The Battle of Neighborhoods – Capstone Project

(Identifying Suitable Location for Business Expansion)

1. Introduction:

1.1 Background:

Your friend is running coffee shop business for past few years in different states of U.S. As a company, they felt that itss time to expand their business. They selected for e.g., Los Angeles, CA city as the location of expansion. Their business model only cares about different neighborhoods in this city which are less competitive in nature along with their proximity to the city/county (assumption). This type of business model worked for them in all their previous scenarios. They thought data can provide solution to their problem. As an experienced data scientist, they approached you for this task.

1.2 Business Problem:

The primary business problem you as a data scientist needs to solve is:

Given a county/city, you have to look for different neighborhoods in this county/city with comparatively small number of coffee shops in them, identify neighborhoods that are in close proximity to the city/county and recommend these neighborhoods to the company team. This helps the company to expand their presence accordingly to their business model.

2. Data:

2.1 Data Source:

All the neighborhoods that are located in *Los Angeles, CA* are taken from a Wikipedia page *[1]*. This page has total of 200 neighborhoods. We use these neighborhoods as a starting point for our project. As all this data only has names of the neighborhoods, to continue we need to extract more information.

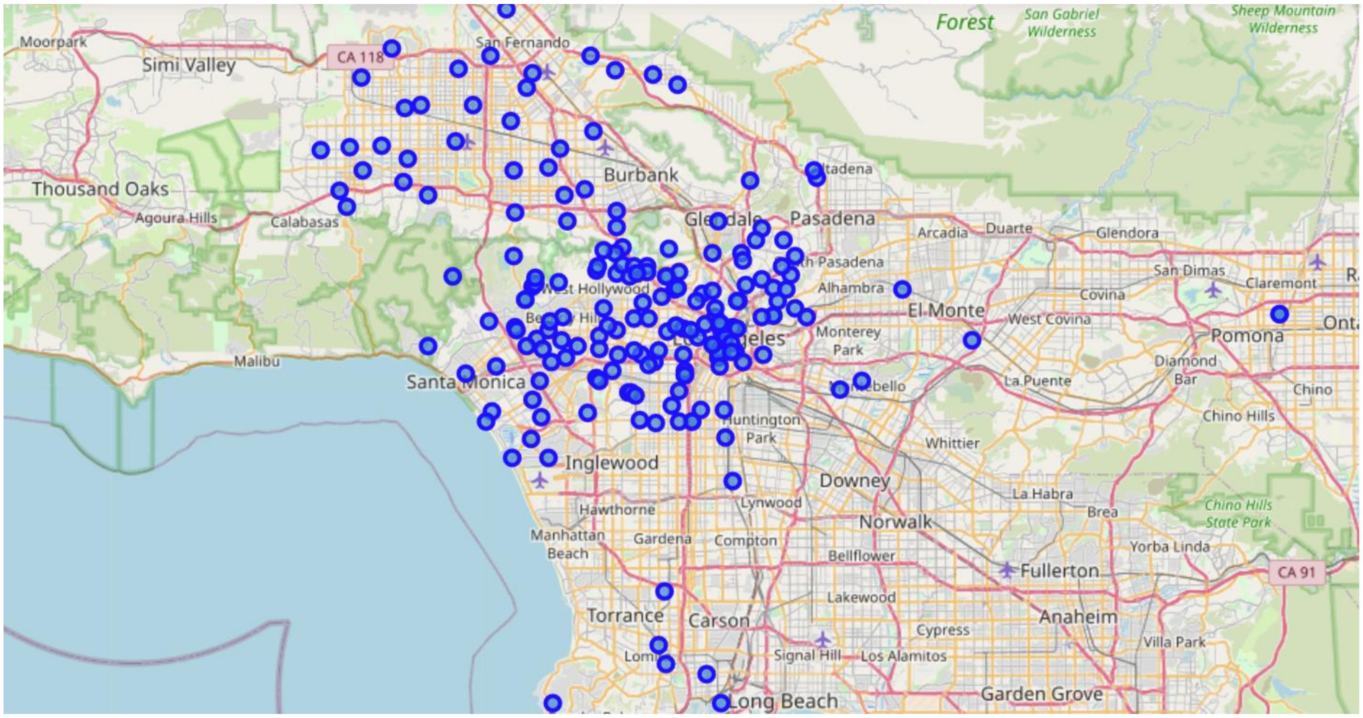
2.2 Data Extraction:

Firstly, all the neighborhood locations in the page can be extracted using Python's BeautifulSoup library. For each neighborhood, we extract its latitude and longitude values using geocoder library. Additionally, we also compute distance from city/county (Los Angeles, CA in this example) to each neighborhood in miles using geopy library. An initial dataframe is created with 'Neighborhood', 'Latitude', 'Longitude' and 'Distance' as its columns. The top 5 rows of dataframe would look like this:

	Neighborhood	Latitude	Longitude	Distance
0	Angelino Heights	34.070290	-118.254800	1.336214
1	Angeles Mesa	2.421100	-76.917380	3438.698701
2	Angelus Vista	34.087575	-118.267156	2.722398
3	Arleta	34.249050	-118.433490	17.342378
4	Arlington Heights	34.039890	-118.325160	4.822006

Visualization of Neighborhoods on Map using Folium:

To visualize geographic details of above neighborhoods on map, we use folium library in Python. I created a map of Los Angeles, CA using its latitude and longitude values. Then, I added markers to this map for each neighborhood location from the dataframe using its latitude and longitude values.



Extracting Nearby Venues using FourSquare API:

Secondly, we use FourSquare API *[3]* to extract nearby venues data for each neighborhood in the initial dataframe. To elaborate on FourSquare API, it explores the neighborhood by taking its name, latitude and longitude information along with the user credentials like Client ID and Access Token in extracting the venues that are nearby to the given neighborhood. It is mainly used to access the venues information like venue name, venue location (both latitude and longitude of the venue) and venue category. All this data is combined to form a new dataframe with columns as 'Neighborhood', 'Latitude', 'Longitude', 'Distance', 'Venue', 'Venue Latitude', 'Venue Longitude' and 'Venue Category'. After all this extraction, the tail of the dataframe would look like below:

	Neighborhood	Latitude	Longitude	Distance	Venue	Venue Latitude	Venue Longitude	Venue Category
4576	Yucca Corridor	34.10392	-118.33	6.084109	Hollywood Burger	34.100978	-118.325924	American Restaurant
4577	Yucca Corridor	34.10392	-118.33	6.084109	Dream Hollywood	34.099879	-118.330173	Hotel
4578	Yucca Corridor	34.10392	-118.33	6.084109	Trejo's Cantina	34.099513	-118.329077	Mexican Restaurant
4579	Yucca Corridor	34.10392	-118.33	6.084109	Mamas Shelter Restaurant	34.099590	-118.331391	Lounge
4580	Yucca Corridor	34.10392	-118.33	6.084109	Wood & Vine	34.101533	-118.326315	American Restaurant

There are total of 4547 venues across all the neighborhoods which are categorized to 363 unique venue categories.

3. Methodology:

We extracted all the data that is required to predict the most suitable neighborhoods for our business expansion. Now, we need to proceed with further steps to make actual prediction.

3.1 Onehot Encoding:

We are looking for different neighborhoods identifying venue categories of type Coffee Shops that are present closer to Los Angeles, CA in estimating our prediction. The venue categories here are the text labels. For any machine learning model to extract meaningful information from these text labels, we need to make them onehot encoded. In this step, we onehot encode all our venue categories of type text labels before passing them to any model. I made use of *pandas* library to serve this purpose.

A part of the head of the onehot encoded dataframe would look like this:

	Neighborhood	ATM	Acai House	Accessories Store	Adult Boutique	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	...	Warehouse Store	Waterfront	Weight Loss Center	Whisky Bar	Wine Bar
0	Angelino Heights	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
1	Angelino Heights	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
2	Angelino Heights	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
3	Angelino Heights	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
4	Angelino Heights	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0

3.2 Model Selection:

Clustering is one better type of technique which we can use in identifying the neighborhoods. The main idea here is that given the venue category (coffee shops in our e.g.) on which we want to cluster, the clustering technique ensures that the neighborhoods are clustered to the given number of clusters based on number of coffee shops available in each neighborhood. All the neighborhoods with fewer coffee shops would be made into a cluster. Now, we can extract required number of neighborhoods from these clusters with fewer coffee shops, also identifying those neighborhoods with in close proximity to the city/county and recommend them to our company for their business expansion.

