**The Battle of Neighborhoods**

**Helping a Client to Identify locations for new business venture**

**Harshal Inamdar**

**26th April’2020**

1. **Introduction**
   1. **Overview of the Problem Statement**

**Client Description --**

The firm is based out of Liverpool UK.

The company was established in Year 1995.

Apart from the main branch, it also has presence in the other major cities of UK like -- London, Manchester, Swansea, Edinburgh, Birmingham.

However, it has no overseas presence.

**Decision to venture into The United States of America**

To try its luck outside of UK, the firm has finalized that the it wants to venture into the United States of America.

**Narrowing down to the specifics**

While the company has decided that the city where they would want to start the 1st Ice Cream Parlor will be California owing to few undisclosed reasons, the company has not been able to figure out in which specific neighborhood the 1st franchise will start.

Also, it should be noted that this 1st franchise will be the company owned franchise. Based on the performance evaluation over a period of 1 year, company will decide on the future course.

**Finalizing the Consultant**

So as to be able to find a suitable neighborhood without much experimentation, the firm has approached a California based consultant specialized in Supply Chain Data Analytics to narrow down to areas where it can start its journey in United States.

Consultant is expected to pinpoint the areas where the 1st Ice Cream Parlor in the United States can start. Consultant can have his/her assumptions based on his experience to define the methodology for arriving at the probable locations.

1. **Methodology to Be Adopted**

**How to gather the data and build the approach to come up with the solution?**

Now that the consultant has been identified to come up with a solution, let us check what approach he can choose –

**Steps in building the required Data –**

1. **Have a detailed list of Postal Codes for California State along with the Latitude and Longitude details for each and every postal code**

There are multiple web sites on internet where you can get the Postal Codes for California State along with Latitude and Longitude Details Once the website is identified, the details can be extracted in one of the two ways --

Scrape the data from the website using html5 library and load the same in pandas

Get the details downloaded in a CSV format. Make necessary changes in the data and then upload the same in the pandas

In case we want to scrape the data using html5 library, following command needs to be executed to install the package - pip install html5lib

Once that is done - we will need lxml package to scrape the contents of the webpage - pip install lxml

Next Step - Import Pandas - import pandas as pd

Using following syntax to get the details in pandas - dfs = pd.read\_html ('URL', header = 0)

Now, we have a data in place that will consist at least following columns -

1. Postal Code

2. Borough

3. City

4. State

5. Latitude

6. Longitude

There can be few changes made like - removing the State column considering for all the columns it will be the same - California

1. **Considering there are multiple cities in the California State, shortlisting only Los Angeles for narrowing down the research part. Assumption is that compared to other Cities, Los Angeles is a better pick.**

This can be replicating the Data frame to a new one with only the details where City is **'Los Angeles'**

**3. Now that we have all the Latitudes and Longitudes details for all the postal codes, next thing that we need is the neighborhood details for all these postal codes**

Neighborhood details actually mean Venues surrounding the Postal Code

We can get the details from Foursquare location data

Getting the Credentials First

CLIENT\_ID = 'Foursquare ID' CLIENT\_SECRET = 'Foursquare Secret' VERSION = 'Version Number'

**Generating URL with which requests can be made to Foursquare APIs for fetching Venues data –**

url = '[https://api.foursquare.com/v2/venues/explore?&client\_id={}&client\_secret={}&v={}&ll={},{}&radius={}&limit={}'.format](https://api.foursquare.com/v2/venues/explore?&client_id=%7B%7D&client_secret=%7B%7D&v=%7B%7D&ll=%7B%7D,%7B%7D&radius=%7B%7D&limit=%7B%7D%27.format)( CLIENT\_ID, CLIENT\_SECRET, VERSION, latitude, longitude, radius, limit) url

**4. Using the Foursquare Location data for getting the details about the nearby Venues**

We already have the Credentials in place and also the URL created. Now we can request Foursquare API for the venue’s details.

For that, 1st step is to import requests library using following command - **import requests**

Once the request is made, the result that we can from Foursquare will be a JSON file. To be able to read it properly and do required operations on it, we will need to get the details from JSON transferred to Pandas Data frame.

