**Finding the Most Affordable Homing in Toronto**

Dhruv Jain

April 11, 2020

# **Introduction**

## Background

Buying a house is one of the toughest decision in a human’s life. Houses don’t come at a cheap price, and it is usually a long-term investment. A person should carefully choose where to buy a house, because if someone doesn’t like their house, it isn’t easy nor cheap to move to a new one. Buying the right house can be a more difficult choice than choosing which university to attend.

The cost of houses in Toronto in 1985 was $109,094. In 2015, that number rose to $566,696. In February 2020, the cost of the same house was $852,900, growing a whopping 60% in the past 5 years. The change in house prices between February 2019 and February 2020, itself, was 10.15% [1][2]. If the same trend continues, the housing market will break the 1 million mark within the next 2 years (if the market does not falls due to the COVID-19 pandemic).

Therefore, selecting the right house is a very important decision and it would be better if the person lives in the neighborhood where whatever they want is at their disposal and the price of his house is affordable.

## Problem

This project aims to categorize the neighborhoods based on the venues and services around it and choose the best neighborhood according to the affordability of the houses and the venues available around it.

## Interest

People who want to buy a house in Toronto and want to compare the neighborhoods on the services they have would be very interested. Researchers who are finding reasons for the growth in market may also be interested.

# **Data Acquisition and Cleaning**

## Data Sources

The housing price of the neighborhoods (HOUSES) in Toronto were obtained from [here](https://toronto.listing.ca/real-estate-prices-by-community.htm), and the coordinates of the neighborhoods (LOCATION) in Toronto were obtained from [here](https://open.toronto.ca/dataset/neighbourhoods/). Foursquare was used to get the data of the venues in the neighborhoods.

## Data Selection and Cleaning

The data was downloaded and scraped from these websites and combined into a table. It was found that one neighborhood has various property types and the properties has various house types. The data scraped from the Real Estate website had a lot of missing data, mainly because some neighborhoods do not have a certain type of property or a certain type of house. For example:

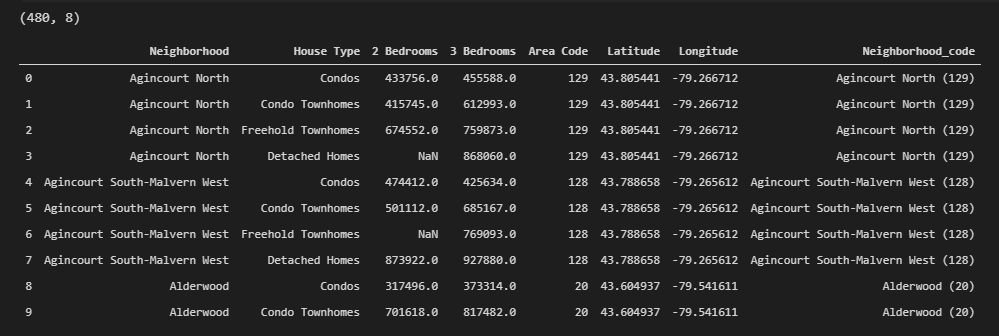
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Neighborhood | House Type | 1 Bedroom | 2 Bedroom | 3 Bedroom |
| A | Condos | 10 | 20 | 30 |
| A | Detached Homes |  |  | 100 |
| B | Condos | 30 | 60 |  |
| B | Town Houses |  | 50 | 75 |

*Fig 1: Example of the Data structure*

To reduce the loss of data when rows with No available values would be drop, only houses with 2 and 3 bedrooms were chosen. This was mainly because people who are looking to buy houses would mostly comprise of individuals and families and they would prefer to live in a 2/3-bedroom house rather than a smaller or larger house. And, it is the most common type of housing layouts.

Secondly, not all neighborhood names in the HOUSES dataset matched with the ones in LOCATION. A script was written to extract the neighborhood names that were not an exact match and it was examined. The names which were similar in both datasets (only differed in symbols and spacing) were changed in the HOUSES (to match the ones in LOCATION). Dissimilar names were researched manually, and it was found that the neighborhoods were same, only the names differed, so that was changed according.

