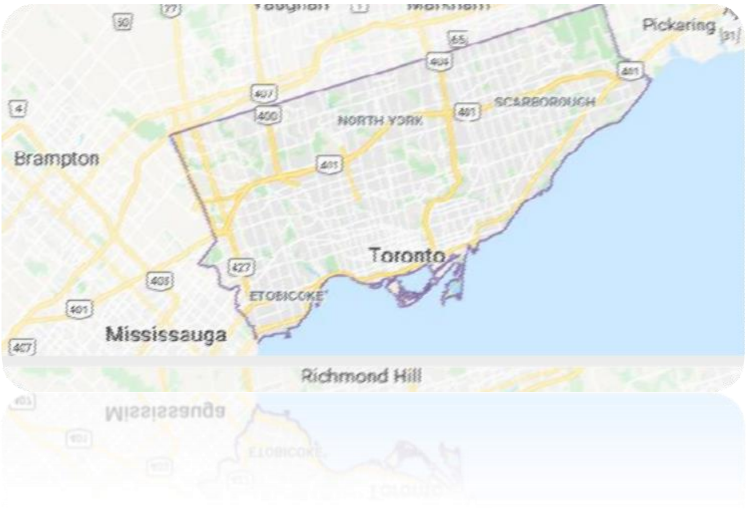
# Coursera Capstone IBM

# Applied Data Science Capstone

### The Battles of Neighborhoods

***Recommendation on the Toronto neighborhoods for XYZ grocery store***



## 1.0 Introduction:

Toronto is a well-developed capital city of Canada, with lots of business opportunities and business friendly environment, it has no issue in attracting many different players into the market. However, that also means the market is highly competitive and as a well-developed city, the cost of doing business is also one of the highest in the country. And thus, any new business venture or expansion in the country needs to be reviewed carefully and strategically targeted so that the return on investment will be sustainably reasonable and more importantly the investment can be considerably less riskier.

## 1.1 Problem Description:

This is clearly a problem that a grocery retailer (i.e. XYZ Grocery) needs to review and resolve as part of their new business venture in the country. As a startup though well-funded, they need to choose their first starting location in the country carefully for the points highlighted above and more importantly, if this is successful, the location should allow them to replicate the same success fairly quickly; so, first mover advantage is critical for this business and thereby the choice of location (i.e. neighborhood) is also important to them.

## 1.2 Target Audience:

To solve this problem, data scientist team led by myself has been engaged by XYZ Grocery. The objective is to locate and recommend to the management which region of the neighborhoods in Toronto will be the best choice to start off their first grocery offering including online capability and delivery services. The management also expects to understand the rationale of the recommendations in the final report.

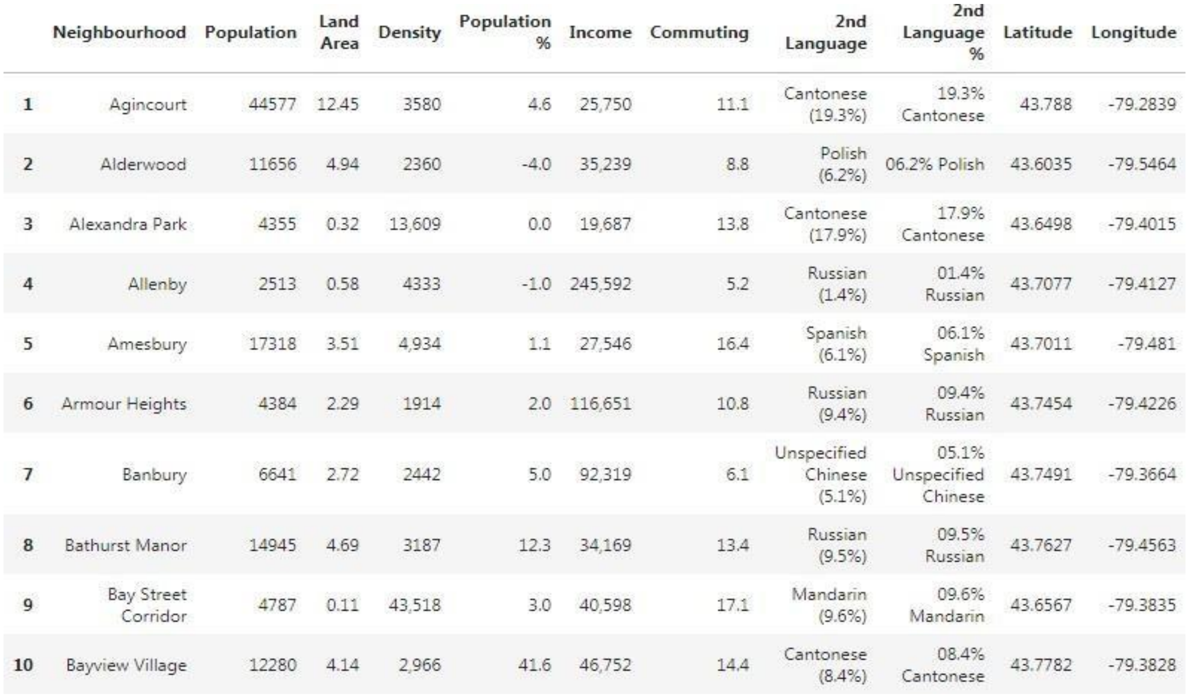
## 1.3 Success Criteria:

The success criteria of this project will be a good recommendation of the neighborhoods choice to the management of XYZ Grocery based on 2 key factors; lack of grocery stores available (less competition) and higher number of residences presented (higher demand) and it should allow easy replication of the business model (similarities among the neighborhoods).

## 2.0 Data Description:

As we need to explore, segment, and cluster the neighborhoods in the city of Toronto, the Toronto neighborhoods data is key for this project. Unfortunately, the data is for the Toronto neighborhood data is not widely available on the Internet in the structured format, hence we need to scrap it through an existing Wikipedia page exists that has all the information we need to explore and cluster the neighborhoods in Toronto. The data should contain the coordinates for each of the neighborhood in Toronto that will help us to further obtaining more information critical for this project. We will also like to obtain the key information like below; such as number of residences information for each neighborhood which is one of the key factors for the neighborhood of choice in the final report. The data needs to be clean up and eventually in a structured format like the example below.

1. Neighborhood Name
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Number of residences in each neighborhood.

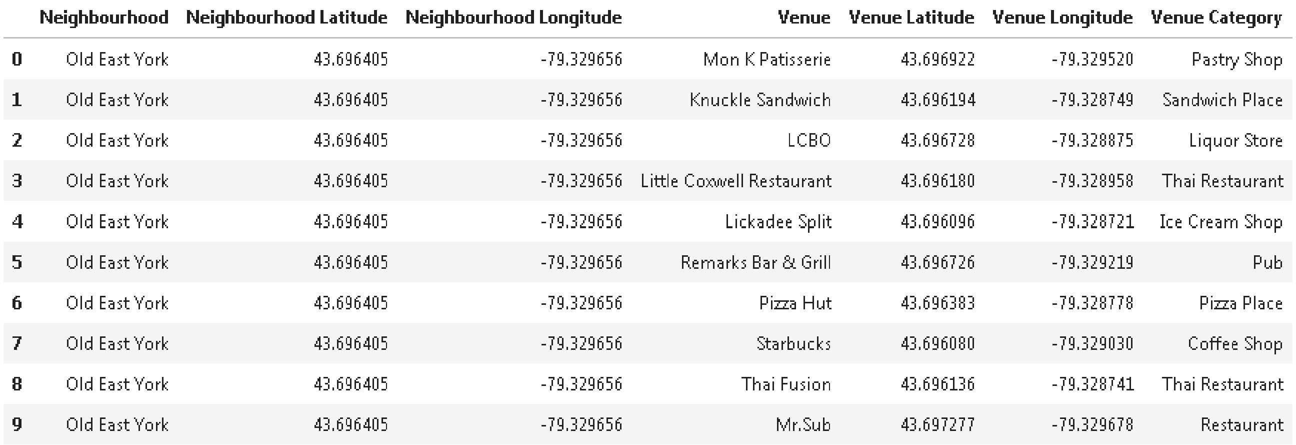


| | Neighbourhood | Population | Land Area | Density | Population % | Income | Commuting | 2nd Language | 2nd Language % | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Agincourt | 44577 | 12.45 | 3580 | 4.6 | 25,750 | 11.1 | Cantonese (19.3%) | 19.3% Cantonese | 43.788 | -79.2839 |
