Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Shopping Mall in Toronto

Gabriele Mineo

May 2020

Introduction

As today, shopping mall are plenty of shoppers, grocery shopping, restaurants and cinemas, etc.
A shopping mall could be a really interesting investment for real estate investor, but it will be really important the chosen location. This criterium will determine whether the mall will be a success or a failure.
Main objectives of this project were to define a business problem, look for data in the web and, use Foursquare location data to compare different neighborhoods of Toronto to figure out which neighborhood is suitable for starting a new shopping mall. In this project, we will go through all the process in a step by step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their decisions.

1. Description of the c & Discussion of the Background:

Problem Statement

In this project we will go through step by step process to make a decision whether it is a good idea to open a shopping mall. We analyze the neighborhoods in Toronto to identify the most profitable area since the success of the shopping mall is the uniqueness in surroundings.

Target Audience

Who will be more interested in this project? What type of clients or a group of people would be benefitted?

1. Business personnel who wants to invest or open a shopping mall in Toronto. This analysis will be a comprehensive guide to start shopping mall.
2. Real estate investors who desire to invest in a property to build a shopping mall
3. Business Analyst or Data Scientists, who wish to analyze the neighborhoods of Toronto using Exploratory Data Analysis and other statistical & machine learning techniques to obtain all the necessary data, perform some operations on it and, finally be able to tell a story out of it.

2. Data acquisition and cleaning

2.1 Data Sources

a) Use of "List of Postal code of Canada: M" (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) wiki page to get all the information about the neighborhoods present in Toronto. This page has the postal code, borough & the name of all the neighborhoods present in Toronto.
b) Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
c) To get location and other information about various venues in Toronto it has been used Foursquare's explore API. Using the Foursquare's explore API, it has been fetched details about the venues up present in Toronto and collected their names, categories and locations (latitude and longitude).
From Foursquare API (https://developer.foursquare.com/docs), it has been retrieved the following for each venue:

- Name: The name of the venue.
- Category: The category type as defined by the API.
- Latitude: The latitude value of the venue.
- Longitude: The longitude value of the venue.

2.2 Data Cleaning

a) Scraping Toronto Neighborhoods Table from Wikipedia
Scraped the following Wikipedia page, "*List of Postal code of Canada: M*" in order to obtain the data about the Toronto & the Neighborhoods in it.

Assumptions made to attain a DataFrame:

- Dataframe will consist of three columns: PostalCode, Borough, and Neighborhood
- Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.
- More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma.
- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

b) Adding geographical coordinates to the neighborhoods
Next important step was to add the geographical coordinates to these neighborhoods and combined with the existing neighborhood dataframe by merging them both based on the postal code

3. Exploratory Data Analysis:

3.1 Folium Library and Leaflet Map

Folium is a python library, it has been used to draw an interactive leaflet map using coordinate data.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the "Shopping Mall" data, we will filter the "Shopping Mall" as venue category for the neighborhoods.

Lastly, it has been performed the 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. The neighborhoods have been clustered into 3 clusters based on their frequency of occurrence for "Shopping Mall". The results allow to identify which neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results

We have made use of Foursquare API to explore the venues in neighborhoods of Toronto, then get good amount of data from Wikipedia which we scraped with help of Wikipedia python library and visualized The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Shopping Mall”:

- Cluster 0: Neighborhoods with low number of shopping malls
- Cluster 1: Neighborhoods with high number of shopping malls
- Cluster 2: Neighborhoods with moderate concentration of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.

Discussion

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of Toronto, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few shopping malls. Therefore, this project recommends real estate investors to capitalize on these findings to open new shopping malls in neighborhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 0 with moderate competition. Lastly, real estate investors are advised to avoid neighborhoods in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, it was considered one factor i.e. frequency of occurrence of shopping malls, there are other factors such as population density and income of residents that could influence the location decision of a new shopping mall. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall.

Conclusion

In this project, it has been conducted 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. real estate investors regarding the best locations to build a new shopping mall. The neighborhoods in cluster 0 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.