Venues Data Analysis of Los Angeles for setting up a new business

Amlan Abhisek Nayak
June 7, 2019

1. Introduction

1.1 Background

Los Angeles ,officially the City of Los Angeles and often known by its initials L.A., is the most populous city in California, the second most populous city in the United States, after New York City, and the third most populous city in North America. The area is a large market in its own right, and it is still growing. There are 17.9 million residents in the five-county area, and nearly 10 million in Los Angeles County. The population increases between 2000 and 2010 were 1.5 million and 283,087 respectively. Census Bureau projections to 2025 indicate significant additional growth. In addition, there is quick access to markets in San Diego and Northern California, as well as Arizona and Nevada.

1.2 Problem

In a city like Los Angeles, there is enormous opportunity to establish a business like restaurant, mall , office etc. One among many problems lies in identifying the place where to setup your business. There are various neighbourhoods in Los Angeles and they are already crowded with venues like restaurants, recreation places, mall etc . Apart from this, the rent price varies in each and every neighbourhood. The challenge here is in understanding the demand , existing competition and infrastructure cost like rent.

1.3 Interest

For a businessman who wants to set up a restaurant, his interest would lie in two factors. First, he would prefer neighbourhoods where intensity of restaurants or that particular category of restaurant, for example, Chinese or Italian , is less and hence the competition will be less. Second, he would also consider the average rent price in that neighbourhood. Whereas, for a company who wants to set up an office would opt for places with certain type of social venues and also consider the rent price of that area.

Keeping in mind these two important factors, we will create a map where we can visualize the neighbourhoods clustered according to venue category , and the average rent price of the neighbourhood is also listed.

2. Data Acquisition and Cleaning

2.1 Data Sources

To consider the problem we can list the datas as below:

A. The rent price for Los Angeles neighbourhoods can be found here . The dataset has the rent price for each and every neighbourhood of Los Angeles from year 2010 to 2016. I couldn't find any reliable source for more recent data than this. Therefore, I used the data from year 2016 for the work. The dataset also contains location coordinates of these neighbourhoods.

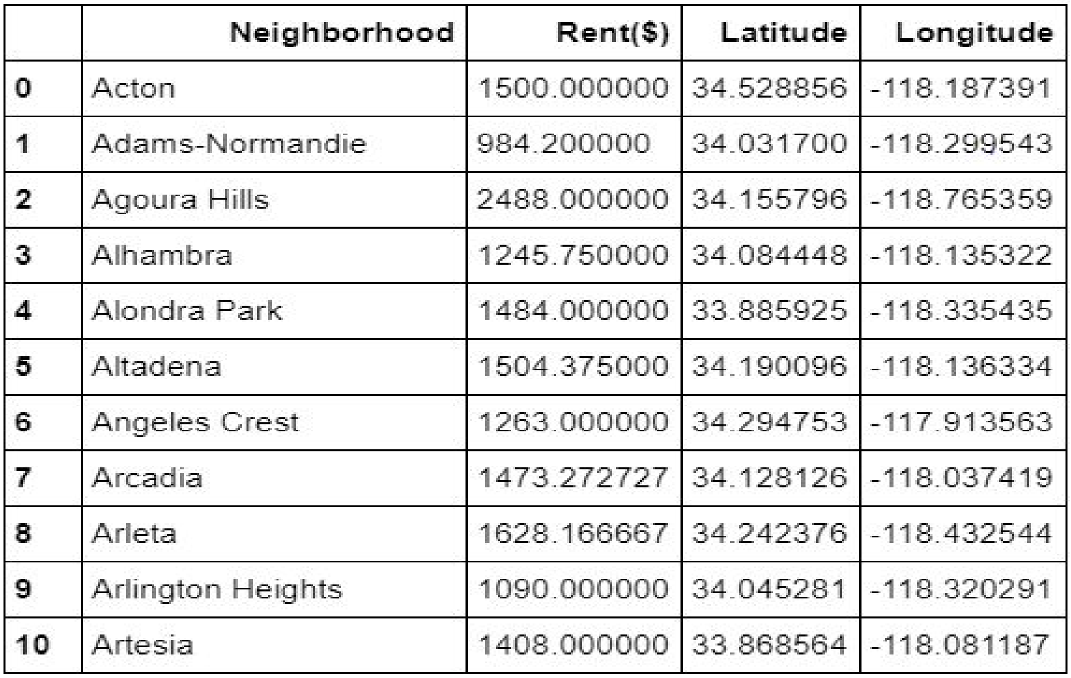
B. I used **FourSquare API** to get the most common venues for each neighbourhood of Los Angeles.

