Selecting the Best Neighborhood to Move in When Relocating into Los Angeles County

Sina Bozorgmehr

June 2020

1. INTRODUCTION

Relocating is always a complicated and hard to do. In this era, finding a better paid job and moving to a new place is constantly happening amongst population especially in younger generation. People are looking for more stable jobs and depending on the industry they are in, there are hot spots that they can find more jobs with better salaries that can persuade them to relocate to a new city or even a new state. Finding the right job is the first step. When people are going to relocate, the first thing they are obsessed with is the neighborhood they are going to live. To choose the right place to live, they consider the housing rate, school ranks if they have children, crime rate, amenities, grocery stores, etc.

Another important factor to choose the right place is the commute distance to work. Specially in more populated cities with more traffic, people tend to live as close as possible to their workplace. So, if you look at it, there are multiple factors involved in the decision-making process which sometimes makes it tough. This problem gets even bigger when we are talking about metropolitan areas like Los Angeles where commuting from north west to the south west pf the city can take up to 3 hours in traffic jam.

It will be much easier if we could cluster the neighborhoods based on the involved factors in the problem and narrow down the choice range. In this Project, it has been tried to do this by:

1. Clustering the Los Angeles county neighborhoods based on their distance to a specific workplace and rent rate and then

2. Based on the user’s input, score each neighborhood in the cluster of choice, and select the one with highest score as the new neighborhood to move in.

With this solution, one can make an easier choice when relocating to Los Angeles county for a new job.

2. Data Collection and Cleaning

2.1 Data Source

In this project, the source dataset was provided by American Community Survey and hosted by University of Southern California. The data set includes median rent price from 2010 to 2016. Median rent price measures gross rent, which includes rent price plus the cost of utilities like electricity, water, gas, and sewage. Many factors can affect the rent price of a housing unit, including its age, condition, size, and design. Additionally, neighborhood characteristics like proximity to amenities, school district performance, and crime rates and safety have a sizable impact on rent prices. The general strength of the economy and housing markets can also cause rent prices to rise or fall. Because the rate of homeownership has declined since the Great Recession, an increased demand for rental units has caused rent prices to rise. Median Gross Rent, MGR, is a measure of the average level of housing affordability in an area. It is a useful tool for comparing affordability on both a neighborhood and national scale.

2.2 Data Cleaning

To this project, only the MGR in 2016 is used to cluster the neighborhoods. Since there are multiple MGRs for each neighborhood in 2016, the maximum rate has been used for clustering. Each neighborhood has its location in a tuple which includes latitude and longitude. Therefore, it was needed to separate these values and put them in their own columns.

The workplace neighborhood is selected randomly from all the neighborhoods in the dataset. Then the miles distance of each neighborhood from workplace is calculated using Haversine formula. Normalization and K-Means clustering of the data is implemented by *scikit-learn*. The cluster number is set to 5 and only MGR and Distance are used as features for clustering.

After clustering all the neighborhoods in 5 distinct groups, the top two clusters with lowest distance to workplace are selected and then the one that has a lower MGR is marked as the cluster of choice for relocation. Then, using the foursquare API, all the venues in these neighborhoods are extracted and using pandas library a data frame is created that consists of top 10 most common venues in each neighborhood. In the next step, user can enter three venues of their choice that are most important to them.

Using a loop, all the neighborhoods are examined and assigned with a score based on the three venues that the user has selected. Finally, the neighborhood with the highest score is selected as a best neighborhood to move in and is shown on the map. If two or more neighborhoods gain the highest score, a list of those will be recommended to move in.

3. Methodology

3.1 Clustering

There are many models for clustering out there. In this study, on of the simplest models has been used to cluster the neighborhoods of Los Angeles county. Despite its simplicity, k-means is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data. Initialization method of the centroids is done by *k-means++*.
k-means++ selects initial cluster centers for k-means clustering in a smart way to speed up convergence. The number of clusters is set to 5 and *n_init* is set to 12 which define number of times the *k-mean* algorithm will be run with different centroid seeds. The results will be the best output of *n_init* consecutive runs in terms of inertia.

3.1.1 Preparation for clustering

The MGR and miles distance values are normalized before clustering. Normalization is implemented using *scikit-learn* package.

