Data Science Capstone Project Report

The Battle of the Neighbourhoods - LA Edition

- Ishaan Vasant

Introduction: Business Problem

Los Angeles, often known by its initials LA, is the most populous city in California and the second-most populous city in the United States. Los Angeles is the cultural, financial, and commercial center of Southern California. The city is known for its Mediterranean climate, ethnic diversity, Hollywood, the entertainment industry, and its sprawling metropolis.

Having lived in LA for over a year now, I can confirm that Los Angeles is also one of the most amazing places to eat, thanks to an incredible variety of international cuisines and some of the most talented chefs in the world. It is a multicultural city, with the biggest communities of several nationalities of people outside of their homelands. They bring their cuisines with them. LA's thriving economy, access to ingredients, and great seasonal produce makes it an ideal place for restaurants to flourish.

The objective of this project is to identify the best potential neighbourhoods where a restaurant can be set up. An international YouGov study of more than 25,000 people in 24 countries found that pizza and pasta are among the most popular foods in the world, as Italian cuisine beats all comers. According to their analysis, 88% of people surveyed in America liked Italian food. Keeping this in mind, the focus of this capstone would be Italian restaurants. Therefore, the analysis and results of this project would interest stakeholders who are interested in opening an Italian restaurant in Los Angeles.

Since there are lots of restaurants in LA, neighbourhoods that are not already crowded with restaurants would be shortlisted. The next filter would be neighbourhoods with the least number of Italian restaurants in its vicinity. Neighbourhoods that are as close to the city center as possible would be preferred. Neighbourhood rent is another factor that would be taken into consideration.

Data

Based on the criteria specified above, the factors that will influence the final decision are: -

- Number of existing restaurants in the neighborhood (any type of restaurant)
- Number of and distance to Italian restaurants in the neighborhood
- Distance of neighborhood from city center
- Average neighbourhood rent

The following data sources will be needed to extract/generate the required information: -

- List of all neighbourhoods in LA - https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_of_Los_Angeles
- Coordinates of all neighbourhoods and venues - GeoPy Nominatim geocoding
- Number of restaurants and their type and location in every neighbourhood - Foursquare API -https://developer.foursquare.com
- LA rent data - https://www.rentcafe.com/average-rent-market-trends/us/ca/los-angeles/

Methodology

In this project, the first step will be to collect data on the neighbourhoods of Los Angeles from the internet. There are no relevant datasets available for this and therefore, data will need to be scraped from a webpage. The location coordinates of each neighbourhood will then be obtained with the help of GeoPy Nominatim geolocator and appended to the neighbourhood data. Using this data, a folium map of the Los Angeles neighbourhoods will be created.

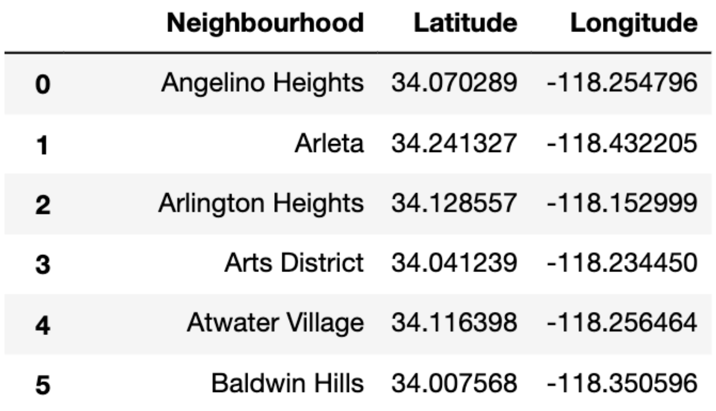
The second step will be to explore each of neighbourhoods and their venues using Foursquare location data. The venues of the neighbourhoods will be analyzed in detail and patterns will be discovered. This discovery of patterns will be carried out by grouping the neighbourhoods using k-means clustering. Following this, each cluster will be examined, and a decision will be made regarding which cluster fits the shareholder's requirements. The factor that will determine this is the frequency of occurrence of restaurants and other food venues within the cluster.

Once a cluster is picked, the neighbourhoods in that cluster will be investigated with regards to the number of Italian restaurants in its vicinity. The ones that fit the requirements will be further explored and shortlisted based on how small their respective distances to the center or Los Angeles are. Finally, if multiple neighbourhoods fit these conditions, Los Angeles rent data can be used to influence the shareholder's decision.

The results of the analysis will highlight potential neighbourhoods where an Italian restaurant may be opened based on geographical location and proximity to competitors. This will only serve as a starting point since there are a lot of other factors that influence such a decision.

