Clustering the San Fernando Valley neighborhoods to set a new restaurant. Analyzing home values in the clustered area.

Francihelena Uzcategui

June 21, 2020

1. Introduction

1.1 Background

A client decided to establish a Restaurant in San Fernando Valley, California. Besides, the client plan to buy a house in the same area where the restaurant will be located.

In the last years, doing business in the San Fernando Valley has been going flexibly. Its neighborhoods earned this reputation, being a stimulus to the economic growth of the region[1].
A similar situation for San Fernando Valley's house prices which has been rising because the home values are sustainable along the time. Thus, it is going to be appropriate to buy a house for business or residence[2].

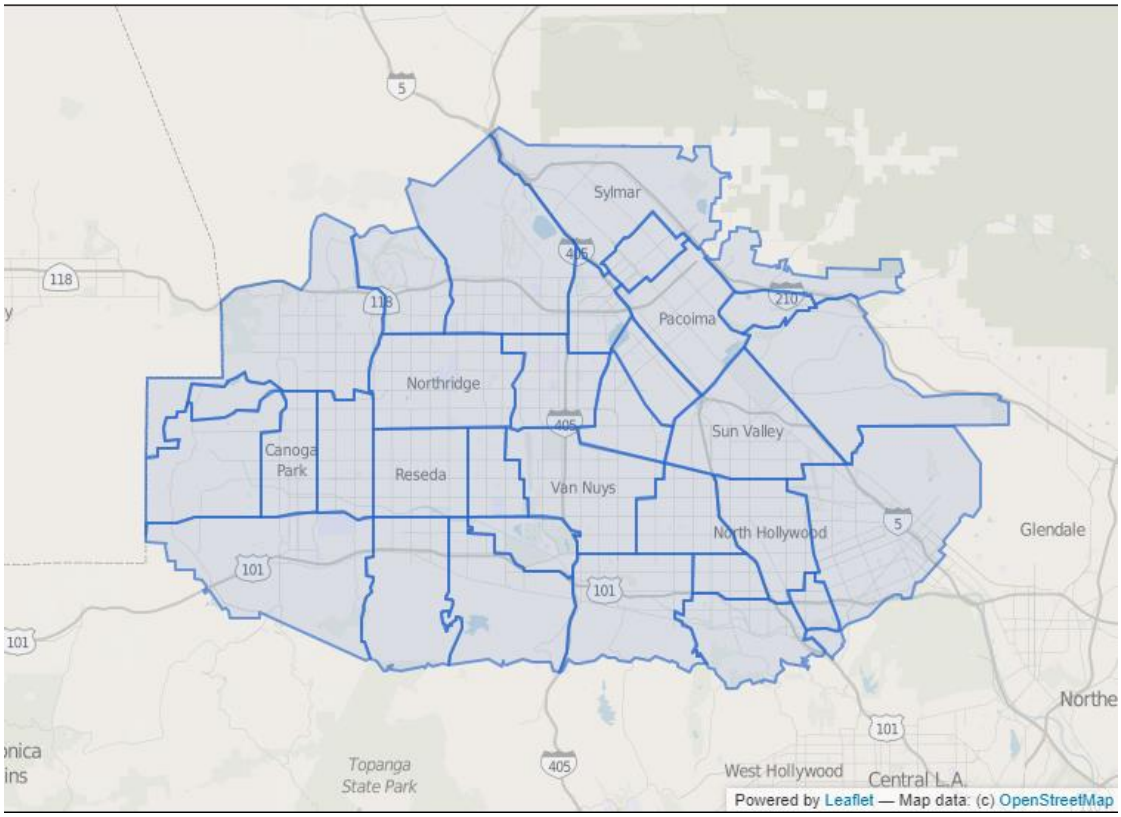
The San Fernando Valley is an urbanized valley in Los Angeles County, California[3], nearly two-thirds of the Valley's land area is part of the LA County.

The San Fernando Valley contains 34 neighborhoods: Burbank, San Fernando, Universal City, Arleta, Canoga Park, Chatsworth, Encino, Granada Hills, Lake Balboa, Lake View Terrace, Mission Hills, North Hills, North Hollywood, Northridge, Pacoima, Panorama City, Porter Ranch, Reseda, Shadow Hills, Sherman Oaks, Studio City, Sun Valley, Sylmar, Tarzana, Toluca Lake, Valley Village, Van Nuys, West Hills, Woodland Hills, Hasem Dam, Winnetka, Sepulveda Basin, Valley Glen, and Chatsworth Reservoir.

---

[1] https://laedc.org/wtc/chooselacounty/regions-of-la-county/san-fernando-valley/
[2] https://www.latimes.com/homeless-housing/story/2020-02-25/southern-california-home-prices
[3] https://en.wikipedia.org/wiki/San_Fernando_Valley

The map below shows San Fernando Valley, our target.



Source: http://maps.latimes.com/neighborhoods/region/san-fernando-valley/

1.2 Problem

San Fernando Valley is a large and diverse region, thus mapping their neighborhoods is a convenient way to deal with this complex scenario. The cluster method helps to select the type of restaurant and its location while defining the area to analyze the home values market.

