Battle of Neighborhoods Capstone Project

Understanding Demographics, Crime and Venues for Neighborhood Segmentation: A k-means approach

Submitted by: Maigha

Introduction

The idea behind this project is to explore the neighborhoods in Toronto for their liveability. A neighborhood may be categorized based on the venues, crime rate, population demographics like total count, immigrations, healthy food index, and other parameters like number of rented vs owned dwellings, average income or average rent. The stakeholders include anyone interested in learning about the neighborhoods not only for liveability but also to understand the business potential of the town. In this project the focus is on understanding the potential of choosing an area of residence.

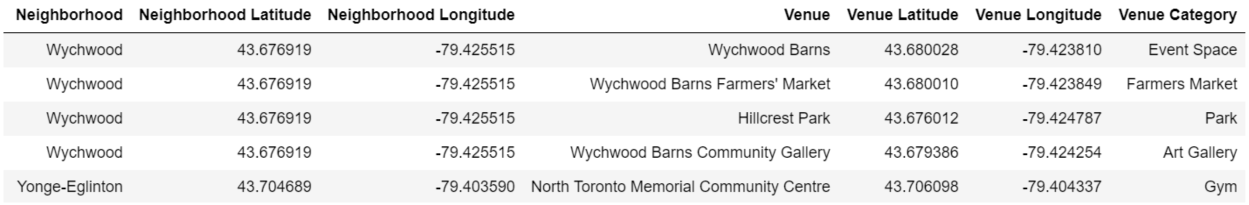
Business Problem

Can we determine the attractiveness of a neighborhoods (top 10) based on the venues, lower crime rates, number of rented dwellings, average rent, etc?

Data Sources and Preparation

The idea behind finding an attractive location for living is based on many factors depending on customer choice. One major requirement is the availability of venues nearby. Fourquare[1] has been used to explore this portion of the project.

Foursquare data



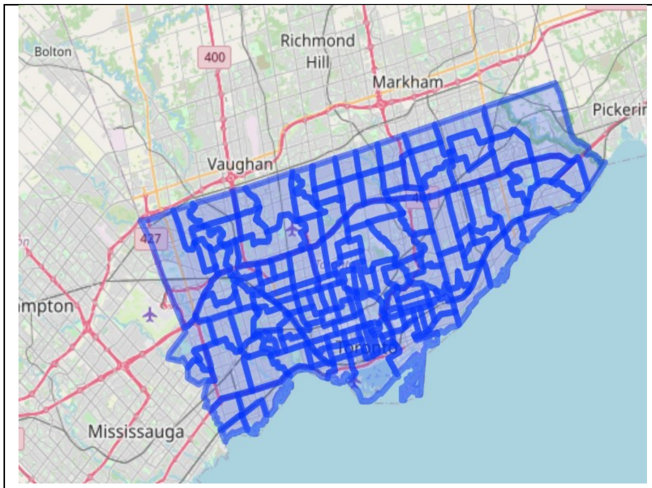
Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Wychwood	43.676919	-79.425515	Wychwood Barns	43.680028	-79.423810	Event Space
Wychwood	43.676919	-79.425515	Wychwood Barns Farmers' Market	43.680010	-79.423849	Farmers Market
Wychwood	43.676919	-79.425515	Hillcrest Park	43.676012	-79.424787	Park
Wychwood	43.676919	-79.425515	Wychwood Barns Community Gallery	43.679386	-79.424254	Art Gallery
Yonge-Eglinton	43.704689	-79.403590	North Toronto Memorial Community Centre	43.706098	-79.404337	Gym

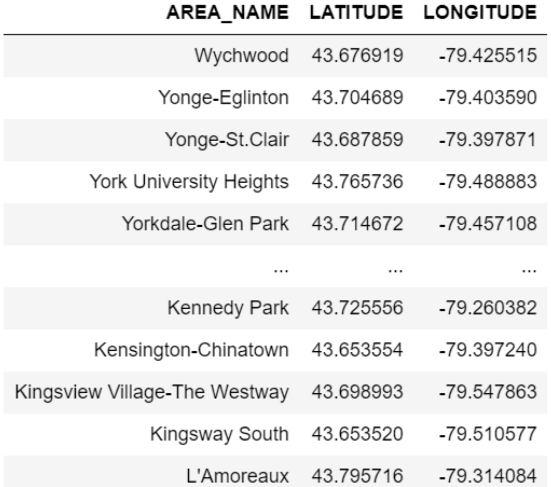
The neighborhoods were categorized based on the number of top venues in the neighborhoods.

