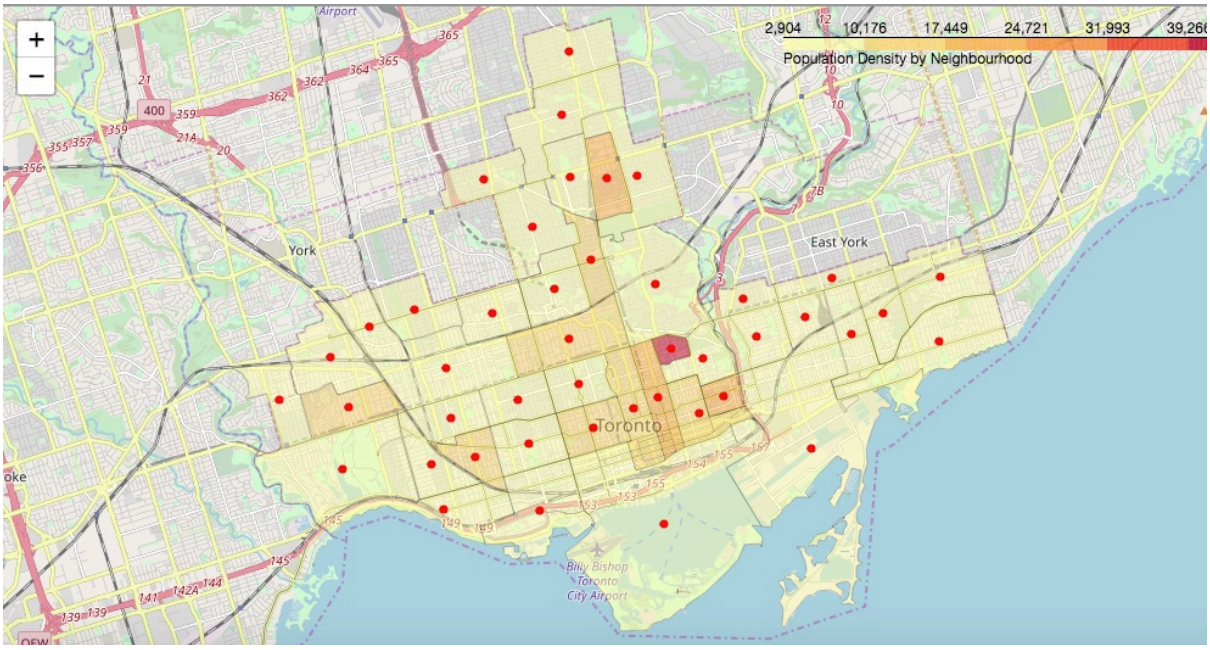
Capstone Project - The Battle of Neighborhoods

Marc van der Valk, February 23rd 2019



Introduction / Business Problem

Finding the right small business location is one of the primary steps in preparing to set up a new business. It is not always an easy task. This project aims to help current and future business owners in the process of selecting business locations. By using data from a location based social network services like Foursquare as well as neighborhood area statistics it should be possible to recommend possible business locations.

As the types of small businesses are manifold, this project will restrict the definition to those businesses that fall under the categories of shops, restaurants, cafes and bars. These types of businesses depend on foot traffic, easy access and good visibility.

There are several of factors that can influence choosing a location:

- Location of similar businesses
  - Businesses are usually located where they are for a good reason,
  - Customers already in the area are more likely to be looking for a similar business
- Consumer statistics for similar business
  - Average number of customer visits.
  - Popularity of a business
- Distance between consumers and business
  - The further the consumer is located from the business the less likely he or she is to visit.
  - Consumer location doesn't necessarily mean domestic location but could also mean job location.
- Locations close to transportation hubs, parking facilities, entertainment centres like theatres, cinemas or public parks and office buildings

  - Locations where there is a large amount of foot traffic: concentration of possible customers
- Population density of the surrounding area
  - More people close by: more possible customers
  - There are statistics available on population by neighbourhood or postal code area.
- Average Income
  - Higher average income: possible customers with more money to spend
  - There are statistics available on average income by neighbourhood.

This project will attempt to combine the above factors to build a clustering and/or recommendation model for the best areas for locating certain businesses. The recommendation(s) given by the model should help the (future) business owner to make a more informed decision

Note: only further analysis in the next stage after gathering the data will prove which machine learning method is better suited to use

Data Section

Statistics Data on Neighbourhoods

I have chosen to look at the neighborhood's in the former city of Toronto for this study. This is based on the fact that the city has a substantially large population with readily available statistics.

1. **Neighbourhoods with central and boundary geo-coordinates**:

- **CDN_Number**: Area code for the neighbourhood, 3 digits
- **Neighbourhood**: Name of the neighbourhood
- **geometry**: collection of geo-coordinates designating the boundary of the neighbourhood
- **Latitude**: the latitudinal coordinate of the center of the area (centroid)
- **Longitude**: the longitudinal coordinate of the center of the area (centroid)

Neighborhood: according to the website of the city of Toronto, the definition of a neighborhood: is an area that respects existing boundaries such as service boundaries of community agencies, natural boundaries (rivers), and man-made boundaries (streets, highways, etc.) They are small enough for service organizations to combine them to fit within their service area. They represent municipal planning areas as well as areas for public service like public health.
A neighborhood has a population roughly between 7,000 and 12,00 people.

	CDN_Number	Neighbourhood	geometry	Latitude	Longitude
0	097	Yonge-St.Clair	POLYGON ((-79.39119482700001 43.681081124, -79...	43.687859	-79.397871
1	027	York University Heights	POLYGON ((-79.505287916 43.759873494, -79.5048...	43.765738	-79.488883
2	038	Lansing-Westgate	POLYGON ((-79.439984311 43.761557655, -79.4400...	43.754272	-79.424747
3	031	Yorkdale-Glen Park	POLYGON ((-79.439687326 43.705609818, -79.4401...	43.714672	-79.457108
4	016	Stonegate-Queensway	POLYGON ((-79.49262119700001 43.64743635, -79....	43.635518	-79.501128

Note: In the case of the neighborhood's geospatial data no data cleansing is necessary, other than removing the CDN number from the description. I have renamed the columns to be consistent. The central geo-coordinates for each neighborhood have also been calculated using a geopandas geometry attribute called centroid.

