**Omar Houssein**

Choosing the best neighborhoods to open a new restaurant in the city of Toronto.

**The Battle of Neighborhoods (Toronto)**

*2019/09/20*

Table of Contents

[1. Introduction 2](#_Toc19970801)

[1.1. Background 2](#_Toc19970802)

[1.2. Business Issue 2](#_Toc19970803)

[1.3. Interest 2](#_Toc19970804)

[2. Data acquisition and cleaning 2](#_Toc19970805)

[2.1. Data sources 2](#_Toc19970806)

[2.2. Data cleaning 3](#_Toc19970807)

[2.3. Feature selection 4](#_Toc19970808)

[3. Exploratory data analysis 4](#_Toc19970809)

[3.1. Home prices 5](#_Toc19970811)

[3.2. Crimes rate 5](#_Toc19970812)

[3.3. Relationship between home prices and crimes rate 6](#_Toc19970813)

[4. Methodology 7](#_Toc19970814)

[K-Means Clustering: 7](#_Toc19970819)

[5. Results 7](#_Toc19970825)

[5.1. K-Means clustering 7](#_Toc19970826)

[5.2. Home prices 8](#_Toc19970827)

[5.3. Crimes rate 9](#_Toc19970835)

[6. Discussion 10](#_Toc19970836)

[7. Conclusion 13](#_Toc19970837)

[APPENDICES 14](#_Toc19970838)

## Introduction

### Background

Toronto city, the capital of Ontario province, is the biggest city in Canada, and is unique by its cosmopolitan and multicultural nature.

Toronto city is an important center for shops, restaurants, hotels, coffee shops, etc., and the profitability of these businesses is highly dependent on where they are located.

Each year we can see many new business starting their activities, however, we also notice the closing of many others at the same time.

It’s well known that to be successful, a retail business, has to be located in an advantageous area.

A business can be the best in its field, in terms of quality products, competitive prices and providing high value service. But if its location is “bad” or not enough evaluated before the settling, it is very common that it won’t have the expected success.

Therefore, it is highly recommended to tackle the location problem for a new business, in order to secure sufficient walk-in traffic and revenues and have the best chances of success.

### Business Issue

This study will focus on finding out what location to choose when opening a new restaurant in the city of Toronto. More specifically, the study aims to identify which neighborhood(s) are best for opening a new restaurant, taking into account its type of cuisine and knowing that neighborhoods known for that type of cuisine will be more attractive. Apart from the aim to benefit from the popularity of the neighborhoods, other factors like real estate prices and crimes rate will also count in our search of the best neighborhoods.

### Interest

This study will be interesting for entrepreneurs aiming to open a restaurant in the city of Toronto.

It can also especially be interesting for immigrant’s entrepreneurs coming to Toronto, who have absolutely no idea about what characterizes each neighborhood, in a city as big as Toronto.

## Data acquisition and cleaning

### Data sources

|  |  |  |  |
| --- | --- | --- | --- |
| **Topic** | **Data** | **Source** | **Format/ Method** |
| Neighborhoods | Name | <https://open.toronto.ca/> | CSV file |
| Latitude |
| Longitude |
| Crimes rate | [http://data.torontopolice.on.ca](http://data.torontopolice.on.ca/) | CSV file |
| Home prices | <https://open.toronto.ca/> | Excel file |
| Restaurants | Name | Foursquare | API call |
| Latitude |
| Longitude |
| Category |
| ID |
| Rating |

### Data cleaning

As mentioned in the previous section, data pertaining to neighborhoods were obtained from a CSV file. This is how I got a dataframe containing the name of the neighborhoods and their coordinates.

Afterwards, I used the Foursquare API to explore the venues of each neighborhood. I established the radius of the search to 500 meters. This means that, for each neighborhood, we will get venues that are within a circle of diameter 1000m, assuming that the center of this circle is the venue in question. I also limited the results to 100 venues per neighborhood.

This operation allowed me to get the latitude, longitude and category of each venue.

I then restrained the venues only to restaurants, by filtering the venue category, and grouped each restaurant by cuisine type to the following 10 categories (cuisine type):

1. African Restaurant
2. American Restaurant
3. South American Restaurant
4. Asian Restaurant
5. European Restaurant
6. Middle Eastern Restaurant
7. Italian Restaurant
8. Japanese Restaurant
9. Indian Restaurant
10. Chinese Restaurant

At the beginning I wanted to group the cuisine type by continent. But when I found a considerable number of Indian, Chinese, Japanese and Italian restaurants, I decided to assign them to separate catehgories, as these types of cuisine were popular.

The following categories were dropped from the dataset because they were too general and could be assigned to many types of cuisine or countries:

* Fast Food Restaurant
* Restaurant
* Vegetarian/ Vegan Restaurant
* Seafood Restaurant
* Halal Restaurant
* Comfort Food Restaurant
* Mediterranean Restaurant
* Indian Chinese Restaurant

In total, 114 restaurants were dropped from an initial dataset of 504 recordings.

