Recommender System for Commercial Warehouse Report

Applied Data Science Capstone by H. Bhattacharya

Executive Summary

This Data Science Capstone project delves in the process of leveraging location data acquired from data providers such as Foursquare to explore the neighborhoods within a targeted city and create clustering models. Using K-means cluster, similar locations with minimum distance shall be grouped into clusters. It is the simplest form of unsupervised machine learning algorithm and it helps in grouping similar data points. Utilizing this model, I intend to create a solution for business start-ups who are exploring ideal locations to establish their enterprise in an urban locality.

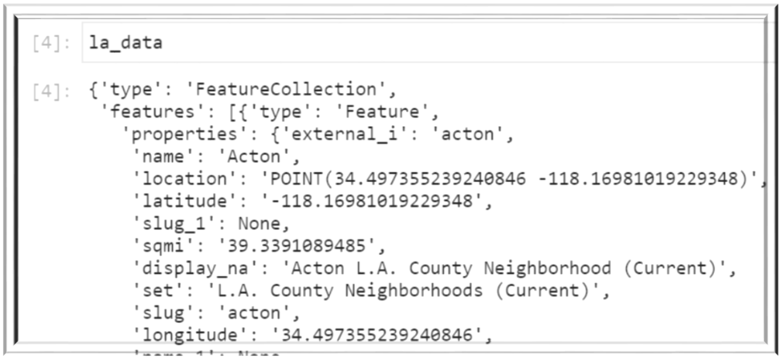
Introduction and Business Problem

Los Angeles is a historical city that is diverse, vibrant and full of opportunities. However, because of its popularity and appeal, can be an expensive place to start a business and thrive. Many businesses want to open a venue here but need to ensure that high startup cost is quickly recovered. Within such a competitive market and high real estate prices, it is challenging for upstarts to find a place to establish themselves. My intention is to use Data Science techniques learnt to make this an intelligent choice based on sold data.

In this scenario, there is a Commercial Cold Storage Warehouse provider looking to open a new Warehouse somewhere in Los Angeles. The deciding factor includes whether there are Restaurants and Grocery Stores in surrounding communities that will use this service in a positive trend.

Data Description

As we need to explore, segment, and cluster the neighborhoods in the city of Los Angeles, Los Angeles neighborhood data is key for this project. Fortunately, I was able to find the data in the appropriate structured format here https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr



```
[4]: la_data

[4]: {'type': 'FeatureCollection',
      'features': [{'type': 'Feature',
        'properties': {'external_i': 'acton',
         'name': 'Acton',
         'location': 'POINT(34.497355239240846 -118.16981019229348)',
         'latitude': '-118.16981019229348',
         'slug_1': None,
         'sqmi': '39.3391089485',
         'display_na': 'Acton L.A. County Neighborhood (Current)',
         'set': 'L.A. County Neighborhoods (Current)',
         'slug': 'acton',
         'longitude': '34.497355239240846',
```

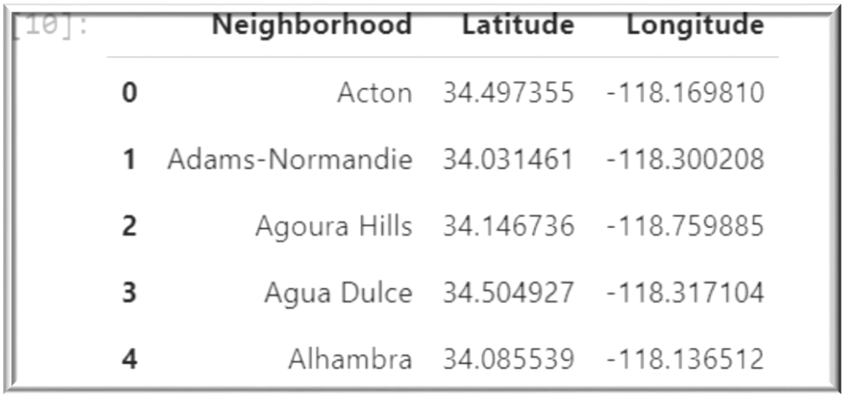
JSON snip

I was able to explore and cluster the neighborhoods in Los Angeles. The data contained the coordinates for each of the neighborhoods that helped to pull critical venue information for this project using Foursquare API. This was done using these key fields:

1. Neighborhood Name

2. Neighborhood Latitude

3. Neighborhood Longitude



[10]:

	Neighborhood	Latitude	Longitude
0	Acton	34.497355	-118.169810
1	Adams-Normandie	34.031461	-118.300208
2	Agoura Hills	34.146736	-118.759885
3	Agua Dulce	34.504927	-118.317104
4	Alhambra	34.085539	-118.136512

Data Features

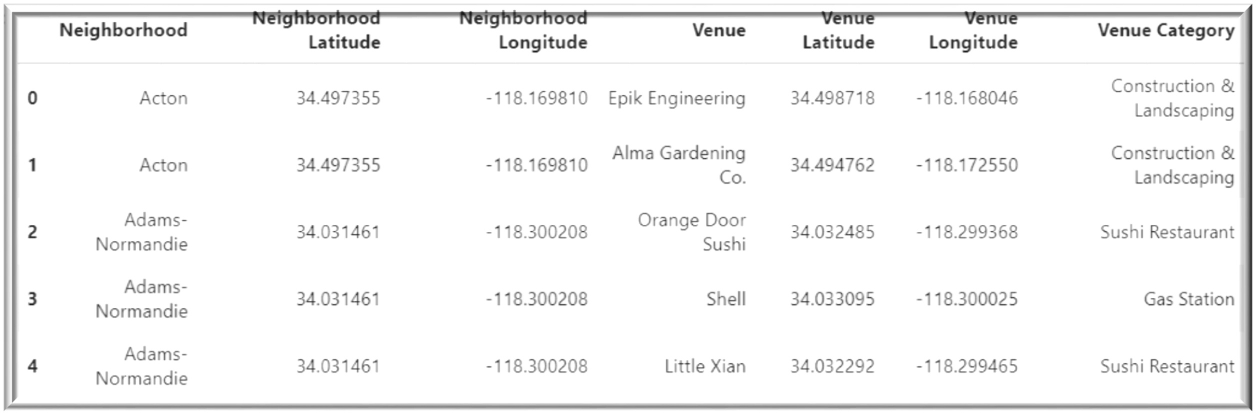
We will be leveraging on features in a reliable location information provider such as the Foursquare.com to explore the various types of venues and its categories available in each neighborhood. We will also need to understand the trending of these venues in the respective neighborhood. The information obtained per neighborhood will be as such like below and must be in a structured format:

4. Venue Name

5. Venue Category

6. Venue Latitude

7. Venue Longitude



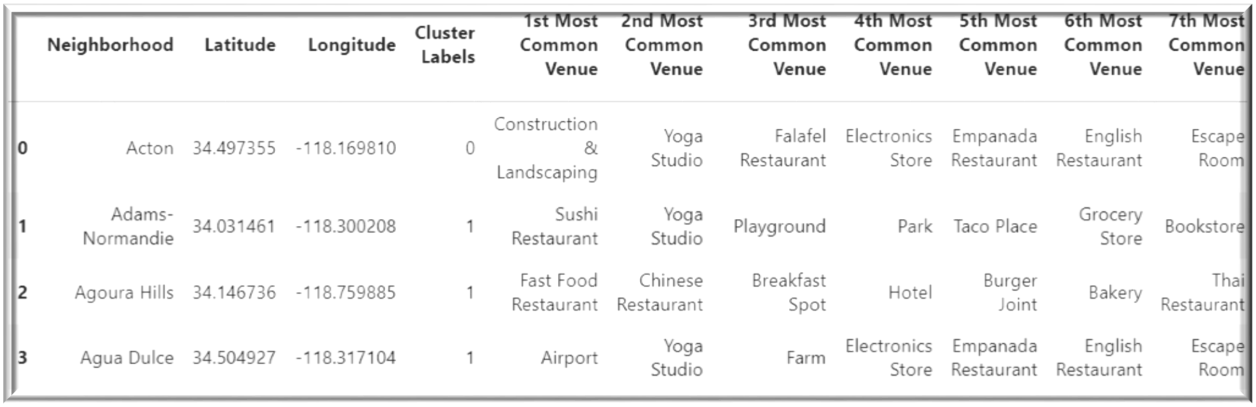
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Acton	34.497355	-118.169810	Epik Engineering	34.498718	-118.168046	Construction & Landscaping
1	Acton	34.497355	-118.169810	Alma Gardening Co.	34.494762	-118.172550	Construction & Landscaping
2	Adams-Normandie	34.031461	-118.300208	Orange Door Sushi	34.032485	-118.299368	Sushi Restaurant
3	Adams-Normandie	34.031461	-118.300208	Shell	34.033095	-118.300025	Gas Station
4	Adams-Normandie	34.031461	-118.300208	Little Xian	34.032292	-118.299465	Sushi Restaurant

