Final Report Capstone Project
Battle of Neighborhoods- Finding New Place to Stay at Toronto, Canada

Introduction

Finding new place to stay is a hard thing to do. Several parameters need to be considered to find the best place to stay at, for example: house prices, crime rate, population density, and nearby venues. House price is important because usually people who want to find a new house have certain budgets and they want to buy (or rent) a new house within their budgets. Even if they have huge budgets, they would want to find the best place with minimum price. House price may be affected by several variables such as crime rate, population density, location, the number of floors, house materials, etc.

Besides the house price, they also need to find place with low crime rate, low population density or high population density depending on preference, and nearby venues that fit their daily activities. The parameters mentioned above can be correlated to each other.

In this project, I use Toronto as the object. Toronto's Neighborhoods are clustered based on the parameters above and the best neighborhoods will be chosen. Although there are more parameters that need to be considered, for the sake of simplicity and data availability, I limit the parameters considered to house prices, crime rate, population density, and nearby venues only. Moreover, the final decision is decided based on preferences of individuals and more research should be made. If possible, visiting the neighborhood in person will be better. Therefore, the results of this project can only be used to narrow the options of neighborhood in Toronto to stay at.

Business Problem

There are around 140 neighborhoods in Toronto. Choosing place to stay in Toronto can be hard because of this. Therefore, reducing the options of neighborhood is important not only for the potential buyer (person who wants to relocate to Toronto), but also for the real estate company. For the potential buyers who have budgets and other preferences, reducing the number of neighborhood options is important because it helps them to reduce the time and effort needed to look into the neighborhood one by one. On the other hand, for the real estate companies, reducing the number of neighborhood options is important because they can focus to promote certain neighborhoods to their potential buyers (or renters).

Data

Several parameters that are used in this study to narrow down the Toronto's neighborhood options are:

1. House price
2. Crime rate
3. Population density
4. Nearby venues

House price is affected by many parameters such as locations, number of bedrooms, conditions, etc. However, in this project, I only consider the average house price in each neighborhood. The source for this data is https://www.zolo.ca/toronto-real-estate/neighborhoods. This website provides average house price, percentage of house sold after 10 days of listing, the number of active listings, and percentage of house sold after the potential buyers asked about the unit in each neighborhood. Based on the source, the data is based on the last 28 days data.

The source of crime rate and population density data is Toronto Public Service Public Safety Data Portal (https://opendata.arcgis.com/datasets/af500b5abb7240399853b35a2362d0c0_0.csv?outSR=%7B%22latestWkid%22%3A26717%2C%22wkid%22%3A26717%7D). This source provides the number of crimes occured in each neighborhood every year from 2014-2018 complete with the types, the number of population and area for each neighborhood. The population density can be calculated using number of populations divided by the neighborhood area.

Nearby venues can be determined using Foursquare API (https://foursquare.com/). Foursquare API provides many data including venue name, type, photos, tips, and many others.

	Neighbourhood	Average sale price	Population	Shape__Area	Total_Crime_2018	density	Latitude	Longitude
0	Agincourt North	703000	31528	7.261857e+06	513.0	0.004342	43.808038	-79.266439
1	Humber Summit	648000	14188	7.966905e+06	59.0	0.001781	43.760078	-79.571760
2	Kingsview Village	619000	14306	5.063371e+06	263.0	0.002825	43.699539	-79.556346
3	West Hill	663000	36371	9.625440e+06	139.0	0.003779	43.768914	-79.187291
4	Agincourt South	676000	28410	7.873163e+06	154.0	0.003608	43.785353	-79.278549

Methodology

To be able to fulfil the objective of this project which is to narrow down the neighborhood options to live at Toronto, several things need to be done with the data mentioned in previous section. The steps are:

1. Exploratory data analysis

   This step can be done by plotting the boxplot or distribution of each variables. In this project, I use the boxplot. By plotting the boxplot, we can get the information of skewness of the data, outliers, quartiles, maximum and minimum values. We can remove the outliers if necessary. We also need to look whether the variables are correlated to each other or not to determine which machine learning method to use and to possibly reduce the dimensions. This can be done by plotting scatter plots between each variable or determine the pearson coefficient and the confidence level. In this project, we will determine the correlations visually.

