Applied Data Science Capstone Project

**Introduction:**

This is the Capstone project for IBM Data Science Professional Certificate. The reason behind to choose this use case, myself and friend are planning to start a Grocery Store business , so I thought of check my skills which has been learned so far. I live in Sacramento, California. In my place there are not enough Grocery Stores. Therefore it would be a great opportunity for an entrepreneur like my friend. With this in mind, finding the location to open such a grocery store is one of the most important decisions, hence I am designing this project to help him or other entrepreneurs to find the most suitable location.

**Business Problem:**

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Grocery store in Sacramento, California. By using data science methods and tools along with

Machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question:

In Sacramento, if an entrepreneur wants to open an Grocery store, where should they consider opening it?

**TARGET AUDIENCE:**

The entrepreneur who wants to find the best location to open Grocery store or similar businesses.

**DATA:**

To solve this problem, we will need below data:

* List of neighborhoods in Sacramento, California
* Latitude and Longitude of these neighborhoods
* Venue data related to Grocery store. This will help us find the neighborhoods that are more suitable to open a Grocery store.

**EXTRACTING THE DATA:**

* Extracted neighborhood details from City of Sacramento website
* Getting Latitude and Longitude data of these neighborhoods
* Using Foursquare API to get venue data related to these neighborhoods

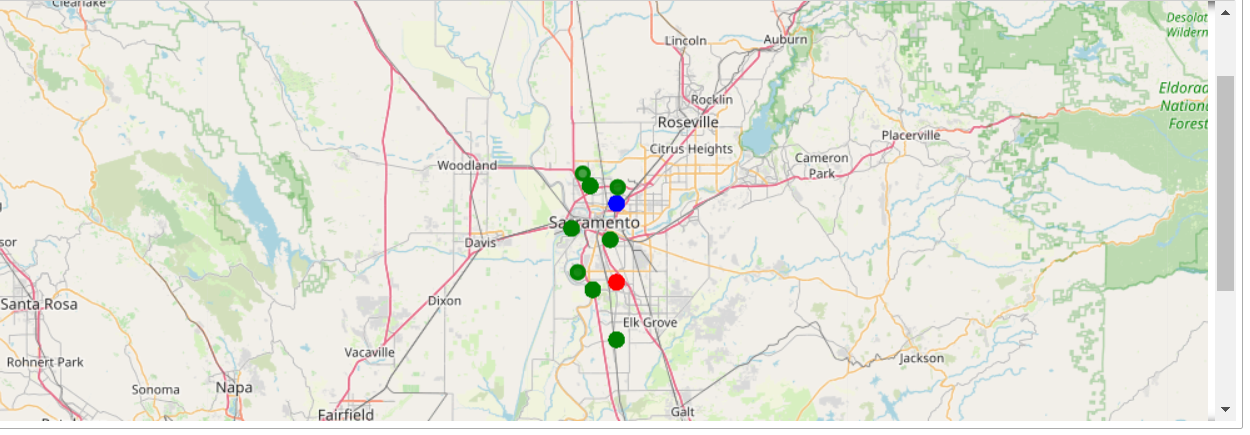
**METHODOLOGY**

First, I need to get the list of neighborhoods in Sacramento, CA USA. This is possible by extracting the list of neighborhoods from City of Sacramento website and loaded into the data frame. However, it is only a list of neighborhood names and zip codes. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I used the CSV file to match the coordinates of Sacramento neighborhoods. After gathering these coordinates, I visualize the map of Sacramento using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 50 venues within 1000 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. With this data, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I made a justification to specifically look for “Grocery Store”. Lastly, I performed the clustering method 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 it is highly suited for this project as well. I have clustered the neighborhoods in Sacramento into 3 clusters based on their frequency of occurrence for “Grocery Store”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the Grocery store.

**RESULT**

CLUSTERS



The results from k-means clustering show that we can categorize Sacramento neighborhoods into 3 clusters based on how many Grocery stores are in each neighborhood:

**RECOMMENDATIONS**

● Cluster 0: Neighborhoods with no Grocery stores

● Cluster 1: Neighborhoods with slightly more Grocery stores.

● Cluster 2: Neighborhoods with less number of Grocery stores.

The results are visualized in the above map with Cluster 0 in green, Cluster 1 in blue, Cluster 2 in red. Slightly more Grocery stores are around Deerfield/Mesa Grande Neighborhood Association, Reith Park Neighborhood Association and Valley Hi Neighborhood Association in cluster 1 neighborhoods, lowest in cluster 2 areas which are in Ben Ali Community Association, Del Paso Boulevard Partnership, and Sacramento Roundtree Homeowners Association areas. Also, there are good opportunities to open near **Creekside Natomas Neighborhood Association, Natomas Community Association, North Natomas Community Association and Robla Park Community Association** at nearby venues it seems cluster 0 might be a good location as there are not Grocery stores in these areas. Therefore, this project recommends the entrepreneur to open an Grocery stores in these locations.