# Clustering and Segmenting

# Los Angeles Communities Through K-Means Unsupervised Machine Learning

Cameron Olson

Github.com/olson996/Coursera\_Capstone

Jan 2019

## Introduction

This report aims to showcase some of the many skills learned throughout the IBM Professional Data Scientist Certification course powered through Coursera. The report coincides with a Jupyter Notebook that can be found on [github](https://www.github.com/olson996/Coursera_Capstone). Instructions on how to run the Jupyter Notebook can be found in the readme file.

Could I find the neighborhoods that match my level of income in Los Angeles and choose one that matches my venue interests? Answering this question will provide a lot of insights and descriptive cluster visualizations of the neighborhoods in Los Angeles. We aim to segment communities through un-supervised machine learning.

We'll be using Foursquare's API to get venues around a community, and cluster Los Angeles communities with the housing price of the area and it's top venues nearby.

The venue categories can be found [here](https://developer.foursquare.com/docs/resources/categories).

## Data

The data used in this research report is provided by the Los Angeles Almanac, https://www.laalmanac.com. This is a great resource that provides lots of economic, financial, geographic, and climatic information surrounding Los Angeles communities and zip codes.

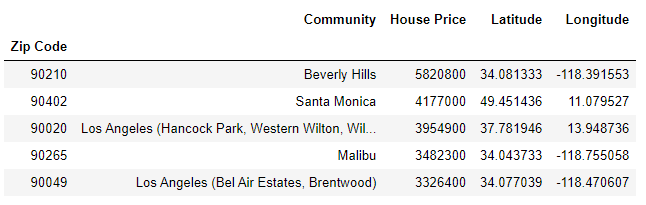
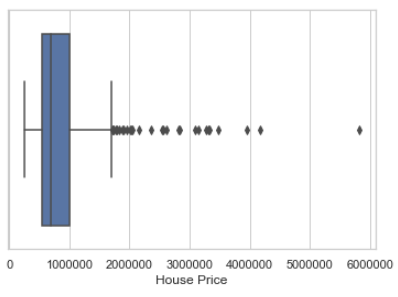
The data includes the zip code, community name, and average price of a single family home. We use geopy to convert zip code into coordinates.

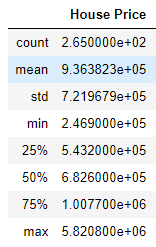
We use the requests, beautiful soup, and panda libraries to load in the data from the almanac into a data frame.

The coordinates of the neighborhoods will be needed to research the top venues for each zip code and geopy handles this for us.

## Methodology

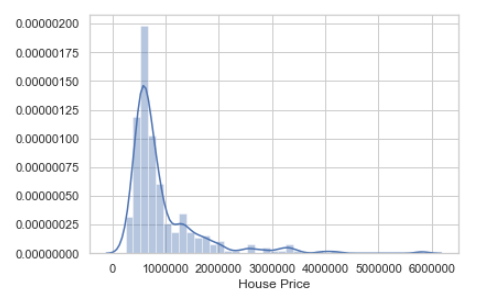
If you follow the link [here](http://www.laalmanac.com/economy/ec37b.php) you’ll see the web page provide by the Los Angeles almanac. The table is scraped using requests, beautiful soup, and cleaned with pandas. The resulting data frame that we’ll be working with only requires the most recent average single family housing price for Los Angeles zip codes, so we can drop the other years. One important pre-processing task is using a regular expression to convert the housing price into an integer. Table 1 shows the initial data frame after we resolve the coordinates. This table currently shows the top 5 average single family housing price out of 265 observations and coordinates for that community.