3.2 Haversine Formula

The haversine formula determines the great-circle distance between two points on a sphere given their longitudes and latitudes. Important in navigation, it is a special case of a more general formula in spherical trigonometry, the law of haversines, that relates the sides and angles of spherical triangles. Using *numpy* library, latitudes and longitudes values are converted to radians to be used in the Haversine formula:

$$d = 2rsin^{-1}\left(\sqrt{sin^2\left(\frac{\Phi_2-\Phi_1}{2}\right) + cos(\Phi_1)cos(\Phi_2)sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\ \right)$$

Where d is the shortest distance between two points on a sphere. r is the radius of earth which is 3,958.8 miles. Φ1 and Φ2 are the latitudes of the points and λ1 and λ2 are the longitudes.

3.3 Scoring each neighborhood

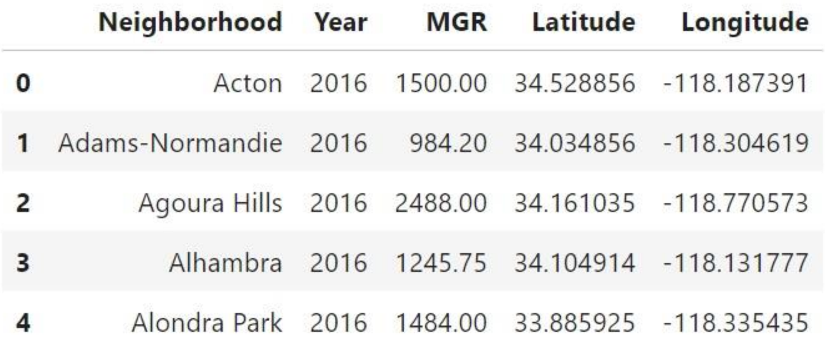
Using the foursquare API, the venues in 500 meters radius of each neighborhood is extracted and inserted in a data frame. All venues are categorized and then the frequency of each category in each neighborhood is merged into the fata frame. Based on these frequencies, a new data frame is created that show the 10 most common venue category in each neighborhood.

Based on three venue categories that uses selects, each neighborhood gets a final score. For example, if the three categories entered by user are: Park, Gym and Bank; then the application goes through all ten columns and if find any of these three categories in the first column, a 10 points will be added. If it finds the categories in the second most common venues, then 9 points will be added. Application goes till the last column and then all the points are added up to create the final score for the neighborhood. At the end of the loop, the neighborhood with highest score is shown as the select place to relocate. If more than one neighborhood gets the same high score, a list of those neighborhood is shown to relocate.

4. Results

4.1 Cleaned Data

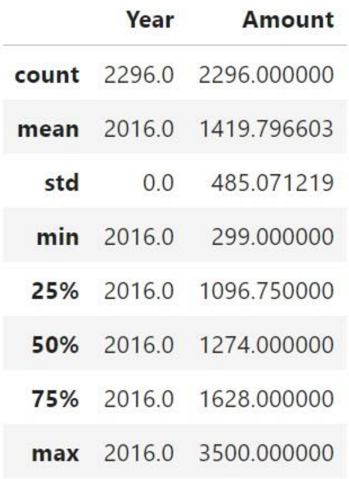
After getting the dataset from the source and cleaning it, the main dataset looks like the data frame in *Table 1*. This data frame consists of the neighborhood name, year in which the MGR has been collected, the amount or MGR, and the latitude and longitude of each neighborhood. This data frame has 257 rows or neighborhoods.



	Neighborhood	Year	MGR	Latitude	Longitude
0	Acton	2016	1500.00	34.528856	-118.187391
1	Adams-Normandie	2016	984.20	34.034856	-118.304619
2	Agoura Hills	2016	2488.00	34.161035	-118.770573
3	Alhambra	2016	1245.75	34.104914	-118.131777
4	Alondra Park	2016	1484.00	33.885925	-118.335435

Table 1. The main dataset after cleaning

Statistical analysis shows that the mean MGR in L.A. county in year 2016 was $1419.80 and that maximum MGR in all neighborhoods where $3500 while the minimum MGR was $299. For more statistical data, please see the *Table 2*.



	Year	Amount
count	2296.0	2296.000000
mean	2016.0	1419.796603
std	0.0	485.071219
min	2016.0	299.000000
25%	2016.0	1096.750000
50%	2016.0	1274.000000
75%	2016.0	1628.000000
max	2016.0	3500.000000

Table 2. Statistical analysis of main data frame

4.2 Data Distribution

The frequency distribution of MGR has been visualized by histograms in *Figure 1.* As you see, more than 80 neighborhoods out of all 257, have an MGR in range of $1200-1488 and only less than 40 neighborhoods have an MGR of more than $2062.

