The city of coffee shops: Clustering Toronto neighborhoods by income

By Marcelo Ocampo

November 2020

1. Introduction

The objective of this analysis is to provide an initial guidance to a potential business owner looking to open a location in one of Toronto's neighborhoods. This initial guidance gives insights on the possible types of businesses that are common in a neighborhood and how the income profile of a certain community might affect preferences reflected on the venues present. Customer income level is one of the most important traits to take into account when opening a physical store as it also serves as an initial guidance to a price point for the business. This is why we decided to explore how these income levels affect the various venues throughout Toronto and combine all this data to provide meaningful insights.

Moreover, by clustering the different neighborhoods based on their income level and their most popular venues we also get a profile of the current state for the neighborhoods of Toronto. This can also prove to be helpful since it shows a picture of what the competition currently looks like and the possible preferences of the population in those neighborhoods. In general, having the overall picture of commerce dynamics in a neighborhood paired with its income level is a powerful tool for any potential business owner.

2. Data

For our analysis we will be using income information for the city of Toronto segmented by neighborhood as well as information on the different venues classified by their neighborhood locations.

First, for the information on Toronto neighborhoods as well as the income level for each neighborhood we will be using an open dataset produced by Statistics Canada for the 2016 census, although the income data dates from 2015. This data is free to use and is compiled by official Canadian governmental sources. For more information and to extract or check the data please visit the following link:

https://open.toronto.ca/dataset/neighbourhood-profiles/

Second, we will also be using the Foursquare API to extract venue information such as latitude, longitude, category, etc. in order to perform our analysis. For more information on the Foursquare API, please visit the official developer website at Foursquare.

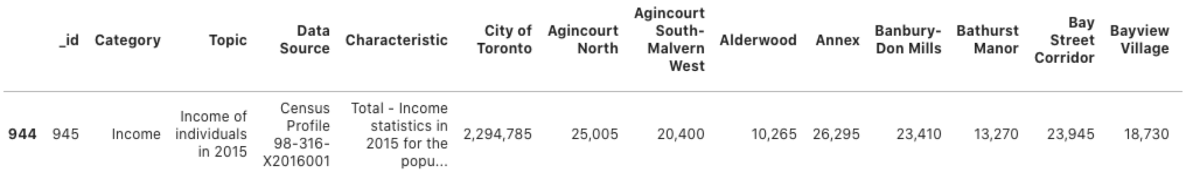
Finally to obtain information on the coordinate for each neighborhood in Toronto we will be using a python geolocator. For more information on this please visit the geolocator official website.

3. Methodology

3.1. Extracting the data and formatting

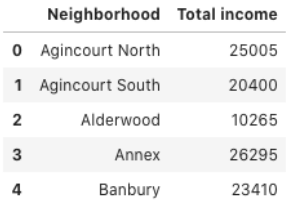
The following dataset contains information on the 140 neighborhoods that comprise the city of Toronto, CA. It includes information on the names of the neighborhoods and other demographics such as median and total income, population, etc. For more information on the dataset please visit the source: https://open.toronto.ca/dataset/neighbourhood-profiles/

The csv file extracted from the data contains various geo-coded data points for the city of Toronto that go from crime-related information to income and population counts. For our purposes, the information we will use is the Total Income by neighborhood from 2015. The final table we will extract should look like this:



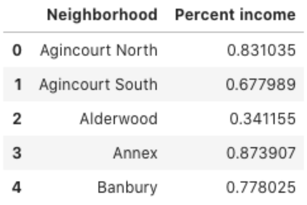
	_id	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	Banbury-Don Mills	Bathurst Manor	Bay Street Corridor	Bayview Village
944	945	Income	Income of individuals in 2015	Census Profile 98-316-X2016001	Total - Income statistics in 2015 for the popu...	2,294,785	25,005	20,400	10,265	26,295	23,410	13,270	23,945	18,730

Where each row represents a data point or topic for the city of Toronto and the neighborhoods are ordered horizontally. We will rearrange this data to keep the neighborhood names and the total income by neighborhood in vertical fashion. The final product should look like this:



	Neighborhood	Total income
0	Agincourt North	25005
1	Agincourt South	20400
2	Alderwood	10265
3	Annex	26295
4	Banbury	23410

