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.

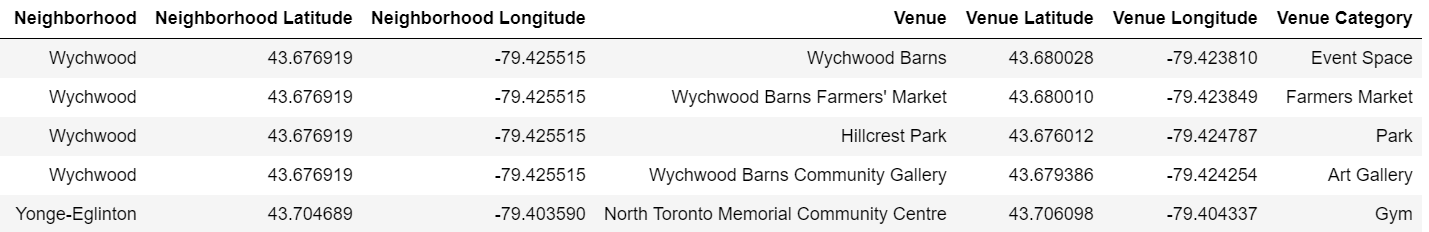
## Business Problem

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]](#footnote-1) has been used to explore this portion of the project.

Foursquare data



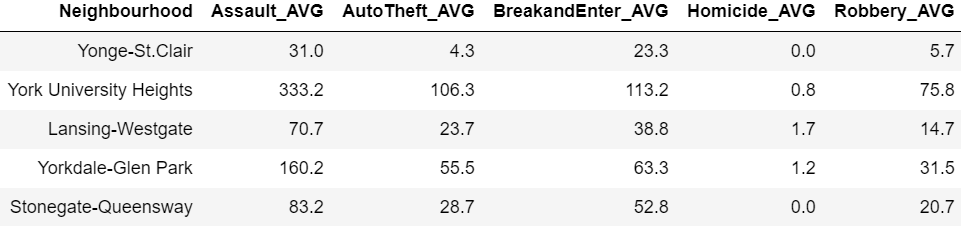
The neighborhoods were categorized based on the number of top venues in the neighborhoods.

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]](#footnote-2). This helps us explore the neighborhoods individually.

|  |  |
| --- | --- |
| Toronto neighborhoods2 | Processed location data |
|  |  |

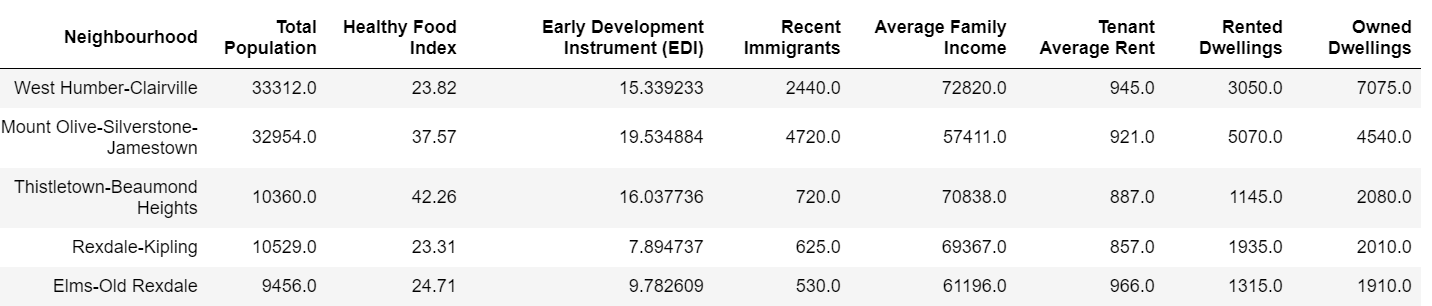
Crime data[[3]](#footnote-3) was obtained from openly available data sources. This included the numbers for different categories like theft, abuse, etc. for the last 5 years. The analysis in the study uses the average for each of these categories for the last 5 years. The data had to be processed for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

Crime data



A great source of data was found at ‘Wellbeing Toronto’[[4]](#footnote-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 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



