Capstone Project - The Battle of Neighborhoods

Identify an ideal Neighborhood to start a Restaurant in Toronto

By

Abhijith Antony

Contents

Chapter 1 Introduction

1.1 Background

Investing in a restaurant can be a very risky venture without a proper plan and strategy. Of all the decisions to be made the location of the restaurant is one of the most important one. Some of the factors to be considered while deciding the location is the competition in a neighborhood and the footfall of people in an area. The popularity or footfall of a place can be judged by the number of stores and avenues in an area and that is the basis of the study. We have a lot of information online related to different avenues in an area, in this study we will be using Foursquare API to fetch these details from all neighborhoods in Toronto and decide on neighborhoods to invest in.

1.2 Goal of the Thesis

Goal of the thesis is to identify an ideal neighborhood in Toronto where we can invest in, to open a restaurant. We will leverage the information available online regarding restaurants and other stores in an area and use K means clustering to divide the region into different areas with different degrees of competition. This information will help us judge which are the best neighborhoods to invest in i.e. areas with least competition and comparably larger footfall. We assume that there is a high footfall of people in an area based on presence of other avenues.

Chapter 2 Data

2.1 Data Gathering

In this project, we will gather the required information of neighborhoods in Toronto by Web scraping the Wikipedia page:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This has the Postal code, borough, and neighborhood details from Toronto. We will use the BeuautifulSoap package to perform web scraping and fetch only the required details from the Wikipedia page (we will also use an HTM parser to the read the page).

Along with the neighborhood details to properly plot the clusters and to get a map view of neighborhoods we need the latitude and longitude information of each corresponding neighborhood. We would be using geocoder API package to fetch this information.

Once the basic Neighborhood information along with its corresponding latitude and longitude is available next step is to fetch information regarding all the avenues like restaurants, grocery stores, salons in the area. To achieve this information, we will use the Foursquare API and fetch details of all stores and confectionaries in a radius of 1000m around the neighborhood. Foursquare API will take the postal code information we fetched from the Wikipedia page as input and return different Avenues circling the area.

2.2 Data Cleaning

It is important that we have the data in clean and standard format before proceeding to further analysis. Data cleaning steps followed in the project:

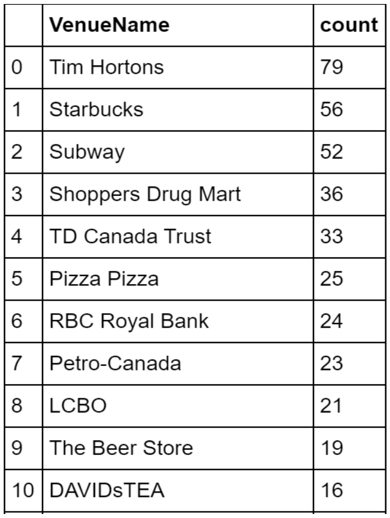
- Remove all the *'Not assigned'* neighborhoods from the list.
- More than one neighborhood can exist in one postal code area, these different neighborhoods must be combined in a single row, ensuring unique postal codes.
- Filter out avenues with *'Restaurant'* in its name.

Once these steps are followed, we can proceed to the Exploratory data analysis step to understand the data further.

Chapter 3 Exploratory Data Analysis

3.1 Visualization

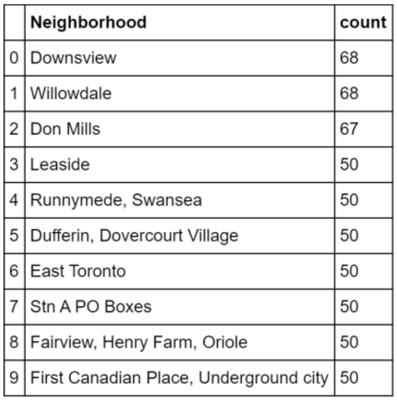
Top 10 avenues in Toronto:



	VenueName	count
0	Tim Hortons	79
1	Starbucks	56
2	Subway	52
3	Shoppers Drug Mart	36
4	TD Canada Trust	33
5	Pizza Pizza	25
6	RBC Royal Bank	24
7	Petro-Canada	23
8	LCBO	21
9	The Beer Store	19
10	DAVIDsTEA	16

Figure 1: Top 10 Avenues

Top 10 Neighborhoods in Toronto (With highest number of avenues):



	Neighborhood	count
0	Downsview	68
1	Willowdale	68
2	Don Mills	67
3	Leaside	50
4	Runnymede, Swansea	50
5	Dufferin, Dovercourt Village	50
6	East Toronto	50
7	Stn A PO Boxes	50
8	Fairview, Henry Farm, Oriole	50
9	First Canadian Place, Underground city	50

Figure 2: Top 10 Neighborhoods

Observations:

- Top Avenue: Tim Horton
- Top Neighborhood: Downsview and Willowdale (68)
- Neighborhood with least number of Avenues: Rouge Hill, Port Union, Highland Creek (1)

Understanding these different neighborhoods will help us in our final decision-making process.

