# Comparing House Prices in Different Neighborhoods of Toronot, Canada

Edison Murairi

June 2020

## 1 Introduction

### 1.1 Background

Toronto will experience a rapid population growth over the next 20 years. Even worse, this growth is accelerating, and is expected to double by the year 2041 [1]. This rapid increase in the population number significantly affects the real estate in Toronto, leaving the low-income households most vulnerable to inadequate housings or to a lack of any at all.

This challenge requires that the real estate industry adaptates to the economic conditions of every social class in Toronto. Otherwise, it will not provide profitable and comprehensive service to meet this challenge. The first to do this is to understand the current real estate conditions in Toronto. Therefore, in this project, we will compare the house prices in neighborhoods of Toronto.

As this challenge will surely affect various areas - urban planning, environment, finance, etc. - the findings are intended to multiple stakeholders. These stakeholders who will benefit from these findings are the urban planning office of the Canadian government, the Ministery of environment, and also the real estate companies operating in the country.

### 1.2 Problem Statement

How does the price of houses vary between different neighborhoods of Toronto?

### 1.3 Data

This project requires first the housing price data. We have obtained house sales data of the Province Ontario [2]. This dataset contains 6 columns: address, areaName (the name of the area where the house is located), the price, the latitude and longitude of the location. Figure 1 shows an example of the house sales data.

| | AreaName | Price ($) | lat | lng |
|---|---|---|---|---|
| 0 | Richview | 999888 | 43.679882 | -79.544266 |
| 1 | Chedoke Park B | 399900 | 43.250000 | -79.904396 |
| 2 | Ainslie Wood East | 479000 | 43.251690 | -79.919357 |

Figure 1: House Sales in the Ontario Province

Secondly, we obtain venues data of each neighborhood from Foursquare using the postal codes obtained from the wikipedia page: "List of all Postal Codes in Canada M". From the Foursquare venues data, we will extract the neighborhood, the neighborhood latitude, the neighborhood longitude, the name of the venue, and the coordinates of the the venue (latitude and longitude) Venue and the category of the venue. We will combine these informations with the the house sales data to compare how the price varies between neighborhoods of different characteristics. Figure 2 shows an example of the venues data from Foursquare.

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Regent Park, Harbourfront | 43.6555 | -79.3626 | Roselle Desserts | 43.653447 | -79.362017 | Bakery |
| 1 | Regent Park, Harbourfront | 43.6555 | -79.3626 | Tandem Coffee | 43.653559 | -79.361809 | Coffee Shop |
| 2 | Regent Park, Harbourfront | 43.6555 | -79.3626 | Figs Breakfast & Lunch | 43.655675 | -79.364503 | Breakfast Spot |
| 3 | Regent Park, Harbourfront | 43.6555 | -79.3626 | Morning Glory Cafe | 43.653947 | -79.361149 | Breakfast Spot |
| 4 | Regent Park, Harbourfront | 43.6555 | -79.3626 | The Yoga Lounge | 43.655515 | -79.364955 | Yoga Studio |

Figure 2: Venues data in Toronto

## 2 Data Cleaning

We obtained house sales data in the Ontario province from Kaggle [2]. The dataset contains 25351 entries and 5 columns (Address, AreaName, Price, Latitude and Longitude).

Out of the 25351 entries, 483 or 1.9% of the dataset were missing AreaName. Because this is a small portion of the dataset, we remove all these rows with missing area names. Finally, we remove all the rows with a house outise Toronto. The final dataset hence obtained contains 5092 entries. Figure 3 shows the first 3 rows of the final dataset.

To obtain the venues data, we first scrapped the list of neighborhoods in

| | Address | AreaName | Price ($) | lat | lng |
|---|---|---|---|---|---|
| **0** | 86 Waterford Dr Toronto, ON | Richview | 999888 | 43.679882 | -79.544266 |
| **4** | #1409 - 230 King St Toronto, ON | Downtown | 362000 | 43.651478 | -79.368118 |
| **5** | 254A Monarch Park Ave Toronto, ON | Old East York | 1488000 | 43.686375 | -79.328918 |
| **6** | 532 Caledonia Rd Toronto, ON | Fairbank | 25 | 43.691193 | -79.461662 |

Figure 3: Preview of the Toronto House Sales Dataset

Canada M and their postal codes from wikipedia. This list contains 180 entries, 77 of which have no assigned neighborhoods. We remove these entries fromt he dataset. To limit this list only to neighborhoods in Toronto, we remove all the entries where the borough does not contain the name Toronto. The final list contains 39 entries. We use the postal codes to retrieve the latitude and longitude of each location in the list.