1.3 Interest

To make wise investments is necessary to research the target market, through reviewing historical data and correlate variables. In this case, the client wants to invest in a restaurant and a house; thus, this project aims to define specific groups of types of restaurants and locations to set the restaurant and home.

2. Data gathering and cleansing

2.1 Data sources

The essential information required: region, neighborhood, ZIP code, longitude, and latitude. The geographic coordinates of each community in Los Angeles County can be found by downloadable csv or excel files on government webpages here and

here. Highlighting that Los Angeles County is a complex and massive unity, it does not have boroughs; instead, it has unincorporated communities, incorporated cities, and neighborhoods of the city of Los Angeles - that conforming to a region such as San Fernando Valley.
The secondary information required to analyze the home prices includes historical data of home values, downloading the csv from here. And additional factors that can affect the home prices, such as the number of bedrooms - between 1 and +5.

2.2 Data cleansing
The three datasets were downloaded and scraped, then they were joined into one, to obtain a single dataset. Lastly, we cleaned and filtered by the 34th target neighborhoods.

The first join included two data frames df1 and df2 to create a consolidated data frame with the required information: 5 columns: ZIP Code, Latitude, Longitude, City, and Community; and several communities by each zip code.

- **df1:** Congressional Districts Los Angeles County - By Zip Code

First, split 'Zip Code – City/Community' column to obtain two columns, one for Zip Code and another for City/Community, this last one requires splitting actions. Also, renaming the columns.
Next, remove unwanted columns.
Finally, convert the data type of the ZIP Code column to an integer to avoid future errors on the join method.

- **df2:** Los Angeles Zip Codes and Latitude/Longitude by each neighborhood in Los Angeles County.

First, drop unwanted columns. Then, split the location column, because it has mixed information and contains the geographic coordinates and zip codes in the same row. And, erase the duplicate zip codes. Lastly, drop duplicate columns and remove unwanted characters ",".

- **df_LAcounty** is the join between df1 and df2.

First, apply the inner join to the two data frames. Next, include by .loc method three missed neighborhoods into the Neighborhoods column. This lack happened due to several communities share the same zip code.
Later, drop unwanted column: Community because it contains duplicate information. Thus, we kept the Neighborhoods column because it has more details of the neighborhoods and includes the format required.

As we mentioned in the Introduction section, our target is the San Fernando Valley rather than Los Angeles city; remember that San Fernando Valley is a conglomerated of communities within Los Angeles County.

Therefore, we searched each SFV's neighborhoods within the LA county's data frame through a lambda operation to get True: San Fernando Valley or False: Los Angeles.

Finally, the data frame df_LAcounty has the latitude and the longitude coordinates of each neighborhood in Los Angeles County, including San Fernando Valley. It is the data structure wanted to use for the clustering process.

- **df_target_forecast**: Forecast home values by SFV neighborhoods

For the analysis of the home market around the cluster zone, we downloaded data already cleaned. As a consequence, we obtained the information required by a simple filter task for Los Angeles County and SFV neighborhoods.

- **merged_df_bed:** All single-family homes data by the number of bedrooms

The data downloaded was cleaned, but was not unified, because there were five files, each by bedroom size, from 1 to +5 bedrooms. Thus, applying filters to each bedroom size applied to get Los Angeles County and SFV neighborhoods.
In the end, there were five data frames to concatenate into one data frame (all the bedrooms.)

- **neighbordhood_and_bedroom**: for better understanding the dataset with the bedroom size, we grouped by RegionName and number of bedrooms.

2.3 Feature selection
After the data cleaning process, there were four main data frames ready to use:

df_LAcounty contains the geographic coordinates of LA County, including SFV neighborhoods-our target. It has 278 samples.

df_target_forecast includes a general forecast of home values in SFV neighborhoods during April 2020. It has 20 samples.
After checking the columns, it was clear there was redundancy in a feature - CityName. It can be dropped.

merged_df_bed has the information of home values by the number of bedrooms of LA County, including SFV neighborhoods-our target.
It has 98 samples and 294 features. Upon check the meaning of each element, it was clear that there were some non-necessary features, such as, from 1996 to 2020 years; we used only one month of forecast April 2020.

selected_neighbordhood_bedroom is the result of Slicing the neighborhoods that we want to select, according to the number of bedrooms. For our investigation, we focus on the last month - April 2020, therefore we dropped 291 features.