2. **Wikipedia table containing neighbourhoods by former city / borough**

This table is used to filter the neighborhood's by the former city area of Toronto: https://en.wikipedia.org/wiki/List_of_city-designated_neighbourhoods_in_Toronto

- **CDN_Number**: Area code for the neighbourhood, 3 digits
- **City-designated-area**: Name of the neighbourhood
- **Borough**: Former city or borough

Note: In the case of the Wikipedia list of neighborhoods' in Toronto, there are now missing values. To be consistent, I have reformatted the CDN number to a zero-fill 3-digit number. Just to make sure I compared the CDN numbers and neighborhood names to the neighborhood geospatial file and there were no differences. The number of rows (read neighborhood's is the same)

	CDN_Number	City-designated area	Borough
0	129	Agincourt North	Scarborough
1	128	Agincourt South-Malvern West	Scarborough
2	020	Alderwood	Etobicoke
3	095	Annex	Old City of Toronto
4	042	Banbury-Don Mills	North York

3. **Toronto Population Statistics by Neighbourhood**

Neighborhood population , area and household income from 2014.This can be retrieved from the city of Toronto neighborhood wellbeing app at https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/wellbeing-toronto/

The file contains the following columns:

- **Neighbourhood**: Name of the neighbourhood
- **CDN_Number**: Three-digit neighbourhood code
- **TotalPopulation**: Total population for the neighbourhood based on 2014 data
- **TotalArea**: Area of the neighbourhood in square kilometers
- **After_TaxHouseholdIncome**: Average household income after tax in Canadian dollars
- **PopulationDensity**: Density of the population by square kilometers

This excel file will be loaded into a pandas dataframe

Example of the data:

	Neighbourhood	CDN_Number	TotalPopulation	TotalArea	AfterTaxHouseholdIncome	PopulationDensity
0	West Humber-Clairville	001	33312	30.09	59703	1107.0
1	Mount Olive-Silverstone-Jamestown	002	32954	4.60	46986	7164.0
2	Thistletown-Beaumond Heights	003	10360	3.40	57522	3047.0
3	Rexdale-Kipling	004	10529	2.50	51194	4212.0
4	Elms-Old Rexdale	005	9456	2.90	49425	3261.0

Note: There were no empty values in this table and the number of rows compared with the previous neighborhood files. To be consistent, I have reformatted the CDN number to a zero-fill 3-digit number. Just to make sure I compared the CDN numbers and neighborhood names to the neighborhood geospatial file and there were no differences. A population density column was calculated by dividing the total population by the total area of the neighborhood.

4. **Combined table with neighbourhood as key**

The three above mentioned tables will be loaded and joined based on FSA code to form a dataframe containing the following columns:

- **CDN_Number**: Three digits designating a neighbourhood (data 1.)
- **Neighbourhood**: Name of the neighbourhood (data 1.)
- **Latitude**: the latitudinal coordinate of the center of the area (data 1.)
- **Longitude**: the longitudinal coordinate of the center of the area (data 1.)
- **geometry**: a list of latitude - longitude coordinates forming the boundaries of the neighbourhood (data 1.)
- **TotalPopulation**: the total population of the neighbourhood (data 3.)
- **TotalArea**: the total area in square kilometers (data 3.)
- **AfterTaxHouseholdIncome**: average household income after tax for the neighbourhood (data 3.)
- **PopulationDensity**: the population density of the area in persons by square km (TotalPopulation/TotalArea)

  This dataframe named **df_toronto_ven** will form the features for a neighborhood and used for the machine learning algorithm

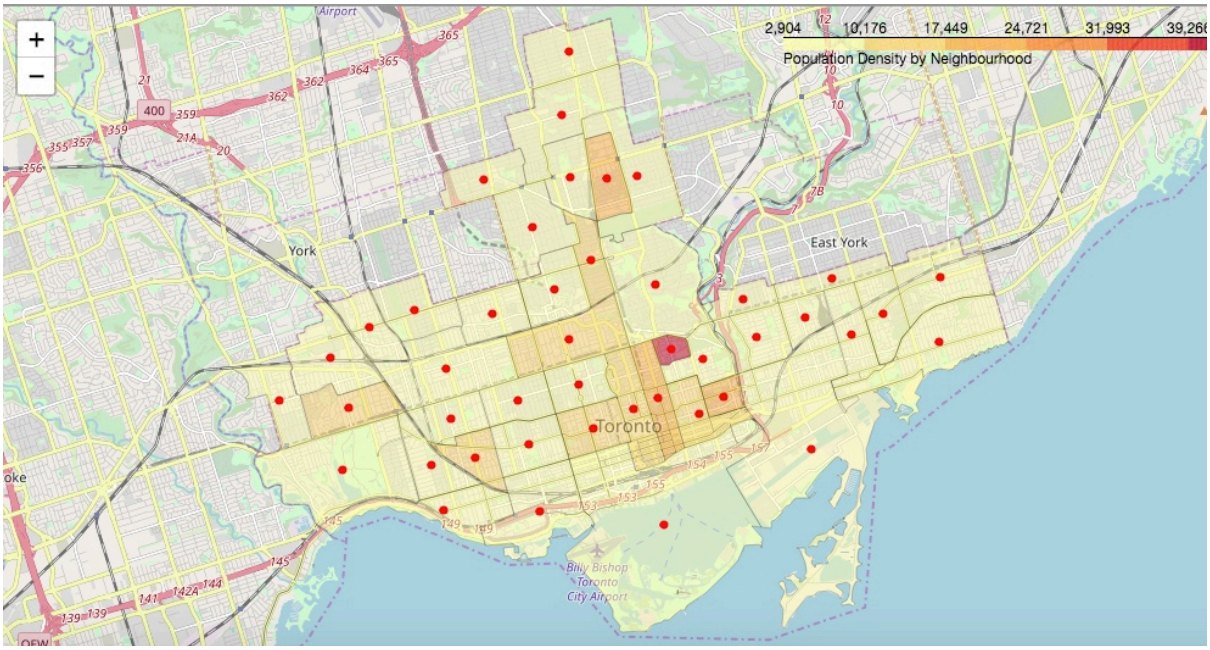
Example data:

Only the neighborhoods' in the former city of Toronto have been retained
After removing several (duplicate) columns, the following columns are available as shown below:

	CDN_Number	Neighbourhood	geometry	Latitude	Longitude	TotalPopulation	TotalArea	AfterTaxHouseholdIncome	PopulationDensity
0	095	Annex	POLYGON ((-79.39414141500001 43.668720261, -79...	43.671585	-79.404000	30526	2.8	49912	10902.0
1	076	Bay Street Corridor	POLYGON ((-79.38751633 43.650672917, -79.38662...	43.657512	-79.385722	25797	1.8	44614	14332.0
2	069	Blake-Jones	POLYGON ((-79.34082169200001 43.669213123, -79...	43.676173	-79.337394	7727	0.9	51381	8586.0
3	071	Cabbagetown-South St.James Town	POLYGON ((-79.376716938 43.662418858, -79.3772...	43.667648	-79.366107	11669	1.4	50873	8335.0
4	096	Casa Loma	POLYGON ((-79.414693177 43.673910413, -79.4148...	43.681852	-79.408007	10968	1.9	65574	5773.0

Display the neighbourhoods on a map by population density

Each neighborhood is shown with a boundary and a color varying from yellow to red, depending on the population density by square kilometer. This is a preliminary exploration into the data we have gathered.



Foursquare API data - Venue Details

The Foursquare API will be used to collect venue data by FSA area. This data can then be combined with the FSA statistical data to be used by the chosen machine learning algorithm to provide insight in business location

5. **Foursquare Venue Categories**:

Each venue on Foursqaure has been assigned to a category. This is the lowest level category that is used by Foursqaure.

Foursqaure usually has two levels of categories, the top level like Food, Arts & Entertainment etc. Under each category there are several sub-categories. For example, Food has a long list of sub-categories including different restaurant types, cafes etc.

There is a special entry point in the Foursqaure API to retrieve all categories and sub-categories. This data will be stored in a table with the following fields:

- **Category**: top level Foursquare venue category
- **Subcategory**: lower level venue category

  The top-level category will be used to categorize venues on a top level as well

Example of the Categories data

	Category	Subcategory
0	Arts & Entertainment	Amphitheater
1	Arts & Entertainment	Aquarium
2	Arts & Entertainment	Arcade
3	Arts & Entertainment	Art Gallery
4	Arts & Entertainment	Bowling Alley

6. **Foursquare Venues by Neighbourhood**

Use the Foursquare Venue Explore API endpoint to gather basic data on venues with a certain radius based on the central coordinates for the area.

The data retrieved in JSON format will be stored in a dataframe with the following columns:

- **CDN_Number**: Three-digit neighbourhood code
- **Neighbourhood**: Name of the neighbourhood the venue is located in
- **Name**: Name of the venue
- **Latitude**: Latitude coordinate of the venue
- **Longitude**: Longitude coordinate of the venue
- **Subcategory**: Lower level category name for the venue
- **Category**: Highest level category, this will be added later

  Note: the venue category will be added to the dataframe using the Foursquare's categories dataframe (5)
  Note: the venue CDN number and neighborhood will be checked against the neighborhood's boundaries

Process retrieving venues by neighbourhood

Loop through the neighborhood dataframe to get the venues within a certain radius of the center coordinates of each neighborhood. Due to the fact that using a radius might cause the API to get venues just outside of the current neighborhood. All the venues found will be verified and if necessary set to the correct neighborhood. The JSON format needs to be parsed and converted to a tabular format to be able to store in a dataframe for further analysis

	CDN_Number	Neighbourhood	Venue	Latitude	Longitude	SubCategory
0	095	Annex	Rose & Sons	43.675668	-79.403617	American Restaurant
1	095	Annex	Ezra's Pound	43.675153	-79.405858	Café
2	095	Annex	Roti Cuisine of India	43.674618	-79.408249	Indian Restaurant
3	095	Annex	Fresh on Bloor	43.666755	-79.403491	Vegetarian / Vegan Restaurant
4	095	Annex	Playa Cabana	43.676112	-79.401279	Mexican Restaurant

Further processing

Process the dataframe by adding each venues main category from the Foursquare Categories data. Also check for the correct neighbourhood to each venue and correct if necessary

The geopandas dataframe has a method to check if a geo-coordinate is within the boundaries of an area, in this case neighborhood boundaries. The df_toronto_nbh dataframe has a column with these boundaries and can be used to verify the venues geo-location.
Note: As reported almost half of neighborhood of all venues has been corrected. This is due to the fact that the Foursquare API endpoint "explore" only accepts a radius from a central point, which can lead to a venue being outside of the neighborhood. 76 venues where entirely outside of the neighborhoods and have been removed

Fixing the categories is proving to be a strenuous process. Foursquare's data is often incomplete so some programmatic fixing is required