I then got the ratings of these restaurants, also from Foursquare. I considered that an acceptable rating should be equal or above 7.

After keeping the restaurants with the required rating, the dataset was reduced to 232 recordings.

I was able then to put in a single table each restaurant with its coordinates, neighborhood, category and rating.

**Home prices:**

The dataset related to home prices, contains data for the year 2012. From this dataset, I just selected the name of the neighborhood and the average home price, and put them in a new dataframe.

**Crimes rate:**

The CSV with crimes data encompassed data from year 2014 to 2018.

As the dataset is very large (167.525 recordings), I decided to work with data only pertaining to the most recent year, ie 2018. This new restrained dataset contained 35.382 crime records.

From this dataset, I counted the total number of crimes grouped by neighborhood.

After grouping by neighborhood, I found a total of 141 neighborhoods whereas I should have found 140.

After analysis of the dataset, I found that one of the neighborhoods was mistakenly written in 2 ways. After correcting its name, I obtained the correct number of 140 neighborhoods.

Home prices and crimes rate were then merged in a second dataset.

### Feature selection

After the data cleaning process, I got 2 datasets:

The first one had 232 samples and 9 features:

* Neighborhood name
* Neighborhood latitude
* Neighborhood longitude
* Venue name
* Venue latitude
* Venue longitude
* Venue ID
* Venue category
* Venue rating

The second one had 140 samples (relative to the number of neighborhoods) and 3 features:

* Neighborhood name
* Home price
* Crimes rate

## Exploratory data analysis

Before processing the data, the dataset showed the existence of 55 restaurant’s categories.

After category modification, Figure 1 shows the number of restaurants per new category.

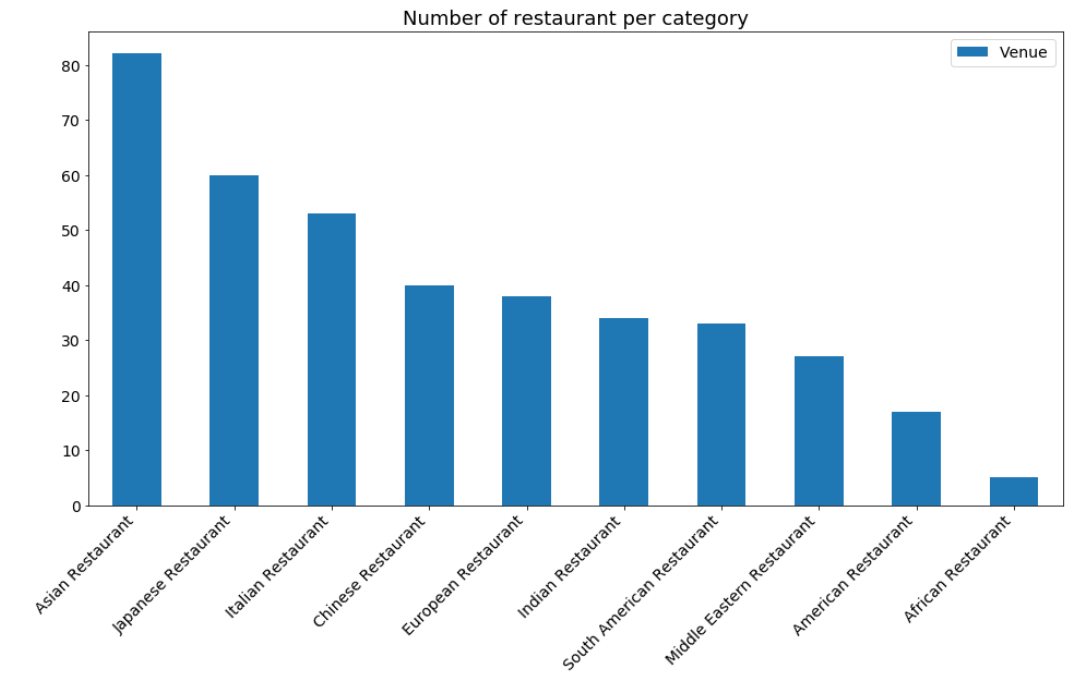


Figure 1. Bar plot of number of restaurants per category



### Home prices

Here is the box plot of neighborhoods’ home prices in Toronto in 2012:

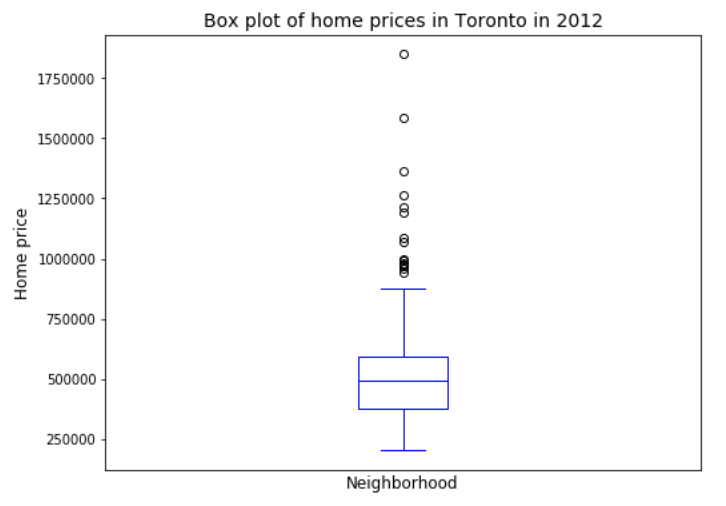


Figure 2. Box plot of Toronto neighborhoods’ home prices in 2012

As we can see, the median price of a house per neighborhood in Toronto in 2012, is about 500.000$.

### Crimes rate

Here is the box plot of neighborhoods’ crimes rate in Toronto in 2018:

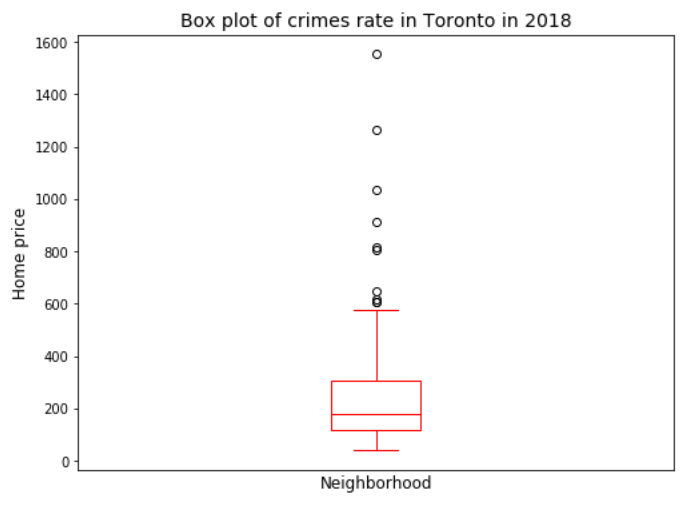


Figure 3. Box plot of Toronto neighborhoods’ crimes rate in 2018

As we can see, the median number of crimes per neighborhood in Toronto in 2018, is about 200.

### Relationship between home prices and crimes rate

One could assume that there is a relationship between home prices and crimes rate, in that when crimes rate increases, home prices decrease.

But when we analyze Figure 3 which shows the scatter plot of crimes rate and home prices, it doesn’t show any linear correlation between these two variables.

For instance, in some neighborhoods of Toronto city, the crimes rate is high, as well as home prices. Rosedale-Moore Park is an example of this affirmation.