2. Clustering analysis I

   This step is done to find the neighborhoods with low average house price, low crime rate, and low population density. The low average house price and low crime rate are obvious parameters. But low population density is a matter of preference. I personally prefer living at not too crowded area, so I will use low population density as one parameter to determine the cluster. For other

preferences the same step can also be done with several modifications. This step can be done using k-means clustering analysis.

3. Finding the nearby venues for chosen neighborhoods

   This step is done using the foursquare API. For the chosen neighborhoods (from step 2), top 5 venues within 500 meters are populated and included into the data frame.

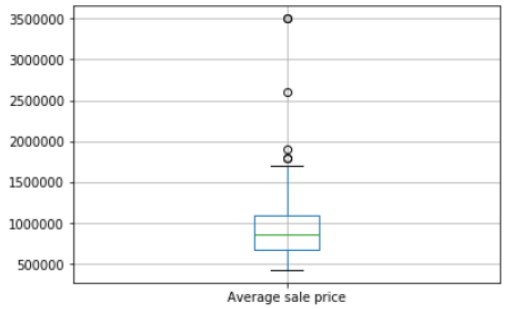
4. Clustering analysis II

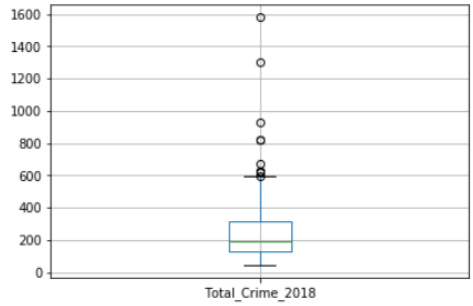
   The second clustering analysis using k-means clustering method is done to determine the clusters of neighborhoods with similar venue types. This is done to further narrow down the neighborhood options based on personal preferences. For example, if the potential buyer is interested in finding a place that is near restaurants, we can suggest the cluster of neighborhoods that has restaurants as the most commonly visited venue.

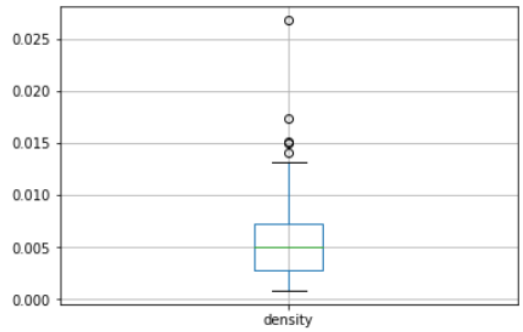
Exploratory data analysis

Boxplots

Based on the boxplots below for each variable, the distributions are right skewed distribution with outliers on the right side of the data. The next step of the exploratory data analysis is to plot the scatter plots of each variable to find any correlations between the variables. Because of this, outliers are removed from the data to better visualize the data in the scatter plot. Moreover, at the end of the project, we are filtering out the high price, high crime rate and high population density anyway, so removing the outliers which are located at high values of each variable helps with this objective.

