Applied Data Science Capstone by IBM/Coursera

“The Battle of the Neighborhoods: L.A. Starbucks/Study Café Edition”

Capstone Project by Meghan Sills

March 28, 2019

Table of Contents

1. Introduction

1.1 Background

Los Angeles is the largest city on the West Coast of North America, with an estimated population of four million. [1] As the cultural, financial, and commercial center of Southern California, it is an attractive area for both students and universities, with three public universities (within city limits) [1] and many more outside city limits. There are also more than 21 private universities in the city and 9 community colleges. Popular universities outside the city limits but still in the Greater Los Angeles area will also be included in this study, as the Claremont Colleges consortium and the California Institute of Technology (Caltech) are some of the most selective universities in the nation.

1.2 Business Problem

Before opening a new study café in L.A., investors are faced with the question of where is the best location to open such a venue. To approach this problem, several aspects should be considered:

a) **Where are the universities located?** Since the new café should be located nearby one or more large universities.
b) **What are the characteristics of different neighborhoods?** Since the new café should be located in a lively neighborhood with other food & drink venues, but not with too many cafés yet (<4).
c) **How high are the housing prices for different neighborhoods?** Since the new café should be located in a neighborhood with low to moderate housing prices (<500 $/sqft) in order to be able to offer low-priced food and drinks to students.

The present project addresses this problem by sourcing different data about L.A. (e.g., housing prices, locations data of universities, venues, and districts) in order to aid investors in finding the optimal location for a new study café.

1.3 Interested Parties

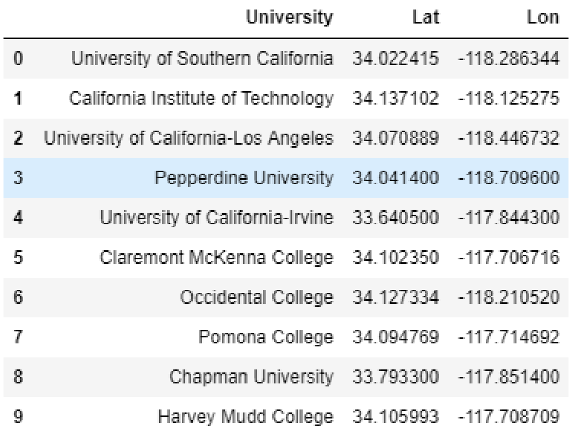
This report will be of particular interest to any investors, private or public, who might want to open a new student venue in Los Angeles and are yet undecided about the best location. Such investors could even include the universities themselves, who may decide to rent nearby off-campus space for their students. Non-local investors will especially profit from this report, as it will give them a nice visual overview of L.A.'s districts and their relevant characteristics.

2. Data Acquisition and Cleaning

To solve the problem, datasets from several sources were combined to answer the three questions referenced in Section 1.2: Business Problem.

2.1 Where are the Universities located?

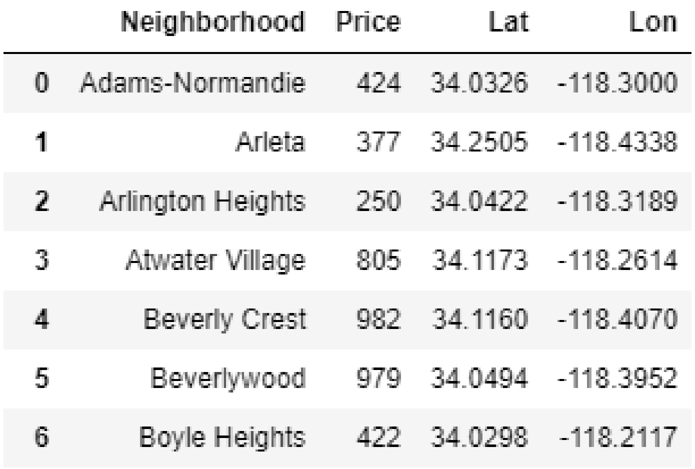
To obtain data on this question, a list of L.A.'s top universities was scraped from the web at: https://www.universities.com/find/los-angeles/best, containing the top 10 in the area. Based on these names, the geopy client (https://pypi.org/project/geopy) was then used with the Nominatim geolocator service to request the geographical coordinates (latitude and longitude) of each university. Since the request returned no result for 3 of the 10 universities, the latitude and longitude of these 3 universities were searched and added manually using Google Maps (https://maps.google.com/). All data was combined into a dataframe including the name, latitude, and longitude for each of the 10 universities, seen below:



	University	Lat	Lon
0	University of Southern California	34.022415	-118.286344
1	California Institute of Technology	34.137102	-118.125275
2	University of California-Los Angeles	34.070889	-118.446732
3	Pepperdine University	34.041400	-118.709600
4	University of California-Irvine	33.640500	-117.844300
5	Claremont McKenna College	34.102350	-117.706716
6	Occidental College	34.127334	-118.210520
7	Pomona College	34.094769	-117.714692
8	Chapman University	33.793300	-117.851400
9	Harvey Mudd College	34.105993	-117.708709

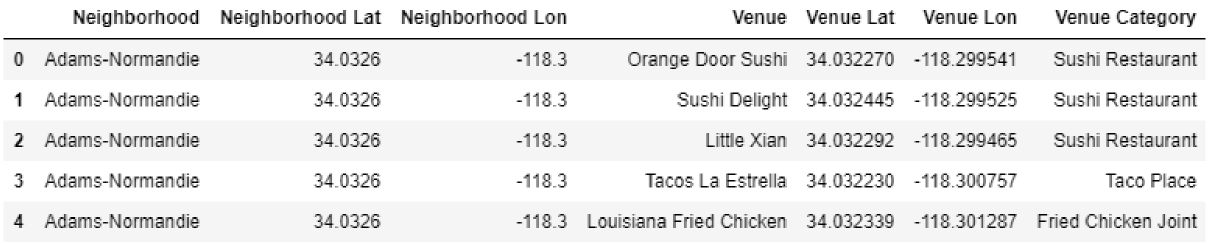
2.2 What are the characteristics of different neighborhoods?

To obtain data on this question, a geojson file of Los Angeles County's neighborhoods was downloaded from http://s3-us-west-2.amazonaws.com/boundaries.latimes.com/archive/1.0/boundary-set/la-county-neighborhoods-v6.geojson, containing the names and geographical borders of 318 neighborhoods. Combined with housing data, the overlap of neighborhoods left us with 80 neighborhoods to reference that were included in both datasets. This information was used for both the choropleth map of housing prices in L.A.'s neighborhoods (see 2.3) and to obtain a list of L.A.'s neighborhood names. This data was combined into a dataframe containing the name, price (per square foot), latitude, and longitude of L.A.'s 80 neighborhoods, an example of which is shown below:



	Neighborhood	Price	Lat	Lon
0	Adams-Normandie	424	34.0326	-118.3000
1	Arleta	377	34.2505	-118.4338
2	Arlington Heights	250	34.0422	-118.3189
3	Atwater Village	805	34.1173	-118.2614
4	Beverly Crest	982	34.1160	-118.4070
5	Beverlywood	979	34.0494	-118.3952
6	Boyle Heights	422	34.0298	-118.2117

Next, a list of venues for each neighborhood was obtained via the Foursquare API, using the "explore" endpoint with a limit of 100 and a radius of 500 meters around a neighborhood's given latitude and longitude. The returned information was combined with the neighborhood data into a dataframe showing both venue and location/neighborhood data, shown here:



	Neighborhood	Neighborhood Lat	Neighborhood Lon	Venue	Venue Lat	Venue Lon	Venue Category
0	Adams-Normandie	34.0326	-118.3	Orange Door Sushi	34.032270	-118.299541	Sushi Restaurant
1	Adams-Normandie	34.0326	-118.3	Sushi Delight	34.032445	-118.299525	Sushi Restaurant
2	Adams-Normandie	34.0326	-118.3	Little Xian	34.032292	-118.299465	Sushi Restaurant
3	Adams-Normandie	34.0326	-118.3	Tacos La Estrella	34.032230	-118.300757	Taco Place
4	Adams-Normandie	34.0326	-118.3	Louisiana Fried Chicken	34.032339	-118.301287	Fried Chicken Joint

This venue data was then used in a k-means clustering analysis to cluster L.A.'s neighborhoods based on their venue characteristics (see methodology section).