It was also found that some neighborhoods in the HOUSES dataset were further divided into 2 parts (for example: Neighborhood ‘Rouge’ was divided to ‘RougeE10’ and ‘RougeE11’). The average of the house prices in these 2 neighborhoods were taken, assigned to the first neighborhood, and the other one was dropped from the dataset. Then the first neighborhood’s name in the HOUSES dataset was changed corresponding to the same one in the LOCATION dataset.

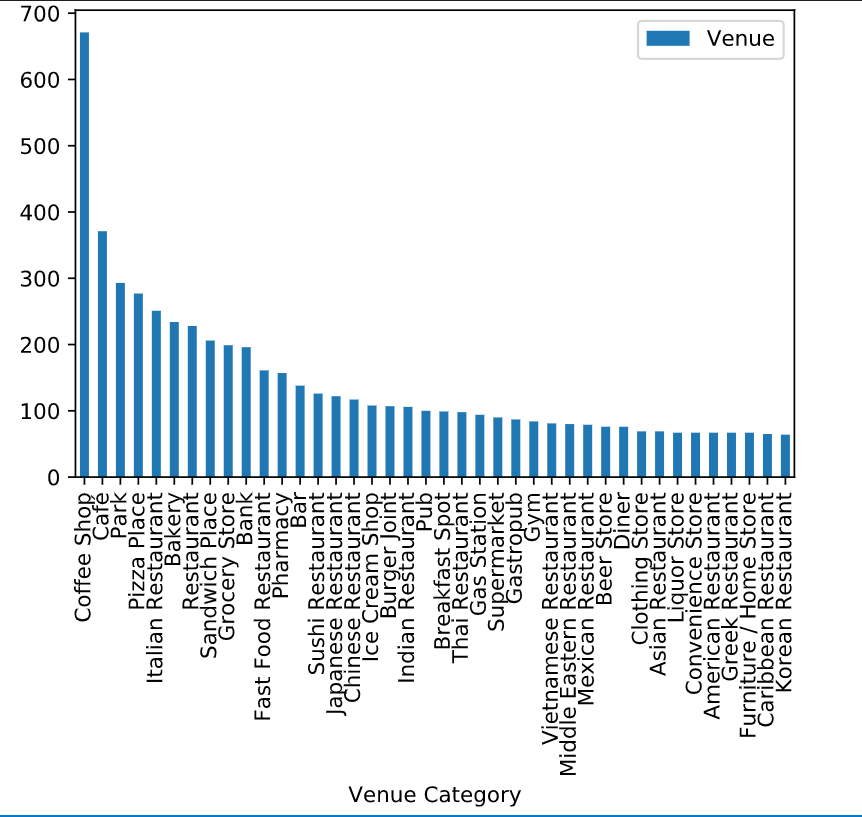


*Fig 2: Merged Houses and Location Data structure*

Foursquare API was used to get and explore the Venues and Venue Categories in each neighborhood. The query was used to get venues in a 1500 meters radius around the neighborhood, because 1.5 kilometer is a reasonable amount for walking and cycling.

