Starbucks Site Selection – Toronto, Canada

IBM Data Science Professional Certificate

Arjun Manjini

April 30, 2021

1. Introduction

1.1. Background

Starbucks Corporation is an American multinational chain of coffeehouses and roastery reserves headquartered in Seattle, Washington. It is the world's largest coffeehouse chain operating around 32,000 stores in 83 countries, of which about 2,000 are in Canada. However, in a recent company announcement in early 2021, Starbucks is shutting down close to 300 coffee shops across Canada in response to a change in customer behaviour and preferences. Considering its commanded substantial brand loyalty, market share, and company value, this study could aid in its restructuring efforts, at least, in Toronto.

1.2. Business Problem

Starbucks Co. is closing stores in downtown core areas and will be focusing on expanding to more pick-up and convenience-led store formats. Therefore, this study is targeted to the global market planning department of Starbucks and Starbucks' local representatives of Toronto, who are looking for new optimal sites to open their next stores while reducing cost and obtaining high footfalls.

Toronto is the capital city of the Canadian province of Ontario and the most populous city in Canada. Unlike most of the grid-plan suburbs dominating in the outskirts of most North American cities, many suburban neighborhoods in Toronto encouraged high-density populations by mixing single-detached housing with higher-density apartment blocks. This kind of diverse cityscape provides ample opportunities to open the new-format Starbucks.

This study could also be replicated to other cities/towns across Canada and maybe to other parts of the world. Other similar chain restaurants/eateries could also benefit with the strategies employed in this study.

2. Data acquisition and cleaning

2.1. Data sources

For our business problem, we will be considering the competition and socioeconomic factors such as population, income, and age groups to determine which groups have high customer demand and less competitors. The population demographics information for each neighborhoods was obtained from Open Data Portal - City of Toronto; OpenCage Geocoding API was used for extracting geographical coordinates of the above neighborhoods; and Foursquare API was used for scraping locations of coffee shops and Starbucks.

2.2. Data cleaning

The population demographics data and geographical coordinates of respective neighborhoods were downloaded and combined into one table (Neighborhoods Profile Data frame). For clarity, some column names were modified. The 45 to 49 years age group feature was missing from the original dataset, but we had total population and other age group values enough to calculate the missing feature. A missing value in the same column was filled by calculating its median.

After combining into one table, we dropped 95 to 99 years age group due to low values in some neighborhoods and fairly assuming their coffee spending will be negligible. Then, we performed boxplot and histogram to understand the distribution of each feature and found all to follow a right-skewed distribution, and some points outside inter-quantile range of the boxplot. So, to ensure to remove any outliers, the values were capped and floored at 99% and 1% percentile, respectively, following the industry standard.

After data cleaning, there were 139 Neighborhoods (samples) and 22 features in the Neighborhoods Profile Data frame (Fig. 1).

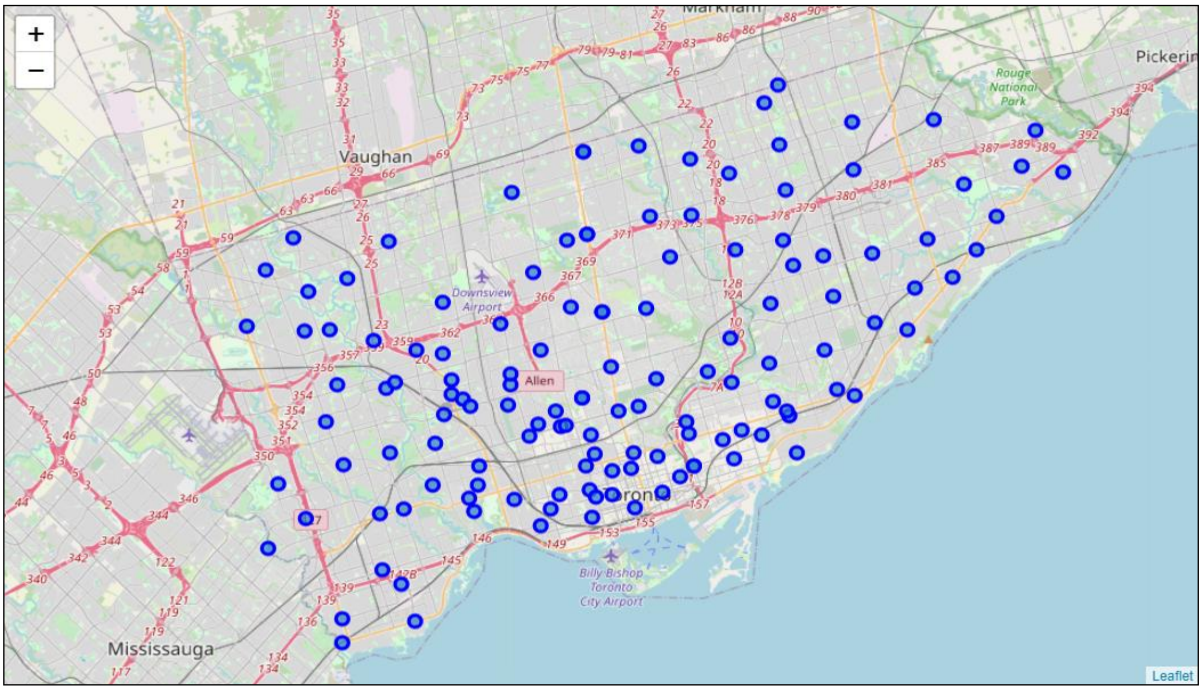

Fig. 1. – Neighborhoods of Toronto, Canada (n=139).

