Applying Clustering Analysis to the Neighborhoods of Los Angeles

Grant Belsterling

We set out to explore the relationships between neighborhoods in Los Angeles. I apply clustering algorithms to gain insight into which Los Angeles neighborhoods form cohesive, similar groups. I then examine what it is that makes them similar, according to the data, and what this says about Los Angeles and, more broadly, urban sociology as a whole.

```r
library(dplyr)
library(readr)
library(factoextra)
library(mice)
library(ggrepel)
library(cluster)
library(ggplot2)
set.seed(156)
```

First let's import the data we scraped from the python script 'scrapeLA.py'. This data comes from city-data.com, a website that aggregates census tract data by city neighborhood.

```r
neighbData <- read_csv("~/Documents/Data Science /Scrape_LA/neighbData.csv")
```

We rename columns to sensible, descriptive, and easy to work with names. This list includes all the variables scraped from city-data.com and serves as a reference for the variables used for the rest of the analysis.

```r
colnames(neighbData) <- c('pct_poverty',
                          'occupations_males_technical',
                          'female_population',
                          'housing_prices_sparse',
                          'neighborhood',
                          'page_year',
                          'housing_prices',
                          'pct_k12',
                          'occupations_males_financial',
                          'avg_num_cars_apartments',
                          'avg_num_cars_houses',
                          'male_median_age',
                          'population_density',
                          'area',
                          'median_rent',
                          'zips',
                          'occupations_males_social_services',
                          'pct_born_in_state',
                          'num_5plus_bedrooms',
                          'avg_household_size',
                          'pct_military',
                          'pct_non_english_speakers',
                          'occupations_males_management',
                          'pct_mortgage',
                          'pct_married_both_working',
                          'pct_never_married_males',
                          'median_income',
                          'occupations_males_education',
                          'male_population',
                          'total_population',
                          'num_bedrooms'
                          )
```

We remove irrelevant and sparse columns.

```
neighbData <- select(neighbData,-c(occupations_males_financial, occupations_males_social_services,occupa
```

We clean the data. Strip commas, percent signs, or dollar signs and convert data to numeric.

```
neighbData$pct_poverty <- as.numeric(gsub('%','',neighbData$pct_poverty))
neighbData$housing_prices <- as.numeric(gsub('\\$|,','',neighbData$housing_prices))
neighbData$pct_k12 <- as.numeric(gsub('%','',neighbData$pct_k12))
neighbData$male_median_age <- as.numeric(gsub(' years','',neighbData$male_median_age))
neighbData$avg_household_size <- as.numeric(gsub(' people','',neighbData$avg_household_size))
neighbData$median_rent <- as.numeric(gsub('\\$|,','',neighbData$median_rent))
neighbData$pct_born_in_state <- as.numeric(gsub('%','',neighbData$pct_born_in_state))
neighbData$pct_military <- as.numeric(gsub('%','',neighbData$pct_military))
neighbData$pct_non_english_speakers <- as.numeric(gsub('%','',neighbData$pct_non_english_speakers))
neighbData$occupations_males_management <- as.numeric(gsub('%','',neighbData$occupations_males_managemen
neighbData$pct_mortgage <- as.numeric(gsub('%','',neighbData$pct_mortgage))
neighbData$pct_married_both_working <- as.numeric(gsub('%','',neighbData$pct_married_both_working))
neighbData$pct_never_married_males <- as.numeric(gsub('%','',neighbData$pct_never_married_males))
neighbData$median_income <- as.numeric(gsub('\\$|,','',neighbData$median_income))
neighbData <- data.frame(neighbData)

#add row names
row.names(neighbData) <- neighbData$neighborhood


#remove two sparse rows
neighbData <- neighbData[!rownames(neighbData) %in% c("Hancock Park", "Los Angeles, California Nei"),]

#scale and standardize the data for use in clustering algorithms
scaledData <- data.frame(scale(select(neighbData, -c(neighborhood,zips))))
```