Instead of the boroughs data used in the assignments, this project uses the detailed boundaries of the neighborhoods in Toronto. A total of 140 neighborhoods have been explored[2]. This helps us explore the neighborhoods individually.

Toronto neighborhoods[2]	Processed location data

[1] https://foursquare.com/
[2] https://open.toronto.ca/dataset/neighbourhoods/

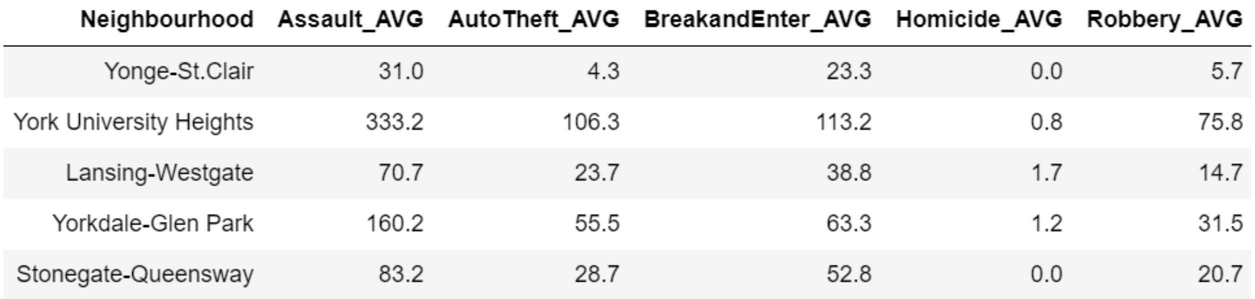




AREA_NAME	LATITUDE	LONGITUDE
Wychwood	43.676919	-79.425515
Yonge-Eglinton	43.704689	-79.403590
Yonge-St.Clair	43.687859	-79.397871
York University Heights	43.765736	-79.488883
Yorkdale-Glen Park	43.714672	-79.457108
...	...	...
Kennedy Park	43.725556	-79.260382
Kensington-Chinatown	43.653554	-79.397240
Kingsview Village-The Westway	43.698993	-79.547863
Kingsway South	43.653520	-79.510577
L'Amoreaux	43.795716	-79.314084

Crime data[3] was obtained from openly available data sources. This included the numbers for different categories like theft, abuse, etc. for the last 5 years. The analysis in the study uses the average for each of these categories for the last 5 years. The data had to be processed for determining any missing data but was generally found to be very consistent.

Crime data



Neighbourhood	Assault_AVG	AutoTheft_AVG	BreakandEnter_AVG	Homicide_AVG	Robbery_AVG
Yonge-St.Clair	31.0	4.3	23.3	0.0	5.7
York University Heights	333.2	106.3	113.2	0.8	75.8
Lansing-Westgate	70.7	23.7	38.8	1.7	14.7
Yorkdale-Glen Park	160.2	55.5	63.3	1.2	31.5
Stonegate-Queensway	83.2	28.7	52.8	0.0	20.7

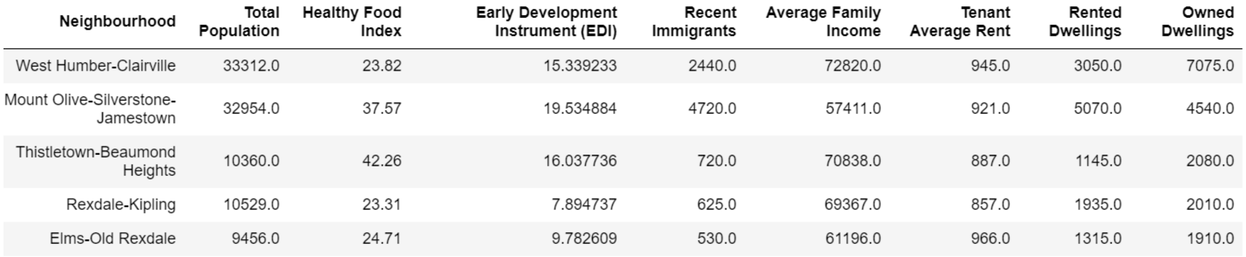
A great source of data was found at 'Wellbeing Toronto'[4]. While this is essentially a website that maps all the data, this data can also be downloaded for analysis. Data was downloaded for a number of categories including the population spread amongst different ethnicity. The ones that were used in