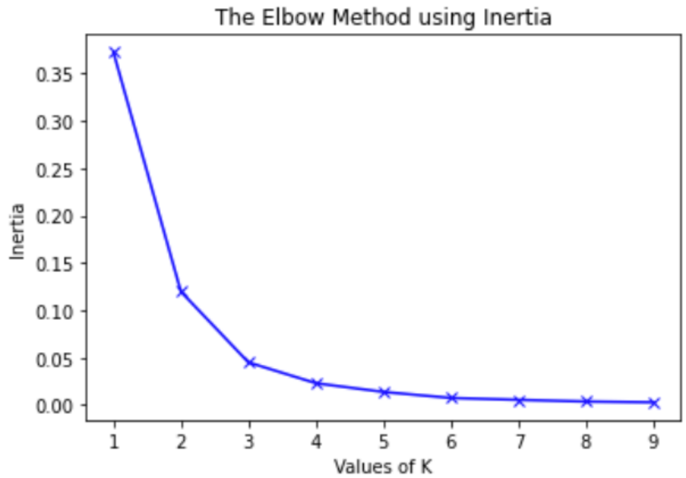
K-Means Clustering:

We will be using K-Means clustering implementation from scikit-learn library. ‘K’ in K-Means clustering represents the number of clusters we want to create.

For efficient model, we need to identify more optimal value for this ‘k’ i.e., we need to identify more optimal number of clusters for our data.

Identifying Optimal ‘K’ Value:

Elbow method *[2]* comes to our rescue in identifying more optimal value for ‘k’. The elbow method identifies this value by plotting different values of ‘k’ with its respective sum of squared distances of samples to their closest cluster center. To determine the optimal number of clusters, we have to select the value of ‘k’ at the “elbow” i.e., the point after which the sum of squared distances start decreasing in a linear fashion.

