Finding the Best Neighbourhoods in Toronto for Renting Apartments

1. Introduction

1.1. Background

Toronto's housing market is a growing rapidly, due to the high influx of population into the city. With the city being so large, there are definitely distinct neighbourhoods within Toronto. Some neighbourhoods are cleaner, some have all the best restaurants, and some are just not recommended to walk alone during the night. However, it is impossible for people that have not lived in Toronto before to know the subtle differences between each of the neighbourhoods. Thus, many people may end up in parts of the city that may not be suitable for them.

1.2. Problem

The goal of this project is to classify various neighbourhoods on their living characteristics. Specifically, the apartment building quality, proximity to amenities, and general neighbourhood safety will be considered

1.3. Interest

This data analysis will be of interest to newcomers trying to find a place to live in Toronto. Also, apartment rental companies may also use the analysis to determine what neighbourhoods are preferable for development.

2. Data

There are five data sources that were used in this project. First, the neighbourhood information with postal codes was obtained by web scraping a Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). The geospatial coordinates of each neighbourhood were obtained from a csv file based on the postal code. Second, the crime data in Toronto was obtained from the Toronto Police Service, with a list of crimes and their locations. Third, the list of apartments in Toronto has been created by the Apartment Building Standards and was obtained from the Toronto Open Data Portal (https://open.toronto.ca/dataset/apartment-building-evaluation/). The dataset also includes a *building evaluation score* that rated the safety and cleanliness of building grounds, entrance, exits and elevators, stairwells, and mechanical systems such as heating and ventilation of the apartment. Fourth, the name and location coordinates of subway stations in Toronto was obtained online. Finally, the venues around each apartment were obtained via Foursquare to evaluate the location of the apartment. A metric named *walking score* was created which measured the proximity of the apartment to nearby restaurants, grocery stores, and subway stations. The *walking score* and the *building evaluation score* will be both used to determine which neighbourhoods contain the best apartments.

3. Methodology

3.1. Neighbourhood segmentation

The Toronto neighbourhood data was first cleaned to remove any postal codes that did not have neighbourhoods assigned to them. The neighbourhoods that shared the same postal codes were grouped into one row and the geospatial data of the neighbourhoods were added via the latitude and longitude positions of the postal codes.

3.2. Neighbourhood crime score

The crime data of Toronto contained a list of all the crimes reported in Toronto for the past 4 years. For each crime, the coordinates of the crime location was matched to the closest neighbourhood, and the total number of crimes per neighbourhood was counted. The total number of crimes was 167525 crimes but only 10% of the crimes were taken into account to save computing time. The assumption was that 16752 crimes were enough to get a good distribution of the crime rate and location in Toronto. A crime score was created between 0 and 1 where 0 was the neighbourhood with the highest number of crimes and 1 was the neighbourhood with the lowest number of crimes.

3.3. Apartment feature selection

The apartment data required location extraction via geopy, which required lengthy computing time for the large dataset (3450 apartments). Also, limiting the apartment search to newer apartments may be favourable to those looking to rent decent quality apartments. However, to ensure the age of the apartments is not biased to specific neighbourhoods (i.e. all the old apartments are in one section of the city), the first hundred apartments were plotted on a map to qualitatively check the distribution of apartment age with respect to the neighbourhoods (Figure 1).