2.2 Data Cleaning and Feature Selection

There were several features or columns in the dataset mentioned above. First, I cleaned the dataset keeping only four features namely Neighborhood, Location, Rent($), Year. After that I extracted the data from year 2016 and used it for further processing. The location data was in a single column enclosed within parenthesis. I extracted the location coordinates from 'Location' feature and added two new columns namely Latitude and Longitude. I removed the 'Year' column since all the rows are from the year 2016. Finally, the dataset contributed to total four features i.e Neighborhood, Rent($), Latitude, Longitude. The rest of the features were related to venues which I got it by using FourSquare API for each neighbourhood. I got atmost 100 top venues for each neighbourhood in Los Angeles and their category. The list of venues were used as features for clustering of neighbourhoods according to venue density.

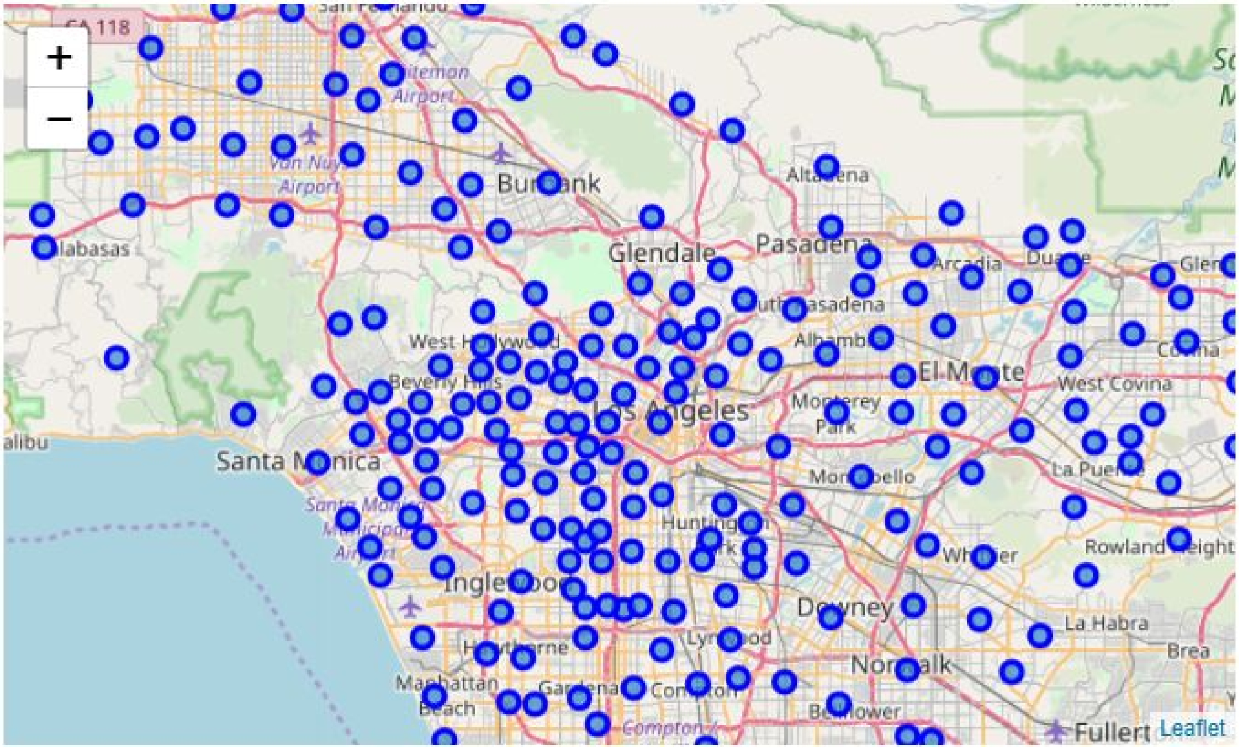
3. Methodology

Initially, the main table has four components Neighbourhood, Rent($), Latitude and Longitude.

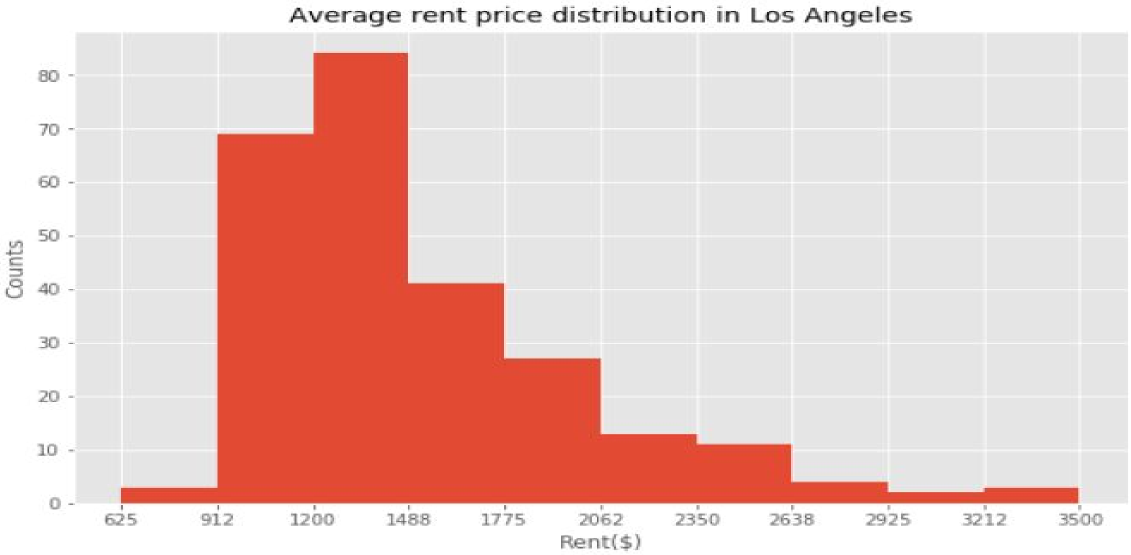


	Neighborhood	Rent($)	Latitude	Longitude
0	Acton	1500.000000	34.528856	-118.187391
1	Adams-Normandie	984.200000	34.031700	-118.299543
2	Agoura Hills	2488.000000	34.155796	-118.765359
3	Alhambra	1245.750000	34.084448	-118.135322
4	Alondra Park	1484.000000	33.885925	-118.335435
5	Altadena	1504.375000	34.190096	-118.136334
6	Angeles Crest	1263.000000	34.294753	-117.913563
7	Arcadia	1473.272727	34.128126	-118.037419
8	Arleta	1628.166667	34.242376	-118.432544
9	Arlington Heights	1090.000000	34.045281	-118.320291
10	Artesia	1408.000000	33.868564	-118.081187