Table 1. Simple feature selection during data cleaning

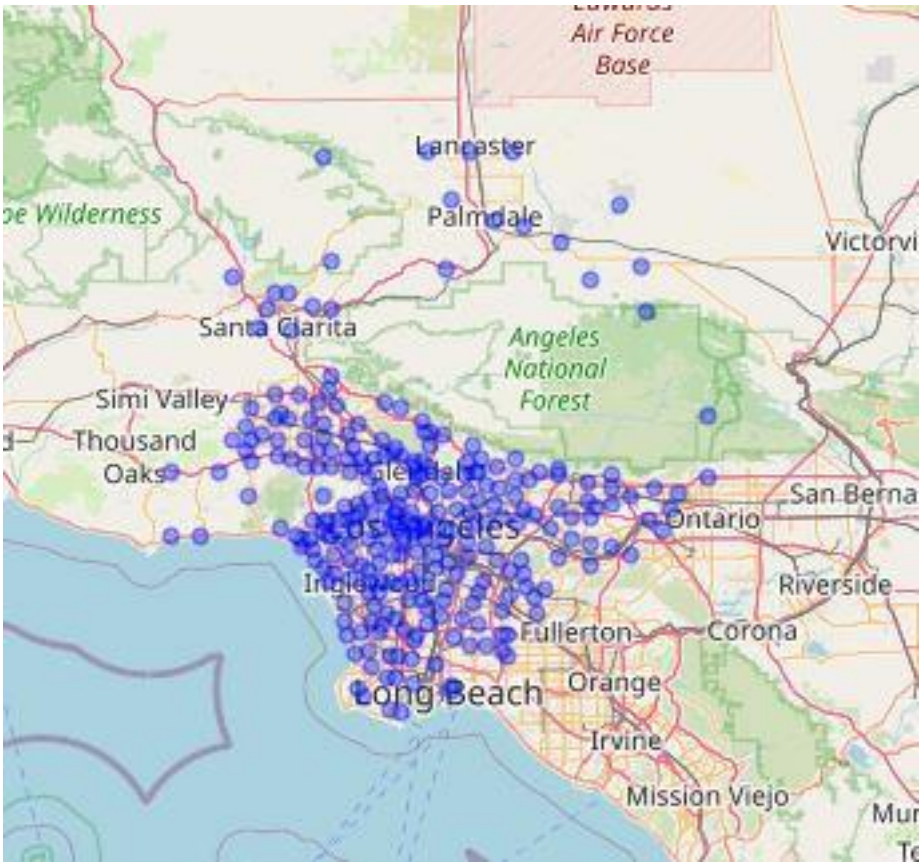
Data frame	Kept features	Dropped features	Reason for dropping features
df_LAcounty	5 features: ZIP Code, Neighborhoods, Latitude, Longitude, Region	2 features: City and TrueFalse	City is redundant with Neighborhoods. TrueFalse is redundant with Region.
df_target_forecast	5 features: Region, RegionName, CountyName, CityName, ForecastYoYPctChange	1 feature: CityName	CityName is redundant with RegionName
merged_df_bed	294: RegionName, bedroom, from 1996-01-31 to 2020-04-30	8 features: SizeRank, RegionID, RegionType, CountyName, City, StateName, State, Metro	These eight features are too general.
selected_neighbordhood_bedroom	3 features: RegionName, bedroom, 2020-04-30	291 features: from 1996-01-31 to 2020-03-31	For our investigation, we want just the last month - April 2020.

3. Exploratory Data Analysis

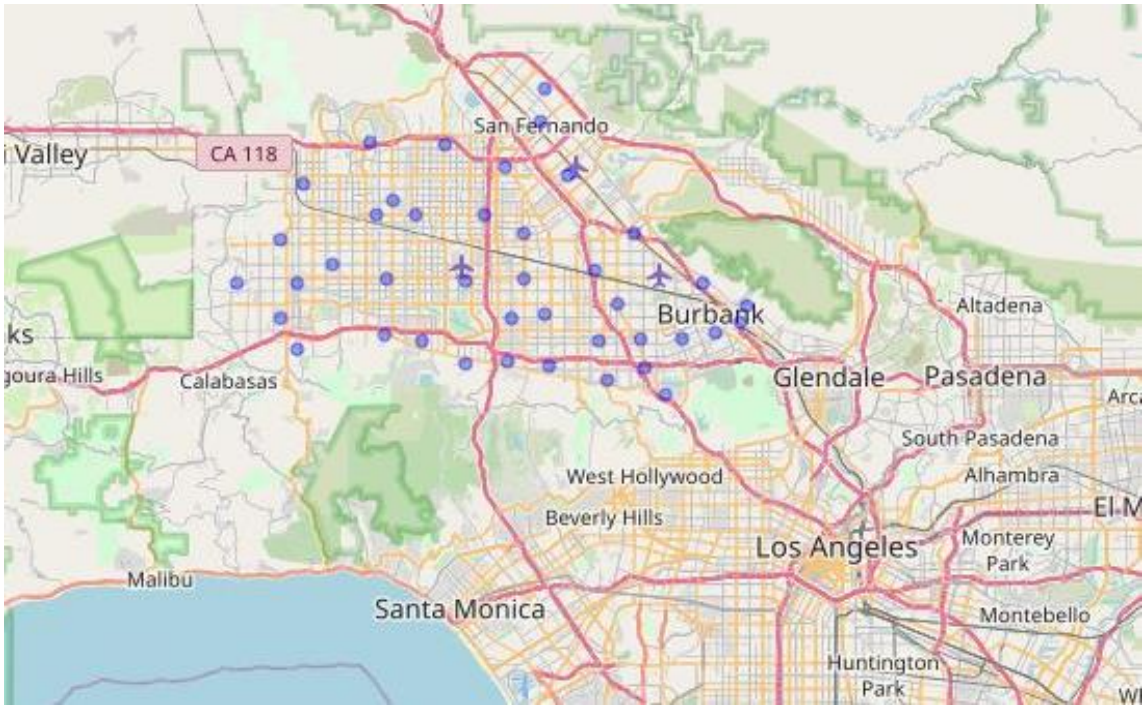
3.1 Exploring geographic coordinates Los Angeles County, and San Fernando Valley

For Clustering, spatial data is essential to define the longitude and latitude of the target variables. Therefore, using the Folium package to map made readable the geographic coordinates of Los Angeles County, and San Fernando Valley.