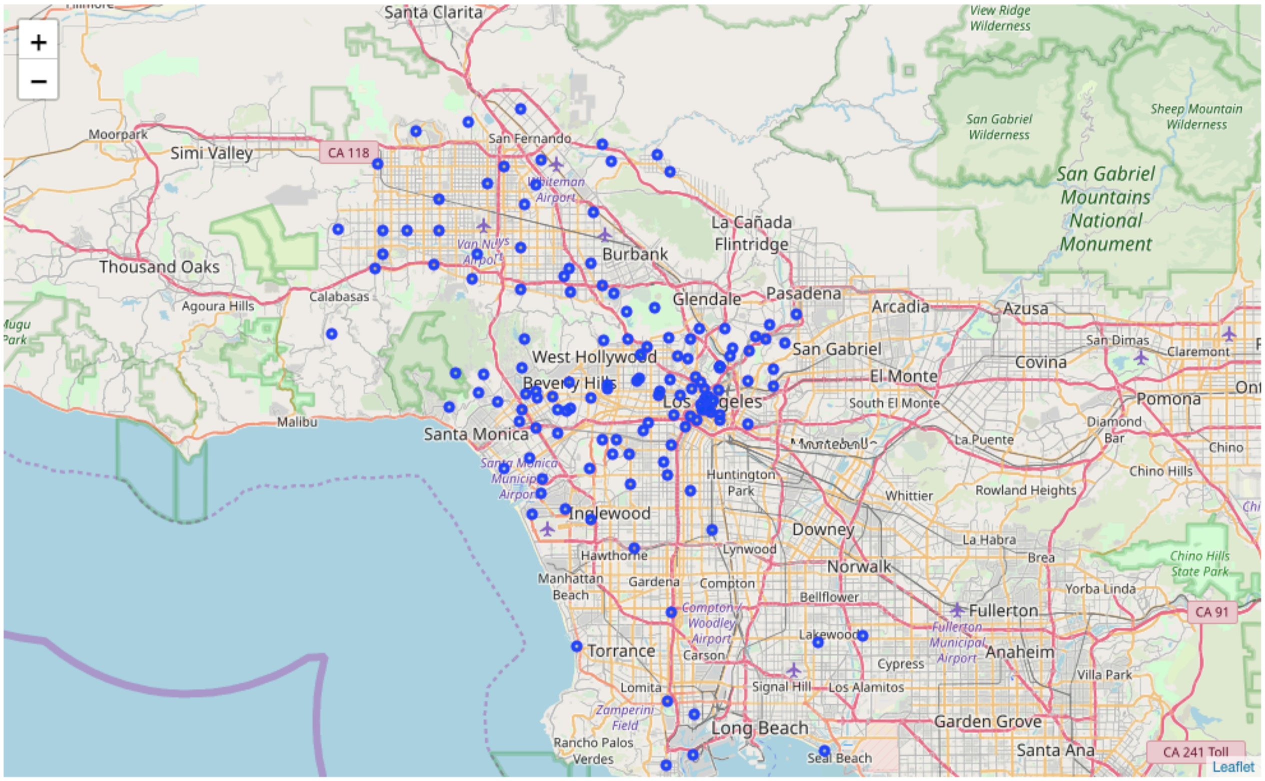
Analysis

- The list of all neighbourhoods in LA is obtained by scraping the webpage specified in the Data section. The data on the webpage is in the form of a list and not a table. Therefore, the data is gathered by searching for all list items and then using a particular characteristic that groups the required items. Using GeoPy Nominatim geolocator, the neighbourhood coordinates are obtained, giving rise to a complete neighbourhood data frame: -

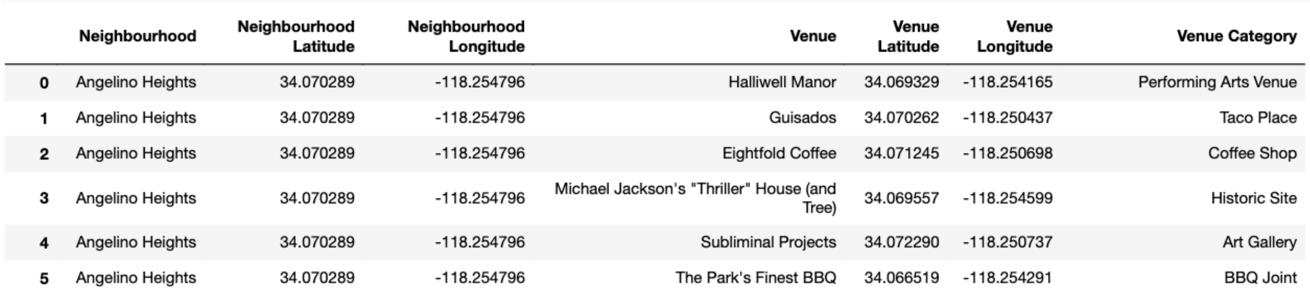


	Neighbourhood	Latitude	Longitude
0	Angelino Heights	34.070289	-118.254796
1	Arleta	34.241327	-118.432205
2	Arlington Heights	34.128557	-118.152999
3	Arts District	34.041239	-118.234450
4	Atwater Village	34.116398	-118.256464
5	Baldwin Hills	34.007568	-118.350596

- A folium map of LA is created with the neighbourhoods superimposed on top: -

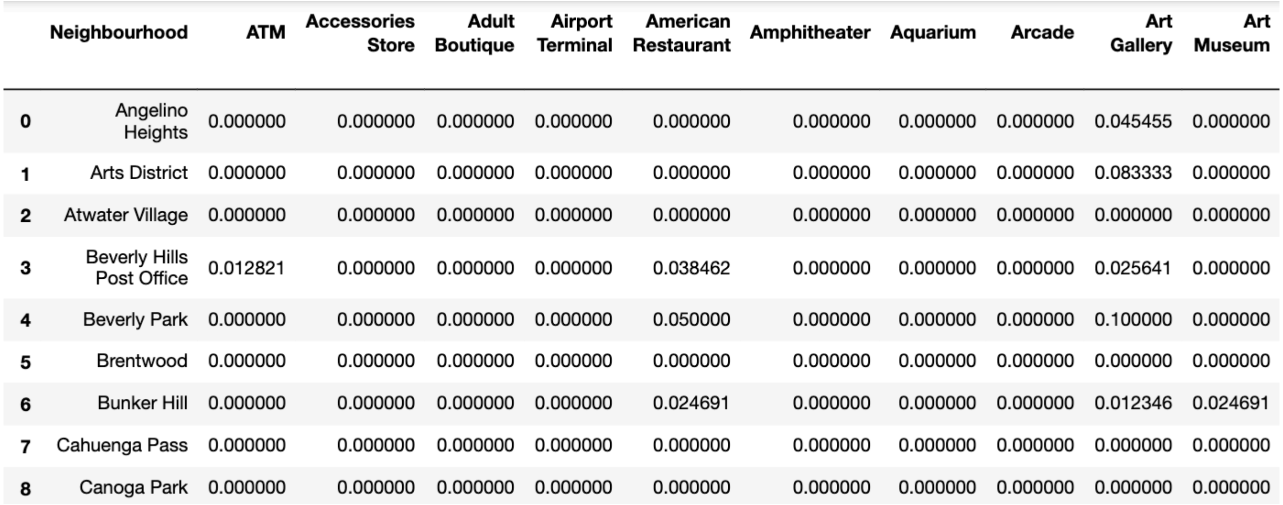


- Foursquare location data is used to obtain nearby venues of all neighbourhoods and load the data into a data frame. It makes sense to set up a restaurant in one of the more popular neighbourhoods so that the restaurant attracts the attention of a lot more people. Therefore, a list of all the popular neighbourhoods i.e. the neighbourhoods with 10 or more venues is obtained. The venues data frame is updated to include only the venues which are in popular neighbourhoods: -