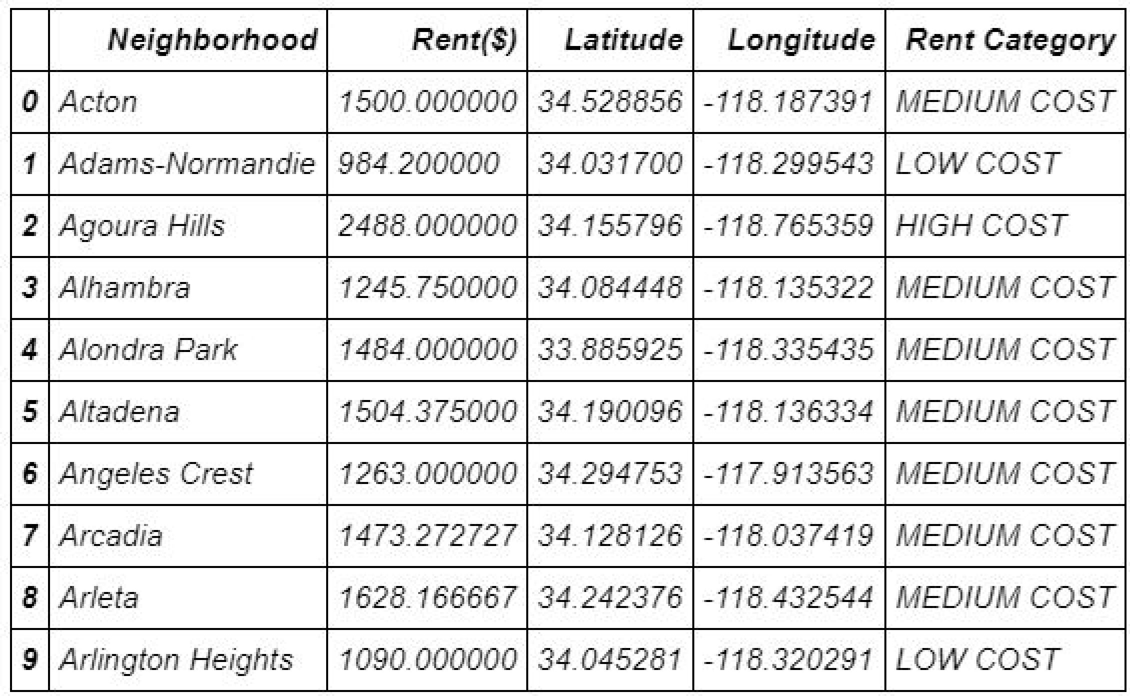
I used python **folium** library to visualize geographic details of Los Angeles and its neighbourhoods and I created a map of Los Angeles with neighbourhoods superimposed on top. I used latitude and longitude values to get the visual as below:



We can also examine that what is the frequency of average housing rent prices in different ranges. Thus, histogram can help to visualization:

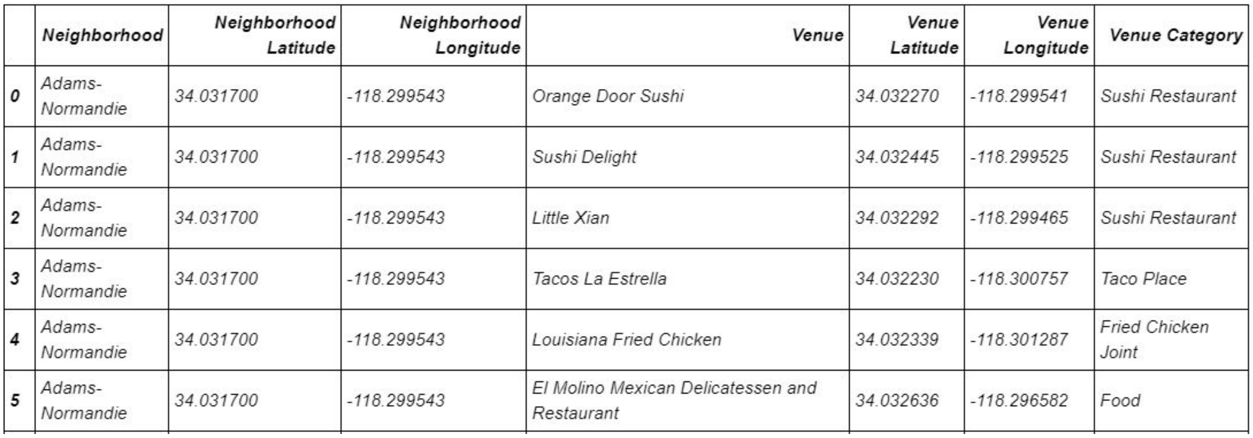


We can see that the minimum rent price in Los Angeles is $625 and maximum rent price is $3500. I have divided the rent prices into three categories high, medium and low. This will allow more insights about a neighbourhood’s rent price standards in comparison to other neighbourhoods in Los Angeles.