We noticed that some of our data is missing. Let's see how bad the problem is.

```
pct_missing <- function(x){sum(is.na(x))/length(x)*100}
apply(scaledData,2,pct_missing)
```

```
##                  pct_poverty            female_population
##                     0.000000                     0.000000
##               housing_prices                      pct_k12
##                    25.675676                     0.000000
##          avg_num_cars_houses              male_median_age
##                     1.351351                     0.000000
##           population_density                         area
##                     0.000000                     0.000000
##                  median_rent            pct_born_in_state
##                     0.000000                     0.000000
##           avg_household_size                 pct_military
##                     0.000000                    54.054054
##     pct_non_english_speakers occupations_males_management
##                     0.000000                     6.756757
##                 pct_mortgage     pct_married_both_working
##                     1.351351                     1.351351
##      pct_never_married_males                median_income
##                     0.000000                     0.000000
##              male_population             total_population
##                     0.000000                     0.000000
```

We see that 27% of neighborhoods are missing housing_prices and 55% of neighborhoods are missing pct_military. We can't impute this large of a portion of missing data without introducing significant bias into the data set.

Let's drop them for now.

```
scaledData <- select(scaledData,-c(pct_military,housing_prices))
```

And impute the few remaining NAs, since NAs cannot be handled by many distance metric implementations.

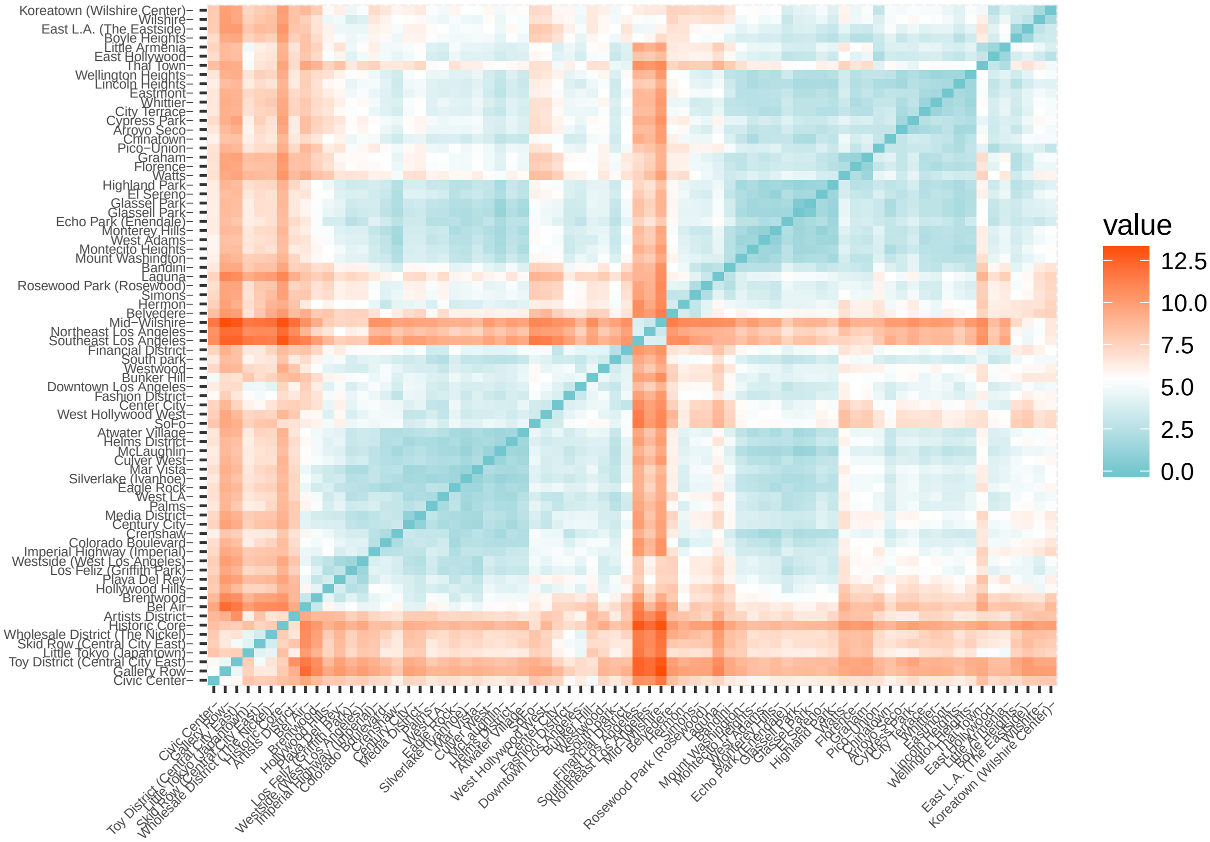
```
imputation <- mice(scaledData,m=5,maxit=20,meth='pmm')
```

```
imputedData <- complete(imputation)
```

Now we're finally ready to explore clustering algorithms. The first step is to determine a distance metric for this data.