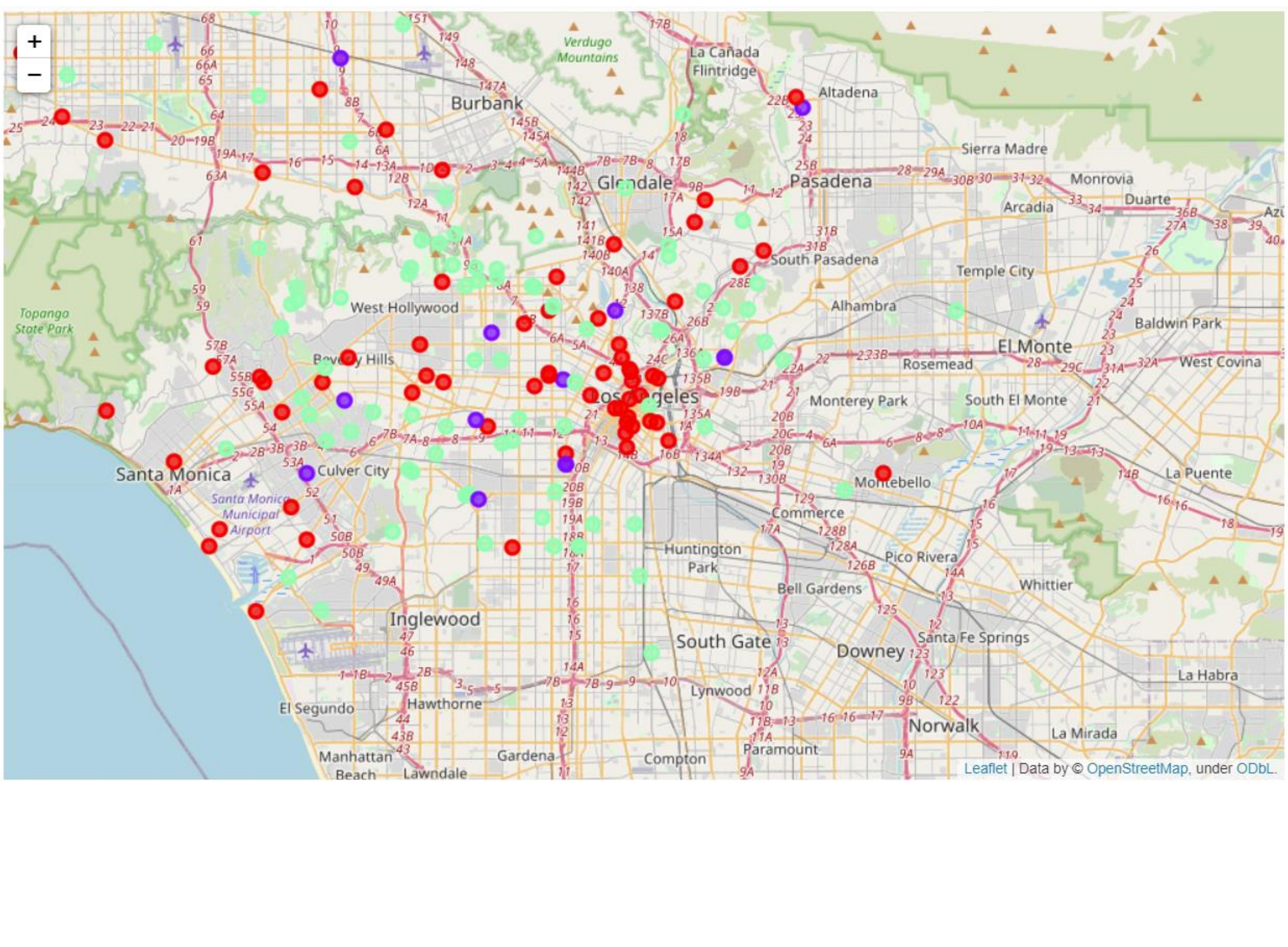


In our case, as shown in figure above, it turns out that the more optimal value for ‘k’ would be 3. So, we divide our data to 3 clusters (Cluster 0, 1 & 2) based on the range of number of coffee shops in each neighborhood.

4. Results:

4.1 Visualization of Clusters on Map using Folium:

To visualize geographic details of created clusters on map, we use folium library in Python. I created a map of Los Angeles, CA using its latitude and longitude values. Then, I added markers to this map for each neighborhood location using its latitude and longitude values with each cluster assigned to a different color.



4.3 Neighborhood Identification for Business Expansion:

In the below figure, we can see that my model predicted 5 neighborhoods in the order of their distance to the city/county. Please note that all the neighborhoods belong to Cluster 0 which does not have venue category of Coffee Shop in its top 10 venues list.

```
Looking for viable neighborhoods for your new business...
Here are my 5 recommendations:
                Neighborhood: Solano Canyon (Cluster 0), Distance from City/County = 2.0 Miles
                Neighborhood: Elysian Park (Cluster 0), Distance from City/County = 2.06 Miles
                Neighborhood: Boyle Heights (Cluster 0), Distance from City/County = 2.08 Miles
                Neighborhood: Lincoln Heights (Cluster 0), Distance from City/County = 2.1 Miles
                Neighborhood: Wilshire Park (Cluster 0), Distance from City/County = 2.14 Miles
```

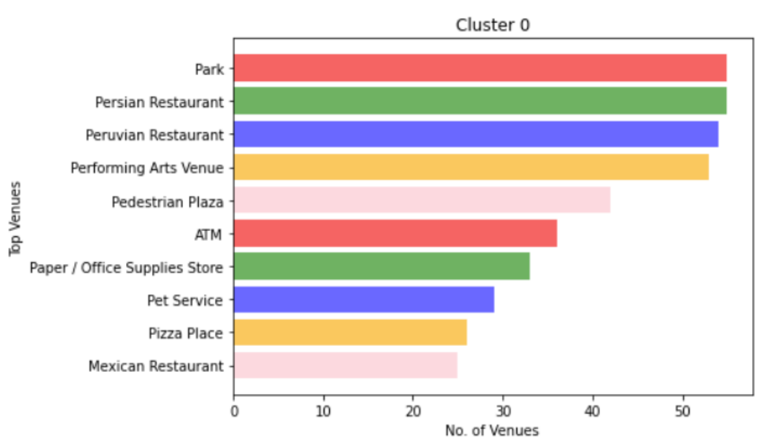
The below data frame gives the complete look into the recommended neighborhoods along with their cluster labels, distance from city/county and their top 10 most common venues.