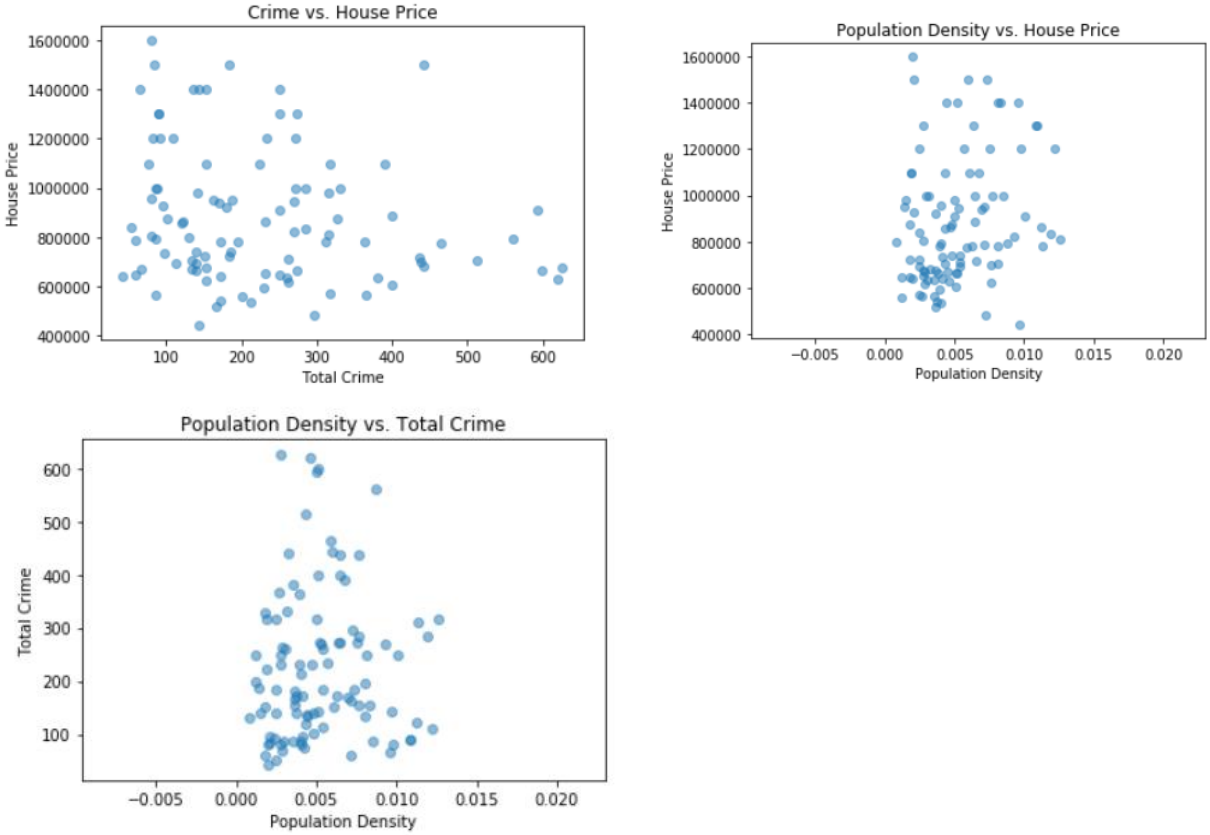






Scatter Plot

Visually, the scatter plots between the 3 variables show that there is no clear correlation between the variables. Therefore, we will use all three variables in the machine learning analysis (clustering). Clustering analysis is done as unsupervised learning because there is no clear result in the dataframe.



Clustering Analysis I

Normalize Data

The data is normalized because the scales of the variables are incomparable. For example, the house price scale is from hundred thousands to millions, while the population densities are less than 1. Conducting machine learning analysis immediately without normalizing the data may heavily be affected by the house price variable. Therefore, normalizing the data needs to be done to make sure that the variables are on the same scale (0-1) and that all variables have same effect to the results.

