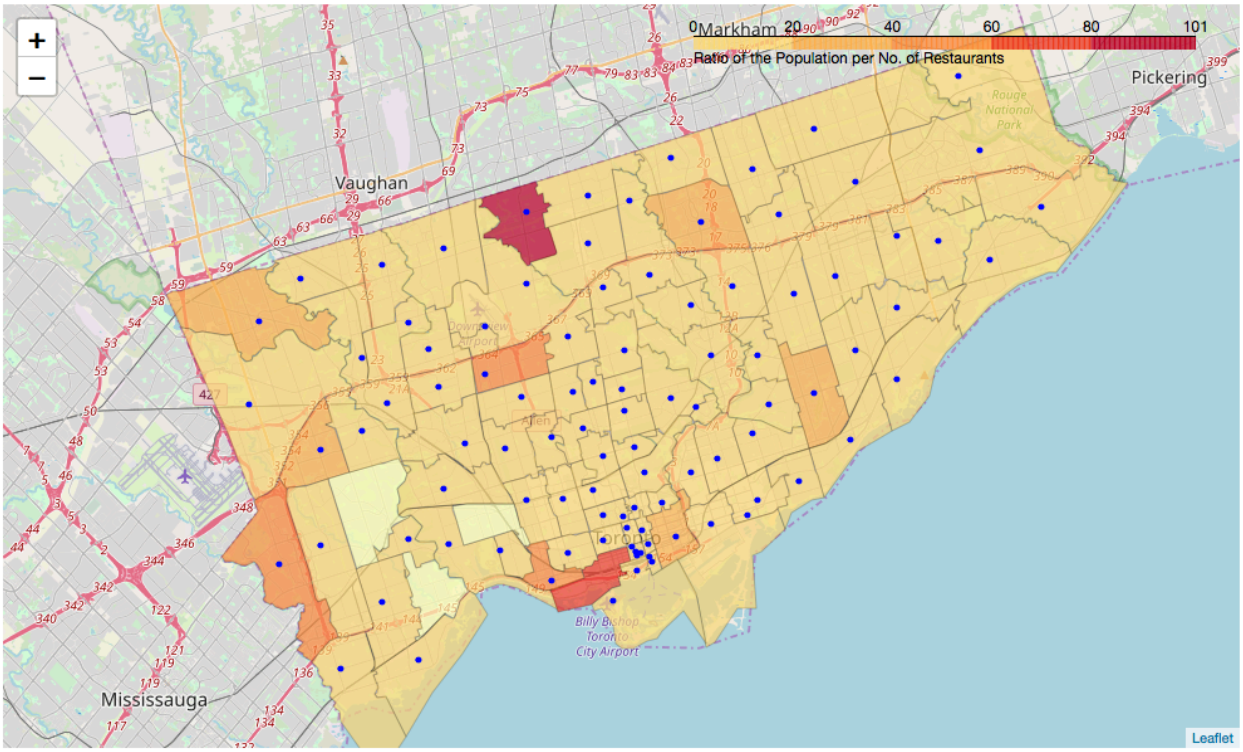


Figure 8 - Ratio of the population per the number of restaurants in each neighbourhood

Clustering the Restaurants

The DBSCAN machine learning tool was used to cluster the restaurants based on their location to identify high density locations of restaurant clusters all over Toronto. As shown in Figure (9), there is a big cluster of restaurants spanning from Downtown Toronto towards the east and north east. Several smaller clusters exist in the north east part. There are several outliers near the borders of Toronto.