We explored Foursquare API and found that 73 neighborhoods have a total of 172 coffee shops, of which, 40 were Starbucks (Fig. 2).

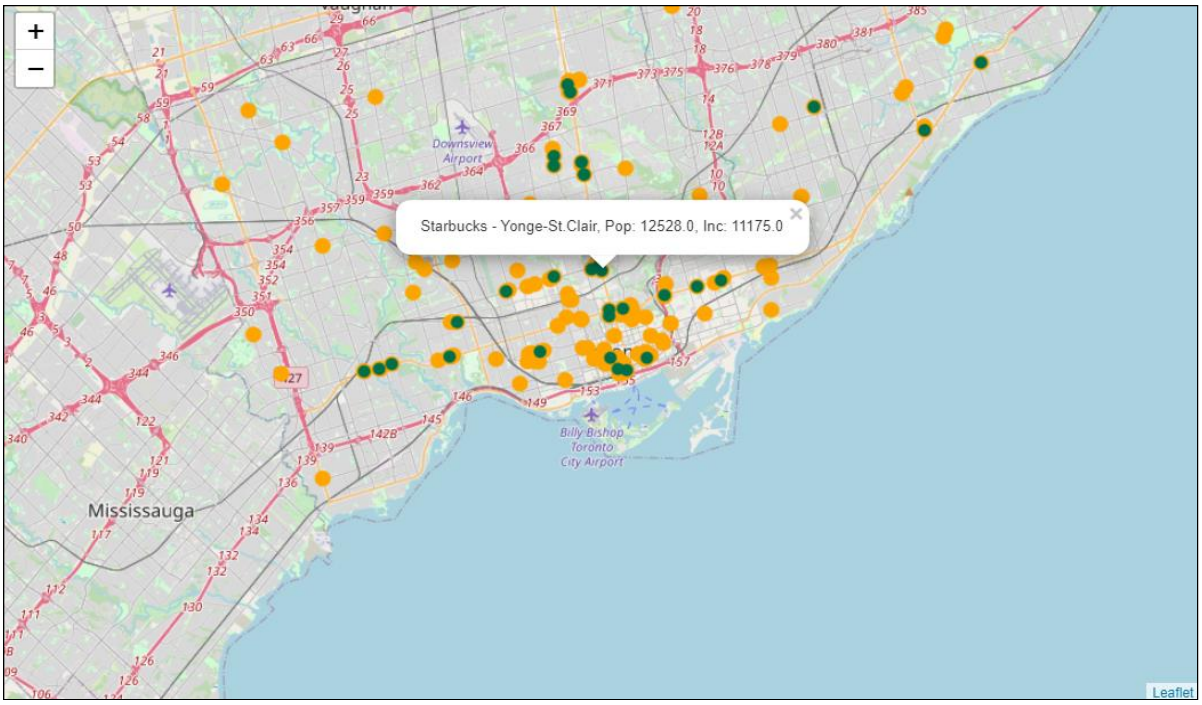
Starbucks - Yonge-St.Clair, Pop: 12528.0, Inc: 11175.0
Downsview Airport
Billy Bishop Toronto City Airport
Mississauga
Leaflet

Fig. 2. – Locations of Starbucks (Green), n=40; and other Coffee shops (Orange), n=172, in Toronto.

3. Methodology

3.1. Feature Normalization and Model selection

For the analysis of our study, we selected 21 features as enlisted in Table-1. Given the different units of the factors, normalization must be conducted to the initial data. Normalization converts the absolute value into relative ones based on the z-score method (also called standardization), thereby obtaining relative size of a value by subtracting the mean of the corresponding feature and then dividing by the standard deviation. The formulation of the normalization operation is shown as follows:

$$x_{std} = \frac{x-\mu}{\sigma}$$

where x represents the original value; μ the mean of the feature; and σ the standard deviation of the feature.

Table-1 – Features (n=21) of the Dataframe for clustering analysis.

Coffee Shop	25 to 29 years	60 to 64 years
Population	30 to 34 years	65 to 69 years
Income	35 to 39 years	70 to 74 years
0 to 14 years	40 to 44 years	75 to 79 years
10 to 14 years	45 to 49 years	80 to 84 years
15 to 19 years	50 to 54 years	85 to 89 years
20 to 24 years	55 to 59 years	90 to 94 years

We can see from Table-1 that there is no target label/feature that act as an output, so, we must determine the groups of neighborhoods with similarities. Thus, Clustering Machine learning algorithm was the most appealing way forward to analyse this study. Furthermore, we are representing this study analysis as a template upon which additional features could be added and perform site selection.