Since we are trying to cluster neighborhoods considering their relative wealth, we will be using the total income column as a percentage of median income in Toronto. Per the dataset we obtained, median total income for 2015 is $30,089. We will use this number for our calculations. And our final dataset with relative wealth (as a ratio between total income by neighborhood and Toronto total median income) should look like this:

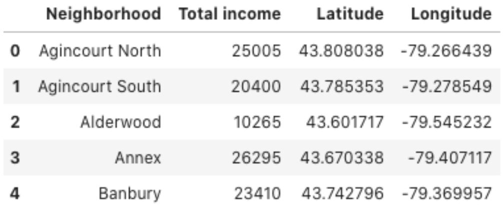


	Neighborhood	Percent income
0	Agincourt North	0.831035
1	Agincourt South	0.677989
2	Alderwood	0.341155
3	Annex	0.873907
4	Banbury	0.778025

3.2. Getting neighborhood coordinates and mapping

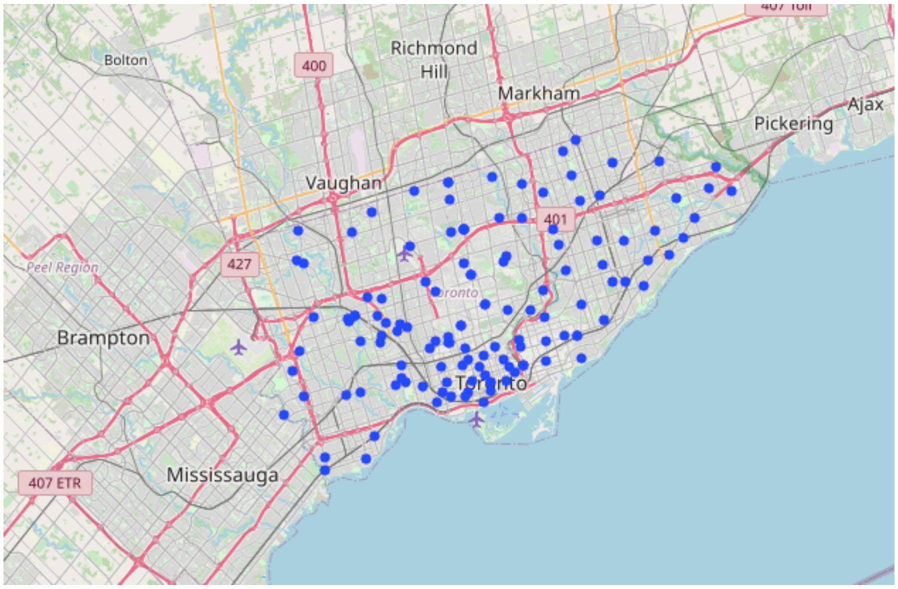
Once we have downloaded the dataset with all the neighborhoods, we will proceed to get the coordinates for each neighborhood and create a map for them. For this we will be using two python libraries, one for extracting coordinates for each neighborhood and create a coordinate table, and another one to create a map with each of the coordinates.

After processing all our neighborhoods into the geolocator tool we will obtain a table with the latitude and longitude information. For our purposes we will also drop any neighborhoods that retrieved no coordinates. In total, we will be dropping 10 neighborhoods with no coordinates information, this leaves us with a total of 130 neighborhoods. Our final table should look like this:



	Neighborhood	Total income	Latitude	Longitude
0	Agincourt North	25005	43.808038	-79.266439
1	Agincourt South	20400	43.785353	-79.278549
2	Alderwood	10265	43.601717	-79.545232
3	Annex	26295	43.670338	-79.407117
4	Banbury	23410	43.742796	-79.369957

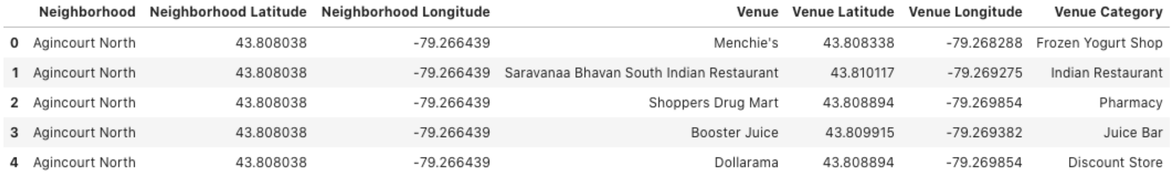
Now, we will proceed to create a Toronto city map with all our results. The map including all neighborhood locations looks like this:



3.3. Using Foursquare API to get venues

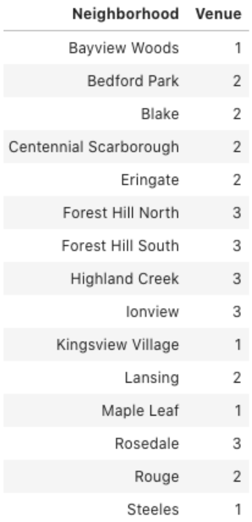
Once we have the geographical information for each neighborhood we can proceed to get the different venues in each of them to get a profile of the businesses in each community. For this, we will be using the Foursquare API, for more information on this, please visit the official Foursquare developer site.

Once we process our data sets we will obtain a new table containing all interesting locations in the vicinity of each neighborhood that will look like the image below. In total, our analysis this time resulted in 3,091 venues for our 130 neighborhoods. The extracted information is useful as it includes longitude and latitude for each venue as well as its category which we will use later on in this analysis.



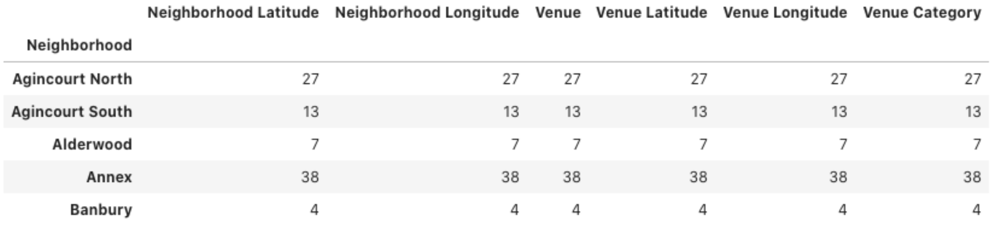
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agincourt North	43.808038	-79.266439	Menchie's	43.808338	-79.268288	Frozen Yogurt Shop
1	Agincourt North	43.808038	-79.266439	Saravanaa Bhavan South Indian Restaurant	43.810117	-79.269275	Indian Restaurant
2	Agincourt North	43.808038	-79.266439	Shoppers Drug Mart	43.808894	-79.269854	Pharmacy
3	Agincourt North	43.808038	-79.266439	Booster Juice	43.809915	-79.269382	Juice Bar
4	Agincourt North	43.808038	-79.266439	Dollarama	43.808894	-79.269854	Discount Store

Additionally, to provide more stability to our analysis of the neighborhoods we will exclude neighborhoods with less than 4 venue results as this is too little information to perform a meaningful analysis. In total we will be excluding 15 neighborhoods. The list of excluded neighborhoods is the following:



Neighborhood	Venue
Bayview Woods	1
Bedford Park	2
Blake	2
Centennial Scarborough	2
Eringate	2
Forest Hill North	3
Forest Hill South	3
Highland Creek	3
Ionview	3
Kingsview Village	1
Lansing	2
Maple Leaf	1
Rosedale	3
Rouge	2
Steeles	1

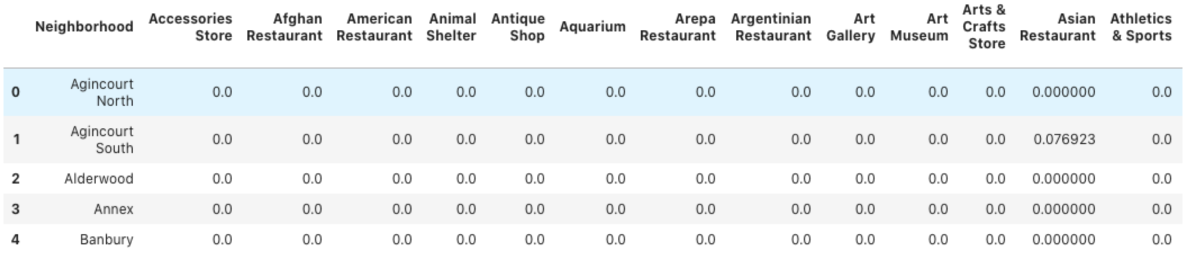
After grouping all venue results into one table we get something like the following, where we see the count for all the venues presented in each neighborhood:



	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Agincourt North	27	27	27	27	27	27
Agincourt South	13	13	13	13	13	13
Alderwood	7	7	7	7	7	7
Annex	38	38	38	38	38	38
Banbury	4	4	4	4	4	4

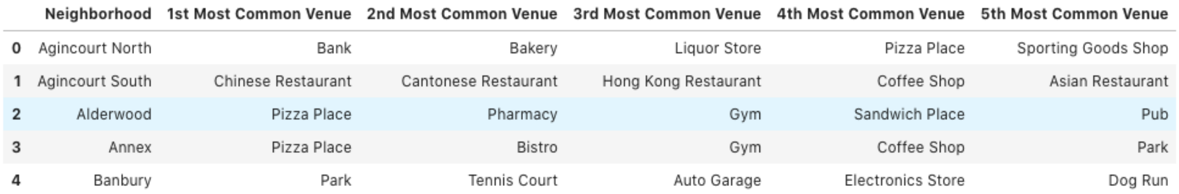
3.4. Analyzing neighborhoods

For this analysis we will turn each venue category in a binary variable which represents whether one venue category exists in a determined neighborhood. Finally, we take the mean of each neighborhood to get the venue profile. The resulting table should look like the following:



	Neighborhood	Accessories Store	Afghan Restaurant	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports
0	Agincourt North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
1	Agincourt South	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.076923	0.0
2	Alderwood	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
3	Annex	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
4	Banbury	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0

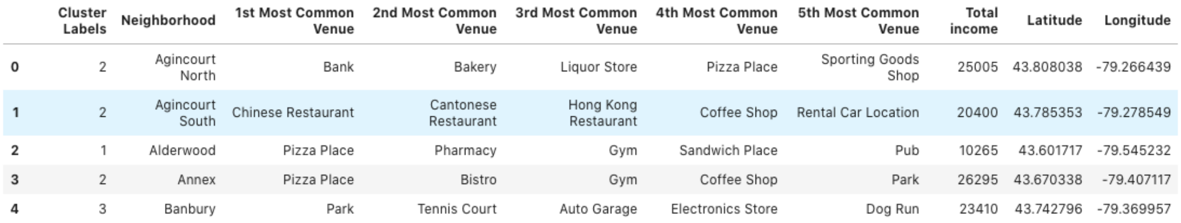
Now, we will proceed to sort out all resulting venues to get the top 5 venues for each neighborhood and arrange all of this in a new table. The result should look like the following:



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Agincourt North	Bank	Bakery	Liquor Store	Pizza Place	Sporting Goods Shop
1	Agincourt South	Chinese Restaurant	Cantonese Restaurant	Hong Kong Restaurant	Coffee Shop	Asian Restaurant
2	Alderwood	Pizza Place	Pharmacy	Gym	Sandwich Place	Pub
3	Annex	Pizza Place	Bistro	Gym	Coffee Shop	Park
4	Banbury	Park	Tennis Court	Auto Garage	Electronics Store	Dog Run

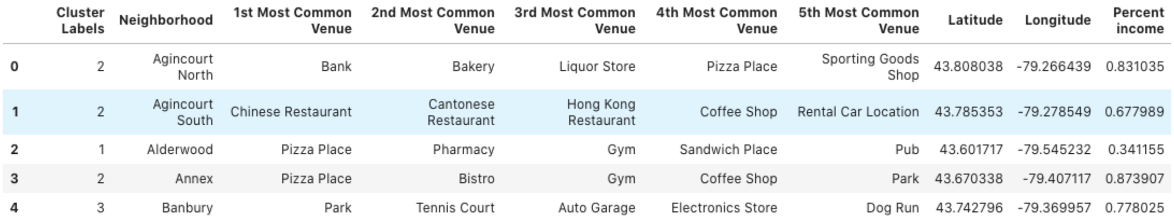
3.5. Clustering

Now that we have the information for all the venues of all neighborhoods including the geolocation we will proceed to the clustering of all these results to get a group profile Toronto. For this we will use a python library for data analysis that includes an AI clustering module. We feed our data from the one-hot encoding as well as the latitude and longitude information to perform the analysis. For our purposes we will categorize the neighborhoods into 4 clusters depending ont their venues and their relative income levels. This will result in cluster labels that we will pair up with our data. The final result where each neighborhood is assigned into a cluster should look something like the following:



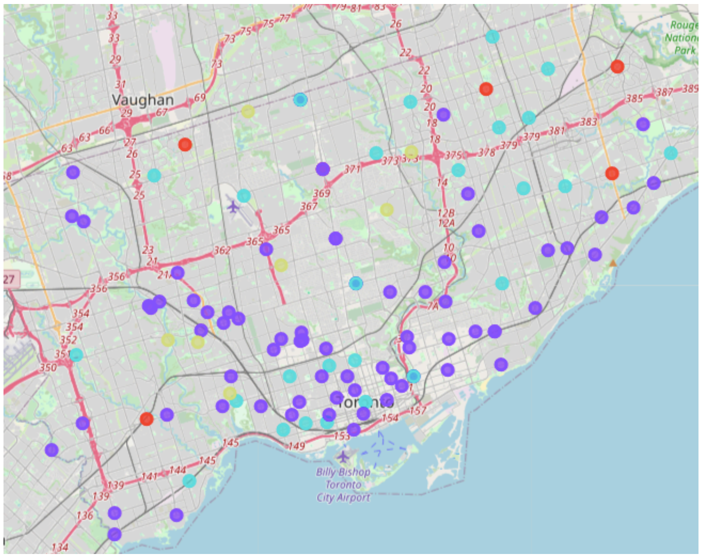
	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Total income	Latitude	Longitude
0	2	Agincourt North	Bank	Bakery	Liquor Store	Pizza Place	Sporting Goods Shop	25005	43.808038	-79.266439
1	2	Agincourt South	Chinese Restaurant	Cantonese Restaurant	Hong Kong Restaurant	Coffee Shop	Rental Car Location	20400	43.785353	-79.278549
2	1	Alderwood	Pizza Place	Pharmacy	Gym	Sandwich Place	Pub	10265	43.601717	-79.545232
3	2	Annex	Pizza Place	Bistro	Gym	Coffee Shop	Park	26295	43.670338	-79.407117
4	3	Banbury	Park	Tennis Court	Auto Garage	Electronics Store	Dog Run	23410	43.742796	-79.369957

We finally have all the parts for the final table we will use in our analysis for the last section. We have: information on the top most common venues, neighborhood name, latitude, longitude, cluster label and income data. However, we need to put all of this together in one table. That is what we will do now. The final result:



	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Latitude	Longitude	Percent income
0	2	Agincourt North	Bank	Bakery	Liquor Store	Pizza Place	Sporting Goods Shop	43.808038	-79.266439	0.831035
1	2	Agincourt South	Chinese Restaurant	Cantonese Restaurant	Hong Kong Restaurant	Coffee Shop	Rental Car Location	43.785353	-79.278549	0.677989
2	1	Alderwood	Pizza Place	Pharmacy	Gym	Sandwich Place	Pub	43.601717	-79.545232	0.341155
3	2	Annex	Pizza Place	Bistro	Gym	Coffee Shop	Park	43.670338	-79.407117	0.873907
4	3	Banbury	Park	Tennis Court	Auto Garage	Electronics Store	Dog Run	43.742796	-79.369957	0.778025

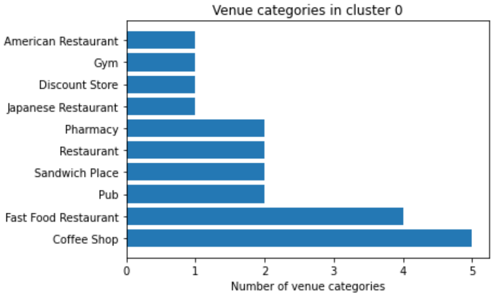
Now, we will also create a map with each neighborhood color-coded to each of their clusters.



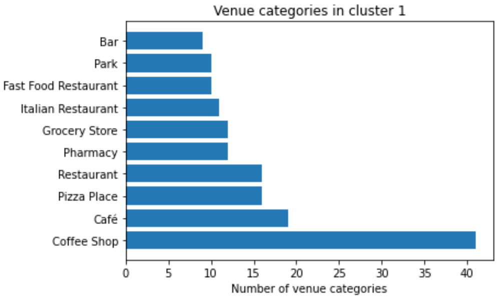
4. Results

Our analysis resulted in 4 clusters which were named according to their relative income: High income, High-medium income, Low-medium income and Low income.