| 2 | Alderwood | 11656 | 4.94 | 2360 | -4.0 | 35,239 | 8.8 | Polish (6.2%) | 06.2% Polish | 43.6035 | -79.5464 |
| 3 | Alexandra Park | 4355 | 0.32 | 13,609 | 0.0 | 19,687 | 13.8 | Cantonese (17.9%) | 17.9% Cantonese | 43.6498 | -79.4015 |
| 4 | Allenby | 2513 | 0.58 | 4333 | -1.0 | 245,592 | 5.2 | Russian (1.4%) | 01.4% Russian | 43.7077 | -79.4127 |
| 5 | Amesbury | 17318 | 3.51 | 4,934 | 1.1 | 27,546 | 16.4 | Spanish (6.1%) | 06.1% Spanish | 43.7011 | -79.481 |
| 6 | Armour Heights | 4384 | 2.29 | 1914 | 2.0 | 116,651 | 10.8 | Russian (9.4%) | 09.4% Russian | 43.7454 | -79.4226 |
| 7 | Banbury | 6641 | 2.72 | 2442 | 5.0 | 92,319 | 6.1 | Unspecified Chinese (5.1%) | 05.1% Unspecified Chinese | 43.7491 | -79.3664 |
| 8 | Bathurst Manor | 14945 | 4.69 | 3187 | 12.3 | 34,169 | 13.4 | Russian (9.5%) | 09.5% Russian | 43.7627 | -79.4563 |
| 9 | Bay Street Corridor | 4787 | 0.11 | 43,518 | 3.0 | 40,598 | 17.1 | Mandarin (9.6%) | 09.6% Mandarin | 43.6567 | -79.3835 |
| 10 | Bayview Village | 12280 | 4.14 | 2,966 | 41.6 | 46,752 | 14.4 | Cantonese (8.4%) | 08.4% Cantonese | 43.7782 | -79.3828 |

## 2.1 Data Features:

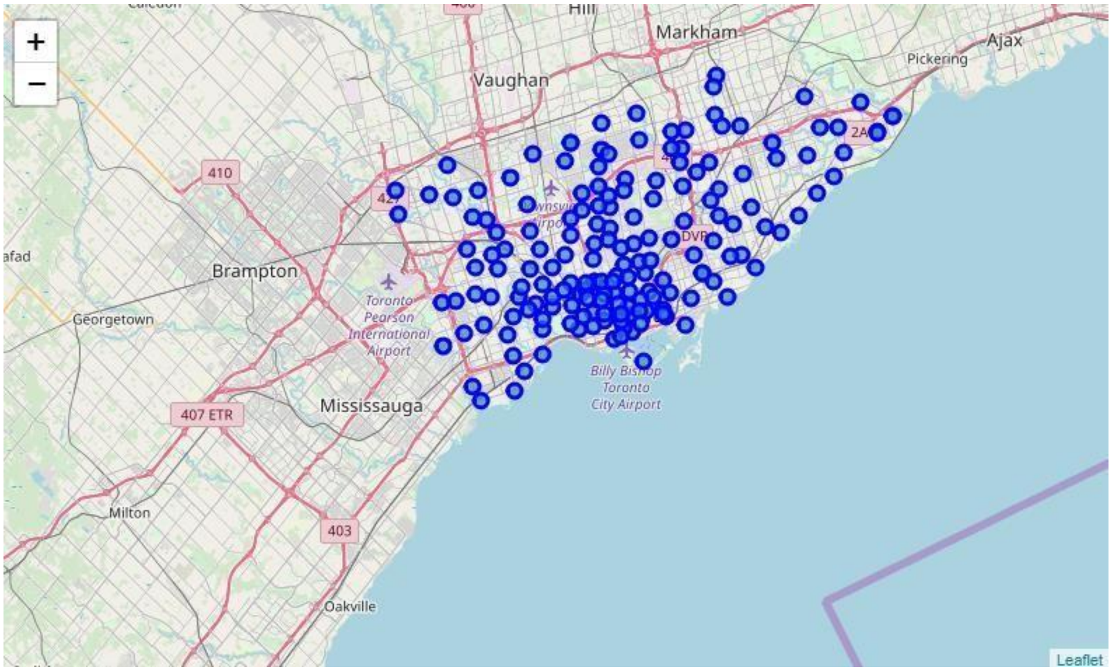
We will be leveraging on features in a reliable location information provider such as the Foursquare.com to explore the various types of venues and its categories available in each neighborhood. We will also need to understand the type of these venues nearby (i.e. within 500M) in each of the respective neighborhood. The information obtained per neighborhood will be as such like below and has to be in a structured format so to allow for further computation:

1. Neighborhood
2. Neighborhood Latitude
3. Neighborhood Longitude
4. Venue Name
5. Venue Category
6. Venue Latitude
7. Venue Longitude



| | Neighbourhood | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Old East York | 43.696405 | -79.329656 | Mon K Patisserie | 43.696922 | -79.329520 | Pastry Shop |
| 1 | Old East York | 43.696405 | -79.329656 | Knuckle Sandwich | 43.696194 | -79.328749 | Sandwich Place |
| 2 | Old East York | 43.696405 | -79.329656 | LCBO | 43.696728 | -79.328875 | Liquor Store |
| 3 | Old East York | 43.696405 | -79.329656 | Little Coxwell Restaurant | 43.696180 | -79.328958 | Thai Restaurant |
| 4 | Old East York | 43.696405 | -79.329656 | Lickadee Split | 43.696096 | -79.328721 | Ice Cream Shop |
| 5 | Old East York | 43.696405 | -79.329656 | Remarks Bar & Grill | 43.696726 | -79.329219 | Pub |
| 6 | Old East York | 43.696405 | -79.329656 | Pizza Hut | 43.696383 | -79.328778 | Pizza Place |
| 7 | Old East York | 43.696405 | -79.329656 | Starbucks | 43.696080 | -79.329030 | Coffee Shop |
| 8 | Old East York | 43.696405 | -79.329656 | Thai Fusion | 43.696136 | -79.328741 | Thai Restaurant |
| 9 | Old East York | 43.696405 | -79.329656 | Mr.Sub | 43.697277 | -79.329678 | Restaurant |

## 3.0 Methodology



Data scrapping from the Wikipedia page that contains the up-to-date population statistics of Toronto neighborhoods has been used. This is critical to understand the population of each Toronto neighborhood which is one of the key elements in the neighborhood of choice in this project.

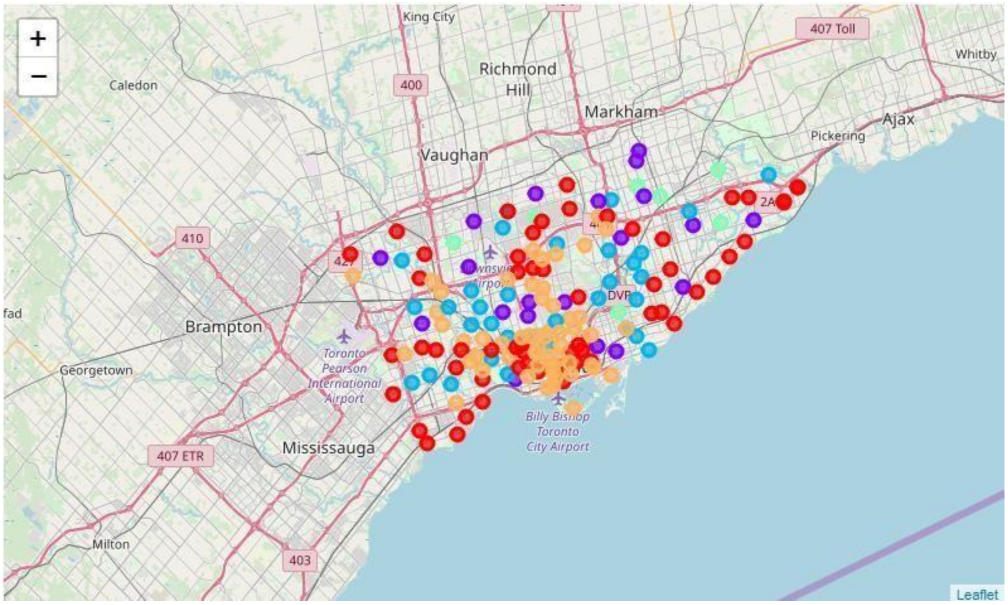
Furthermore, we need to know the coordinates and locations of the neighborhoods, and therefore the Geocoder API has been used for achieving this objective. This is important so that we can input this information into the location information provider such as Foursquare.com to obtain venue information in these neighborhoods, and this is precisely what we have done for it in this project.

