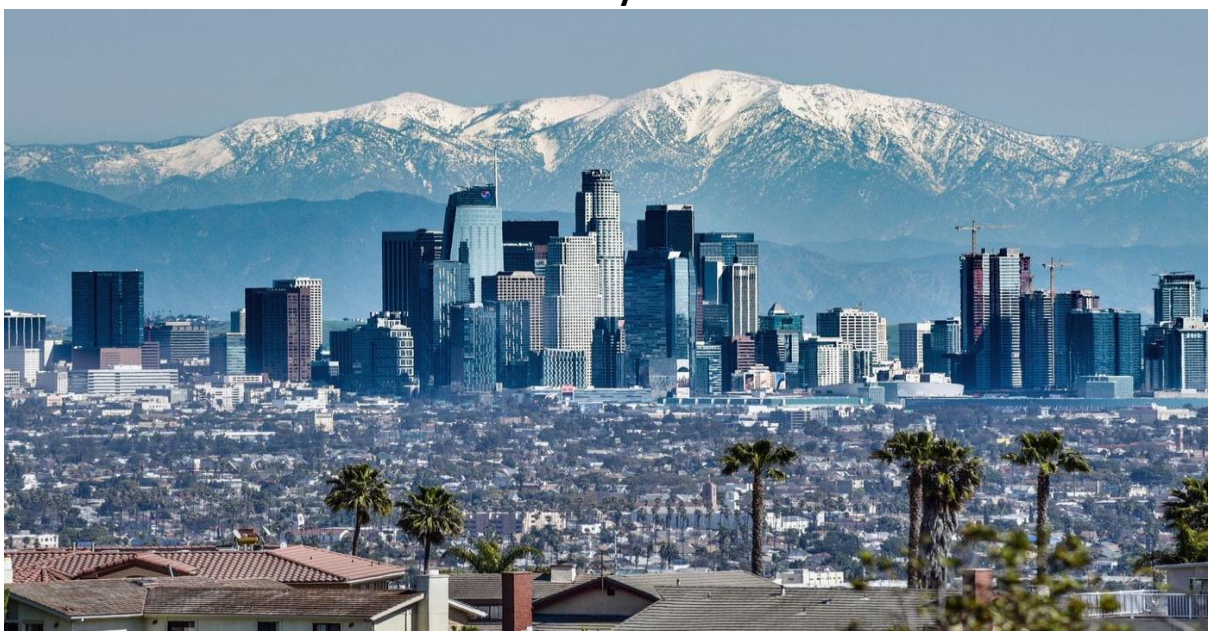
Data Science Capstone Project

Part of Coursera IBM Applied Data Science Capstone

Opening a Supermarket in Los Angeles, California

By: Saurav Arun Nair

22nd May 2020



Introduction:

In today's fast paced world, there is should be a one stop place for everything people would like to purchase rather than having to visit different shops for every particular item on their shopping list as this would be very time consuming. Supermarkets are stores where one can easily purchase all the required items like grocery, toys, food, packaged products, home appliances, clothes etc. Supermarkets are efficient as they provide the customer all they need at one location without having to visit different stores for each particular item. There are various supermarkets in Los Angeles namely Raley's Supermarket, Rosco Supermarket, CostCo, Walmart etc. As Efficient as they are, Supermarkets are costly to build and maintain as they require a big amount of land. For a real estate developer or investor, it is important that the supermarket is opened in such a location where there is less competition from other supermarkets.

Problem Description:

While opening a new supermarket, it is important to ensure that the supermarket is opened in such a place where there are no Supermarkets as this will reduce direct competition. Using Data science techniques, we can solve this problem using location wise venue data from reliable sources and using machine learning techniques to solve this problem.

Data Description:

As per the course, we can gather venue Data from Foursquare to analyse for supermarkets in each neighbourhood and make an informed decision as to where it would be efficient to open a new Supermarket. Neighbourhood data for Los Angeles is collected from Wikipedia from the Wikipedia through the below link. Using this data we will see which neighbourhood suits best to open a new supermarket. Further, we will use Data from Foursquare to get nearby venues details of the neighbourhoods gathered from the Wikipedia page. Foursquare Labs Inc., commonly known as Foursquare, is an American technology company. The company's location platform is the foundation of several business and consumer products, including the Foursquare City Guide and Foursquare Swarm apps. In November 2009, Foursquare opened up access to its API, enabling developers to access data generated by the Foursquare app and build applications on top of that data. We will be using this API to gather venue data for our neighbourhoods. Link:https://simple.wikipedia.org/wiki/List_of_districts_and_neighborhoods_in_Los_Angeles . Neighbourhood names Data will be extracted from Wikipedia in an excel sheet which will be used to collect neighbourhood names into the python notebook. Further, we using those names as an input to gather co ordinate data for each of the neighbourhoods. Foursquare will then provide venue data for each neighbourhood in the Los Angeles area.