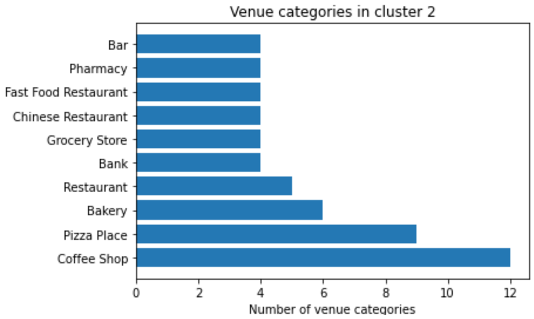
The first cluster is the high income cluster with an average relative income of 1.4 times the median income in Toronto. Below is a snapshot of the most common venues:



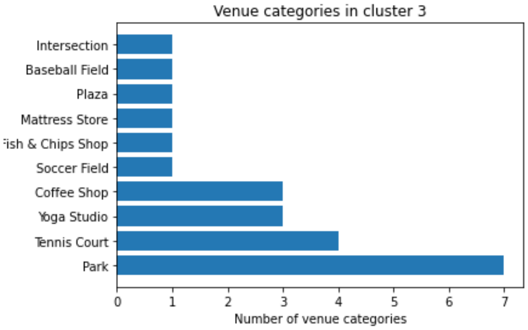
The second cluster is the low income cluster with an average relative income of 0.39 times the median income in Toronto. Below is a quick view of the most common venues here:



The third cluster is the high-medium income with an average relative income of 0.8 times the median income in Toronto. Below is a snapshot of the venue dynamic here:



The final cluster, fourth cluster, is the low-medium income cluster with an average relative income of 0.6 times the median income in Toronto. Below is a quick view of the venues:



5. Conclusions

From our analysis above we can see that classifying the different neighborhoods in the city of Toronto by their income yields interesting results. First, it looks like Toronto is the city of coffee shops! Coffee shops was one of the most popular categories across all income segments. However, besides that there seems to be a lot of differences between the clusters in terms of popular venues.

For Cluster 0, which aggregates the neighborhoods with the highest income relative to the median income in 2015, it looks like there's a tendency to quick eating venues such as fast food restaurants or sandwich places. The reason behind this might be that the neighborhoods included here tend to have a strong commercial sector. All these quick eats places might be there as part of the shopping experience. This can be a good guide for a new business trying to open in this areas to maybe stay away from sit-down experiences in favor of something more for-the-go.

For Cluster 1, the low income cluster, it seems there's the opposite tendency as there're more full-service restaurants. This might be explained as these might also be more residential areas which, in turn, house more families. An interesting next step here would be to research the average price level for these restaurant venues (in case the interest is in restaurant in particular) to get a sense of the level of prices that the population here is used to. Nevertheless, a price level research can also result in a good idea of the price elasticity for this population. Again, the main takeaway here are the most popular categories that resulted from the analysis.

For Cluster 2, the high medium income cluster, it seems the tendency more closely resembles that of a residential-familiar neighborhood with the inclusion of bakeries and grocery stores and even banks in the mix. Given that these areas command a higher income than Cluster 1, for example, it would make sense that price levels might also be higher for entertainment venues such as restaurants. This might also be explained by the age level in these neighborhoods, as if there's an older skew, that might explain the tendency to a more daylight entertainment.

Finally Cluster 3, the low medium income cluster, presents some interesting characteristic with the high prevalence of parks and places for exercising and sports in contrast to dining for the other clusters. A good idea here would be to think about complementary services that go hand in hand with these venues instead of thinking about the same venues already present here.

6. Recommendations

To end this analysis, I would like to finish the way that every business action finishes, not only with an answer but with next steps in light of the insights we have gathered. First off, further work is needed with more data and with better quality datasets in order to determine the best locations for certain businesses. As said in the conclusion paragraph above, data such as age, marital status, labor status, etc. might be useful to determine the probability of success for a specific venture. Additionally, there should also be field research to get to the appropriate price level and to empirically confirm the insights drew from this analysis.

Finally, in determining the success of a business venture a data analysis is never enough to make a full prediction. Even though this analysis might give a business owner the first major insight to take that first step he or she requires, there still is work to be done.