**Table 1**



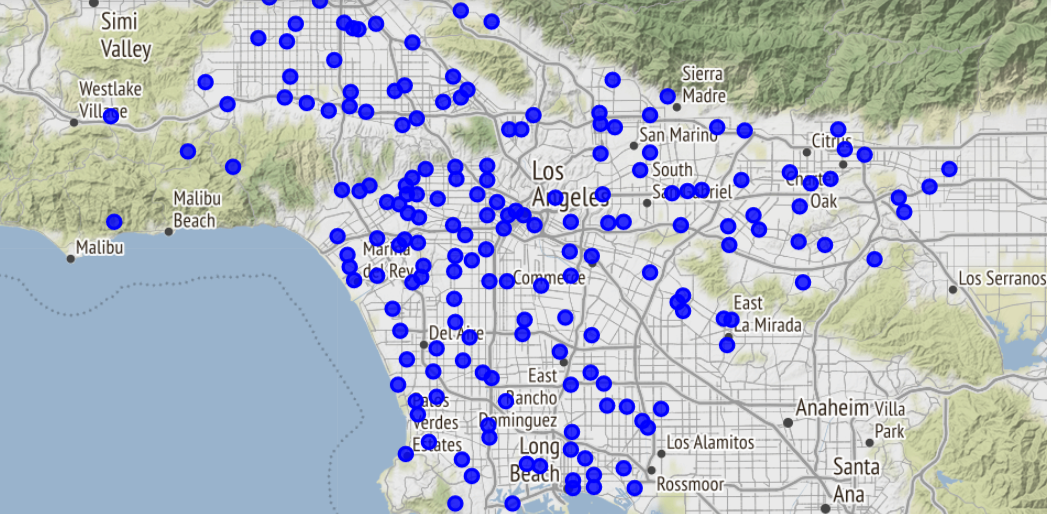
**Table 2 Figure 1**

Table 2 shows the quartiles and the box plot in Figure 1 visually shows some of the distribution of our data. 75% of the data is around one million dollars or less. There are a significant amount of what are statistically referred to as outliers above 1.7 million, because that is above 1.5 multiplied by the interquartile range. More specifically, the inter-quartile range would be 1.5 \* (1.007e+06 – 5.432e+05) = 695,700. Quartile 3 + 695,700 = 1,702,700. There are 28 zip code communities that meet the criteria of an outlier. They’ll remain in our dataset but we will keep them in mind during our analysis. We use a histogram and distribution plot to provide more insight into the underlying data in Figure 2. Seaborn provides a combination of these plots for us.



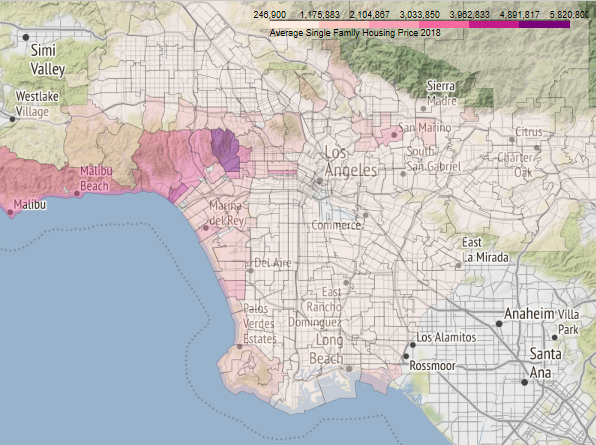
**Figure 2**

The distribution shown in Figure 2 is positively skewed. The majority of zip codes average single family homes around $600,000. A fatter tail on the right side of this distribution plot tells us that the data is positively skewed. It looks to be significantly positively skewed and it is because of well known and high priced real estate in locations like Beverly Hills and Santa Monica.



**Figure 3**

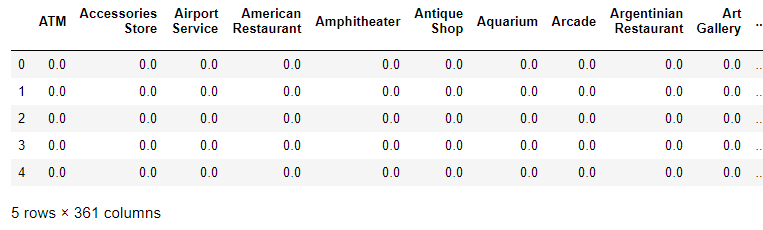
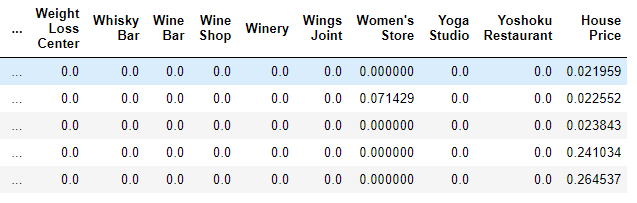
We use Folium to visualize the a scatter plot of communities in our data set in Figure 3. Since we have zip codes and the average single family housing price, we can create a choropleth map and bin the communities based of the price which we do in Figure 4.



**Figure 4**

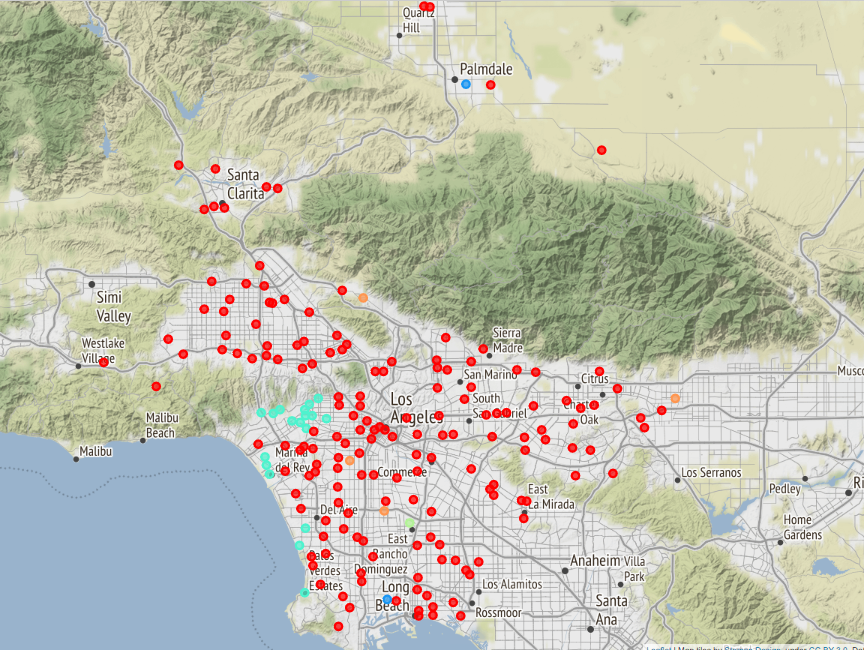
The average housing price for 265 zip codes is $936,382.26 and Figure 4 is yet another representation of how the average single home pricing varies across parts of LA. From the map we can see Beverly Hills in it's own bin (or the only one shaded purple. We know that Beverly Hills is an outlier in our data and it is significantly higher than the rest of the area. However, and it is intuitive, that the more expensive single family home housing prices are higher along the west coast and appear to increase with altitude around the hills and over the valley as well. We’ll come back to this choropleth map after visualizing the output of our K-Means machine learning model.

We use Foursquare’s API to pass coordinates and retrieve all venues within a specified range. Out of all venues found for a zip code we create a frequency table for each venue category. We will cluster and label these communities using this venue frequency table and the housing price that can be viewed in Table 3.

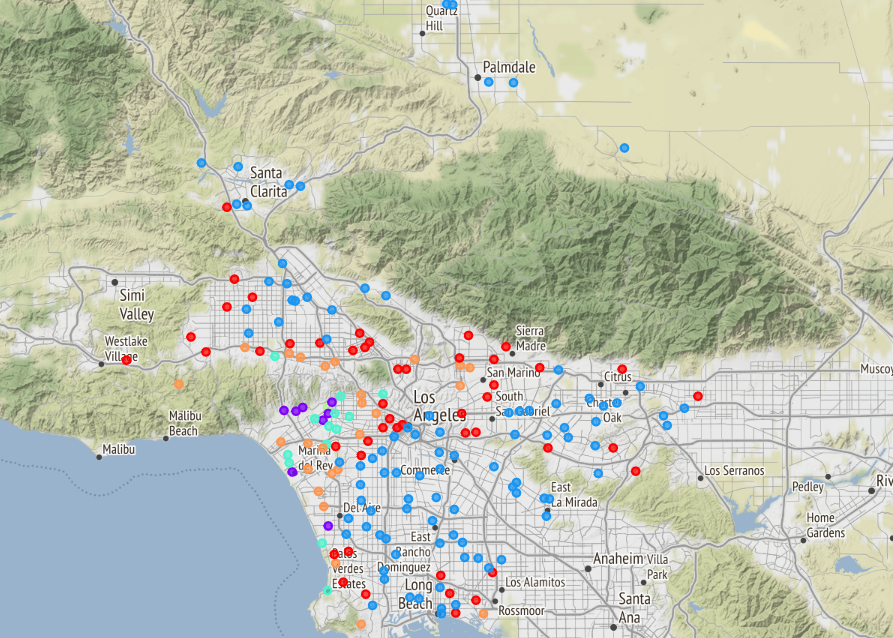


**Table 3**

As you can see in the last column of Table 3, this data frame contains the housing price now normalized. Below in the next images the clustering that performed in each one is using traditional K-Means Clustering. The only differences between them are the features that the algorithm uses. The features we are using in this case are the frequency of venues in an area and that area’s average single-family housing price. We will run K-Means with the housing price normalized, without it normalized and without the housing price at all and only using the most frequently found venues in an area.



**Unsupervised K-Means Clustering; K=6; Housing Price Normalized**



**Unsupervised K-Means Clustering; K=6; Housing Price Non-Normalized**

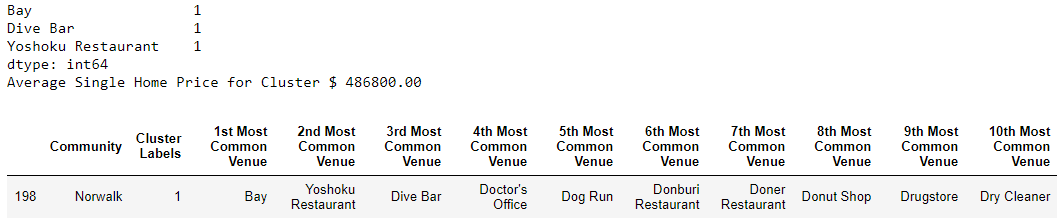
## Results

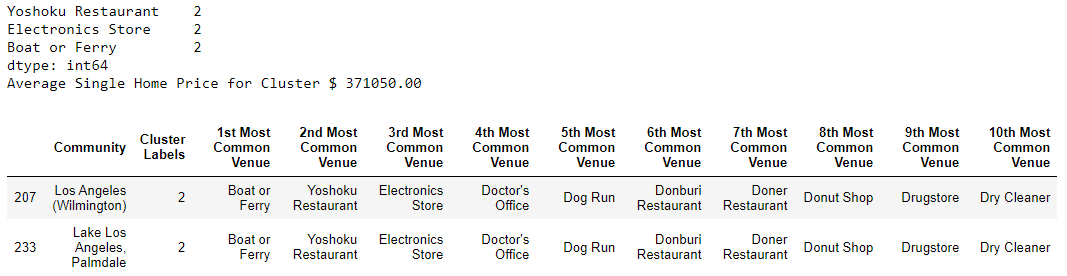
The following results show the segmented cluster data. The cluster tables are limited to 5 for space purposes but you can view more by downloading and running the Jupyter Notebook associated with this report.

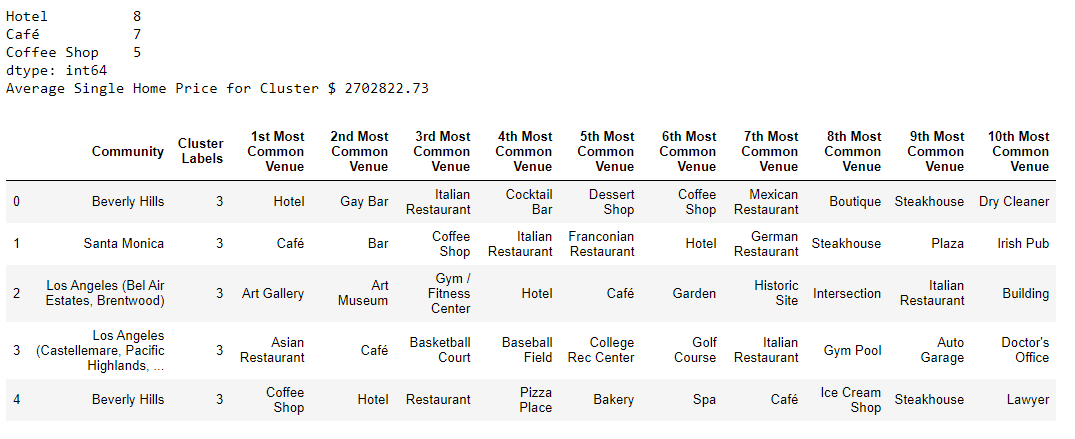
Each cluster also lists the value counts for the top three venue categories that appear most often among neighbors.

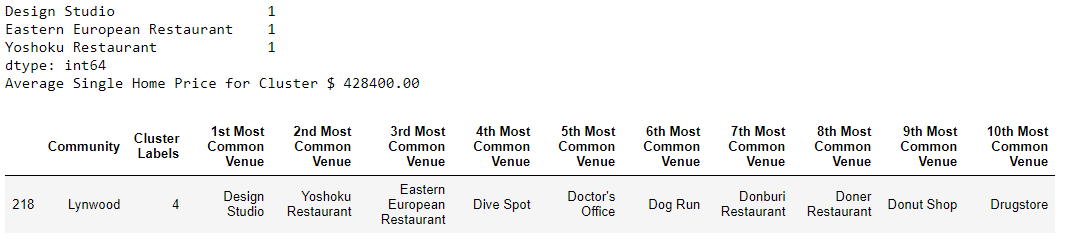
**Housing Price Normalized Clusters 0-5 Results:**

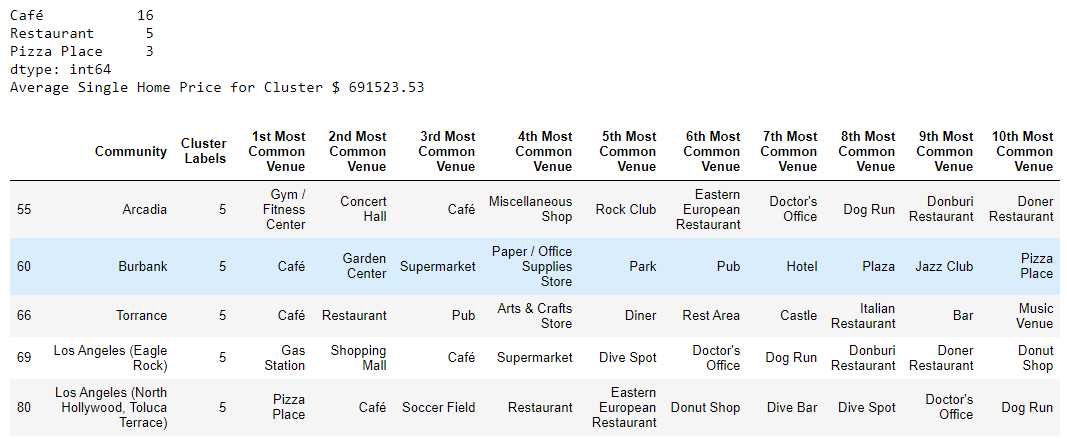




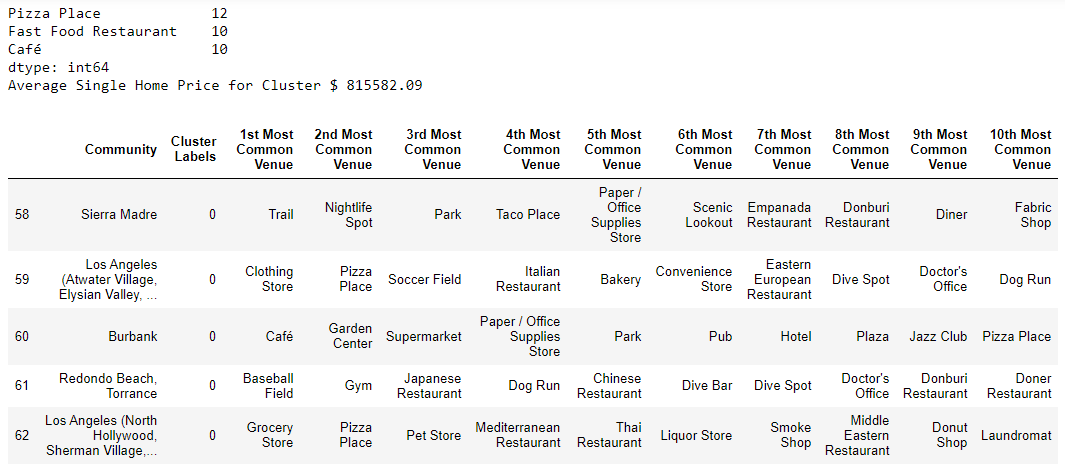




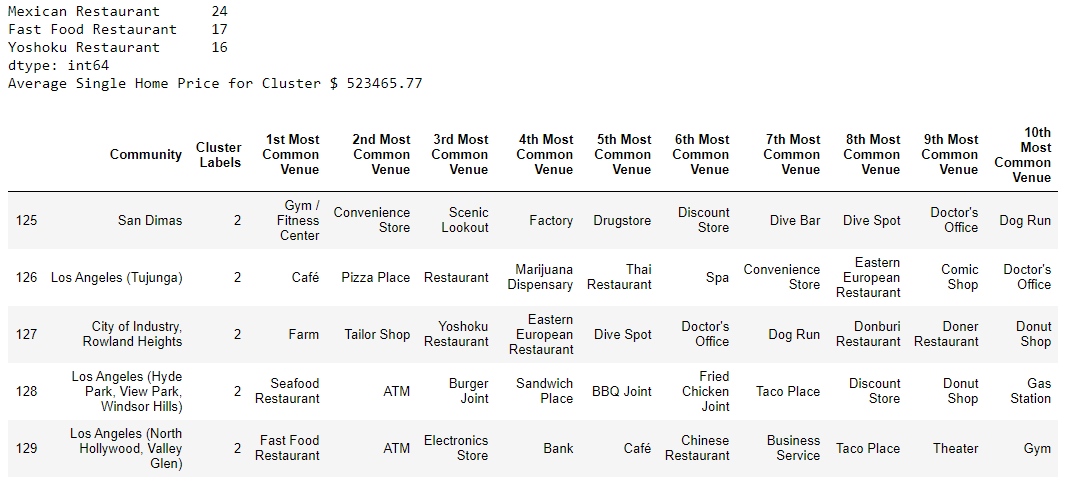


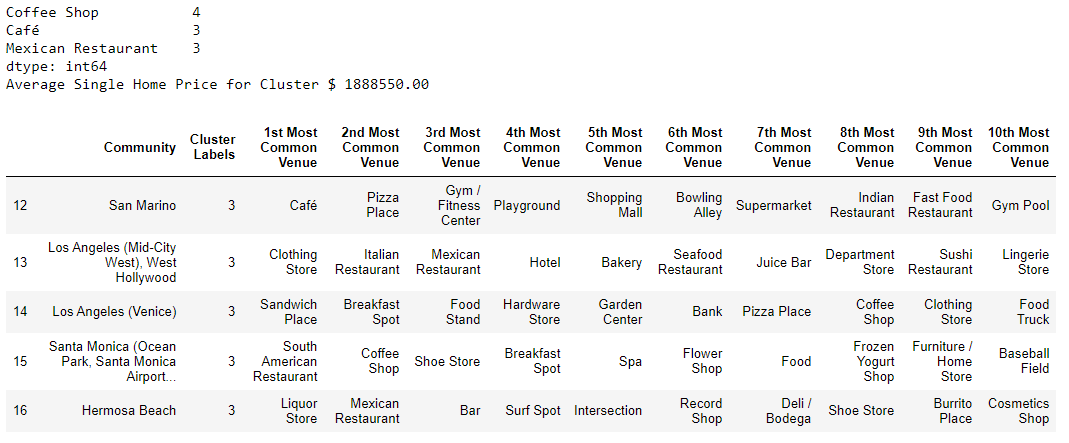


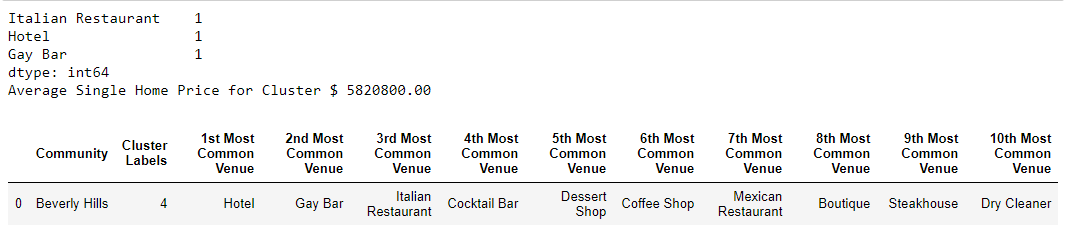
**Housing Price Non-Normalized Clusters 0-5 Results:**













## Discussion

When the data is not normalized K-Means assigns greater weight to the housing price in regard to its distance measure or similarity to other areas. In the non-normalized example above, the areas of higher housing price were clustered together, and lower ones were as well. The pattern of clustering seemed to resemble the choropleth map we viewed earlier in which the west coast and communities around Hollywood Hills and over the value had a higher mean single-family housing price. The normalized example doesn’t quite cluster in to delineated groups. The normalized example does cluster the higher mean priced housing areas, it fails to cluster the lower mean priced ones.

## Conclusion

Earlier when analyzing the data, we know that there is a disproportionality among the data we are viewing. The data is skewed and when normalized this significantly changes the way that K-Means clusters the data. Since the frequency table and the housing price are both between 0 and 1, K-Means doesn’t give the housing price more weight than any other venue. When the housing price isn’t normalized the algorithm is seems to be a weighted K-Means in which the housing price completely outweighs all of the venue frequency inputs. More analysis is needed, and we’d likely return to the K-Means algorithm to cluster inputs with only the housing price as a feature. That will be for another notebook! Feel free to look at the non-normalized and normalized results to see what top venues for each community are.