Let's visualize what the euclidean metric looks like; its a standard metric with an intuitive geometric interpretation: in 3D space, it's simply a straight line between two points (which forms the shortest path as we intuitively understand it).

```
dists_euclidean <- get_dist(imputedData, stand = TRUE, method = "euclidean")
fviz_dist(dists_euclidean,
   gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"), lab_size=5)
```



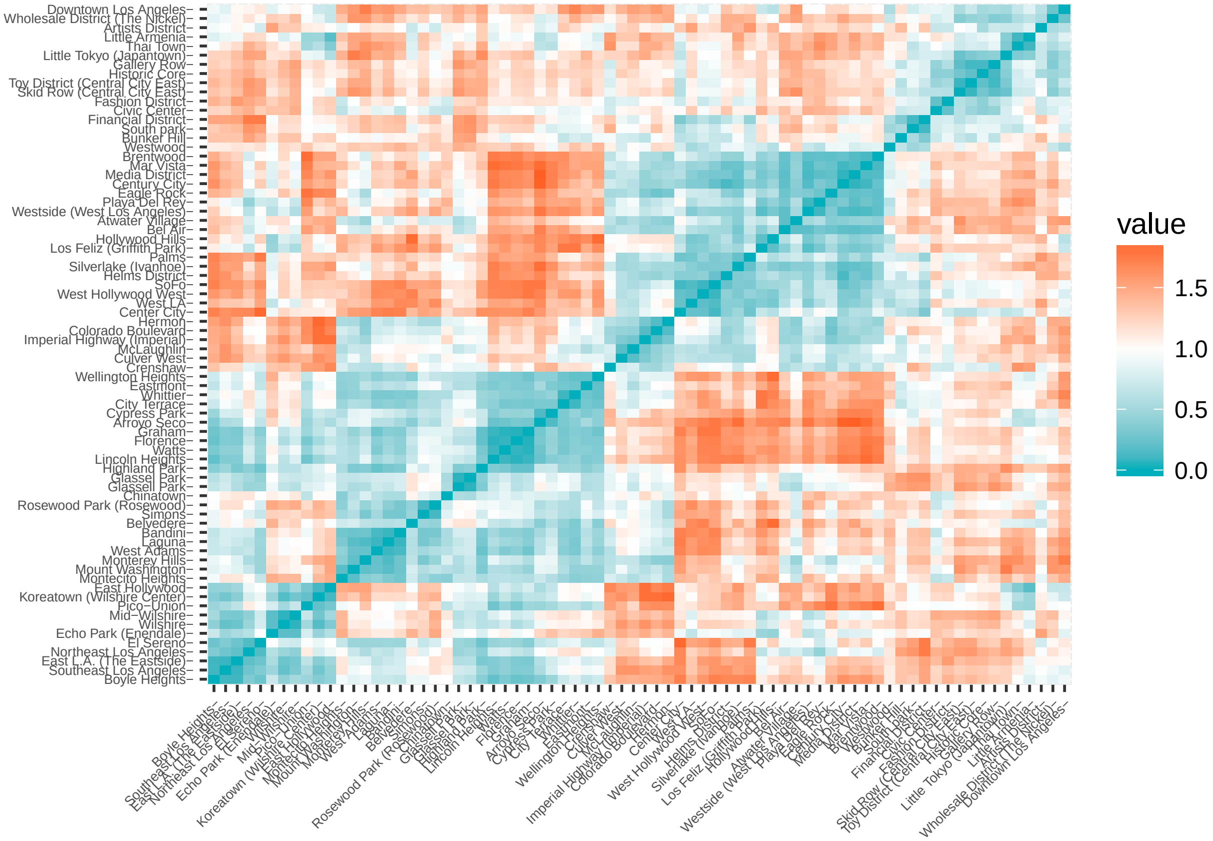
We notice an odd cross pattern of red formed by Mid-Wilshire, Northeast LA, and Southeast LA. By the euclidean metric, these three neighborhoods are very distant (and there 'dissimilar') from every other neighborhood. They're outliers in some way. We have the prior knowledge of LA to know that these neighborhoods are actually larger sections of the city rather than neighborhoods in their own right. Each section contains multiple sub-neighborhoods. Mid-Wilshire contains Little Ethiopia, Park La Brea, the Miracle Mile area, etc; Northeast LA contains Highland Park and Eagle Rock; Southeast LA contains Watts, Florence, etc. This hunch is supported by the data: the area of Mid-Wilshire, Northeast LA and Southeast LA are 23.3 , 17.1, and 15.9 square miles, respectively. These are drastically bigger than the average area of all LA neighborhoods which sits at a compact 3.2 square miles.