	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Angelino Heights	34.070289	-118.254796	Halliwell Manor	34.069329	-118.254165	Performing Arts Venue
1	Angelino Heights	34.070289	-118.254796	Guisados	34.070262	-118.250437	Taco Place
2	Angelino Heights	34.070289	-118.254796	Eightfold Coffee	34.071245	-118.250698	Coffee Shop
3	Angelino Heights	34.070289	-118.254796	Michael Jackson's "Thriller" House (and Tree)	34.069557	-118.254599	Historic Site
4	Angelino Heights	34.070289	-118.254796	Subliminal Projects	34.072290	-118.250737	Art Gallery
5	Angelino Heights	34.070289	-118.254796	The Park's Finest BBQ	34.066519	-118.254291	BBQ Joint

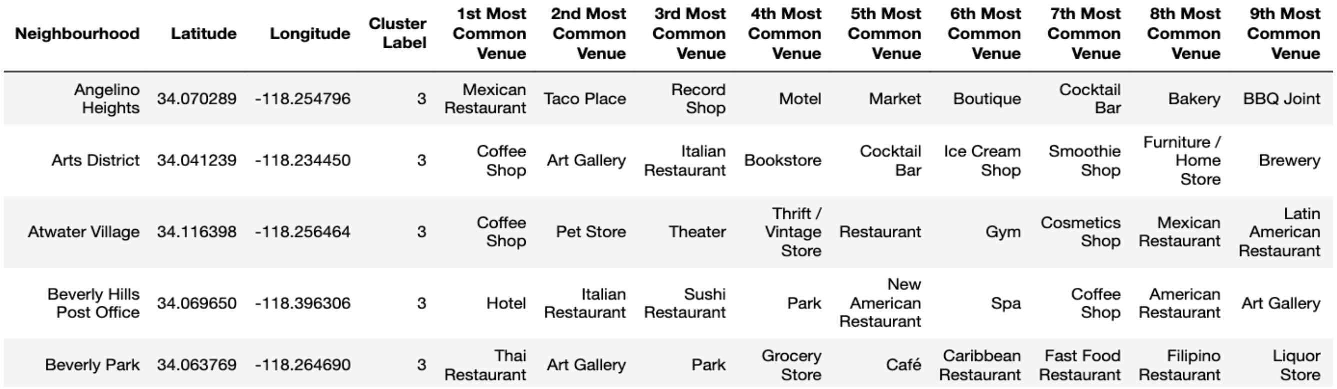
- The rows in the above data frame are grouped by neighbourhood, taking the mean of the frequency of occurrence of each category: -



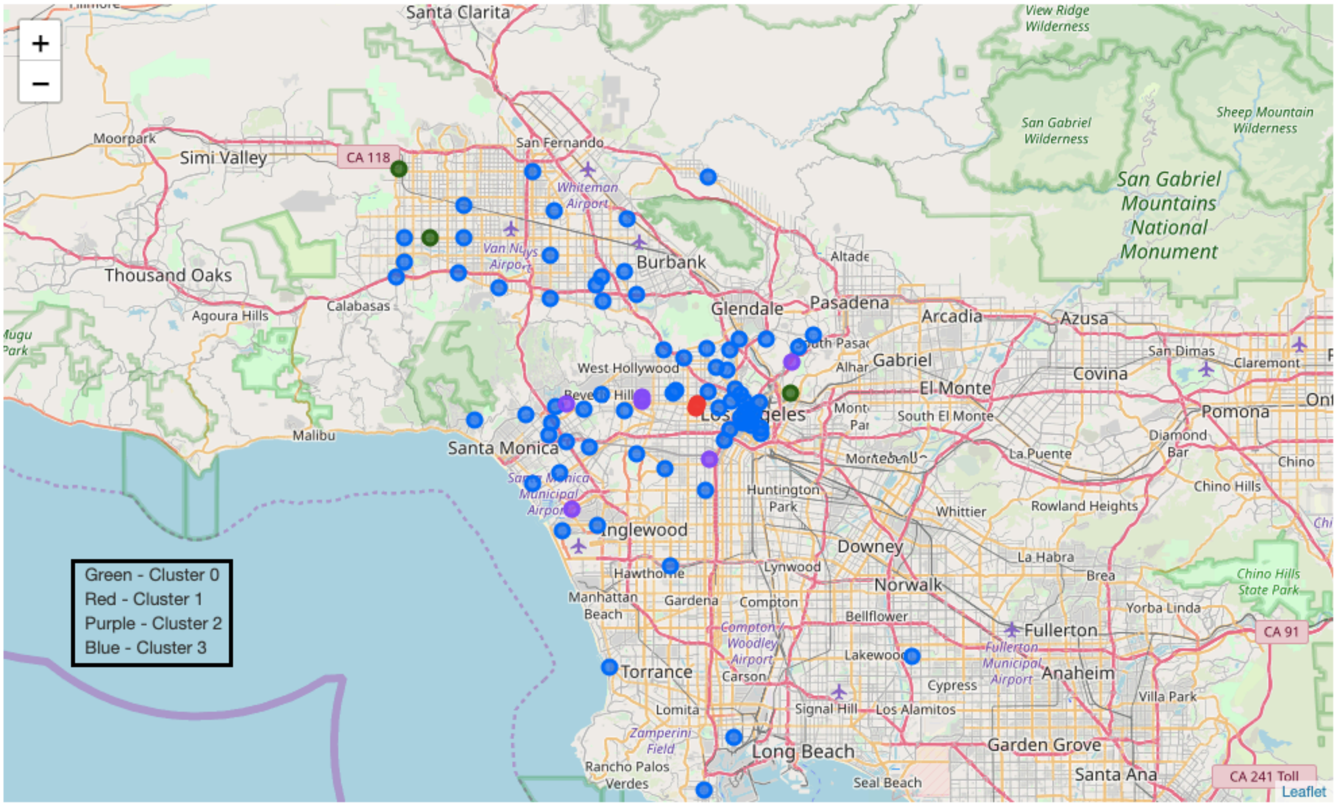
	Neighbourhood	ATM	Accessories Store	Adult Boutique	Airport Terminal	American Restaurant	Amphitheater	Aquarium	Arcade	Art Gallery	Art Museum
0	Angelino Heights	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.045455	0.000000
1	Arts District	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.083333	0.000000
2	Atwater Village	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Beverly Hills Post Office	0.012821	0.000000	0.000000	0.000000	0.038462	0.000000	0.000000	0.000000	0.025641	0.000000
4	Beverly Park	0.000000	0.000000	0.000000	0.000000	0.050000	0.000000	0.000000	0.000000	0.100000	0.000000
5	Brentwood	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	Bunker Hill	0.000000	0.000000	0.000000	0.000000	0.024691	0.000000	0.000000	0.000000	0.012346	0.024691
7	Cahuenga Pass	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	Canoga Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

- The first step in clustering the neighbourhoods is to determine the optimal value of K for the dataset using the Silhouette Coefficient Method. The Silhouette Coefficient is found to be the highest when the number of clusters is 4. Therefore, the neighbourhoods shall be grouped into 4 clusters (k=4) using *k*-means clustering.