We pass the coordinates of each region to the Foursquare API to retrieve venues data for the locations. We obtain 1535 entries, none of is missing values. Furthermore, all entries have the correct format. Hence no further processing was needed for this dataset.

The final step is to combine the housing price dataset with the venue dataset from Foursquare. In other words, we need to find the venues that are in the neighborhood of each house sold. To do this, we assign to each house a neighborhood of 0.015° Latitude wide, and 0.025° Longitude wide. Then, for each house, we count the number of venue categories (Restaurant, Bakery, Spa, ...) falling into its neighborhood. An example of the final dataset that we will analyze is shown in Figure 4.

| | AreaName | Price ($) | lat | lng | Bakery | Coffee Shop | Breakfast Spot | Yoga Studio | Pub | Spa | Thai Restaurant | Event Space | Restaurant | Italian Restaurant | Distribution Center |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Richview | 999888 | 43.679882 | -79.544266 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | Downtown | 362000 | 43.651478 | -79.368118 | 9 | 38 | 8 | 2 | 10 | 1 | 3 | 1 | 9 | 4 | 1 |
| **2** | Old East York | 1488000 | 43.686375 | -79.328918 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | Fairbank | 25 | 43.691193 | -79.461662 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | Wallace Emerson | 113 | 43.664101 | -79.439751 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 4: Preview of the Final Toronto House Sales Dataset

# 3 Exploratory Analysis

We first use Folium to visualize where the geographical distribution of the houses sold in Toronto, Figure 5. The orange marks represent the houses with price in the first quartile. The blue marks represent houses with prices in the second and third quartile. The red marks are the houses with prices above the third quartile. We observe that the houses in all three price ranges are distributed uniform across the city.

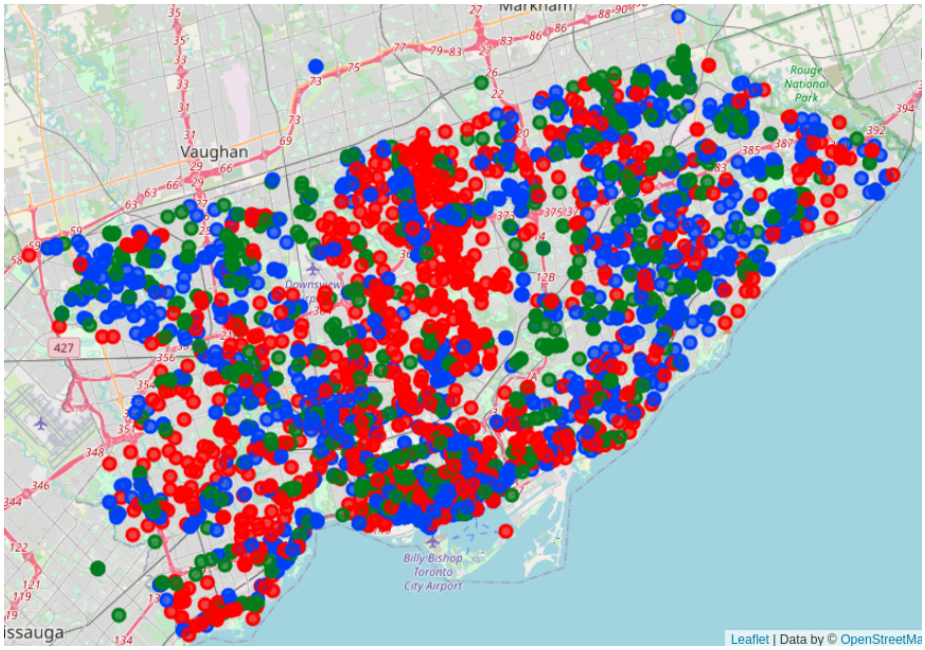


Figure 5: Geographical Distribution of House Sales in Toronto. The orange marks represent the houses with price in the first quartile. The blue marks represent houses with prices in the second and third quartile. The red marks are the houses with prices above the third quartile.