Data exploration / Methodology

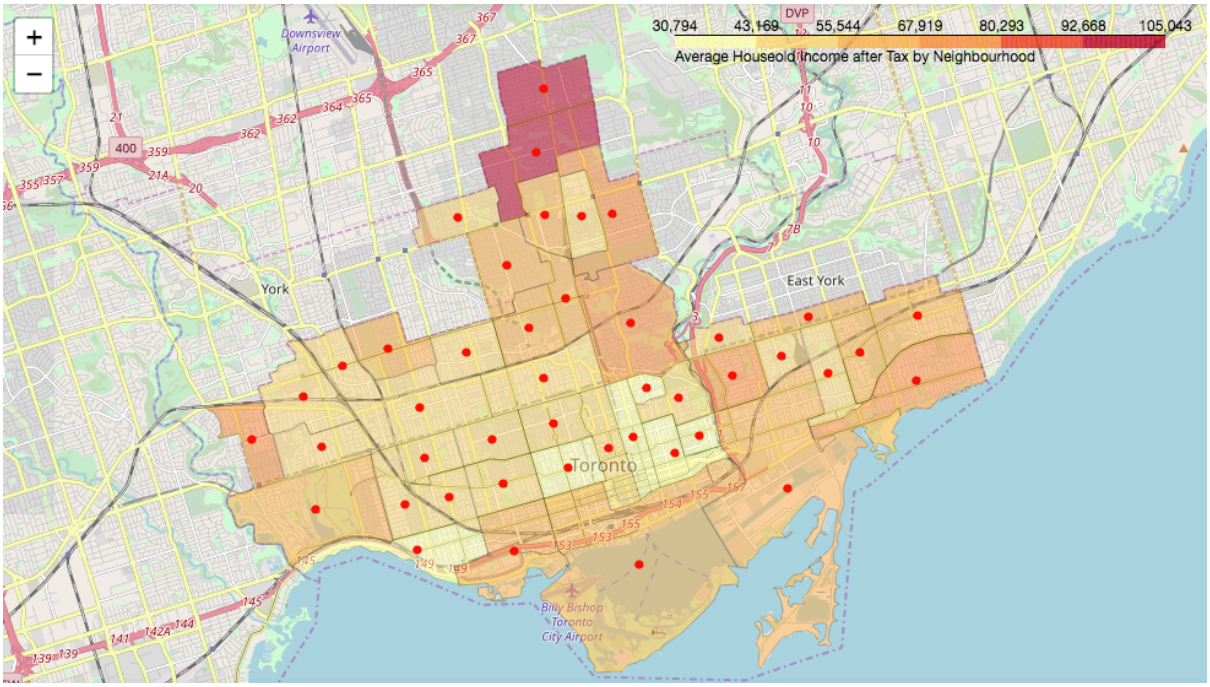
Analysis of the data gathered

To get a general idea, let's see how many venue categories we have found by neighborhood:

8 top level categories with 281 unique venue categories found across 44 neighborhoods and 33 43 venues in total

To get a better idea of the neighborhood's spending power, I used a Choropleth map to visualize the average household income after tax by neighbourhood

- Build a folium choropleth map of the area to show the average incomes by neighbourhood
- This should give some insight for the analysis of the K-Means clustering further on down

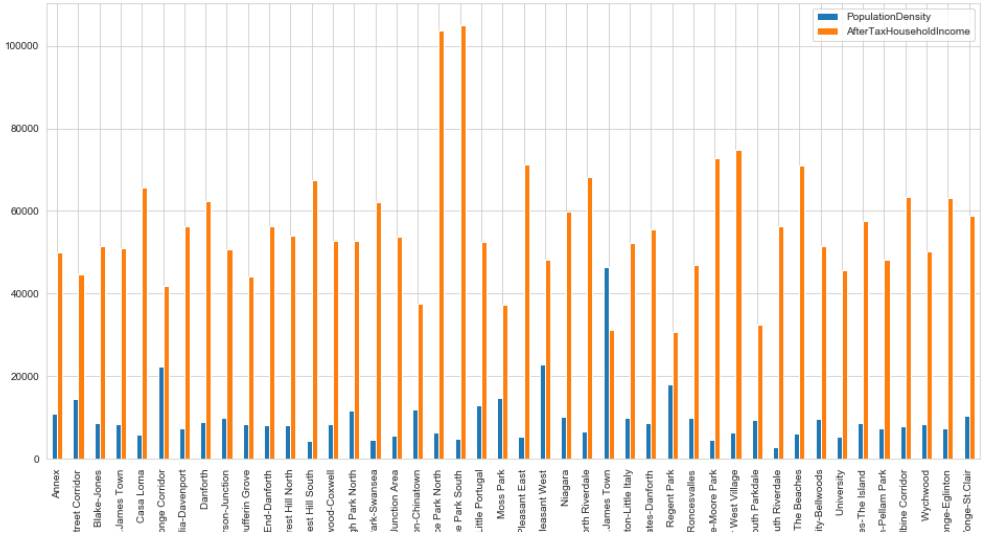


Average household income after tax by neighbourhood

The neighborhoods in the north of Toronto like Lawrence Park South and North are the high-income neighborhoods. Also visible is that neighborhoods closer to the lakeside have a higher average income as well as those on the edges of the city. In the central part of Toronto there are neighborhoods with lesser average income.

For the small businesses within the central neighborhoods of Toronto this doesn't necessarily mean that there is lesser spending power, as there are potentially more offices in the area. The average neighborhood income would not be reflected in the incomes of the people working in these areas.

Let's have a look at a graph of the population density and after-tax income by neighborhood