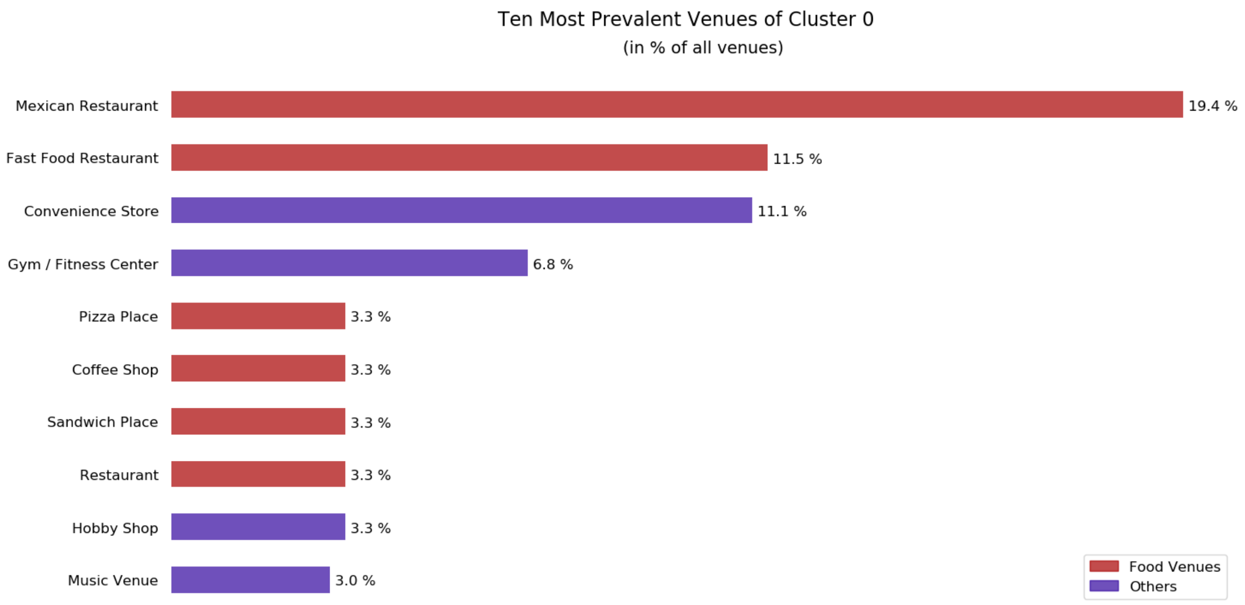
- A new data frame that includes the cluster, as well as the top 10 venues for each neighbourhood, is created. The resulting neighbourhood clusters are also visualized on the map: -



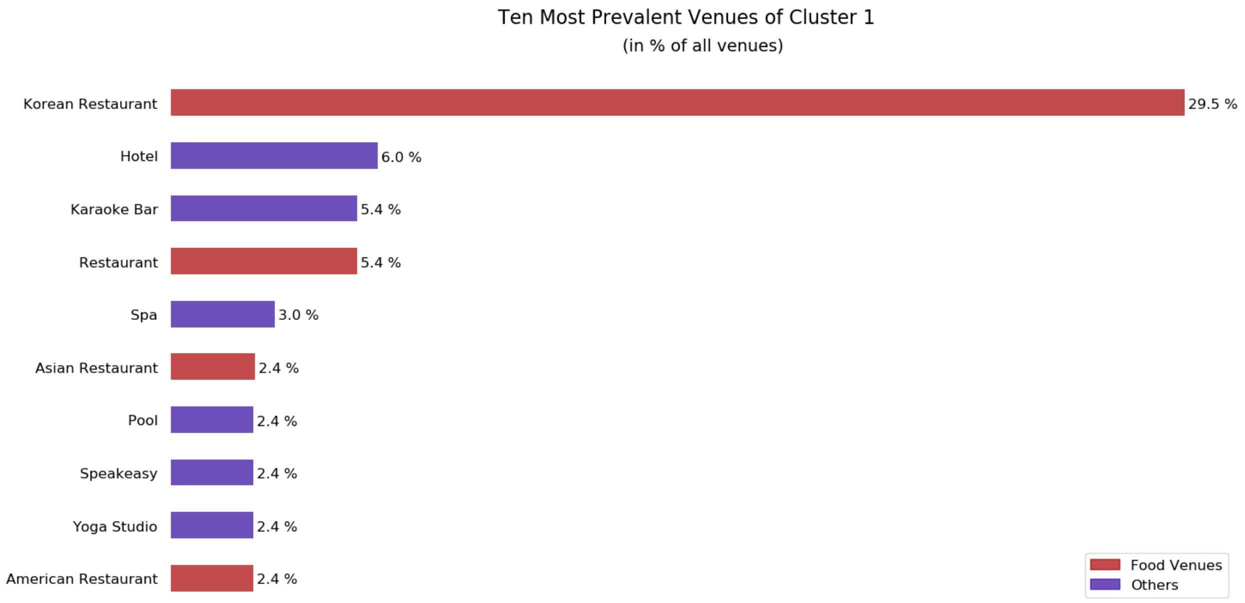
Neighbourhood	Latitude	Longitude	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
Angelino Heights	34.070289	-118.254796	3	Mexican Restaurant	Taco Place	Record Shop	Motel	Market	Boutique	Cocktail Bar	Bakery	BBQ Joint
Arts District	34.041239	-118.234450	3	Coffee Shop	Art Gallery	Italian Restaurant	Bookstore	Cocktail Bar	Ice Cream Shop	Smoothie Shop	Furniture / Home Store	Brewery
Atwater Village	34.116398	-118.256464	3	Coffee Shop	Pet Store	Theater	Thrift / Vintage Store	Restaurant	Gym	Cosmetics Shop	Mexican Restaurant	Latin American Restaurant
Beverly Hills Post Office	34.069650	-118.396306	3	Hotel	Italian Restaurant	Sushi Restaurant	Park	New American Restaurant	Spa	Coffee Shop	American Restaurant	Art Gallery
Beverly Park	34.063769	-118.264690	3	Thai Restaurant	Art Gallery	Park	Grocery Store	Café	Caribbean Restaurant	Fast Food Restaurant	Filipino Restaurant	Liquor Store