[3] https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/

[4] http://map.toronto.ca/wellbeing/#eyJ0b3Itd2lkZ2V0LWNsYXNzYnJlYWsiOsSAcGVyY2VudE9wYWNpdHnElzcwfSwiY3VzxIJtYcSTYcSXxIBuZWlnaGJvdXJob29kc8S2fcSrxIHEg8SFxIfEicSLdGFixYXEmCLEo3RpdmVUxZBJZMSXxYnEhMWPYi1pbmRpY2HEgnLFhcWIxaTFpsWoxarFksSAxZjFq2lvbsSXMsSsc8WkZ2xlxLbErcS%2FxJPEn1RpbWXFnMSoxKzFlsaIxbIiN8aBxa7Fp8WpxIPFnHNBxaVXxLnEu3TFklvEgMSHxZ43MyLErHfGnGh0xJcxxKzEk8W0c2VQb8SOcsSlxKc6ZmFsxrHEq8ahxZ06IjE2xqYixqjEusaqxqzGrmXGsMayxrTEs8a3xJfGusa8Zca%2BIsaix4EzxqXGp8apxqs6xq0ixq%2FEm8axxrPGtceRxrnGu8a9LMa%2FxZ4zMseFx4fGnceKx6HHjMejx47HpsSmx5LHqceVx6vHl8eAIjM0x7DHncezx6LFq8e3x5DHuceox5THlseYyIA1yIPHiMeex6DIhsekx4%2FGtsiKx5PHqsesx5nHhMecyJLIhce1yIfHpciJxrjImse8yJzIgMaRyJ%2FHssefx4vHjcikyJjIpse7yI3HvzM4yJHIrciUyKLIlse4yLPIjMe9yI4zOci5x4nIrse0yLDIl8enyKfItcWeMcmFyJPIr8e2yLHJi8i0yYHHvzHJhMisyYbIu8mJyL7HusiMXcWHxYjGjWXGsca2yabFhsSsxK5yxoR0ScWlxpTFqk3Fg8aAx4HFvG7FvsaAxYhhZ3NNYXDGgXrFgm3GrDPErHjEly04ODM3NzYzLjXKkTcyN8Ssxrg1NDEyOTMxLjI0ypAyODXFhw%3D%3D

analysis are listed below. It must be noted here that given a customer is interested n knowing about the concentration of an ethnicity in an area, that can be easily determined using the data.

Wellbeing Data



Neighbourhood	Total Population	Healthy Food Index	Early Development Instrument (EDI)	Recent Immigrants	Average Family Income	Tenant Average Rent	Rented Dwellings	Owned Dwellings
West Humber-Clairville	33312.0	23.82	15.339233	2440.0	72820.0	945.0	3050.0	7075.0
Mount Olive-Silverstone-Jamestown	32954.0	37.57	19.534884	4720.0	57411.0	921.0	5070.0	4540.0
Thistletown-Beaumond Heights	10360.0	42.26	16.037736	720.0	70838.0	887.0	1145.0	2080.0
Rexdale-Kipling	10529.0	23.31	7.894737	625.0	69367.0	857.0	1935.0	2010.0
Elms-Old Rexdale	9456.0	24.71	9.782609	530.0	61196.0	966.0	1315.0	1910.0

