# City of Los Angeles Housing Recommendation System

### Summary

Through this project, we wanted to address the growing need for individuals relocating to Los Angeles without the necessary information and recommendation system that allowed them to properly locate a neighborhood that fit their preferences. Our team used K-means clustering to identify characteristics of different areas in LA, and presented a ranking application that allowed users to identify ideal neighborhoods for relocation developed from user input.

### Data Collection and Manipulation

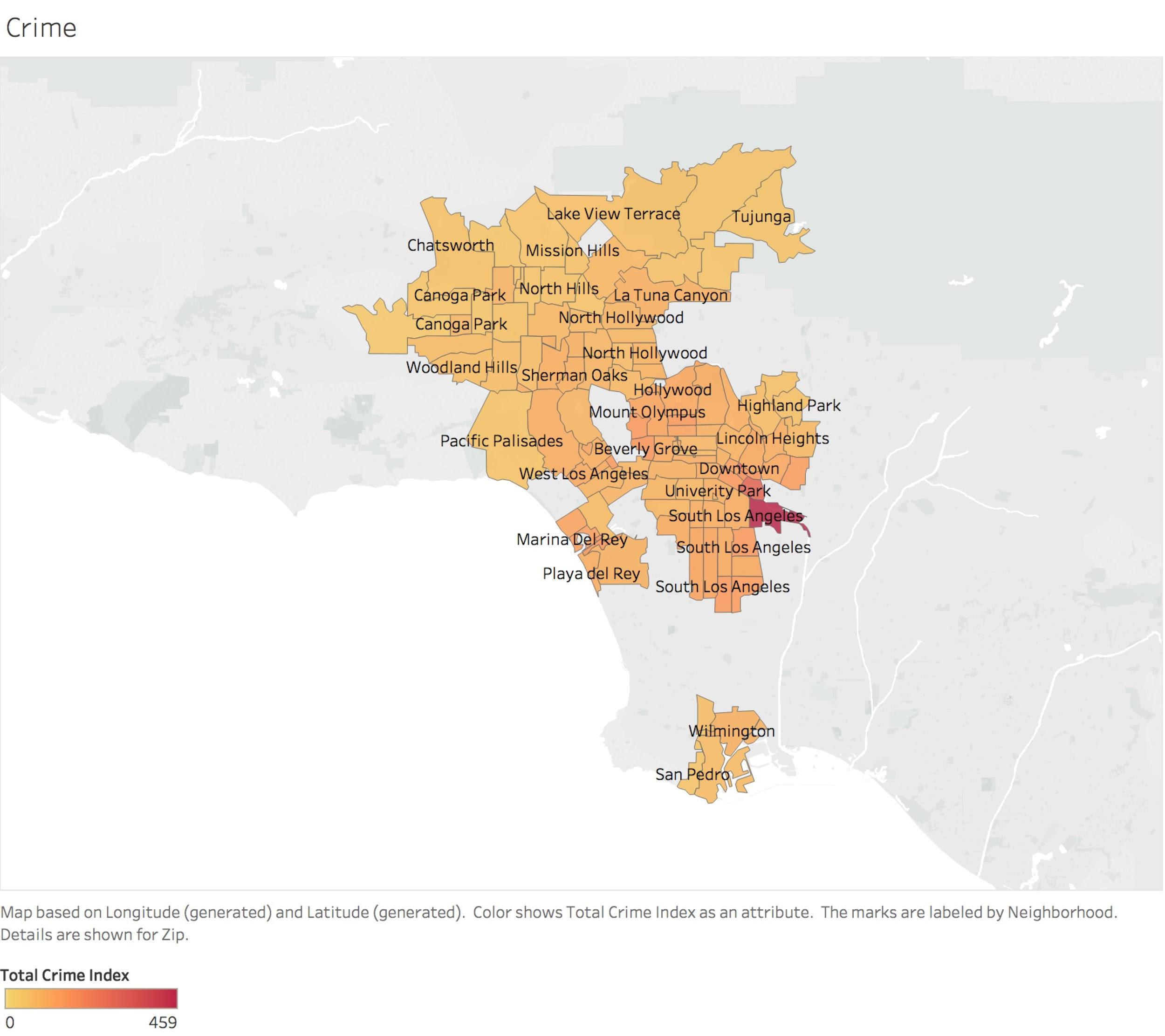
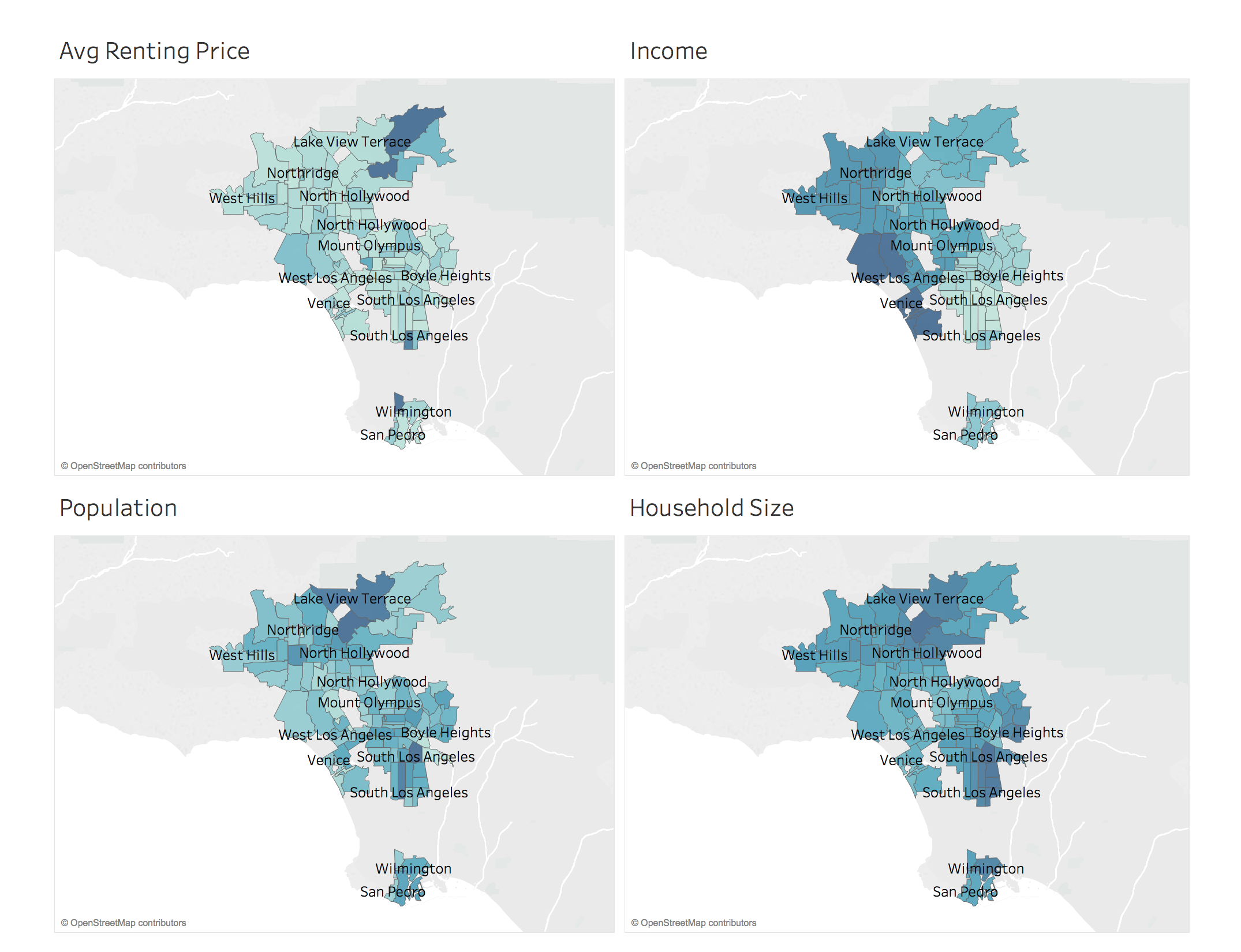
We identified three dimensions that are most important to our project, which are safety, demographics, and vitality. We collected most of the data from City of Los Angeles Open Data, crime index from our alumina working in City of Los Angeles, average rental price from Zillow, and walkscore from Walkscore.com. Following table shows details of our datasets.

