Locating the Desirable Locations for New Business to Setup Office

YuanCheng Zhou

August, 12, 2020

1. Introduction

1.1 Description & Discussion of the Background

2020 is a difficult year as everyone is affected by the pandemic of the Covid-19. This deadly virus has caused thousands of deaths around the globe, which it forced people to self-isolate and businesses to shut down. As a recently graduated university student who is also living through this global disaster, has experienced difficulty to find a job in the Greater Toronto Area in Canada, which I believed this is true to most people as well. I was fortunate to received several call interviews from the new Tech Startup CEOs about position in data analyst or their data team, but was also unfortunate to not offered the role because my lack of professional experience at that time. I was surprised to know that even everyone is trying to social-distancing and the economy is declining, there are still many passionate entrepreneurs that aren't discouraged by the global pandemic. However, despite the great business ideas from the Startup founders and the huge number of talented people in Greater Toronto Area, when consider starting business in Greater Toronto Area, the budget management will always be a problem to deal with at the beginning because of the equipment setup, salary, and most importantly the office setup. In Canada, Ontario is the province that has the highest average house rental price, and the Greater Toronto Area lies within Ontario that new Startup will face high prices when trying to lease an office at the beginning. It is possible that new Startup can still find cheap prices to lease an office, but other elements are needed to consider are whether the office location's surrounding traffic system and the venue types can provide employees fast travel to work, and convenient dining to replenish energy quickly that less time be wasted during the work hours. Therefore, it is essential to find the most desirable office location for business to success.

1.2 Problem

The data that might help to determine the desirable office locations in Greater Toronto Area might include regional average house rental price and area venue categories. This project aims to find the locations with low leasing prices, sufficient public transports and restaurants based on these data.

1.3 Interest

The Startup people who are looking for opportunities in Greater Toronto Area would be interested in finding the desirable office location for saving budget, attract employees and better work hour management.

2. Data acquisition and cleaning

2.1 Data sources

The average house rental prices are found in the Canada Mortgage and Housing Corporation dataset here. The dataset found from the Canada Mortgage and Housing Corporation provided average house leasing price for every neighbourhoods in Canada, however, the dataset does not provide the coordinates for each neighbourhoods and the neighbourhoods are named with different complexity, which it caused the python library Geocoder unable to accurately search the neighbourhood's coordinates based on neighbourhood's name. To fix this issue, I manually searched the coordinates for each Greater Toronto Area neighbourhoods by using Google Map and created a new dataset based on the one founded in the Canada Mortgage and Housing Corporation website. The new dataset can be found here where I uploaded to my Github repository. The final venue category data are acquired by using Foursquare API to explore the first 100 venues around neighbourhoods within a range of 1000 meters.

2.2 Data Cleaning

The Data downloaded from the Canada Mortgage and Housing Corporation were included with rows of data descriptions and corporate information, and thus these rows were removed from the dataset. Furthermore, the whole data were filled with both English and French, but only the English data were needed, therefore removed the French parts from the dataset.

After that, I changed the name of columns such as "Total" to "Average Price", the names of province "Ont." to "Ontario", the Zones like "Etobicoke (South)" to "Etobicoke" that are inside the Greater Toronto Area and split the Neighbourhoods with a slash "/" that has multiple neighbourhoods in the same row and shares the same average rental price, then created new rows for each split neighbourhoods in order to support accurate venue search and consistent data structure.

2.3 Feature Selection

From examining each feature, there were some redundancy in the features. The redundant features are “Dwelling Type” and “Bedroom Sizes”, there are no specific rules on selecting building types for office as people may select any type of buildings. Since this project is only aiming at the desirable office locations but not building types, if trying to specify every single dwelling types, bedroom or room sizes will cause the result harder to acquire and create unnecessary confusions, therefore I only focused on “Total” for all the dwelling types and “Total” for all the bedroom sizes to get an overall average house rental price for the neighbourhoods in Greater Toronto Area, then dropped the rest of the dwelling types, bedroom sizes and neighbourhoods that are not inside Greater Toronto Area. Eventually there are 5 features and 155 Neighbourhoods were selected.