Next, we investigate the relation between the price of the house and the number of venues its neighborhood. Figure 6 shows the plot of the price against the number of venues in its neighborhood. We can observe that the most expensive houses have the least venues around them. The price falls rapidly as the number of venues in the neighborhood increase. It is as if "rich" people live in places where they are "alone".

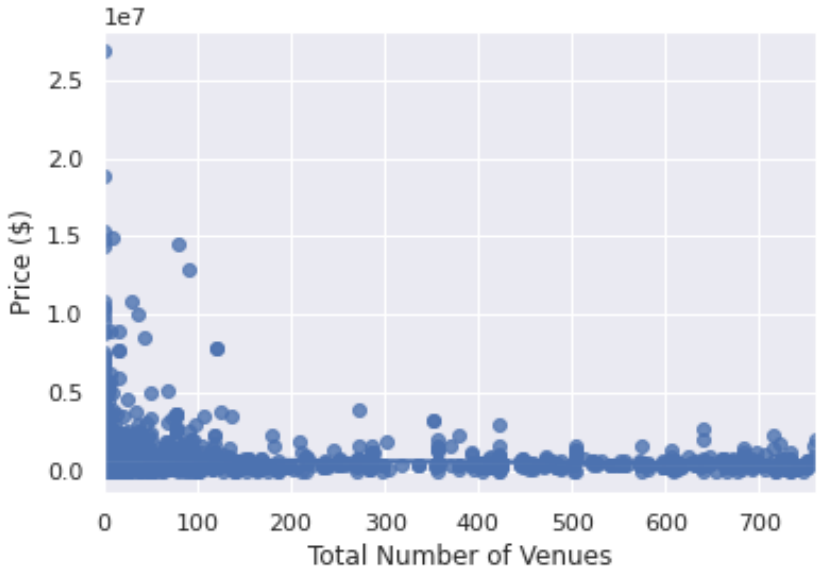


Figure 6: Price of houses against the numbe of venues in the neighborhood

# 4 Analysis

We have observed that the low price, medium price and high price houses are distributed uniformly across the city (Figure 5). We have also seen that the price falls off rapidly as the number of venues in the neighborhood increases. These two observations could be somewhat contradictory. To resolve this issue, we need a deeper analysis of how the price depends on the availability of venues in its neighborhood. We run a K-means clustering with 3 clusters to group houses with categories of different prices and different neighborhoods.

Figure 7 shows the map of Toronto with the houses grouped in three diffferent clusters after the K-means clustering. The houses marked in red (first cluster) corresppond to the cheapest prices (price below 1.728 Million USD). The ones marked in green are the intermediate ones (price between 1.749 Million and 6.995 Million USD), and the ones marked purple are the most expensive ones (prices above). In the section below, we present the results in more details.

# 5 Results

From Figure 7, we can observe that the most expensive houses (purple) are in the middle of the city. The ceneter of Toronto is probably the most expensive one. Another observation is that most houses fall into lower price category, and