## 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

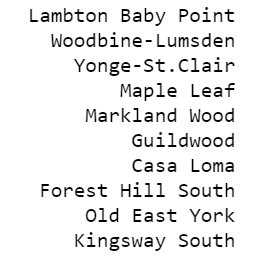
## Analysis based on venues

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:

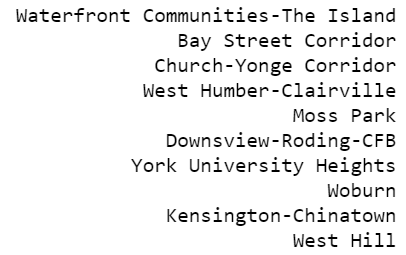


## Neighborhoods with lowest average crime rates

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

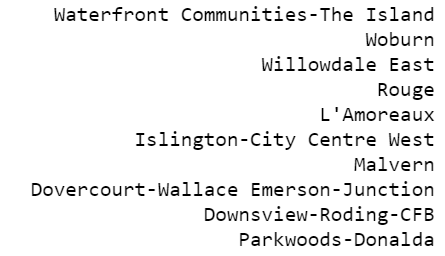


Looking at the top 10 crime focus centers:



## Most populated areas

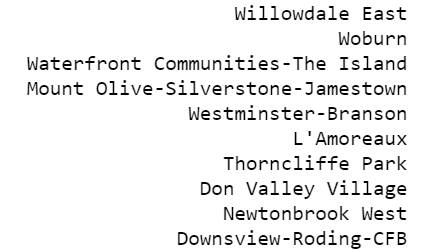
Next, we wish to understand the population density of the neighborhoods. The top 10 most dense neighborhoods are:



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

## Recent immigration areas

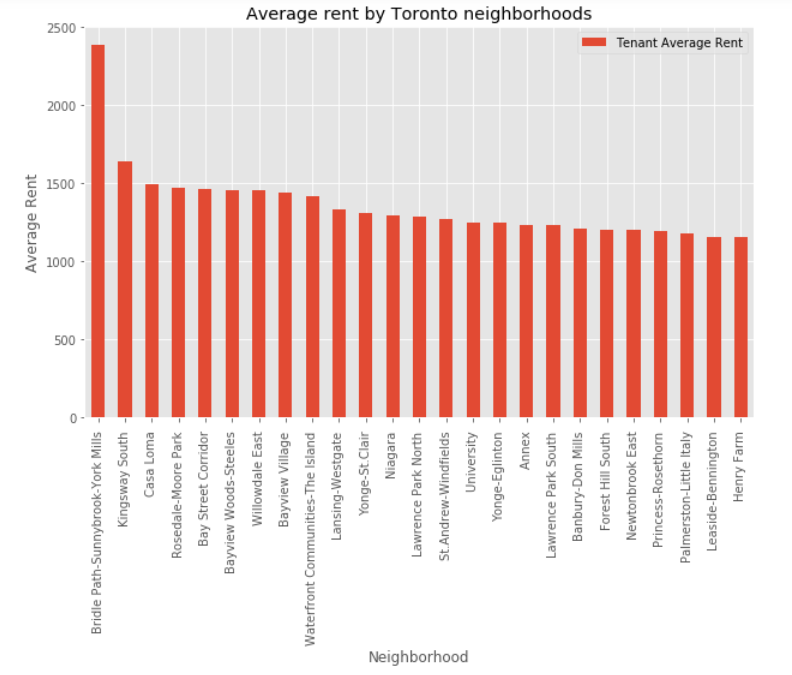
The top 10 neighborhoods with immigrants were found.

