Professional data science program – IBM

Week 5 assignment

Opening a New Supermarket in Amsterdam, the Netherlands

By Ali Soleymani

March 2020

Introduction

Having access to the supermarkets can be a key feature for choosing the right place to live. This can be more important for elderly and people with disability to access supermarkets easily. It is not only important to determine which area of the city have a better access to supermarkets by customers, but also it can be very important for policy makers, city municipality and investors to understand the people's needs. In the big cities like Amsterdam it is crucial to choose the right place to live and an important factor can be the access to a supermarket.

Problem

The objective of this capstone project is to analyze and select the best locations in the city of Amsterdam to open a new shopping mall. Also, the results of project can help the people to choose the suitable place to live (where they can have a fast and easy access to supermarkets). Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Amsterdam, if an investor is looking for opening a new supermarket, where would you recommend that they open it? And, if someone is looking for a suitable place to live with a good access to supermarket where would be the best place?

Target Audience of this project

This project is particularly useful for investors looking to open or invest in a new supermarket in Amsterdam. And people who currently looking for suitable accommodation.

Data

For solving this problem and answer the questions we are addressing here, we need the following data.

- List of neighborhoods in Amsterdam.
- Latitude and longitude coordinates of those neighborhoods. This is required to plot the map and to get the venue data.
- Venue data, particularly data related to the supermarkets. We will use this data to perform clustering on the neighborhoods.

Source of data and extracting methods

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Amsterdam) contains a list of neighborhoods in Amsterdam, with a total of 106 neighborhoods. 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

neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates.
Then, we will use Foursquare API to get the venue data. Foursquare has one of the largest databases 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 supermarkets 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

First, we received the list of neighborhood from the following Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Amsterdam). We scraped the web and extracted the data using Python requests and beautifulsoup packages. As this result only provide us a list of names, we also need to get the geographical coordinates in the form of latitude and longitude to use Foursquare API. To do so, we used the wonderful Geocoder package that allowed us to convert address into geographical coordinates in the form of latitude and longitude. Gathered data, data transferred into a pandas data frame and then the neighborhoods visualized using Folium package on a map. This allows us to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Amsterdam.

second, Foursquare API used to get the top 100 venues that are within a radius of 2000 meters. We then made API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare returned the venue data in JSON format and we extracted the venue name, venue category, venue latitude and longitude. Using this, 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 analyzed each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. In this way, we are also preparing the data for use in clustering. Since we are analyzing the “supermarket” data, we will filter the “supermarket” as venue category for the neighborhoods.

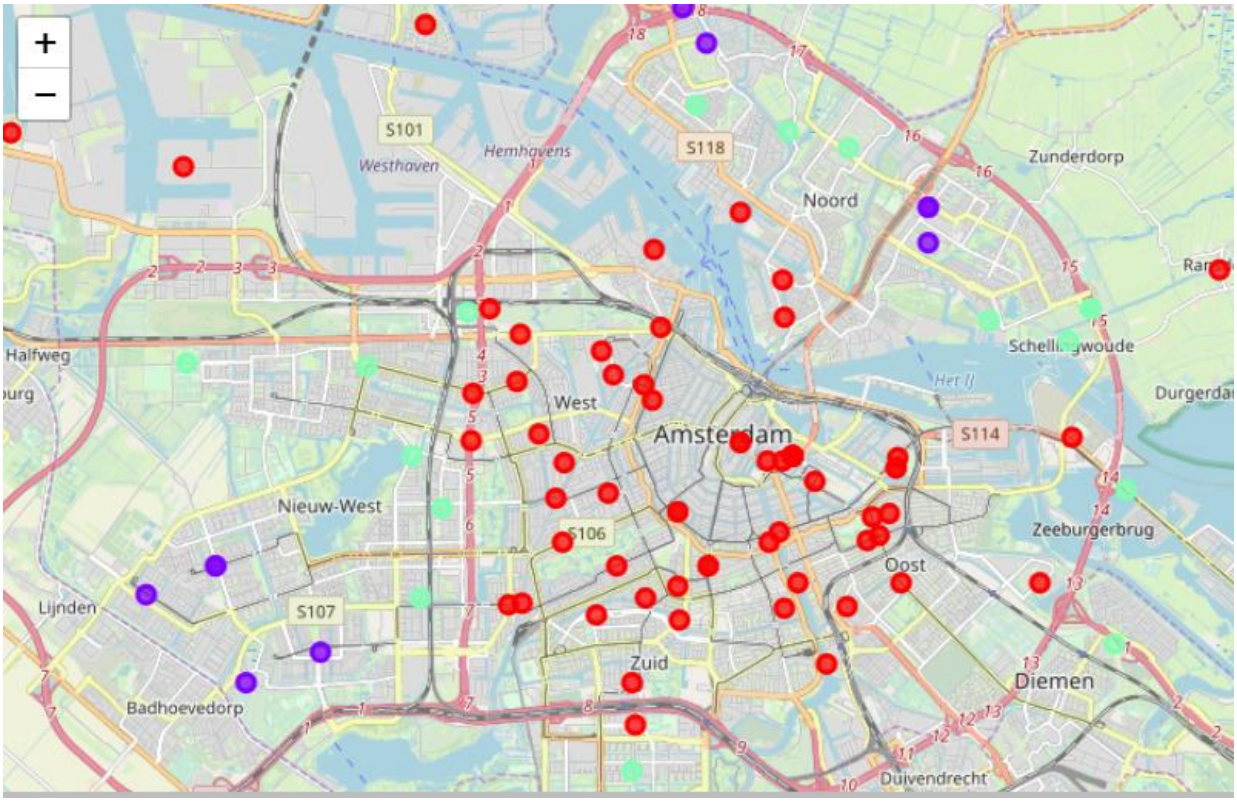
Finally, we performed clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “supermarket”. The results will allow us to identify which neighborhoods have higher concentration of supermarkets while which neighborhoods have fewer number of supermarkets. Based on the occurrence of supermarket in different neighborhoods, it help us to answer the question as to which neighborhoods are most suitable to open new shopping malls and choose a place to live.

Results

k-means clustering the neighborhoods illustrate 3 clusters based on the frequency of occurrence for supermarkets.

- • Cluster 0: Area with **moderate** number of supermarkets - Red
- • Cluster 1: Area with **low number to no existence** of supermarkets - Purple
- • Cluster 2: Area with **high** concentration of supermarkets – Mint green

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 supermarkets are concentrated in the central area of Amsterdam, 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 supermarket in the area. This represents a great opportunity and high potential areas to open new supermarket as there

is very little to no competition from existing supermarkets. Meanwhile, supermarkets in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of supermarkets. Therefore, this project recommends property developers to capitalize on these findings to open new supermarkets 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 supermarket in area cluster 0 with moderate competition. Lastly, investors are advised to avoid areas in cluster 2 which already have high concentration of supermarkets and suffering from intense competition.

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 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new supermarket. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new supermarket. 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 invest on opening a new store.