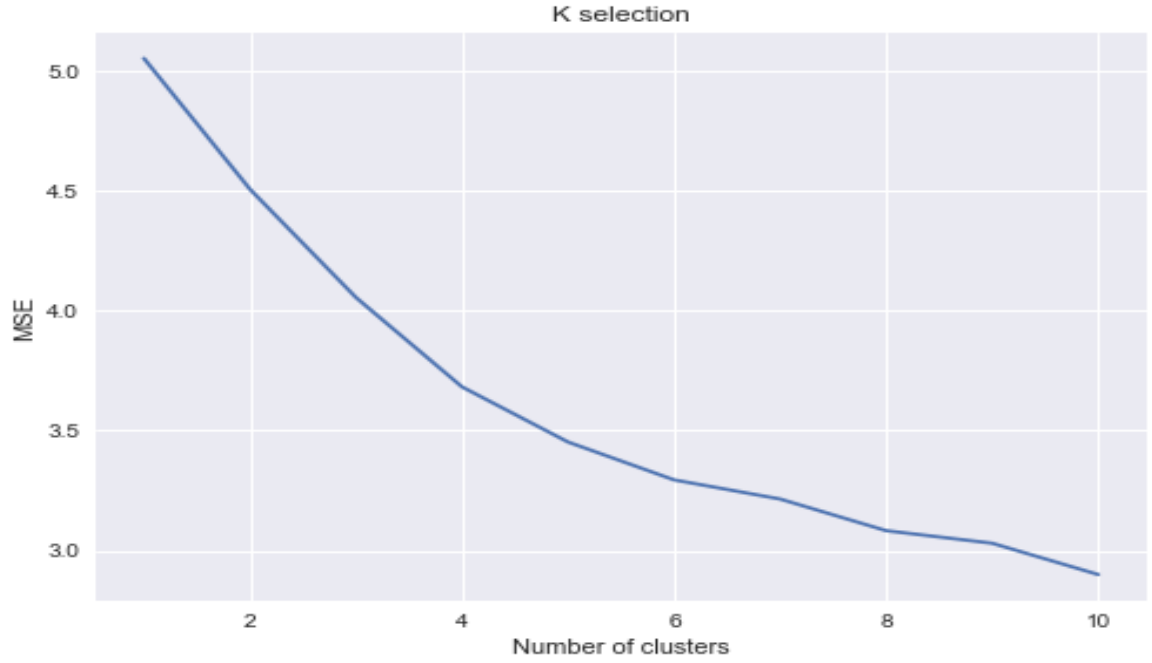
Methodology:

1. Analytic approach: We use K means clustering method to divide the neighbourhoods into various clusters based on the presence of supermarkets. This will help us to understand how many supermarkets are present in a neighbourhood. The cluster with the least amount of supermarkets will be our target cluster.

2. Data Requirements: Neighbourhood names will gathered from Wikipedia from the source mentioned in the data description section. We will also require longitude latitude details of all the neighbourhoods in our data which can be obtained through the geocoder package in python. Venue data for each neighbourhood will also be required to see what are the different outlets in a particular neighbourhood. For this purpose, we will use Foursquare API which will allow us to get nearby venue data for each neighbourhood.
3. Data Collection: Once the data requirements are met we can check the data for possible outcomes. This initial phase will help us determine if we need more data and if the data has enough parameter to make a decision. This will also help us understand if we have enough data related to supermarkets to make a prediction as to where a new supermarket can be opened. If not we can acquire more venue data from Foursquare API to make a better decision.
4. Data Understanding and Preparation: using descriptive statistics and visualization we can determine how certain parameters or variables are related to each other to better understand the link between the data by using techniques like correlation. In terms of visualization, We can use Folium package in Python to plot our data using the longitude latitude collected through geocoder. It will help us verify if our location data is correct or not. Once the data is check we can go further by using machine learning techniques to gather insights.
5. Modelling and Evaluation: K means clustering will be used to divide the data into clusters which will help us identify location with no or less supermarkets. Of the 4 clusters created, we can see that neighbourhoods in cluster 1 has no supermarkets. These neighbourhoods can serve as places where a new supermarket can be opened.

Results:

After using the K means algorithm we can see that the data can properly clustered into 4 clusters as per the image below.



Dividing the neighbourhood into 4 clusters we can see that cluster 1 has no supermarkets followed by some supermarkets in cluster 2, cluster 0 and cluster 3 respectively. This tells us that cluster 1 can be used further to determine a location for our supermarket. As there are no Supermarkets in the neighbourhoods present in cluster 1, there will less competition. This will ensure that the business will run efficiently.

Discussion:

Now that we have determined the cluster with no supermarkets, we can finalize one location from the list of neighbourhoods present in Cluster 1. As per the notebook, all these show promise for new supermarkets but we may need to consider other factors as well. We will need to take into account the cost of land in each neighbourhood as this will highly affect the Return on investment for prospective investor or owner. Furthermore, we will also have to check the population in each of these neighbourhoods as more population will mean that more people will visit the supermarket which will directly increase business income.

Future Scope:

Based on the above discussion section, we can further continue our analysis by including data for cost of land in each neighbourhood. We will also need to include data for population in each neighbourhood as these two factors will play a key role in determining the neighbourhood that will be the most suitable for opening a supermarket. Although, Supermarkets may not be present in the neighbourhood that we finalize, we may also need to consider the smaller establishments that sell the same products our supermarket is willing to sell. This may affect the business as people may rely on such stores to get their items. We can gather more data from Foursquare API to check for more similar establishments in the neighbourhood.

Conclusion:

Based on the Analysis, we can confirm that neighbourhoods present in Cluster 1 can serve as the best locations to open a new supermarket. We can further analyse base on cost of land and population data of these neighbourhoods to pick out one neighbourhood that will best suit to open a new supermarket.