Coursera Capstone

IBM Applied Data Science Capstone

By Shreyak Vashisht

2020

Introduction

Supermarkets are the most important aspect of food production and distribution because they are the interface between supply and demand. There are alternatives such as farmers markets, but not in reach. So, when a supermarket such as WalMart decides to offer milk only produced without certain hormones, the entire milk production industry begins to change practices. A farmer's competitive advantage is no longer "produces more milk", it is "produces rBST free milk". In a similar vein, supermarkets nominally compete with farmers markets. Shop at farmers markets and it will make supermarkets begin to seek out and offer the organic foods you can only find at the farmer stand.

We've all heard the familiar aphorism "location location location" for emphasising the importance of this factor in real estate, but how do we determine the most Important Location?

Business Problem

I've learned is that location is indeed the single most important factor but its relative importance is determined by the type of store.

Grocery shopping in particular is an activity where the consumer behaves quite logically in both convenience shopping and in supermarket (destination) shopping. Mr Shopper will go to his nearest convenience store or possibly second nearest (if it's much nicer) but won't generally pass four or five to get to one that's "just right". For supermarkets, he will travel further but again will most likely go to one of his four or five closest supermarkets, his patronage being determined by the trade-off between the pain of travel and the rewards of store attractiveness

By using spatial analysis methods such as Clustering, , we can assess the importance of location and the results are pretty clear. Along with value store location in grocery shopping is THE most important factor in determining the success of a grocery store.
The Problem we are trying to solve is the Optimum Location of a Supermarket

Target Audience of this project

Target Audience of this Project is investors who are looking to open or invest in the Capital city of California, San Francisco this project is aimed at giving the investors a perfect location to open a Supermarket considering Multiple Factors

Data

To solve the problem, we will need the following data:

- List of neighbourhoods in San Francisco. Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco) contains a list of neighbourhoods in San Francisco.

We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.