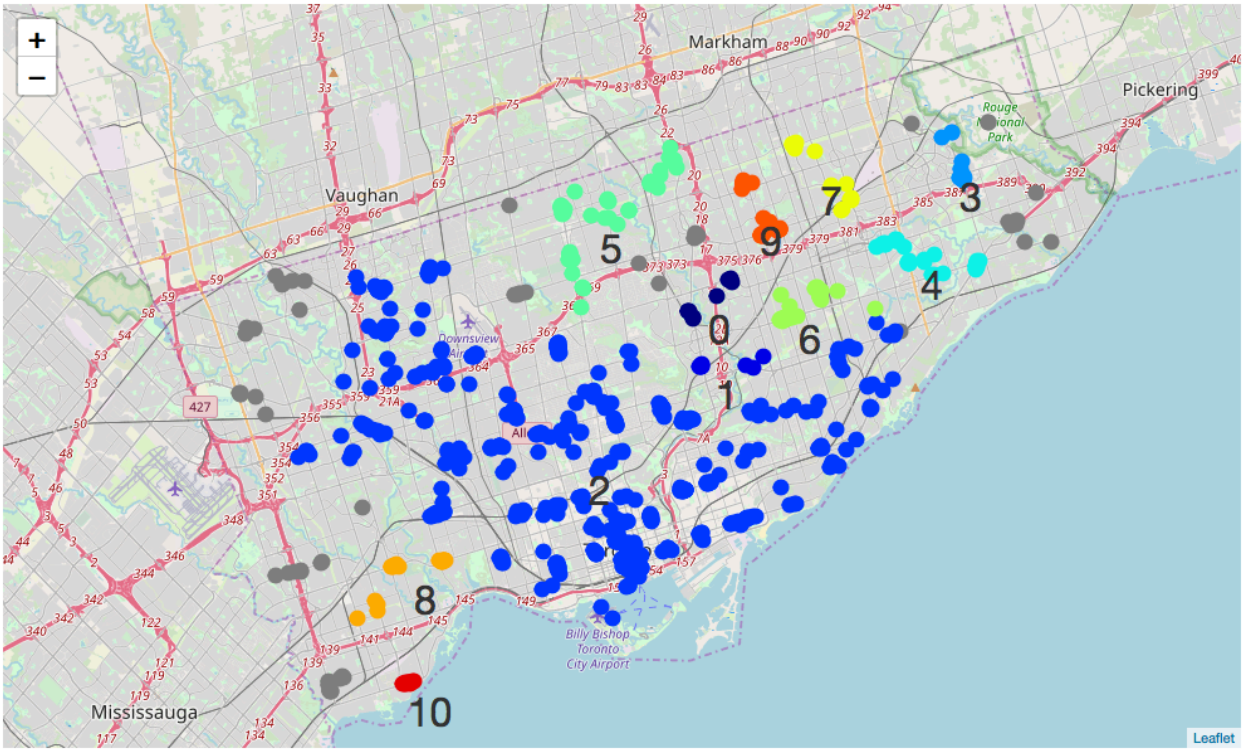


Figure 9 - Clustering of Restaurants using DBSCAN based on their locations

To get more meaningful insights, the restaurants were clustered based on their location in addition to their rating, likes, tips and price tier. Figure (10) shows 5 clusters based on the aforementioned features. The average value for each feature is tabulated in Table (2). The largest cluster #0 appears to be very similar to the largest cluster# 2 in Figure (9). Within the largest cluster #0, there is another smaller cluster#4 that is concentrated near the Toronto Downtown area. This is a highly competitive area because, as shown in Table (2), it has relatively a high like count, a high tip count and a low-price tier. On the contrary, cluster #3 seems to be a much less competitive area owning to the low average ratings, like count and tip count. Also, the price tier for cluster #3 is 2.0 which is nominal among all the restaurants, but it can be lower. Cluster #3 is located near the Scarborough Centre area which can be ideal to start a new restaurant with a high quality and an affordable menu.