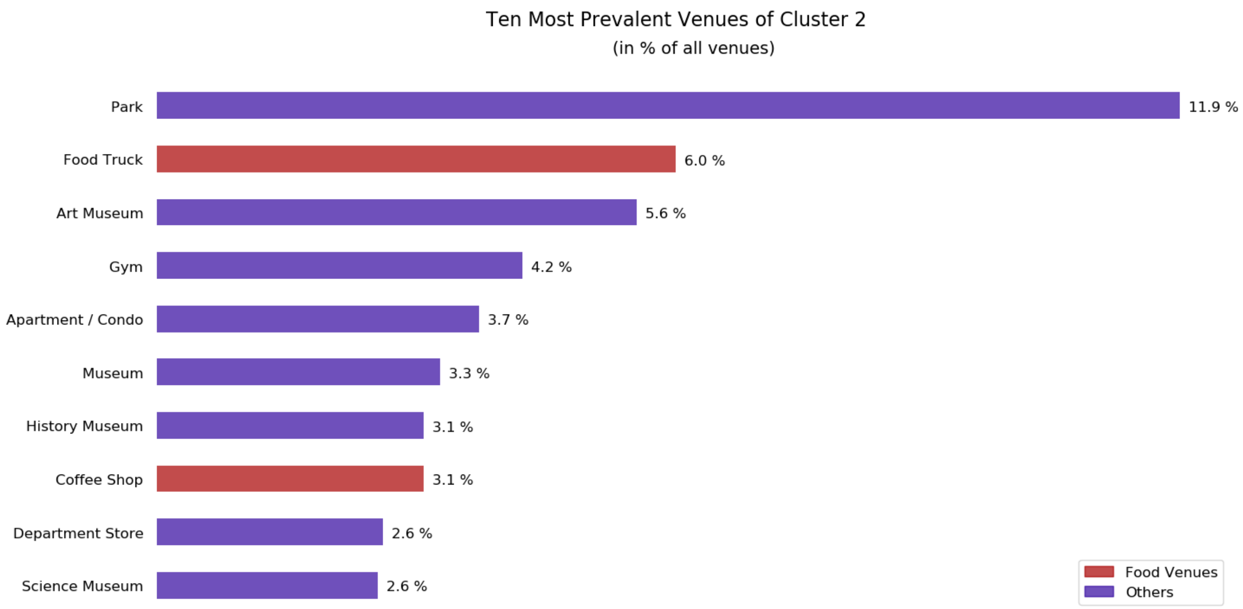
- A horizontal bar plot visualizing the top 10 venues for each cluster is generated, highlighting the food venues: -



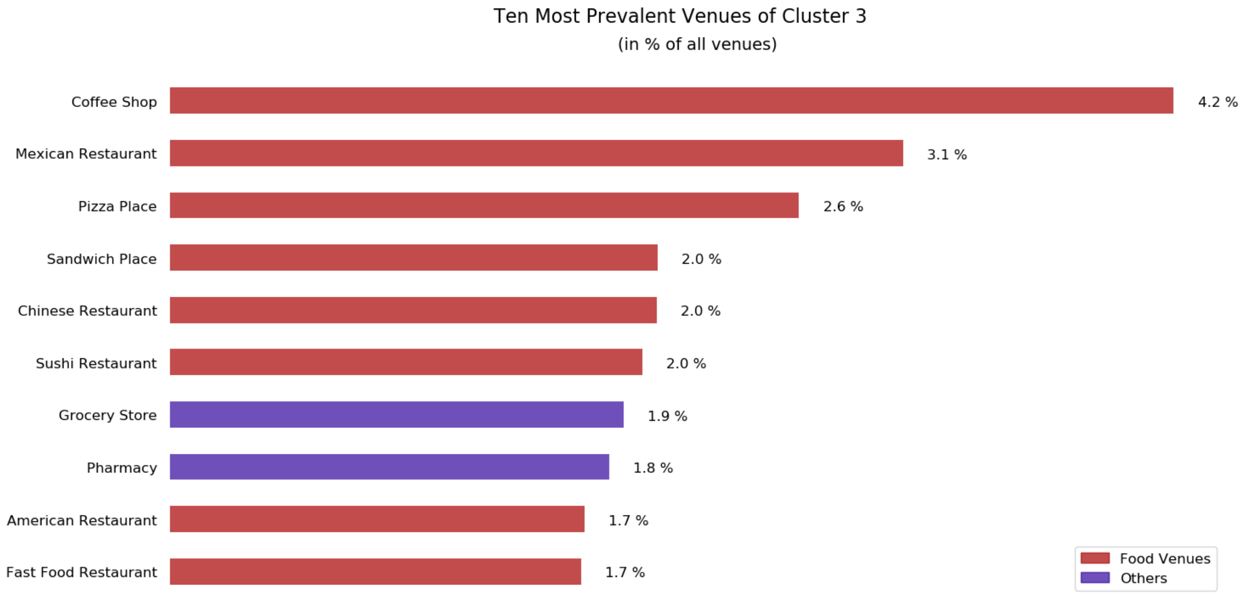
- There are 6 food venues in the top 10 venues of Cluster 0 with Mexican Restaurants making up nearly 20% of all venues. These facts indicate that Cluster 0 would not be the best one to explore further in terms of setting up a new restaurant.



- There are 4 food venues in the top 10 venues of Cluster 1 with Korean Restaurants making up a huge majority (nearly 30%) of all venues. This is unsurprising as Cluster 1 consists of only two neighbourhoods, one being Koreatown and the other (Mid-Wilshire) also having a lot of Korean Restaurants. While there are only 4 food venues in the top 10, the complete dominance of Korean Restaurants in the area indicates the fact that Cluster 1 need not be looked into any further.

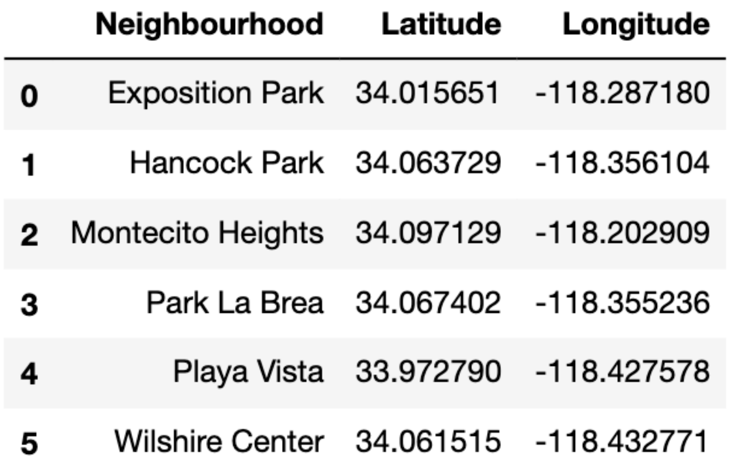


- There are only 2 food venues in the top 10 venues of Cluster 2. To add to that, the two venues are Food Trucks and Coffee Shops as opposed to proper restaurants. There are a lot of public venues in this cluster - venues that see high footfall such as parks, museums, gyms, and department stores. The presence of condominium complexes in this list also suggests that the population per square unit of these neighbourhoods is high. All of these observations point in the direction of Cluster 2 being nominated as the cluster to explore further. Having said that, the decision to explore Cluster 2 can only be confirmed after examining Cluster 3: -



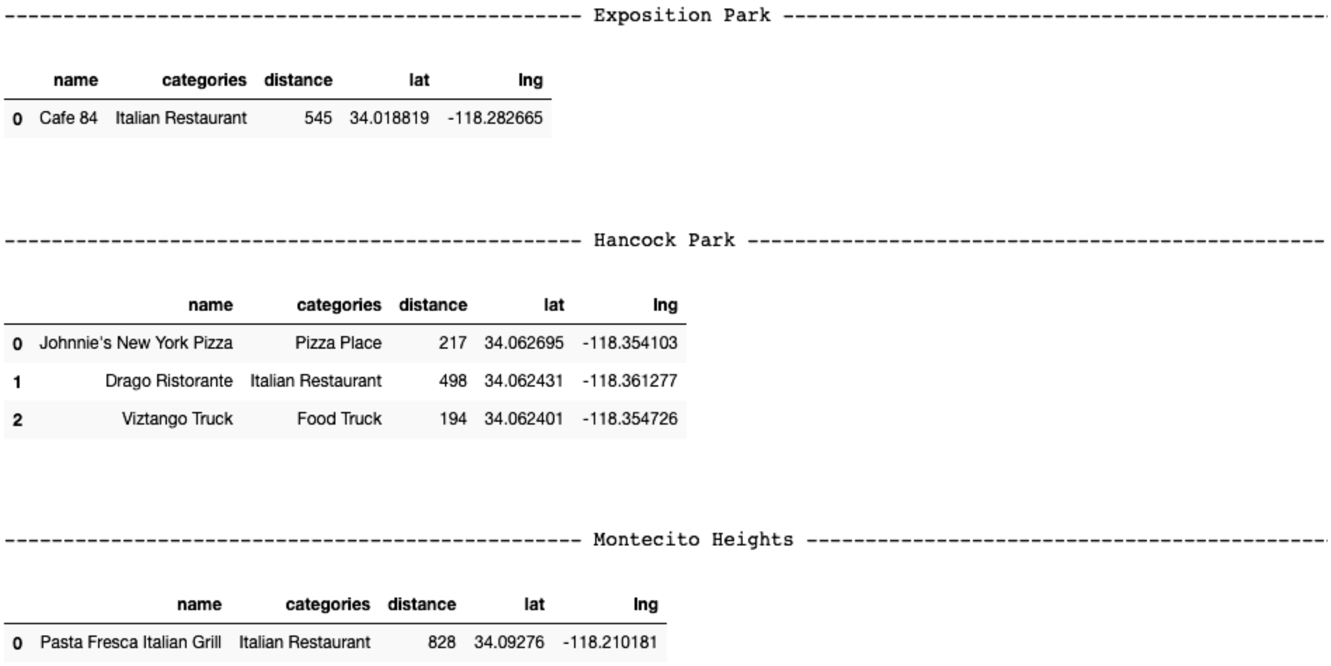
- There are 8 food venues in the top 10 venues of Cluster 3 which is a huge percentage. Except for the number 1 venue (Coffee Shops), all other food venues are proper restaurants. This indicates that the neighbourhoods in Cluster 3 are saturated with restaurants already and need not be considered when opening a new restaurant. It is now safe to confirm the decision of investigating **Cluster 2** further and eliminating all other clusters.

- The neighbourhoods in Cluster 2 along with their coordinates are displayed below: -



	Neighbourhood	Latitude	Longitude
0	Exposition Park	34.015651	-118.287180
1	Hancock Park	34.063729	-118.356104
2	Montecito Heights	34.097129	-118.202909
3	Park La Brea	34.067402	-118.355236
4	Playa Vista	33.972790	-118.427578
5	Wilshire Center	34.061515	-118.432771

- The closest Italian restaurants from each neighbourhood in Cluster 2 are displayed below, along with the corresponding distances: -



```
-------------------------------------------------- Exposition Park ----------------------------------------------
```