We will also use machine learning techniques such as the K-Means to segment and cluster these neighborhoods so that we can group them together to understand their similarities. This is critical as we need to recommend to the management the regions of the neighborhoods of the choice in our recommendation so that XYZ Grocery can easily replicate their business model across multiple neighborhoods of similarities easily and quickly as part of their business growth plan.

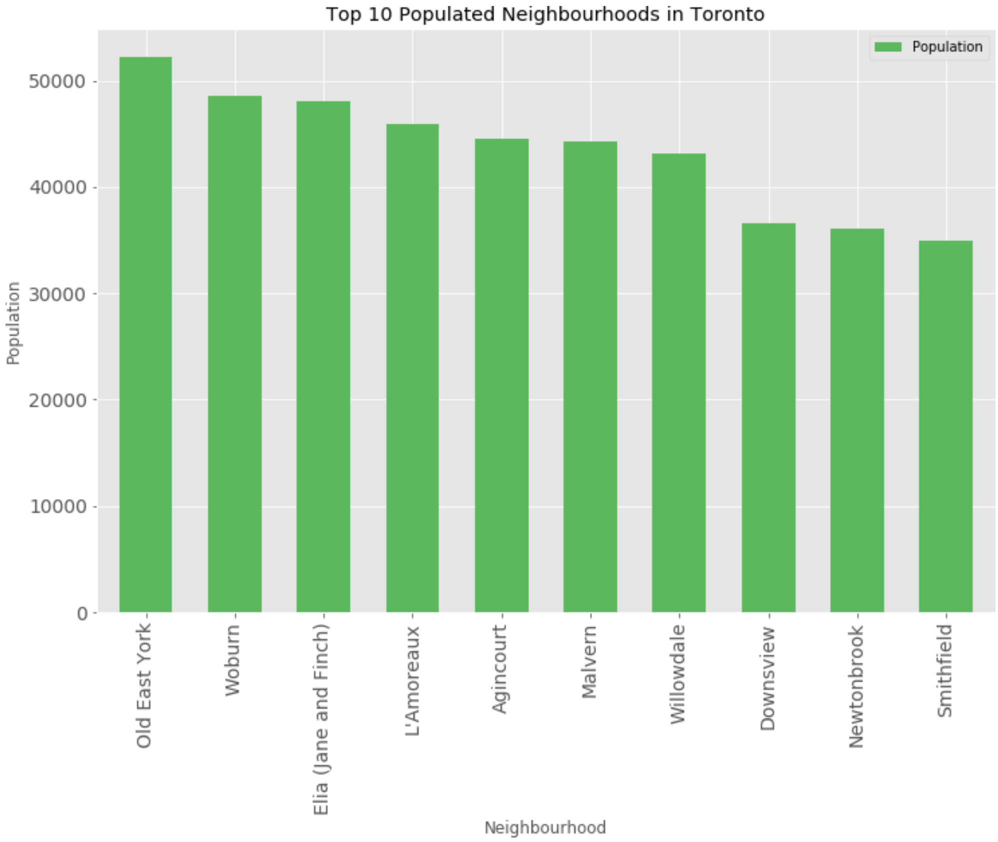
Finally, with all these methodologies, we will then be able to come up with a best recommendation to the management of XYZ Grocery to their problem which is where is the best regions of neighborhoods for them to first start off to offer their services based on neighborhoods similarities, high population and low competition (i.e. fewer grocery stores). In other words, we will not want to recommend to the management to enter a neighborhood whereby there is already a high concentration of grocery stores available and lower demand in the neighborhood.

