Hailin Wang

Dec 28, 2020

Abstract

Evaluating the best neighbourhoods to build a sports complex in Toronto based on the venues using a k-means clustering Algorithm

Data Science Capstone project

Sport Complex in Toronto

## Introduction

Toronto is the largest cities in Canada. Downtown Toronto is urban and very densely populated with a variety of residential and commercial buildings. An estimated 2.83 million people live in Toronto as of 2020 and this number is rapidly increasing. As such there are plenty of people who eat, sleep, and navigate the city daily. One of the key necessities for more people is exercise and entertainment. In this scenario we are approached by a development company that wants to build a sports complex in one of the neighbourhoods in Toronto. They wish to find the optimal place to live to help minimize competition and maximize revenue. Therefore, our problem of choice will be trying to decide where we should build a new sports complex based on existing venues.

## Data

The main dataset we will be focused on is the foursquare dataset. From the foursquare dataset we will obtain location data and points of interests that are related to our gyms. Some of the most important factors will be derived from the following data in foursquare are as follows:

1. Neighbourhoods

Neighborhood data is key as this is how we will divide the candidate areas for the developments. As from the lab in week 3 the data of the neighbourhoods in Toronto was scraped from <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> The is read into a pandas data frame using the read\_html () method. This gives us the data in format that can be easily processed. Using the head method, we can view the data as follows.

Graphical user interface, text

Description automatically generated

Figure 1 Neighbourhoods in Toronto

1. Location Data from GeoPy

From the geocoder library and GeoPy we obtain data on the geographical position of the neighbourhoods. This allows us to make visualization with Folium and the locations are also important in determining their relative distance to other points of interest. The first few rows are shown below. Graphical user interface, text, application, email

Description automatically generated

Figure 2 - Neighbourhoods in Toronto with Longitude and Latitude

1. Foursquare Venue Data

Venue data from Foursquare will help to determine the types of amenities that already exist in the areas surrounding our sports complex. This data can help to predict the expected value of our condo location depending on which neighbourhood we decide to build in. The venue data has been extracted using the Foursquare API. This data contains venue recommendations for all neighborhoods in Toronto and is used to study the popular venues of different neighborhoods as well as build the unsupervised learning model to cluster neighborhoods. Sample data is below. From the venue we can see that there are distinct categories that can be either beneficial or a detriment to the value of our apartment development.

Table

Description automatically generated

Figure 3 – Venues in each Toronto Neighbourhood

## Methodology

### Exploratory Data Analysis

To gain a better understanding of our data we begin by creating a visualization of candidate neighbourhoods in Toronto. After dropping data that was unavailable from our dataset we matched the neighbourhoods with location data from GeoPy. Next we generated a map with folium depicting the various neighbourhoods denoted with markers.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 4 - Neighbourhoods in Toronto

To process our exploratory data, we transformed the foursquare data to a table that listed common venues in each neighborhood. This allows us to see both the categories of locations and how the data we input to our algorithm is organized. In our application, the focus is to reduce competition as such we can see that venues such as gyms, yoga studios and fitness centers are areas of interest. By performing this exploratory data analysis, we adjusted the sample weights in our algorithm for greater emphasis on these types of venues.

Table

Description automatically generated

Figure 5 - Most common venues in each Neighbourhood

Before running the algorithm, we represented each type of venue in one hot-code. That way we could use the skLearn library’s k-means function.

### Algorithm Tuning

K-means unsupervised learning technique was used to cluster the neighborhoods. We segmented our neighbourhoods based on the category of venues nearby. One important aspect of the k-means model is to determine the number of clusters to use in the model. A metric for the performance of k-means is the silhouette score. Thus, we performed the calculation of the silhouette score for a range of ‘k’ values as shown below.

Chart, line chart

Description automatically generated

Figure 6 - Silhouette scores

Clearly, we achieve the highest silhouette score when we segment our neighbourhoods into 4 clusters. Therefore, when we are building our model this is what we will set that parameter to.

## Results

From our K-means model we derived a label for each Toronto neighborhood representing the cluster it belongs to. Below is a summary of the first few rows of the data. In each row we have kept the information related to the most common venues to retain a general understanding of how the results relate to the original data.

Graphical user interface, diagram, application

Description automatically generated

Figure 7 - Neighbourhoods with cluster labelled

To visualize the results, we plot each type of neighbourhood in a different colour.

Map

Description automatically generated

Figure 8 - Map of Toronto with neighbourhoods clustered

## Discussion

By analyzing the four clusters obtained we can see that some of the clusters are more suited for a location for a new sports complex. Neighborhoods in clusters 1,2 contain a greater portion of their top ten venues relate to fitness and exercise (figure 9,10,11). This includes yoga studios, gyms, fitness centers and outdoor parks/sports fields. If we were to build a new sports complex in these areas, the surrounding competition would be greater.



Figure 9 - Cluster 2 labels

Table

Description automatically generated

Figure 10 - Cluster 3 labels

Table

Description automatically generated

Figure 11 - Cluster 1 Labels

In neighbourhood three however, the main venues are restaurants shops and neighborhoods in clusters 1 and 2 contain a much higher degree of restaurants, hotels, cafes, and bars (figure 11). Thus, these neighbourhoods would entail lesser competition when opening a sports complex. As such our recommendation is to scope out these neighbourhoods for developing the proposed sports complex.

Table

Description automatically generated

Figure 12 - Cluster 3 Labels

## Conclusion

In this project, we successfully determined the best neighborhoods in Toronto to build a sport complex. Based on the segmentation analysis carried out, neighborhoods in cluster 3 are optimal locations for the proposed project. This corresponds to the light blue markers in the results map. This has also been plotted in the map in section 5, discussion. The stakeholders and investors can further converge on a location by performing analysis on other factors like building codes and costs associated with a particular neighbourhood. These were out of the scope for this project and thus were not considered in this report.