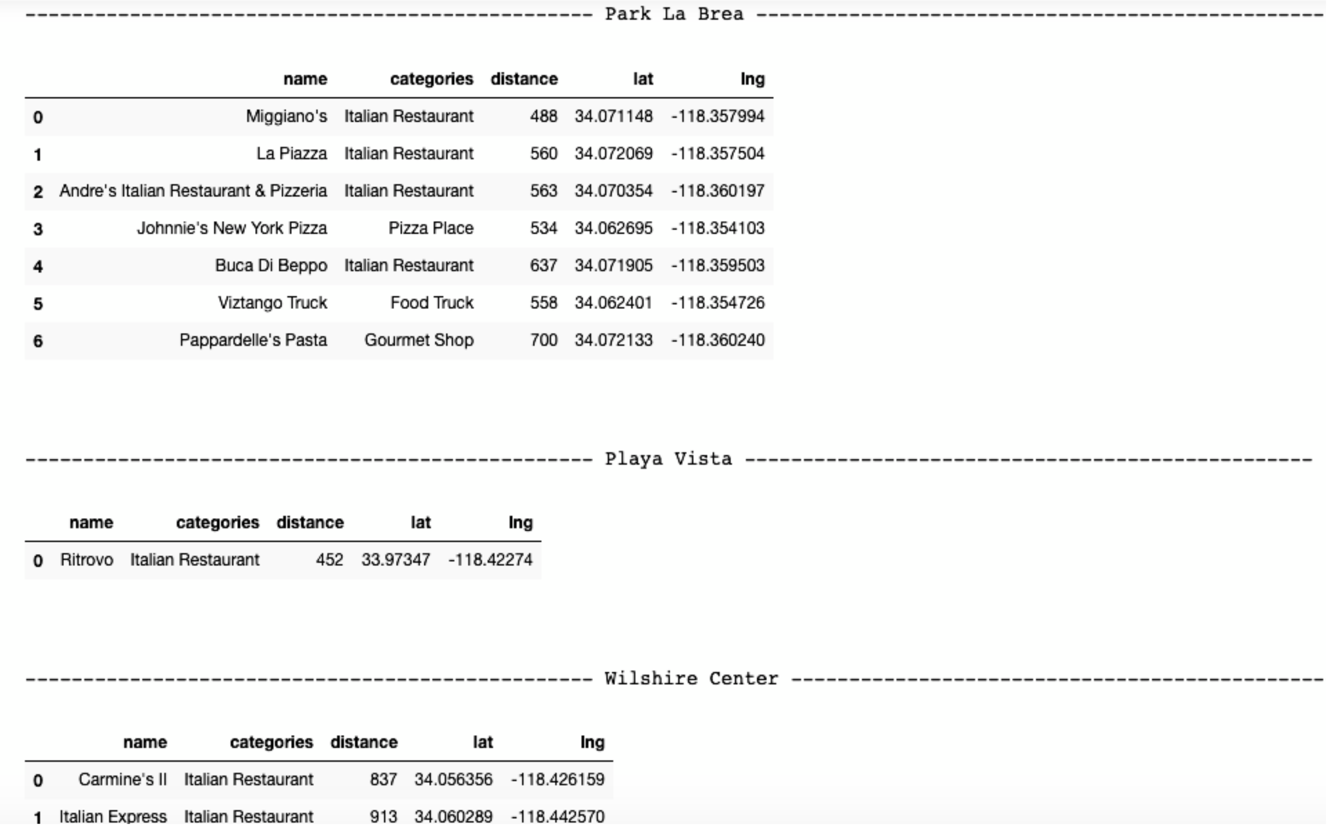
	name	categories	distance	lat	lng
0	Cafe 84	Italian Restaurant	545	34.018819	-118.282665

```
-------------------------------------------------- Hancock Park --------------------------------------------------
```

	name	categories	distance	lat	lng
0	Johnnie's New York Pizza	Pizza Place	217	34.062695	-118.354103
1	Drago Ristorante	Italian Restaurant	498	34.062431	-118.361277
2	Viztango Truck	Food Truck	194	34.062401	-118.354726

```
-------------------------------------------------- Montecito Heights --------------------------------------------
```

	name	categories	distance	lat	lng
0	Pasta Fresca Italian Grill	Italian Restaurant	828	34.09276	-118.210181



```
----------------------------------------------- Park La Brea -----------------------------------------------
```

	name	categories	distance	lat	lng
0	Miggiano's	Italian Restaurant	488	34.071148	-118.357994
1	La Piazza	Italian Restaurant	560	34.072069	-118.357504
2	Andre's Italian Restaurant & Pizzeria	Italian Restaurant	563	34.070354	-118.360197
3	Johnnie's New York Pizza	Pizza Place	534	34.062695	-118.354103
4	Buca Di Beppo	Italian Restaurant	637	34.071905	-118.359503
5	Viztango Truck	Food Truck	558	34.062401	-118.354726
6	Pappardelle's Pasta	Gourmet Shop	700	34.072133	-118.360240

```
----------------------------------------------- Playa Vista -----------------------------------------------
```

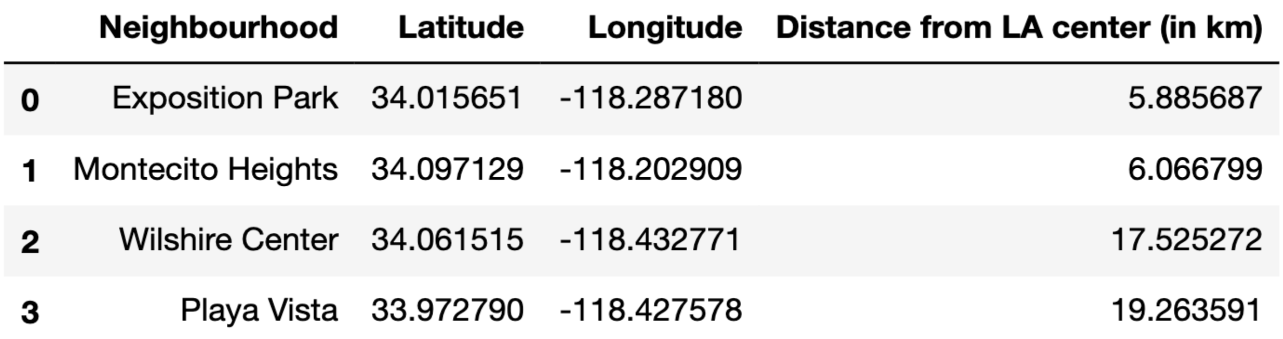
	name	categories	distance	lat	lng
0	Ritrovo	Italian Restaurant	452	33.97347	-118.42274

```
----------------------------------------------- Wilshire Center -----------------------------------------------
```

	name	categories	distance	lat	lng
0	Carmine's II	Italian Restaurant	837	34.056356	-118.426159
1	Italian Express	Italian Restaurant	913	34.060289	-118.442570

- From the data frames above, it is observed that Park La Brea has 7 Italian Restaurants within 700 meters from its center. Hancock Park has fewer (3) but two of them are less than 250 meters away from its center. This indicates that Park La Brea and Hancock Park would not be suitable neighbourhoods to open an Italian Restaurant and can therefore be eliminated.