## 4.0 Result

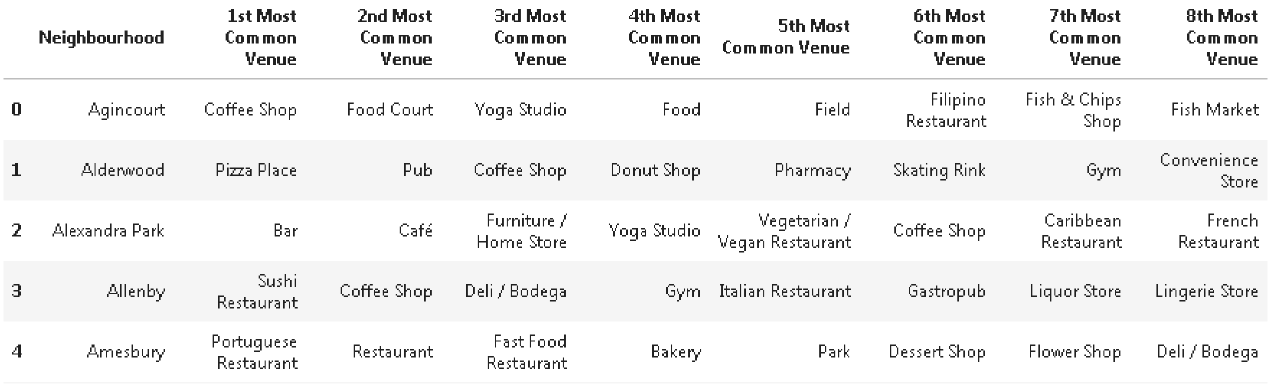
With K-Means clustering technique, the top 5 clusters of similar neighborhoods have been apparent in the result, see below. These clusters are group together based on the similar nearby venues in each of the neighborhoods. This information is critical so that we can target on the cluster that offer the largest business expansion and growth opportunity as the management of XYZ Grocery is interested to replicate their business model fairly quickly upon success in their first service offering in the selected neighborhood.



With bar chart visualization technique, we can easily tell what is the top population (i.e. higher number of residences) in the neighborhood cluster. This is also critical as we will like to recommend to the management of XYZ Grocery of the neighborhood with the higher number of population so that there will be a higher demand for their service offering. The top 10 neighborhoods with highest number of populations are as follows.



With Foursquare.com API, we are also able to leverage on the data to find out the top common nearby venues and their categories in each of these neighborhoods. This is critical as we want to recommend a neighborhood whereby the supply is low (lower competition). As shown below, these neighborhoods have fewer grocery choices available giving  XYZ Grocery a higher advantage and chance to succeed upon entry.