We conclude that these three neighborhoods are outliers in the data. This necessitates finding a more robust distance metric to achieve a useful clustering analysis.

We consider alternatives and settle on using Spearman's correlation coefficient. Essentially, this metric measures the degree of correlation between variables. This is done using the rank ordering of the variables, not their actual values. This makes it much more robust against outliers/extreme values, which we have seen in the three neighborhoods analyzed above.

Let's visualize the distances between neighborhoods using Spearman's coefficient as a distance metric.

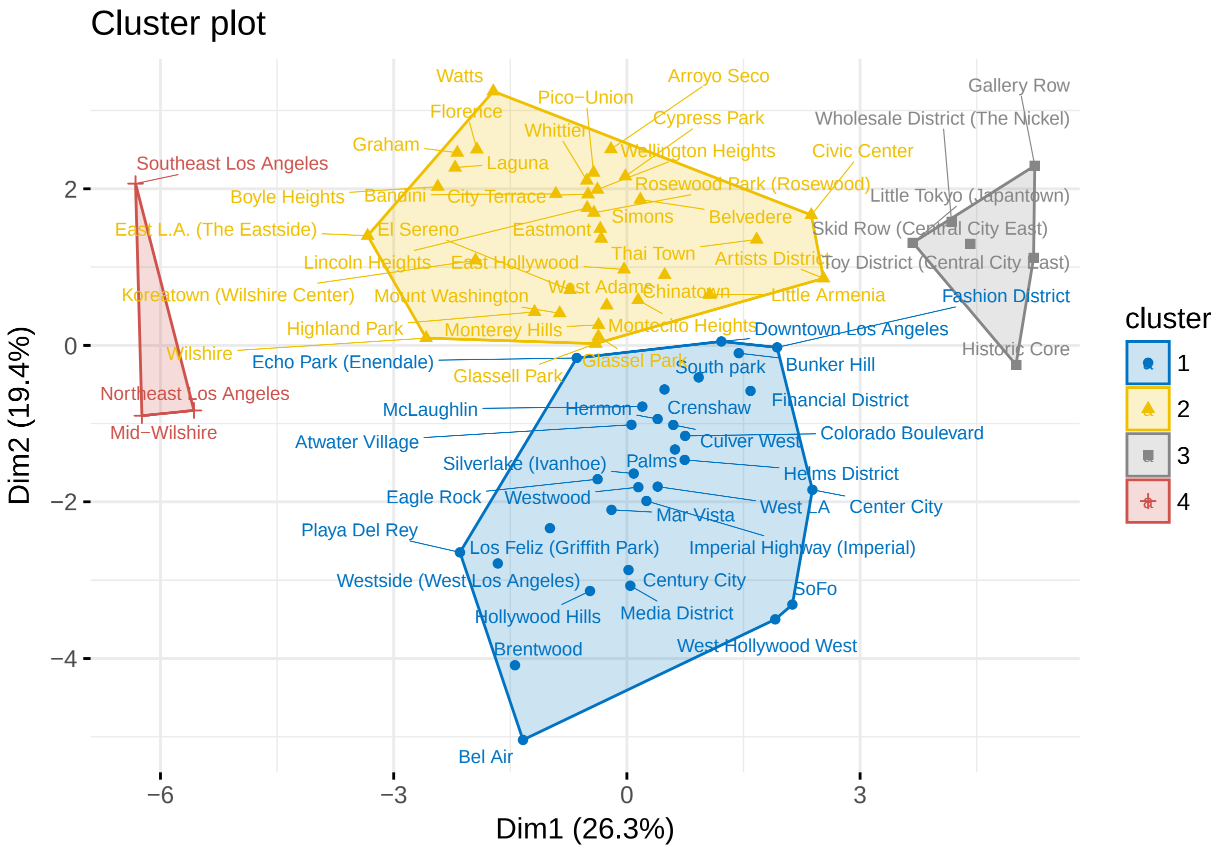
```
dists_spearman <- get_dist(imputedData, stand = TRUE, method = "spearman")
fviz_dist(dists_spearman,
   gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"), lab_size=5)
```



We notice the cross pattern formed by those three neighborhoods has largely disappeared. Instead, we see a lot of other structure emerge from the data. This structure (the patches of red and blue) are likely to lead to robust and interesting clusters.

