To find the best location to set up a Pizzeria

in the Toronto Neighborhood

By Joseph Lwere

July 2021

1.Introduction

1.1 Background

Maestro foods is the leading pizzeria business in North and Central America. This leading pizzeria business wants to extend line to Toronto the number one city in Ontario Province in Canada and a major city along the Great lakes region of North America. Maestro foods wants to know the best possible location to set up a new pizzeria business before investing. They are look at a strategic location that will lead to profitable venture and a worthy investment.

1.2 Problem Statement

In this project we want to find the best possible location in Toronto to set up a pizzeria that will kick off with good sales. This location will be one that is strategically located in the city and one capable of attracting people to dine out and enjoy the long list of pizza varieties.

2. Data Collection and Cleaning

2.1 Data Collection

We will open up the pizzeria to have the capacity to supply pizzas using food vendors and online delivery orders system from online stores and a restaurant. We will also want to know the number of nearby restaurants within a 5km radius so we use the neighborhood to determine the number of potential vendors. For orders to be delivered, the volume ordered for each

neighborhood's will be considered. The volume of orders is determined by the number of people (population) and average house hold income.

2.2 Data Sources:

- Location data - Wikipedia Website
- Restaurants in the neighborhood - Foursquare Website
- Socioeconomic data (population size by neighborhood, average after tax household income by neighborhood) - Toronto Wellbeing page

3. Method of Work

1. Import data sets and combine them into a single data frame.
2. We then use the Foursquare API and determine the number nearby restaurants within a 5km radius. We get a list of restaurants and then appended to the combined data frame.
3. To identify the best possible location, neighborhoods will be clustered into different regions via k-means clustering.
4. We then generate the statistical data of each cluster.

4. Scope of Work

4.1 Web Scraping Data

To get the the whole data layout about the Toronto Neighborhood, data is downloaded and scraped from Wikipedia. After we download that piece of data, we add more columns to the scraped data. Data added includes coordinates and postal code.

4.2 Download the Social Economic Data

From the Toronto well being website we import the social economic data of Toronto neighborhoods. We are looking for the livelihoods of these neighborhoods because this information is very vital before setting up the business. We want to know the average household income of the people in these neighborhoods. We are also looking at finding out

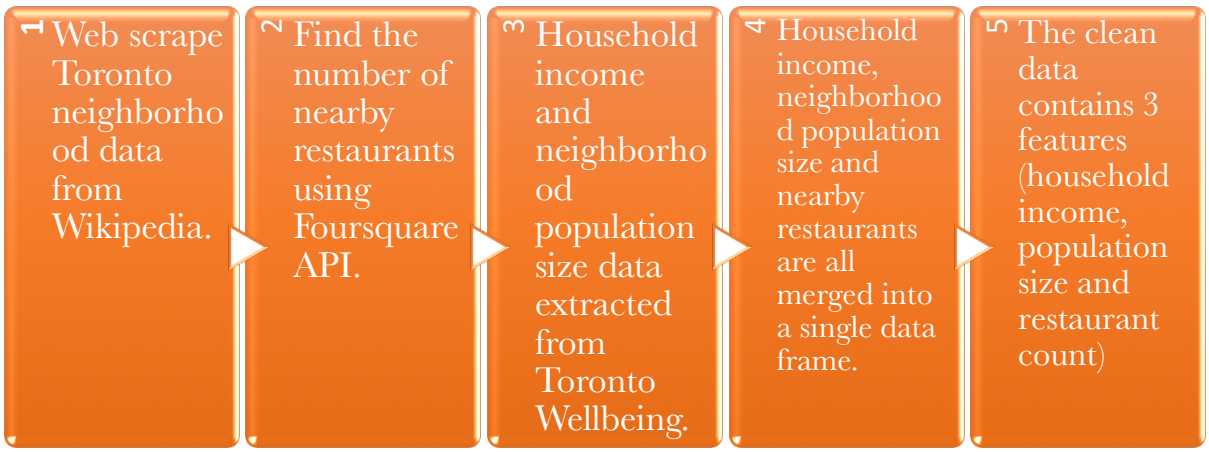
the population size of these neighborhoods as this is vital information before opening the pizzeria business.

4.3 Find out the number of Restaurants

We use the Four Square website to find out the number of restaurants in the Toronto neighborhood. This information will give us a clearer picture of where other restaurant business owners place their shops. Which neighborhoods have the highest number of restaurants and then try to figure out why. Is it because of the social economic well being of the population in those neighborhoods?

We also use four square to determine the number of potential vendors in the different neighborhoods. One of the determinants of a successful pizzeria business is the number of sales that go out every single day. So getting information about vendors we get a clearer picture of another avenue to make sales from the pizzeria. We are also relying on vendors because then we don't have to get a big space for the restaurant.