For this we will need JSON Normalization package by executing following command - **from pandas.io.json import json\_normalize**

Making the request be using following code - **results = requests.get(url).json()**

**5. What to expect from Neighborhood Data and Why do we need that?**

Which Neighborhoods should be more appealing for an Ice Cream Parlor?

When we want our Ice Cream Parlor to be in an easy reach to the customers, first level of filtering the venues that we have will be to identify areas with a good population density.

How to figure out whether the Density is low or high – This we can figure out based on the existing venues in the localities – The places where there are many restaurants, shopping malls, schools, colleges, universities, firms, corporate/business parks.

In general, this is all about finding areas that are always filled with potential customers rather than choosing the areas which are too residential and lacks the human activity needed to sustain the business.

**Important venues to be tracked from the Foursquare Location Data -**

1. Restaurants/Pubs
2. Malls
3. Multiplexes, Cinema Theatres, Local Operas
4. Schools
5. Universities/Colleges
6. Industries/Firms/Business Parks
7. Airports/Bus Terminals/Metro Stations

Apart from these, once we have the venue details, we can add few more to the list.

**6. Finding Top Venues for all the Postal Codes to check the areas of our interest, the areas that will suit an Ice Cream Parlor**

This will be done by Finding 10 common venues in order for all the postal codes that we have. This will be done using One Hot Coding - Having a counter set for all the venues and counting the number of such venues for each of the neighbourhood. Once that is done, sorting it in a descending order and coming up with the Top 10 Venues for each Borough/Postal Code.

Here, it is now important, to check the venues that are not of interest and removing those from the list.

Now we will have a reduced list of Boroughs having the venues or localities that have interest for us.

**7. Using the Top Venues Data, cluster will be formed of different areas based on similarity of venues and then it will be presented to the Client - IceStorm**

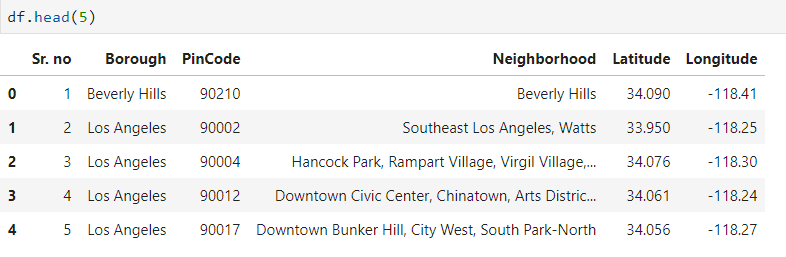
Clusters can be formed using the similarity of the top venues for each of the Borough. Clustering can be done using K-Means.

This now can be mapped on a graph using the geopy plotting function

**8. Once we have details on the venues of the clusters, we can the prioritize the details based on the list of neighborhood in point no. 4 and come up with 2-3 Postal Codes where IceStorm can start its Ice Cream Parlor.**

We can then present the shortlisted clusters to the Client and then see if any further narrowing is needed and if we can get more specifications to narrow down the result.

1. **Data Section**
2. **Importing a CSV file having Latitude Longitude and Neighborhood Details –**

