|  |
| --- |
| Battle of the Neighborhoods |
| An Analysis Using Data Science Methodology |
| Ryan Belfer |

**Introduction**

Every city has its own unique culture, and even within each city, neighborhoods can wildly differ. From decade to decade, this culture can change drastically. In 1950, New York City was over 90% white; in the 2010s, New York continues to grow as one of the most racially diverse cities on the planet. In the ‘70s and ‘80s, crime was rampant in many parts of the city, but by the turn of the century, the crime rate had dropped below the national average. The change in the character of a city affects the culture, politics, and economy of the neighborhoods. However, it is difficult to tell which neighborhoods will change, how they will change, and over what timeframe the change will occur.

One of the trends of the 21st century still going on is the reversal of urban flight. The upper and upper-middle classes are moving back into cities, and causing gentrification of many of the poorer neighborhoods. The ethics of gentrification can be debated on, but there is no doubt that the property values in gentrified neighborhoods rise steeply. Getting a foot in the door early, whether it be by opening a strategically located business, or more popularly by buying cheap real estate, can lead to huge dividends if the neighborhood ends up becoming gentrified.

This study will take a look at Los Angeles, California, a city that is generally not known for its gentrification, and attempt to predict which neighborhoods will become gentrified. With these predictions, people interested in the real estate market could buy properties that currently sell for cheap but would end up highly valuable. Alternatively, city planners and other interested parties could use the analysis to decide which neighborhoods to focus on protecting against gentrification.

**Data Gathering**

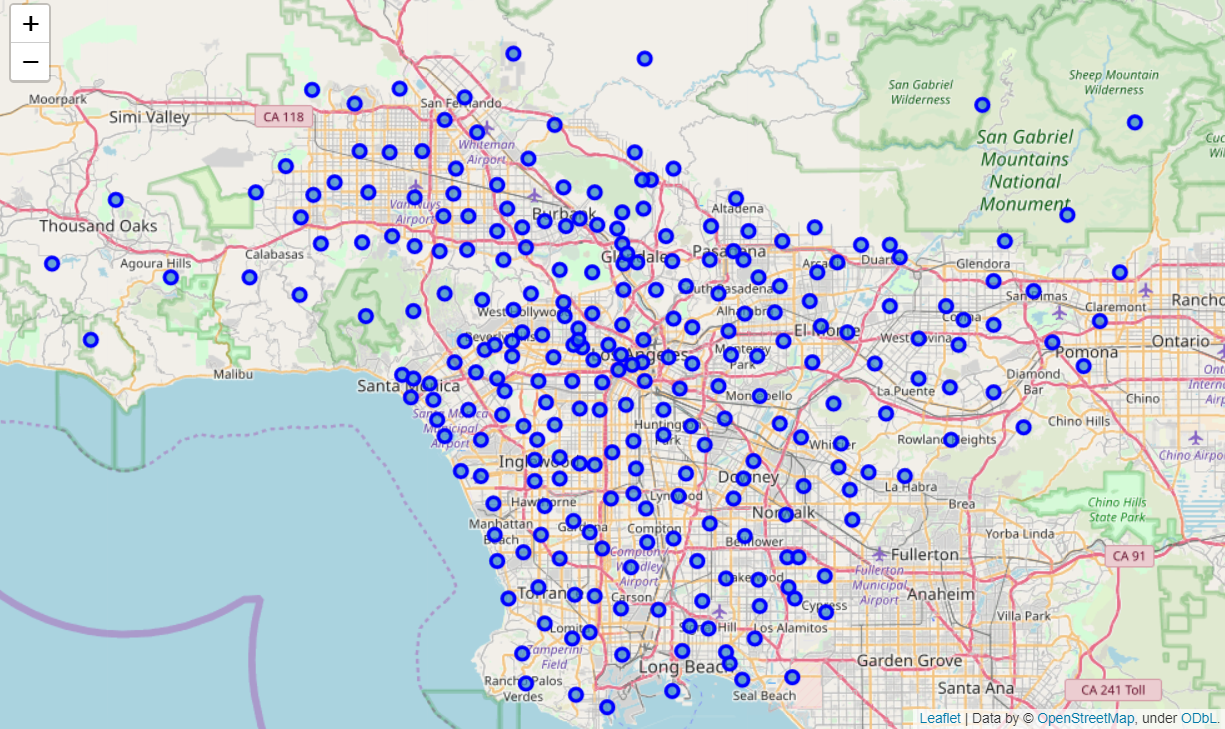
We will use data from multiple sources in order to perform the analysis. In order to perform clustering, we need to make sure we have the ZIP codes of Los Angeles. We will get these ZIP codes from the [LA Almanac Median Income by ZIP Code Table](http://www.laalmanac.com/employment/em12c.php). Alternatively we could get from [LA Almanac Community Table](http://www.laalmanac.com/communications/cm02_communities.php), but the Median Income will be useful for later. This table is useful to see any ZIP codes that are omitted from our analysis. We will also need to know where each ZIP code is geographically. We will use a JSON file that can be downloaded on the [LA Times website](http://boundaries.latimes.com/set/zip-code-tabulation-areas-2012/) that marks the latitude and longitude of each ZIP code. The data that the Foursquare API retrieves can be joined to the ZIP Code/Neighborhood data using the latitude and longitude. Lastly, we will use the LA Almanac’s [Median Income by ZIP Code Table](http://www.laalmanac.com/employment/em12c.php) again to build a picture of how wealth is distributed in the Los Angeles area. We can use a choropleth visualization to see which neighborhoods qualify for gentrification. The median income for all of Los Angeles is approximately $70,000, so we would take a look at ZIPs below that income that have the most similarity to the richer ZIP codes.