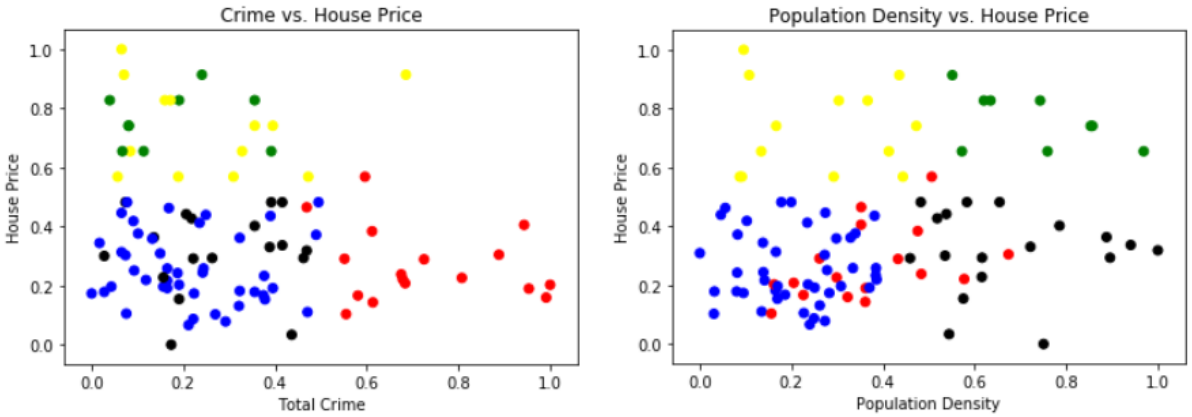
	Neighbourhood	Average sale price	Population	Shape__Area	Total_Crime_2018	density	Latitude	Longitude
0	Agincourt North	0.226057	31528	7.261857e+06	0.807560	0.298687	43.808038	-79.266439
1	Humber Summit	0.178602	14188	7.966905e+06	0.027491	0.080287	43.760078	-79.571760
2	Kingsview Village	0.153581	14306	5.063371e+06	0.378007	0.169372	43.699539	-79.556346
3	West Hill	0.191544	36371	9.625440e+06	0.164948	0.250673	43.768914	-79.187291
4	Agincourt South	0.202761	28410	7.873163e+06	0.190722	0.236159	43.785353	-79.278549

K-Means Clustering

K-means clustering is one of unsupervised learning where the code divides the data into k number of clusters based on their similarities. In this project, the number of clusters used is determined using trial

and error. After each trial, the results are visualized. The number of clusters that will be used is 5. It is because using 5 clusters, we can fulfil our objective to determine neighborhoods with low price, low crime rate and low population density. Using number of clusters lower or higher than 5 result in unclear cluster properties.

Based on the clustering result, we will proceed with the neighborhood cluster 2, which has low price, low crime rate, and low population density. In the scatter plots, the neighborhoods are marked with blue color.



Finding the nearby venues for chosen neighborhoods

Foursquare API is used to find nearby venues for the chosen neighborhoods. In this project, we are using 500 meters radius from the location data of each neighborhood. After that, the top 5 venues are populated and included into the dataframe.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Agincourt South	Chinese Restaurant	Rental Car Location	Coffee Shop	Asian Restaurant	Hong Kong Restaurant
1	Banbury	Park	Auto Garage	Tennis Court	Falafel Restaurant	Food & Drink Shop
2	Bayview Village	Pizza Place	Bank	Outdoor Supply Store	Clothing Store	Sandwich Place
3	Bayview Woods	Dog Run	Yoga Studio	Farmers Market	Food Court	Food & Drink Shop
4	Beechborough	Furniture / Home Store	Dessert Shop	Italian Restaurant	Discount Store	Auto Garage

Clustering Analysis II

The second K-means clustering analysis is done to cluster the neighborhoods with similar top 10 nearby venues. The number of clusters are also determined by trial and error. After each trial of number of clusters, cluster exploratory is done by observing each cluster. Based on the cluster exploratory, I determine to use 5 clusters.

Cluster Exploratory

First Cluster

First cluster has park as the most common venue. These neighborhoods are suitable for people who enjoy spending time outside, especially when the weather is nice.