3.2. Clustering analysis

There are other unsupervised learning methods such as Principal component analysis (PCA) to figure out what are the core characteristics of a neighborhood that provides the most benefit for opening and running a Starbucks. However, PCA uses covariance matrix to describe the variance in the data and ranked accordingly – meaning it shows the linear relationships between features to reduce the number of features to a smaller number while losing as little information as possible. In our case, most of the features consists of age groups where in this study, we were not interested in their linear relationship among each feature. Thus, cluster analysis was more suited for this study to discover groups (cluster) in the data and understand the core characteristics.

We used KMeans algorithm, a most used clustering algorithm, to get a meaningful intuition of the structure of the data we are dealing with. Elbow curve method and Silhouette curve method were used to find the optimal value for K (number of clusters).

3.2.1. Elbow curve method

The intuition behind the Elbow curve method is that the explained variation changes rapidly until the number of groups you have in the data and then it slows down leading to an elbow formation in the graph. In Fig. 3 we can see that the elbow forms at both cluster 3 as well as on 5.

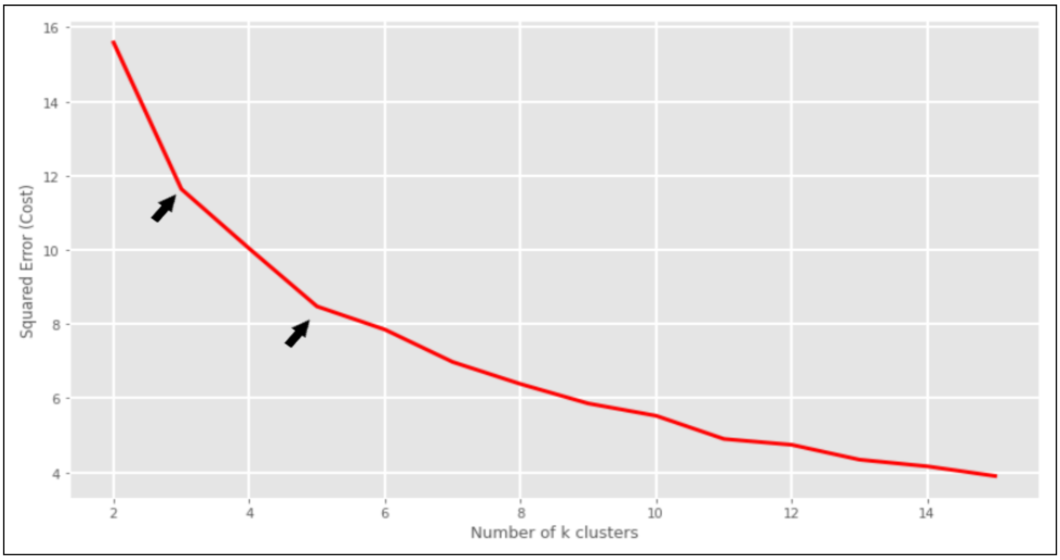

Fig. 3. – Elbow curve method: elbow forms at both K=3 and K=5.

3.2.2. Silhouette curve method

To clarify the above situation, we employed the Silhouette curve method. The silhouette coefficient calculates the density of the cluster by generating a score for each sample based on the difference between the average intra-cluster distance and the mean nearest-cluster distance for that sample. We observed from Fig. 4 that K=5 is the most optimal number of clusters for this study due to relatively higher silhouette scores with even distribution of clusters.