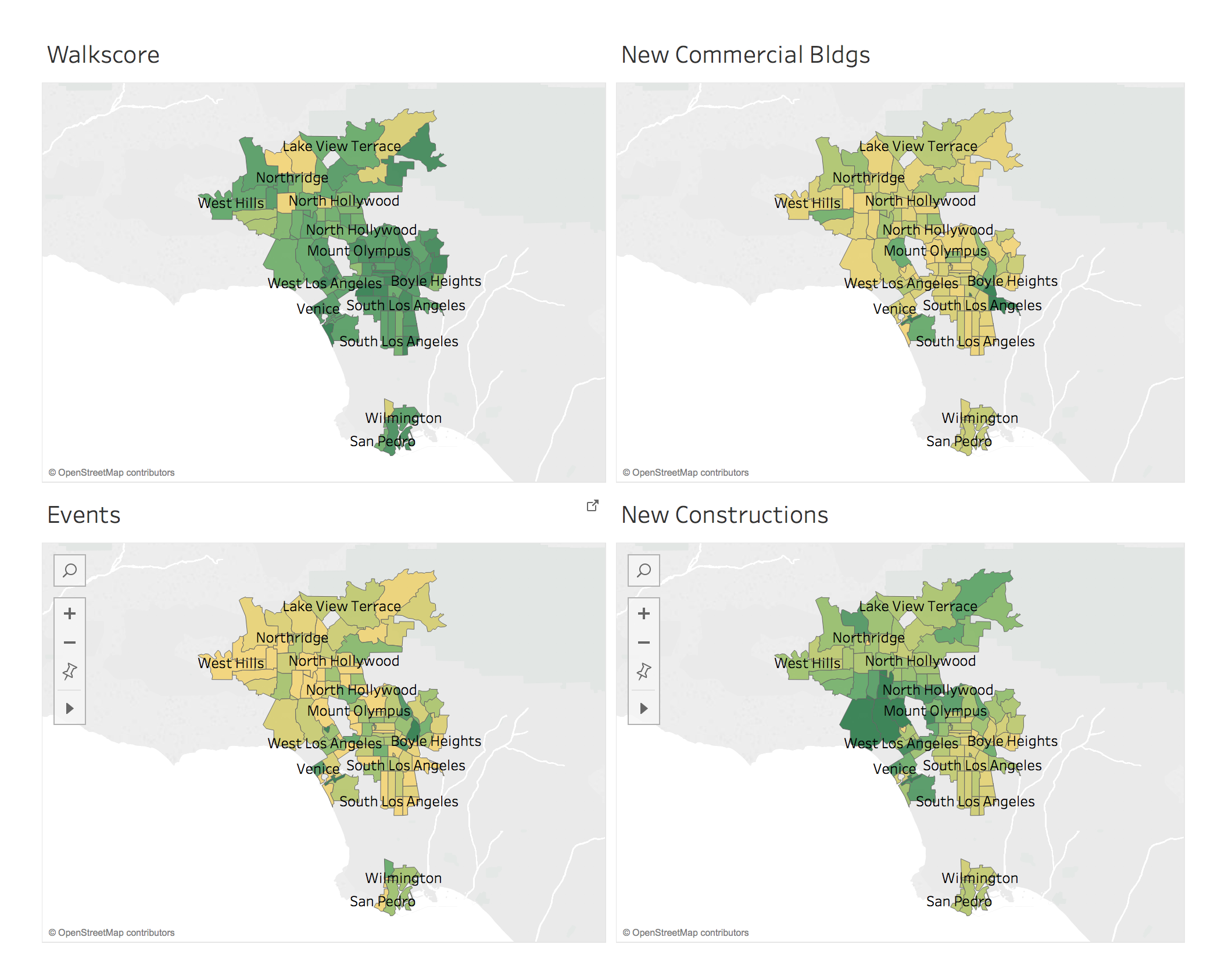
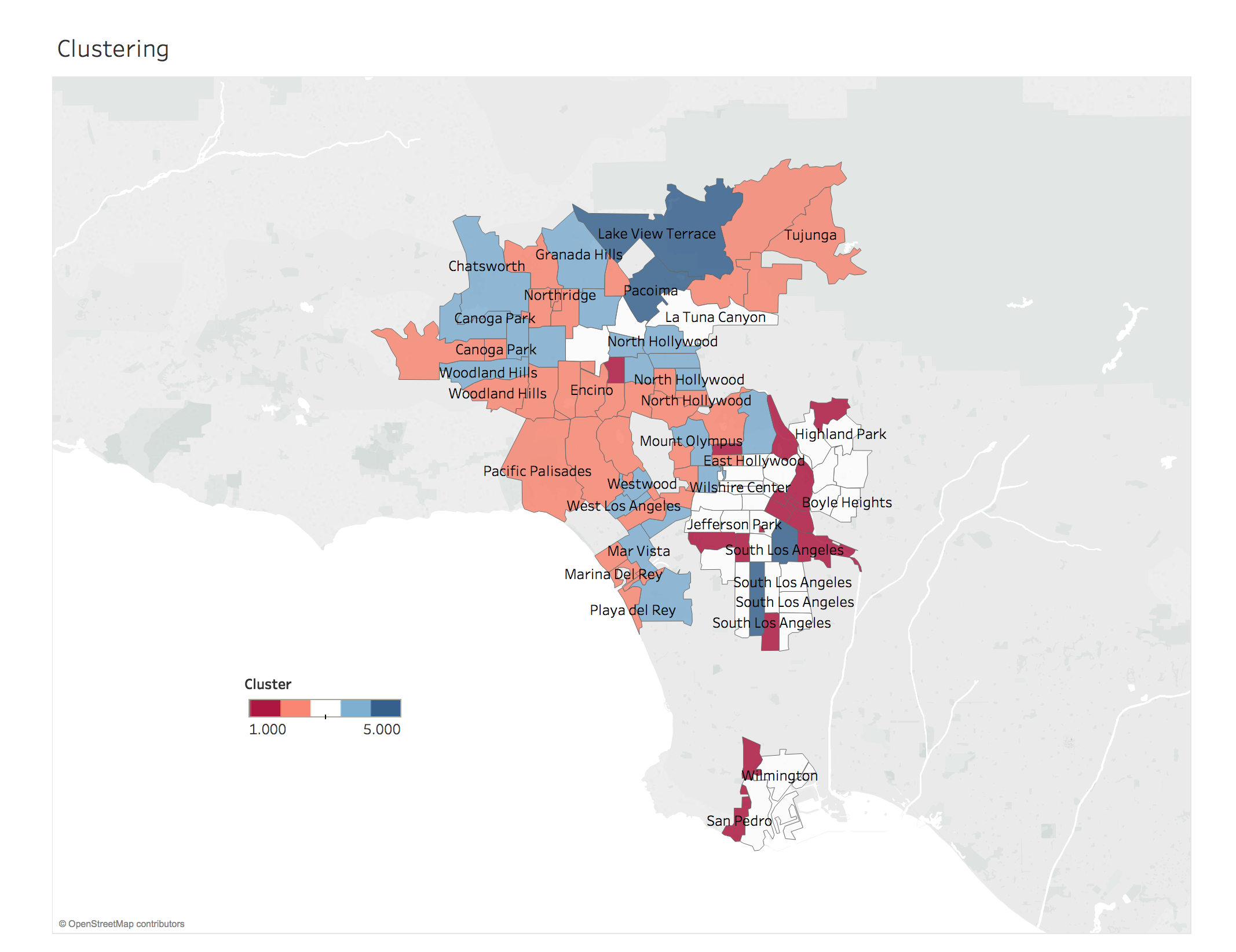
|  |  |  |
| --- | --- | --- |
| **Safety** |  |  |
| Total Crime Index | 2016 | LA City Crime Index |
| **Demographics** |  |  |
| Avg. Household Size | 2010 | 2010\_Census\_Populations\_by\_Zip\_Code |
| Total Population | 2010 | 2010\_Census\_Populations\_by\_Zip\_Code |
| Median Age | 2010 | 2010\_Census\_Populations\_by\_Zip\_Code |
| Race | 2014 | LA\_City\_General\_Population\_by\_Race\_-\_2014 |
| Occupation | 2013 | Employment\_by\_Occupation\_by\_Gender\_for\_Council\_Districts |
| Income | 2013 | Median\_Household\_Income\_by\_Council\_Districts |
| Average Rental Price | 2017 | Zillow |
| **Vitality** |  |  |
| New Commercial Bldgs | 2013-2017 | New\_Building\_Commercial |
| New Constructions | 2015-2017 | New\_Construction\_Permits\_2015 |
| Events | 2015-2017 | What\_s\_Happening\_LA\_Calendar\_Dataset |
| Walkscore | 2017 | Walkscore.com |