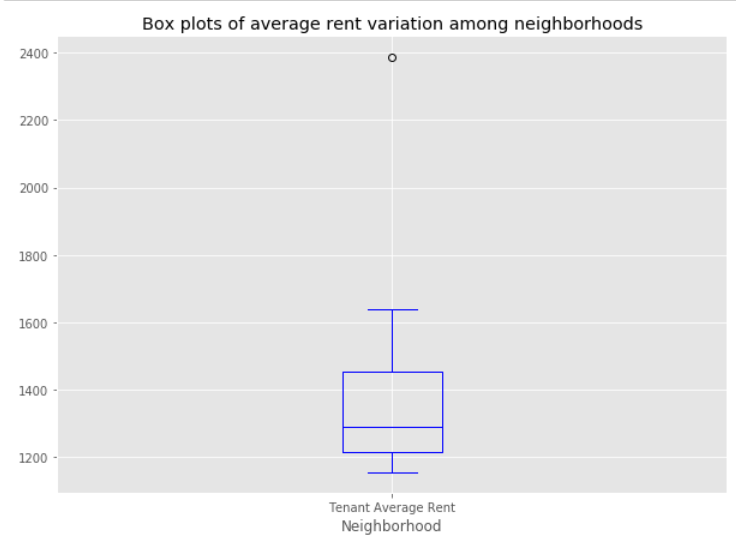


#### An interesting observation is that five of the most populated areas were also the ones with large immigrant population.

## 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

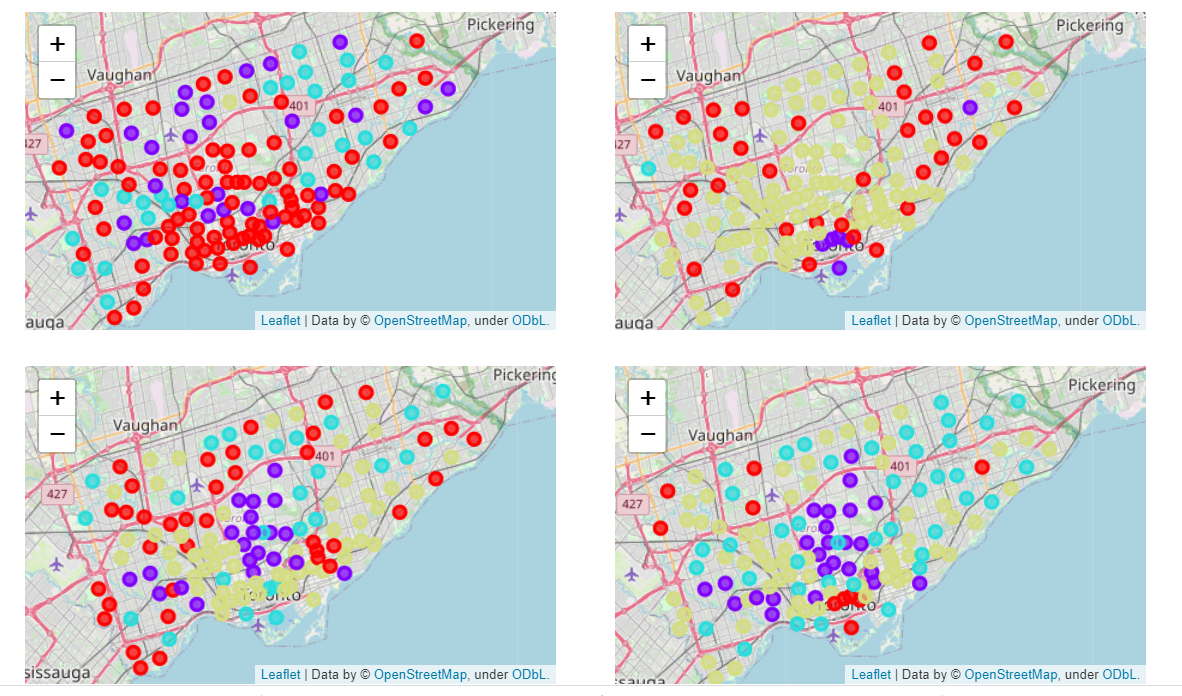
The analysis was performed in four steps:

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.

