Identifying Quiet Cities in Los Angeles County, California

Samuel Chodur

June 2, 2020

1. Introduction

1.1 Background

Los Angeles is the largest city in the state of California. It has an estimated population of nearly four million people making it tough to find ‘quiet’ places that lack cultural noise.

1.2 Problem

There are several factors that may contribute to whether a city can be considered relatively quiet compared to those around it. Some of these factors are the population density, the venue density and the surface area of the city. This project attempts to identify clusters of quiet cities in Los Angeles county based on the aforementioned data.

1.3 Target Audience

The target audience for this project can be potential business owners. Potential business owners would most likely not want to target quiet areas of the county if they are in need of foot traffic to sustain their business. On the other side, homeowners may want to live in areas that could be considered quiet so they can have “peace and quiet” while in their home.

2. Data acquisition and cleaning

2.1 Data sources

Data will be collected from the following sources:

- Los Angeles county city data that contains **city name**, and **population**.
    - *Data Source:* Wikipedia
- Los Angeles county city data that contains **land area** in square miles.
    - *Data Source:* Census.gov
- Coordinate data for the cities in Los Angeles county which includes **latitude** and **longitude**.
    - *Data Source:* The Geocoder Python library
- The number of retail establishments will be fetched using the Foursquare API.
    - *Data Source:* Foursquare API

2.2 Data cleaning

Some of the data obtained from the sources listed in the previous section needed to be cleaned. The procedure of obtaining and cleaning the data was as follows in the subsequent paragraphs.

The city data from Wikipedia included information about when the city was incorporated which was of no use to this study and was subsequently dropped from the data set. After obtaining the latitude and longitude values by using the Geocoder Python library, the data was merged with the city and population data.

Surface area data was obtained from the Census.gov website which makes the data available in a standard text file delimited by tabs. After this data was read into a Pandas DataFrame, the columns that would be of use to this study were each of the city's land surface area. We dropped the rest of the data and merged what was left with the original DataFrame.

Finally, Los Angeles county venue data needed to be obtained via the Foursquare API. Venue data was obtained by using a function which found nearby venues within a certain radius of a latitude and longitude combination. After all venues that met the criteria were accumulated, they were grouped by city and counted. The resulting DataFrame was then merged with the DataFrame holding all of the previously acquired data.

3. Methodology

3.1 Approach

The approach to resolve the issue of identifying the most quiet cities in Los Angeles county was as follows:

- Collect Los Angeles county city data from wikipedia.
- Collect Los Angeles county city surface area data from Census.gov
- Utilize the Geocoder Python library to collect city coordinate data.
- Use the Foursquare API to get the approximate number of retail establishments in each City.
- Exploratory data analysis
- Preprocess the data
- Analyzing by using Clustering (K-Means).
- From results, infer which areas would be the quietest and draw conclusions.

3.2 Exploratory data analysis

Once the data was obtained and merged together, we identified that a city's population and its venue count would be used to obtain city clusters. The describe method of a Pandas DataFrame was used to generate the descriptive statistics of our data and can be seen in Table 1.

	Population	Latitude	Longitude	Land_Area_sqmi	ALAND	Venue
count	7.700000e+01	77.000000	77.000000	77.000000	7.700000e+01	77.000000
mean	1.098218e+05	34.012020	-118.197845	17.887039	4.632716e+07	26.727273
std	4.302547e+05	0.183215	0.224700	55.067444	1.426242e+08	19.115051
min	1.120000e+02	33.344110	-118.818750	0.948000	2.455587e+06	1.000000
25%	2.025600e+04	33.895690	-118.353740	3.577000	9.263209e+06	12.000000
50%	4.636100e+04	34.011580	-118.154940	7.236000	1.874171e+07	23.000000
75%	8.429300e+04	34.112980	-118.063700	13.472000	3.489115e+07	35.000000
max	3.792621e+06	34.698900	-117.750030	468.956000	1.214591e+09	95.000000

Table 1: Descriptive statistics of the acquired and merged data.