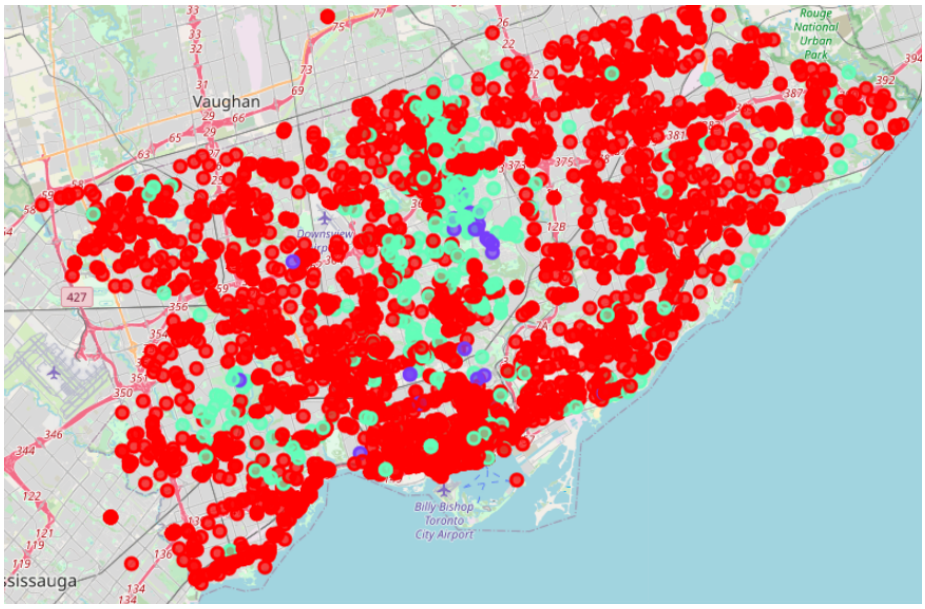


Figure 7: Map of Toronto with Houses Grouped in 3 Clusters

are distributed almost everywhere. This consistent with the earlier observation that the houses were distributed quasi-uniformly across the city regardless of their prices. Another interesting feature is that the outskirt of the ciy has almost only the lower price category houses. Contrary to the uniform distribution claimed earlier, the medium price houses are mostly in the Central Western side of the city while the Eastern side mostly has the lower price ones.

Figures 8, 9 and 10 show plots of the prices against the number of venues for the three different categories. We can see in each category that the price falls as the number of venues increases. Now, we investigate the distribution of venues in these three categories. From Figure 6, we expect that expensive houses will have fewer venues in their neighborhoods.

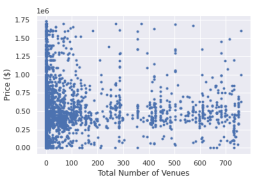


Figure 8: Category 1: Low Price

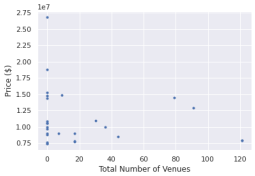


Figure 9: Category 2: Medium Price

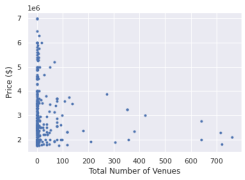


Figure 10: Category 3: High Price

Figure **??** shows the average number of venues in each category. We can see

that the average falls linearly as we go through the categories. If we consider the category labels as the x-axis, we predict that the average number of venues as compared to the low prices, medium prices and high prices category follow a linear model with slope $-17.5$ and x-intercept 83.2, or as given by the equation below:

$$\text{Average Number of Venues} = -17.5 \times \text{Price Category} + 83.2 \qquad (1)$$

where the price category is an integer. Therefore, if we know the number of venues in the neighborhood, we can predict whether the house will be cheap, medium price or expensive.

# 6 Conclusion and Recommendations

In this exploratory analysis, we have studied how the price of houses vary in neighborhoods of Toronto. We have ran a K-means clustering with 3 clusters to group the houses in 3 price and neighborhood categories. We have first found that the houses naturally fall into low, medium and high prices. The houses in the center of Toronto tend to be the expensive ones, suggesting that the center of Toronto is a rich area. The medium price houses tended to be in the West while the cheap houses are distributed across the entire city. Furthermore, most houses in the outskirt are the among the cheap ones.

Analyzing the distribution of the venues in the neighborhood of each house, we see that the number of venues falls rapidly as the number of venues increases. In other words, rich people live in areas where they are "alone". This suggests that most of venues in Toronto are probably small businesses. Furthermore, we derived a model to predict whether a given house is cheap, medium or expensive given the number of venues in its neighborhood.

In lights of these findings, we suggest further studies into the distribution of wealth and habitation in Toronto. This categorization could be the footprint of the social classes in the city. The future findings on the distribution of wealth will allow the city to provide affordable accomodation to most people in the future. This is crucial given that the population in Toronto is growing rapidly. Furthermore, we also see here an opportunity for employment research. It is important to confirm the types and wealth of businesses around the city since they seem related to the housing as shown in this study. Therefore, the city can understand what jobs they should provide to sustain its growing population.

# 7 Acknowledgement

We thank the Coursera IBM team for providing the training necessary to conduct this study through the Coursera Data Science Specializatio and Professional Certificate.

# References

[1] Canadian Urban Institute Canadian Center for Economic Analysis. Toronto housing market analysis. jan 2019. This report is published by the Canadian Center for Economic Analysis and the Canadian Urban Institute.

[2] Mahdy Nabaee. House sales in ontario. jan. This is a dataset available in Kaggle.