	Neighborhood	Rent($)	Latitude	Longitude	Rent Category
0	Acton	1500.000000	34.528856	-118.187391	MEDIUM COST
1	Adams-Normandie	984.200000	34.031700	-118.299543	LOW COST
2	Agoura Hills	2488.000000	34.155796	-118.765359	HIGH COST
3	Alhambra	1245.750000	34.084448	-118.135322	MEDIUM COST
4	Alondra Park	1484.000000	33.885925	-118.335435	MEDIUM COST
5	Altadena	1504.375000	34.190096	-118.136334	MEDIUM COST
6	Angeles Crest	1263.000000	34.294753	-117.913563	MEDIUM COST
7	Arcadia	1473.272727	34.128126	-118.037419	MEDIUM COST
8	Arleta	1628.166667	34.242376	-118.432544	MEDIUM COST
9	Arlington Heights	1090.000000	34.045281	-118.320291	LOW COST

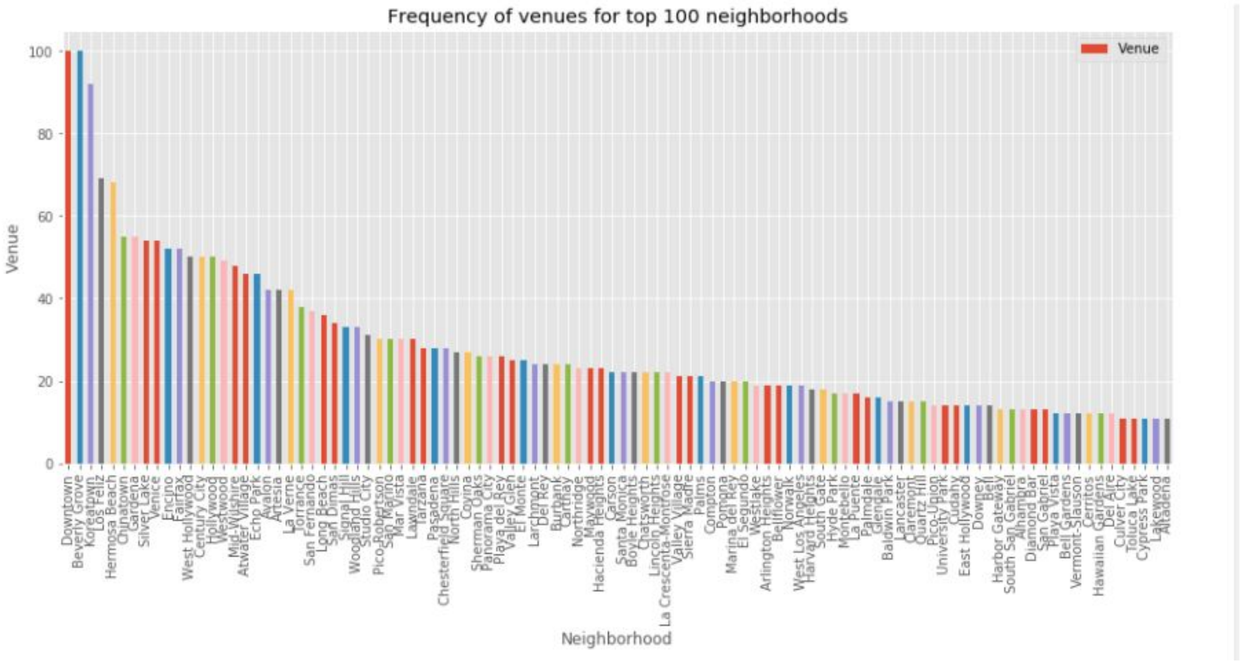
I utilized the Foursquare API to explore the neighbourhoods and segment them. I designed the limit as 100 venues and the radius 500 meter for each neighbourhood from their given latitude and longitude informations. In summary, 3478 venues were returned by Foursquare. Here is a merged table of neighbourhoods and venues.