2.3 How high are the housing prices for different neighborhoods?

To obtain data on this question, a list of housing prices for L.A.'s neighborhoods was gathered from Zillow (downloaded to a local machine), containing the names and average housing prices in $/sqft for February 2019 (see 2.2 for example).

To integrate this data with the neighborhood's location and venue data from above, it was checked if the spelling of all neighborhoods included in this dataset were identical to the districts' spelling in the GeoJSON file. Some neighborhoods were combined if the only difference was punctuation, and many neighborhoods that weren't included in both datasets were dropped. This data was then used to create a map of housing prices and universities in L.A.'s neighborhoods (see methodology section).

3. Methodology

This project collected and combined several data sources about Los Angeles in order to recommend ideal locations for opening a new study café to potential investors. In the first step, data from each source was acquired and cleaned (see Section 2), resulting in the following datasets:

- A dataframe including the name, latitude, and longitude of L.A.'s top 10 universities (10 rows),
- A dataframe including the name, latitude, longitude, and price per sqft of L.A.'s neighborhoods (80 rows),
- And a dataframe including the name, latitude, longitude, and category of nearby venues for each of L.A.'s neighborhoods (1340 rows).

We then used this data for further analysis, including data exploration (3.1), k-means clustering (3.2), and data visualization (3.2).

3.1 Data Exploration

L.A.’s venue data was further explored before using the data for clustering. First, it was checked to see how many unique venue categories were returned in total (=238). Also, the number of returned venues per neighborhood was plotted in a histogram (Figure 1) in order to get a first idea about the venue richness of L.A.’s neighborhoods.

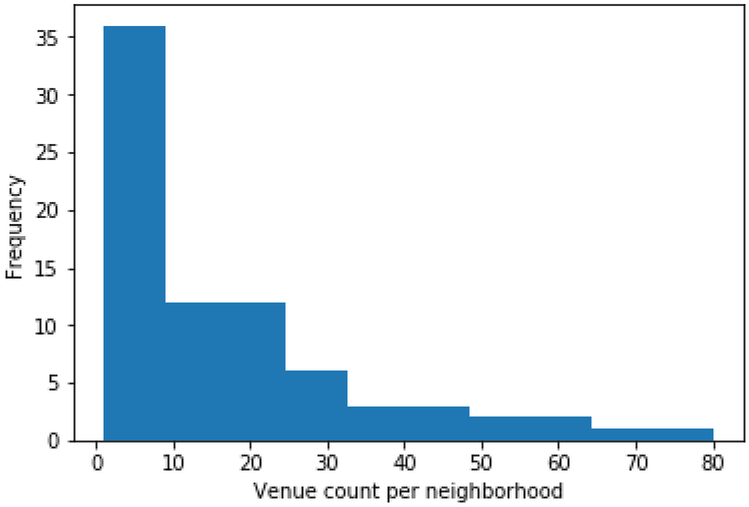
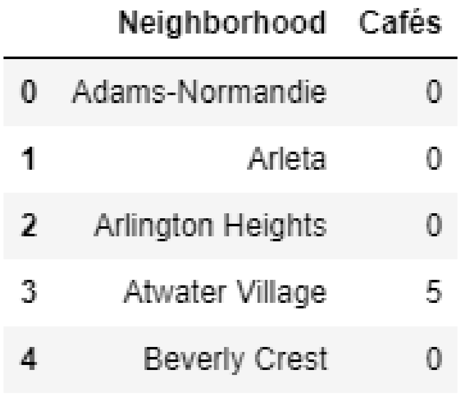


Figure 1.

Furthermore, the names of all venue categories were explored to identify the categories that describe coffee shops. These included “café” and “coffee shop,” for which the combined total number per neighborhood was calculated and saved in a new dataframe for later data visualization.



	Neighborhood	Cafés
0	Adams-Normandie	0
1	Arleta	0
2	Arlington Heights	0
3	Atwater Village	5
4	Beverly Crest	0

From the histogram in Figure 2, it is made clear that most neighborhoods have either no or very few cafés yet. However, some neighborhoods seem to already have many cafés (i.e., Westwood with 7 cafés) and would thus be less suitable for opening the new study café.

