Calgary Neighborhoods Comparison

1. Introduction

One important key to have a successful business is to analyze the location in terms of safety. When a location is safer, customers feel more comfortable and this brings more profit to a business. This study helps a new business to be started in a proper location by analyzing Calgary venues and crime count in different neighborhoods.

Unsupervised learning is a Machine Learning Algorithm which is used to find patterns and similarity between samples in data. K-Means is an Unsupervised learning algorithm that helps finding patterns in Convex Sets of data.

Foursquare is a social networking service that allows people to find attractions, leaving reviews and connect with friends. In this study, Foursquare is used to extract venues in two neighborhoods in Calgary, Canada and the aim is to compare both venues and committed crime in these two neighborhoods.

2. Data

In this study, Calgary Crime Data was downloaded from https://data.calgary.ca. The data was filtered and downtown and east neighborhoods and their respective location in longitude and latitude coordinates were taking into account. After having longitude and latitude coordinates of neighborhoods, the Foursquare API is used to extract venues in both locations. The API sent a GET request to acquire the venues that are within a radius of 500m. The venues then were formatted using one hot encoding and were grouped by neighborhoods computing the mean of each feature. The frequency of the categories found in the neighborhoods was used to find the similarities. The similarities were an important element to help a user to choose a location for a new business.

3. Methodology

As it mentioned before, One Hot Encoding is used for feature extraction of categories. To indicate each feature as a category that belongs to a venue, each feature becomes binary. In other words, 1 shows that the category is found in the venue and 0 means that the category is not found in the venue. In the next step, all the venues are grouped by the neighborhoods and the mean which results in calculating the frequency of occurrence of that particular category. In order to find similarities between neighborhoods, a clustering algorithm, specifically K-Means was implemented. K-Means is a clustering algorithm that searches clusters within the data to minimize the data dispersion for each cluster. To find a proper the number of clusters the elbow method was implemented and used. After indicating the number of clusters, each cluster was analyzed separately.

3. Results

After clustering data, the result may help people to start a business in a location. Commited crimes in a particular location has an important effect in terms of safety for a new business to be started. For instance, in Calgary, downtown is a active business location but on the other hand it is unsafe in terms of crimes. Cluster one in downtown has multiple active businesses including pharmacy, restaurant, gas station, school, supermarket, coffee shop, baseball stadium, pool, bank and etc. The most crime has been committed in this area with the number of 260679 out of 268014 since 2012 which is huge and it might be helpful to have it considered. On the contrast, the active business location of the east side of Calgary is cluster two, with 57717 number of crimes out of 67704 since 2012.