	Neighbourhood	density	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Kingsview Village	0.169372	43.699539	-79.556346	0	Park	Yoga Studio	Falafel Restaurant	Food & Drink Shop	Flea Market
13	Banbury	0.274121	43.742796	-79.369957	0	Park	Auto Garage	Tennis Court	Falafel Restaurant	Food & Drink Shop
61	Eringate	0.000000	43.662273	-79.576516	0	Park	Yoga Studio	Falafel Restaurant	Food & Drink Shop	Flea Market
67	Henry Farm	0.240083	43.769509	-79.354296	0	Park	Tennis Court	Yoga Studio	Falafel Restaurant	Food & Drink Shop

Second Cluster

Second cluster has pizza place and fast food restaurants as the most common venues. Several neighborhoods that have venues other than pizza place or fast food restaurants in the most common venue, have fast food restaurants (pizza, fried chicken, sandwich, falafel, etc.) in their top 5 most common venues. This cluster is suitable for busy people who need to grab fast foods frequently.

	Neighbourhood	density	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	West Hill	0.250673	43.768914	-79.187291	1	Pizza Place	Breakfast Spot	Fast Food Restaurant	Coffee Shop	Smoothie Shop
5	Keelesdale	0.384223	43.690158	-79.474998	1	Museum	Discount Store	Sandwich Place	Yoga Studio	Falafel Restaurant
8	Bayview Village	0.334206	43.769197	-79.376662	1	Pizza Place	Bank	Outdoor Supply Store	Clothing Store	Sandwich Place
18	Elms	0.030735	43.696998	-79.521883	1	Pizza Place	Park	Skating Rink	Grocery Store	Soccer Field
22	Brookhaven	0.385946	43.700778	-79.494522	1	Pizza Place	Vietnamese Restaurant	Grocery Store	Supermarket	Electronics Store
28	York University Heights	0.165104	43.758781	-79.519434	1	Pizza Place	Discount Store	Fast Food Restaurant	Gas Station	Coffee Shop
31	Morningside	0.227095	43.782601	-79.204958	1	Park	Coffee Shop	Sandwich Place	Supermarket	Pharmacy
32	Weston	0.308177	43.700161	-79.516247	1	Coffee Shop	Train Station	Pharmacy	Fried Chicken Joint	Convenience Store
34	Hillcrest Village	0.166166	43.799664	-79.365019	1	Pharmacy	Pool	Shopping Mall	Park	Sandwich Place
35	Malvern	0.273170	43.809196	-79.221701	1	Pizza Place	Gym / Fitness Center	Fast Food Restaurant	Pharmacy	Grocery Store
42	Bendale	0.370557	43.753520	-79.255336	1	Intersection	Chinese Restaurant	Grocery Store	Dog Run	Fast Food Restaurant
43	Cliffside	0.297947	43.711170	-79.248177	1	Pizza Place	Breakfast Spot	Coffee Shop	Pub	Sandwich Place
45	Briar Hill	0.141443	43.703045	-79.451344	1	Coffee Shop	Mediterranean Restaurant	Trail	Bike Shop	Fast Food Restaurant
50	Clairlea	0.329067	43.708823	-79.295986	1	Fast Food Restaurant	Sandwich Place	Diner	Grocery Store	Restaurant
60	Eglinton East	0.095264	43.739622	-79.232290	1	Fast Food Restaurant	Indian Restaurant	Train Station	Sandwich Place	Restaurant
69	High Park	0.381446	43.653867	-79.466864	1	Convenience Store	Pizza Place	Pool	Mexican Restaurant	Pub
97	Parkwoods	0.272589	43.758800	-79.320197	1	Pizza Place	Chinese Restaurant	Shopping Mall	Liquor Store	Bank
106	Rustic	0.234548	43.713366	-79.504504	1	Fast Food Restaurant	Gas Station	Caribbean Restaurant	Yoga Studio	Farmers Market
119	Woburn	0.169811	43.759824	-79.225291	1	Fast Food Restaurant	Coffee Shop	Pizza Place	Bank	Gym

Third Cluster

Third cluster contains the neighborhoods that have non-American restaurant (chinese, korean, mexican, asian, italian, etc.) as the top 5 most common venues. These neighborhoods are suitable for people who loves non-american foods and frequently order them at restaurants.