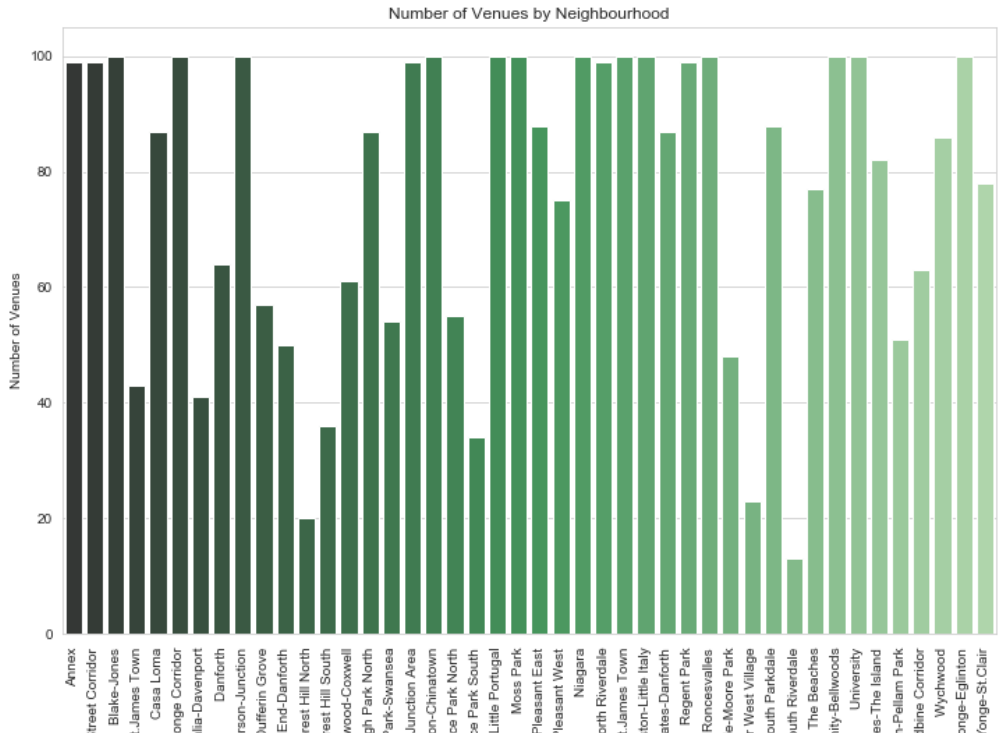


Note: Also note that some of the neighborhoods with a high average income have a lower population density, which could mean a spacious suburb with large housing plots

Number of Venues by Neighbourhood

The graph below visualizes the number of venues by neighborhood. Looking at two of the neighborhoods with the highest incomes, Lawrence Park South & North, we notice that the number of venues is relatively small compared to the others. Forest Hill North & South are similar neighborhoods.

Note: Due to the cap of 100 venues in the Foursquare API endpoint "explore", it is not possible to retrieve more.



Which machine learning algorithm to use?

The goal of for this project was to provide a (future) business owner with business location information for making a more informed decision. Looking a possible suitable machine learning algorithms, I have chosen to focus on either a recommender system or using K-Means clustering for the solution.

After long thought on which machine learning algorithm to use, I have decided to use the K-Means clustering algorithm to provide better insight. Along with the other exploratory data analysis, it should be possible to categorize the clusters as found by the K-Means clustering algorithm.

The first step in preparing for the K-Means algorithm:

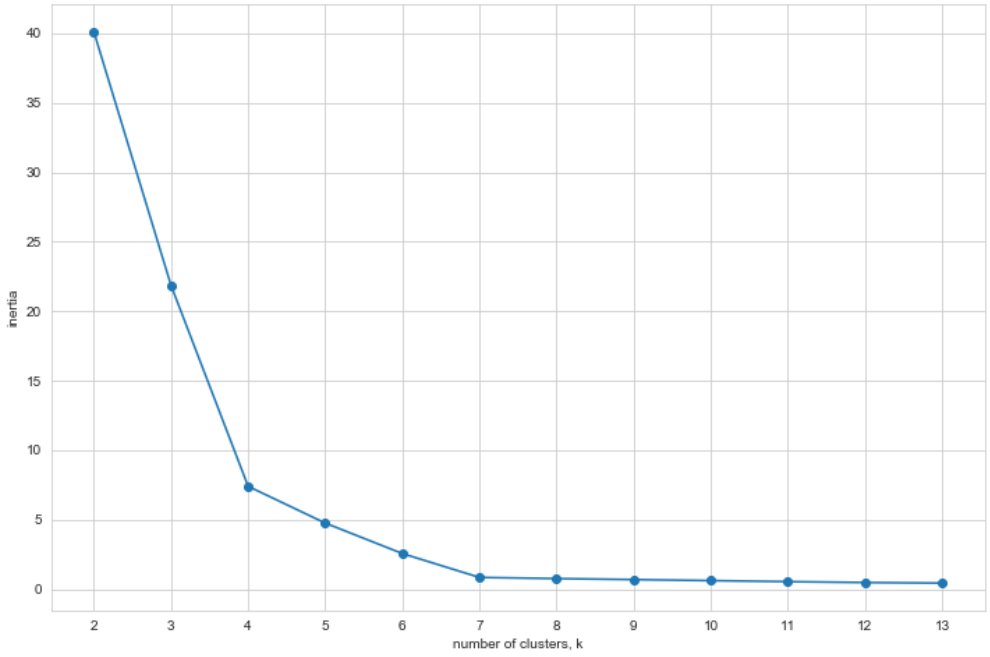
- By using the pandas get_dummies method we are creating a dataframe with a column for each category.
- Then used this dataframe to create a dataframe representing the percentage of venues for a category by neighborhood (they are normalized)
- Add the columns population density and average household income after tax to the dataframe after normalizing the figures

  Note: all data used by the K-Means algorithm needs to be normalized otherwise some features will standout more than others when determining the clustering

Neighbourhood	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport	PopulationDensity	AfterTaxHouseholdIncome
Annex	0.070707	0.000000	0.656566	0.040404	0.040404	0.030303	0.151515	0.010101	0.183297	0.257485
Bay Street Corridor	0.080808	0.010101	0.616162	0.030303	0.040404	0.010101	0.212121	0.000000	0.261906	0.186130
Blake-Jones	0.020000	0.000000	0.720000	0.090000	0.020000	0.000000	0.140000	0.010000	0.130220	0.277270
Cabbagetown-South St.James Town	0.046512	0.000000	0.581395	0.046512	0.209302	0.000000	0.116279	0.000000	0.124467	0.270428
Casa Loma	0.057471	0.000000	0.609195	0.011494	0.057471	0.011494	0.206897	0.045977	0.065751	0.468424

We need to run the K-Means algorithm several times to determine the optimal number of clusters to use when using the model

- Once run the model returns a value of inertia: model.inertia_.
- We are looking for a number of clusters where the inertia visibly flattens out



The elbow diagram visualizes the optima number of clusters to use, in this case 7.