Kept features	Dropped features	Reason for dropping features
Dwelling Type: Total	Dwelling Type: Row, Apartment & Other	Focusing on overall average house rental price for all the building types to better interpret the result.
Average Rental Price: Total	Average Rental Price: Bachelor Studios, 1 Bedroom, 2 Bedroom, 3 Bedroom	Focusing on overall average house rental price for all the bedroom or room sizes to better interpret the result.
Province, Zone: Ontario, Greater Toronto Area	Province, Zone: Other Provinces in Canada	This project only interested in searching the desirable office setup locations in Greater Toronto Area.

3. Exploratory Data Analysis

3.1 Setting the price level for the average house rental prices

Used linspace() function from Numpy package to bin the average house rental prices with 10 different price levels, this helped to analyze the average prices to see if it is relatively at low price, or higher price compared with all other average prices within Greater Toronto Area.

The 10 price levels came from the following histogram where there are 10 sections of price ranges in between lowest average price and highest average price.

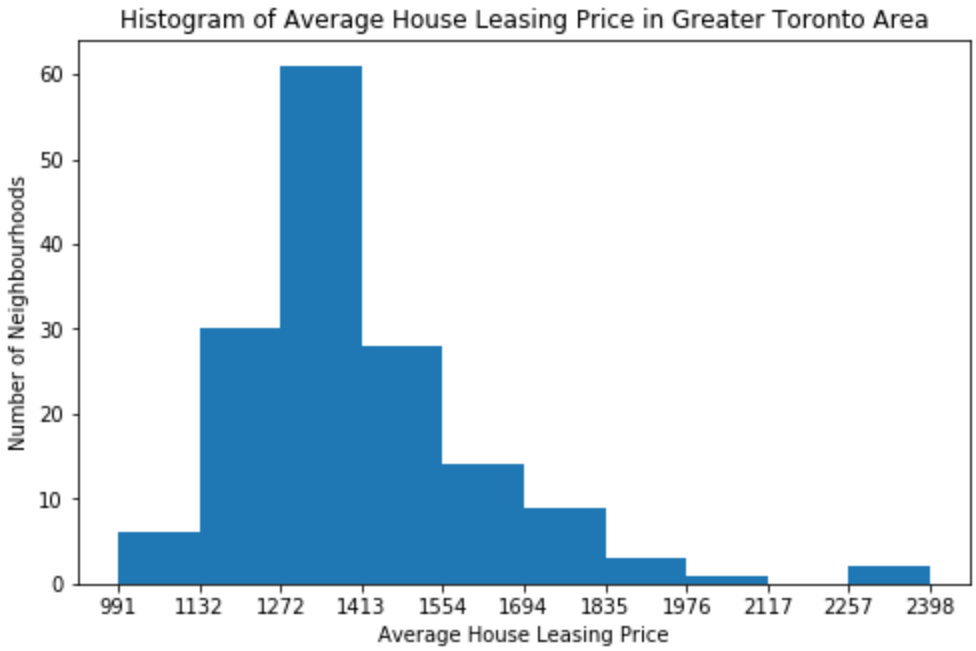


Figure 1. Histogram of Average Housing Leasing Price in Greater Toronto Area

We can see that most neighbourhoods have the average house rental price between 1272 and 1413.

After using the information gained from histogram and labeled the price levels on all the neighbourhoods, I created a bar chart to look for more information.