	Neighbourhood	density	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Humber Summit	0.080287	43.760078	-79.571760	2	Pharmacy	Park	Bakery	Empanada Restaurant	Gift Shop
4	Agincourt South	0.236159	43.785353	-79.278549	2	Chinese Restaurant	Rental Car Location	Coffee Shop	Asian Restaurant	Hong Kong Restaurant
15	L'Amoreaux	0.282766	43.799003	-79.305967	2	Chinese Restaurant	Coffee Shop	Athletics & Sports	Bus Stop	Yoga Studio
16	Newtonbrook West	0.199539	43.793886	-79.425679	2	Korean Restaurant	Coffee Shop	Vietnamese Restaurant	Hardware Store	Middle Eastern Restaurant
24	Yorkdale	0.185932	43.724642	-79.447503	2	Clothing Store	Coffee Shop	Cosmetics Shop	Fast Food Restaurant	Toy / Game Store
26	Etobicoke West Mall	0.135088	43.643549	-79.565325	2	Hotel	Coffee Shop	Café	Bank	Farmers Market
36	West Humber	0.031865	43.678692	-79.483427	2	Furniture / Home Store	Seafood Restaurant	Park	Locksmith	Yoga Studio
41	Beechborough	0.045986	43.695030	-79.471683	2	Furniture / Home Store	Dessert Shop	Italian Restaurant	Discount Store	Auto Garage
54	Dorset Park	0.249233	43.752847	-79.282067	2	Indian Restaurant	Gaming Cafe	Asian Restaurant	Bakery	Beer Store
56	Downsview	0.081515	43.749299	-79.462248	2	Coffee Shop	Metro Station	Gym Pool	Gym / Fitness Center	Playground
70	Highland Creek	0.055583	43.790117	-79.173334	2	Neighborhood	Yoga Studio	Ice Cream Shop	Food & Drink Shop	Flea Market
74	Junction Area	0.082492	43.665478	-79.470352	2	Coffee Shop	Italian Restaurant	Bar	Thai Restaurant	Café
85	Mimico	0.277772	43.616677	-79.496805	2	Playground	Skating Rink	Bar	Bakery	Yoga Studio
87	Mount Dennis	0.261845	43.686960	-79.489551	2	Coffee Shop	Furniture / Home Store	Pizza Place	Grocery Store	Pub
90	New Toronto	0.137747	43.600763	-79.505264	2	Mexican Restaurant	Indian Restaurant	Breakfast Spot	Gym	Italian Restaurant
95	O'Connor	0.339629	43.750275	-79.317901	2	Coffee Shop	Medical Center	Restaurant	Tennis Court	Yoga Studio
109	South Riverdale	0.177865	43.665470	-79.352594	2	Vietnamese Restaurant	Chinese Restaurant	Bakery	Light Rail Station	Bar
116	Victoria Village	0.138696	43.732658	-79.311189	2	Middle Eastern Restaurant	Bus Line	Thai Restaurant	Spa	Yoga Studio

Fourth Cluster

The fourth cluster contains only one neighborhood. This neighborhood contains playground as the most common venue. This neighborhood is suitable for family with kids.