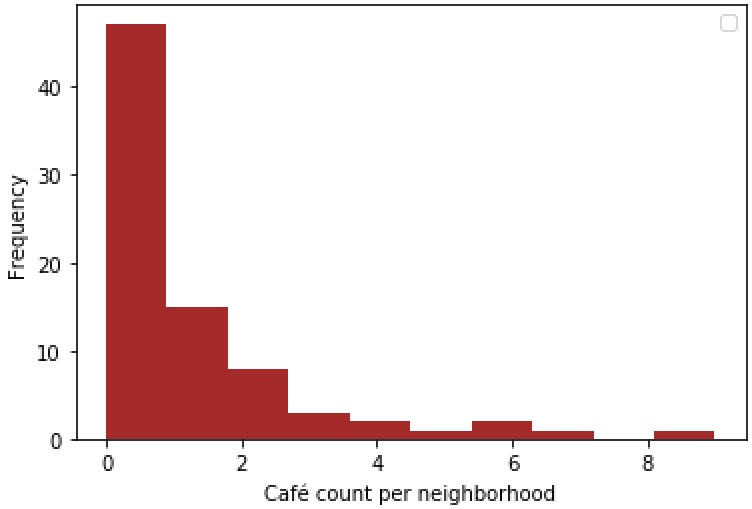


Figure 2.

Finally, a histogram of the housing prices is shown in Figure 3 to get an idea about the housing price distribution in L.A.'s neighborhoods. Some neighborhoods seem to have very high housing prices and would thus be less suitable for opening a new study café. For example, Brentwood, Beverly Crest, and Pacific Palisades are a couple of neighborhoods with extremely high housing prices.

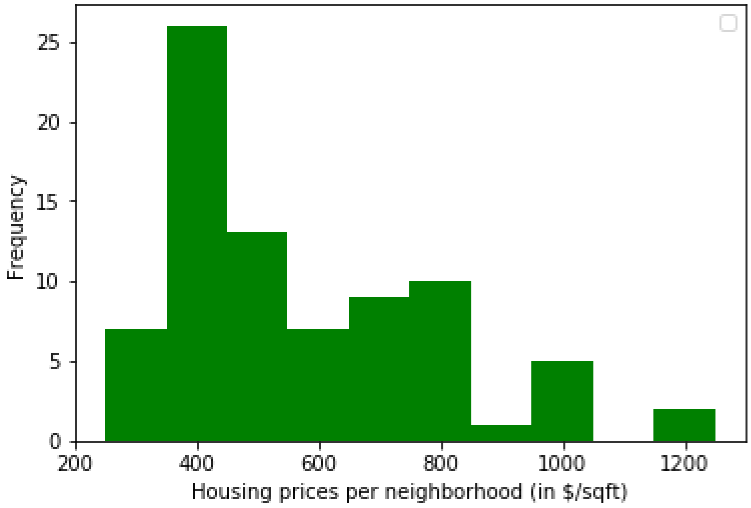
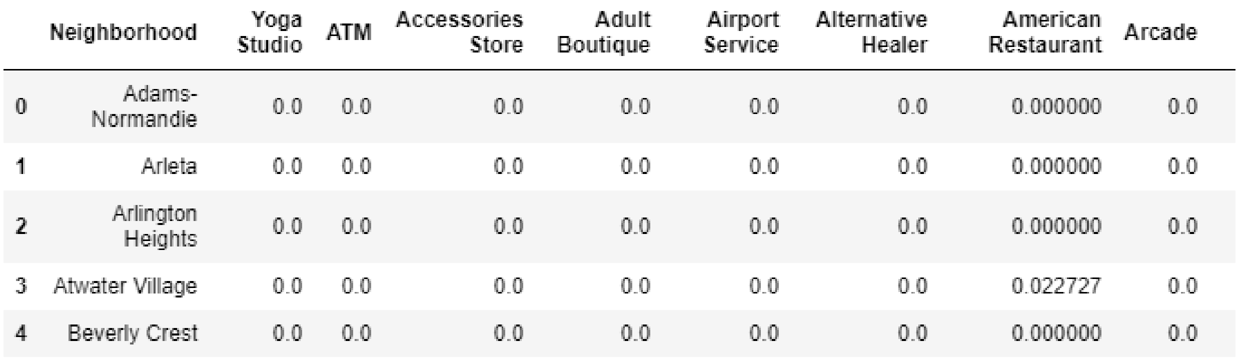


Figure 3.