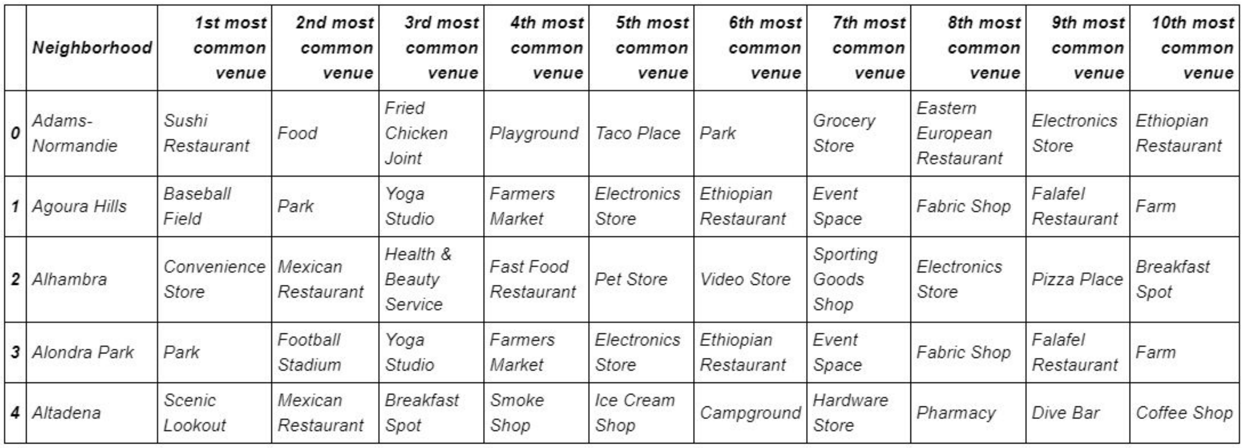
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adams-Normandie	34.031700	-118.299543	Orange Door Sushi	34.032270	-118.299541	Sushi Restaurant
1	Adams-Normandie	34.031700	-118.299543	Sushi Delight	34.032445	-118.299525	Sushi Restaurant
2	Adams-Normandie	34.031700	-118.299543	Little Xian	34.032292	-118.299465	Sushi Restaurant
3	Adams-Normandie	34.031700	-118.299543	Tacos La Estrella	34.032230	-118.300757	Taco Place
4	Adams-Normandie	34.031700	-118.299543	Louisiana Fried Chicken	34.032339	-118.301287	Fried Chicken Joint
5	Adams-Normandie	34.031700	-118.299543	El Molino Mexican Delicatessen and Restaurant	34.032636	-118.296582	Food

We can see that Downtown and Beverly Grove reached the limit of 100 venues. On the other hand; Lakewood, University Park, San Gabriel Culver city, West Los Angeles etc are below 20 venues in our given coordinates with Latitude and Longitude, in below graph.

The result doesn't mean that inquiry run all the possible results in neighbourhoods. Actually, it depends on given Latitude and Longitude informations and here is we just run single Latitude and Longitude pair for each neighbourhood. We can increase the possibilities by increasing the radius for each neighbourhood.



In summary of this graph 321 unique categories were returned by Foursquare, then I created a table which shows list of top 10 venue category for each neighbourhood in below table.