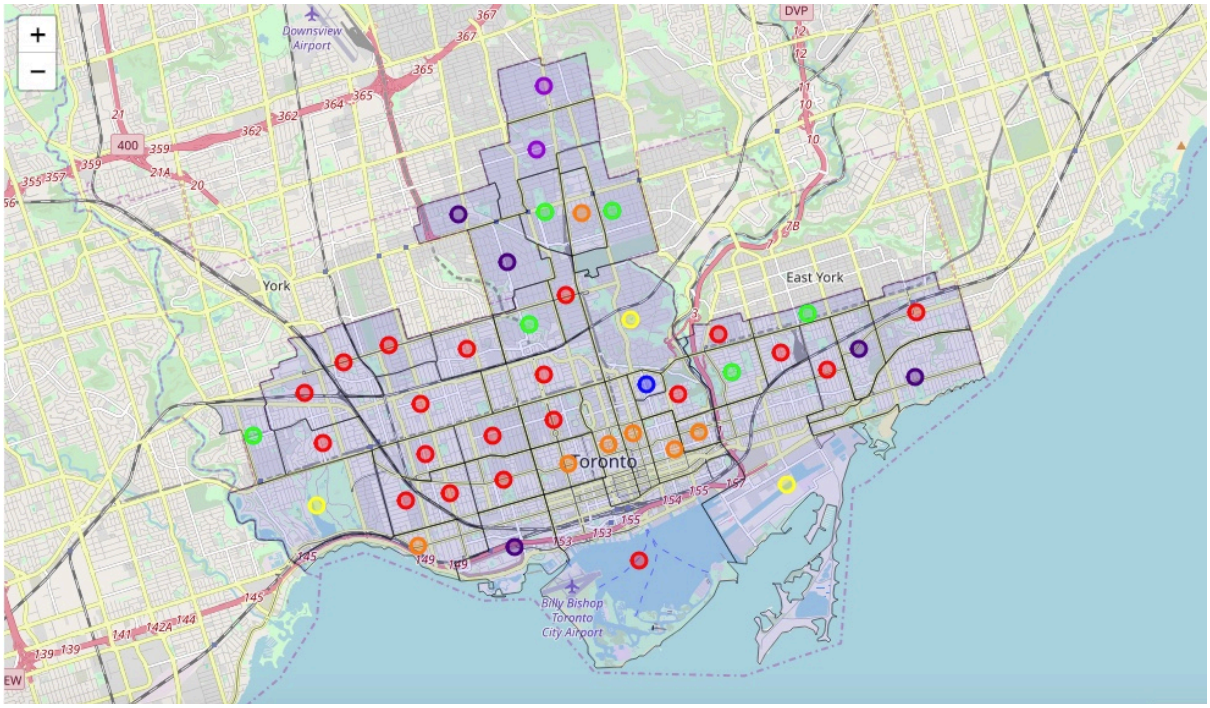
Merge the venues grouped by category dataframe along with the neighborhoods dataframe

- Add the cluster labels column found by the K-Means model which returns a list of labels as the variable K-Means-labels_. These will be added to the dataframe.
- Merge the venues grouped by category dataframe with the neighborhoods dataframe
- Plot the merged dataframe using Choropleth to view the results of the K-Means clustering

Analysis

By visualizing the K-Means clustering by neighborhood on a map we can get a better idea of the geo-spatial aspect. Steps are:

- Map the outlines of the neighborhood boundaries
- Plot the assigned cluster of each neighborhood using a different color for each cluster
- With this plot is should be easier to discover the patterns in the clustering assignment



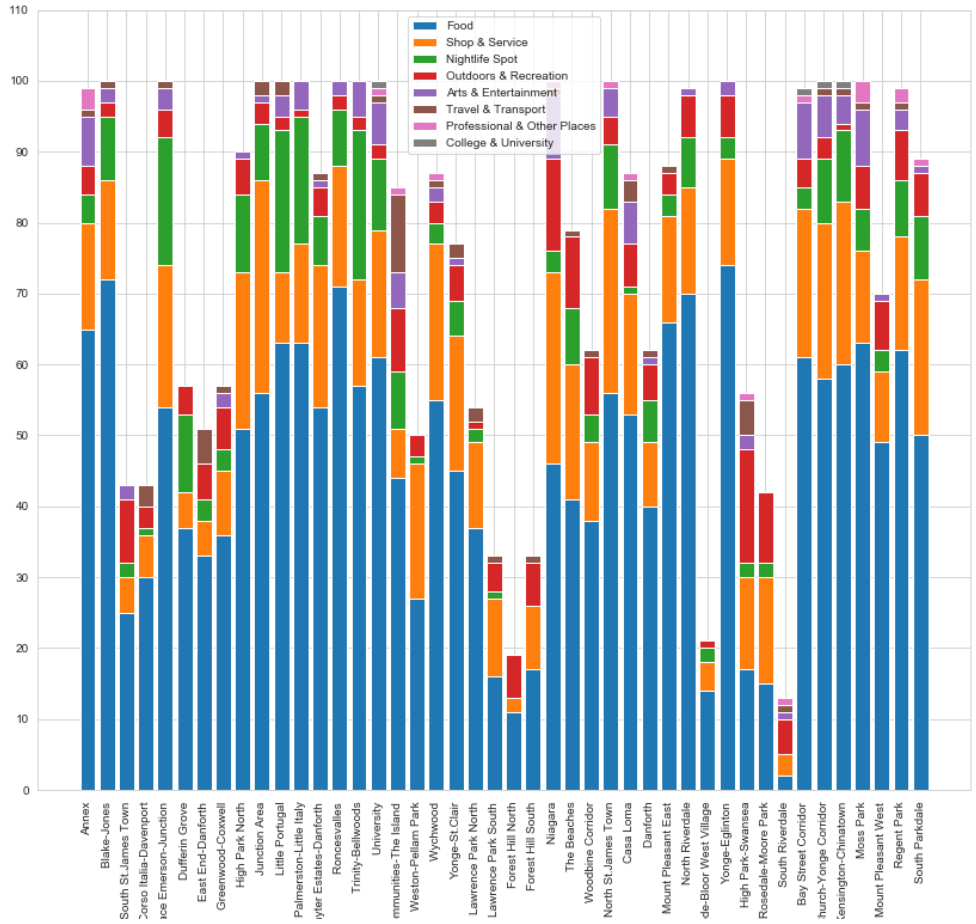
Legend: (including initial analysis on the clustering results after looking at the map)