Key Features for analysis

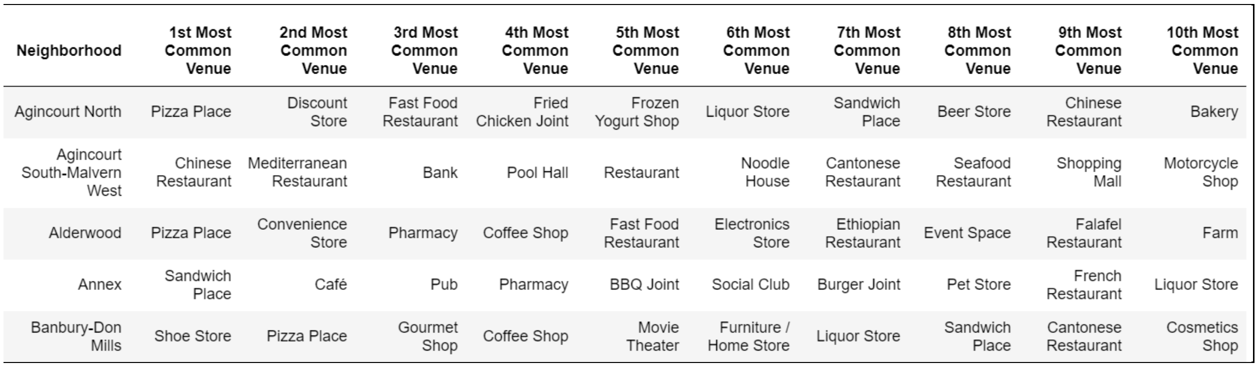
The key features used in the analysis individually or in the combined form are listed below. The color coding is the differentiate them based on the data sources.

Venues	Total population	Healthy food index
Early development instrument	Recent immigrants	Average family income
Tenant average rent	Rented Dwellings	Owned Dwellings
Assault	Auto Theft	Breaking and Entering
Homicide	Robbery	

Exploratory Data Analysis

Analysis based on venues

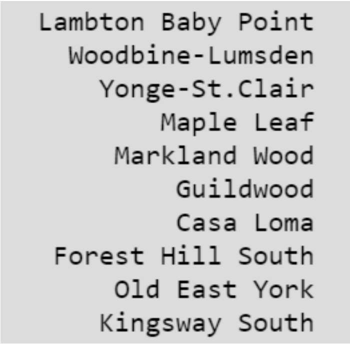
This part of the analysis follows the assignment and re-uses part of the code from the assignments. Fourquare data was used to determine the venues for each neighborhood. The neighborhoods were then sorted based on the top 10 venues in each. A snapshot of the result is below:



Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Agincourt North	Pizza Place	Discount Store	Fast Food Restaurant	Fried Chicken Joint	Frozen Yogurt Shop	Liquor Store	Sandwich Place	Beer Store	Chinese Restaurant	Bakery
Agincourt South-Malvern West	Chinese Restaurant	Mediterranean Restaurant	Bank	Pool Hall	Restaurant	Noodle House	Cantonese Restaurant	Seafood Restaurant	Shopping Mall	Motorcycle Shop
Alderwood	Pizza Place	Convenience Store	Pharmacy	Coffee Shop	Fast Food Restaurant	Electronics Store	Ethiopian Restaurant	Event Space	Falafel Restaurant	Farm
Annex	Sandwich Place	Café	Pub	Pharmacy	BBQ Joint	Social Club	Burger Joint	Pet Store	French Restaurant	Liquor Store
Banbury-Don Mills	Shoe Store	Pizza Place	Gourmet Shop	Coffee Shop	Movie Theater	Furniture / Home Store	Liquor Store	Sandwich Place	Cantonese Restaurant	Cosmetics Shop

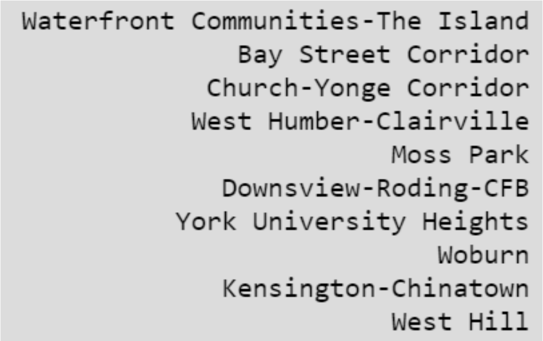
Neighborhoods with lowest average crime rates

The neighborhoods were sorted in the ascending order for crime. The top 10 in the list are given below:



```
Lambton Baby Point
  Woodbine-Lumsden
    Yonge-St.Clair
        Maple Leaf
     Markland Wood
         Guildwood
         Casa Loma
 Forest Hill South
     Old East York
    Kingsway South
```