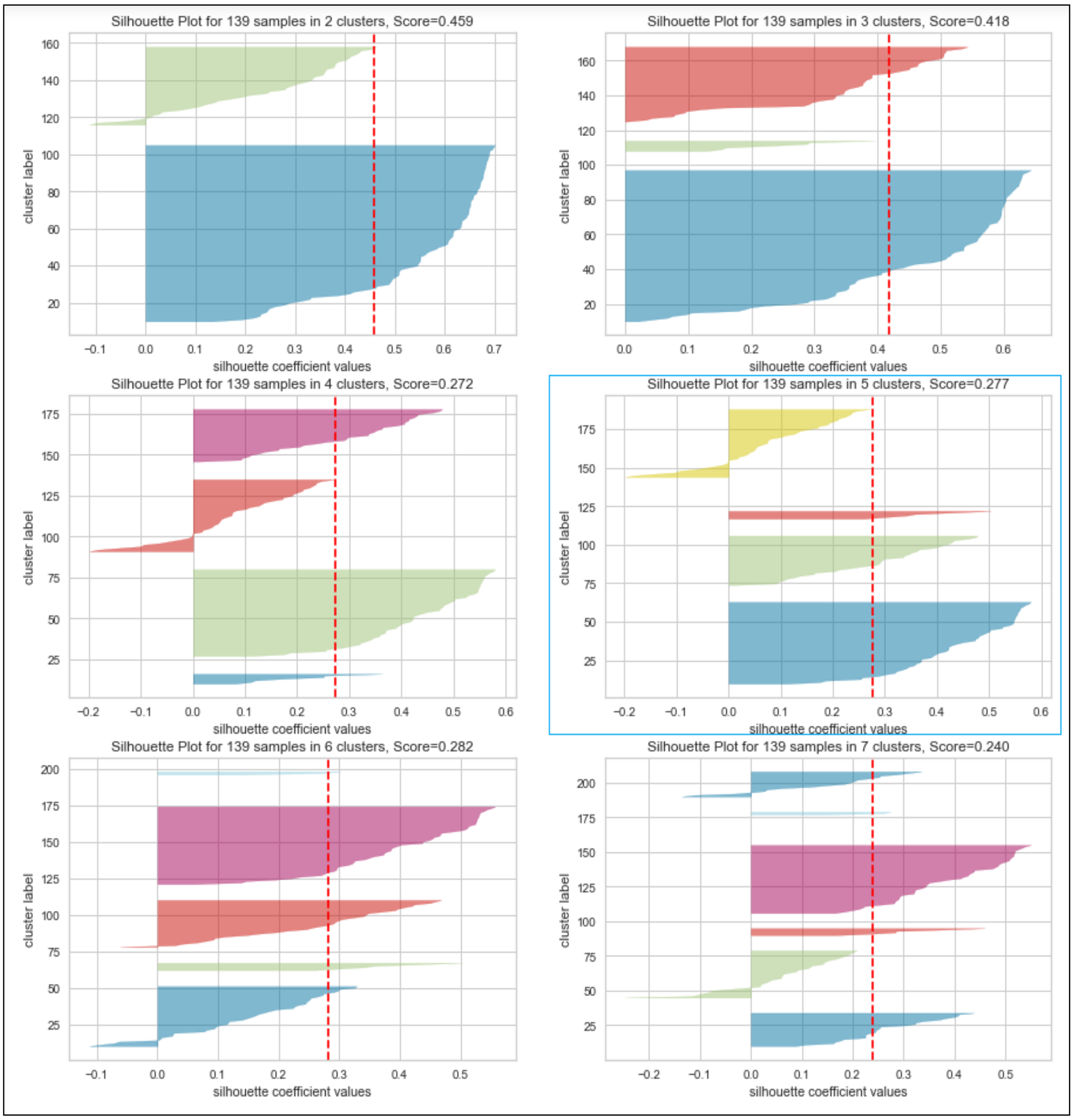

Fig. 4. – Silhouette curve method: The clusters distribution is more uniform in K=5.

4. Results and Discussions

4.1. Cluster examination

KMeans clustering was performed for K=5 and Toronto neighborhoods clusters were visualized on the map (Fig. 5). Each location characteristic such as population, income, number of competitors (other coffee shops), and different age groups of each cluster were examined and compared to the same location characteristics of neighborhoods with Starbucks to select the optimal neighborhood clusters. We compared the clusters by taking the median of individual feature of their respective clusters.

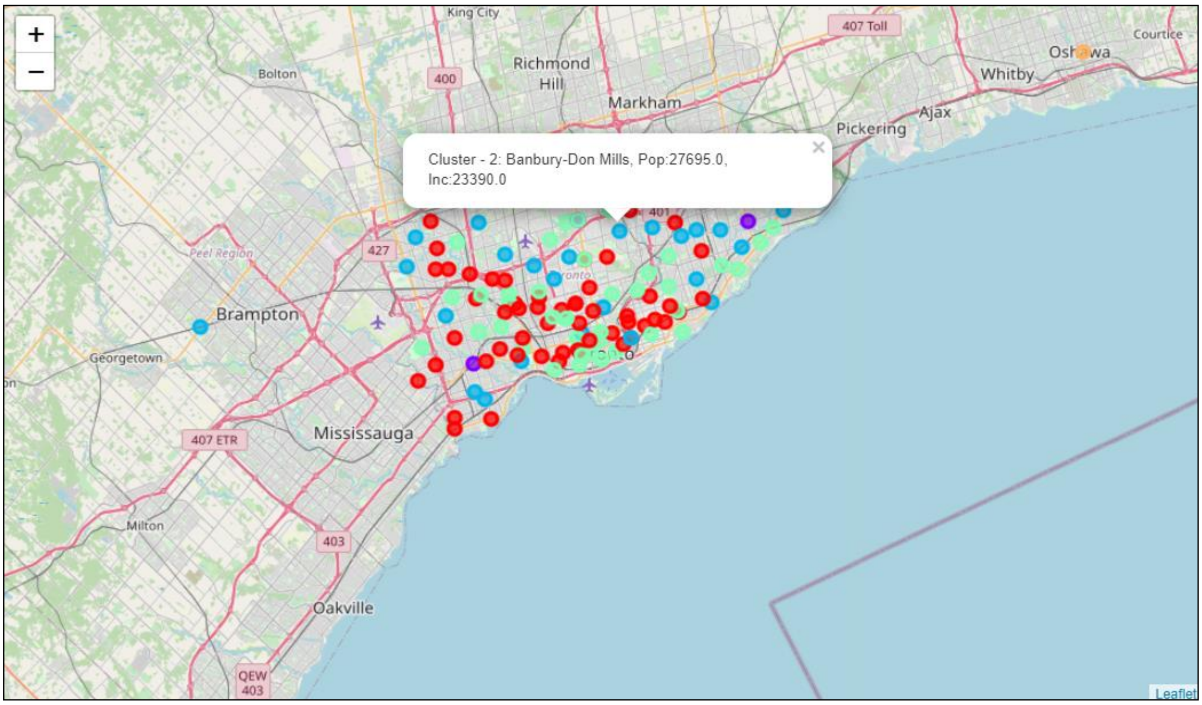
Cluster - 2: Banbury-Don Mills, Pop:27695.0, Inc:23390.0