1. [https://foursquare.com/](about:blank) [↑](#footnote-ref-1)
2. [https://open.toronto.ca/dataset/neighbourhoods/](about:blank) [↑](#footnote-ref-2)
3. [https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/](about:blank) [↑](#footnote-ref-3)
4. [http://map.toronto.ca/wellbeing/#eyJ0b3Itd2lkZ2V0LWNsYXNzYnJlYWsiOsSAcGVyY2VudE9wYWNpdHnElzcwfSwiY3VzxIJtYcSTYcSXxIBuZWlnaGJvdXJob29kc8S2fcSrxIHEg8SFxIfEicSLdGFixYXEmCLEo3RpdmVUxZBJZMSXxYnEhMWPYi1pbmRpY2HEgnLFhcWIxaTFpsWoxarFksSAxZjFq2lvbsSXMsSsc8WkZ2xlxLbErcS%2FxJPEn1RpbWXFnMSoxKzFlsaIxbIiN8aBxa7Fp8WpxIPFnHNBxaVXxLnEu3TFklvEgMSHxZ43MyLErHfGnGh0xJcxxKzEk8W0c2VQb8SOcsSlxKc6ZmFsxrHEq8ahxZ06IjE2xqYixqjEusaqxqzGrmXGsMayxrTEs8a3xJfGusa8Zca%2BIsaix4EzxqXGp8apxqs6xq0ixq%2FEm8axxrPGtceRxrnGu8a9LMa%2FxZ4zMseFx4fGnceKx6HHjMejx47HpsSmx5LHqceVx6vHl8eAIjM0x7DHncezx6LFq8e3x5DHuceox5THlseYyIA1yIPHiMeex6DIhsekx4%2FGtsiKx5PHqsesx5nHhMecyJLIhce1yIfHpciJxrjImse8yJzIgMaRyJ%2FHssefx4vHjcikyJjIpse7yI3HvzM4yJHIrciUyKLIlse4yLPIjMe9yI4zOci5x4nIrse0yLDIl8enyKfItcWeMcmFyJPIr8e2yLHJi8i0yYHHvzHJhMisyYbIu8mJyL7HusiMXcWHxYjGjWXGsca2yabFhsSsxK5yxoR0ScWlxpTFqk3Fg8aAx4HFvG7FvsaAxYhhZ3NNYXDGgXrFgm3GrDPErHjEly04ODM3NzYzLjXKkTcyN8Ssxrg1NDEyOTMxLjI0ypAyODXFhw%3D%3D](about:blank#eyJ0b3Itd2lkZ2V0LWNsYXNzYnJlYWsiOsSAcGVyY2VudE9wYWNpdHnElzcwfSwiY3VzxIJtYcSTYcSXxIBuZWlnaGJvdXJob29kc8S2fcSrxIHEg8SFxIfEicSLdGFixYXEmCLEo3RpdmVUxZBJZMSXxYnEhMWPYi1pbmRpY2HEgnLFhcWIxaTFpsWoxarFksSAxZjFq2lvbsSXMsSsc8WkZ2xlxLbErcS%2FxJPEn1RpbWXFnMSoxKzFlsaIxbIiN8aBxa7Fp8WpxIPFnHNBxaVXxLnEu3TFklvEgMSHxZ43MyLErHfGnGh0xJcxxKzEk8W0c2VQb8SOcsSlxKc6ZmFsxrHEq8ahxZ06IjE2xqYixqjEusaqxqzGrmXGsMayxrTEs8a3xJfGusa8Zca%2BIsaix4EzxqXGp8apxqs6xq0ixq%2FEm8axxrPGtceRxrnGu8a9LMa%2FxZ4zMseFx4fGnceKx6HHjMejx47HpsSmx5LHqceVx6vHl8eAIjM0x7DHncezx6LFq8e3x5DHuceox5THlseYyIA1yIPHiMeex6DIhsekx4%2FGtsiKx5PHqsesx5nHhMecyJLIhce1yIfHpciJxrjImse8yJzIgMaRyJ%2FHssefx4vHjcikyJjIpse7yI3HvzM4yJHIrciUyKLIlse4yLPIjMe9yI4zOci5x4nIrse0yLDIl8enyKfItcWeMcmFyJPIr8e2yLHJi8i0yYHHvzHJhMisyYbIu8mJyL7HusiMXcWHxYjGjWXGsca2yabFhsSsxK5yxoR0ScWlxpTFqk3Fg8aAx4HFvG7FvsaAxYhhZ3NNYXDGgXrFgm3GrDPErHjEly04ODM3NzYzLjXKkTcyN8Ssxrg1NDEyOTMxLjI0ypAyODXFhw%3D%3D) [↑](#footnote-ref-4)