# **Exploratory Data Analysis**

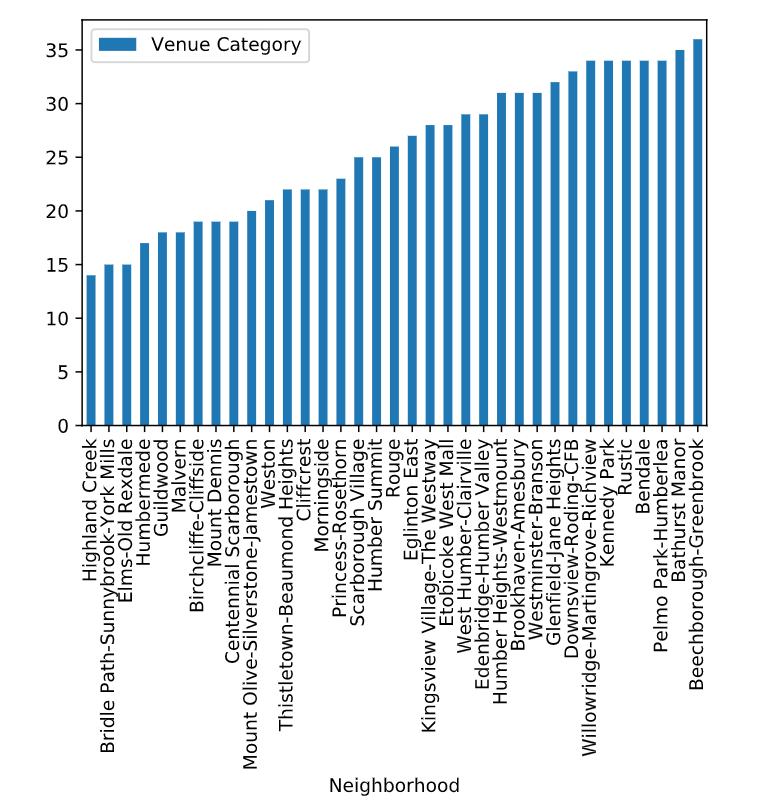
The extracted venues were counted by their venue category and the corresponding graph was plotted. Only the top 40 venue categories (out of 353) were selected because more categories would have made the graph unreadable.



*Fig 3: Number of Venues vs Venue Category Graph*

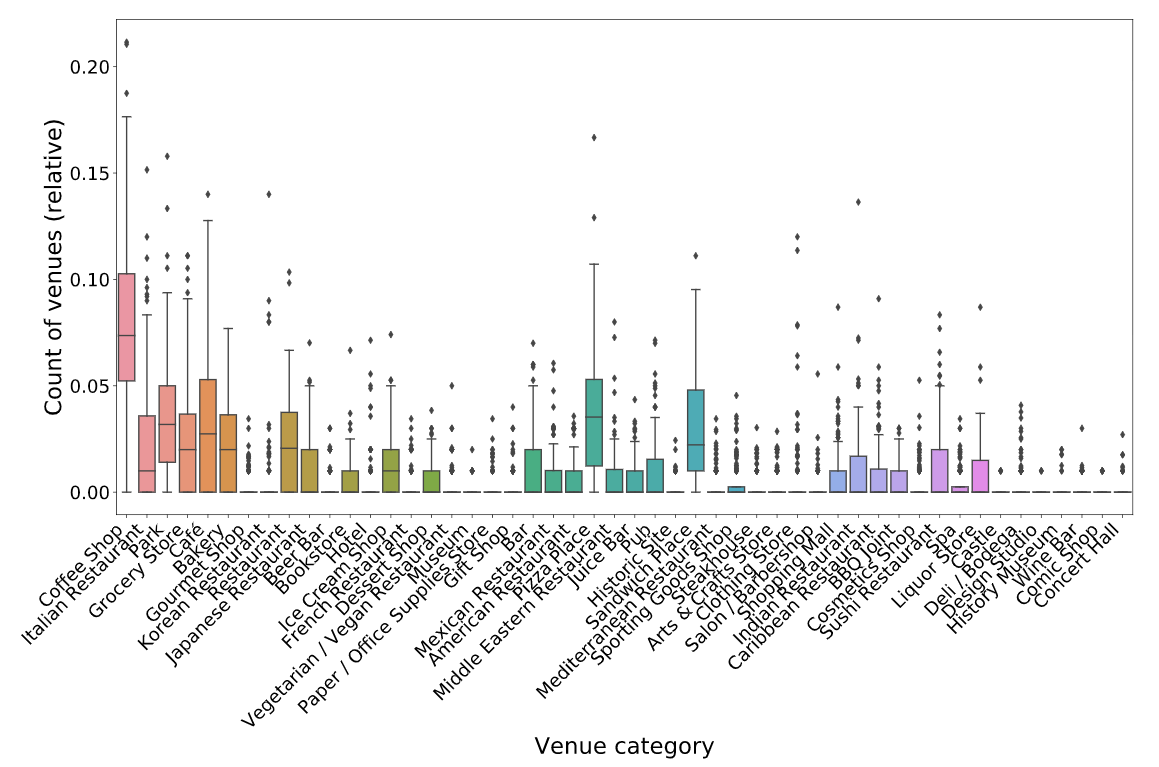
This may not be the true value of the of total number of venues in the venue category as it does not account for duplicate venues. (considering that 1500-meter radius is quite large and might extend the boundaries for the small neighborhoods, resulting in one venue being included in 2 or more neighborhoods)