Fig. 5. – Clustered neighborhoods of Toronto based on similarities in features: Cluster-0 (Red), Cluster-1 (Purple), Cluster-2 (Blue), Cluster-3 (Cyan), and Cluster-4 (Orange)

4.1.1. Population

We see that in Fig. 6, the median population of neighborhoods with Starbucks currently located is around 15,000. The next closest neighborhoods to that value are in clusters 0, 2, and 3. Although clusters 1 and 4 have the highest median population which would give us more footfalls, we also need to consider other factors.

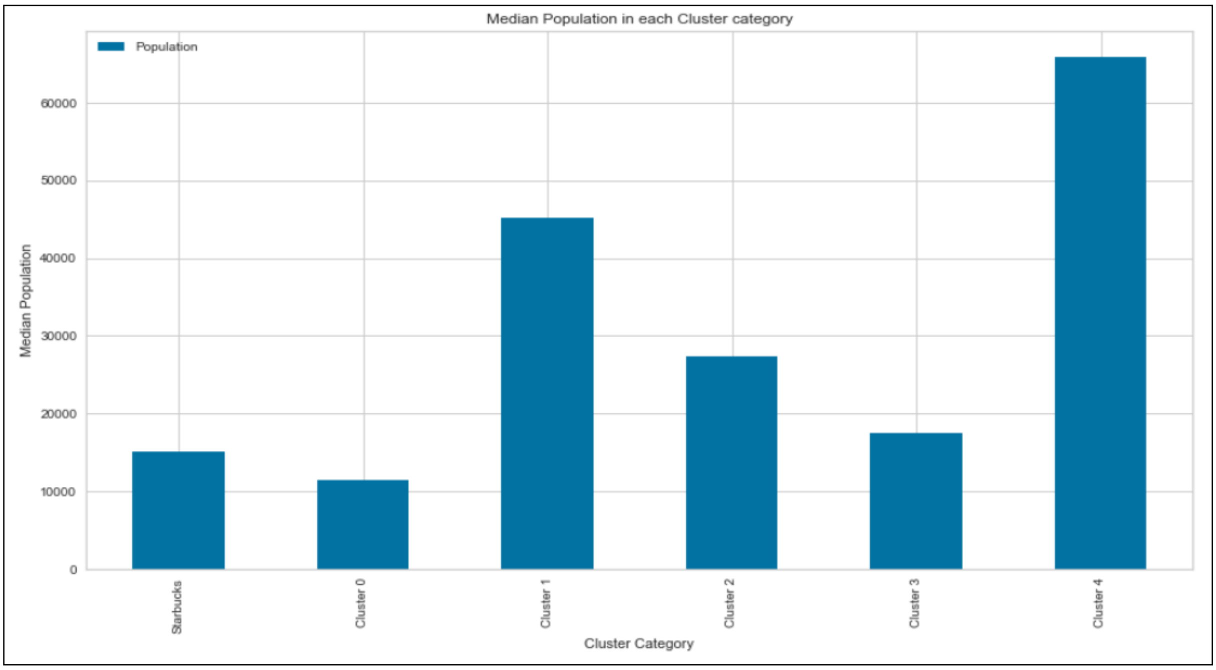

Fig. 6. – Median Population of neighborhoods with Starbucks and of each cluster.

4.1.2. Income

We observed (Fig. 7) the median income of neighborhoods with Starbucks currently located is around 12,000 Canadian Dollars (CAD). So, we are looking for clusters with low- or mid-income, and clusters 0, 2, and 3 fit well into that criteria.