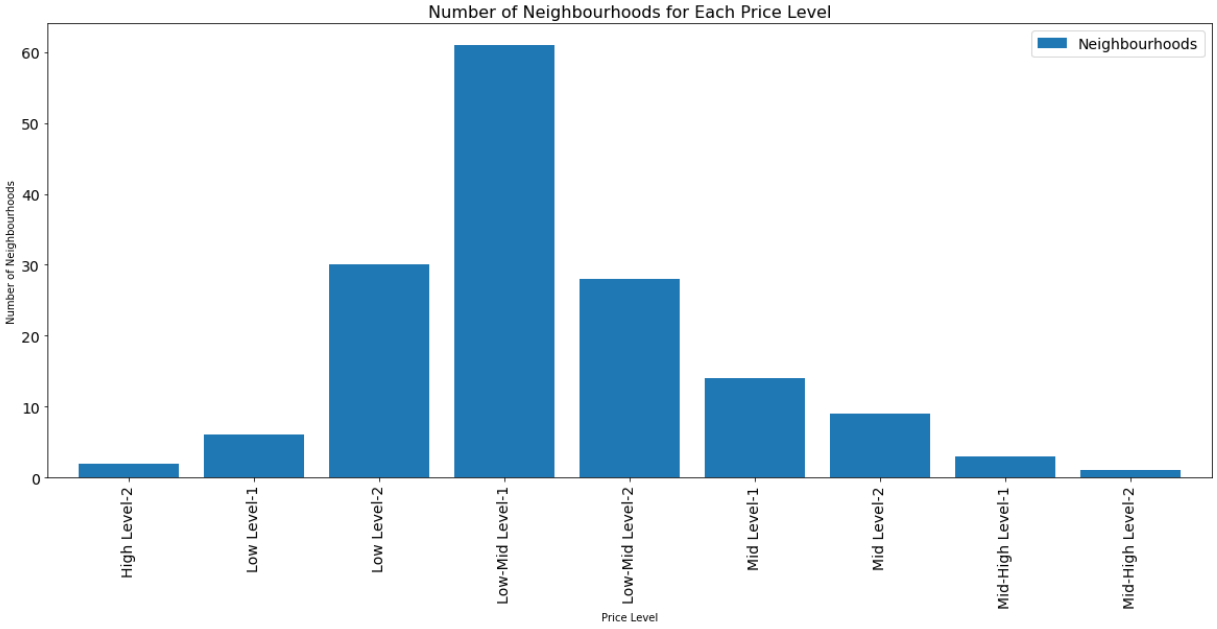


Figure 2. Number of Neighbourhoods for Each Price Level

The bar chart shows that price level at Low-Mid Level-1 has most number of neighbourhoods, therefore most neighbourhoods have this price level that is between 1272 and 1413.

3.2 Correlation between average house rental price and venue category

The house rental price can be affected by surrounding types of venues, then I created a boxplot to visualize the correlation between average house rental price and venue category in order to prove my assumption is true.

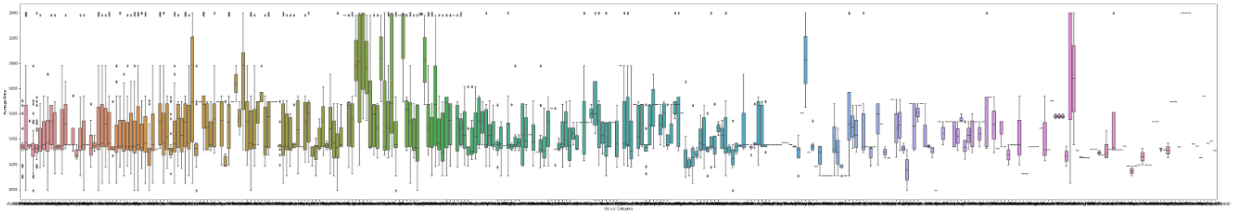


Figure 3. Boxplot of Average House Rental Price VS. Venue Category

There are 330 unique venue categories for all the neighbourhoods, therefore the boxplot might be crowded with many boxes. From the boxplot, the boxes are not aligned together and their mediums are varied as some of the boxes are higher and some are lower in the boxplot. Now it is certain to say the venue category have an affect on the average house rental price.

4. Cluster Modeling

The cluster model is the ideal solution to find the desirable office location, by using the venue categories from each neighbourhood, the cluster model can group neighbourhoods with similar venue categories into same cluster. It is helpful to cluster same venues together, that it allows me to look through the prices of neighbourhoods within a cluster. For example, if a cluster has of neighbourhoods with venues such as restaurant, café, or public transport. Since I am looking for desirable locations with sufficient public transport and restaurant, this gives me advantage to simply look through low prices in that single cluster to choose the desirable neighbourhoods as the locations to setup office.

