Where would I build a Starbucks(™)?

Introduction

The Toronto Foursquare data set provides community preference data related to venues. One of the main questions asked for data science in geographic data is where to deploy resources to make the most profit or return on investment (ROI). By conducting this study, a data scientist is exploring the main methodologies that can be used for any business application to try and determine the best location for business opportunities.

In this project I will attempt to find out where a coffee company, such as Starbucks, would want to explore new business opportunities in the city of Toronto based upon customer preference, population, and geographic location. There could be many other factors related to this decision such as demographics, customer traffic, and competition, but this simplified model is a good starting point for this problem.

Data

The data I will be using starts with the Foursquare data set but adds data from two new sources. The first is a Kaggle database that provides the Latitude and Longitude of existing Starbucks coffee shops. The second is for the City of Toronto for the population estimate for each neighborhood. In order to use this dta I need to join the data sets through neighborhood names and postal codes. The final data set includes the neighborhoods where coffee shops or cafe are the 1st or 2nd most preferred venue, the population estimate, and the number of starbucks locations per resident of the community.

Neighborhood: A string or multiple strings that were sorted using the Toronto Foursquare project.

Population: Census information provided as floating number values

Starbucksperperson: Population / locations per neighborhood (new calculated field as a floating point value)

1st Most Common Venue: A string containing standardized names. Note that this data set includes coffee shop and cafe that I will use interchangeably.

2nd Most Common Venue: A string containing standardized names. Note that this data set includes coffee shop and cafe that I will use interchangeably.

The file Starbuck_locations2 contains the store locations

The file torontostarbuck6 includes population data added to the neighborhood data.

Methodology

The initial Foursquare data set used in the assignment shows the top ten venues by neighborhood. All of this data is not required and to simplify the analysis only those neighborhoods with a coffee shop and/or a cafe in the top 2 venues will be targeted for this analysis. This will rescue the number of dta row but should eliminate some of the noise from the analysis.

The Starbucks locations dataset from Kaggle list the latitude and longitude for all locations in the greater Toronto area. For the join with the Foursquare data I had to use the geographic data to determine the postal code for each coffee shop. This data set is useful because it will show which areas already have shops located in specific geographies and can be used to determine how many potential customers are in each neighborhood per location.

The third data set is adding the population data to the data set. The City of Toronto open data project lists a large demographic file but does not organize the neighborhoods the same way as the Fourquare data. This data was filtered and combined with the original data to provide an estimate for the population of each neighborhood and to calculate the average population supported by each starbucks in the targeted locations. Some neighborhoods could not be correlated with population due to the different area divisions, and these neighborhoods will be removed from the analysis.

After combining and organizing the data test I will perform mapping and see if a Kmeans analysis can provide some unsupervised learning insights into the best groupings for where the most popular ensure are and most opportunities for a new Starbucks location. The data can also be sorted to find underserved markets based upon favorable conditions in each row.

Results

After filtering for 1st and 2nd most popular venues including only coffee shop and cafe there were 32 neighborhoods identified as potential targets for a new Starbucks coffee shop location. Figure 1. Displays the location on a folium map for remaining neighborhoods.