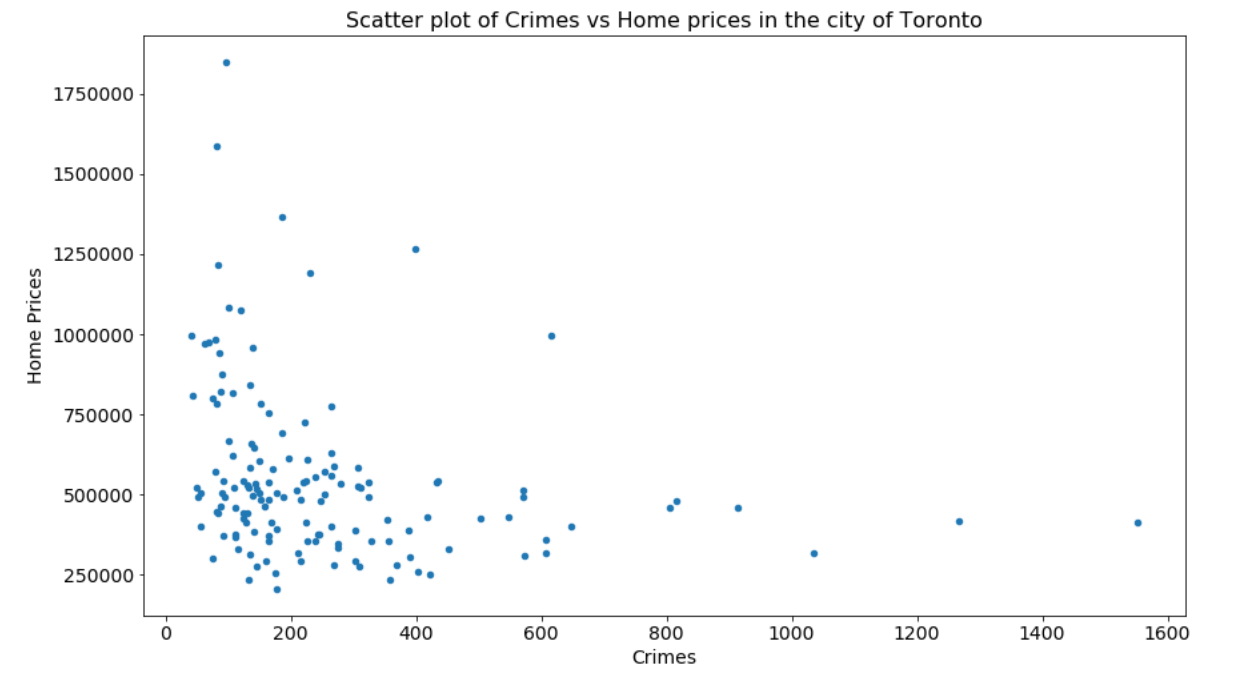


Figure 4. Scatter plot of neighborhoods’ crimes rate and home prices

## Methodology

In order to respond to our business problem, which is finding the best neighborhood(s) to open a new restaurant with a certain type of cuisine, it is obvious that we need to group all the restaurants in the city of Toronto according to their category (or type of cuisine), and then determine in which neighborhoods they are located.

To do so, I decided to use clustering machine learning algorithms.

First, I will use k-Means algorithm to a dataset that includes all selected types of restaurants located within Toronto’s neighborhoods, based on the number of popular restaurants per neighborhood. By popular, I mean restaurants that have a rating equal or greater than 7 in Foursquare.

Then, I will refine my search of the best neighborhood(s) by studying neighborhood’s homes prices as well as crimes rate, and confront the results to the findings of K-Means clustering.

To study neighborhood’s homes prices and crimes rate, I will use a data visualization technique which is choropleth map, in order to visually determine which neighborhoods are less expensive and have lower crime rates than others.

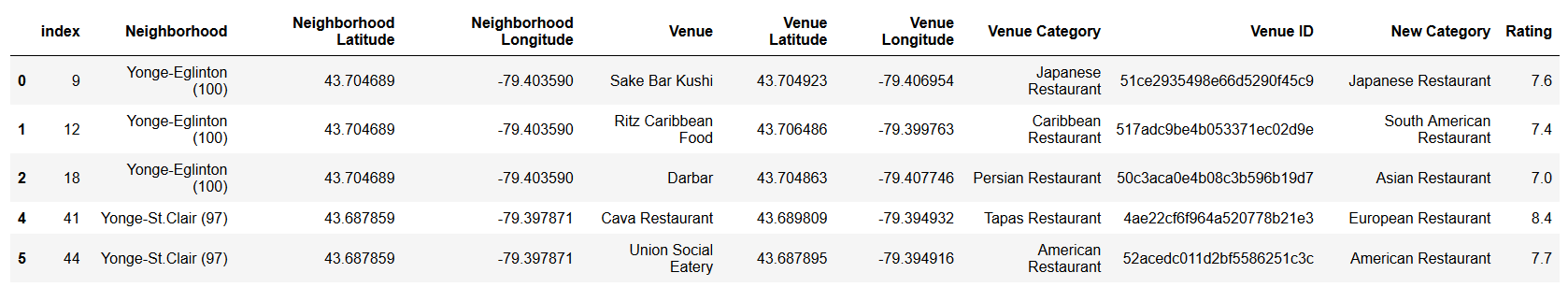


### K-Means Clustering:

K-Means is considered one of the simplest models amongst clustering algorithms, and is vastly used for clustering in many data science applications.

The methodology chosen is to segment Toronto city’s neighborhoods into clusters that have similar characteristics, in our case, based on the number of restaurants with the same category.

Here is a snapshot of the dataset our work will be based on:



I started my work by doing one hot encoding to the categories, in order to have the neighborhoods as rows, and the restaurant’s categories as columns. I then grouped all rows by neighborhood, such as I have the mean of occurrence of each restaurant category for each neighborhood.

The result showed that some neighborhoods didn’t have any restaurant matching the required categories. As using K-Means requires assigning each neighborhood to a cluster, I had to first drop these neighborhoods from the dataset. After this operation, the total number of neighborhoods became 54.

I then applied k-Means algorithm on the dataset. I tuned the number of cluster parameter to 10 clusters, which is equal to the number of chosen restaurant’s categories.



## Results

## K-Means clustering

K-Means clustering gave us 10 neighborhoods’ clusters based on the number of restaurants with the same category.

The details of these clusters can be seen in the appendices section.