4.1 Finding the best K for K-Means Clustering

I used K-Means Clustering to build a cluster model is because outliers are not a significant factor when I am only looking for neighbourhoods with sufficient public transport and restaurant. In other words, neighbourhoods with only one public transport and two or three restaurants will suffice to be a good area for office, the remaining relies on the average price.

To pick the best K, I used the Elbow Method and visualizes the distortion for K ranging between 1 and 10 in a line graph.

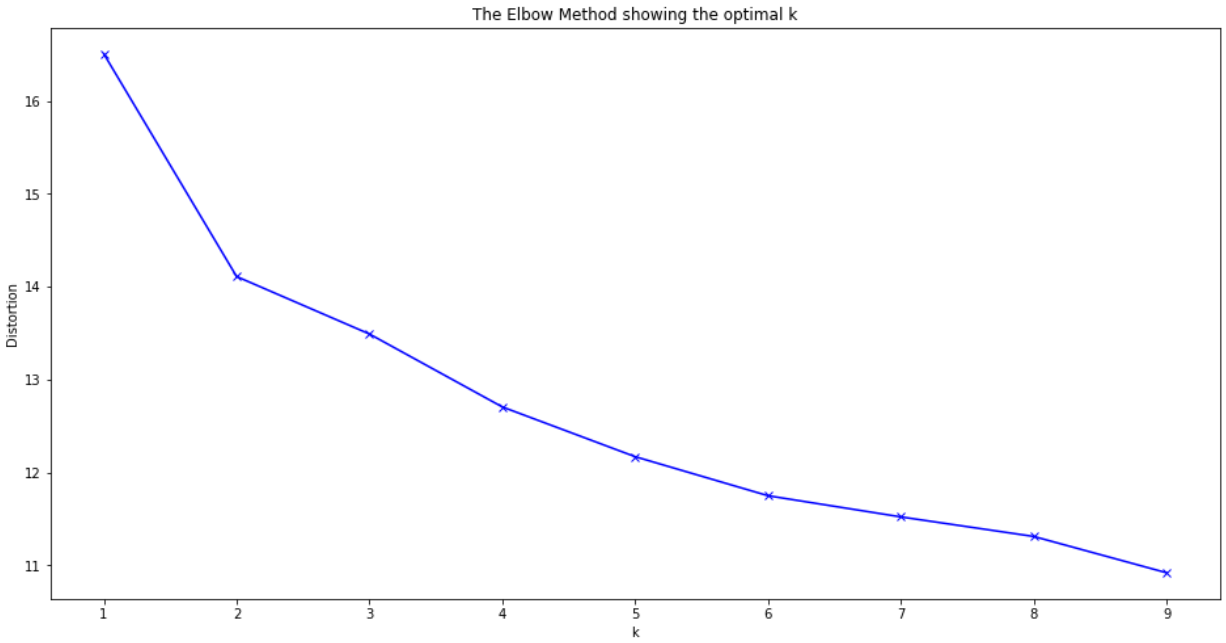


Figure 4. The Elbow Method showing the optimal K

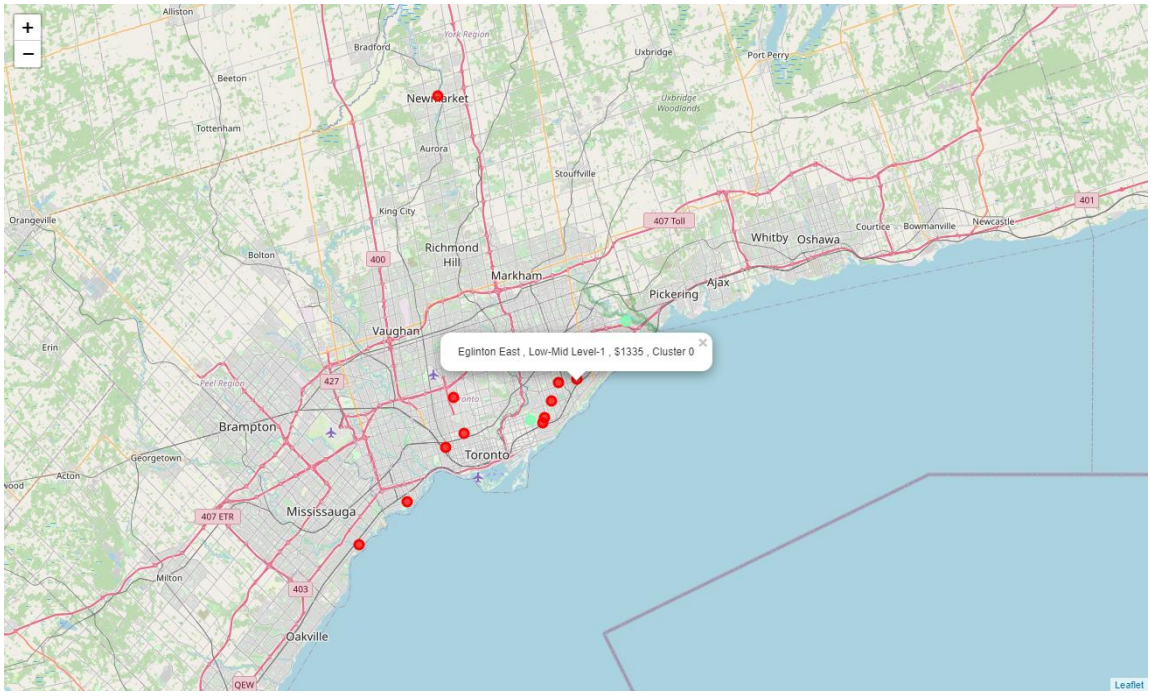
The graph of Elbow Method shows a turning point at K=2 where there is a significant drop on the distortion. The distortion continues to drop when K is getting bigger, however the idea of clustering is to group the neighbourhoods