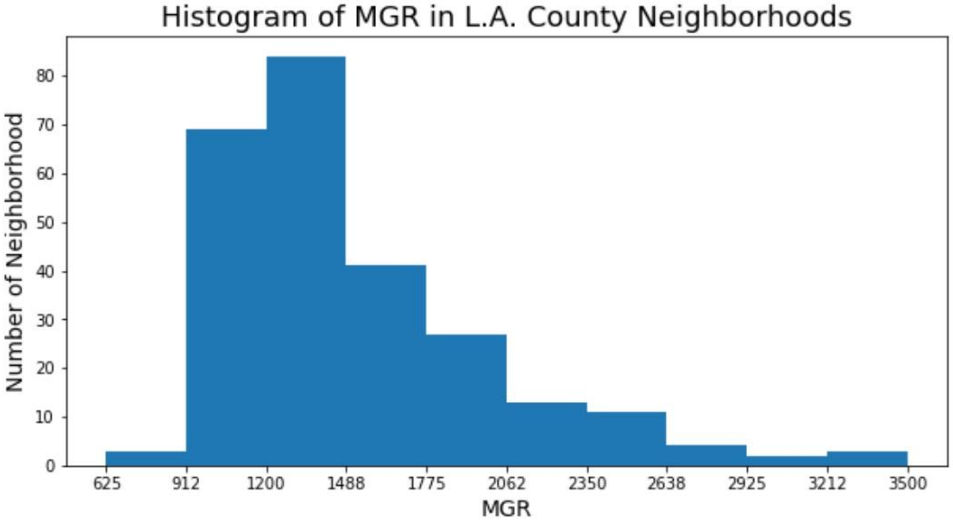


Figure 1. Frequency Distribution of MGR in all Neighborhoods of L.A. county

Figure 2 shows all the neighborhoods in L.A. county superimposed in the map. Using pandas library, 10 most expensive neighborhoods have been extracted from the main data frame and you can see them on the map on *Figure 3.* Also, *Figure 4* shows the bar chart for these neighborhoods.