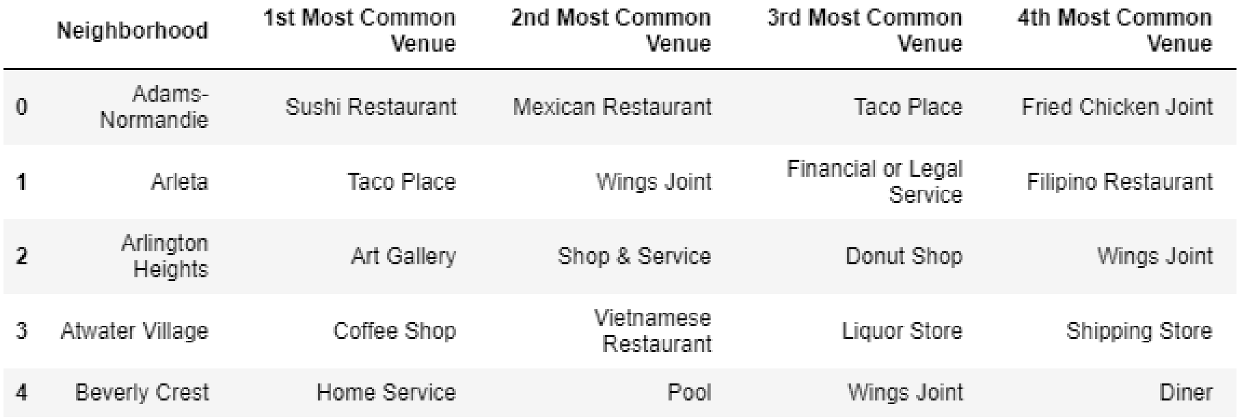
3.2 Clustering Analysis of L.A.'s Neighborhoods by Venues

One problem that needed to be addressed by this project was to identify neighborhoods that are most suitable for opening a new study café based on their current venue characteristics. In particular, we were looking for lively neighborhoods in which several food and drink venues were already located and which might be attractive places for students to meet and socialize in between or after classes. To approach this problem, a k-means clustering algorithm was used on L.A.'s venue data. Specifically, we first used the venue data acquired from Foursquare (see below) to obtain the mean frequencies of venue categories for each neighborhood. Then, we used this data to cluster the districts via the k-means clustering algorithm.



	Neighborhood	Yoga Studio	ATM	Accessories Store	Adult Boutique	Airport Service	Alternative Healer	American Restaurant	Arcade
0	Adams-Normandie	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
1	Arleta	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
2	Arlington Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
3	Atwater Village	0.0	0.0	0.0	0.0	0.0	0.0	0.022727	0.0
4	Beverly Crest	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0

Based on this data, another dataframe was created that showed the top 10 venue categories for each neighborhood:



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Adams-Normandie	Sushi Restaurant	Mexican Restaurant	Taco Place	Fried Chicken Joint
1	Arleta	Taco Place	Wings Joint	Financial or Legal Service	Filipino Restaurant
2	Arlington Heights	Art Gallery	Shop & Service	Donut Shop	Wings Joint
3	Atwater Village	Coffee Shop	Vietnamese Restaurant	Liquor Store	Shipping Store
4	Beverly Crest	Home Service	Pool	Wings Joint	Diner

Next, the districts were clustered by their venue frequencies using the k-means clustering algorithm from the sklearn library (https://scikit-learn.org). As the optimal k for clustering is unknown, the k-means clustering algorithm was first run with different values for k and an elbow plot (Figure 4) was created that shows the within-cluster sum of squares (i.e., inertia or distortion) for each value of k.

