Vancouver Grocery

1. Introduction

1.1. Background

From grocery shops customers buy everyday necessities for them and their families.

There's a wide range of products which for each there is a wide range of qualities and as a consequence price.

The task of finding a perfect location for a new retail shop is a very demanding one, let alone for a grocery shop.

There are several methods of finding the perfect spot, starting from the more traditional such as wandering through the roads of some neighborhoods which have been picked up either by chance or based upon some advice of some expert

More advanced method is gathering and studying publicly available data about the neighborhoods in the city of interest. Population, neighborhood demographics, visibility, the amount of traffic that goes by and local competition are all factors taken into consideration This data will be used to study each of the following 3 aspects about what makes a location good for a retail shop (2).

1.2. Problem Description

The board members of our company, FreshGems, have assigned our team to research the best location for our new high-end grocery shop in Vancouver city, Canada. This new shop would help us target customers looking for high quality groceries, harvested according to the latest bio standards.

This is a new hype with many potential customers as people are increasingly aware of the benefits of proper healthy nutrition with bio ingredients. Based on our company's ethics, healthy nutrition should be within the reach of all levels of society, thus the result of our work should be a location where there is a fairly high concentration of retail shops that target the lower half of the market.

2. Solution approach and data mining

2.1. Data Requirements

From the 4 aspects that we'll study we derive the following sets of data that will be used during the study:

- Customers: popularity of grocery shops in contrast to supermarkets
- Find out which neighborhoods are high exposure ones
- Popularity of each grocery shop per neighborhood
- Combining all of the above to make a suggestion

2.1.1. Customers

We'll gather data using Foursquare API about the popularity of the grocery shops and the supermarkets in each of the neighborhoods. Based upon the coordinates of each neighborhood, a "SEARCH" request is sent to the foursquare API with radius set to 500 m and a list of categories of shops that deal with food and beverages

2.1.2. High exposure streets / neighborhoods

We'll use the K-Means library to cluster the popularity of the neighborhoods based on the popularity of the existing shops

2.1.3. Popularity of groceries per neighborhood

To be able to determine centers of competition we'll study the concentration and popularity of grocery shops within 500m radius. From the data gathered in part 1 and for each of the acquired venues, a "VENUE" request is sent to the foursquare API from which the venues' details are returned along with it rating

3. Methodology

Assumptions

Due to lack of recourses the data of the city of Toronto is used to conduct the study, the "SEARCH" requests of each neighborhood are merged into one dataframe which is then exported to a csv file to conserve foursquare API request quota.
Due to the big number of venues recorded and the fact that foursquare quota is not enough to get each venue's rating, a random 2 decimal number between 0 and 10 is assigned to each.

Steps of the analysis

First the required libraries are imported that will be used throughout our data processing session.
Next the list of Toronto's postal codes is scrapped from Wikipedia and loaded into a dataframe.
Next the coordinates of Toronto city are acquired, and the neighborhoods are plotted on the map to visualize the area being studied.
At this point venues' data is requested from the foursquare APIs for each neighborhood and for each venue a rating is assigned.
One hot encoding is used to remove the numerical bias that's added by the count of similar venues in each neighborhood.
Then the data is grouped by neighborhood summing over the number of similar venue categories, aggregating the count of venues in each category.
As a control measure the top 5 venues of each neighborhood are displayed.
The mean value of the ratings per venue type for each neighborhood is calculated.
With the aid of a helper function a dataframe with the top 10 venues of each neighborhood.
K-means clustering is run on the dataset that would cluster the venues into 5 groups based on each group's rating.
On a map the clusters are plotted each in a separate color to visualize them.