into few clusters, the fewer the better, therefore choosing the k at the turning point as the number of cluster, but this time I want to add one more cluster to increase a bit of diversity because there are 330 types of venues.

4.2 Solution to the Problem

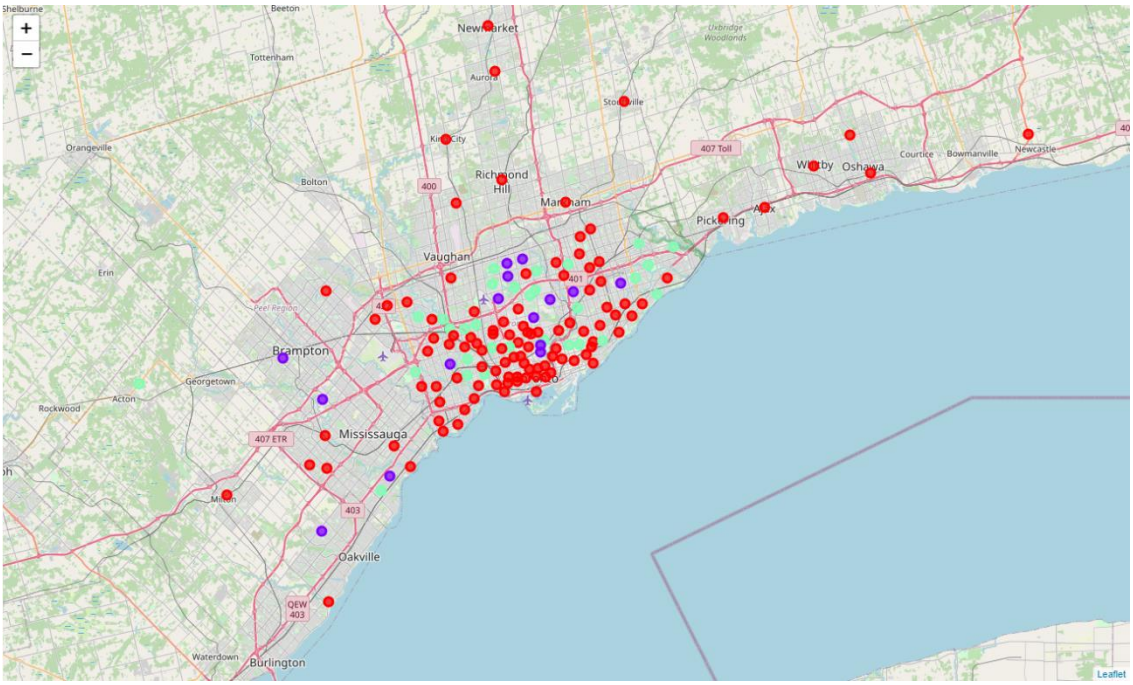
The main problem for locating the right places is the difficulty to find a match is both low price and venues sufficient on the overwhelming information provided. The K-Means Cluster Model tried to group all the neighbourhoods with similar venue categories together, which it created different lists of neighbourhoods sharing the same venue categories. It is extremely convenient that I don't need to go through the whole dataset, but simply check the low average price neighbourhoods in the cluster with most number of restaurant and public transit would find myself a desirable location for office setup.