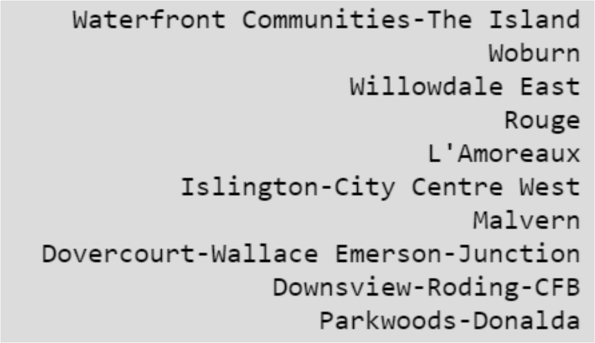
Looking at the top 10 crime focus centers:



```
Waterfront Communities-The Island
              Bay Street Corridor
            Church-Yonge Corridor
           West Humber-Clairville
                        Moss Park
             Downsview-Roding-CFB
          York University Heights
                           Woburn
             Kensington-Chinatown
                        West Hill
```

Most populated areas

Next, we wish to understand the population density of the neighborhoods. The top 10 most dense neighborhoods are:

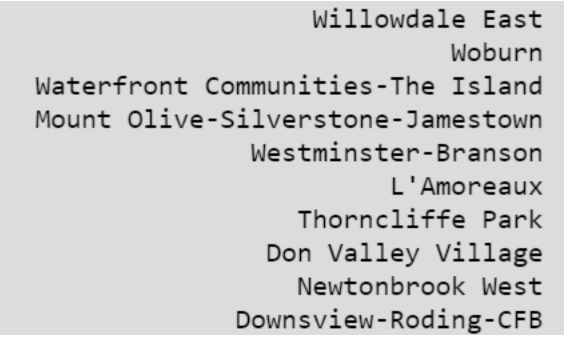


```
  Waterfront Communities-The Island
                             Woburn
                    Willowdale East
                              Rouge
                         L'Amoreaux
         Islington-City Centre West
                            Malvern
Dovercourt-Wallace Emerson-Junction
               Downsview-Roding-CFB
                  Parkwoods-Donalda
```

It is interesting to note here that the dense neighborhood was also on the top crime list as well.

Recent immigration areas

The top 10 neighborhoods with immigrants were found.