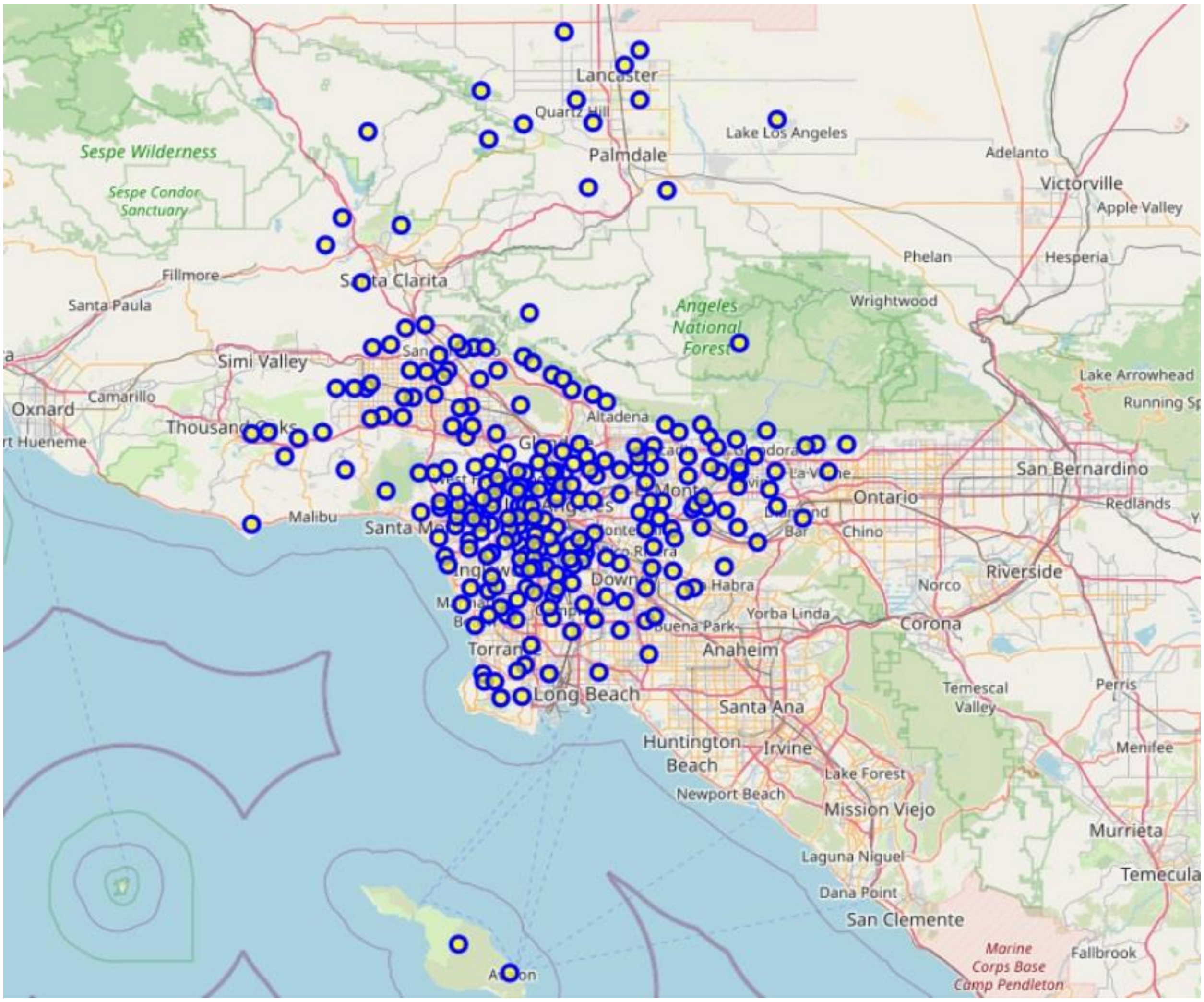


Figure 2. Los Angeles county neighborhoods superimposed on the map.