So let's finally apply a clustering algorithm! We use the method partition around medoids, as it is more robust than k-means classification and is a routinely used clustering algorithm that usually produces good results. It's a good starting point.

```
cluster_results <- pam(imputedData, 4, metric='spearman')
fviz_cluster(cluster_results, data = imputedData,
             ellipse.type = "convex",
             palette = "jco",
             ggtheme = theme_minimal(),
             repel=TRUE,
             labelsize=7)
```



To visualize the results (which are high dimensional), we plot against the first two principal components of the data.

In cluster 4, we notice the three neighborhoods we discussed earlier as outliers. Makes sense.

In cluster 1 (in blue, at the bottom of the plot), we see neighborhoods like Brentwood, West Hollywood, Los Feliz, and the Westside. The neighborhoods in this cluster are all regarded as ‘nice neighborhoods’ and are desirable places to live. They are largely affluent, centrally located in the urban core, and are predominantly white.

In cluster 2 (in yellow, at the top of the plot), we note neighborhoods like Boyle Heights, East Hollywood, Watts, and Koreatown. These neighborhoods are peripheral to the urban core, less affluent, and composed largely of people of color.

We notice some interesting findings in the borderlands between cluster 1 and 2. Here, we notice Echo Park, DTLA, and Glassell Park. They form the edges of the two clusters, and in my exploration they’ve switched back and forth (especially when using kmeans, which is sensitive to the initial random placement of clusters). These neighborhoods occupy a liminal space in the analysis; they similarly occupy a liminal space in the mind of the Angeleno. Namely, they are some of the most prominently and rapidly gentrified neighborhoods in LA. We see this in how they’ve moved from cluster 2 to cluster 1 and now inhabit the area between them; in fact, they define the border between them.

In cluster 3 (gray, to the right of the plot), we note the neighborhoods that compose Skid Row and its environs. These areas are marked by large homeless populations and extreme poverty.

We’ve mapped the mindscape of how Angelenos conceive of their neighborhoods and their city. These groupings are relatively intuitive, and uncover relationships that are latent but powerful in influencing how Angelenos interact with their city. It’s not a stretch to say that current residents of neighborhoods in cluster 1 would be open to the idea of moving to other neighborhoods within that cluster; in fact, they probably work, shop, or play in those neighborhoods already. Similarly, it’s conceivable that residents in other clusters move mostly within that same cluster and have their social network built locally within those clusters.

With that, we also uncover the systemic inequality, segregation, and racism that lies beneath the clustering of these neighborhoods. We turn to the principal component analysis to understand this more clearly. Principal components are combinations of other variables that explain the the variation seen between the neighborhoods. So the second principal component (on the y axis of the cluster plot) explains the second largest proportion of variance between neighborhoods: it's very good for telling them apart. The biggest drivers for the second component are pct_poverty (with a coefficient of .30), median_rent (-.40), pct_non_english_speakers (.39), and median_income (-.44). We interpret the positive/negative signs to mean that neighborhoods that score low for this component (and are plotted low on the y-axis in the chart) have high incomes, high rents, few non-native English speakers, and few people living in poverty. First, it's telling that the second biggest divider among neighborhoods is wealth (evidenced by median_rent, pct_poverty, and median_income). This speaks to the high level of segregation due to social status and earning power within Los Angeles and American cities as a group. Second, it's troubling that the percentage of non-native English speakers is included in this component. This speaks to the high level of correlation between income and being a native English speaker. It's important to note that race and ethnicity were not included in the data, so the variable of being a non-native English speaker is likely serving as a proxy for ethnicity in this analysis. We observe that ethnicity and income are so incredibly inter-related, and that they form one of the most informative bases for clustering neighborhoods. Through this analysis, we have also uncovered an insidious strain of racism and inequality at work in the geography and mindscape of Los Angeles.

Overall, our analysis has confirmed things we've intuitively known about LA. It's interesting that these heuristics are supported by data, and are so easily produced solely by the data. The analysis has given us numerical, data-driven insight into such enduring urban problems like gentrification, segregation, and income inequality.