Although we would have liked to use the neighborhood names to do this analysis, we cannot, as each neighborhood does not fit nicely into one ZIP code. It will be much more useful to do an analysis using the ZIP codes themselves as pseudo-neighborhoods, and make a recommendation based on the ZIP code. The almanac table omits a couple dozen ZIP codes, so the recommendation is not as robust as it could be. The analysis is also limited by having to set a constant radius for the API call; the ZIP code areas vary wildly, meaning that a 500 meter radius would cover a small portion of one ZIP, while it might extend beyond the boundaries of another.

**Methodology and Analysis**

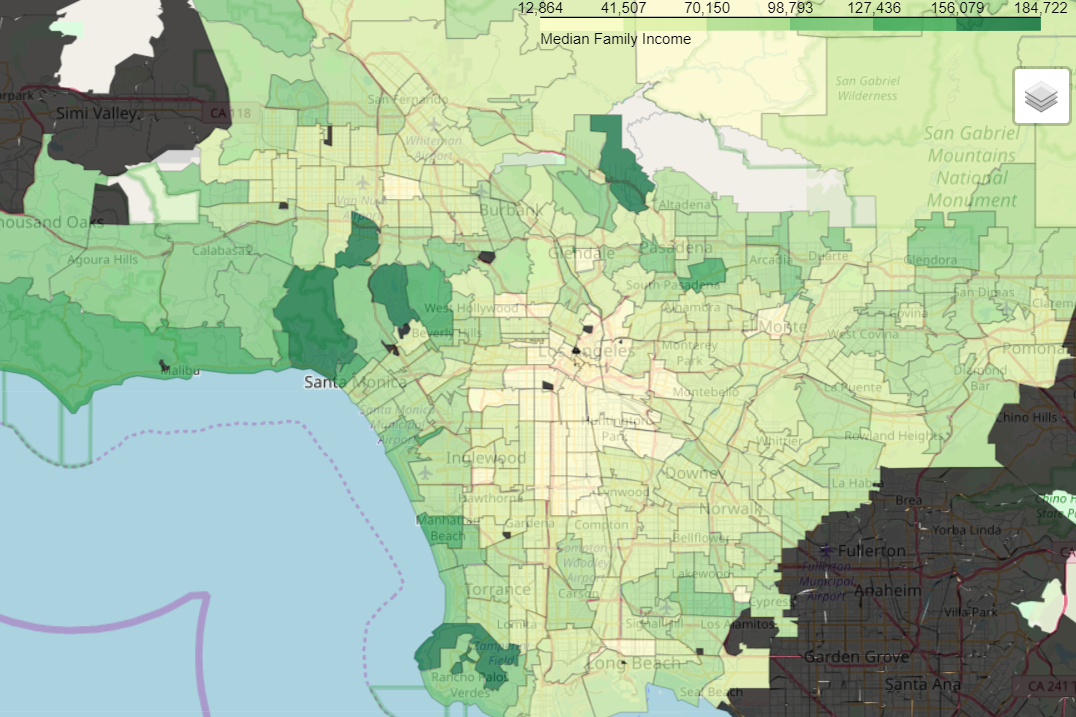
Before starting the analysis, the data had to be munged and processed into usable tables. For the data on the ZIP codes, neighborhood names, and median incomes, we created a list for each from the LA Almanac page, and then combined the lists into a Pandas dataframe. We also created a list of ZIP codes that contain the term “Los Angeles” by using regular expressions; although we did not use this list for any dataframe construction, it may have been useful for analysis of just the city area, instead of including surrounding places such as Burbank or Chino. Lastly, the median incomes from this list had to be converted into integers for further analysis.