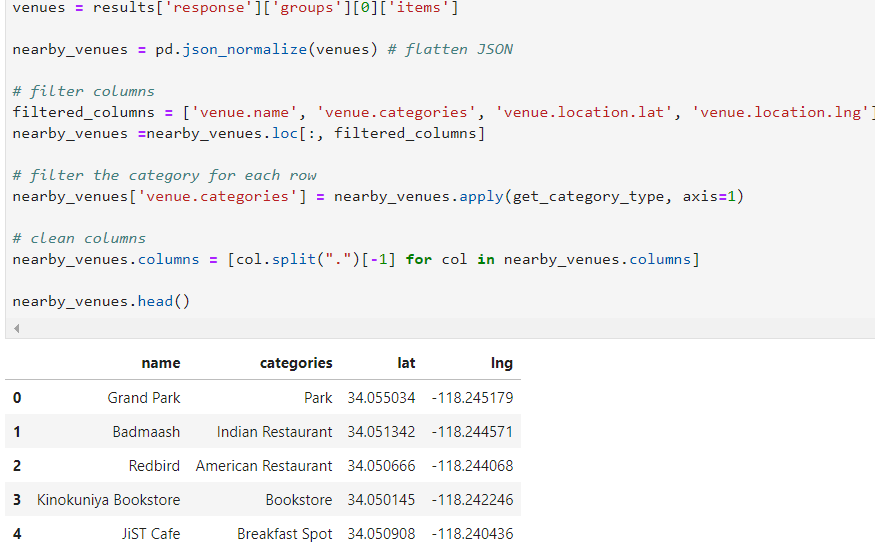


1. **Setting up Foursquare Credentials and generate Latitude and Longitude for Los Angeles City**
2. **Importing Features by making API Call to the Foursquare data**



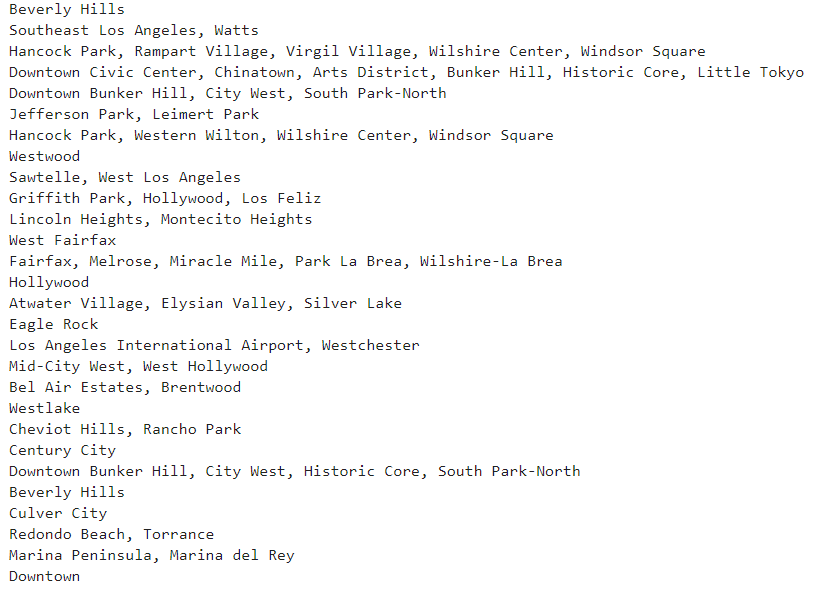
1. **Narrowing down the data and choosing the features that are needed –**

**Choosing only Venue Names, Venue Categories and Location Data**



1. **Now that we have the data from foursquare, aligning this with the Neighborhood data that we already have –**

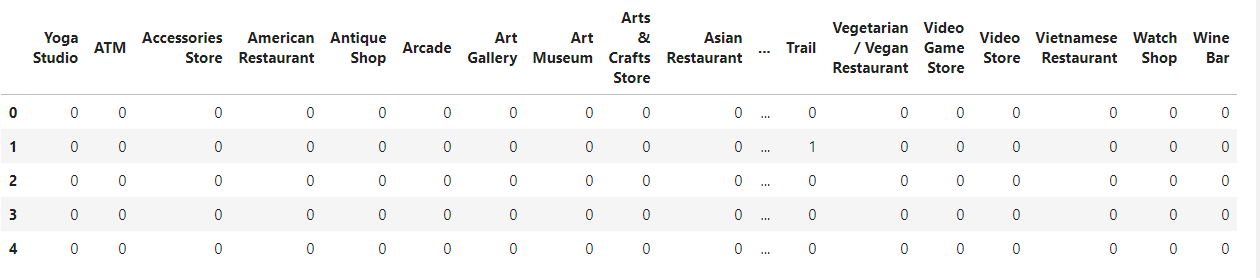
**Getting Nearby Venues for each of the Neighborhood by defining a function for the same.**