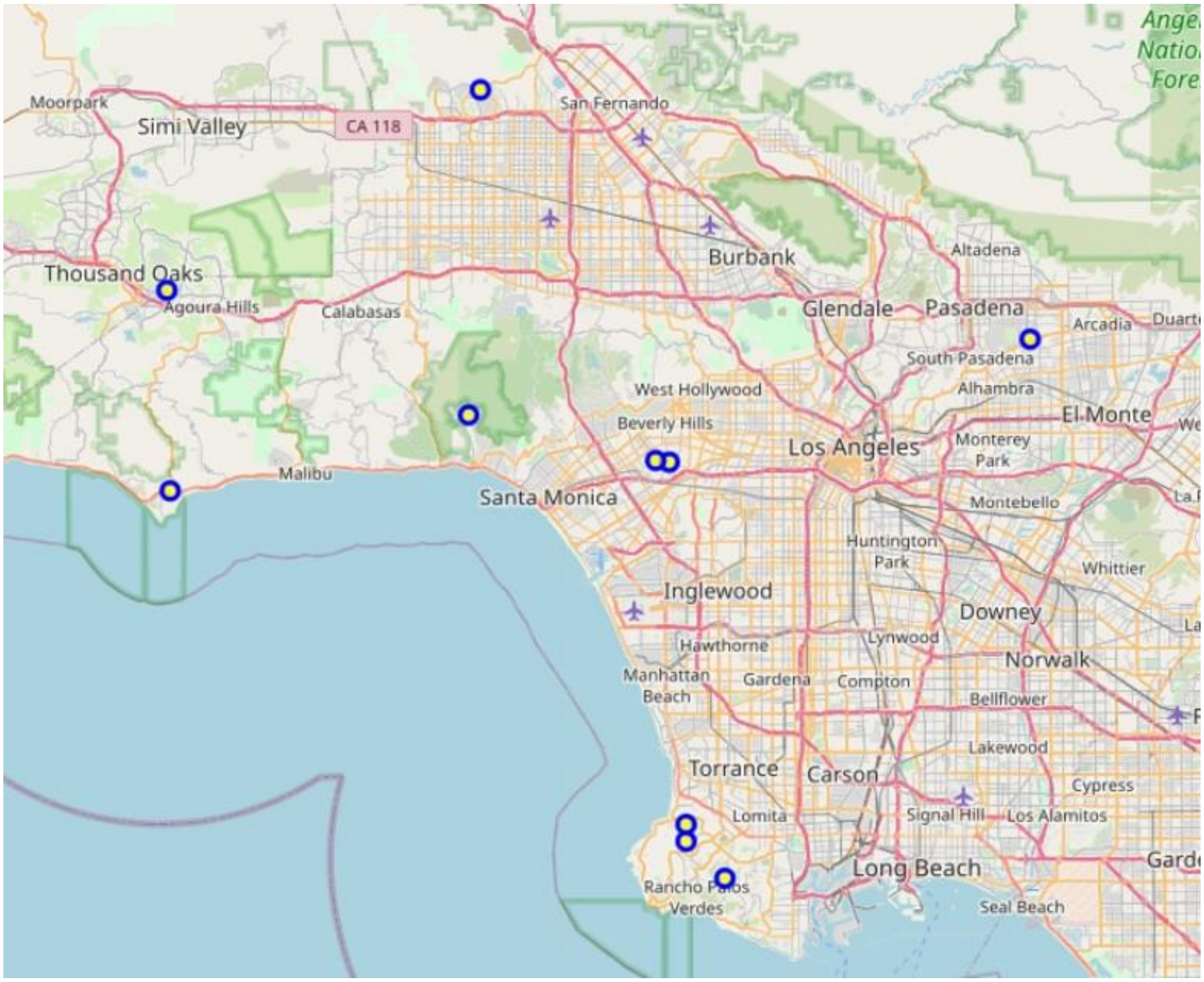


Figure 3. Top 10 most expensive neighborhoods

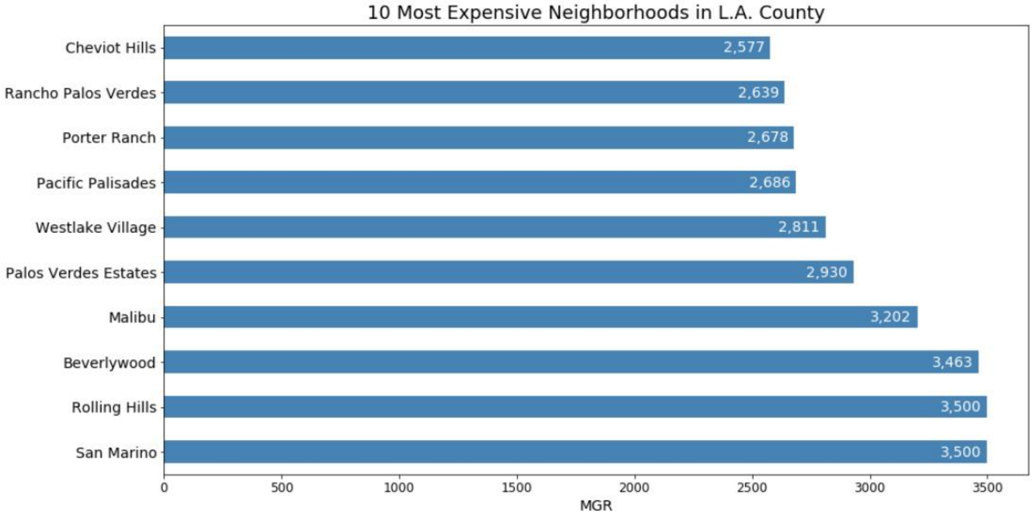
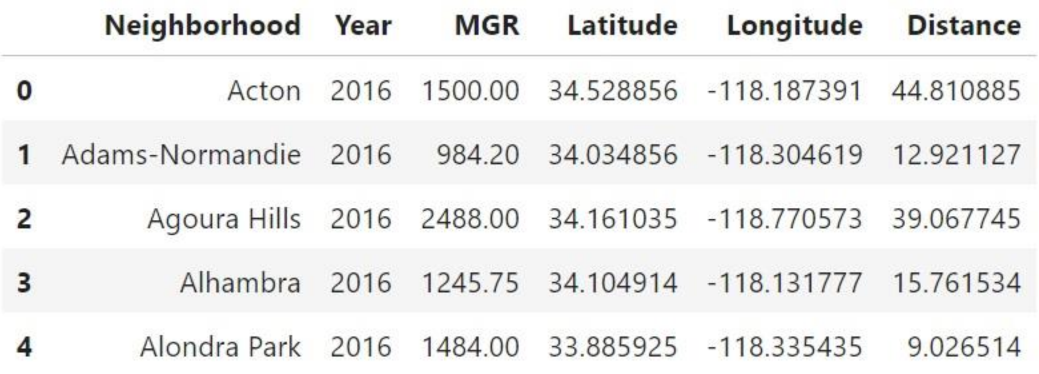


Figure 4. Bar chart of 10 most expensive neighborhoods in Los Angeles county.

