Capstone Project

Opening a New Pub in Richmond Hill, Ontario

August 8, 2020

Introduction

The history of pubs can be traced to Roman taverns in Britain in 43 AD ("Historic UK", n.d., para 3). It means pubs have been existed for almost 2,000 years. What reasons that make pubs last so long? Tribute Cornish Pale Ale conducted a poll about reasons for visiting a pub, and the results showed that catching up with friends, the atmospheres and the opportunity to have a few drinks were the top three reasons ("Beertoday", 2017). For many peoples, going to a pub is a way to relax and enjoy themselves after work, at weekends or during holidays. On the other hand, the profitability of pubs has strong correlation with sports events ("Financial Time", n.d.). For example, large number of residents in the neighbourhoods is attracted to nearby pubs on game nights. There are hundreds of pubs in Richmond hill. Therefore, selecting location of a pub is crucial for the pub owner's investment decision.

Business problem

The main purpose of this project is to find an optimal location for a new pub in Richmond hill, Ontario by using data science methodology, specifically machine learning techniques.

Who interested in this project

This project is particularly useful to potential pub owners and investors looking to open or invest in new pubs in Richmond hill, Ontario.

Data Section

- ***The following data is what we need for this project:***

  1. List of neighbourhoods including postal code in Richmond hill, Ontario.
  2. Latitude and longitude coordinates of the neighbourhoods.
  3. venue data, particularly data related to pubs on the neighbourhoods.

- ***Sources of the data and methods for manipulating***

  Richmond hill is a city in south-central York Region, Ontario, Canada. The city contains four major neighbourhoods: Richmond hill (Southeast), Richmond hill (Southwest), Richmond hill (Oak Rideges/Lake Wilcox/Temperaneceville) and Richmond hill (Central). The webpage https://en.wikipedia.org/wiki/Richmond_Hill,_Ontario#Communities includes a list of neighbourhoods in Richmond hill.

  Since the number of neighbourhoods is small, so we can just build a dataframe using pandas of Python data analysis library. Then, we will get the geographical coordinates of the neighbourhoods using Python Geocoder package. Next, we will use Foursquare API to obtain the venue data for the neighbourhoods. Foursquare API will provide a lot of categories of the venue data. However, we are interested in the pub category data which can help us to solve the business problem as mentioned above. Finally, we will clean the dataset, apply machine learning skill K- means clustering and map visualization by using Folium package.

Methodology

Step 1: we need to get the list of neighbourhoods in the city of Richmond hill. The information is available in the webpage https://en.wikipedia.org/wiki/Richmond_Hill,_Ontario#Communities .

Step 2: we make a dataframe using pandas of Python data analysis library to contain neighbourhood's data.

Step 3: we need to know the geographical coordinates in the form of latitude and longitude by using Geocoder package. Then, we put the data into the dataframe and visualize the neighbourhoods in a map by using Folium package.

Step 4: we use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. Foursquare will return the venue data in JSON format. After extracting the venue name, category, latitude and longitude, we can check how many venues were returned for each neighbourhood and examine how many unique categories.

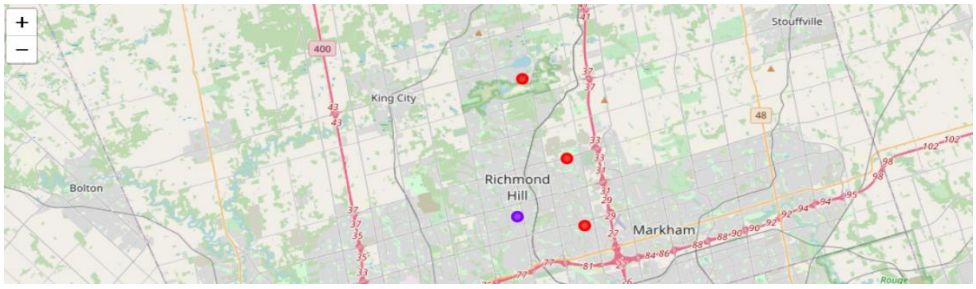
Step 5: we analyse each neighbourhood by grouping Postal code and taking the means of the frequency of occurrence of each venue category. Then, we filter the pub category for the neighbourhoods.

Step 6: we perform K-means clustering. We cluster the neighbourhoods into clusters based on their frequency of occurrence for pub category.

Results

The results from the K-means clustering show that we have 2 clusters: neighbourhoods without pubs (red) and neighbourhoods with pubs (purple).

The following map can clearly show that there are 2 clusters.



Discussion

Based on what has been presented in the results section, all the pubs are concentrated in the southwest of Richmond hill according to the dataset we obtained. The reason may be that the downtown of Richmond hill is in southwest of the city along Yonge street.

Therefore, there is a great opportunity to open new pubs in the cluster which no pubs in, such as southeast, central, Oak Rideges/ Lake Wilcox / Temperaneceville areas of the city.

Conclusion

According to what we analyzed using machine learning techniques, we find that we could choose a location in the southeast, central, Oak Rideges/ Lake

Wilcox / Temperaneceville areas of Richmond hill. However, there are some limitations in the project.

We only have four major neighbourhoods. Each neighbourhood has a large area in the city, so it is not a best way to find an optimal location for a pub. If we have enough data sources, it is better to use small communities as our data frame. Also, there are many factors need to be considered for finding an optimal pub location, such as population, incomes, age, education and occupation. However, we only considered one factor which is frequency of occurrence of pubs. Therefore, for further better research, we would better to build a model that contains more correlated factors to obtain a more accurate result.