First, we processed Los Angeles county, the whole dataset, including the San Fernando Valley.



Second, access to San Fernando Valley geographic coordinates by filter SFV neighborhoods.



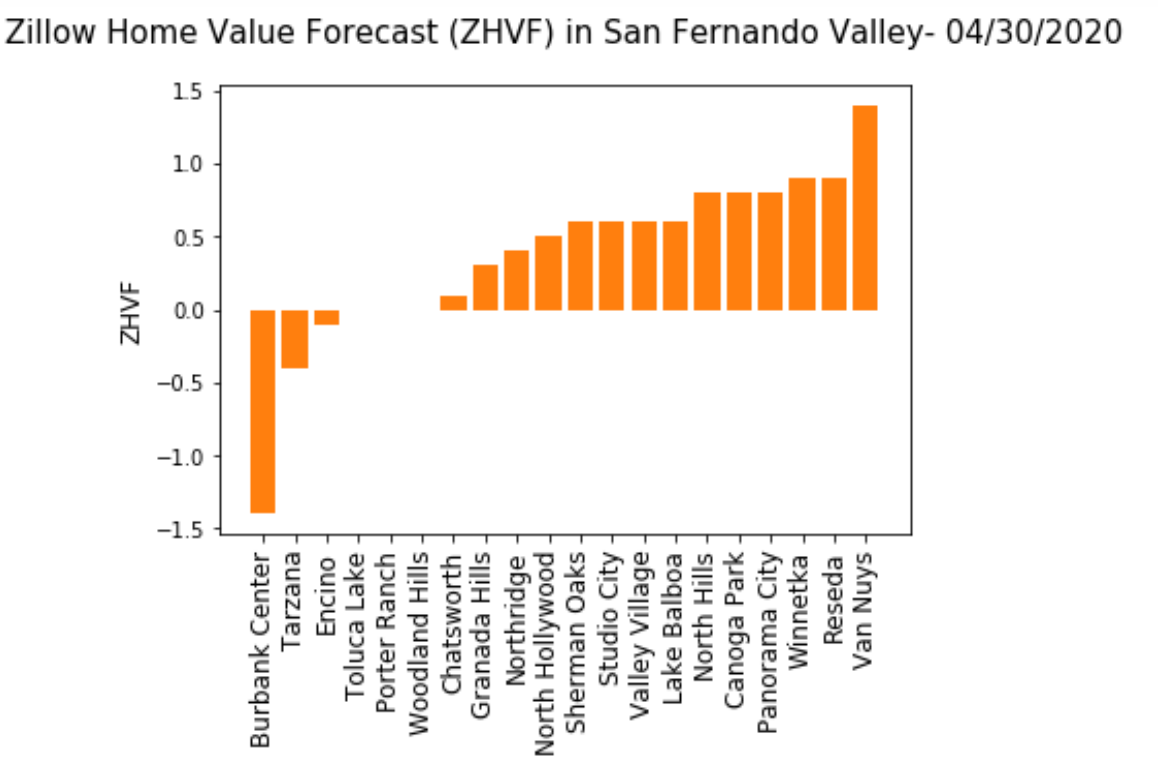
3.2 Explore the San Fernando Valley neighborhoods and segment them
Foursquare is a local search-and-discovery mobile app, which helps to get recommendations of places to go and its locations based on users browsing. In our case, the Restaurants in the San Fernando neighborhoods.

To use the Foursquare API required credentials.

The query's result brought 208 unique categories of venues, including the Restaurant venues, and 28 San Fernando neighborhoods. Naturally, it is convenient to reduce the number of groups, therefore searching the top ten most common venues by each community. On further process, we filtered it one more time by the Restaurant venues (our object.)

3.3 Explore the Zillow Home Value Forecast (ZHVF) in San Fernando Valley- 04/30/2020

The forecast shows a positive and negative tendency by each neighborhood. To invest, we focused on down tendency because of looking at a competitive home price.



To get more significant finds, we focused on these neighborhoods with negative forecasts and evaluated the home values of the last month of April 2020. These neighborhoods: Woodland Hills, Porter Ranch, Toluca Lake, Encino, Tarzana, and Burbank Center.

3.4 Explore single-family homes data by the number of bedrooms

For the data that contains all single-family homes by the number of bedrooms, we applied grouping by RegionName, the number of bedrooms, and the month. Therefore, focusing on Cluster 2's neighborhoods with negative forecasts for April 2020.

This information brings to the client a readable output about the neighborhood where to invest and the home values for the number of bedrooms during the last month.

