Introduction

Finding new place to stay is a hard thing to do. Several parameters need to be considered to find the best place to stay at, for example: house prices, crime rate, population density, and nearby venues. House price is important because usually people who want to find a new house have certain budgets and they want to buy (or rent) a new house within their budgets. Even if they have huge budgets, they would want to find the best place with minimum price. House price may be affected by several variables such as crime rate, population density, location, the number of floors, house materials, etc.

Besides the house price, they also need to find place with low crime rate, low population density or high population density depending on preference, and nearby venues that fit their daily activities. The parameters mentioned above can be correlated to each other.

In this project, I use Toronto as the object. Toronto's Neighborhoods are clustered based on the parameters above and the best neighborhoods will be chosen. Although there are more parameters that need to be considered, for the sake of simplicity and data availability, I limit the parameters considered to house prices, crime rate, population density, and nearby venues only. Moreover, the final decision is decided based on preferences of individuals and more research should be made. If possible, visiting the neighborhood in person will be better. Therefore, the results of this project can only be used to narrow the options of neighborhood in Toronto to stay at.

Business Problem

There are around 140 neighborhoods in Toronto. Choosing place to stay in Toronto can be hard because of this. Therefore, reducing the options of neighborhood is important not only for the potential buyer (person who wants to relocate to Toronto), but also for the real estate company. For the potential buyers who have budgets and other preferences, reducing the number of neighborhood options is important because it helps them to reduce the time and effort needed to look into the neighborhood one by one. On the other hand, for the real estate companies, reducing the number of neighborhood options is important because they can focus to promote certain neighborhoods to their potential buyers (or renters).

Target Audience

1. Potential House Buyer
   Due to huge number of neighborhoods in Toronto, looking for the best place to live at is quite exhausting. The potential house buyer may spend a lot of time and efforts to look into the neighborhoods. Because of this, narrowing down the neighborhood options is really important, and this is what this project does. This project helps the potential buyers to find the neighborhoods that fit into their preferences and they can start searching for new house in those neighborhoods.
2. Real Estate Agency/Company
   On the other hand, for the agencies, this project helps them to focus on promoting houses at certain neighborhoods that fit into the potential buyer preferences. This may help increasing the house purchasing rate.

Data

Several parameters that are used in this study to narrow down the Toronto's neighborhood options are:

1. House price
2. Crime rate
3. Population density
4. Nearby venues

House price is affected by many parameters such as locations, number of bedrooms, conditions, etc. However, in this project, I only consider the average house price in each neighborhood. The source for this data is https://www.zolo.ca/toronto-real-estate/neighborhoods. This website provides average house price, percentage of house sold after 10 days of listing, the number of active listings, and percentage of house sold after the potential buyers asked about the unit in each neighborhood. Based on the source, the data is based on the last 28 days data.

The source of crime rate and population density data is Toronto Public Service Public Safety Data Portal (https://opendata.arcgis.com/datasets/af500b5abb7240399853b35a2362d0c0_0.csv?outSR=%7B%22latestWkid%22%3A26717%2C%22wkid%22%3A26717%7D). This source provides the number of crimes occured in each neighborhood every year from 2014-2018 complete with the types, the number of population and area for each neighborhood. The population density can be calculated using number of populations divided by the neighborhood area.

Nearby venues can be determined using Foursquare API (https://foursquare.com/). Foursquare API provides many data including venue name, type, photos, tips, and many others.

	Neighbourhood	Average sale price	Population	Shape__Area	Total_Crime_2018	density	Latitude	Longitude
0	Agincourt North	703000	31528	7.261857e+06	513.0	0.004342	43.808038	-79.266439
1	Humber Summit	648000	14188	7.966905e+06	59.0	0.001781	43.760078	-79.571760
2	Kingsview Village	619000	14306	5.063371e+06	263.0	0.002825	43.699539	-79.556346
3	West Hill	663000	36371	9.625440e+06	139.0	0.003779	43.768914	-79.187291
4	Agincourt South	676000	28410	7.873163e+06	154.0	0.003608	43.785353	-79.278549

How the data can be used to solve the problem

To be able to fulfil the objective of this project which is to narrow down the neighborhood options to live at Toronto, several things need to be done with the data mentioned in previous section. The steps are:

1. Exploratory data analysis

   This step can be done by plotting the boxplot or distribution of each variables (house price, crime rate, and population density). In this project, I use the boxplot. By plotting the boxplot, we can get the information of skewness of the data, outliers, quartiles, maximum and minimum values. We can remove the outliers if necessary. We also need to look whether the variables are correlated to each other or not to determine which machine learning method to use and to possibly reduce the dimensions. This can be done by plotting scatter plots between each variable or determine the pearson coefficient and the confidence level. In this project, we will determine the correlations visually.

2. Clustering analysis I

This step is done to find the neighborhoods with low average house price, low crime rate, and low population density. It is done by clustering the neighborhoods using the 3 variables (house price, crime rate, and population density). The low average house price and low crime rate are obvious parameters. But low population density is a matter of preference. I personally prefer living at not too crowded area, so I will use low population density as one parameter to determine the cluster. For other preferences the same step can also be done with several modifications. This step can be done using k-means clustering analysis.

3. Finding the nearby venues for chosen neighborhoods

   This step is done using the foursquare API. For the chosen neighborhoods (from step 2), top 5 venues within 500 meters are populated and included into the data frame.

4. Clustering analysis II

   The second clustering analysis using k-means clustering method is done to determine the clusters of neighborhoods with similar venue types. This is done to further narrow down the neighborhood options based on personal preferences. For example, if the potential buyer is interested in finding a place that is near restaurants, we can suggest the cluster of neighborhoods that has restaurants as the most commonly visited venue.