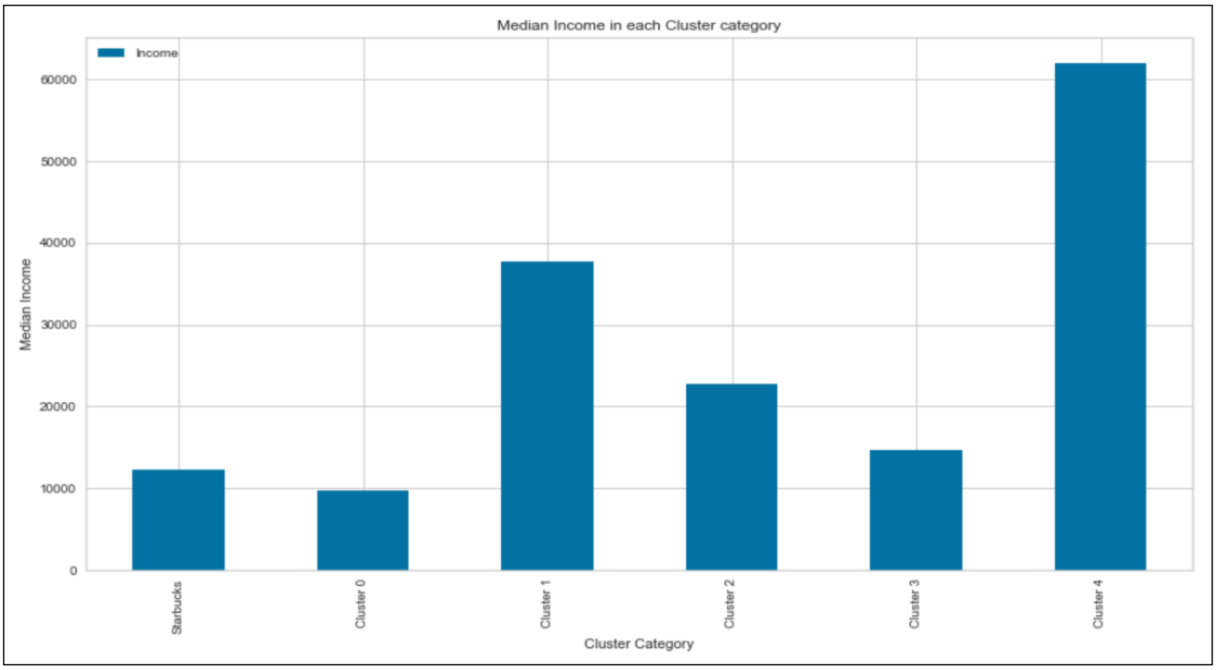

Fig. 7. – Median Income of neighborhoods with Starbucks and of each cluster.

4.1.3. Age groups

Here, we saw that except for cluster 4, all the other clusters showed similar age groups “pattern” to that of neighborhoods with Starbucks currently located, and cluster 1 looked the best followed by clusters 2, and 3 (Fig. 8).

4.1.4. Number of competitors (Coffee shops)

Clusters 0, 2, and 3 were clearly seen to have the lowest number of coffee shops (Fig. 9). Moreover, these three clusters matched well with other factors – population, income, age groups too, therefore, clusters 0, 2, and 3 were the best sites to open our new format Starbucks.