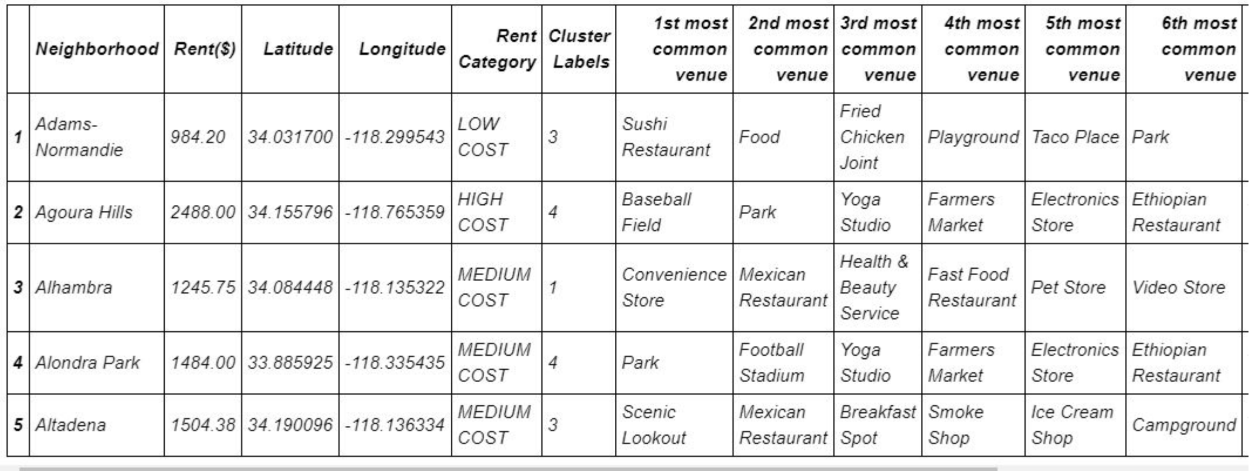


	Neighborhood	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue	6th most common venue	7th most common venue	8th most common venue	9th most common venue	10th most common venue
0	Adams-Normandie	Sushi Restaurant	Food	Fried Chicken Joint	Playground	Taco Place	Park	Grocery Store	Eastern European Restaurant	Electronics Store	Ethiopian Restaurant
1	Agoura Hills	Baseball Field	Park	Yoga Studio	Farmers Market	Electronics Store	Ethiopian Restaurant	Event Space	Fabric Shop	Falafel Restaurant	Farm
2	Alhambra	Convenience Store	Mexican Restaurant	Health & Beauty Service	Fast Food Restaurant	Pet Store	Video Store	Sporting Goods Shop	Electronics Store	Pizza Place	Breakfast Spot
3	Alondra Park	Park	Football Stadium	Yoga Studio	Farmers Market	Electronics Store	Ethiopian Restaurant	Event Space	Fabric Shop	Falafel Restaurant	Farm
4	Altadena	Scenic Lookout	Mexican Restaurant	Breakfast Spot	Smoke Shop	Ice Cream Shop	Campground	Hardware Store	Pharmacy	Dive Bar	Coffee Shop

We have some common venue categories in neighbourhoods. In this reason I used unsupervised learning K-means algorithm to cluster the neighbourhoods. K-Means algorithm is one of the most common cluster method of unsupervised learning.