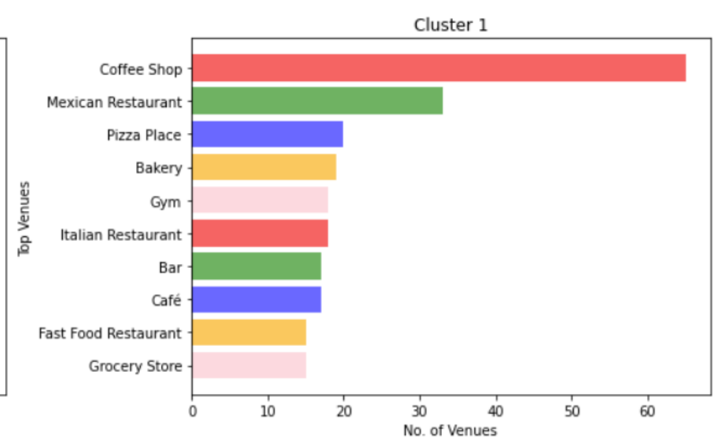
	Neighborhood	Cluster Labels	Distance	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Solano Canyon	0	2.004575	Playground	Baseball Field	Park	Basketball Court	Disc Golf	Noodle House	Paper / Office Supplies Store	Pet Store	Pet Service	Peruvian Restaurant
1	Elysian Park	0	2.060077	Park	Playground	Disc Golf	ATM	Peruvian Restaurant	Persian Restaurant	Performing Arts Venue	Pedestrian Plaza	Paper / Office Supplies Store	Outdoors & Recreation
2	Boyle Heights	0	2.076601	Grocery Store	ATM	Video Store	Ice Cream Shop	Fast Food Restaurant	Pizza Place	Cosmetics Shop	Café	Sushi Restaurant	Bank
3	Lincoln Heights	0	2.103666	Mexican Restaurant	Convenience Store	Fried Chicken Joint	Burger Joint	Music Venue	Fast Food Restaurant	Food Truck	Sandwich Place	Pizza Place	Outdoor Sculpture
4	Wilshire Park	0	2.137273	Latin American Restaurant	Fast Food Restaurant	Karaoke Bar	Mexican Restaurant	Korean Restaurant	Park	Chinese Restaurant	Theater	Seafood Restaurant	Convenience Store

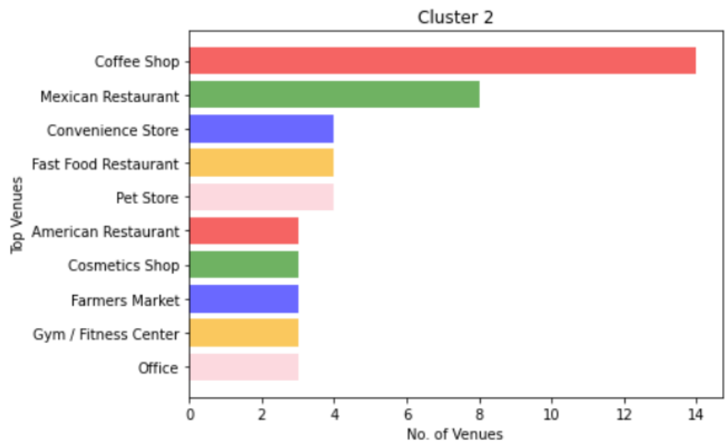
5. Discussion:

5.1 Bar Chart Visualization:

Let us visualize the top venue categories in each cluster using a bar chart.







In these bar charts, we can see that Coffee Shop is one of the top 10 venue categories in both Cluster 1 and Cluster 2. The Coffee Shop does not make into the list of top 10 venue categories in Cluster 0. Thus, our most of the recommended neighborhoods are from Cluster 0 and are in close proximity to the city/county.

Additionally, my code is more generic and works for any venue category by passing the required venue category while applying clustering. For neighborhood prediction, we can pass variable number of neighborhoods we are looking for and my code ensures that all the required conditions are met and return these neighborhoods. With respect to bar chart visualizations, we can pass variable number of top venues that we want to plot. In this example, I limited number of clusters to 3, venue category to 'Coffee Shop', number of neighborhoods to recommend to 5 and number of top venues to 10 for bar chart visualization.

6. Conclusion:

In this project considering myself as a Data Scientist, I recommended the neighborhoods for a company who wants to expand their Coffee Shop business in Los Angeles, CA. I used K-Means clustering technique to identify three different clusters of neighborhoods based on the range of number of coffee shops in each neighborhood. Finally, after meeting all the required conditions that the company has mentioned I recommended five neighborhoods. These five neighborhoods are in close proximity to the Los Angeles, CA in comparison to the other neighborhoods of the same cluster. The effectiveness of these recommendations has been verified by plotting the top ten venues of all the clusters using a bar chart. Appropriate visualizations are also provided whereever required in the project.

7. References:

1. https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_in_Los_Angeles
2. https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/
3. https://developer.foursquare.com/