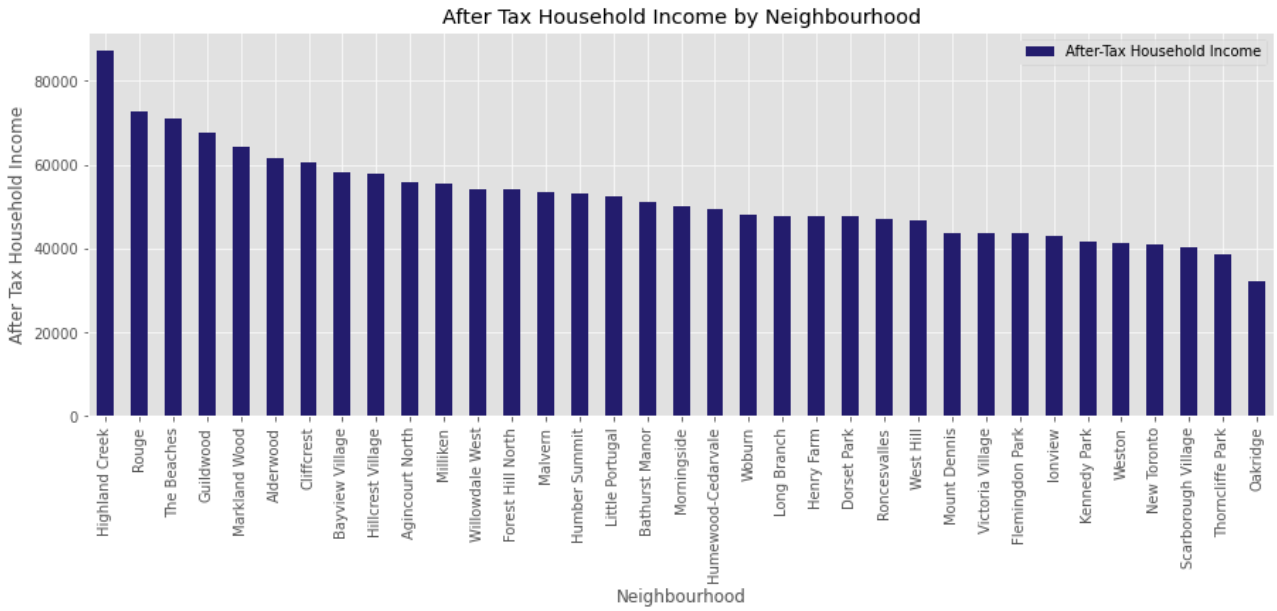
The figure below shows the scope of work in a nutshell.



5. Data Analysis

5.1 Household Income

On the data analysis stage we start by finding the relationship between the household income and the different neighborhoods. Quickly we find that *Highland Creek* neighborhood has the highest income earner as seen in the graph below.

