IBM data science capstone project

Vivek Mankar
vivek.mankar18@vit.edu

Opening a new supermarket in Pune, Maharashtra

April 29, 2020

Overview

A **supermarket** is a self-service shop offering a wide variety of food, beverages, and household products, organized into sections. It is larger and has a wider selection than earlier grocery stores, but is smaller and more limited in the range of merchandise than a hypermarket or big-box market.

There has been a rapid transformation of the food sector in developing countries like India. With growth, has comes considerable competition and some amount of consolidation. The growth has been driven by increasing affluence and the rise of a middle class; the entry of women into the workforce; with a consequent incentive to seek out easy-to-prepare foods; the growth in the use of refrigerators, making it possible to shop weekly instead of daily; and the growth in car ownership, facilitating journeys to distant stores and purchases of large quantities of goods. Due to all of these, the need for supermarkets will increase over the period of time.

Goal

The objective of the project is to analyze and select the best location in the city to open a new supermarket. Using data science and machine learning techniques like clustering, this project aims to solve the following business Question:

Que: In the city of Pune if a company wants to open a new supermarket where would you recommend that they open it?

Target audience

The project is particularly useful to the property developers and investors looking to open or invest in new supermarkets. With the increase in no of middle-class people in India with women working in the family, there would be consequent incentive to seek out easy-to-prepare foods. Having a supermarket nearby is the best possible solution to cope with the changing lifestyle.

Data

To solve this problem we will need the following data :

- List of neighborhoods in Pune.
- Latitude and Longitudinal coordinates of those neighborhoods.
- Venue data for each neighborhood.

Sources

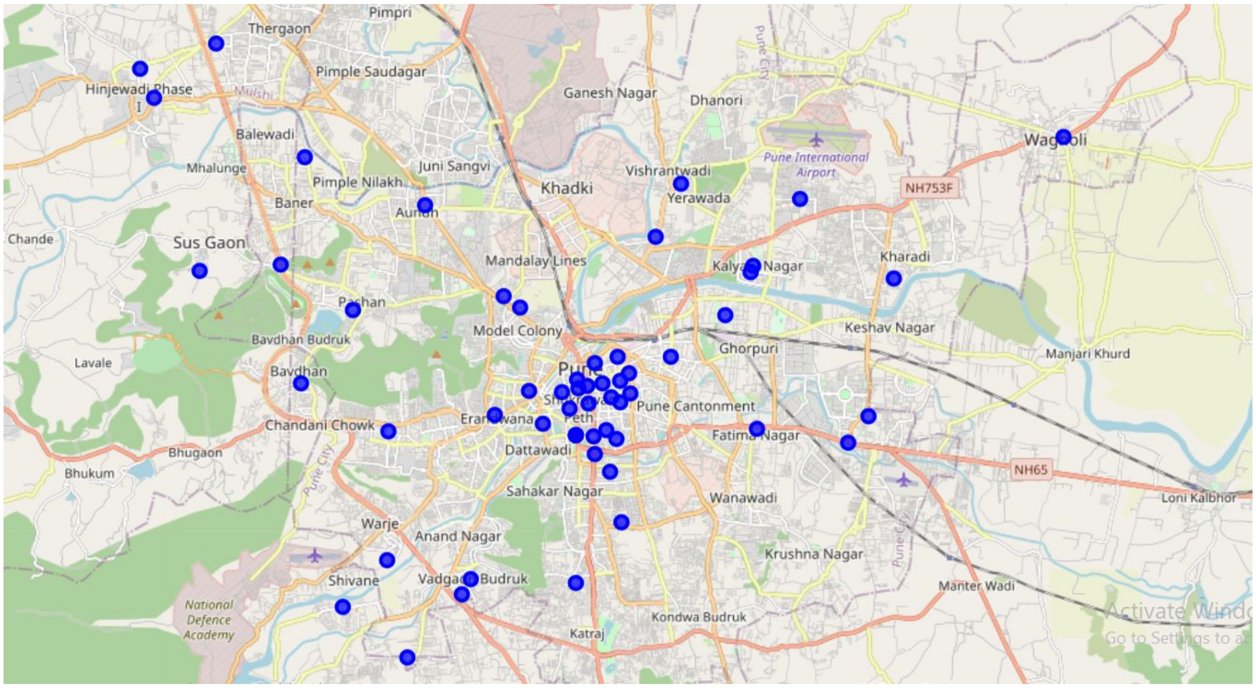
- For the list of neighborhoods, I used (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Pune)
- For Latitude and Longitudinal coordinates: Python Geocoder Package (https://geocoder.readthedocs.io/)
- For Venue data: Foursquare API (https://foursquare.com/)

Methods to extract data from Sources

To extract the data we will use python packages like requests, beautifulsoup, and geocoder.