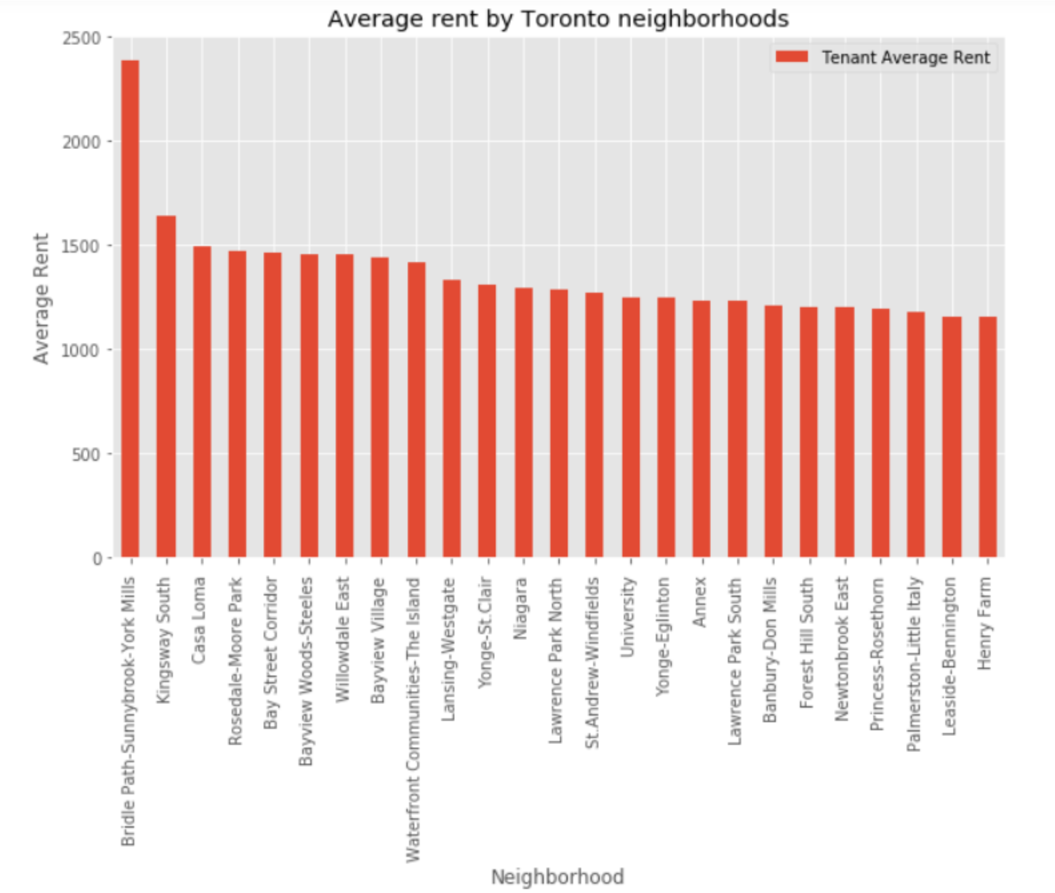


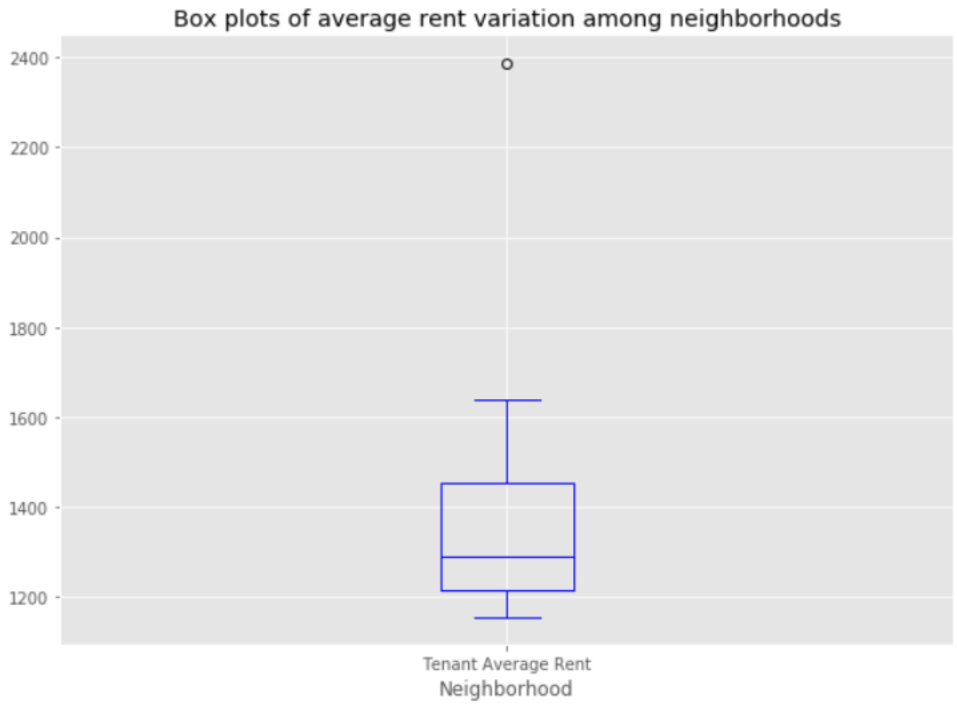
```
                       Willowdale East
                                Woburn
    Waterfront Communities-The Island
    Mount Olive-Silverstone-Jamestown
                    Westminster-Branson
                              L'Amoreaux
                        Thorncliffe Park
                      Don Valley Village
                        Newtonbrook West
                     Downsview-Roding-CFB
```

An interesting observation is that five of the most populated areas were also the ones with large immigrant population.