1. **Check how many unique venue categories we have –**



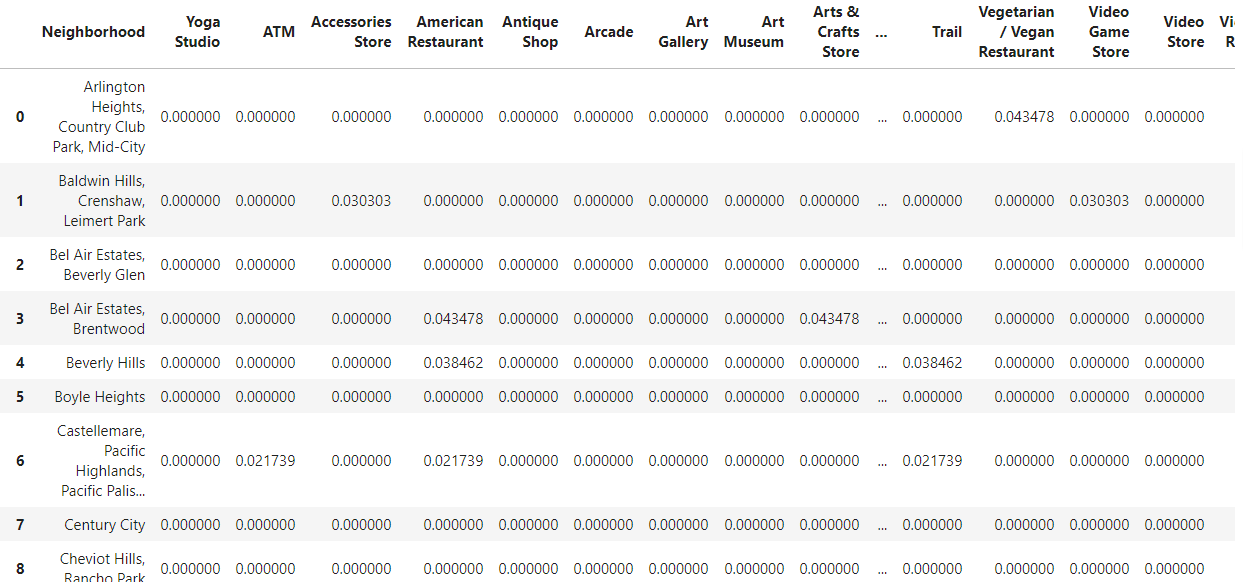
1. **Counting the similar kind of venues in each neighborhood. Setting up a counter for the same.**

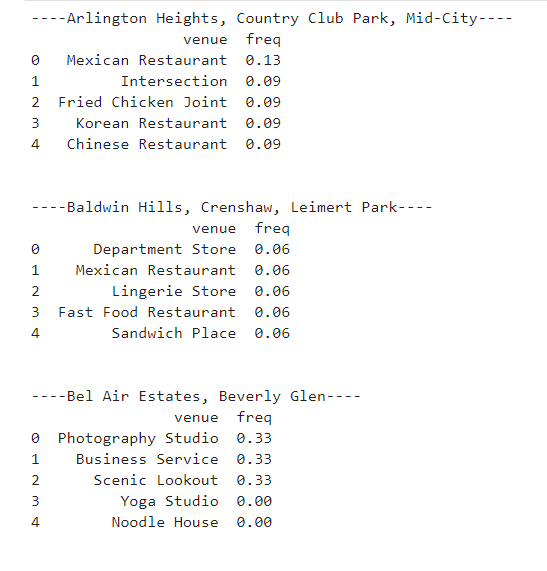


1. **Finding Top Venues for each neighborhood –**

Calculating frequency of similar venues in each neighborhood.

Grouping the neighborhood such that similar Neighborhoods can then be classified in same cluster.

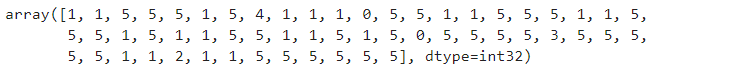


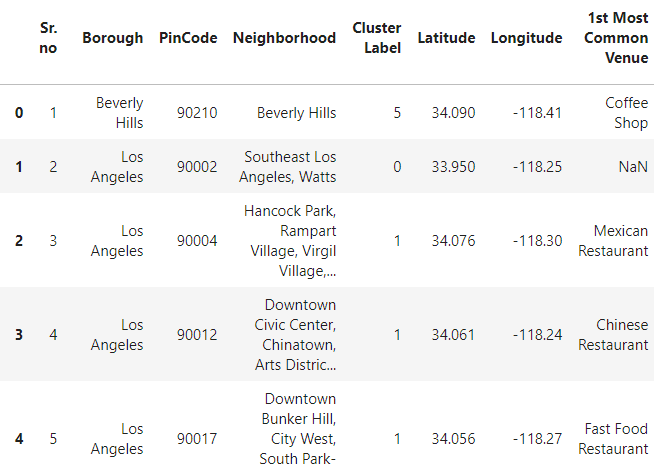


1. **Deriving at the most common venues for all the Neighborhoods and arranging in order**