For the same reason, another graph of Venue Category vs Neighborhood is plotted. Here, it does not matter if the venues are duplicate as a particular venue can be accessed by 2 different neighborhoods. For the same reason as above, only 35 Neighborhoods (of 140) were selected.



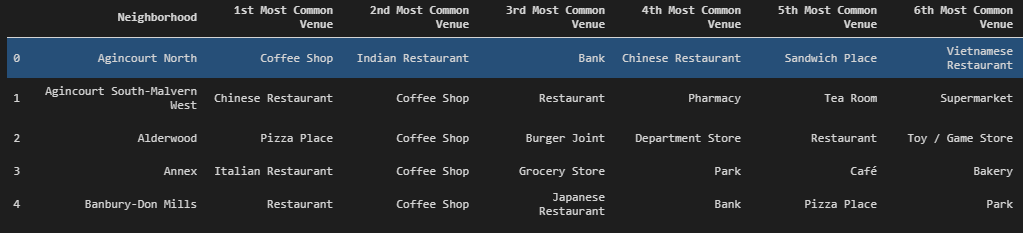
*Fig 4: Number of Venues vs Neighborhood Graph*

The dummy variables of each Venue Category are assigned and the means of the frequency of those variables is calculated. That calculated mean is then used to graph a box plot showing the distribution of the venue categories in the neighborhoods.



*Fig 5: Distribution of the categories in the Neighborhoods*

Finally, the most visited Venue Categories in each Neighborhood were found by sorting the Neighborhood on frequency of the categories and the top 10 Venue categories in each Neighborhood was printed. (Only the first 6 is shown below)

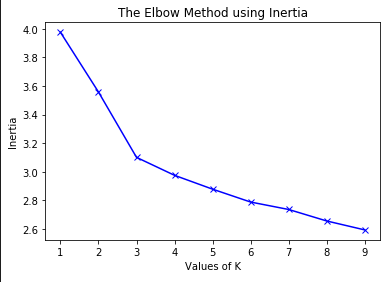
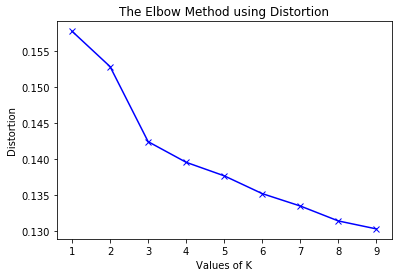


*Fig 6: Top 6 Venues in Each Neighborhoods*

# **Predictive Modeling**

## Classification Model

K-means clustering was used to model the problem. The optimum number of clusters was determined by the Elbow Method using Distortion and Inertia. Distortion is the average of the squared distances from the cluster centers of the respective clusters and Inertia is the sum of squared distances of samples to their closest cluster center[3].



*Fig 7: Elbow Method for K-Means Clustering*

Both the graphs show a huge elbow when the K value is 3 and a smaller elbow when the K value is equal to 6. Mapping the model when the number of clusters was taken as 3 provided with the results shown in Fig 8, which mainly divided the clusters to Downtown Toronto and the suburban areas of Toronto.

A K value of 5 was then taken, and the results of the model are shown in Fig 9. Here the suburban areas of Toronto were divided into smaller clusters, but Downtown was still one big cluster which had to be further divided.

Therefore, a K value of 6 was then taken, results shown in Fig 10, and that model provided with better results, dividing Downtown into 2 clusters.