3.2 Preprocessing of the data

From the descriptive statistics, we can see that the columns corresponding to a city's population and that of venue count are of different units. They also have significantly different means and standard deviations when compared to each other.

To make this data more useful for analysis, we calculated population and venue densities per square mile for each city. After these densities were available, the values were normalized using MinMaxScaler from sklearn. This step ensured that the densities were equally used when calculating clusters and not letting one have more influence than the other.

4. Results

4.1 Clustering the cities

After the previous preprocessing steps were completed, the calculation of clusters was ready to begin. We used a 'kcluster' value of 5 as that is what was used during the lesson on clustering. We could have developed a method for identifying the best value for the number of clusters but due to time constraints we did not address that during this study. The resultant clusters are shown in Figure 1.

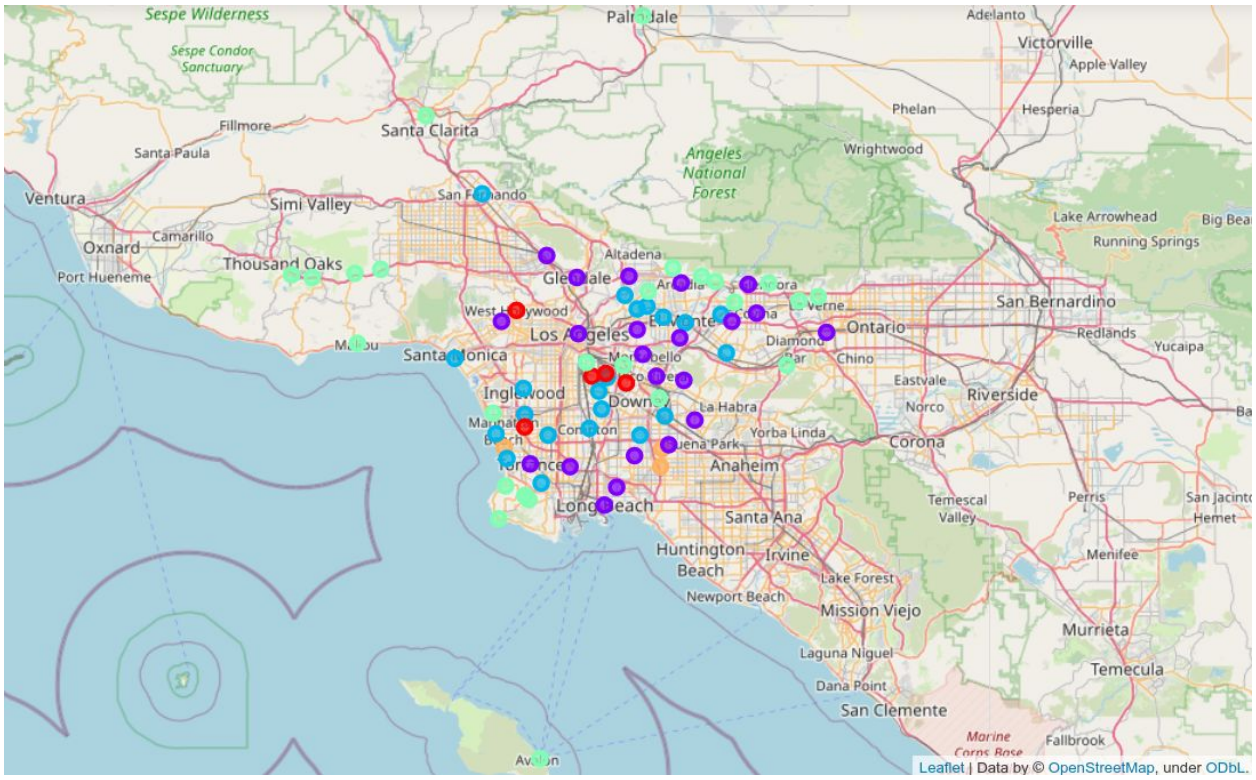

Figure 1: The resultant Los Angeles County city clusters

4.2 Identifying the quiet and the loud clusters

We plotted the means of both the population density and the venue density to get an idea of why different cities were clustered together. From the bar graph in Figure 2, we can get a better idea of why cities were clustered together. Based on the mean values of population and venue densities, we could say that cities in Cluster 0, 2 and 4 (Figure 3) could be considered relatively non-quiet areas while cities in Cluster 1 and 3 (Figure 4) may be considered quiet.