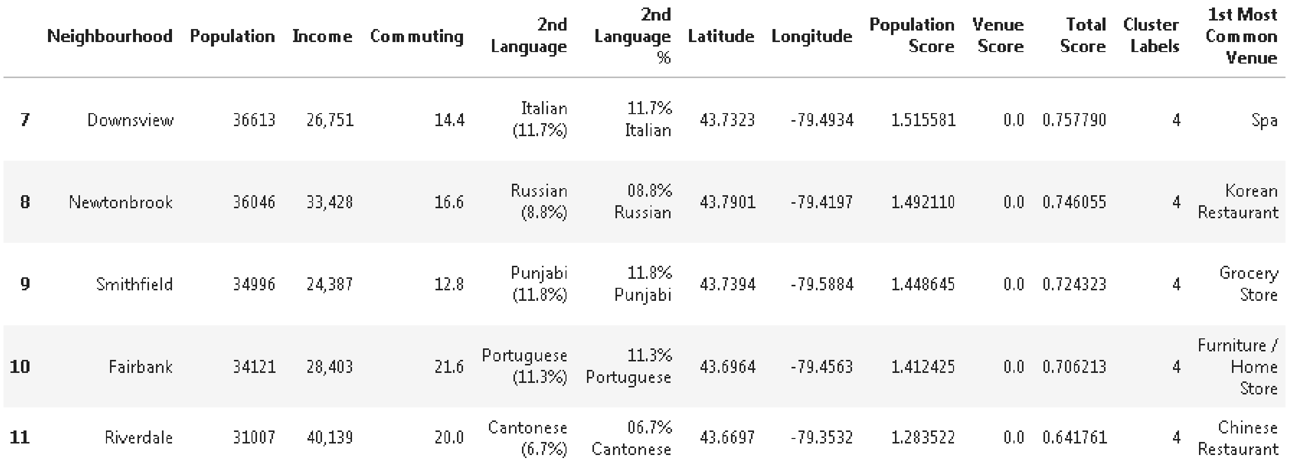


| | Neighbourhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Agincourt | Coffee Shop | Food Court | Yoga Studio | Food | Field | Filipino Restaurant | Fish & Chips Shop | Fish Market |
| 1 | Alderwood | Pizza Place | Pub | Coffee Shop | Donut Shop | Pharmacy | Skating Rink | Gym | Convenience Store |
| 2 | Alexandra Park | Bar | Café | Furniture / Home Store | Yoga Studio | Vegetarian / Vegan Restaurant | Coffee Shop | Caribbean Restaurant | French Restaurant |
| 3 | Allenby | Sushi Restaurant | Coffee Shop | Deli / Bodega | Gym | Italian Restaurant | Gastropub | Liquor Store | Lingerie Store |
| 4 | Amesbury | Portuguese Restaurant | Restaurant | Fast Food Restaurant | Bakery | Park | Dessert Shop | Flower Shop | Deli / Bodega |

## 5.0 Discussion

Based on the result above, the *fifth* cluster looks to offer a higher number of similar neighborhoods and allow XYZ Grocery to replicate their business offering quickly (due to the similarities in these neighborhoods) as part of their growth plan.

Within the *first* cluster, we will like to recommend a neighborhood with higher demand and lower supply to give XYZ Grocery a higher advantage and chance to succeed upon their first service offering. Hence, with this in mind, it is apparent that neighborhood *Downsview* looks to be the choice as it is the highest populated (i.e. 36,613) and very few Grocery stores in the neighborhood (i.e. close to none for the first few most common venues in this neighborhood).



| | Neighbourhood | Population | Income | Commuting | 2nd Language | 2nd Language % | Latitude | Longitude | Population Score | Venue Score | Total Score | Cluster Labels | 1st Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | Downsview | 36613 | 26,751 | 14.4 | Italian (11.7%) | 11.7% Italian | 43.7323 | -79.4934 | 1.515581 | 0.0 | 0.757790 | 4 | Spa |
| 8 | Newtonbrook | 36046 | 33,428 | 16.6 | Russian (8.8%) | 08.8% Russian | 43.7901 | -79.4197 | 1.492110 | 0.0 | 0.746055 | 4 | Korean Restaurant |
| 9 | Smithfield | 34996 | 24,387 | 12.8 | Punjabi (11.8%) | 11.8% Punjabi | 43.7394 | -79.5884 | 1.448645 | 0.0 | 0.724323 | 4 | Grocery Store |
| 10 | Fairbank | 34121 | 28,403 | 21.6 | Portuguese (11.3%) | 11.3% Portuguese | 43.6964 | -79.4563 | 1.412425 | 0.0 | 0.706213 | 4 | Furniture / Home Store |
| 11 | Riverdale | 31007 | 40,139 | 20.0 | Cantonese (6.7%) | 06.7% Cantonese | 43.6697 | -79.3532 | 1.283522 | 0.0 | 0.641761 | 4 | Chinese Restaurant |

It is also apparent that there is a high number of *Italian* people in that neighborhood and hence, we will also like to encourage the management of XYZ Grocery to offer Punjabi food or related supplies in their service offerings in that neighborhood.

## 6.0 Conclusion

With that, we have concluded that the best recommendation for XYZ Grocery to first offer their services in Toronto will be neighborhood *Downsview* with the key factors to consider such as higher demand, lower competition, and easy replication for business expansion. See the recommendation summary below.

1. Region: Fifth Cluster.
2. Neighborhood: Downsview
3. Additional Offering: Italian food or related supplies.

It is also recommended to the management of XYZ Grocery to re-run this data science program to get the updated result and use the result into consideration as part of the business growth plan in selecting the next neighborhood to offer their services. This is critical not only to make sure that they got the updated result for better decision making, but also to make sure that they can re-validate the findings from this project. Finally, thank you for the opportunity in this project and we wish you the best success in your business.