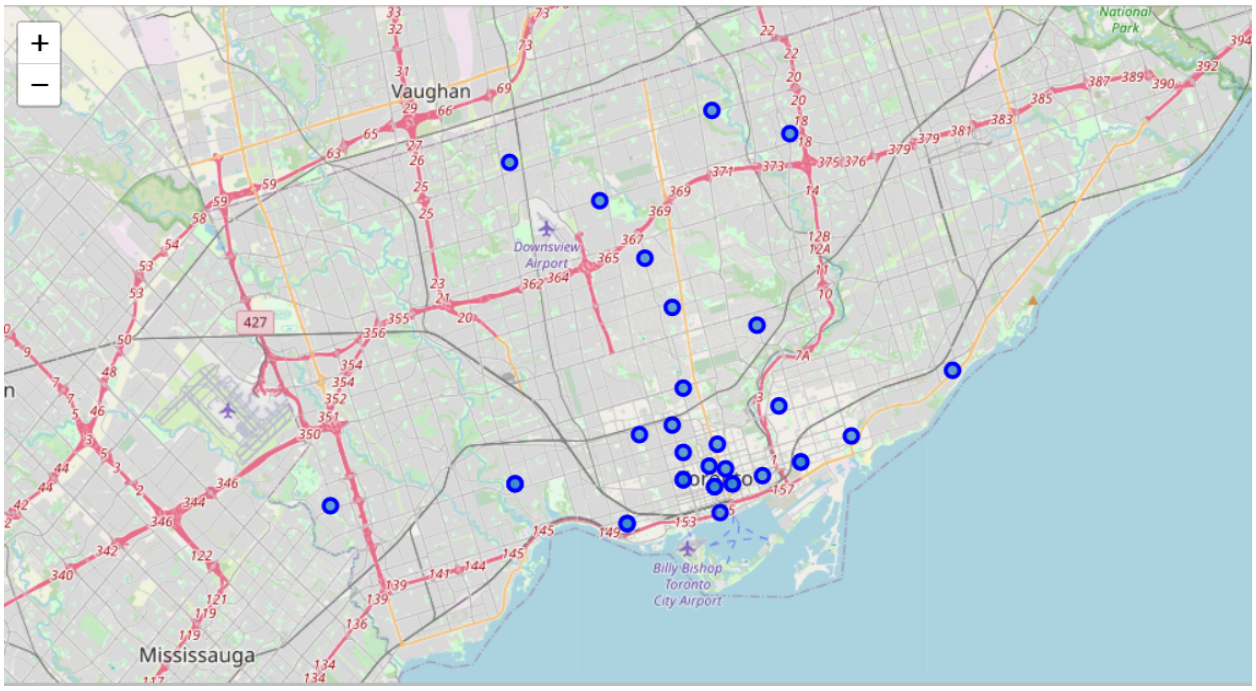

Figure 1. Map of best neighborhoods for a new Starbucks location.

Best neighborhoods:

To find the top locations for a new starbucks the best neighborhoods where coffee shop/cafe was the most popular venue were reviewed. All line items that were listed as cafes were converted inside the columns to a coffee shop string. This data set was then sorted to find the highest value of starbucks per person, indicating where the lowest location per capita was located, See Table 1.

	Neighborhood	Population	Starbucksperperson	1st Most Common Venue	2nd Most Common Venue	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Starbucks
0	Bathurst Manor, Wilson Heights, Downsview North	65290.0	32645.000000	Coffee Shop	Bank	M3H	North York	Bathurst Manor, Wilson Heights, Downsview North	43.754328	-79.442259	2.0
1	Brockton, Parkdale Village, Exhibition Place	37406.0	9351.500000	Coffee Shop	Coffee Shop	M6K	West Toronto	Brockton, Parkdale Village, Exhibition Place	43.636847	-79.428191	4.0
2	Brockton, Parkdale Village, Exhibition Place	37406.0	9351.500000	Coffee Shop	Coffee Shop	M6K	West Toronto	Brockton, Parkdale Village, Exhibition Place	43.636847	-79.428191	4.0
3	Central Bay Street	25767.0	398.916667	Coffee Shop	Italian Restaurant	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383	12.0
4	Church and Wellesley	13397.0	2232.833333	Coffee Shop	Sushi Restaurant	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	6.0

Table 1. Data sorted for lowest per capita of starbuck locations in each neighborhood.

Using this analysis it appears that the top neighborhoods are the cluster of Bathurst Manor, Wilson Heights, and Downsview North. This geographic area has a low starbucks location count and coffee shops are the 1st most popular venue.

As a further analysis an unsupervised Kmeans analysis was completed and the Dendogram in Figure 2 was the result. There were essentially four (4) main clusters with different attributes for each. There was a main orange cluster, #1, and three others that were examined to understand how the algorithm sorted the neighborhoods, #2, 3 ,& 4.