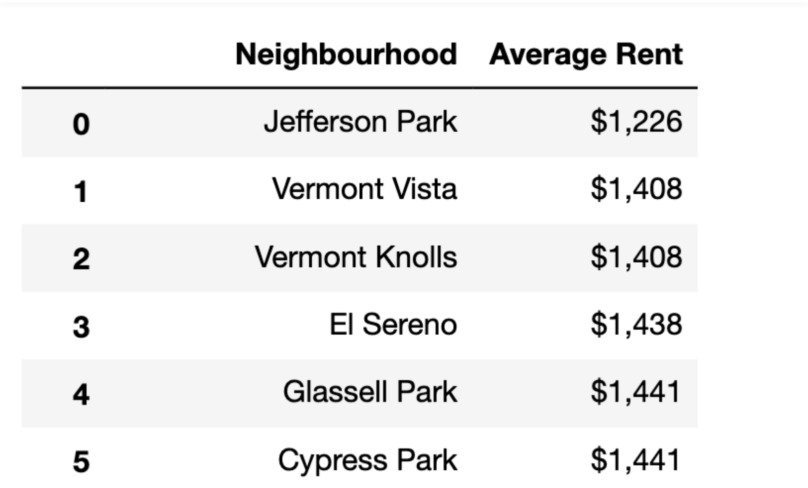
- The distance of each of the remaining neighbourhoods from the center of LA is computed and added as a column to the existing data frame: -



	Neighbourhood	Latitude	Longitude	Distance from LA center (in km)
0	Exposition Park	34.015651	-118.287180	5.885687
1	Montecito Heights	34.097129	-118.202909	6.066799
2	Wilshire Center	34.061515	-118.432771	17.525272
3	Playa Vista	33.972790	-118.427578	19.263591

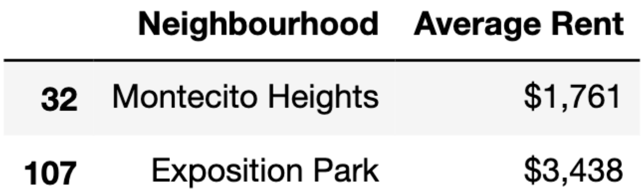
- It is clear from the data frame above that Exposition Park (~6km) and Montecito Heights (~6km) are much closer to the center of Los Angeles than Wilshire Center (~17.5km) and Playa Vista (~19km). Since the distance from LA center is a criterion in choosing the optimal neighbourhood, Wilshire Center and Playa Vista would not be appropriate choices.

- Following this, the list of average rent of all neighbourhoods in LA is obtained by scraping the relevant webpage: -



	Neighbourhood	Average Rent
0	Jefferson Park	$1,226
1	Vermont Vista	$1,408
2	Vermont Knolls	$1,408
3	El Sereno	$1,438
4	Glassell Park	$1,441
5	Cypress Park	$1,441

- The 2 neighbourhoods in question can be identified from the table and their average rents displayed: -



	Neighbourhood	Average Rent
32	Montecito Heights	$1,761
107	Exposition Park	$3,438

- The average rent in Exposition Park is nearly two times the average rent in Montecito Heights. This means that Exposition Park is a significantly more expensive neighbourhood. The implications of this observation are discussed in the next section.

Results and Discussion

At the beginning of the analysis, the data frame of Los Angeles neighbourhoods was trimmed to include only the ones that had 10 or more venues. This decision was taken as it made sense to set up a restaurant in one of the more popular neighbourhoods, thereby attracting the attention of a lot more people.

When clustering the neighbourhoods, the optimal value of k (k=4) for the dataset was arrived at using the Silhouette Coefficient Method. As a consequence, all neighbourhoods were grouped into 4 clusters using k-means clustering. To examine the deterministic characteristics of each cluster, a data frame for each cluster was created that included their most frequently occurring venues in descending order. A horizontal bar plot was generated showing the top 10 venues for each cluster, highlighting the food venues. This helped in determining the optimal cluster for further analysis. All of the observations pointed in the direction of Cluster 2 being that cluster. It had only 2 food venues amongst the top 10 - food trucks and coffee shops - which were not full-fledged restaurants. The cluster also had apartment complexes and a lot of public venues which meant that the neighbourhoods in it see a lot of people.

The following step was to obtain and display the closest Italian restaurants from each neighbourhood in Cluster 2 and their corresponding distances. It was observed that Park La Brea has 7 Italian Restaurants within 700 meters from its center. Hancock Park had fewer (3) but two of them were less than 250 meters away from its center. This indicated that Park La Brea and Hancock Park would not be suitable neighbourhoods to open an Italian Restaurant in and were eliminated.

The next criterion was the distance of each of the remaining neighbourhoods from the center of the city. It was found that Exposition Park (~6km) and Montecito Heights (~6km) are much closer to the center of Los Angeles than Wilshire Center (~17.5km) and Playa Vista (~19km). Therefore, it was understood that Wilshire Center and Playa Vista would not be appropriate choices.

The table of average rent of all neighbourhoods in LA was obtained by scraping the relevant webpage. The two neighbourhoods that remained in contention were identified from the table and their average rents displayed. It was detected that the average rent in Exposition Park is nearly two times the average rent in Montecito Heights, implying that Exposition Park is a significantly more expensive neighbourhood. However, this does not automatically mean Montecito Heights is the better option. A factor to consider is the type of restaurant the shareholder is interested in setting up. If, for example, a high-end fine dining restaurant needs to be set up, a neighbourhood that has a low average rent would not work. The reason for this is that such a neighbourhood would generally be home to people with lower income and a high-end fine dining restaurant may not see a healthy influx of people. On the other hand, if a fast-casual/casual dining restaurant needs to be set up, a high-rent neighbourhood would not be ideal simply because the restaurant will not be able to afford the rented space. While average rent can point in the direction of the right neighbourhood, a final decision cannot be made without all the required information.

Conclusion

The objective of this project was to identify the best potential neighbourhoods in Los Angeles where an Italian restaurant can be set up. All the required neighbourhood data was either scraped off the internet or obtained using a geolocator. After the neighbourhoods were visualized on a folium map, their venues were explored using Foursquare location data. Based on the frequency of occurrences of different venue types, the neighbourhoods were divided into four groups with the help of k-means clustering. The clusters were examined and the best one in which a restaurant could be set up was chosen. The neighbourhoods were filtered further based on proximity to existing Italian restaurants and distance from the center of the city. The analysis brought the number of contenders down to two neighbourhoods - Exposition Park and Montecito Heights. Average neighbourhood rent data was called upon and while it provided interesting insights, it could not influence the decision only with the information at hand. As touched upon earlier, the results of the analysis highlight potential neighbourhoods where an Italian restaurant may be opened solely based on geographical location and proximity to competitors. This will only serve as a starting point in the overall investigation since there are a lot of other factors - availability of commercial spaces, the appeal of each location, proximity to major roads, access through public transport, etc. - that influence such a decision.