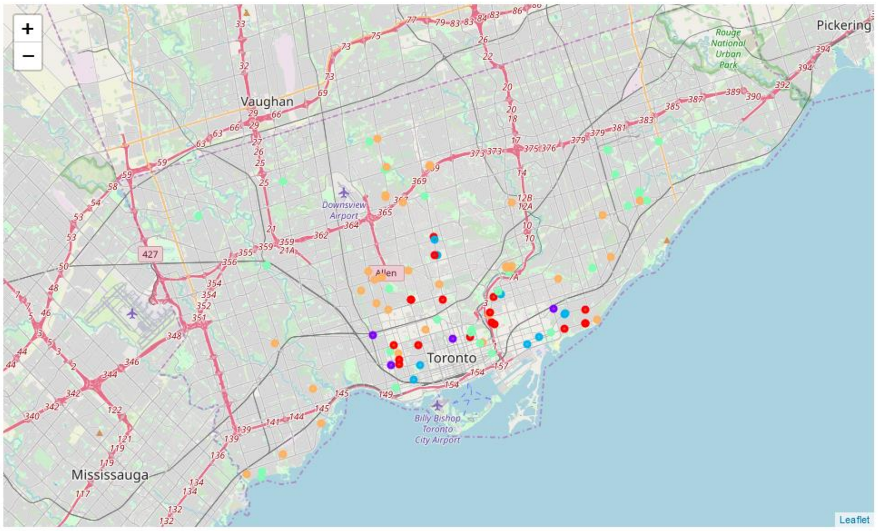

Figure 1 Apartment segmented by age

No clear distinction can be made about the age of the apartments and thus only the apartments built from the year 1975 and up were considered, limiting the number of apartments to 504. The coordinates for all the new apartments were searched and added to the dataset.

3.4. Foursquare venues

For each of the 504 apartments in the dataset, the closest 100 venues in a radius of 1 km was searched via the Foursquare API. The radius was chosen as an appropriate distance an adult is likely to walk (10-15 minutes). Of the 100 venues, the categories were compared with typical keywords restaurants will have (i.e. restaurant, bar, bakery etc.) to select just the restaurant type venues. A restaurant score was calculated as the following:

$$\mathit{restaurant\_score} = 1 - (\mathit{avg\_dist\_to\_restaurants}) / (\mathit{number\_of\_restaurants\_in\_radius}) / 0.2$$

Next, the categories were compared to keywords grocery stores will have (i.e. grocery, supermarket etc.) to select grocery stores. A grocery score was calculated as the following:

$$\text{grocery\_score} = 1 - (\text{dist\_to\_nearest\_store}) * 0.3$$

3.5. Transit score and walking score

The distance to the closest subway station was calculated for each apartment as well to get a transit score:

$$\mathit{transit\_score} = \mathit{1} - (\mathit{dist\_to\_nearest\_train\_station}) * 0.1$$

Finally, the transit score, restaurant score, and grocery score were all averaged to get a walking score. The apartments were grouped by the neighbourhoods they belonged in and added to the mean walking score and apartment evaluation score was added to the neighbourhood dataset.

3.6. K-means Clustering

The neighbourhoods were clustered using the mean apartment evaluation score, neighbourhood crime score, and mean apartment walking score. K-means clustering was used to group neighbourhoods that had similar characteristics. For example, one cluster may be safer (high crime score) but may not be suitable to travel by foot (low walking score). Depending on the needs of each individual, a cluster can be selected based on the characteristics, and apartments in those neighbourhoods are more likely to be suitable.

4. Results and Discussion

4.1. Optimal k for k-means clustering

When using k-means clustering, the value of k has a significant impact on the clustering outcome but the selection of k can be arbitrary. Here, the elbow method (Figure 2) is used to determine the range of k where the distortion (distance of point to center of cluster) decrease is minimal.

Between k values of 3-4, the distortion seems to start to plateau. For this case, a k of 4 will be used.