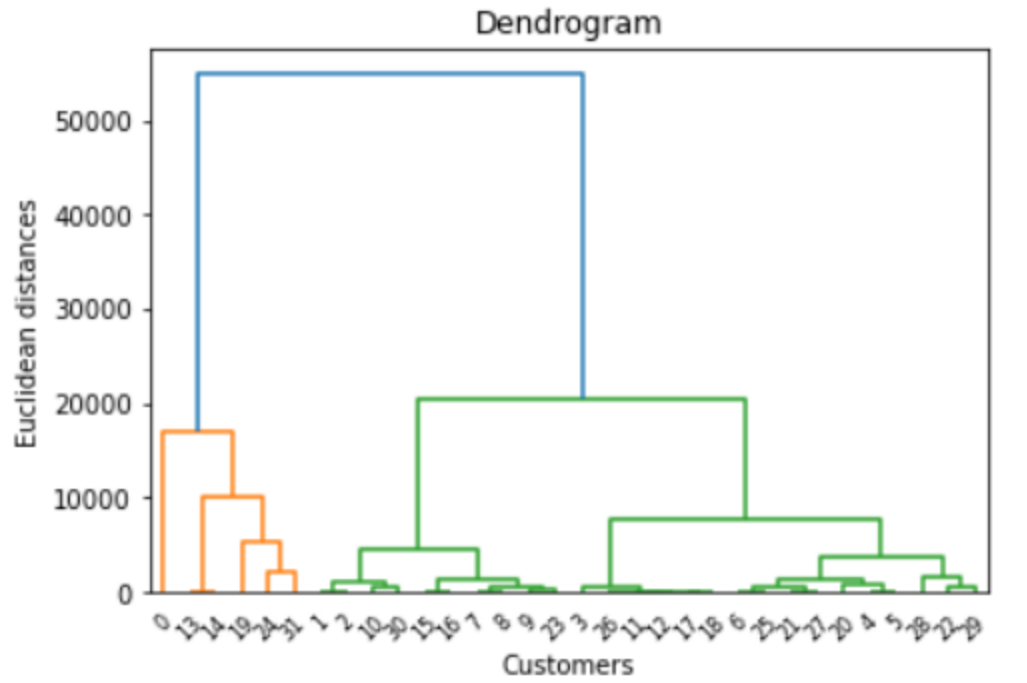

Figure 2. Dendogram from K Means cluster analysis

Cluster #1: Orange Cluster (large population, low starbucks count) - These are the best choices with high populations and a low number of starbucks per location. The initial target area of Bathurst/Wilson/Downsview is listed in this cluster.

Cluster #2: Green Cluster (medium population, multiple locations) - This consists of communities with a medium quantity of locations such that the Starbucks per location count is less than 10,000.

Cluster #3 (Data Errors) Some communities in the data set did not return a population. For the purpose of this analysis we will remove these from the choices.

Cluster #4 (Small Populations) - This final cluster includes neighborhoods with a reported low population and coffee shops were the 2nd most popular venues.

Discussion

As with any realistic situation, the main work in developing this analysis was attempting to find ways to connect the data bases together using dissimilar datasets. The neighborhoods, wards, and names of burroughs varied according to the data set. After the data was collected and connected appropriately it was able to determine some preferential neighborhoods that would be targets for a new coffee shop location.

In addition to the main target area of Bathurst Manor, Wilson Heights, and Downsview North, cluster #1 also recommended Runnymede/Swansea, SummerhillWest/Ratheny/Southhill, Birchcliff/Cliffside, Westm and Danforth West/Riverdale.

Reviewing these target neighborhoods SummerhillWest/Ratheny/Southhill in central Toronto may also be a good target for expansion considering its high population and the popularity of coffee shops.

Conclusion

This analysis utilized geographic data combined with external data sets to try and determine expansion locations for a Starbucks (™) coffee shop. The main factors used included the number of locations already present in the area and the population of the neighborhoods. There are many other marketing considerations that can be used including traffic patterns, demographics, etc. but this initial analysis provides some insight into how we can use existing data sets to provide guidance. Two potential target neighborhoods were identified that are high potential opportunities.

S.Dilney

3/14/2021