|  |
| --- |
| 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 Methodology**

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.