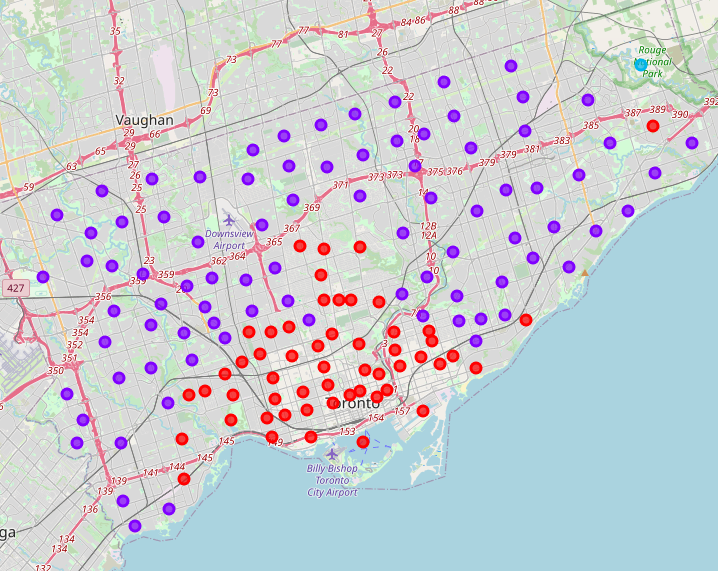


Fig 8: Map of Toronto with 3 clusters

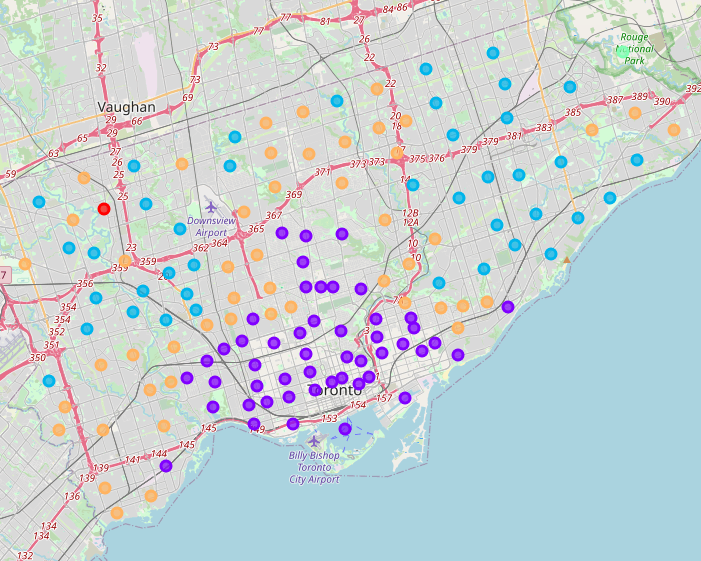


Fig 9: Map of Toronto with 5 clusters

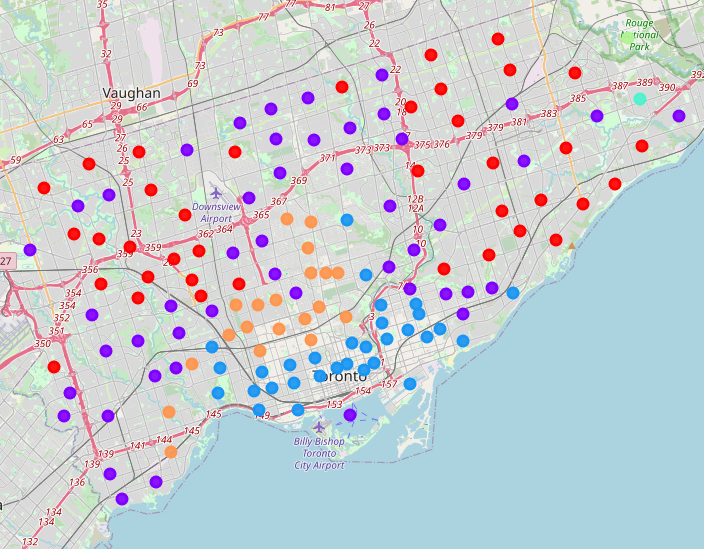