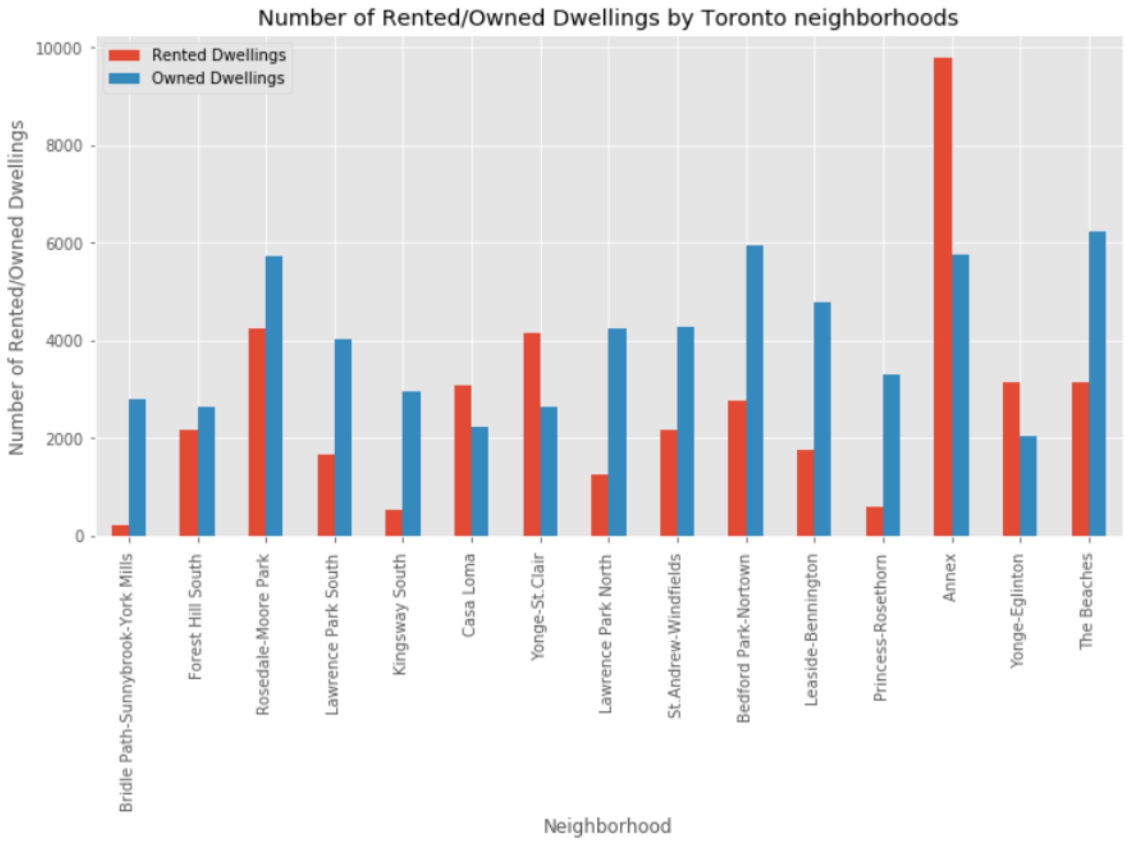
Average tenant rents

Next, we wanted to explore the average tenant rents in the neighborhood. A bar chart of the top 25 neighborhoods based on their average tenant rent have been shown below. *The box plot that follows corroborates the fact that except for the first neighborhood in the chart, the rents of the other neighborhoods are distributed in a narrow range.* Bridle Path seems to be the most expensive for rentals, obviously nit the most populated.





Interestingly, a bar chart representing the number of rented and owned dwellings sheds some more light on this observation.



It is clear from the above graph that the neighborhood with the highest rental average has the fewest rented units. It seems that mostly owned dwellings exist in the area.

The above data analysis showed a great potential in understanding the underlying dynamics of the neighborhoods that may make them attractive residing spots based on customer criteria. The next part of the project is associated with clustering the neighborhoods based on these parameters.