How the problem will be solved

We use K-Clustering techniques to segment and cluster these neighborhoods so that we can group them together to understand their similarities and what best we can do in these types of neighborhoods.

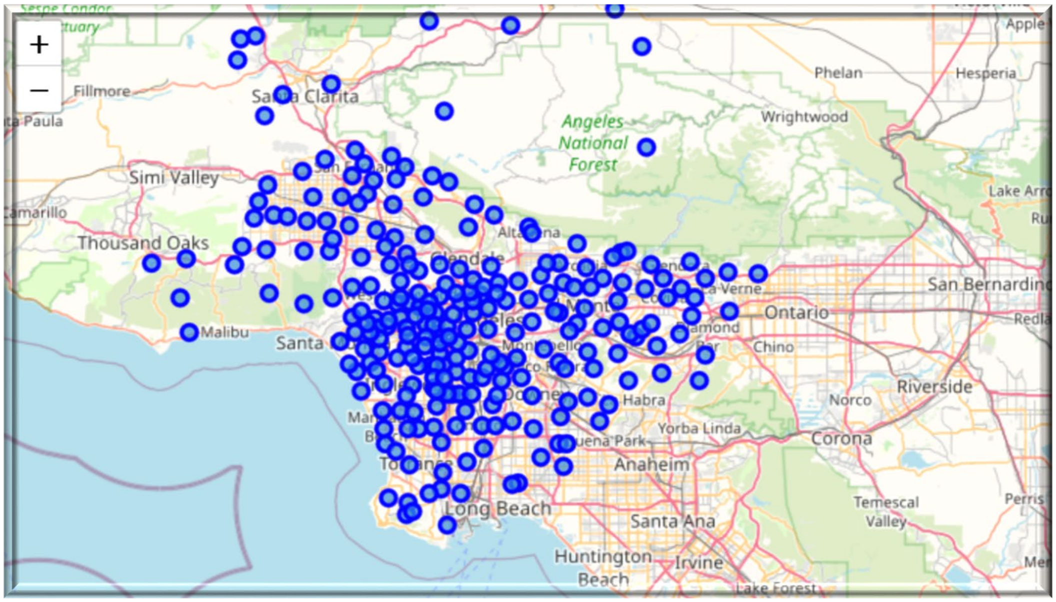
With all these features, techniques and data, we will then be able to come up with a best recommendation for the Commercial Warehouse, that is where is the optimal neighborhood for them to build and base their services. For an example, we will want to enter a neighborhood where there is a high concentration of grocery stores and restaurants; where we know that there will be a higher demand of such storage services in that area.

This is project will make use of many data science tools, working with API (Foursquare), data cleaning, data wrangling, machine learning (K-means clustering) and map visualization (Folium}.



	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Acton	34.497355	-118.169810	0	Construction & Landscaping	Yoga Studio	Falafel Restaurant	Electronics Store	Empanada Restaurant	English Restaurant	Escape Room
1	Adams-Normandie	34.031461	-118.300208	1	Sushi Restaurant	Yoga Studio	Playground	Park	Taco Place	Grocery Store	Bookstore
2	Agoura Hills	34.146736	-118.759885	1	Fast Food Restaurant	Chinese Restaurant	Breakfast Spot	Hotel	Burger Joint	Bakery	Thai Restaurant
3	Agua Dulce	34.504927	-118.317104	1	Airport	Yoga Studio	Farm	Electronics Store	Empanada Restaurant	English Restaurant	Escape Room

Dataframe with cluster labels



Folium used to superimpose neigborhoods on map

Methodology

Exploratory Data Analysis - Data Preparation

The data for this project will be extracted, processed and analysed by integrating the information for Los Angeles extracted from a dataset found in the web and venue related information acquired through Foursquare API. The data extraction will involve downloading the data set and transforming this data of nested Python dictionaries into a Pandas dataframe using the Pandas Python library.