	Neighbourhood	density	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
6	Steeles	0.384632	43.816178	-79.314538	3	Playground	Deli / Bodega	Food & Drink Shop	Flea Market	Fish Market

Fifth Cluster

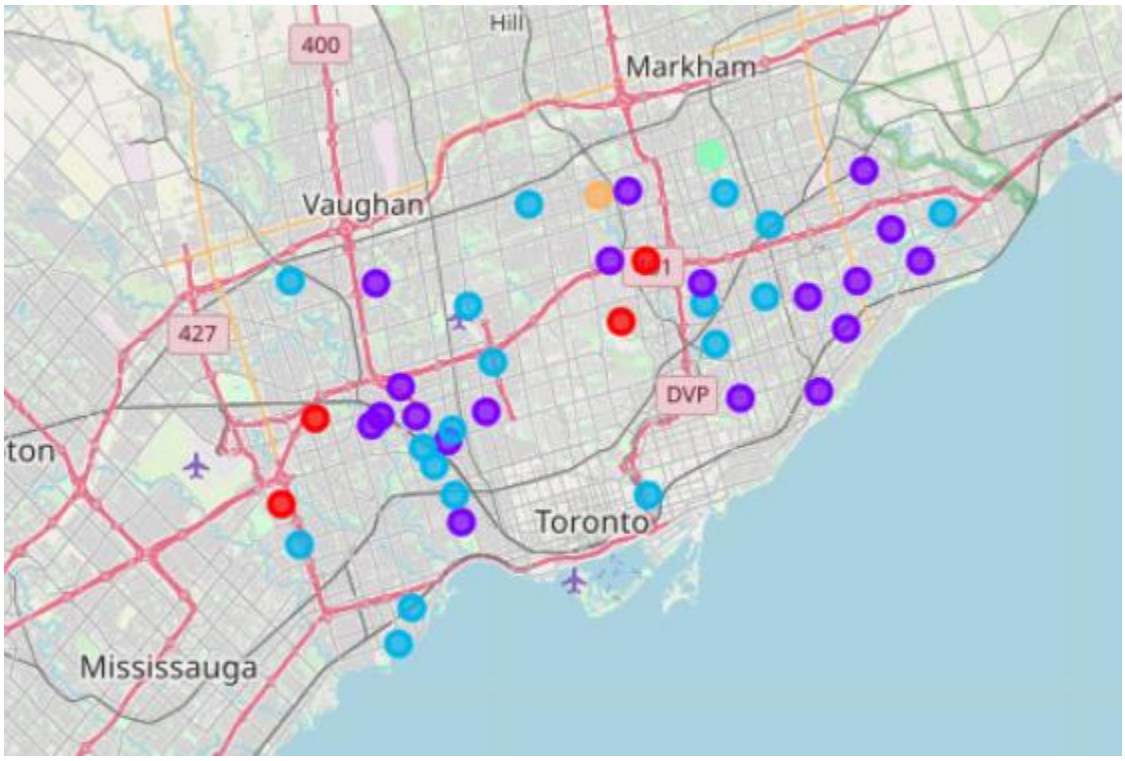
The fifth cluster also contains only one neighborhood. This neighborhood contains dog run as the most common venue. This neighborhood is suitable for people with dogs.

	Neighbourhood	density	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
29	Bayview Woods	0.102567	43.798127	-79.382973	4	Dog Run	Yoga Studio	Farmers Market	Food Court	Food & Drink Shop

Result and Discussion

Visualizing the Neighborhood Clusters in Map

The map provides the locations of neighborhoods and the clusters. Red dots are the first cluster, purple dots are second cluster, light blue dots are third cluster, light green dot is fourth cluster, and yellow dot is the fifth cluster.



From the results of the first clustering, which is based on house price, crime rate, and population density, we can determine the neighborhoods with low house price, low crime rate and low population density (blue dots), the list of the neighborhoods is provided below. However, if high population density is prefered, neighborhoods that clustered with black dots can be chosen, the list of neighborhoods is provided below. It depends on preferences.

Moving on from the first clustering results, the top 5 nearby venues for neighborhoods with low house price, low crime rate and low population density are populated and included in the dataframe. After that the second clustering is done with 5 clusters.