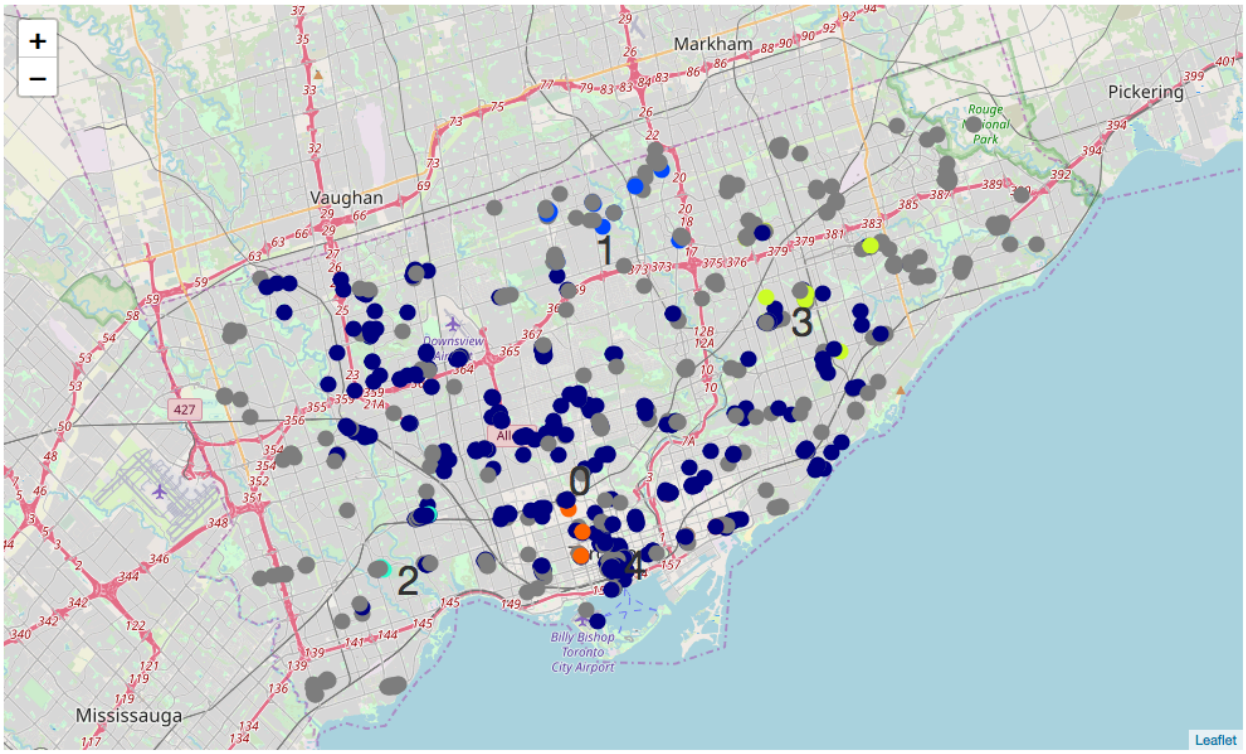


Figure 10 - Clustering of Restaurants using DBSCAN based on their locations, Rating, Like Count, Tip Count and Price Tier

Table 2 - Clusters' values for the DBSCAN in Figure (10)

Cluster #	Avg Rating	Avg Like Count	Avg Tip Count	Avg Price Tier
0	6.83	1.51	1.03	2.0
1	6.49	1.4	0.53	2.0
2	7.42	0.0	0.0	2.0
3	5.12	0.6	0.3	2.0
4	6.92	1.9	1.2	1.0

Conclusions

We explored different datasets from Foursquare and Canada Statistics to identify optimum locations to start a new restaurant. We defined the optimum locations as either those that have very low number of restaurants compared to their population, or those that have some restaurants, but they are of low quality/unpopular. The results suggested that Willowdale West in North York has relatively a very few restaurants compared to its population. Also, the restaurants located near Scarborough Centre area have relatively low ratings and are less liked by their visitors. Consequently, these two locations are considered ideal to start a new restaurant.

References

[1] https://www.prnewswire.com/news-releases/global-food-service-market-report-2019-2024-market-is-expected-to-reach-a-value-of-us-4-2-trillion-300907559.html
[2] https://www.restaurantscanada.org/resources/foodservice-industry-forecast/#preview
[3] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
[4] https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2016-eng.cfm
[5] http://qgis.com
[6]https://medium.com/dataexplorations/generating-geojson-file-for-toronto-fsas-9b478a059f04