Coursera Applied Data Science Capstone Project

Exploring the opportunity for opening a new recreational center in Calgary, Alberta

By: Praise Okwa

May 2020

Introduction

For many cities like Calgary, recreational centers provide an opportunity for active living and recreation in a safe, inclusive environment. For cities they are incredibly important for a healthy, vibrant community. By creating a positive atmosphere, these facilities become essential to personal health and wellness, thereby reducing reliance on healthcare and other costly social services. This in turn boosts the local economy and can also help contribute to overall economic development. For the city of Calgary, planning a development project like a recreational center, therefore requires details analysis in order to determine the success or failure of any proposed projects.

Business Problem

My objective for this project is to analyse the city of Calgary using data science and propose the optimum location for a new recreational center. By using Segmentation and Clustering techniques, my aim is to provide insights to a city council/property developer considering the feasibility of such a development.

Target audience for this project

The target audience for this project are developers or Calgary city officials exploring the idea of building a recreational center and along with other analyses are interested in looking at data led analysis of possible locations to house such a facility.

Data

In order to solve this problem, I will be making use of datasets available online. These datasets include:

- A list of all neighbourhoods within the City of Calgary including residential communities, industrial areas, major parks and residual areas by electoral ward.
- Datasets that show names and addresses for current recreation facilities, including amenities available at each location.
- Longitudinal and Latitudinal data for locations of neighbourhoods as well as current recreational facilities, as this will aid in collecting more data relating to those locations as well as allow for plotting a map.

Wikipedia provides a list of the neighbourhoods in Calgary along with information such as the type, population and dwellings for each neighbourhood. Web scraping will be used to collect this data and with LXML and beautifulsoup packages I will organize this data into a Pandas data frame. Python Geocoder will then be used to assign longitude and latitude coordinates to the neighbourhoods.

Once this process is completed the Foursquare API and datasets from the City of Calgary will then be used to explore the recreational venues in the city. Using machine learning; K-means clustering, and by visualizing the results on a map using Folium, I will be able to make a proposal to meet my objectives.

Methodology

Firstly, we need to get the list of neighbourhoods in Calgary and data on current recreational centers in Calgary. Fortunately, the list is of neighbourhoods in Calgary can be got from Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary) and data regarding recreational facilities from Calgary data page (https://data.calgary.ca/Recreation-and-Culture/Recreation-Facilities/hxfu-6d96/data). Web scraping using Python requests and LXML packages will then be used to extract the list of neighborhoods data. Once a list of neighborhoods are placed in the Pandas frame, we need to get the geographical coordinates in the form of latitude and longitude coordinates so we can use the Foursquare API.

The Geocoder package will allow us to convert the address or neighborhood name into geographical coordinates, latitude and longitude. After gathering the new data in a Pandas frame, we can visualize the neighbourhoods in a map using Folium.