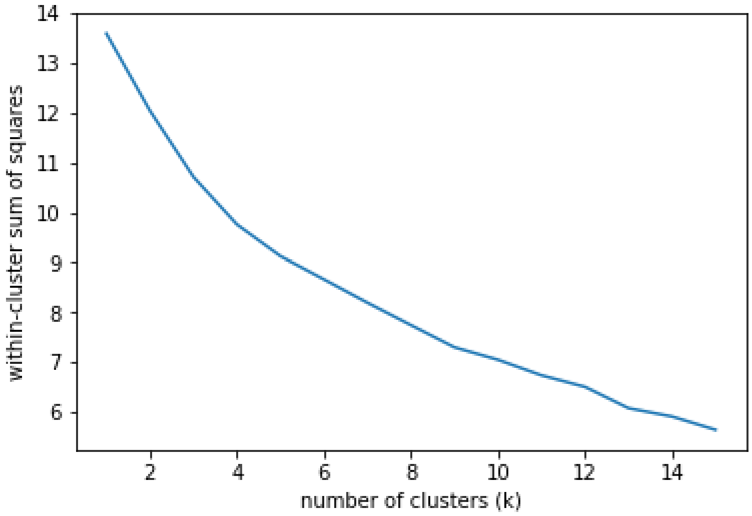
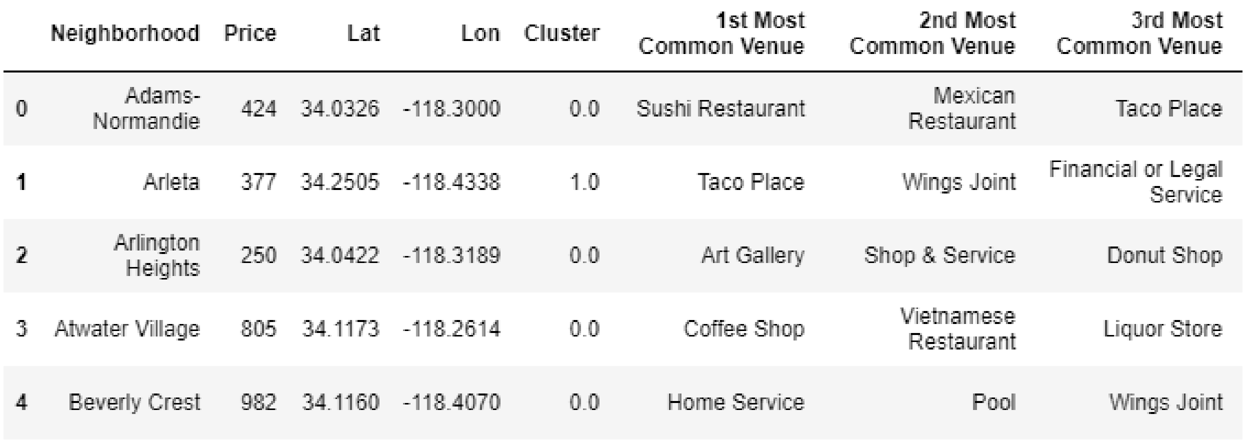


Figure 4.

As we see in the plot, there is no clear "elbow" that shows us the optimal value of k for clustering. Note that I also tried alternative methods to determine the optimal k, e.g. the silhouette coefficient or the gap statistic (code not shown to keep the notebook cleaner). However, these methods also yielded no clear answer about the optimal value for k. Hence, the value for clustering was set to **k = 4**, based on the following rationale: On the one hand, we want more than 1-2 clusters to better distinguish the large number of neighborhoods based on their different venue characteristics. Yet, we also do not want too many clusters to keep the resulting cluster structure easily understandable (e.g. we do not want many unique clusters that only contain 1-2 neighborhoods each).

Next, the resulting cluster labels were included in the below dataframe containing the top 10 venues for each neighborhood.