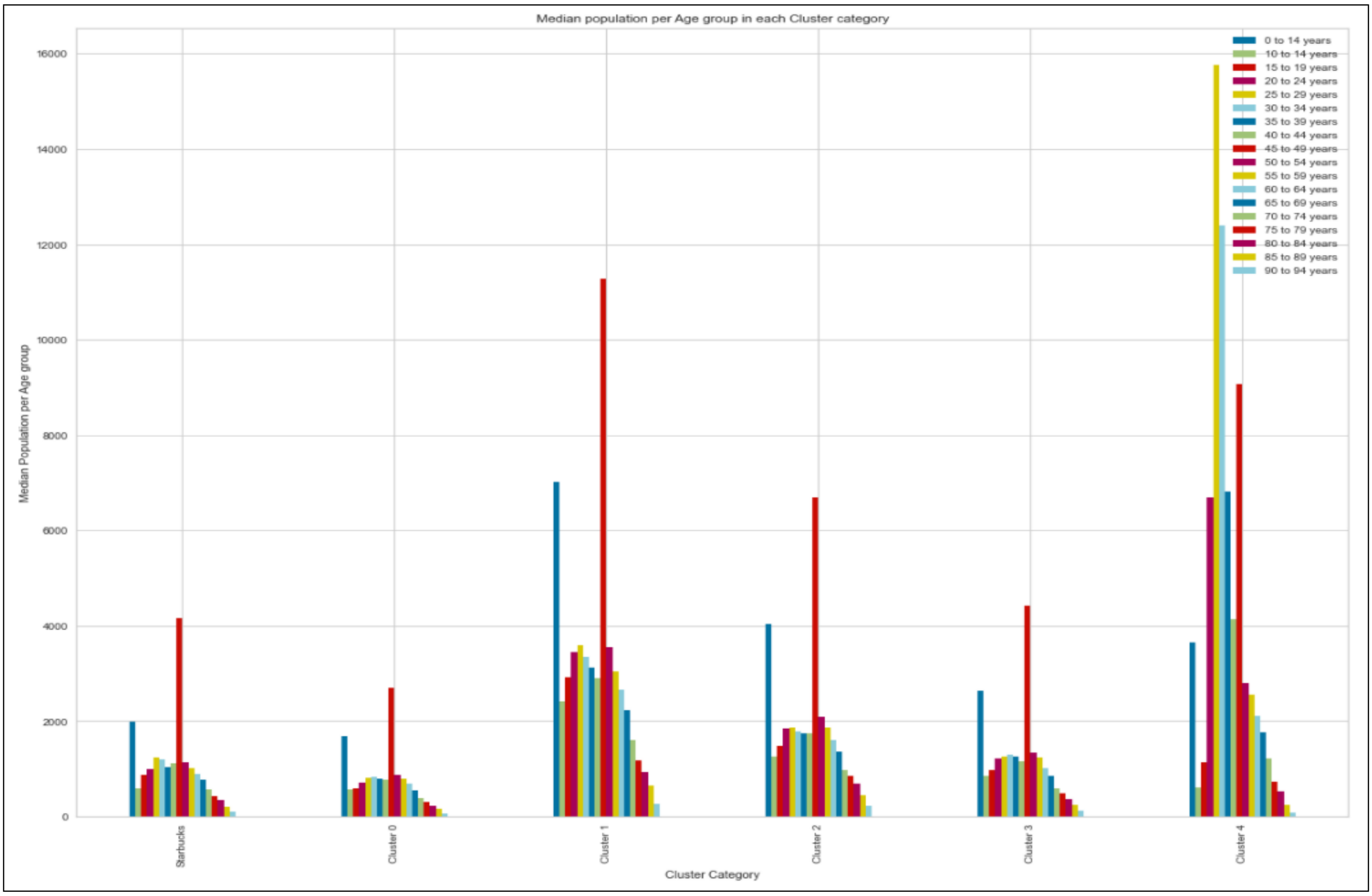

Fig. 8. – Median population grouped by age of neighborhoods with Starbucks and of each cluster.

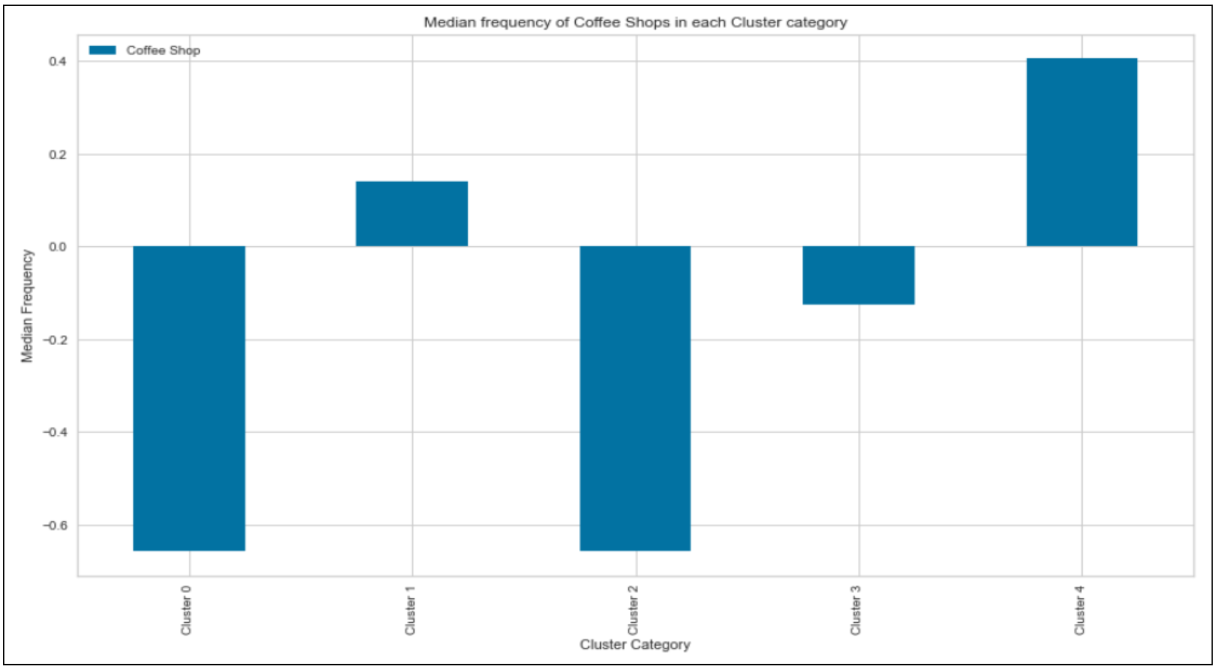

Fig. 9. – Median frequency of coffee shops in each cluster.

4.2. Optimal sites for the new-format Starbucks

In Fig. 10, the shaded neighborhood clusters that are not in the vicinity or overlapping with the current Starbucks locations (dark green), are the optimal sites to open the new format Starbucks store(s). Most of the locations are away from core Downtown Toronto are which was our primary interest.