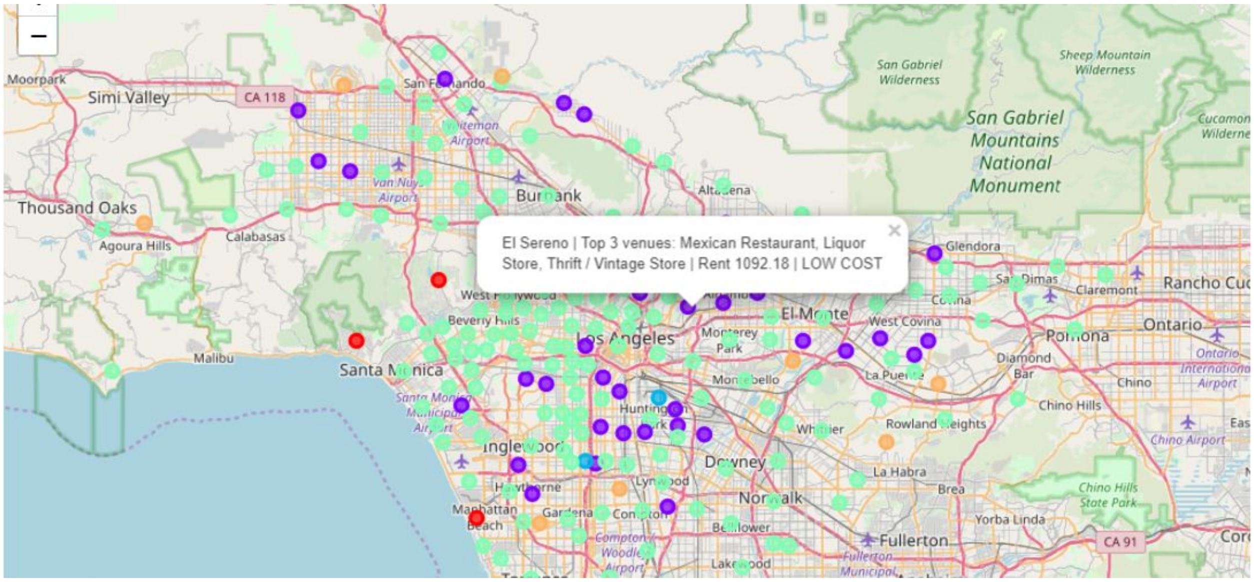
4. Results

Here is my merged table with cluster labels for each neighbourhood.



	Neighborhood	Rent($)	Latitude	Longitude	Rent Category	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
1	Adams-Normandie	984.20	34.031700	-118.299543	LOW COST	3	Sushi Restaurant	Food	Fried Chicken Joint	Playground	Taco Place	Park
2	Agoura Hills	2488.00	34.155796	-118.765359	HIGH COST	4	Baseball Field	Park	Yoga Studio	Farmers Market	Electronics Store	Ethiopian Restaurant
3	Alhambra	1245.75	34.084448	-118.135322	MEDIUM COST	1	Convenience Store	Mexican Restaurant	Health & Beauty Service	Fast Food Restaurant	Pet Store	Video Store
4	Alondra Park	1484.00	33.885925	-118.335435	MEDIUM COST	4	Park	Football Stadium	Yoga Studio	Farmers Market	Electronics Store	Ethiopian Restaurant
5	Altadena	1504.38	34.190096	-118.136334	MEDIUM COST	3	Scenic Lookout	Mexican Restaurant	Breakfast Spot	Smoke Shop	Ice Cream Shop	Campground

You can now see above Neighbourhood name, rent, top 3 venues, cluster labels for each neighbourhood in Los Angeles. You can also see a clustered map of Los Angeles neighbourhoods below.



El Sereno | Top 3 venues: Mexican Restaurant, Liquor Store, Thrift / Vintage Store | Rent 1092.18 | LOW COST

The different colored circles signifies different clusters in the map. When you click on any colored circle in the map it provides the following information:

1. Neighbourhood name
2. Top 3 venues
3. Rent price
4. Rent Category

5. Discussions

There are various neighbourhoods in Los Angeles and they are already crowded with venues like restaurants, recreation places, mall etc . Apart from this, the rent price varies in each and every neighbourhood. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results for this city.

I used K means algorithm for clustering the neighbourhoods of Los Angeles. I had a dataset of 257 neighbourhoods and venues limit of 100 for a radius of 500 metres. For more detailed list of venues we can increase the limit or the radius for each neighbourhood. I completed the analysis by visualizing the data and clustering information on the map of Los Angeles.

6. Conclusion

Los Angeles is a big city with a huge population density and enormous opportunity to setup a business. One among many problems lies in identifying the place where to setup your business. This study will provide a platform for businessmen or companies to achieve better outcomes in decision making and certainly make their lives a lot easier.