1. Red = Cluster 0
   - Most prominent, grouped around the downtown area
2. Light Purple = Cluster 1
   - Most northern neighborhoods with a low population density and high avg. income
3. Dark Purple = Cluster 2
   - Appear to be located on outer neighborhoods of the research area or close to a recreational area, needs further investigation
4. Blue = Cluster 3
   - Only one neighborhood represented, needs some further investigation

5. Green = Cluster 4
   - Also appear to be located on outer neighborhoods of the research area
6. Yellow = Cluster 5
   - Seems to be neighborhoods with a larger park or recreational facilities
7. Orange = Cluster 6
   - Mainly grouped in the downtown area

We can do further analysis on the clusters by looking at most prominent venue categories by neighborhood

- Create a cross table dataframe with number of venues by category by neighborhood
- We can use this to create a stacked bar plot showing the proportion of venues by category for each neighborhood
- This plot should also be helpful in detecting patterns behind the clustering assignment



Several observations here:

Neighborhoods by cluster label number:

0. **From Annex to Wychwood (red)**
   Cluster 0 neighborhoods are in the majority. They have a relatively large number of food related venues. And shops, services and nightlife venues are also prominent. The locations are located to the west/north-west
   of the downtown area, as well as to the east

1. **Lawrence Park North and South (light purple)**

   Both neighborhoods have a high average income and low population density which would lead to conclude that these are mainly residential areas

2. **Forest Hill North to Woodbine Corridor (dark purple)**

   There appears to be a relatively large number of shops and services in these neighborhoods And also relatively less nightlife spots (bars, clubs etc.) compared to the cluster 0 (red) neighborhoods. The outdoors and recreational venues are also prominent.

3. **North St.James Town (blue)**

   This neighborhood has been separately clustered due to the fact that there it has the highest population density of all the researched neighborhoods. It has a relatively high number of shops and service venues.

4. **From Casa Loma to Yonge-Eglinton (green)**

   The cluster 4 neighborhoods are located in the north and eastern part of the research area with one exception located in the far west. Mainly on the outskirts. There are a relatively large number of shops and services as well as outdoor and recreational venues. Nightlife venues are also prominent.

5. **High Park-Swansea to South Riverdale (yellow)**

   These neighborhoods are located in recreational areas like parks or close to the waterfront and have a high percentage of outdoor and recreational venues as well as travel and transportation venues (hotels, transportation hubs: bus, metro or train stations)

6. **From Bay Street Corridor to South Parkdale (orange)**

   These are neighborhoods in the downtown area of Toronto with a high number of venues in the food category like restaurants. Shops and services are also prominent as well as venues in the nightlife spot category

In general:

- In almost all neighborhoods venues within the food category (restaurants, coffee shops etc.) are the most prominent
- Shops & Services are the second most prominent all-round (stores, shops, fitness studios etc.)
- Nightlife Spots are the third most prominent all-round (bars, speakeasy's, clubs etc.)

Create a dataframe to visualize the venue categories by venue count by neighborhood

In order to do further analysis on the data I decided to create a dataframe where the cluster numbers, population density, average household income are included as well as the categories and number of venues by category. These will be displayed in order of categories having the most venues. This will make it easier to visualize any features that stand out regarding the cluster assignment by the K-Means model.

	Neighbourhood	Cluster	AvgIncome	PopDensity	1st Category	1st # Venues	2nd Category	2nd # Venues	3rd Category	3rd # Venues	4th Category	4th # Venues
0	Annex	0	49912	10902.0	Food	65	Shop & Service	15	Arts & Entertainment	7	Nightlife Spot	4
2	Blake-Jones	0	51381	8586.0	Food	72	Shop & Service	14	Nightlife Spot	9	Arts & Entertainment	2
3	Cabbagetown-South St.James Town	0	50873	8335.0	Food	25	Outdoors & Recreation	9	Shop & Service	5	Arts & Entertainment	2
6	Corso Italia-Davenport	0	56345	7438.0	Food	30	Shop & Service	6	Outdoors & Recreation	3	Travel & Transport	3
8	Dovercourt-Wallace Emerson-Junction	0	50741	9899.0	Food	54	Shop & Service	20	Nightlife Spot	18	Outdoors & Recreation	4
9	Dufferin Grove	0	44145	8418.0	Food	37	Nightlife Spot	11	Shop & Service	5	Outdoors & Recreation	4
10	East End-Danforth	0	56179	8223.0	Food	33	Outdoors & Recreation	5	Shop & Service	5	Travel & Transport	5

Also look at the mean and medians of average income and population density

	AvgIncome	PopDensity
count	44.000000	44.000000
mean	55981.727273	9972.568182
std	15130.504763	7034.026401
min	30794.000000	2904.000000
25%	48171.000000	6336.750000
50%	53315.500000	8461.000000
75%	62678.250000	10153.500000
max	105043.000000	46538.000000

Results

Results of analyzing K-Means clustering of neighborhoods

- **Cluster 0 (red)**

  These neighborhoods are in the majority. The neighborhoods have a similar average household income as those of cluster 2 which are around the overall median of 53,315 Canadian dollars. With two exceptions, High Park North and Little Portugal, the population density is close to the overall median populating density.
  They are located around the downtown area, mainly to the west/north west and several to the east as well as one below the downtown area close to the lake. On average there are a higher number of venues in the food, shops and service and nightlife spot categories.