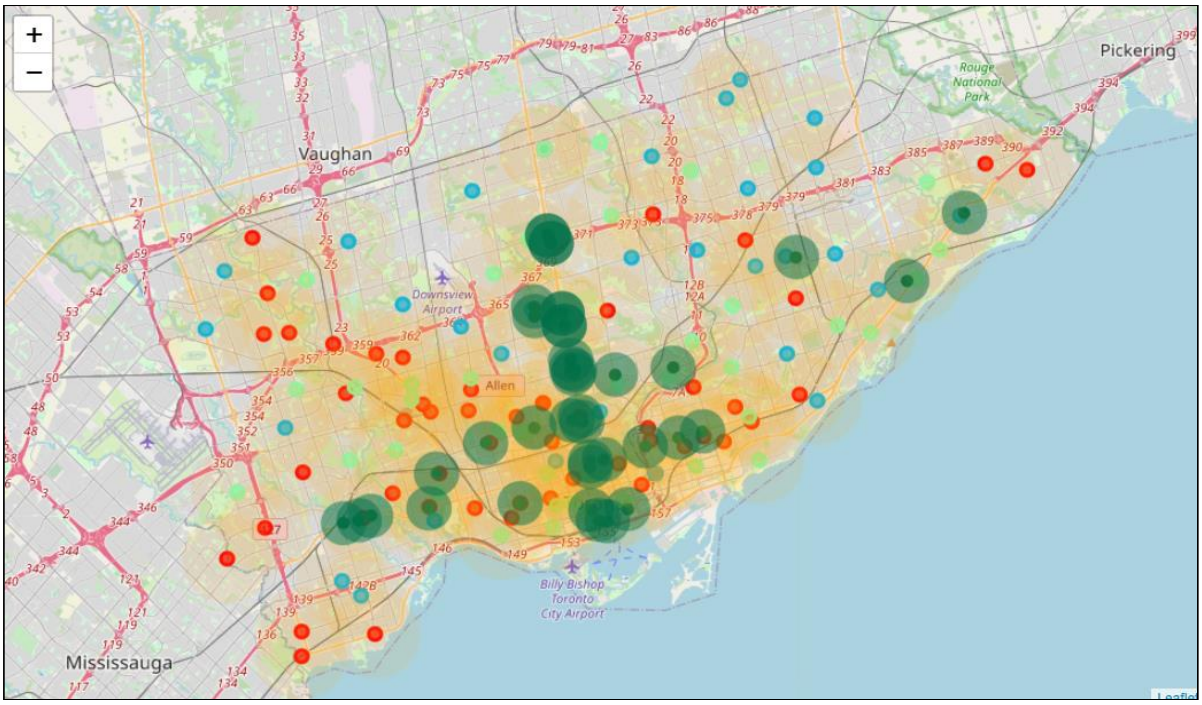

Fig. 10. – Locations of current Starbucks (Green), and potential sites for future expansion (Orange).

The selected neighborhood clusters were identified by comparing with the neighborhood characteristics of current Starbucks locations. The neighborhood characteristics described below represent the Starbucks target population or potential customers:

- **Medium** size population with **high** number of Middle-aged adults and **medium** number of young adults
- **Medium** spending power (income)
- **Low** number of competitors

5. Conclusions

In this study, we identified the clusters of Toronto neighborhoods based on the similarity in their features/characteristics – population, spending power (income), and number of competitors, and then chose potential neighborhood clusters for setting up new Starbucks stores. The most promising neighborhoods belonged to clusters 0, 2, and 3, excluding the current Starbucks locations.

The target customers are the working population comprised mostly middle-aged adults followed by young adults. This makes sense as the current Starbucks are located closely to Office buildings intending for customers to have a cup of coffee before or after work or during breaks. Now, with the change in customer behaviours and preferring to work from home mostly, the chosen neighborhoods are well placed within the residential are or suburbs. This would help the global market planning department of Starbucks and Starbucks' local representatives of Toronto to open up more pick-up and convenience-led store formats.

As these mostly residential neighborhoods, number of competitors are scarce, space rental would expect to be lower, and the pick-up format Starbucks can be easily placed within the existing hypermarkets, discount department stores, and grocery stores. There are also some potential locations on the Ontario Highway 401 which would be perfectly suited for travellers crossing the city or leaving the city for vacations.

In conclusion, the study extensively provides optimal neighborhoods for the stakeholders of Starbucks Co., to open their new format stores with high likelihood success. Similarly, we can interchangeably use this study to expand the potential Starbucks locations in other cities and towns across Canada.

6. Future Directions

The model has certainly some limitations in capturing the target population accurately, as the core Downtown Toronto, where most of the Starbucks coffeeshops are located, are usually comprised of commercial areas and office buildings. Therefore, the observed demographic features would not reflect well in the census. This could potentially be resolved by exploring the Starbucks' customer data for the future studies. The customer data could give us more accurate target population and the potential neighborhoods could be narrowed down.

A possible direction for future research in this area could be towards improving the recommendation performance such as precision, recall and F1 score by carefully exploiting different preferences among users, proximity to landmarks or public areas, business areas, and real estate availability. Additionally, using these factors and more we could create a hybrid recommendation engine to predict or suggest specific Starbucks store format (traditional, pick-up, express, or store tie-ups) for different neighborhoods.