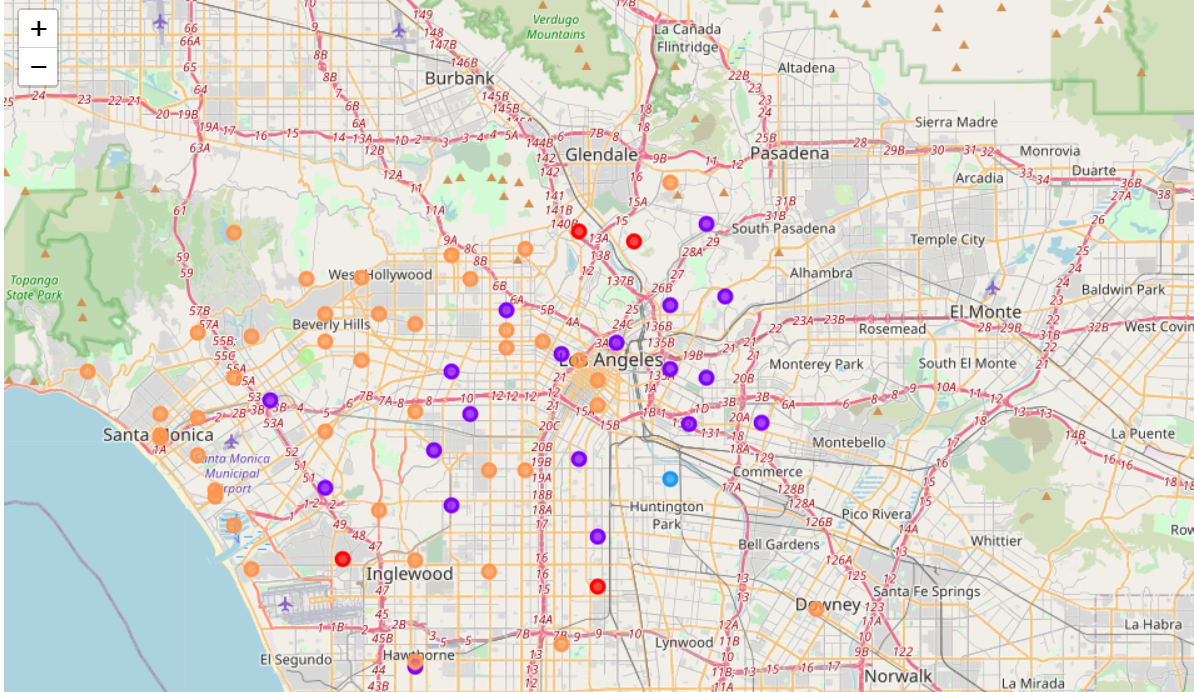


1. **Import KMeans from Sklearn.cluster to assign cluster labels for each Neighborhood**