We will use Requests and beautifulsoup packages for web scraping(https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Pune) to get the list of neighborhoods in Pune and geocoder package to get the latitude and longitude coordinates of each neighborhood.

Then we will use Folium to plot these neighborhoods on the map. The output is something like this.



After that, we will use the foursquare API to get the venue data of those neighborhoods. Foursquare is a social location service that allows users to explore the world around them. Foursquare has one of the largest databases of 105+ million places and it is used by over 125,000 developers. Foursquare API will provide many categories of the venue data we are particularly interested in the supermarket category in order to help us to solve the business problem put forward.

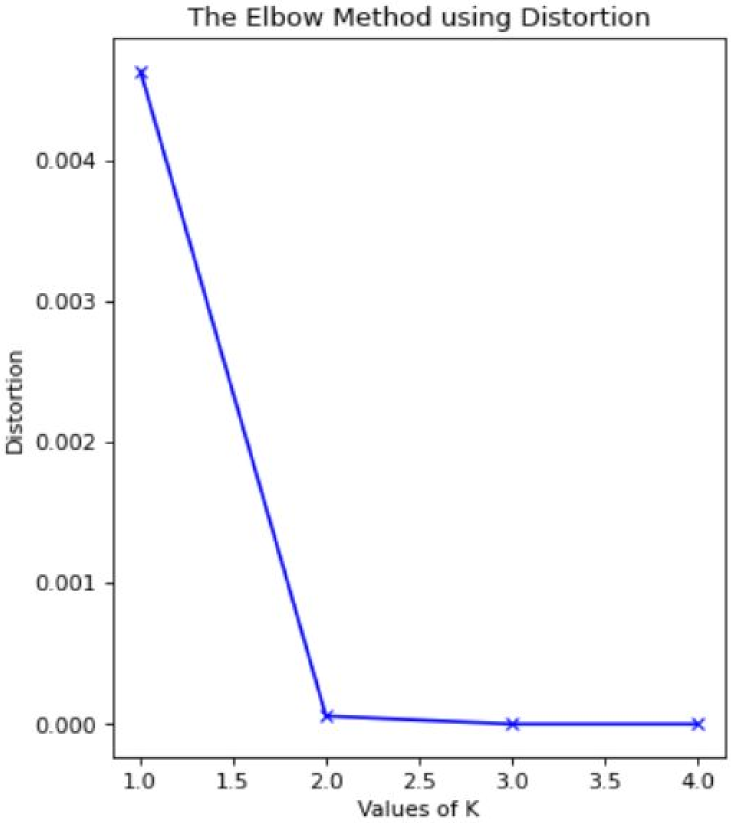
Methodology

The data science skills/techniques used in this project are :

- web scraping(Wikipedia)
- working with API(Foursquare)
- data cleaning
- data wrangling
- machine learning(K-Means Clustering)
- map visualization(Folium)

Firstly we need to get the list of neighborhoods in the city of Pune and the list is available on the Wikipedia page(https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Pune). We will do web scraping using Python requests and beautiful soup package to extract the list of neighborhood data. Then we will use the Geocoder package that will allow us to convert the address into geographical coordinates in the form of latitude and longitude. These data can be used for four square API.

After getting the data we will populate the data into a pandas data frame and then visualize the neighborhoods in a map using folium package. This will allow us to perform a center check to make sure that the geographical coordinates data returned by geocoder are correctly plotted in the city of Pune. Next, we will use the Foursquare API to get the top 150 values that are within the radius of 3000. Four square will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude, and longitude. Then we will analyze the data by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. Then we will plot a 3D plot by using matplotlib with x and y co-ordinate as latitude and longitude respectively and the Z coordinate as the mean of frequency of occurrence of supermarket category.