We joined datasets by zip code and council district. After cleaning and manipulating data, we had a final dataset to use for analysis which consisted of 105 unique observations and 24 variables including: zip code, neighborhood, average rental price, council district, total population, median age, average household size, race % with 5 categories (1. Asian, 2. Black, 3. Hispanic, 4. Other, 5. White), median income, occupation % with 5 categories (1. management, business, art, sciences, 2. Natural resources, construction, maintenance, 3. Production, transportation, materials moving, 4. Sales and office, 5. Service), total crime index, # new constructions per capita, # new commercial buildings / capita, number of events / capita, and walkscore.

### Data Visualization

Through data visualization, we can get the basic idea of Los Angeles. For example, South Los Angeles has bigger household size, larger population, lower income, and lower average renting price, while it’s also the place with the highest crime rate and low vitality. Following pictures show the distributions of some variables.



### K-Means Clustering

We used K-means clustering to illustrate broader groups of this information throughout Los Angeles. We used k = 5 to cluster types of people into neighborhoods based off our dataset, and we identified certain groups of people that would live in specific neighborhoods. Group 2 for instance, had a very high median income, high management, business, art and science occupation, minimal construction, low crime, very few events, relatively low population, higher median age, average household size of 2.3, predominately white, and second highest rental prices. Concluding that this group of people were likely to be younger couples or empty nesters who lived in suburban areas that were safer, less walkable, and likely more educated. Some of the neighborhoods identified were: Brentwood, Canoga Park, Cheviot Hills, Encino, Marina Del Rey. All of which fit the above description. The K-means clustering gave us an ability to take our recommendation system to the next level by identifying a group of people that may be associated with the user and illustrate neighborhoods that match this group. Following charts list neighborhood in each group and the center of each group.

|  |  |
| --- | --- |
| **Clustering** | **Neighborhoods** |
| **1 (19 neighborhoods)** | Atwater Village, Baldwin Hills, Downtown, Eagle Rock, Harbor City, Hollywood, San Pedro, South Los Angeles, USC, Van Nuys, Wilshire Center |
|
| **2 (33)** | Beverly Glen, Beverly Grove, Brentwood, Canoga Park, Century City, Cheviot Hills, Encino, Hollywood, Marina Del Rey, Mission Hills, North Hollywood, Northridge, Pacific Palisades, Playa del Rey, Playa Vista, Shadow Hills, Sherman Oaks, Studio City, Tarzana, Tujunga, Valley Village, Venice, West Fairfax, West Hills, Westwood, Woodland Hills |
|
| **3 (29)** | Beverly Glen, Beverly Grove, Canoga Park, Century City, Cheviot Hills, Encino, Hollywood, Marina Del Rey, Mission Hills, North Hollywood, Northridge, Pacific Palisades, Playa del Rey, Playa Vista, Shadow Hills, Sherman Oaks, Studio City, Tarzana, Tujunga, Valley Village, Venice, West Fairfax, West Hills, Westwood, Woodland Hills |
|
| **4 (20)** | Boyle Heights, Cypress Park, East Hollywood, Echo Park, Hancock Park, Hyde Park, Jefferson Park, Koreatown, La Tuna Canyon, Lake Balboa, Lincoln Heights, Mid-City, Monterey Hills, Panorama City, Pico Heights, San Pedro, South Los Angeles, University Park, West Adams, Westlake, Wilmington |
|
| **5 (4)** | South Los Angeles, Pacoima, Lake View Torrance |
|