However, this will be a list of names but also coordinates that we need in the form of latitude and longitude in order to be able to use Foursquare API later.

Using the coordinates we will populate the data into a pandas DataFrame and then visualize the LA neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned are correctly plotted in the city of Los Angeles.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 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 5 clusters. The results will allow us to identify which neighbourhoods have higher concentration of grocery stores and restaurants. Based on the occurrence of these in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open the Commercial Cold Storage Warehouse.

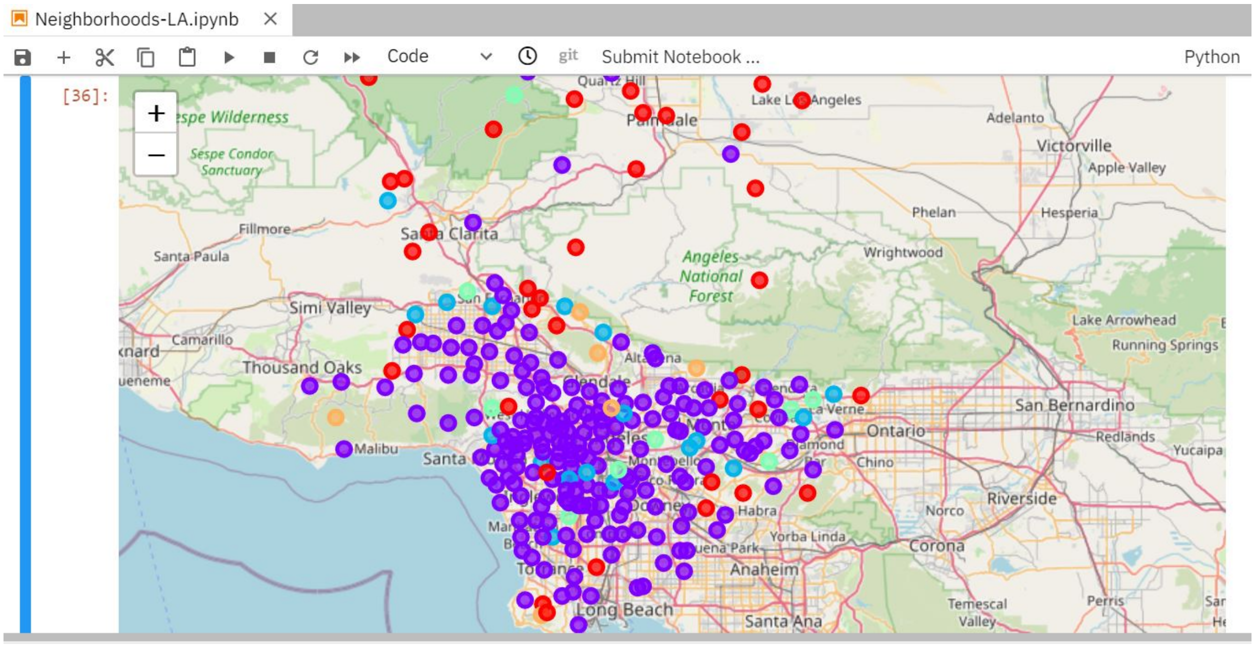
Results

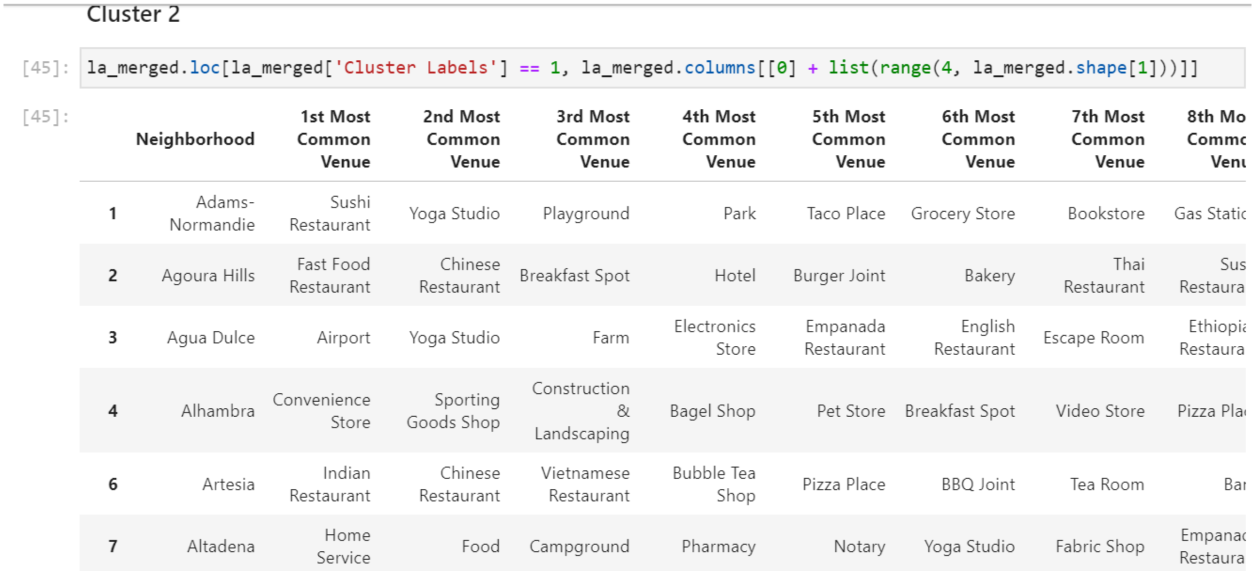
The results from the k-means clustering show that we can categorize the neighbourhoods into 5 clusters based on the frequency of occurrence for Grocery Stores and Restaurants:

- Cluster 1: Neighbourhoods with low number to zero existence of Grocery Store and Restaurants
- Cluster 2: Neighbourhoods with **high** concentration of Grocery Store and Restaurants
- Cluster 3: Neighbourhoods with low number to zero existence of Grocery Store and Restaurants
- Cluster 4: Neighbourhoods with moderate concentration of Grocery Store and Restaurants

- Cluster 5: Neighbourhoods with low number to zero existence of Grocery Store and Restaurants

The results of the clustering are visualized in the map below.





Cluster 2

[45]:
```
la_merged.loc[la_merged['Cluster Labels'] == 1, la_merged.columns[[0] + list(range(4, la_merged.shape[1]))]]
```

[45]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Mo Commc Venu