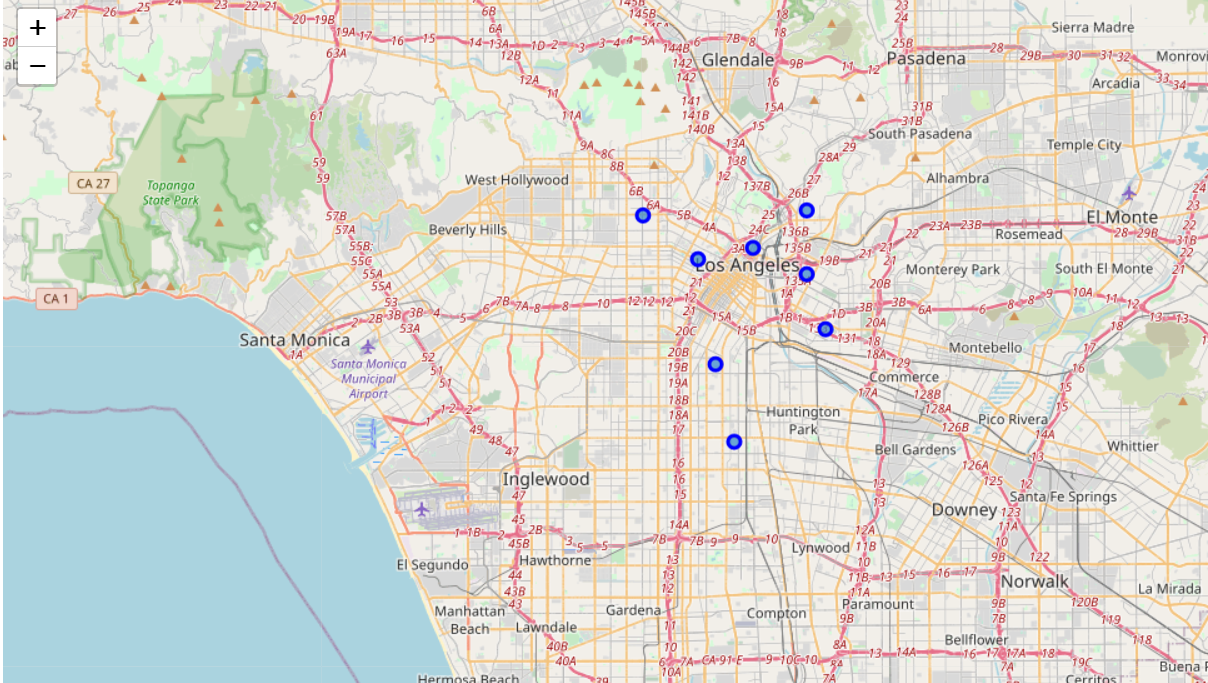
1. **Checking the clusters on a map using Folium**



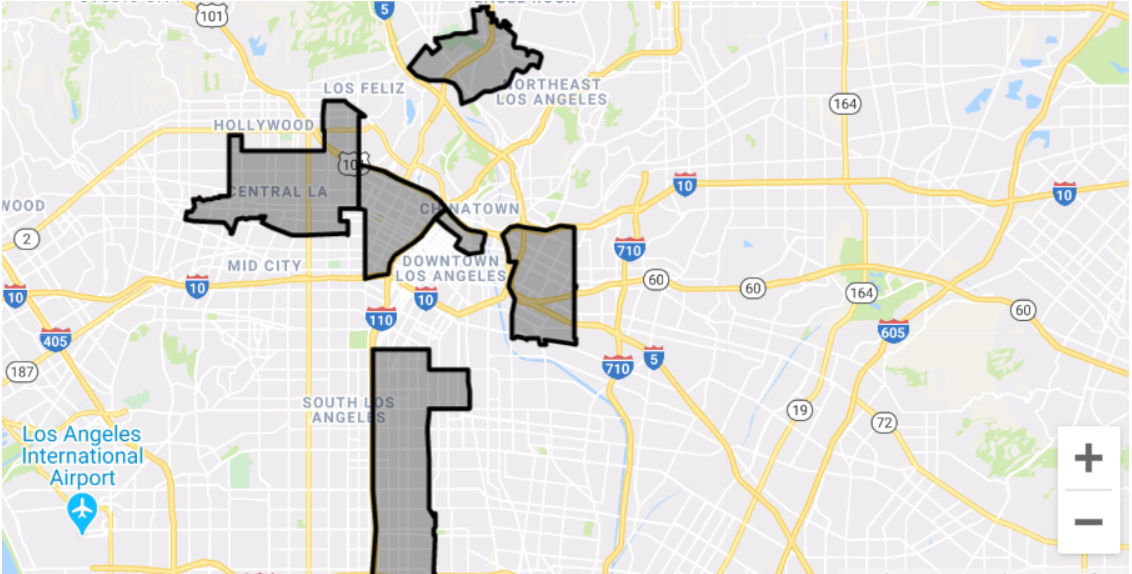
1. Taking into account the city centre latitude and longitude, only the co-ordinates falling in the following bracket are kept and others are excluded from the data –



Co-ordinates between 33.95 N and -118.30 W and 34.20 and -118.20 kept. Others are removed.