Then we will use the elbow method to check the suitable value of k for our k mean clustering. From the graph, it can be seen that k = 2 is the most suitable value because it is the. When the curve starts to flatten out. Lastly will perform clustering on the data by using k-means clustering.

k means clustering algorithm identifies k number of centroid 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 neighborhood into two clusters based on the frequency of occurrence for the supermarket. The results will allow us to identify which neighborhoods which have a higher concentration of supermarkets while which neighborhoods have a lower concentration of supermarkets.

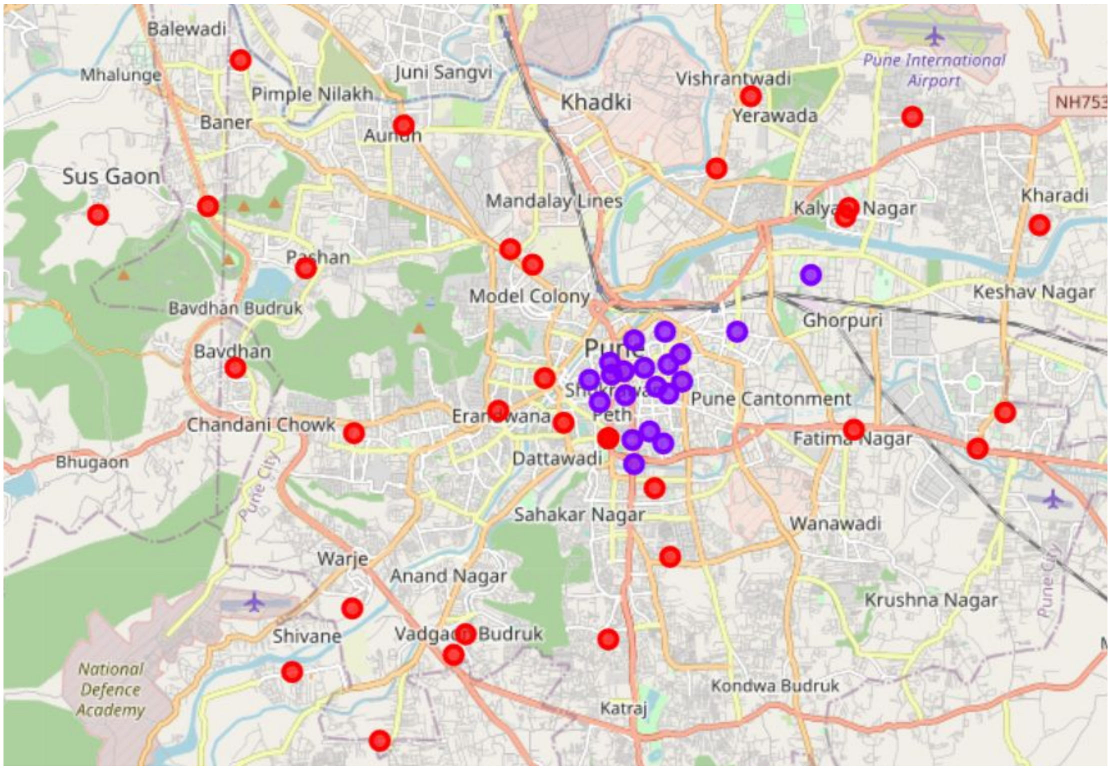
Results

The results from the k-means clustering show that we can categorize the neighborhoods into two clusters based on the frequency of occurrence for a supermarket.

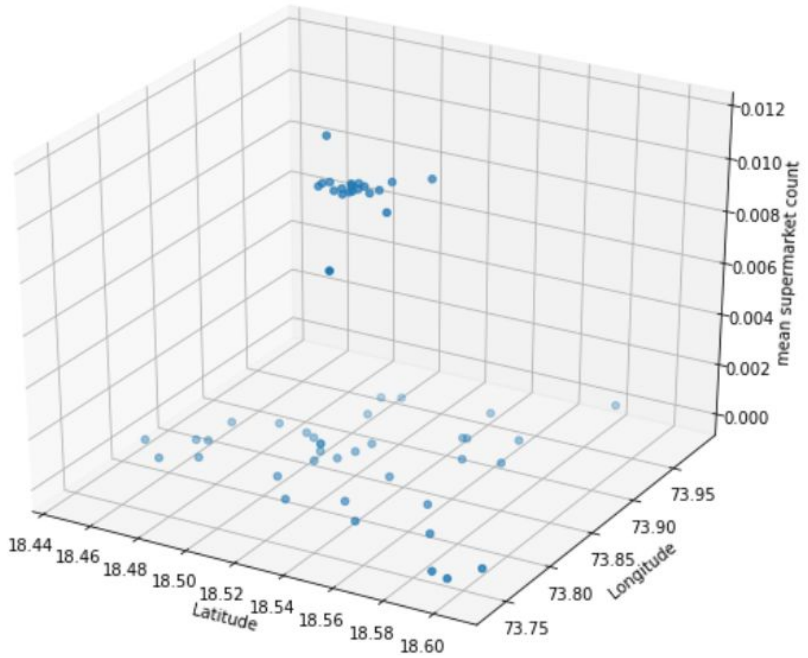
Cluster 0: Neighborhoods with very less or no number of supermarkets.