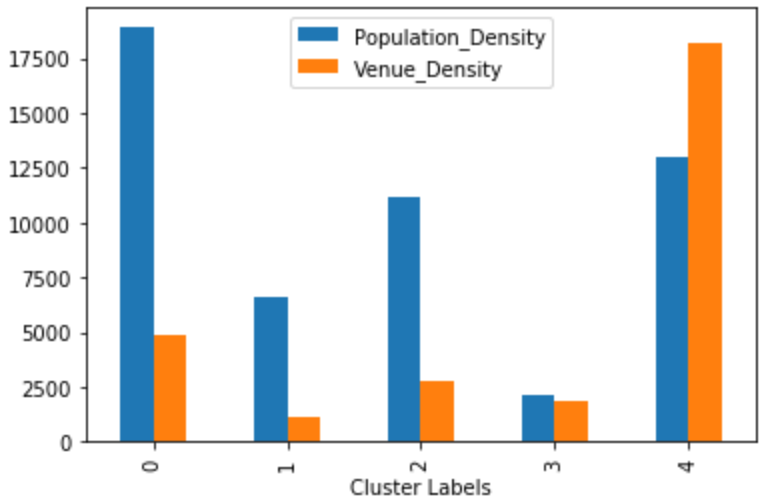

Figure 2: The mean population and venue density values for each cluster. Venue density was scaled by 500.

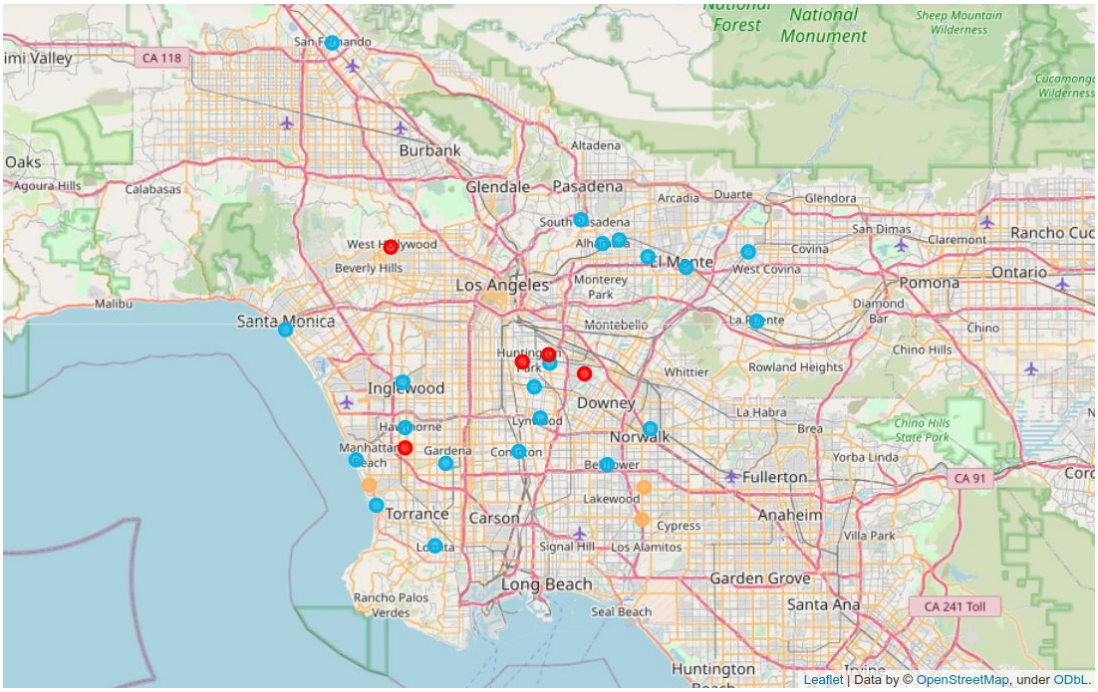

Figure 3: The relatively non-quiet cities of Los Angeles County.