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Total Population | Median Age | Asian | Black/ African American | Hispanic/ Latino | Other | White |
| 1 | 18669.1 | 36.1 | 10.4 | 10.9 | 64.6 | 2.4 | 11.8 |
| 2 | 21953.2 | 40.5 | 13.0 | 4.5 | 26.6 | 3.6 | 52.2 |
| 3 | 51839.0 | 31.9 | 11.4 | 12.0 | 64.4 | 2.1 | 10.2 |
| 4 | 47977.0 | 35.3 | 12.3 | 4.7 | 29.1 | 3.5 | 50.4 |
| 5 | 97271.3 | 29.1 | 5.0 | 14.5 | 68.8 | 1.6 | 10.1 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Average Household Size | Median Income | Management Business Science & Arts | Sales & Office | Service | Natural Resources Construction & Maintenance | Production Transportation & Materials Moving |
| 1 | 2.3 | 37885.5 | 24.8 | 23.6 | 24.4 | 10.2 | 17.0 |
| 2 | 2.3 | 65575.1 | 50.4 | 23.2 | 14.7 | 5.2 | 6.4 |
| 3 | 3.3 | 37400.0 | 22.1 | 23.4 | 26.5 | 10.3 | 17.8 |
| 4 | 2.6 | 63541.1 | 47.3 | 23.4 | 16.2 | 6.0 | 7.2 |
| 5 | 4.2 | 40257.3 | 18.3 | 23.6 | 26.1 | 11.9 | 20.1 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Crime Index | New Constructions | New Commercial Bldg | Events | Walkscore | Avg Rental Price |
| 1 | 141.8 | 46.8 | 4.9 | 2724.5 | 67.4 | 3536.1 |
| 2 | 90.0 | 13.2 | 0.4 | 33.5 | 57.3 | 3494.6 |
| 3 | 95.6 | 3.4 | 0.2 | 3.8 | 74.8 | 2803.2 |
| 4 | 79.3 | 6.9 | 0.3 | 3.0 | 52.2 | 2793.5 |
| 5 | 90.8 | 3.6 | 0.2 | 2.5 | 71.3 | 2894.0 |

### Algorithm and App

To take it a step further, we had the idea of starting our own company throughout this process and felt that a ranking system was the start of an overall neighborhood guide. The goal was to construct a ranking system that allowed users to input their preferred weights for the designated grouped variables and answer a few questions related to demographics and give them a recommendation of what neighborhood best met these criteria based off our data.

By giving each category a value, which independently ranked them from highest to lowest relative to that category, we had a specified rank of each variable respective to the zip code and neighborhood. Then we weighted each variable to get a total of 1 for that group. The demographics grouped variables were weighted at race (0.2), income (0.15), rent (0.4), occupation (0.05), age (0.1), and household size (0.1). This allowed us to put more weight on the demographic grouped variable based on affordability and ethnicity which we felt was more important when individuals were looking for neighborhoods. For vitality grouped variables, we weighted new commercial buildings (0.2), new construction (0.2), events (0.4), walkscore (0.2). We feed this information into shiny R and develop a “survey” that would let users input their own information and calculate a ranking of each neighborhood by zip code.

The “survey” allowed users to input the desired weights (slider input scaled 0 - 100) of each of the three grouped variables to accurately reflect their personalized ranking of most important factors for neighborhoods. After that, the user is asked to answer 5. “What is your age?” (slider input scale 0 - 70), “What is your annual household income?” (slider input scale $0 - $200,000), “How many people live in your household?” (slider input scale 0 - 10), “Which best describes your ethnicity?” (drop down of race), “Which best describes your occupation?” (drop down of occupation). We then used these inputs in our algorithm to construct an overall ranking.

The algorithm constructed for our ranking system took the rank of each grouped variable and multiplied it by the input weight the user distinguished from the “survey”. Once these were calculated, we re-ranked them in decreasing order from 1 - 105 (1 being the highest).

In conclusion, the ranking system gives a whole new perspective into house hunting. We felt there were many resources that allowed users to research specific neighborhoods, homes, and demographic information, but none of them allows for a personalized search. We can take user input combined with all open source data for a personalized ranking of neighborhoods creating a targeted search without the groundwork. We can build this into a business that can combine numerous open source datasets and allow for user interaction to identify a more personalized house hunting approach to the real estate market. The target demographic in mind would be people relocating to LA for the first time and need a more robust approach. We believe this can provide this experience and help people identify what neighborhood they can call home.