1	Adams-Normandie	Sushi Restaurant	Yoga Studio	Playground	Park	Taco Place	Grocery Store	Bookstore	Gas Static
2	Agoura Hills	Fast Food Restaurant	Chinese Restaurant	Breakfast Spot	Hotel	Burger Joint	Bakery	Thai Restaurant	Sus Restaura
3	Agua Dulce	Airport	Yoga Studio	Farm	Electronics Store	Empanada Restaurant	English Restaurant	Escape Room	Ethiopia Restaura
4	Alhambra	Convenience Store	Sporting Goods Shop	Construction & Landscaping	Bagel Shop	Pet Store	Breakfast Spot	Video Store	Pizza Plac
6	Artesia	Indian Restaurant	Chinese Restaurant	Vietnamese Restaurant	Bubble Tea Shop	Pizza Place	BBQ Joint	Tea Room	Bar
7	Altadena	Home Service	Food	Campground	Pharmacy	Notary	Yoga Studio	Fabric Shop	Empanac Restaura

Discussion

As observations noted from the map in the Results section, most of the Grocery Stores and Restaurants are concentrated in the central area of LA, with the highest number in cluster 2 and moderate number in cluster 4. On the other hand, cluster 1 and 3 have very low concentrations of Grocery Stores and Restaurants in the neighborhoods.

For the Commercial Cold Storage Warehouse provider, establishing the business in cluster 2 has the best potential as there are the highest concentration of Grocery Stores and Restaurants. This would mean that it is likely that food businesses would use the Cold Storage facilities being offered.

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 5 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 Commercial Cold Storage Warehouse.

The answer to the question raised and proposed by this project is: The neighborhoods in **cluster 2** are the most preferred locations to open a new Commercial Cold Storage Warehouse.