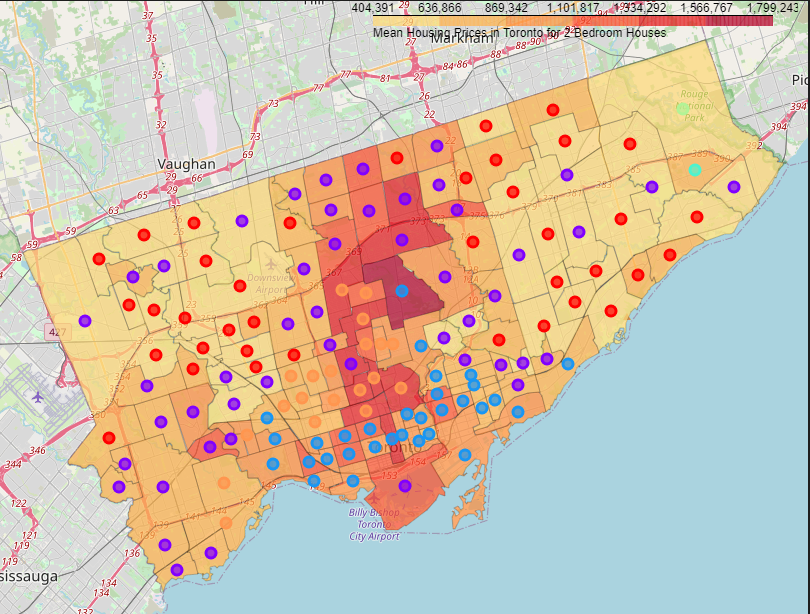


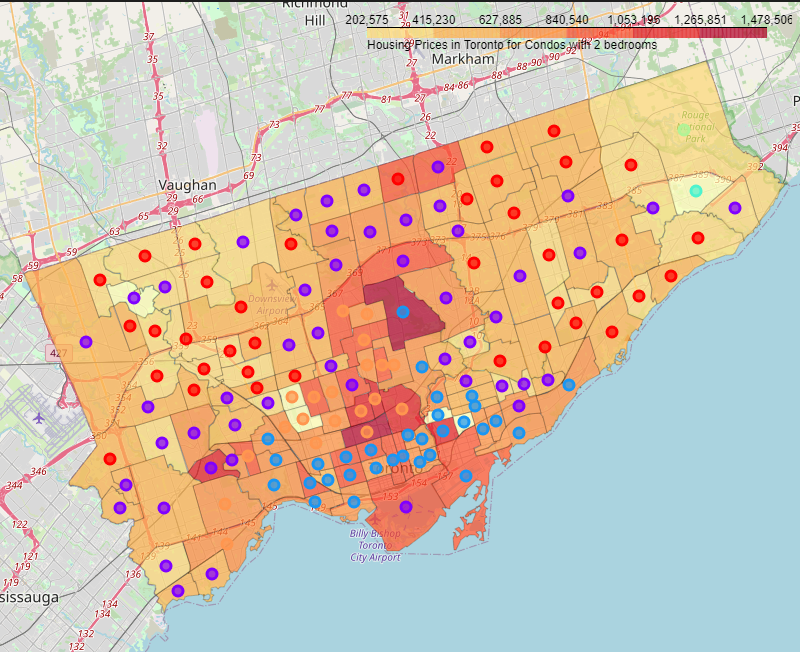
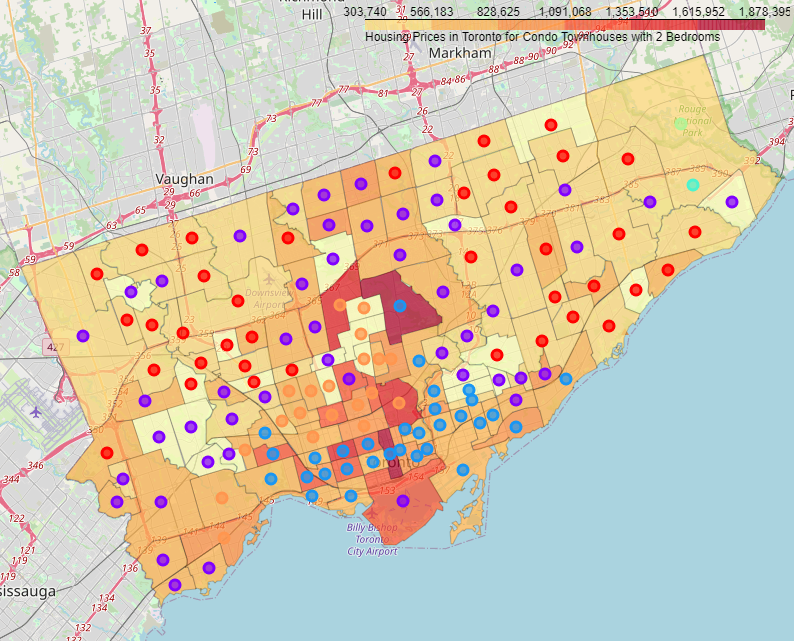
Fig 10: Map of Toronto with 6 clusters