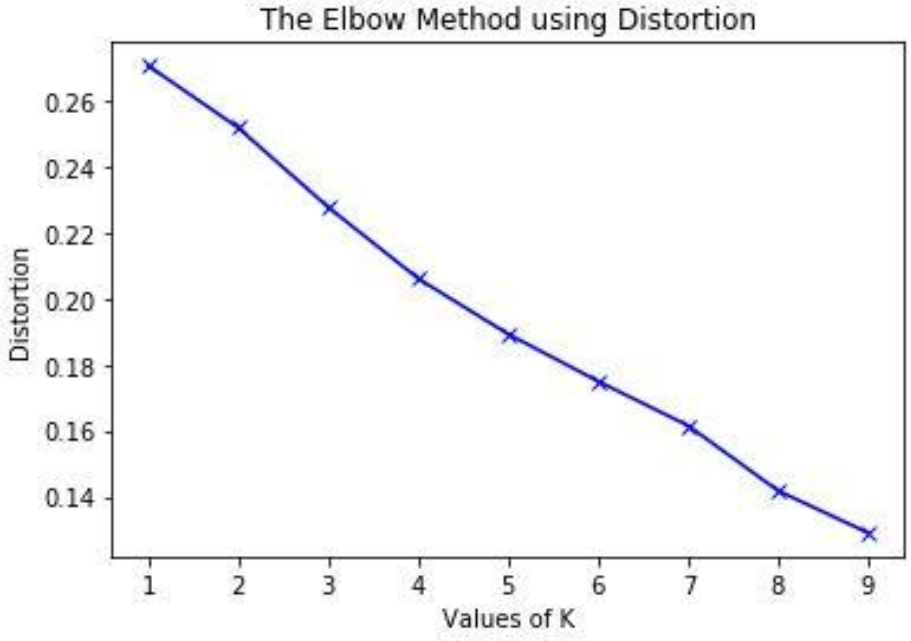
4. Predictive Modeling

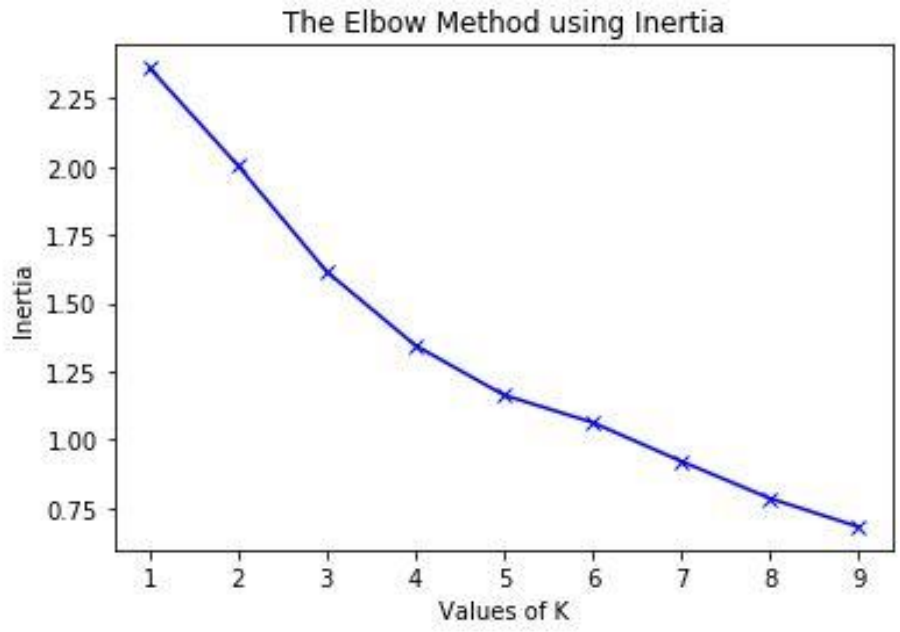
4.1 Classification model

There are two types of models, regression, and classification. The classification model is appropriate to apply the clustering used to define the kind of restaurants and their locations. The regression model is not convenient to use because the goal is not predicting the variety of restaurants. Although for the second part of the study- home value price, we can apply the regression model, but we would not carry that now. At this point, it is just an exploratory task.

K - means algorithm helps to cluster the features: typical venues and neighborhoods. This algorithm looks for similar group venues within each neighborhood of San Fernando Valley.