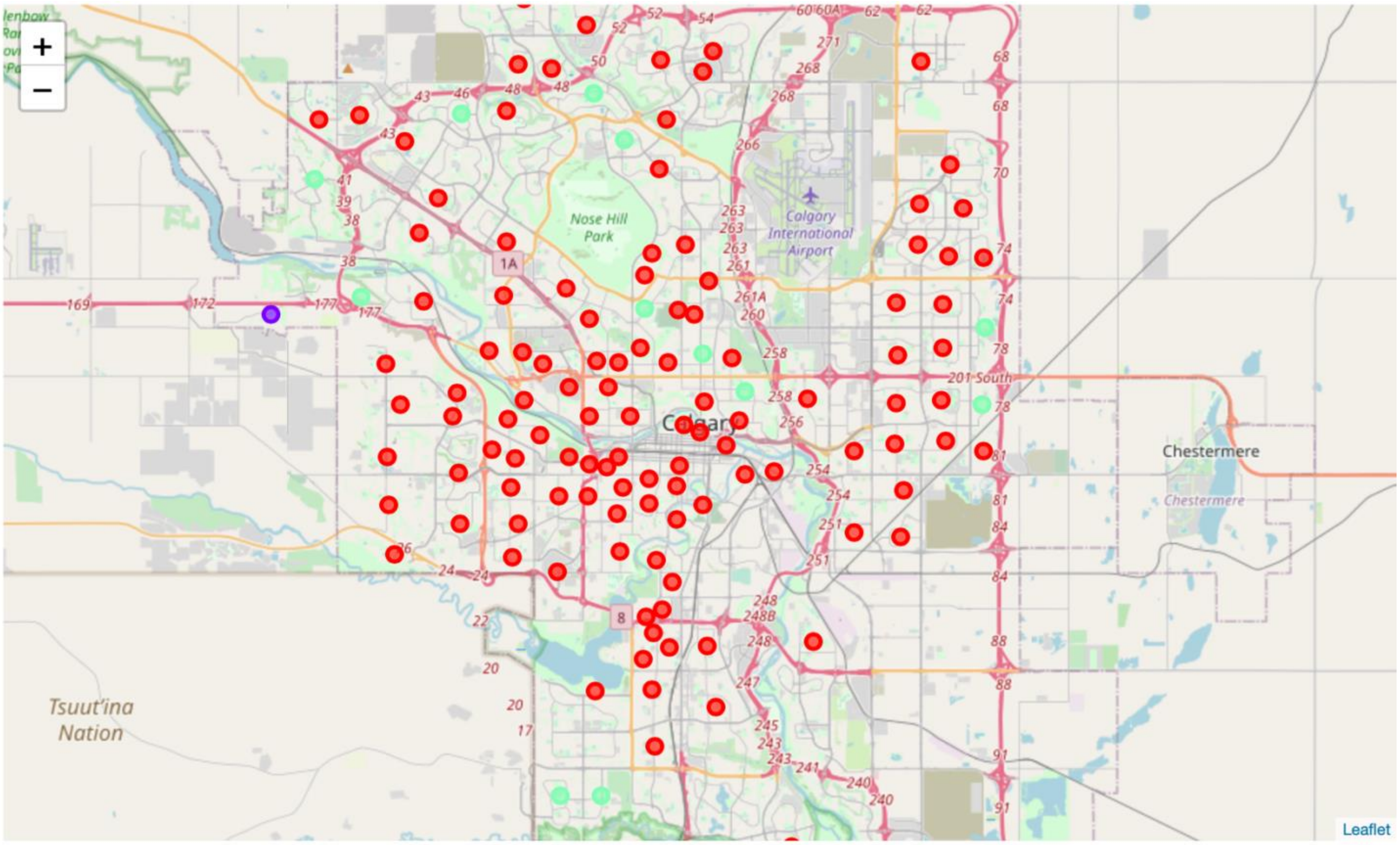
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We make API calls to Foursquare passing the geographical coordinates of the neighbourhoods with a Python loop and Foursquare will return the venue data in JSON format from which we extract the venue name, category, latitude and longitude. With this data, we can check how many venues were returned for each neighbourhood and examine how many unique categories that can be curated from all the returned venues. Each neighbourhood is then analysed further by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing this, we prepare the data for clustering. We are particularly interested in “recreational facilities and so we filter the Recreation centers as the venue category for the neighbourhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for recreation centers. The results will allow us to identify which neighbourhoods have higher concentration of recreation centers and with some exploratory analysis determine, based on the occurrence of recreation centers in different neighbourhoods, the most suitable to open new center.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for recreation centers:

- Most of the Recreation centers are located centrally in Calgary with the vast majority in Cluster 0.
- The location in cluster 1 is isolated on the outskirts of the city and can be neglected from consideration
- Cluster 2 offers the most promising location for a new center. There are a limited number of centers, especially in the NW Quadrant



Discussion

As can be seen from the map in results section, most of the recreation centers are concentrated in the central area of Calgary, with the highest number in cluster 0 and an outlier in cluster 1 On the other hand, cluster 2 has a relatively low number or recreation centers near the neighbourhoods and this represents a great opportunity and high potential area for opening a new recreational facility. Meanwhile, a facility in cluster 0 are likely to suffer from oversaturation from nearby residents.

From another perspective, the results also show that the oversupply of recreation centers mostly occurred in the central area of the city, with the outskirts of the city having very few facilities. Therefore, this project recommends property developers capitalize on these findings to open new facilities in neighbourhoods in cluster 2 with little to no competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 0 which already have high concentration of recreation centers and are most likely suffering from over saturation.

Limitations and Suggestions for Future Research

In this project, we only consider the frequency of occurrence of recreational facilities, however there are other factors such as population and income of residents that could influence the location decision of a new recreation center. As more data on these relationships becomes available more detailed research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations however for the purpose of this project, the analysis was limited. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. A more detailed research approach could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new recreation center.

To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to open a new recreational center. These findings will assist stakeholders in making a decision to capitalize on opportunities and high potential locations while avoiding overcrowded areas in their decisions to open a facility

References

Category: List of Neighborhoods in Calgary. *Wikipedia*. Retrieved from (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary

Foursquare Developers Documentation. *Foursquare*. Retrieved from https://developer.foursquare.com/docs

(Recreation - Facilities, 2020) (https://data.calgary.ca/Recreation-and-Culture/Recreation-Facilities/hxfu-6d96/data