The first cluster most common venue is park. The neighborhoods in this cluster are suitable for people who enjoy outdoor activities especially when the weather is nice. The second cluster most common venue is fastfood restaurants. The neighborhoods in this cluster are suitable for people who like to order fastfood probably because they are busy. The third cluster most common venue is non-american restaurants. The neighborhoods in this cluster are suitable for people who enjoy non-american foods such as Asian, Mexican, Italian, Middle Eastern foods. The fourth and fifth clusters contain only one neighborhood each. The fourth cluster's neighborhood most common venue is playground which is suitable for families that have kids, while the fifth cluster's neighborhood most common venue is dog run, which is suitable for people who own dogs. The lists of neighborhoods for each cluster can be found below.

Neighborhoods with low price, low crime rate, and low population density

Humber Summit	Elms	West Humber	Henry Farm	South Riverdale
Kingsview Village	Brookhaven	Beechborough	High Park	Victoria Village
West Hill	Yorkdale	Bendale	Highland Creek	Woburn
Agincourt South	Etobicoke West Mall	Cliffside	Junction Area	
Keelesdale	York University Heights	Briar Hill	Mimico	
Steeles	Bayview Woods	Clairlea	Mount Dennis	
Bayview Village	Morningside	Dorset Park	New Toronto	
Banbury	Weston	Downsview	O'Connor	
L'Amoreaux	Hillcrest Village	Eglinton East	Parkwoods	
Newtonbrook West	Malvern	Eringate	Rustic	

Neighborhoods with low price, low crime rate, and high population density

Westminster	Little Portugal
Maple Leaf	Long Branch
Mount Olive	Mount Pleasant West
Willowdale East	Oakridge
Bay Street Corridor	Tam O'Shanter
Don Valley Village	University
Caledonia	Weston
Dovercourt	Woodbine
Flemingdon Park	
Greenwood	

First Cluster, Common Place = Park

Kingsview Village
Banbury
Eringate

Henry Farm

Second Cluster, Common Place = Fastfood Restaurants

West Hill	Bendale
Keelesdale	Cliffside
Bayview Village	Briar Hill
Elms	Clairlea
Brookhaven	Eglinton East
York University Heights	High Park
Morningside	Parkwoods
Weston	Rustic
Hillcrest Village	Woburn
Malvern	

Third Cluster, Common Place = Non-American Restaurants

Humber Summit	Highland Creek
Agincourt South	Junction Area
L'Amoreaux	Mimico
Newtonbrook West	Mount Dennis
Yorkdale	New Toronto
Etobicoke West Mall	O'Connor
West Humber	South Riverdale
Beechborough	Victoria Village
Dorset Park	
Downsview	

Fourth Cluster, Common Place = Playground

Steeles

Fifth Cluster, Common Place = Dog run

Bayview Woods

Conclusion

Toronto has around 140 neighborhoods and it is hard to find the best place to move to. To narrow down the options, we use house price, crime rate, population density, and nearby venues data. Clustering analysis is done using house price, crime rate, population density data to find neighborhoods with low house price, low crime rate, and low population density. We can also determine neighborhoods with low house price, low crime rate and high population density if potential buyer prefers to live at crowded place.

Moving on with the low house price, low crime rate, and low population density neighborhoods, Foursquare API is used to find top 5 most common venue within 500 meters. Another clustering analysis is done and 5 clusters are determined. The first cluster contains park as the most common place with is suitable for people who enjoy outdoor activities, the second cluster's most common venue is fastfood restaurants, the third cluster's most common venue is Non-American restaurant, the fourth and fifth cluster's most common venues are playground and dog run respectively. Depending on preferences, one can choose to look into neighborhoods in preferred nearby venues.

Having said that, caution needs to be applied when looking into the Neighborhood because the analysis done in this study is simplified where the clusters are determined only on 3 variables which may affect each other even though the scatter plots show no direct correlation between each variable. Moreover, it is best for everyone who wants to move to a certain neighborhood to directly visit the neighborhood to directly experience the living condition there.