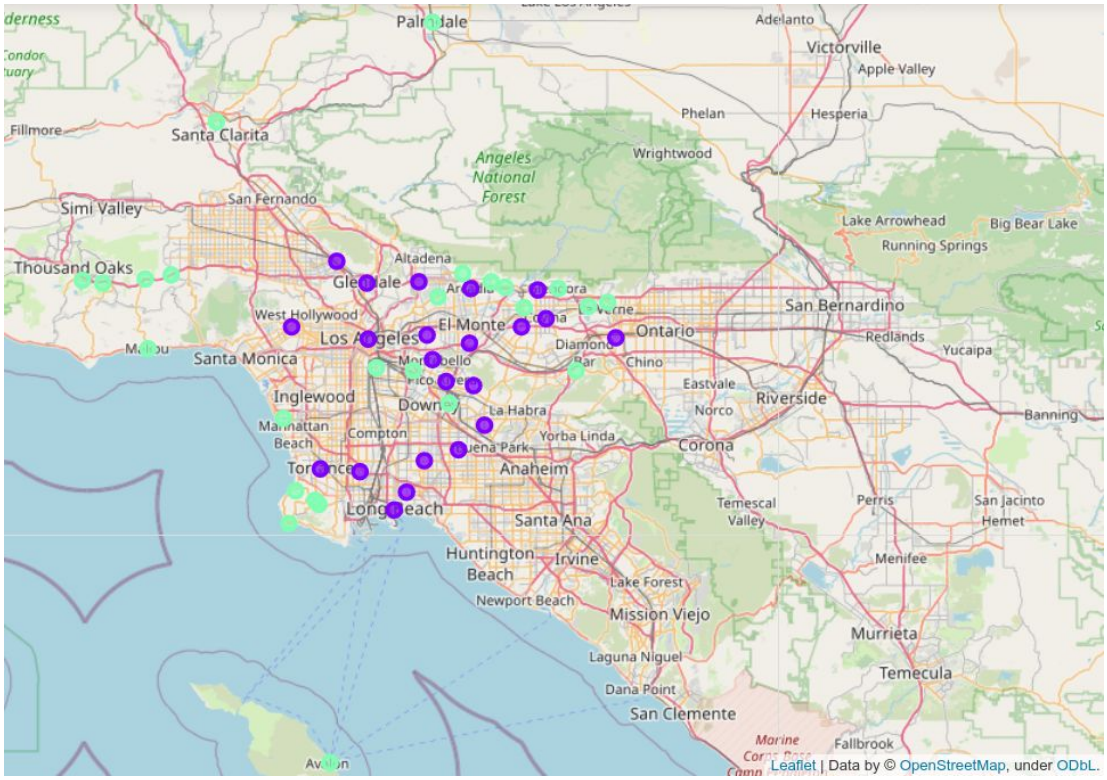

Figure 4: The relatively quiet cities of Los Angeles County.

4. Discussion

4.1 Observations and recommendations

When looking more closely at the resulting clusters, some may not intuitively make sense. One example of this is that the city of Los Angeles ends up in Cluster 1. Los Angeles should not be considered a relatively quiet city, so we can see something is incorrect about our methodology. We believe that the function that acquires near-by venues could be updated in some way so that the radius is based on the size of the city. In essence, we are misrepresenting how many venues are in each of the cities. Since cities vary in size and area, radius is not an adequate way to obtain information about venue counts in each city.

5. Conclusion

In this study we used population counts, venue counts and land areas of cities in Los Angeles County, CA in an attempt to identify areas of low cultural noise. We obtained this data through python libraries to scrape city information and the use of Foursquare's API to tie venue counts to each of the cities. Since we have no available data to target, we made the assumption that cities with low population density and low venue density were to be considered quiet. Under this assumption, we were able to identify five different clusters of cities in Los Angeles County and identified which of these would be best described as a quiet area.