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.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of San Francisco . Fortunately, the list is available in the Wikipedia page ((https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco).

We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of San Francisco.

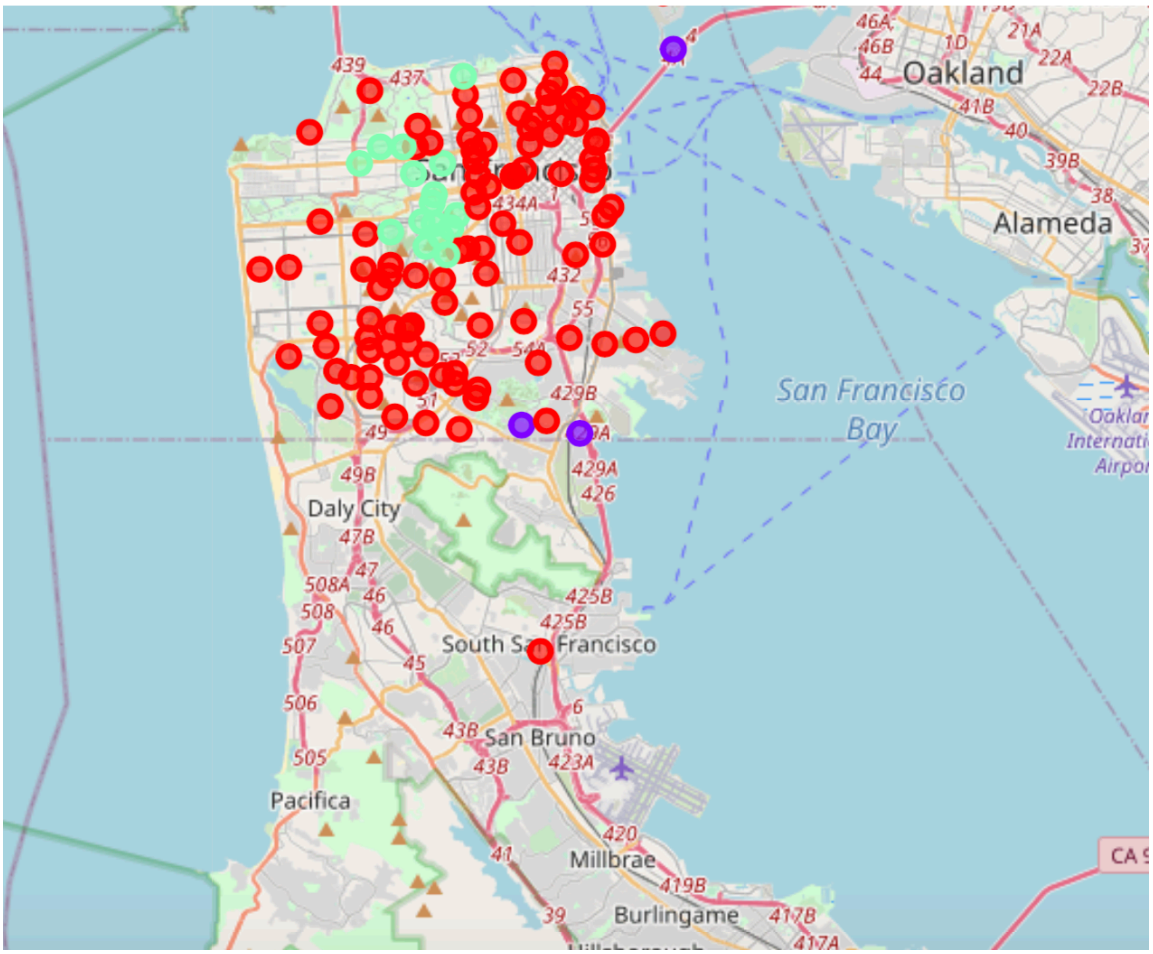
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods 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 neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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

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 “Super Market” and ‘Bus Station’. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence

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 Super Market" and 'Bus Station'.

- Cluster 1: Neighbourhoods with low number to no existence of Super Market and no Bus stations
- Cluster 2: Neighbourhoods with low concentration of Super Market and Prescence of Bus stations(ideal)
- Cluster 0: Neighbourhoods with moderate number of shopping malls and low concentration of bus stations



Conclusion/Insights

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

The whole idea behind this Project is to find the top Location(s) for a Supermarket
we Tried to Cluster neighborhoods based on the occurance of Supermarket and proximity to Bus station

Cluster1,did not have any Supermarket nor had any Bus stations,so ideally not our top Choice but can still be considered because competition will be less

Cluster2,is our **top choice** becasue it had comparitively more Bus Station and no Supermarkets which means opening a supermarket here, will face no competition and easy access to Transportation resulting in higher Foot fall

Cluster3,is not our choice becasue of the prescence of other Supermarkets ie higher competition

The top locality are: (Cluster2)
1.Little Hollywood
2 Sunnydale
3 Yerba Buena