The following figures shows clusters’ distribution on Toronto map:

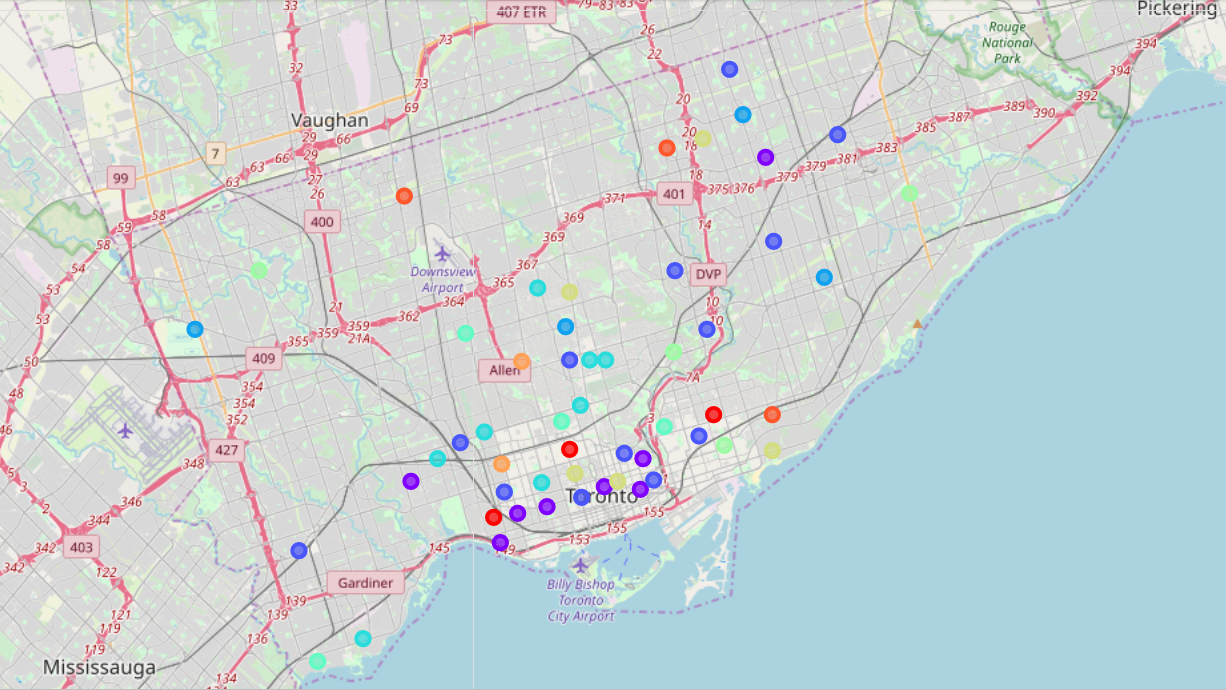


Figure 5. K-Means clusters distribution on Toronto map

After K-Means clusters analysis, I could assign almost each cluster with a specific type of cuisine as follow:

|  |  |
| --- | --- |
| Cluster | Cuisine type |
| 1 | American |
| 2 | Not specific |
| 3 | Asian |
| 4 | Chinese |
| 5 | Italian |
| 6 | European |
| 7 | Indian |
| 8 | Japanese |
| 9 | Middle Eastern |
| 10 | South American |

### Home prices

As mentioned in the feature selection section, the features of the second dataset are: neighborhood name, home prices and crimes rate.

From this dataset, I just used the name of the neighborhood and the average home price.

In order to get an idea about the neighborhoods in terms of housing prices, I used a choropleth map.

The choropleth map allows to easily distinguish between the expensive neighborhoods and the cheap ones, knowing that the darker colors relates to more expensive and lighter colors relates to less expensive.

Here is the choropleth map of neighborhoods’ home prices in Toronto in 2012:

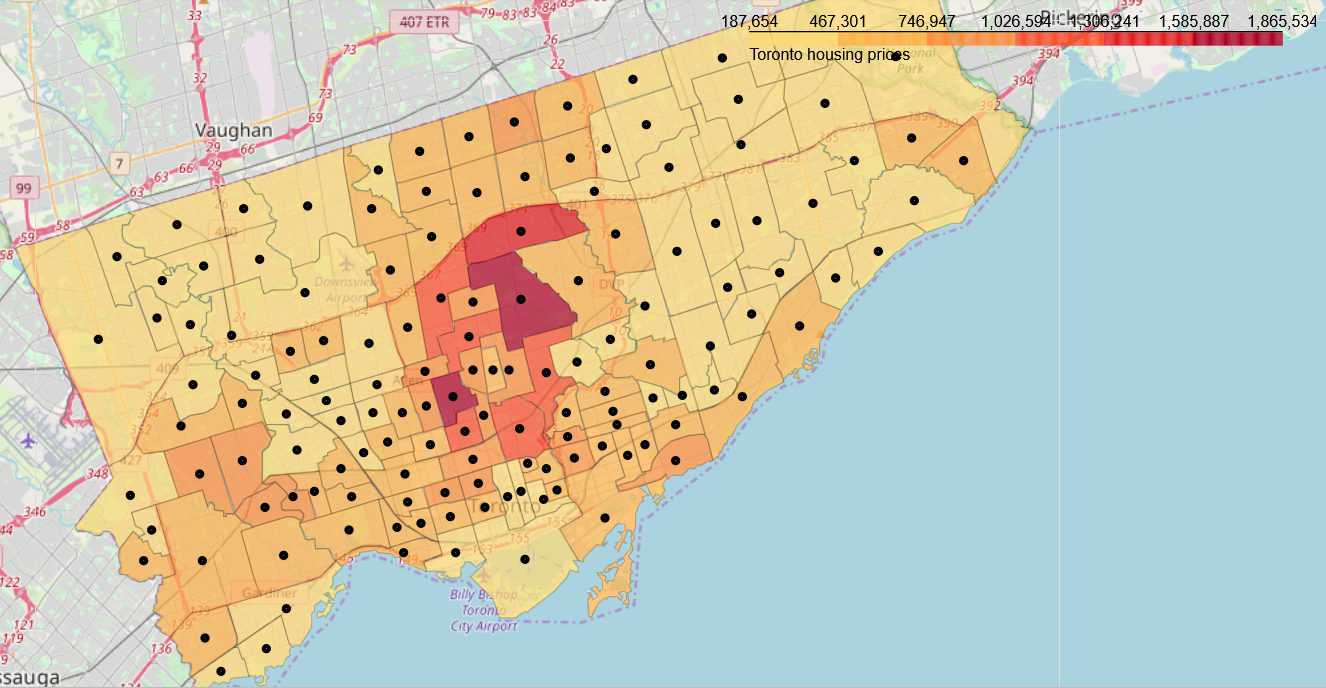


Figure 6. Choropleth map of neighborhoods’ home prices in Toronto in 2012



### Crimes rate

As mentioned in the feature selection section, the features of the second dataset are: neighborhood name, home prices and crimes rate.

From this dataset, I just used the name of the neighborhood and the crimes rate.

As for of home prices, we use a choropleth map in order to get an idea about the neighborhoods in terms of crime rate.

Here is the choropleth map of neighborhoods’ crime rate in Toronto in 2018:

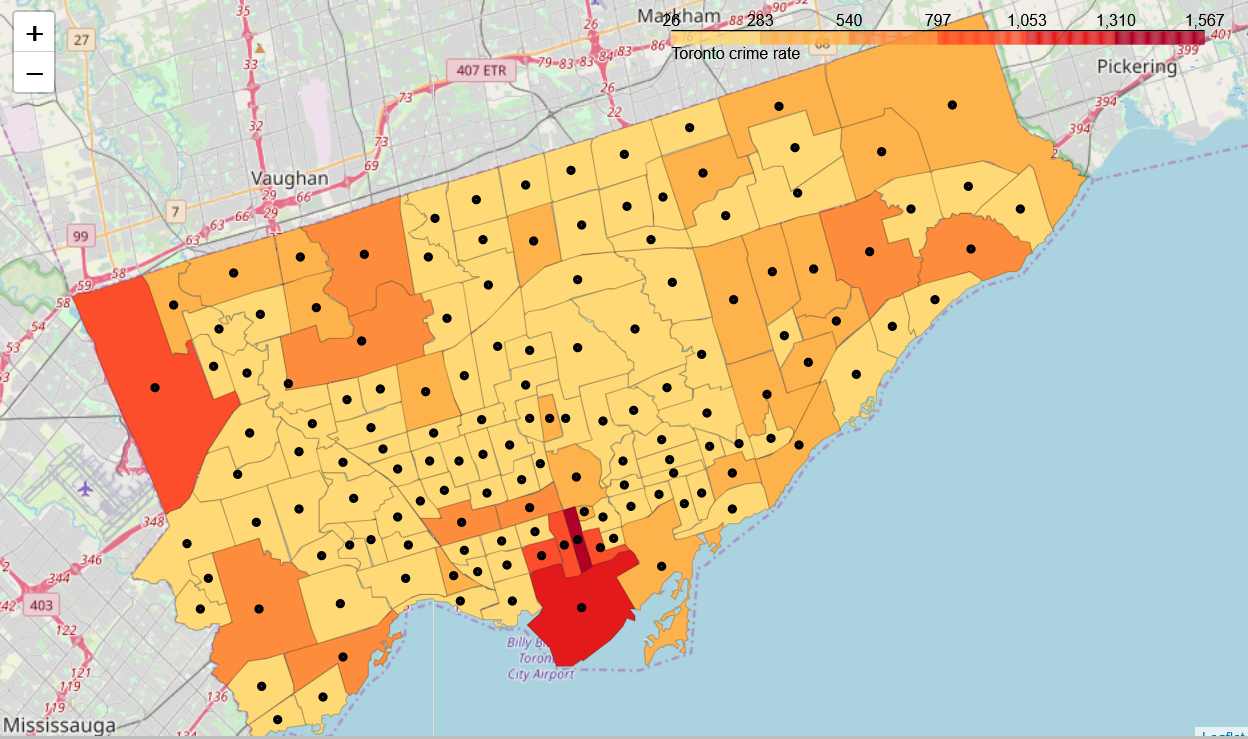


Figure 7. Choropleth map of neighborhoods’ crimes rate in Toronto in 2018

## Discussion

In order to answer our business question, I first took the findings of K-Means clustering in order to identify the neighborhoods related to each type of cuisine. Then I confronted these results to the findings related to housing prices crimes and rates in Toronto.