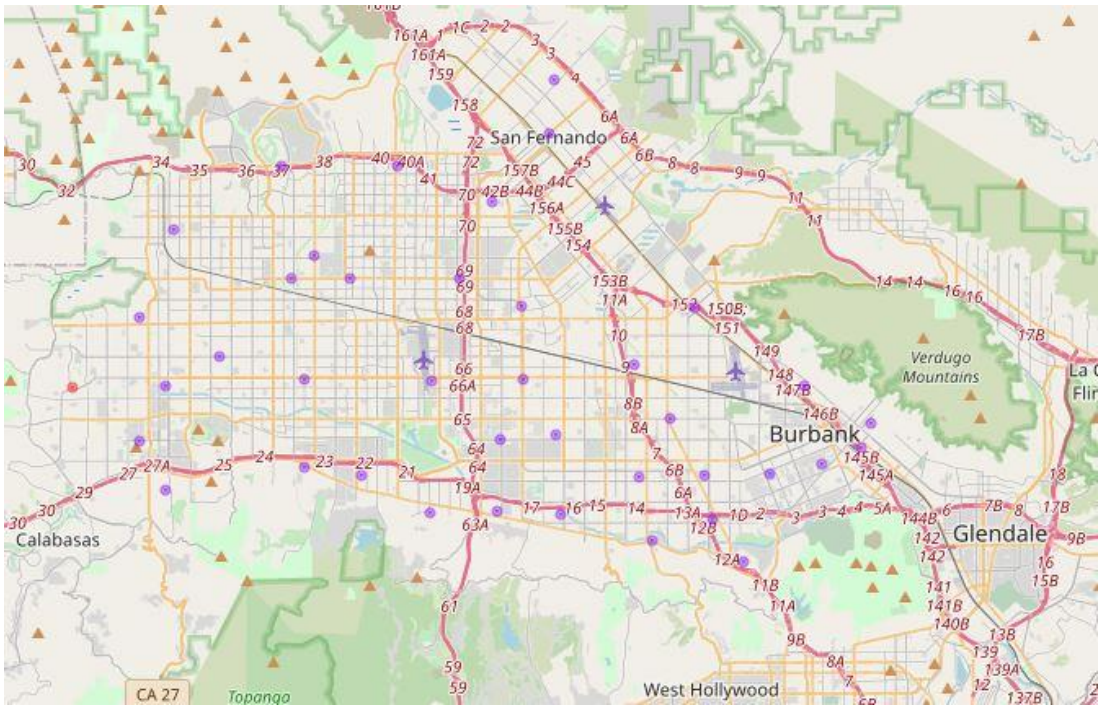
The Elbow Method helps to define the optimal value of k in K - Means, using a range from 1 to 10.





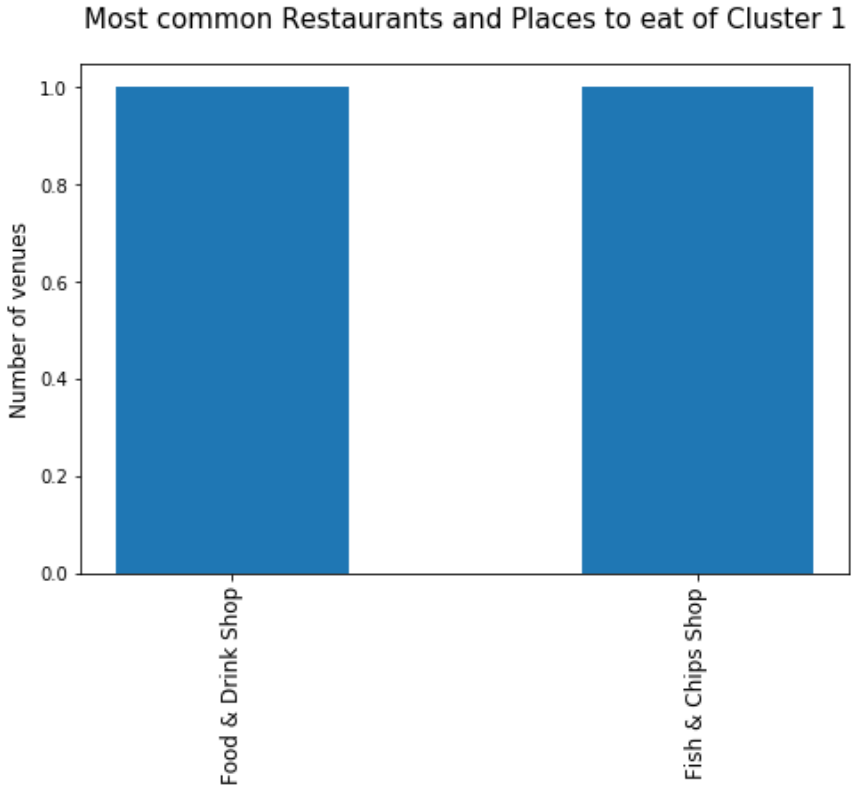
The above Distortion/Inertia plots show elbows at place 2. But, at 3 the line started to decrease. The elbows are not sharply shifting, but they depict that point 3 is the best value of K. Remember, increasing the K will always reduce the error. Since setting the number of clusters: three.

The map below shows the distribution of the clusters and their labels in the San Fernando Valley.

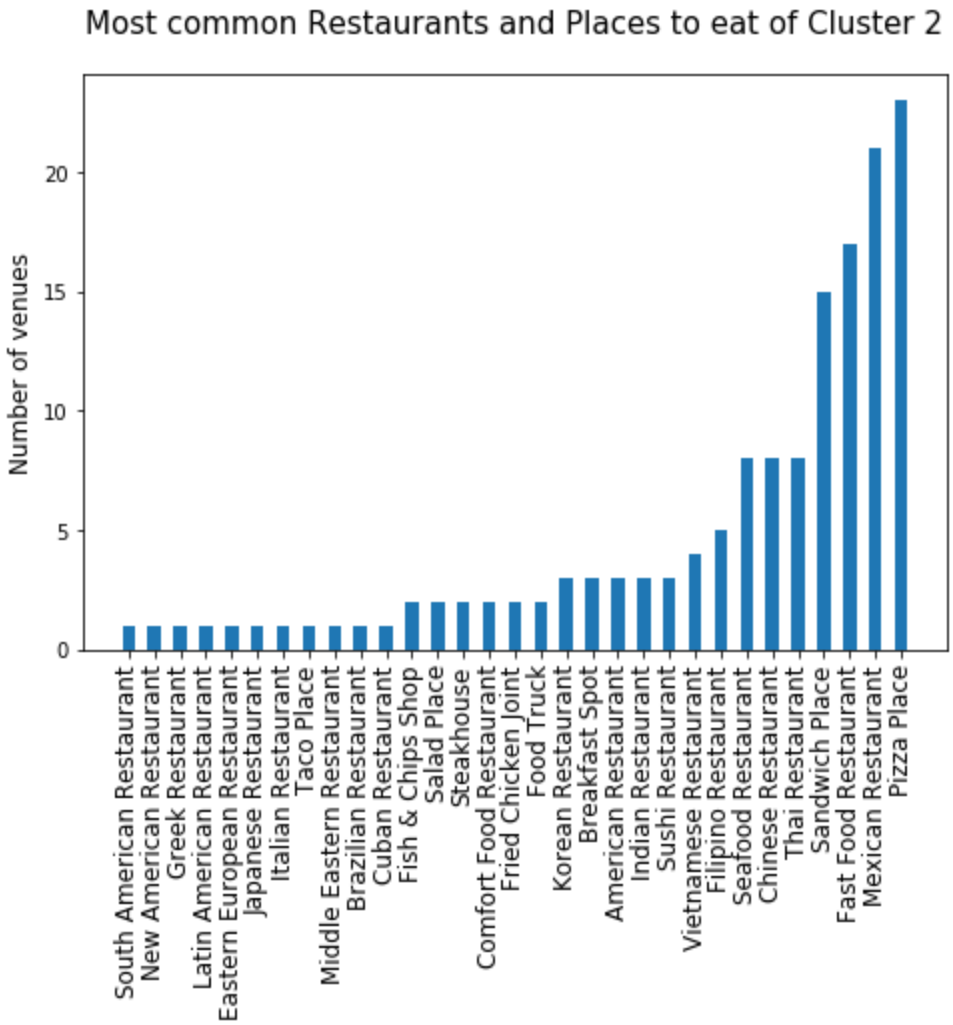


5. Discussion

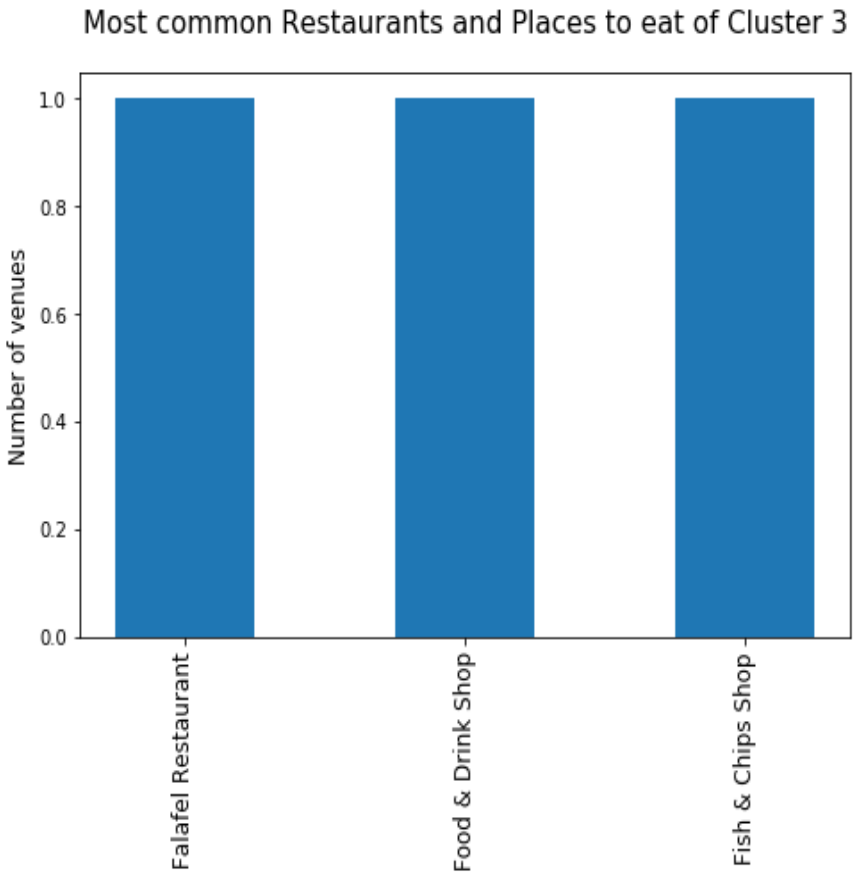
The type of food for the restaurant is not defined yet, either the location. Although at this stage, the client can have the information required to decide.
The results of the most popular venues help to know more about the Restaurant industry and customer preferences.
According to the Cluster's findings, San Fernando Valley has a large variety of restaurants that offer international cuisines, such as Latin, Asian, and European.