Neighborhood Clustering

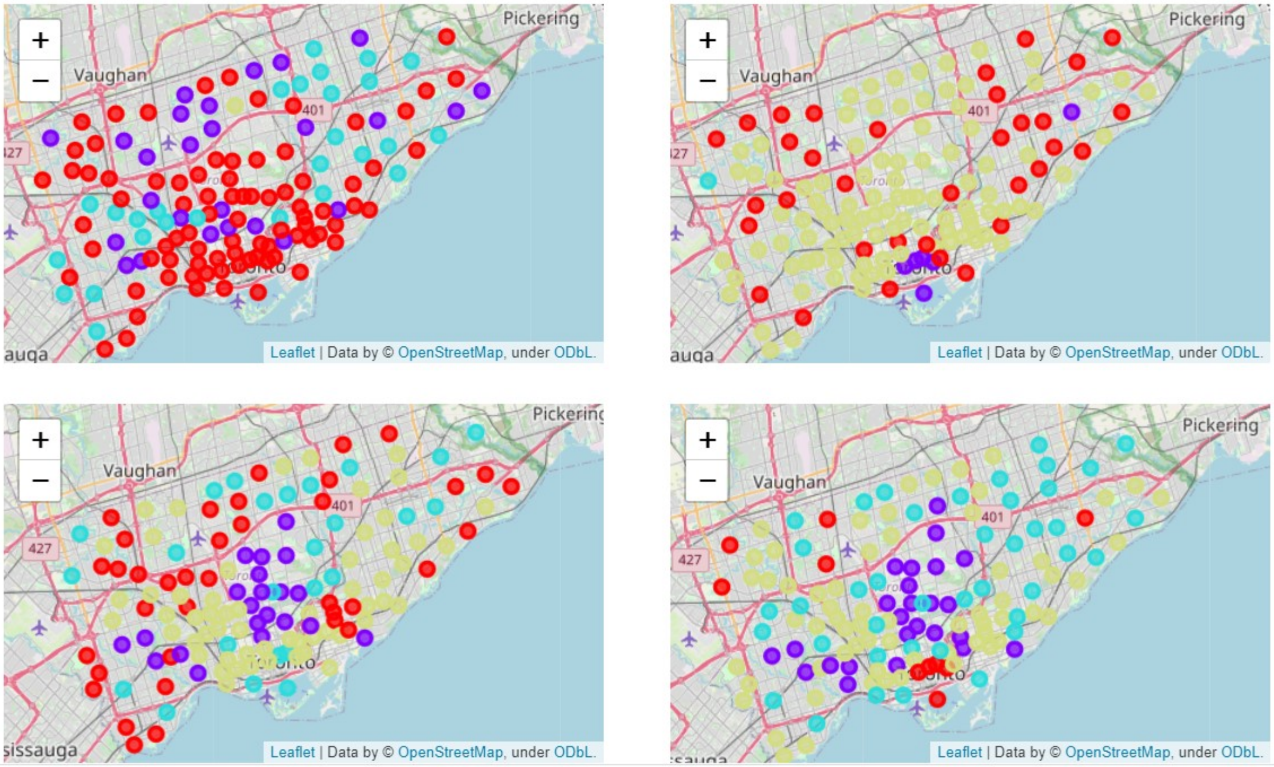
The analysis was performed in four steps:

1. The neighborhoods were clustered based on venues only
2. Clustering based on crime data – each feature taken individually
3. Clustering based on the features extracted from wellbeing data
4. Clustering based on combining the above features

Each time the analysis was performed, data was normalized carefully. This is very important because the features vary on a large scale. While one hot encoding was used for venues data, standard scaler was used to normalize the rest of the features.

Results and Discussion

The clustering of the neighborhoods based on the above criteria provide a first look into how the areas may be related to each other. The results shown below need to read row wise from left to right for the cases 1-4 listed above.



The top left is the clustering based on venues, top right is based on crime, bottom left is based on wellbeing data and bottom right combines crime and wellbeing data. Due to the large number of features in wellbeing data and probably their higher influence, the final combines clustering looks very similar to the one obtained for wellbeing data.

Conclusion

Using the clustering and initial data exploration, it may be possible to categorize and shortlist neighborhoods based on personalized criteria. The objective of this analysis was to show the possibility and some initial visualization. It is evident that a lot needs to be explored in depth for driving decisions based on this.

Future directions

Based on the above analysis, fewer features will be chosen for in-depth analysis to determine the underlying similarities between neighborhoods, their key features and characteristics. This study has probably just scratched the surface within a tiny scope.