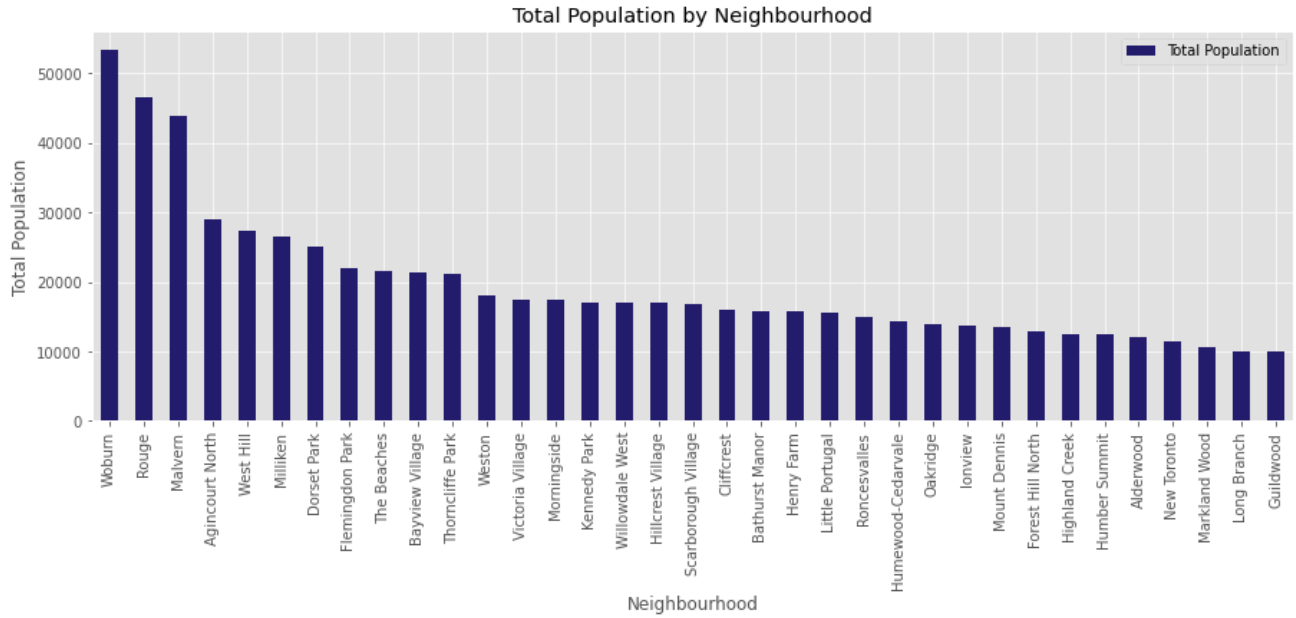


However, household income after tax in it's self is not enough to determine where we set up the pizzeria. We must find more valid reasons to inform our decision because getting started.

Now that we don't want to use household income after tax we move to the next available information.

5.2 Populations Figures of the Neighborhoods

Let's consider population of the Toronto neighborhoods. From the graph below we note that *Woburn* has the highest population. What we see from the first two graphs is we cannot base on one piece of information to make the decision. As we see *Highland Creek* neighborhood that had the highest income earners has one of the lowest population. And *Woburn* that has the highest population from the sample had an average level in term of income earners. The question is, do we then base on this two great information to make the decision?

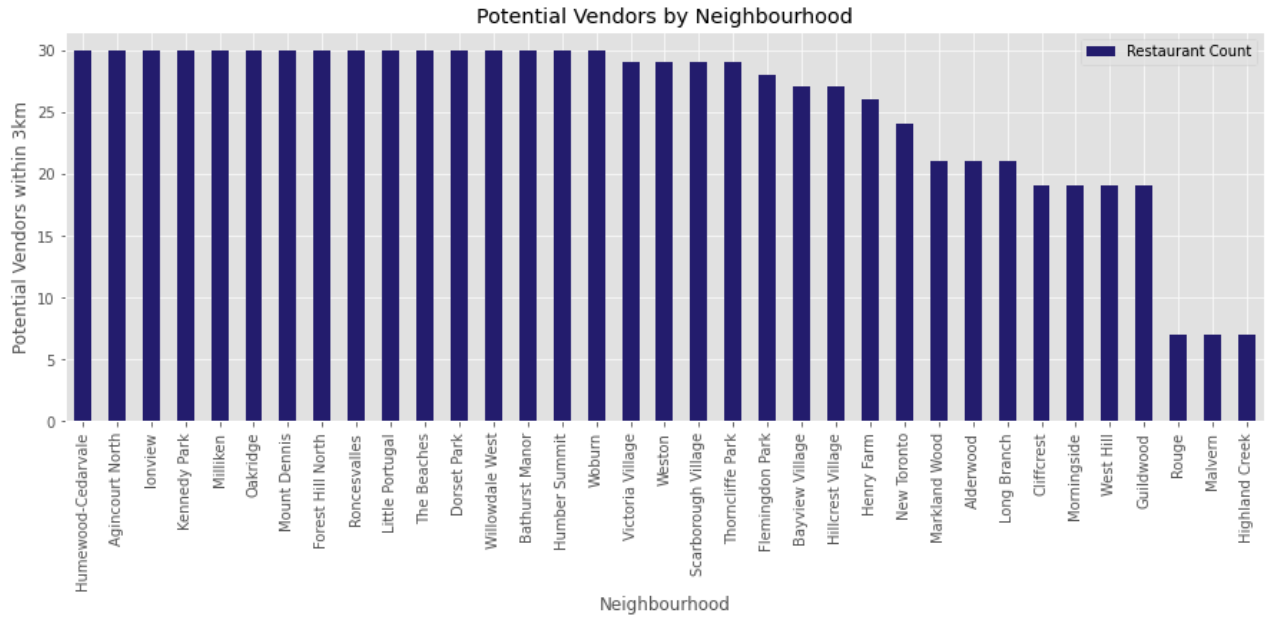


Looking at income levels and populations of the different neighborhoods in Toronto city, gives us vital information but not enough to make the decision. We will need to consider another source of information to make an informed decision.

5.3 Food Vendors

Looking at the scope of work one vital information that we can add to household income after tax and population figures of the Toronto neighborhoods is the food vendors that are existing in the different neighborhoods. As we are looking at food vendors to upscale our sales this is vital information that we have to find out. It's is said that 2.5 billion people buy food from food vendors daily which is an average of 100 million people every hour.