3.2 Neighborhood Map

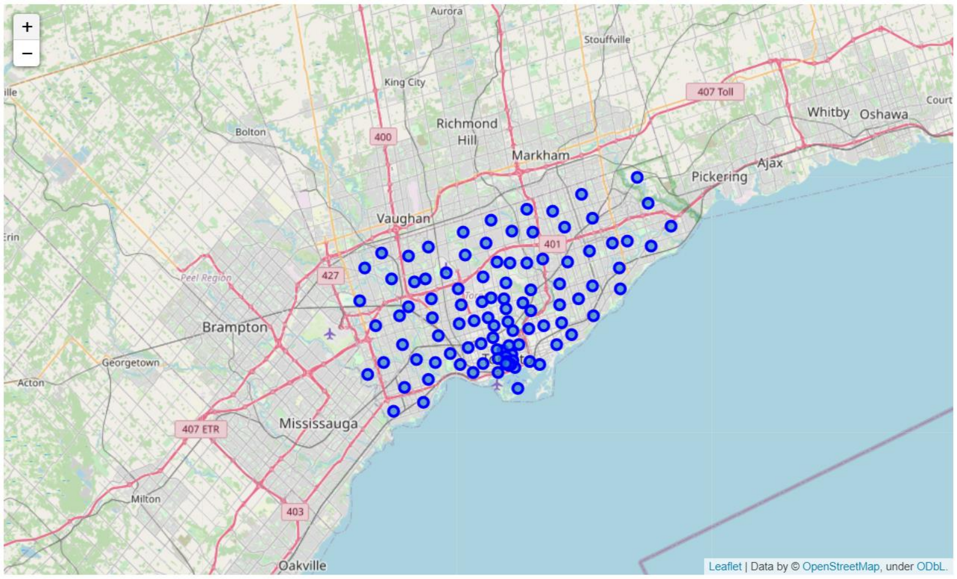


Figure 3: Neighborhood location

Chapter 4 K Means Clustering

There are many models for clustering out there. In this project, we will be presenting the model that is considered one of the simplest models amongst them. Despite its simplicity, the K-means is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data**.** We will be grouping the neighborhoods based on the number of restaurants in each of them and then cluster based on this count information. This will give us different segments based on the number of restaurants in an area helping us to assess competition in a region.

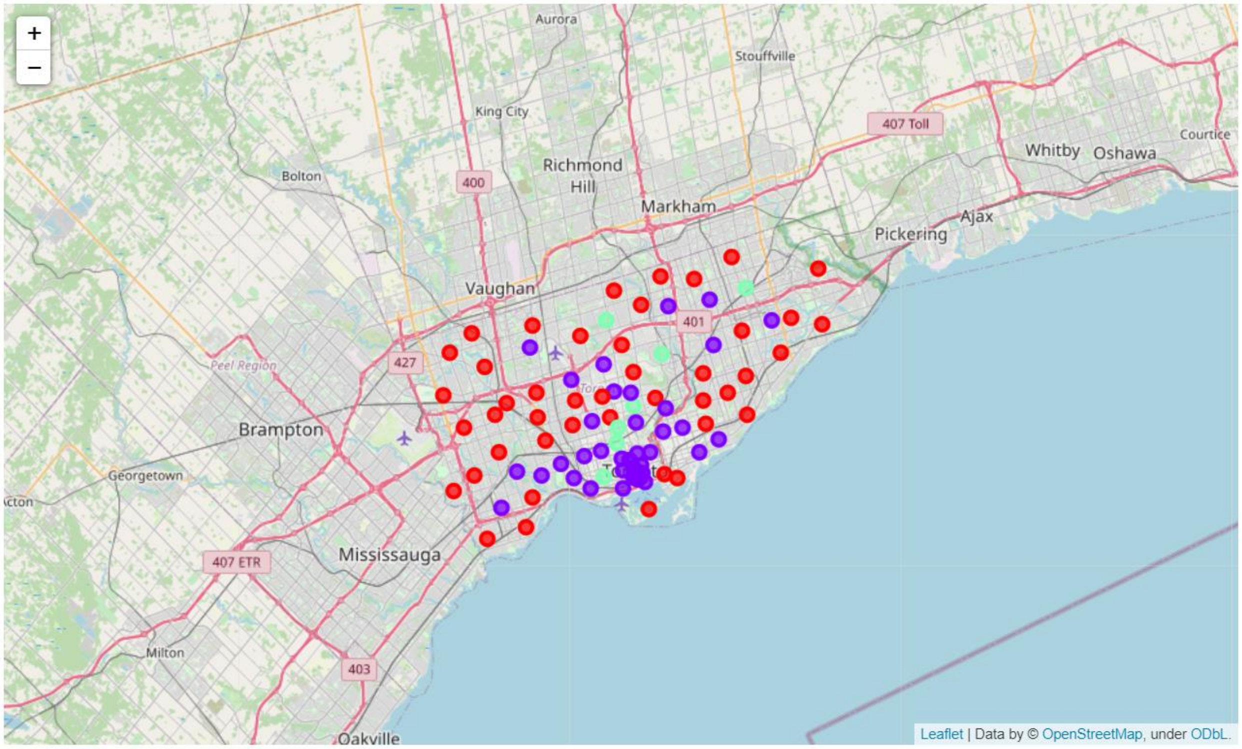


Figure 4: Clustering Results

Observations from the clustering results:

- Red: Cluster 0: With least number of restaurants in the surroundings.
- Green: Cluster 1: With Average number of restaurants in the surroundings.
- Blue: Cluster 2: With Maximum number of restaurants in the surroundings (Will be the most competitive regions).
- We see that the least amount of competition is faced by the red or cluster 0 is mostly on the outskirts of the city. But we do see few neighborhoods closer to city coming in this segment which can be potential investment opportunity.

Chapter 5 Recommendation and Conclusion

- We see that Cluster 0 is the least competitive and in cluster 0 neighborhood Woodbine Heights has the maximum footfall and other avenues in the region hence this would be the prime target followed by Regent Park, Harbourfront, Leaside, Runnymede and The junction North neighborhoods.
- In cluster 1 Downsview is most optimum neighborhood to invest in, this has the highest number of avenues in the region and since its in cluster1 there is only average competition which can be leveraged here.
- Cluster 2 is highly competitive with already a very high number of restaurants in the region hence I would not recommend investing here for the time being unless we have any further information which might change the current scenario.