# **Results**

Now that the clusters were defined, a choropleth map for the housing prices was created with the cluster markers. Fig 11 shows the mean housing price for 2 Bedroom houses. This map can be further divided to show the housing price for each house type, (Fig 12)



*Fig 11: Mean Housing Prices for 2 Bedroom Houses*



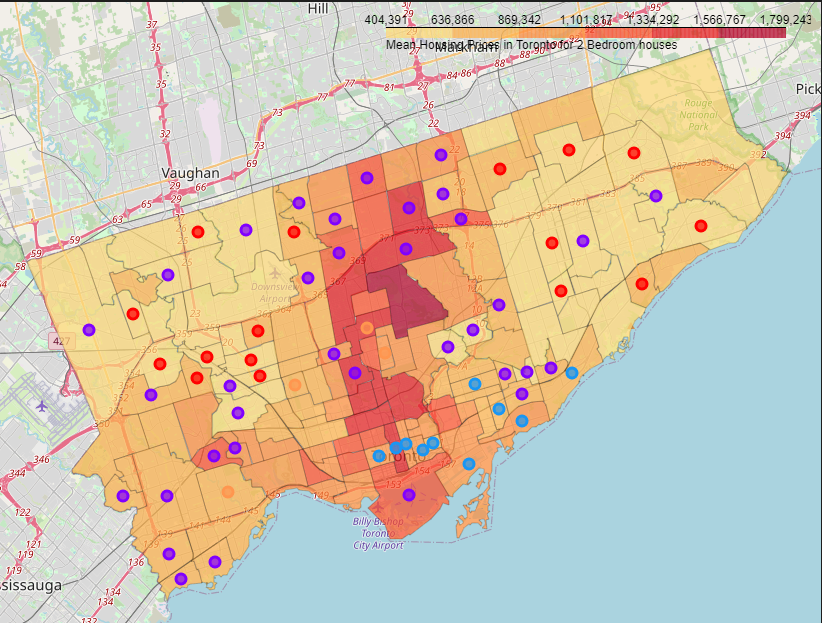
*Fig 12. Housing Prices for 2 Bedroom Condos and Condo Town Houses Respectively*

After examining each cluster, it was found that the clusters were created upon the following

|  |  |  |
| --- | --- | --- |
| Cluster | Color | Popular Venues |
| 0 | Red | Coffee Shop, Fast Food Restaurant, Sandwich Place |
| 1 | Purple | Coffee Shop, Bank, Pharmacy |
| 2 | Blue | Café, Park, Bakery, Restaurant |
| 3 | Greenish-Blue | Burger Joint, Coffee Shop, Pub |
| 4 | Green | Zoo, Trail, Fast Food Restaurant, Dessert Shop |
| 5 | Orange | Italian Restaurant, Bakery, Coffee Shop, Café |