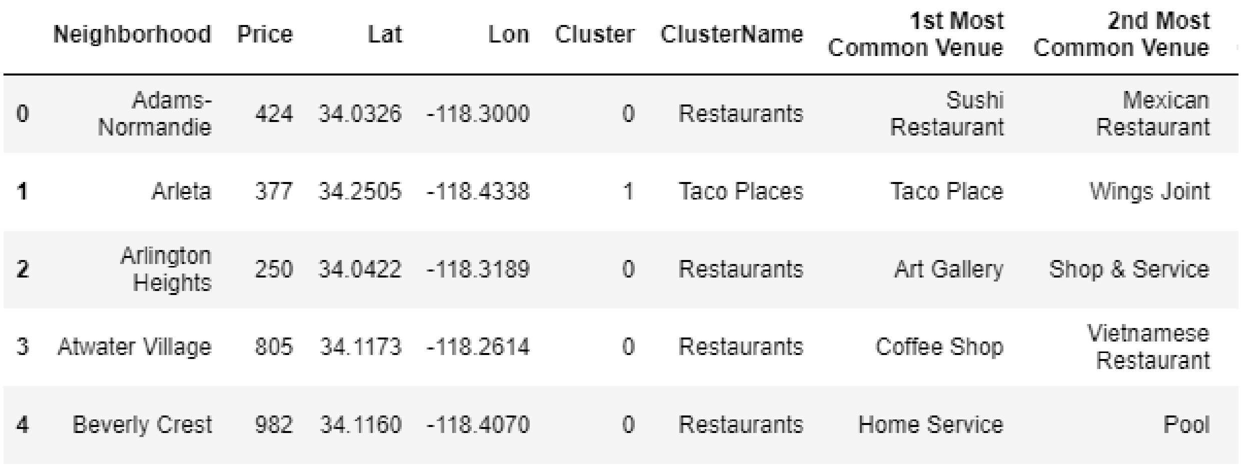


	Neighborhood	Price	Lat	Lon	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Adams-Normandie	424	34.0326	-118.3000	0.0	Sushi Restaurant	Mexican Restaurant	Taco Place
1	Arleta	377	34.2505	-118.4338	1.0	Taco Place	Wings Joint	Financial or Legal Service
2	Arlington Heights	250	34.0422	-118.3189	0.0	Art Gallery	Shop & Service	Donut Shop
3	Atwater Village	805	34.1173	-118.2614	0.0	Coffee Shop	Vietnamese Restaurant	Liquor Store
4	Beverly Crest	982	34.1160	-118.4070	0.0	Home Service	Pool	Wings Joint

After obtaining the different cluster labels, the resulting clusters were further examined to determine the discriminating venue categories that distinguish each cluster. Based on the 5 most common venue categories, a name was assigned to each cluster resulting in the following cluster names:

- Cluster 0: “Restaurants” (74 neighborhoods)
- Cluster 1: “Taco Places” (2 neighborhoods)
- Cluster 2: “Parks” (1 neighborhood)
- Cluster 3: “Grocery Stores” (1 neighborhood)

Finally, these cluster names were added to the below dataframe:



	Neighborhood	Price	Lat	Lon	Cluster	ClusterName	1st Most Common Venue	2nd Most Common Venue
0	Adams-Normandie	424	34.0326	-118.3000	0	Restaurants	Sushi Restaurant	Mexican Restaurant
1	Arleta	377	34.2505	-118.4338	1	Taco Places	Taco Place	Wings Joint
2	Arlington Heights	250	34.0422	-118.3189	0	Restaurants	Art Gallery	Shop & Service
3	Atwater Village	805	34.1173	-118.2614	0	Restaurants	Coffee Shop	Vietnamese Restaurant
4	Beverly Crest	982	34.1160	-118.4070	0	Restaurants	Home Service	Pool

3.3 Data Visualization Using Folium

In the final step, an interactive map of Los Angeles was created that combines all relevant data to solve the initial problem of finding an optimal location for the new study café. The map was created using the Folium library (https://pypi.org/project/folium/) and included the following data:

- The locations of L.A.'s universities, each marked by a small blue circle on the map with a popup label showing the university's name.
- The locations of L.A.'s neighborhoods, each marked by a larger grey circle on the map with a popup label showing the neighborhood's name, cluster label (0-3), its average housing price, and the number of cafés already located in that neighborhood.

As the map represents the main result(s) of this project, a figure of this map is presented in the following Results section.

4. Results

In this project, several data sources about L.A. (housing prices, location of neighborhoods, universities, and venues) were combined and visualized in an interactive Folium map in order to find good locations for opening a new study café. Figure 5 gives a static impression of the resulting Folium map.