According to the choropleth maps obtained, I cloud classify home prices and crimes rate according to the color code as follow:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Low | Low to medium | Medium | Medium to high | High | Very High |

The selection process was based on choosing neighborhoods having the lowest home price and crimes rate. According to the type of cuisine of the future restaurant, here are the recommended neighborhoods:

**American cuisine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **1st most common venue** | **Home price** | **Crimes rate** | **Selection** |
| Annex (95) | Indian Restaurant | - | - | No |
| Danforth (66) | American Restaurant | Low to medium | Low | Yes |
| Roncesvalles (86) | American Restaurant | Low to medium | Low to medium | No |

**Asian cuisine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **1st most common venue** | **Home price** | **Crimes rate** | **Selection** |
| Agincourt South-Malvern West (128) | Asian Restaurant | Low | Low | Yes |
| Banbury-Don Mills (42) | Asian Restaurant | Low to medium | Low | No |
| Blake-Jones (69) | Chinese Restaurant | - | - | No |
| Dufferin Grove (83) | Asian Restaurant | Low to medium | Low | No |
| Flemingdon Park (44) | Asian Restaurant | Low | Low | Yes |
| Islington-City Centre West (14) | Asian Restaurant | Low to medium | Medium | No |
| Kensington-Chinatown (78) | Asian Restaurant | Low to medium | Medium to high | No |
| North St.James Town (74) | South American Restaurant | - | - | No |
| Regent Park (72) | Asian Restaurant | Low to medium | Low | No |
| Steeles (116) | Chinese Restaurant | - | - | No |
| Weston-Pellam Park (91) | South American Restaurant | - | - | No |
| Wexford/Maryvale (119) | Asian Restaurant | Low | Low to medium | 2nd choice |
| Yonge-Eglinton (100) | South American Restaurant | - | - | No |

**Chinese cuisine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **1st most common venue** | **Home price** | **Crimes rate** | **Selection** |
| Ionview (125) | Chinese Restaurant | Low | Low | Yes |
| L'Amoreaux (117) | Chinese Restaurant | Low | Low to medium | 2nd choice |
| Lawrence Park South (103) | Chinese Restaurant | Medium to high | Low | No |
| West Humber-Clairville (1) | Chinese Restaurant | Low | Medium | No |

**Italian cuisine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **1st most common venue** | **Home price** | **Crimes rate** | **Selection** |
| Bedford Park-Nortown (39) | Italian Restaurant | Medium to high | Low | No |
| Corso Italia-Davenport (92) | Italian Restaurant | Low to medium | Low | 2nd choice |
| Junction Area (90) | Japanese Restaurant | - | - | No |
| Mount Pleasant East (99) | Italian Restaurant | Medium | Low | No |
| Mount Pleasant West (104) | Japanese Restaurant | - | - | No |
| New Toronto (18) | Italian Restaurant | Low | Low | Yes |
| Palmerston-Little Italy (80) | Italian Restaurant | Medium | Low | No |
| Yonge-St.Clair (97) | Italian Restaurant | Medium | Low | No |

**European cuisine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **1st most common venue** | **Home price** | **Crimes rate** | **Selection** |
| Casa Loma (96) | European Restaurant | Medium | Low | No |
| Long Branch (19) | European Restaurant | Low | Low | Yes |
| Playter Estates-Danforth (67) | European Restaurant | Medium | Low | No |
| Yorkdale-Glen Park (31) | European Restaurant | Low | Low to medium | 2nd choice |

**Indian cuisine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **1st most common venue** | **Home price** | **Crimes rate** | **Selection** |
| Greenwood-Coxwell (65) | Indian Restaurant | Low to medium | Low | 2nd choice |
| Thistletown-Beaumond Heights (3) | South American Restaurant | - | - | No |
| Thorncliffe Park (55) | Indian Restaurant | Low | Low | Yes |
| Woburn (137) | Indian Restaurant | Low | Medium | No |

**Japanese cuisine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **1st most common venue** | **Home price** | **Crimes rate** | **Selection** |
| Church-Yonge Corridor (75) | Japanese Restaurant | Low | Very high | No |
| Lawrence Park North (105) | Japanese Restaurant | Medium | Low | No |
| Pleasant View (46) | Japanese Restaurant | Low | Low | Yes |
| The Beaches (63) | Japanese Restaurant | Medium | Low | No |
| University (79) | Japanese Restaurant | Medium | Low | No |