The data frame was then merged with the LA Times file to add latitude and longitude to the table. We had to convert the latitude and longitude into floats. We then mapped the data using Folium to see the geographic distribution of the data. There were a few outliers that were north of the city, such as 30 miles from downtown near Santa Clarita, or 60 miles away at Edwards Airforce Base. We restricted the latitude to south of Santa Clarita to get rid of those northerly neighborhoods. The final map looked like the following:



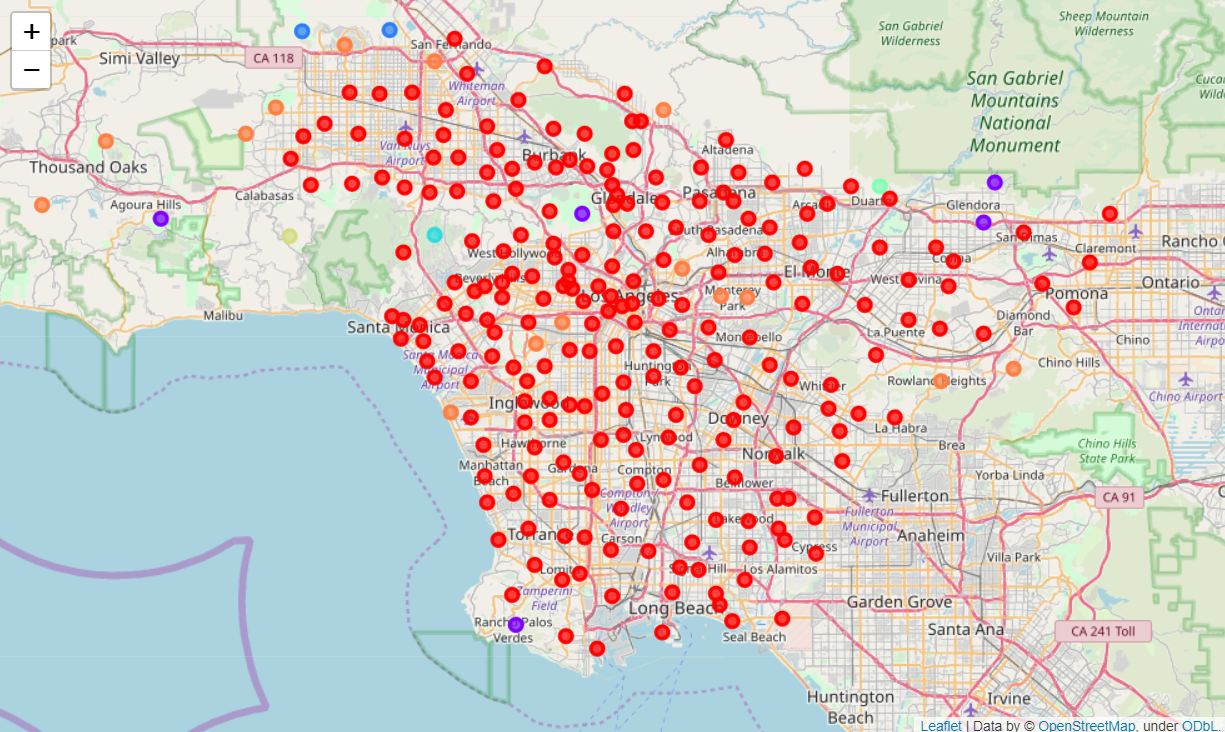
There are still some geographical outliers, such as in Thousand Oaks or up in the San Gabriel Mountains, but none of them are in the Inland Empire region to the east.