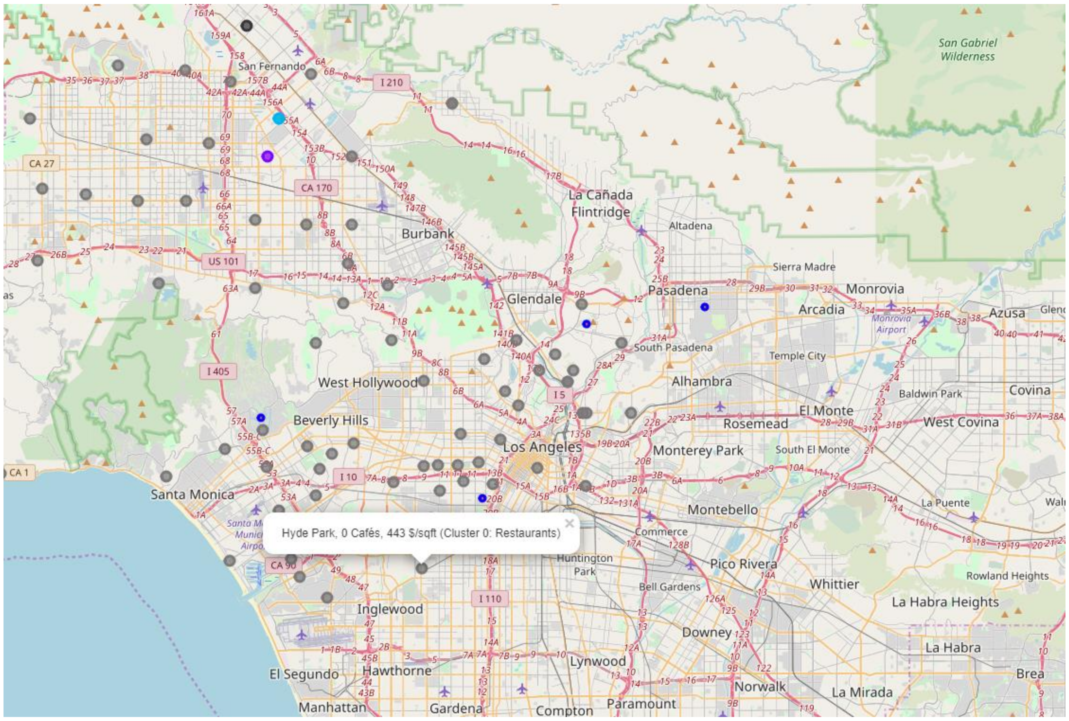


Figure 5.

From the map in Figure 5, we can already make two important observations: First, most universities are scattered, thus leaving investors to pick a neighborhood near one university in particular. Second, we can see that most neighborhoods belong to **Cluster 0 "Restaurants,"** which means that most neighborhoods have a high number of food and drink places already there (but not necessarily a high number of cafés).

In the next section, we will come back to the business problem and use these results to address the initial question of where to best open the new study café.

5. Discussion

This project addressed the problem of finding the ideal neighborhood for opening a new study café in Los Angeles. Specifically, we were looking for a neighborhood with the following characteristics:

- The study café should be located nearby one or more universities.
- The study café should be located in a lively neighborhood with other food and drink venues, but not with too many cafés yet (<4).
- The study café should be located in a neighborhood with low to moderate housing prices (<500 $/sqft) in order to be able to offer low-priced food and drinks to students.

By zooming in on the map and exploring the districts, we can identify some districts that fulfill the above criteria: **University Park, Adams-Normandie**, and **Jefferson Park**. We can now present the map and list of selected neighborhoods to make a best-fit recommendation to potential investors about where to open the new study café. Based on this list and my own experiences from living in Los Angeles, I would specifically recommend the **University Park** neighborhood, as it is a very lively and nice area, is easy to walk to from the nearby University of Southern California, and is not run down compared to the other two neighborhoods selected with low to moderate housing prices.

6. Conclusion

The aim of this project was to help investors in finding the optimal location for opening a new study café by giving them a nice visual overview of L.A.'s neighborhoods and their relevant characteristics. To this end, the project collected and combined data about L.A. from several sources (housing prices, location of universities, venues, and neighborhoods) and visualized this data on a geographical map of Los Angeles. Based on this map, the project identified three neighborhoods and recommended one neighborhood in particular that met the required criteria for the new study café. These neighborhoods should be considered by potential investors as places to open a new study café or other off-campus venue, which may greatly enrich student life in Los Angeles.

7. References

1. https://en.wikipedia.org/wiki/Los_Angeles#Colleges_and_universities
2. https://www.universities.com/find/los-angeles/best
3. https://pypi.org/project/geopy
4. https://maps.google.com/
5. http://s3-us-west-2.amazonaws.com/boundaries.latimes.com/archive/1.0/boundary-set/la-county-neighborhoods-v6.geojson
6. https://scikit-learn.org
7. https://pypi.org/project/folium/