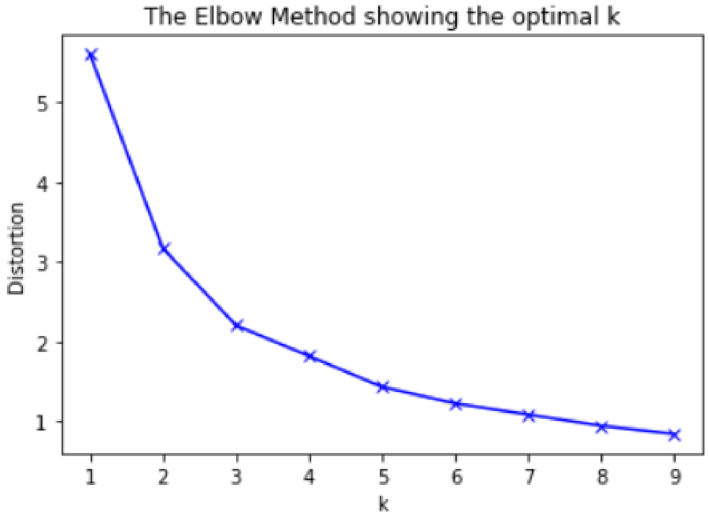

Figure 2 Elbow method to find the optimal k for k-means clustering

4.2. Clustering results

The neighbourhoods are clustered into four clusters and plotted on the map of Toronto (Figure 3). For each cluster, the mean apartment evaluation score, walking score, and crime score is also summarized in Table 1. The characteristics of the four clusters can be determined from the mean scores of each cluster. Most neighbourhoods in cluster 0 (purple) are located roughly near the center of Toronto. They have low apartment evaluation scores but the best walking score and an average crime score. Most neighbourhoods in cluster 1 (light blue) are located near the central north area. They have the best apartment evaluation score, an average walking score, and a high crime score. Most neighbourhoods in cluster 2 (yellow) are located on the west side of the city. The have the worst scores in all three categories. Most neighbourhoods in cluster 3 (red) are located on the east side of the city. They have average scores in all categories.

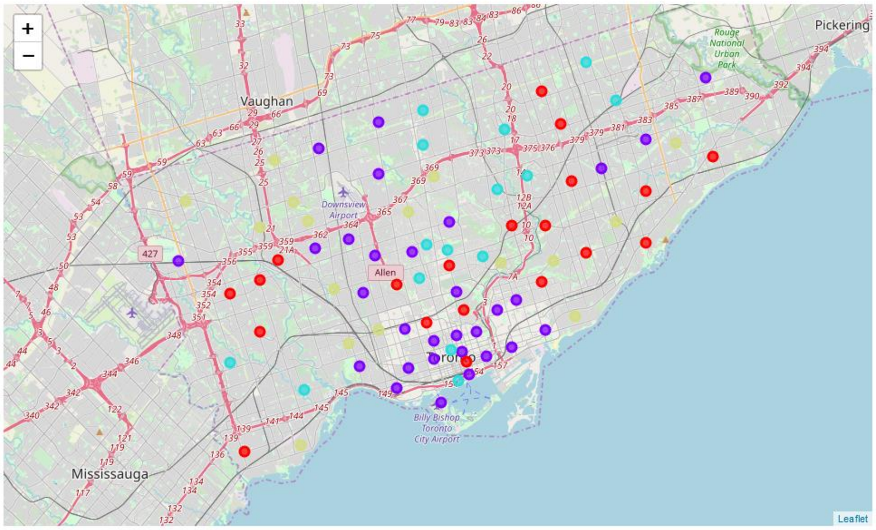

Figure 3 Neighbourhood clustered by characteristics

Table 1 Average neighbourhood scores for each cluster

Cluster	Apartment Evaluation Score	Walking Score	Crime Score
0 (purple)	0.763	0.801	0.604
1 (light blue)	0.900	0.749	0.746
2 (yellow)	0.700	0.721	0.566
3 (red)	0.821	0.745	0.650

Depending on the individual’s priorities, different neighbourhood clusters should be selected for renting apartments. Those who have a top priority in being close to necessities should select neighbourhoods in cluster 0. Those who have top priority in neighbourhood safety and building quality should select cluster 1. People looking to rent in cluster 2 should use caution to ensure that their apartments are in safe locations. Finally, cluster 3 is well rounded in all three categories and should be suitable to most people.

5. Conclusion

The neighbourhoods in Toronto have been clustered based on the apartment building quality, the apartment proximity to necessities, and the safety of the neighbourhood. Each cluster had its own characteristics, and newcomers to Toronto can select the cluster that will suit their need the most. Also, apartment rental companies can use this data to select certain neighbourhoods for apartment development, depending on their target renters as well. However, the model is still at a preliminary stage and needs further work. First of all, each of the scores calculated in this project may need some fine tuning. The sensitivity of changing the equations has not been comprehensively investigated yet. Also, adding crucial data such as average apartment rental costs and apartment vacancy is a required step before the results can be more useful.