We can also generate a choropleth map using the median income data. The data ranges from about $12,000 to over $180,000, though there are only a handful of communities with a median income that high. The median income for the region is approximately $70,000, so we would expect most communities to be within the $40,000 to $100,000 range, since the standard deviation is about $30,000. This can be seen in the choropleth map itself, where all of the neighborhoods within a standard deviation are a pale green color:

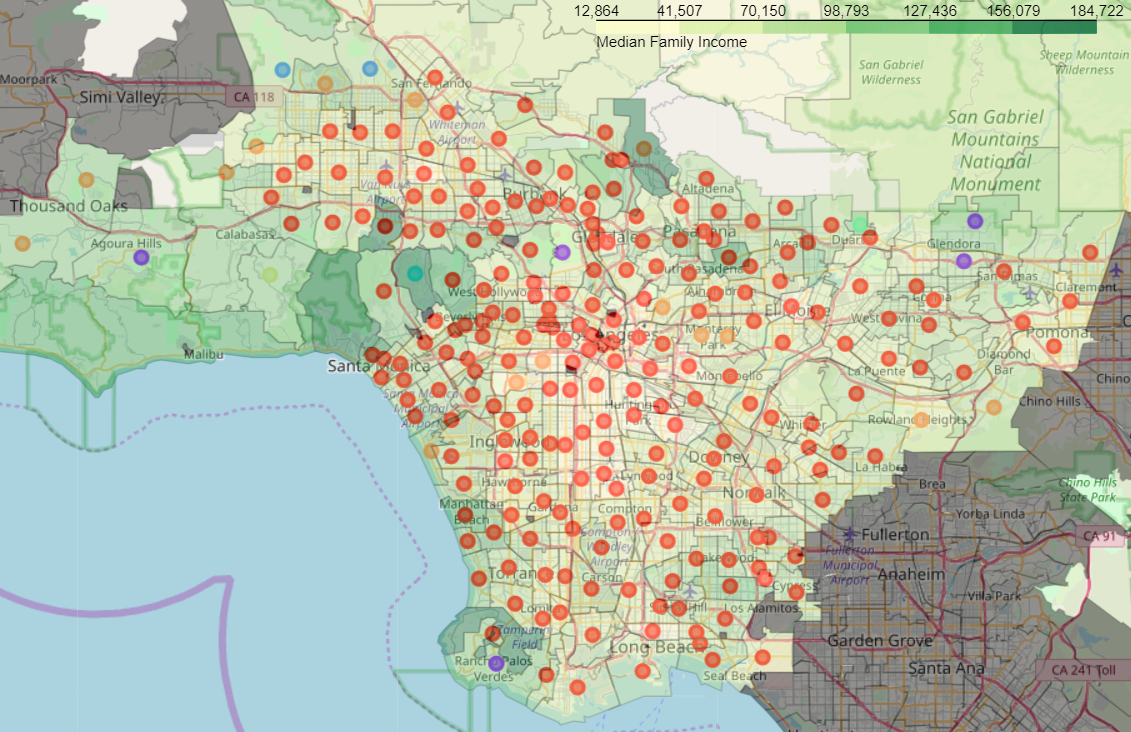


We then brought in the Foursquare data using the Foursquare API. The latitudes and longitudes of each venue is used to group them into the neighborhoods. We used a radius of 800 meters, or about half a mile, since the majority of ZIP codes of Los Angeles are at least a mile in length. We had to use a larger radius for Los Angeles’ over 500 square miles than for New York’s, where all of Manhattan’s few dozen ZIP codes are contained within 22 square miles.

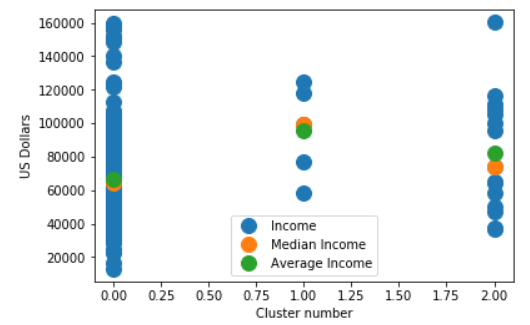
From Foursquare, we found 410 unique types of venues spread over nearly 200 neighborhoods. To group the neighborhoods by venue similarity, we used **K Means clustering**. We chose 7 clusters because the hope is to divide the neighborhoods into groups of around 30. Having too many clusters would create clusters of size 1, and having too few would mean that there is not enough differentiation. After splitting into clusters and plotting another map, we get the following:



By combining this map with the choropleth, we can get a sense of what the median income might be for each cluster visually:



Before proceeding further with the analysis, we made a decision to reduce the amount of data by removing any clusters that did not have enough data. We chose a threshold of five points to keep the cluster, because any less and we could not even make a boxplot. Out of the 7 clusters, only 3 of them had five or more points. This is represented on the map by the red, orange, and purple clusters. We chose to visualize the data using a scatter plot instead of a box plot because it gives a better sense of the distribution of the data above and below the mean/median than just five lines plus outliers, especially for the smaller clusters. We are also not concerned with outliers in the data.



The plot shows that there is a large range of incomes in each cluster. In clusters 0 and 2, the income is skewed upwards, shown by the average being higher than the median.

After counting the number of neighborhoods per cluster and the range in incomes, it was clear that cluster 0 was much too large to make any sort of conclusions on, and cluster 1 was too small; additionally the income was not skewed upwards, meaning that that we were not grouping a low-income area with many high-income areas. So, we reduced the data further to just the last cluster, which contained 15 neighborhoods. From these 15, we filtered down to the neighborhoods that were below the median of the group, as gentrification is much less likely to occur in already-high-income areas.

|  |  |  |
| --- | --- | --- |
| **ZIP** | **Neighborhood** | **Income** |
| 90008 | Los Angeles (Baldwin Hills, Crenshaw, Leimert Park) | 36641 |
| 90018 | Los Angeles (Jefferson Park, Leimert Park) | 37341 |
| 90032 | Los Angeles (El Sereno, Monterey Hills) | 47370 |
| 91755 | Monterey Park | 49755 |
| 91754 | Monterey Park | 58056 |
| 91304 | Los Angeles (Canoga Park, West Hills), Box Canyon | 64367 |
| 91748 | City of Industry, Rowland Heights | 64983 |
| 91345 | Los Angeles (Mission Hills) | 74193 |

The above table contains the 8 neighborhoods that fall within the parameters of our analysis. They are similar to 7 neighborhoods that have high median incomes.

**Results and Discussion**

By clustering the neighborhoods using K-Means, we were able to find the neighborhoods that were most similar. The idea is that gentrified neighborhoods would start out as lower income neighborhoods and slowly become more similar to higher income neighborhoods. So, lower income areas that are already somewhat similar to higher income ones would be a good place to assume that gentrification has either already started or has the potential for such.

For calculating the average of the median incomes to check the skewness, we assumed that the populations of the neighborhoods are equal. We could have found a more accurate average by finding a table of population by ZIP code or the [U.S. Census Population Factfinder](https://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml?src=bkmk), which shows that the populations of ZIP codes are not at all constant. Even in the final ZIP code results, the population ranges between 20,000 and 50,000. The median of the median incomes is more useful in this case, which allowed us to split the cluster into richer and less rich areas.

To attempt further reduce the list of possible neighborhoods to invest in or protect, we can look at the population densities of each area, using a source such as the [Statistical Atlas](https://statisticalatlas.com/zip/91755/Population). For example, the density of 90018 is 17,000 people per square mile, while 91755 is about half that. We can conclude that 90018 is much more urbanized, meaning that it fits more the trend of gentrification in urbanized areas. However, heavily suburbanized areas on the edge of urbanization, which we can classify 91755 as, can also become gentrified.

While population density may not reduce the number of results, we can use geographic location, such as miles from the city center. Towns and cities closer to downtown Los Angeles would be more attractive to invest in, since there would be more human activity. The below table shows the distance from the city center:

|  |  |
| --- | --- |
| **ZIP** | **Distance (miles)** |
| 90008 | 6 | |
| 90018 | 4 | |
| 90032 | 5 | |
| 91755 | 7 | |
| 91754 | 6 | |
| 91304 | 25 | |
| 91748 | 20 | |
| 91345 | 20 | |

There are clearly two groups of ZIP codes: one group is within 10 miles of city center, while the second group is 20 miles or further. The second group also contains the higher median incomes of the group, which means there might not be as much room for gentrification.

**Conclusion**

In this study, we synthesized income and geospatial data of neighborhoods in Los Angeles. A city chosen for its relatively low rate of gentrification compared to other large population centers such as New York, Boston, or Seattle. We used venue data from the Foursquare API to cluster neighborhoods based on venue similarity, with the logic that the poorer neighborhoods clustered with richer neighborhoods would have the possibility of gentrification. We then considered population density and distance from city center to reduce the resulting table even further. After all of this analysis, the final group of neighborhoods that it is recommended to invest in or protect due to gentrification is:

**Leimert Park**

**El Sereno**

**Monterey Park**

There are many other factors that cause gentrification that cannot be gleaned from simple geographical and income data. A better analysis might include neighborhoods from other cities, with venue data from the present and 10-15 years ago so we can see which neighborhoods are currently most similar to those neighborhoods from the past.