Cluster 1: Neighborhoods with a high concentration of supermarkets.

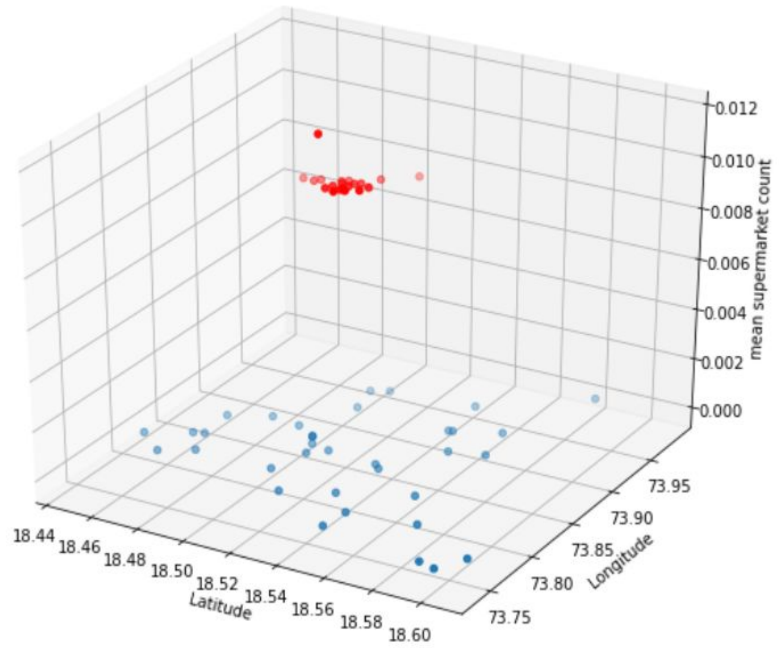
The results of the clustering are visualized in the map below with cluster 0 in red color and cluster 1 in blue color and a 3D plot showing cluster visualization.



3D plot



Clusters



Discussions

Most of the supermarkets are concentrated in the central area of the city. From the result, it is clear that there are only two kinds of neighborhoods i.e. neighborhoods with supermarkets and without supermarkets. Oversupply of supermarkets mostly happened in the central area of the city while the suburban area still have very few or no supermarkets. The neighborhoods which are in the cluster zero has very little or no supermarket this represents a great opportunity and high potential areas to open a new supermarket as it is very little or no competition from existing supermarkets.mean by the supermarkets in cluster one is likely suffering from insane competition due to oversupply and high concentration of supermarkets.from the other perspective the results also show that the power supply of supermarkets mostly happened in the central area of the city with the suburb area still have very few supermarkets.there for this project requirements the property developers to capitalize on these findings to open a new supermarket in neighborhoods in cluster zero with little or no competition. property developers with unique selling propositions to stand out from the competition and also open news supermarkets in the neighborhood in cluster 0 with the least competition.

Limitations

In this project, we only considered one factor that is the frequency of occurrence of supermarkets, there are other factors such as population and income of residence that could influence the location decision of a new supermarket.

This project uses the free stand box trial account of four square API that comes with limitations as to the number of API calls and result returned. one could make use of the paid accounts to bypass this limitation.

Conclusion

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 2 classes based on their similarities, and lastly providing recommendations to relevant stakeholders.

The answer proposed by this project for the business question is

The neighborhoods in the cluster zero are most preferred locations to open a new supermarket the findings of this project will help the relevant stakeholders to get to capitalize on the opportunities on high potential locations while awarding overcrowded areas in their decisions to open a new supermarket.

Reference

- https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Pune
- https://foursquare.com/
- https://www.ibm.com/developerworks/library/os-foursquare/index.html