**Middle Eastern cuisine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **1st most common venue** | **Home price** | **Crimes rate** | **Selection** |
| Dovercourt-Wallace Emerson-Junction (93) | Middle Eastern Restaurant | Low to medium | Medium | No |
| Forest Hill North (102) | Middle Eastern Restaurant | Medium | Low | Yes |

**South American cuisine:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Neighborhood** | **1st most common venue** | **Home price** | **Crimes rate** | **Selection** |
| Don Valley Village (47) | South American Restaurant | Low to medium | Low | Yes |
| East End-Danforth (62) | South American Restaurant | Low to medium | Low to medium | No |
| York University Heights (27) | South American Restaurant | Low | Medium | No |

To recap, my recommendation for someone who wants to choose in which neighborhood to open a new restaurant in Toronto city, according to the type of cuisine and taking into consideration real estate prices and crimes rate, is as follow:

|  |  |
| --- | --- |
| *Cuisine type* | *Neighborhoods* |
| American | Danforth (66) |
| Asian | Agincourt South-Malvern West (128), Flemingdon Park (44) |
| Chinese | Ionview (125) |
| Italian | New Toronto (18) |
| European | Long Branch (19) |
| Indian | Thorncliffe Park (55) |
| Japanese | Pleasant View (46) |
| Middle Eastern | Forest Hill North (102) |
| South American | Don Valley Village (47) |

## Conclusion

In this study, I analyzed the specifications of the neighborhoods of Toronto based on the category of restaurants present within their boundaries. I then, tried to refine the findings with data related to neighborhoods, such as real estate prices and crimes rate.

So this is like the first-order solution to the question “Where should we open a new restaurant in Toronto?”

However, there is certainly is lot of room for improvement. In fact, this study should be completed by a marketing study to analyze other factors in choosing the best neighborhoods to open a new restaurant.

These factors include demographic data related to future customers as well as competitors’ location. Demographic data refers to socio-economic data such as population, ethnicity, purchasing power and education which represent specific neighborhoods.

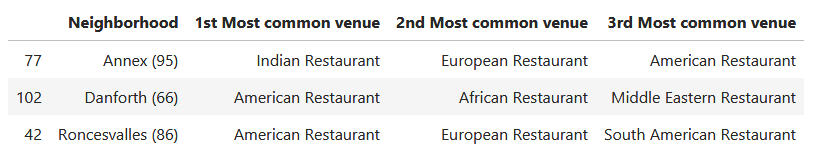
The competitor analysis can also be done by clustering the restaurants within each neighborhood, in that we certainly don’t want too many competitors in the same sector.

This marketing study should further help the new restaurant to choose its location, and also to define its positioning, pricing strategy, and offerings to perfectly meet future customers demand.

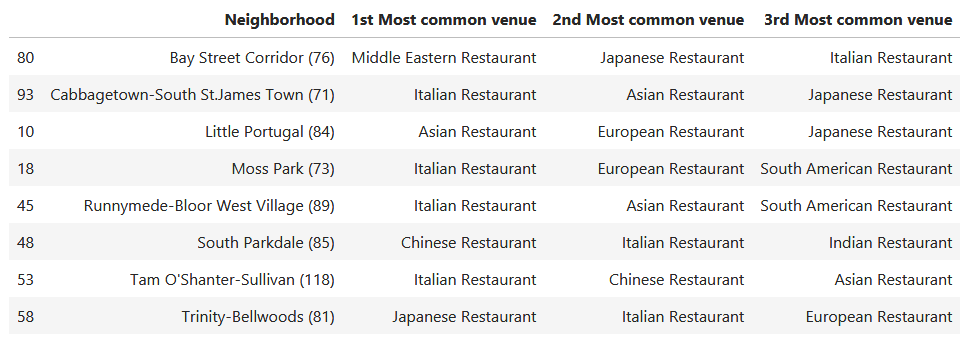
### APPENDICES

K-Means clusters:

Cluster 1:



Cluster 2:



Cluster 3:



Cluster 4:



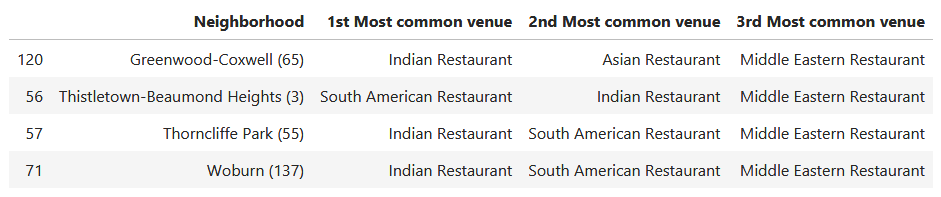
Cluster 5:



Cluster 6:



Cluster 7:



Cluster 8:



Cluster 9:



Cluster 10:

