Finding Gentrification Areas within LA County using K-means Clustering

Introduction

There is a problem that is as old as property has been available for purchase to investors. This problem is, how can an investor know which property is the best one to purchase? There are many strategies to buying homes and many ways to be profitable. An investor could look for homes that are beat up, do some renovation, and sell the homes for a premium. An investor could look for homes that are underpriced and flip them. An investor could look for homes that have a low mortgage compared to how much rent could be brought in. And there are various other methods to achieving monetary gain. At the end of the day, the investor's goal is to make money, and this is the problem we will address today.

For this project we will focus on the investor who wants to buy a home that is an area that is going to gentrify. Gentrification is "the process of renovating and improving a house or district so that it conforms to middle-class taste" according to google. If an investor can find a house that is in an area that will gentrify, then they can buy it at a certain value, rent it for a few years, and then sell it at a much higher price when the area has improved. However, finding a neighborhood as it is undergoing these changes is a little more difficult than it may seem.

The purpose of this project will be to determine which homes in LA County are likely to undergo gentrification. By defining certain neighborhood types according to the location data, it will be possible to correlate different neighborhoods that may be undergoing a change. These certain neighborhoods should contain a similar types of establishments in near equivalent proportions. We will take a base neighborhood that is undergoing gentrification, like Compton, and then use a clustering algorithm to try and determine other neighborhoods that are undergoing similar changes. This will help the investor to know which areas of LA County to focus on. In a city with over 10 million people, finding the right areas is a great step for the investor.

Data

There will be two primary sources that are used for data. The first source is LA County area location data and the second is Foursquare data.

Location data for LA County will give us geolocations for each of the cities within LA County. The website http://www.laalmanac.com/geography/ge09.php contains the location data for 142 different cities. This data will be used to segment LA County into different sections. This will allow us to do analysis by groupings later. Different areas have different establishments and cultures that make them unique so it is important that we capture this.

Foursquare data is the second source of data that will be used. This can be found here: https://foursquare.com/. Using this data we will be able to see the kinds of establishments that are in areas that are currently gentrifying. The establishments in an area that is gentrifying will be much different than an area that has already reached the improved state. Think of the difference between the

kinds of establishments you would find in Santa Monica vs. Compton. There will no doubt be a noticeable difference. We are looking for the telltale signs that a neighborhood is about to gentrify or is already in the process.

Methodology

The exploratory data we will start with is the LA Locations data. From the LA Almanac webpage I was able to scrape the geolocations of each city in LA, this consists of 142 different cities. The data was presented in the table shown in Figure 1.

Place	Area (Square Miles)			Geographic Coordinates
	Total	Land	Water	
Los Angeles County	4,750.94	4,057.88	693.06	34.196398 -118.261862
Acton*	39.28	39.26	0.02	34.49626 -118.183891
Agoura Hills	7.82	7.79	0.03	34.148925 -118.763917
Agua Dulce*	22.84	22.84	0.01	34.501757 -118.320567
Alhambra	7.63	7.63	0.00	34.083571 -118.136444
Alondra Park*	1.14	1.11	0.04	33.889678 -118.335541
Altadena*	8.73	8.71	0.02	34.192212 -118.135589
Arcadia	11.13	10.93	0.21	34.132689 -118.036347
Artesia	1.62	1.62	0.00	33.867593 -118.080635
Avalon	2.94	2.94	0.00	33.332675 -118.330166
Avocado Heights*	2.84	2.71	0.14	34.036769 -118.001783
Azusa	9.67	9.66	0.01	34.138524 -117.912253
Baldwin Park	6.79	6.63	0.16	34.082825 -117.971286
Bell	2.62	2.50	0.12	33.979704 -118.179249

Figure 1 – LA Almanac Geolocation Data

After scraping this table, I was able to view the table in a Jupyter notebook, as shown in Figure 2.

	Place	Lat	Lon
0	Los Angeles County	34.196398	-118.261862
1	Acton*	34.49626	-118.183891
2	Agoura Hills	34.148925	-118.763917
3	Agua Dulce*	34.501757	-118.320567
4	Alhambra	34.083571	-118.136444
5	Alondra Park*	33.889678	-118.335541
6	Altadena*	34.192212	-118.135589
7	Arcadia	34.132689	-118.036347
8	Artesia	33.867593	-118.080635
9	Avalon	33.332675	-118.330166
10	Avocado Heights*	34.036769	-118.001783
11	Azusa	34.138524	-117.912253
12	Baldwin Park	34.082825	-117.971286
13	Bell	33.979704	-118.179249

Figure 2 - Scraped Geolocation Data

Once the data was in this format I was able to plot it to make sure it was accurate using Folium. This is shown in Figure 3.

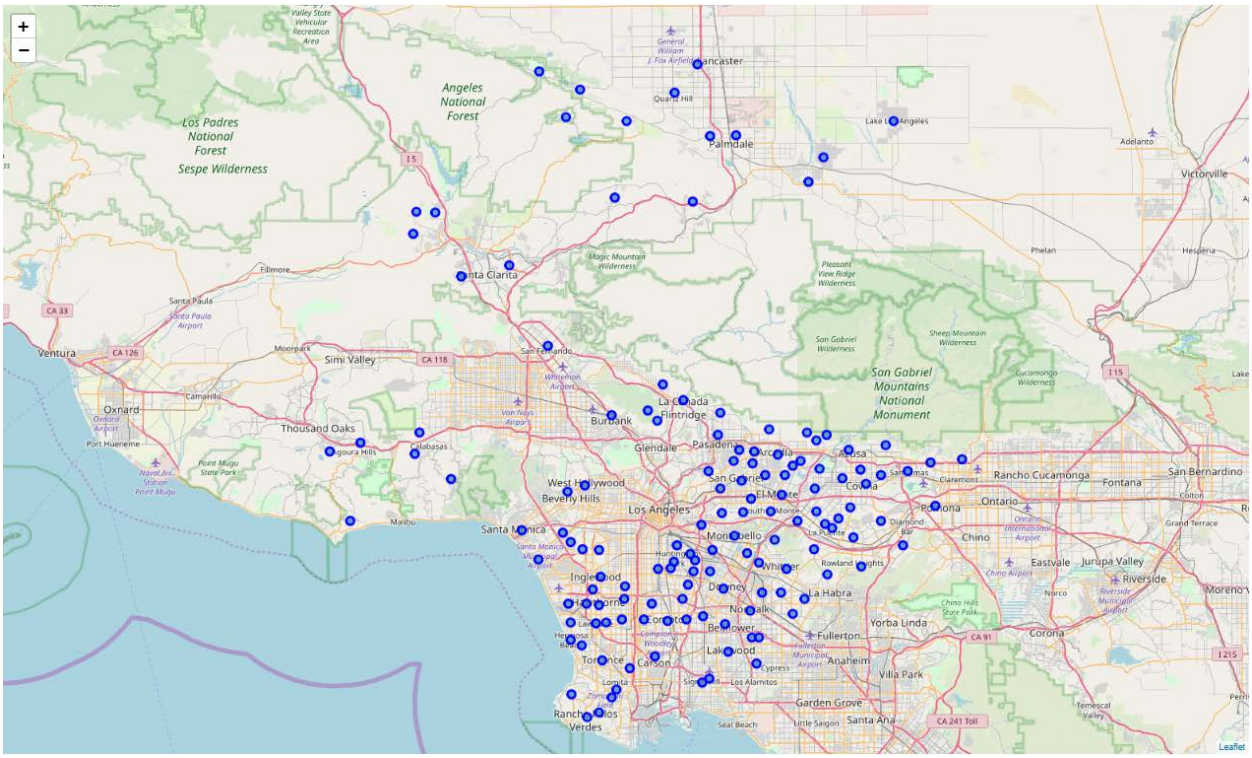

Figure 3 - Plots of geolocations from scraped data

Upon viewing this and seeing that it was accurate, I went forward to grab the foursquare data. From this I set a max of 300 points of interest per location within a radius of 2500. This grabbed a total of 368 unique establishment types. It also grabbed 10,983 different establishments.

From here it was time to choose my machine learning algorithm to cluster the data. Naturally a clustering algorithm was a good choice since similar types of cities are being looked for. I chose the k-means algorithm due to its robust nature. I then ran the data in the K-means clustering algorithm. I used 20 clusters since there are a lot of cities. I tried various amounts of clusters and found that 20 was a good number. I also used 150 recursive attempts from the algorithm to find the best fit. After this was ran, I was returned the labels for the k-mean clusters. Using these labels and the location data, I was able to plot the clusters in Folium as shown in Figure 4.

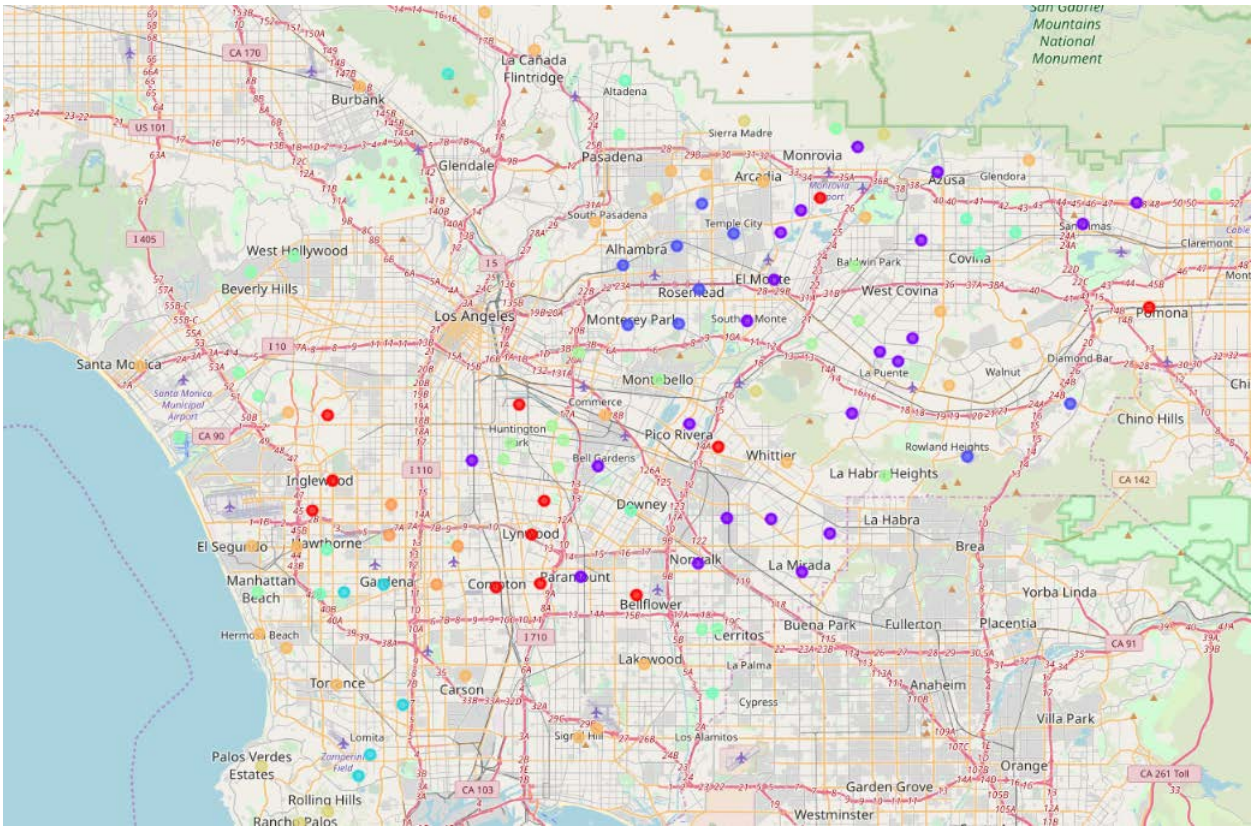

Figure 4 - Cities have been clustered to indicate which ones are similar. Clusters are indicated by color.

Results

As can be seen above, Compton is in the red cluster. So I grabbed all the red cluster points and listed them out as shown in Figure 5.

	Place	Lat	Lon	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
15	Bellflower	33.887821	-118.12725	0	Convenience Store	Sandwich Place	Mexican Restaurant	Burger Joint	Coffee Shop
27	Compton	33.892614	-118.227374	0	Sandwich Place	Fast Food Restaurant	Mexican Restaurant	Burger Joint	Discount Store
39	East Rancho Dominguez*	33.894834	-118.195588	0	Mexican Restaurant	Fast Food Restaurant	Pizza Place	Convenience Store	Sandwich Place
57	Inglewood	33.956068	-118.344274	0	Fast Food Restaurant	Convenience Store	Mexican Restaurant	Grocery Store	Burger Joint
71	Lennox*	33.938064	-118.358543	0	Mexican Restaurant	Pizza Place	Hotel	Convenience Store	Sandwich Place
77	Lynwood	33.923962	-118.201647	0	Mexican Restaurant	Fast Food Restaurant	Burger Joint	Sandwich Place	Pizza Place
93	Pomona	34.058595	-117.761266	0	Mexican Restaurant	Convenience Store	Fast Food Restaurant	Pharmacy	Pizza Place
113	South Gate	33.944159	-118.192761	0	Convenience Store	Fast Food Restaurant	Pharmacy	Mexican Restaurant	Coffee Shop
114	South Monrovia Island*	34.123435	-117.99586	0	Mexican Restaurant	Racetrack	Coffee Shop	Fast Food Restaurant	Breakfast Spot
126	Vernon	34.001123	-118.210869	0	Convenience Store	Fast Food Restaurant	Burger Joint	Mexican Restaurant	Food Truck
127	View Park-Windsor Hills*	33.994473	-118.347762	0	Park	Southern / Soul Food Restaurant	Pizza Place	Mexican Restaurant	Fast Food Restaurant
137	West Whittier-Los Nietos*	33.976001	-118.068925	0	Mexican Restaurant	Fast Food Restaurant	Convenience Store	Pizza Place	Sandwich Place

Figure 5 - Potential areas where gentrification may be occurring.

This is the data we are looking for. Having lived in LA most of my life, I have personally visited Compton and know that it is gentrifying. These other areas are all similar to Compton in the type of establishments that are most common. We can take a closer look at these areas because there is a good chance that they are also gentrifying. This is great information for an investor. Now instead of having to think through 142 different cities in the county, the investor can focus their effort on just 12.

Discussion

Living in LA for most of my life, I can tell you that this analysis is pretty good. These 12 cities are indeed all similar in ways to Compton. I have not travelled to all the parts of LA as it is a very big place, but using this information I am pretty confident that these 12 cities have what an investor would be looking for. Viewing the map above (Figure 4) it can be seen that the clusters are pretty good. They tend to be similar near the ocean, and move outward staying somewhat similar based on distance, though not entirely. The neighborhoods near the ocean tend to be more expensive and have a certain atmosphere about them. This just gives more credence to the analysis.

Conclusion

Using the right data and methodology, an accurate conclusion can be made about what area in LA may be gentrifying. The location data was able to segregate LA into meaningful locations, and the Foursquare data was able to show which establishments were common among different cities. I have a firsthand knowledge that Compton is undergoing gentrification, so viewing the cities that were clustered with Compton is a pretty good sign that those areas also have the potential of gentrification. The 12 cities to keep an eye out for are seen in Figure 5.