1. **Getting the Data around the commercial/retails rental pricing in Los Angeles to help client come up with a decision of choosing the location from the final shortlist of 8.**



Based on the details received on one of a real estate website following is the order of the rates in these localities –

1. Boyle Heights - $4000 Sq. ft

2. Civic Center - Little Tokyo - $6000 per Sq.ft

3. Glassel Park - $5000 per Sq.ft

4. Mid Wilshire - $2000 per Sq.ft

5. South east Los Angeles - $8000 per Sq.ft

6. Westlake - $2000 per Sq.ft.

1. **Decoding and Discussing the results –**

Thus, post doing the clustering, out of the total clusters we had, it was needed to choose from the cluster based on the top venues on which each clusters were based on.

Out of the two highly populated clusters i.e. Cluster no. 1 and Cluster no. 5, below are the general trends, based on which we can make recommendations to the client.

1. Most common location for Cluster 5 is Coffee Shop and for Cluster 1, it is Mexican Restaurant.
2. In terms of density with respect to restaurants, that can be seen as high in Cluster 1, whereas in Cluster 5, it is more with respect to Coffee Shops., Gym, Yoga, though there are numerous restaurants.
3. Though not much difference in the way of clustering, it is ideal to go with the area having more restaurants than the coffee shops.
4. Restaurants attract more crowd owing to the large sitting places that they have compared to coffee shops; thus, it will be okay to assume the density of crowd at the places in Cluster 1 will be more than in Cluster 5.
5. In addition to this, in Cluster 1, there are following locations in the common venues apart from the restaurants –
   1. Grocery Store
   2. Convenience Store
   3. Historic Sites
   4. Museum
   5. Park
   6. Coffee Shops
   7. Discount Stores

Now that we have a list of locations from Cluster 1, next thing will be to classify them further.

Based on one of the inputs received from Client, it is imperative that the location of the Ice Cream Parlor should be in a locality that is not in the suburbs but having proximity to the city centre.

Considering this is the first venture not only in United States but also out of UK i.e. home country, Client management has provided clear-cut inputs stating the location should be such that it should make sure to catch maximum eyeballs and ice cream parlour should make its presence felt in the heart of the city.

Owing to this, next step was to check the Location details of the Los Angeles City Center, the very hub of the city.

Location Details were - 34.0522° N, 118.2437° W

To make sure the neighbourhoods aren’t much far from the city centre, only locations falling within following locations selected –

Latitude – 33.95 N to 34.20 N

Longitude - -118.30 W to -118.20 W

Those locations following out of these ranges are removed. Thus, we have narrowed down the list of locations to 8.

1. **Conclusion**

In case the client wants to go for a cheapest alternative of these, it can start Ice Cream Parlor at either Mid-Wilshire or Westlake.

However, in case there are no limitations with respect to rental pricing, firm can decide the go with Civic Center - Little Tokyo owing to following advantages it seems to have over others --

1. In terms of location it is extremely close to the city centre - 34.061 Lat -118.24 Long

2. Varied Cuisines served in the locality = Chinese, Mexican, Vietnamese, Continental, Seafood - This will attract diverse groups of people to this area which can promote branding of the firm in different communities.

Apart from the Restaurants, there is a huge list of other places that can attract large sections of crowd –

1. El Pueblo Los Angeles Plaza Historical Park

2. Proximity to Santa Anna Freeway increasing the accessibility for people from other localities

3. Metro Station nearby

4. Los Angeles Union Station

5. Los Angeles State Historical Park

6. Los Angeles Public Library

7. Far East Plaza

8. Banks

9. Enormous Parking Space nearby

10. Alpine Recreation Center and Park

Thus, overall a hotspot attracting people from all over the town.

Hence, as a consultant, it is advisable to have the new ice cream parlour setup at one of the 8 locations, with Downtown Civic Center, Chinatown, being the 1st preference.