5. Conclusions

In this study, I analyzed the relationship between average house rental price and venue category. I identified the best number of clusters for the venue category, and created model to select the desire locations to setup office at Greater Toronto Area. Furthermore, I have discovered more information from the final office map I created.



The map shows the desire locations to setup office with most of them to be the red cluster 0 and some few light green cluster 2. The cluster 0 is the Coffee Shop & Pizza Place Cluster where the top most and second most venues founded around the neighbourhoods are coffee shops and pizza places, and all of the cluster 0 are located at the center of the city. This tells that most of the desire or suitable office locations are at the city center and in reality, we can also notice a lot of companies, businesses setup their office in the city center which it explains a lot. Therefore, it is possibly true to say that most of the business owners valued more on the surrounding transit systems and eatery convenience, more than the office rental cost itself when setting up an office. Now let's see the map that shows all the locations of neighbourhoods in Greater Toronto Area.




```
cluster_0['1st Most Common Venue'].value_counts()
```

```
Coffee Shop              29
Pizza Place              10
Italian Restaurant        6
Bar                       6
Chinese Restaurant        5
Café                      5
Bank                      5
Pharmacy                  4
Grocery Store             4
Gas Station               3
Restaurant                3
Bakery                    3
Greek Restaurant          3
Fast Food Restaurant      2
Convenience Store         2
Indian Restaurant         2
Hotel                     2
Beach                     1
Storage Facility          1
Sandwich Place            1
Tennis Stadium            1
Sushi Restaurant          1
Soccer Field              1
Ice Cream Shop            1
Pet Store                 1
Mexican Restaurant        1
Brewery                   1
Pool Hall                 1
Intersection              1
Gym / Fitness Center      1
Temple                    1
Clothing Store            1
Name: 1st Most Common Venue, dtype: int64
```

By observing the red dots on map, we can notice that most of the neighbourhoods in cluster 0 are located around the center of city with many coffee shops, pizza places, banks, restaurant and many other types of venues. Therefore the final conclusion is that areas with many public transit, restaurant places are ideal to setup office, or in other words search the best of office locations at the center of city.

6. Future Directions

The K-Means Clustering model was able to classify the venue category groups with distortion around 13.5 and useful when selecting the best neighbourhoods' locations. The cons here is the overly generlized average rental pirce where there might be a big differences between the actual rental price. If the data of rental price of each single building can be obtained, it will bring significant improvements to the cluster model where the target location will be more precise, and the venue search range reduce closer to the target location's area to extract more detailed venue category. Furthermore, if a proper geojson file of Greater Toronto Area can be provided in the future, then the combination of these datasets will be useful at predicting the changes of rental prices and help real estate leasing services create business decisions.