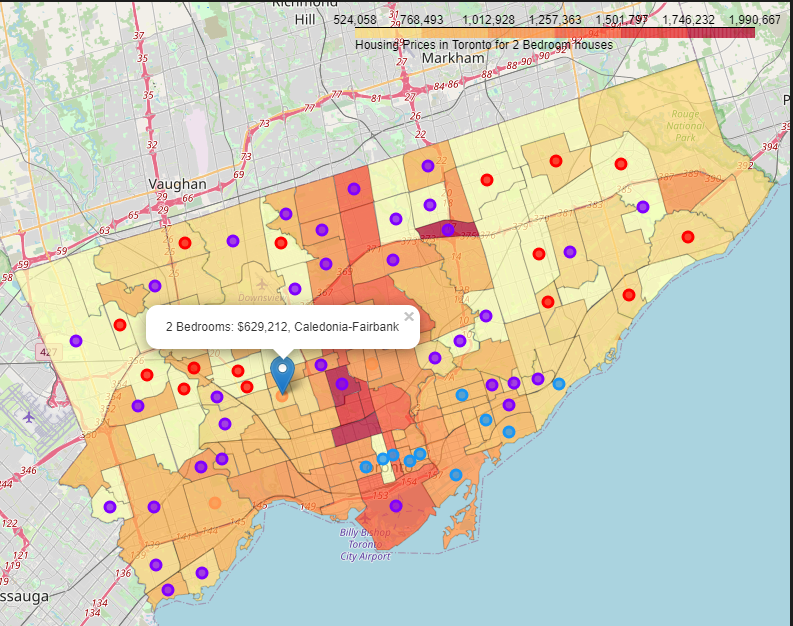
*Fig 13: Results of each Cluster*

To find the best neighborhood, the user needs to input a desired venue. In the following figure, Pizza Place and Coffee Shop were inputted for venue, and only the Neighborhoods in which the same 2 locations were the most popular were given as outputs.



*Fig 14: Clusters showing neighborhoods where Pizza Place or Coffee Shop are the most popular*

From that map, the lowest housing price for each remaining cluster was selected, of all house types (meaning the lowest price could be of any house type)



*Fig 15: Lowest House Price of all House Types of Cluster 5 having Coffee Shop or Pizza Place as the greatest number of venues*

# **Discussion**

This best neighborhood is only found considering the housing price and the desired venues in that neighborhood. It does not take into account the crime rate, the air quality, the traffic, the distance from work, or even the how good the venue is in the neighborhood are. People, generally, would prefer to pay more and live in a safe neighborhood and better rated venues than live in a cheap neighborhood with a high crime rate and bad venues. Therefore, if someone is looking to find the best neighborhood, they should not rely completely on this, without doing any research on the provided neighborhood.

# **Conclusion**

In this study, the neighborhoods of Toronto were analyzed and classified according to the number of Venues in the Venue Categories in the neighborhoods. Then the housing price for each neighborhood was plotted on a Choropleth Map, with the cluster markers added on top of it. The desired Venue Category was inputted, and neighborhoods where the desired venue was not the most popular was filtered out. The lowest housing price for each cluster was found. This was how the best neighborhood was found: of a person’s liking and the most affordable housing.

# **Future Directions**

In this study, only the minimum house price of all the house types was found. Future models could most certainly be divided by House types, providing with 24 more maps. Also, currently the Neighborhood containing the desired Venue Category is only found from the most popular column. In the future, all the neighborhoods where the desired Venue Category is in the Top 5 or 10 popular Venues Categories will be considered, so that more desired Venue Categories could be filtered. Another thing that can be done is considering the crime rate, the air quality, the traffic, the distance from work, the rating of the venue, etc. while finding the best neighborhood.

**References:**

[1] - <https://business.financialpost.com/personal-finance/mortgages-real-estate/now-and-then-do-canadian-homes-really-cost-that-much-more-than-30-years-ago>

[2] - <https://creastats.crea.ca/en-CA/>

[3] - <https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/>