  Based on the fact that the locations are not too far from the center of town and the area isn't a low income area these would be neighborhoods to consider when opening a restaurant or shop. Due to the relatively high number of venues in the nightlife category these neighborhoods should be considered when looking for a location like a bar or nightclub. Venues like these attract other venues because the areas are already known for nightlife with higher visitors numbers. As in any case you still would need to do local research as to exactly where you would prefer to locate.

- **Cluster 1 (light purple)**

  This cluster contains the two neighborhoods with the highest average household income and low population density. This would indicate a richer residential area. The number of venues in total is relatively low compared to other clusters. The venues under category shops and services is high in ratio to the food venues category compared to other neighborhood clusters.

  Based on the fact that these are high income residential areas there should be some possibilities to open a high-end restaurant or delicatessen take away. Looking at the lower population density and the fact that the neighborhoods are farther away from the downtown area, there probably not that much foot traffic. As for setting up a smaller store or shop more local research would be necessary: somewhere with ample parking facilities, easy access and other stores in the vicinity. Another idea would be some types of service venues like a yoga studio or fitness center where people can go to after work or during the weekends.

- **Cluster 2 (dark purple)**

  These neighborhoods have a higher than average household income above the median of 53,315 Canadian dollars. Appear to be located on the outer edge of the boundaries of the research area or close to a recreational area like a park. The total number of venues under the shops and services is percentage-wise higher than most neighborhood clusters. The number of venues in the category outdoors and recreational is higher than average. There also appear to be less venues in the category nightlife spots (bars etc.)

  Based on the fact that there are few nightlife venues it would seem not to be an area to open a new venue of this category. That is unless you have done more local research

and found a location that has enough foot traffic and visitors in the evenings. Because of there are a higher number of venues in the outdoors and recreational category, it might be an idea to open a café or day restaurant where visitors can get some refreshment and a bite to eat. Area's like these usually have a lot of day visitors.

- **Cluster 3 (blue)**

  The neighborhood North St. James Town has a very high population density and much lower than median household income. Presumably this is why it has been assigned a separate cluster. This area might be a good place to open a small shop or simple restaurant or takeaway to attract local residents. The population density is high so there should be enough foot traffic as well. Opening a bar or club could be a possibility as well.

- **Cluster 4 (green)**

  The neighborhoods have a relatively high average income and lower than median population density which leads me to believe that these are mainly residential areas. The neighborhoods are located further away from the center of town. Business in these areas would depend on visitors coming by car, there for ample parking facilities. Having other venues in the nearby vicinity would attract consumers due to the convenience of having several shops, restaurants etc. close by.

  A good restaurant or takeaway with a popular and not too specific cuisine would be possible idea. Also some types of service venues like a yoga studio or fitness center where people can go to after work or during the weekends.

- **Cluster 5 (yellow)**

  The neighborhoods High Park-Swansea, Rosedale-Moore Park and South Riverdale have a higher than median average household income and a much lower than median population density. This is due to the fact that these neighborhood all contain larger parks or beach areas within their boundaries.

  As with the cluster 2 neighborhoods, because of there are a higher number of venues in the outdoors and recreational category, it might be an idea to open a café or day restaurant where visitors can get some refreshment and a bite to eat. Area's like these usually have a lot of day visitors. A business based on recreation could be a possibility considering that there are parks or beaches close by.

- **Cluster 6 (orange)**

  These neighborhoods have a relatively low average household income compared to the overall median average income of 53,351 Canadian dollars. The population density is high. This is most likely due to the fact that these neighborhoods are located in the downtown area where the real estate prices are high leading to more concentration by square kilometer. Having a lower average household income doesn't necessarily mean that there is less spending power for consumers, as office workers tend to go out for a bite to eat and some refreshment. Considering these neighborhoods are in the downtown area with larger office buildings and public administration centers, there should be enough foot traffic during most hours of the day. There are a large number of colleges, schools and university faculty buildings in the area. This means that there are students in the vicinity as well.

I would think a take-away, café or coffee shop would be appropriate in these areas. Possibly a trendy restaurant or bar as well.

Discussion

I decided to use the main venue categories as there are so many different types of venues in the Toronto area not located in one specific area. That is why I decided to group these by category so the choices would not be spread out too thinly. Adding the average household income after tax and the population density proved to be a good idea looking at the way the clustering of neighborhoods was fit by the K-Means model. The K-Means model came up with a surprisingly good categorization of the neighborhoods which has proven to be useful for advising for a business location. Using the map visualizations are very informative and useful for discovering (hidden) patterns which you would struggle to find otherwise.

With more data and time I could imagine coming up with more precise recommendations for business locations.

Data like:

- Real estate rent and sale price
- Business turnover statistics
- Building, housing and business categories with their numbers
  Although Foursquare is useful to a certain degree, I would think that venues like restaurants and cafes where you can sit and meet people, would be more represented that small shops or takeaways where you usually spend a short amount of time.
- Consumer data, like number of daily visits, time spent
  Foursquare only has this data for venue owners and is not available for the other API users.
- Work statistics: number of workers during the day, average working hours with time frame
- Traffic and travel statistics and parking facilities

I would like to try to come up with a recommender system, something along the lines of given the type of business, expected consumer turnover and business expenses, which neighborhoods would the system recommend.

Conclusion

The combination of business location allocation, machine learning and geo-spatial tools are beneficial in providing information for possible business location decisions. It has proven to be really interesting area to get involved in. Even something relatively simple as the K-Means algorithm has come up with some surprising results in this project. The downside of the K-Means clustering is that it doesn't come up with precisely defined and descriptive labels for the categories. You have to come up with your own analysis as to why a neighborhood falls under a certain category. Local government area statistics are readily available and simple to work with. This data has proven to be beneficial to the location analysis as well as being helpful in defining the assigned clustering. The python programming language has a plethora of libraries available to handle all kinds of machine learning algorithms and statistical data.

I have learnt a lot from this project as doing is much more educational than reading.