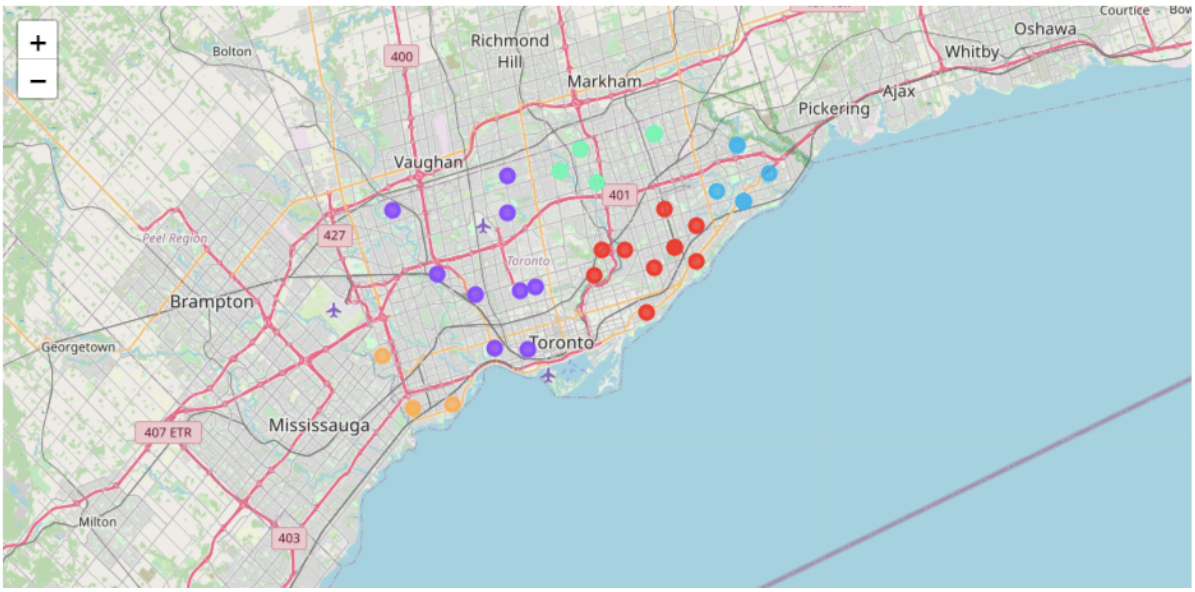
The figure below shows the potential vendors in Toronto neighborhoods. What we learn is *16 neighborhoods* have up to *30 food vendors* and only *three neighborhoods* have less than *10 vendors*. The question still remains would we use the information we have got so far to determine which neighborhood to set up the pizzeria?



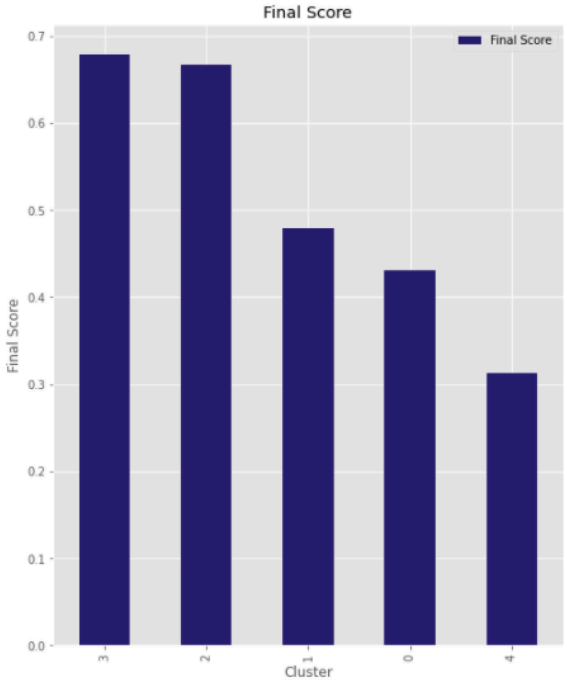
6.0 K-Means Clustering

6.1 Normalisation

Next we use k-mean clustering to normalize each of the categories that we have looked at. We normalize household income after tax, neighborhood population and number of vendors in a neighborhood. With k-means we try to minimize the intra-cluster distances and maximize the inter-cluster distances. Because we have a big number of neighborhoods, we cluster them to find the cluster with the highest likelihood of meeting our target. The map of Toronto below shows the different clusters of neighborhoods.



From the map we see that we have *five* different clusters. We analyze the best cluster by normalizing each individual cluster until we have a good decision on which neighborhood meets the desired decision. After normalizing each individual cluster we find out that the cluster with these neighborhoods has the highest score in meeting our goals of setting up a pizzeria.



Cluster 3 has the highest likelihood of setting up a profitable pizzeria business. The neighborhoods in that cluster include; *Miliken, Agincourt North, Henry Farm, Bayview Village* and *Hillcrest Village*.