Cluster 1 has a small size, being a balanced tendency for each type of food venue: Food & Drink Shop and Fish & Chips Shop. Group 1 is conforming to West Hills neighborhood.



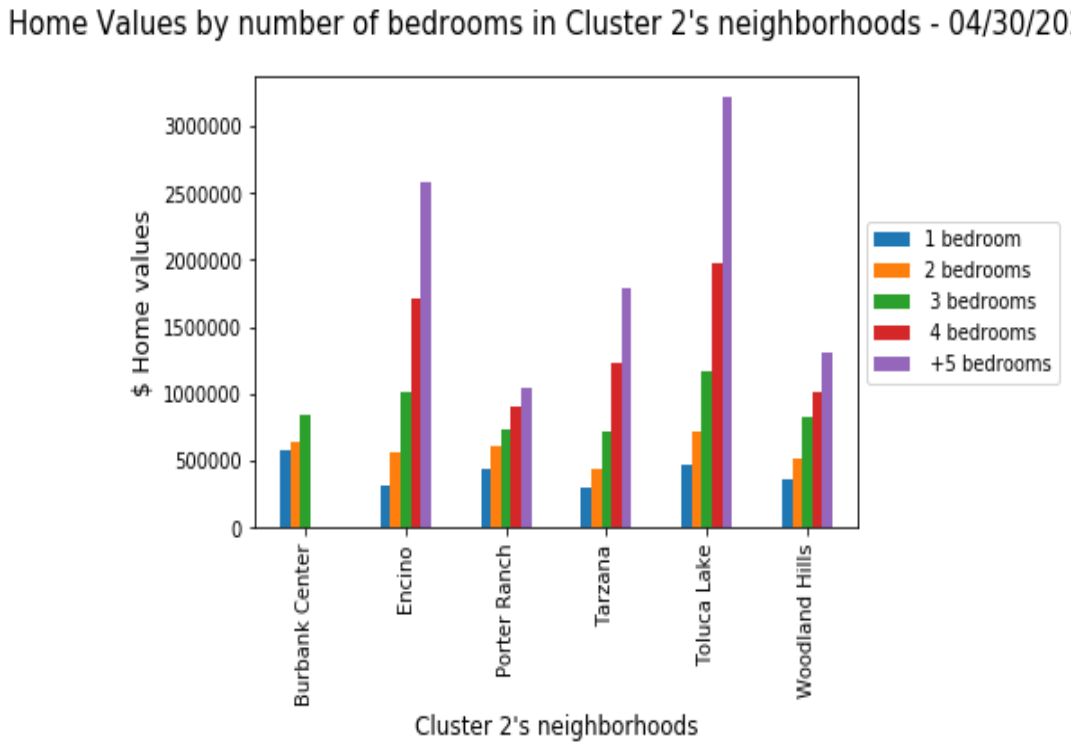
Cluster 2 contains the most massive venue numbers and covers a vast zone of San Fernando Valley, where the leader of the type of Restaurant is the Pizza. Follow by the Mexican restaurants, next, the Fast food—subsequent, the Sandwich places. Lastly, similar preferences for Thai Restaurants, Chinese restaurants, and Seafood restaurants.



Cluster 3 has a small size, being a balanced tendency for each type of food venue: Food & Drink Shop, Fish & Chips Shop, and Falafel Restaurant. Cluster 3 is conforming to the Pacoima neighborhood.

Using the Cluster 2's findings to explore the home values:

Creating a handle summary, we focused on Cluster 2's neighborhoods with negative forecasts for April 2020.



The side by side graph shows that Toluca Lake, Encino, and Tarzana neighborhoods have higher home values, for 4 and +5 bedrooms.

Further, Toluca Lake, Encino, and Burbank Center neighborhoods have higher home values for 3 bedrooms size.

There are no relevant differences between the home values with 1 or 2 bedrooms, except for Burbank Center, in which 1-bedroom size is the highest.

6. Conclusion

According to these finds, we suggest working on the top five of the cuisine: Pizza, Mexican restaurants, Fast food, Sandwich places, and Asian food (Thai Restaurants, Chinese restaurants, and Seafood restaurants.)

Cluster 2 has almost the total of the venues; thus, the Restaurant's location should be placed within this area.

In sum, the client can launch the new Restaurant within these established market-top five venues and place it at any location in Cluster 2's neighborhoods. These neighborhoods are Winnetka, San Fernando, Tarzana, Granada Hills, North Hollywood, Encino, Northridge, Chatsworth, Lake View Terrace, Van Nuys, North Hills, Universal City, Lake Balboa, Sherman Oaks, Toluca Lake, Reseda, Studio City, Burbank, Porter Ranch, Panorama City, Sun Valley, Valley Village, Canoga Park, and Woodland Hills.

According to the selection of the Restaurant location, the client can proceed to decide his Home location. For the short term, he can follow the forecast and stick to the neighborhoods with negative home value prices: Burbank Center, Encino, Porter Ranch, Toluca Lake, Woodland Hills, and Tarzana.