4.3 Clustered Neighborhoods

Before starting to cluster the neighborhoods, their linear distance from workplace neighborhood, Long Beach, was calculated using Haversine formula and the data were added to the data frame (*Table 3*).



	Neighborhood	Year	MGR	Latitude	Longitude	Distance
0	Acton	2016	1500.00	34.528856	-118.187391	44.810885
1	Adams-Normandie	2016	984.20	34.034856	-118.304619	12.921127
2	Agoura Hills	2016	2488.00	34.161035	-118.770573	39.067745
3	Alhambra	2016	1245.75	34.104914	-118.131777	15.761534
4	Alondra Park	2016	1484.00	33.885925	-118.335435	9.026514

Table 3. Data frame with linear distance

For clustering, only MGR and Distance columns were kept and then normalized using *scikit-learn* package. *Figure 5* shows the scatter plot of all 5 clusters together while distance to workplace is shown on X-axis and the MGR is shown on Y-axis.

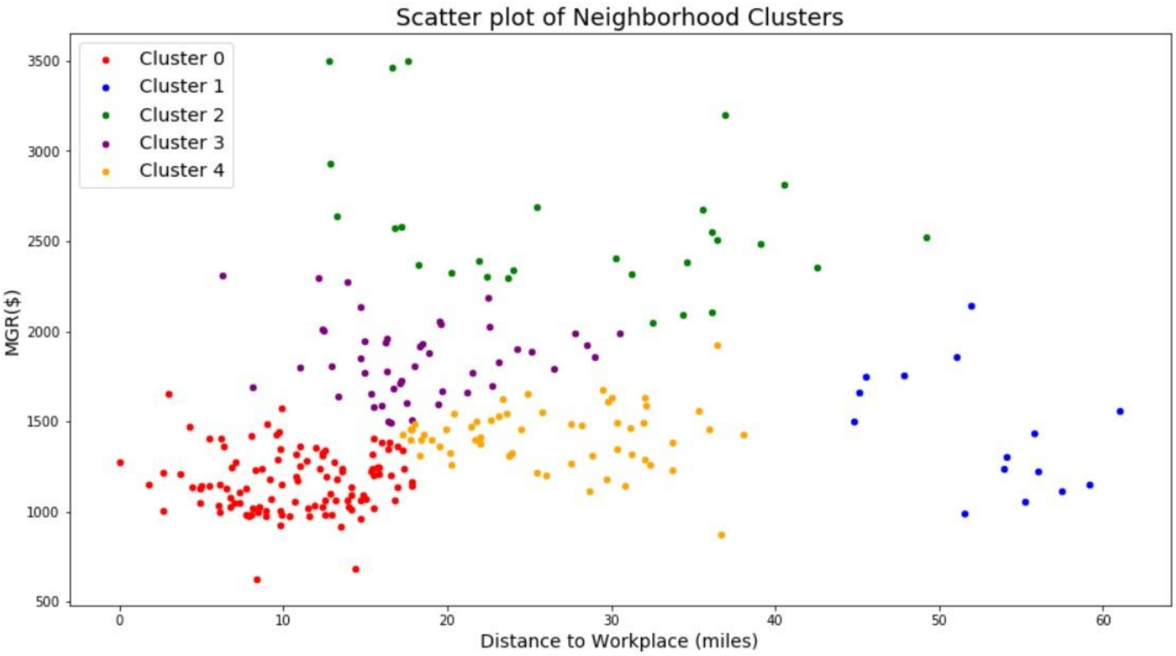


Figure 5. Scatter plot of all clusters

As you see, cluster 0 is the closest to the workplace and has MGR range of $1000-1500 while cluster 1 is farthest and has a wider range of MGR, $1000-2200. Cluster 2 is more spread in respect to its distance from workplace and is the most expensive cluster amongst all 5. While cluster 3 and 4 share almost a similar range of distance, the former has a little higher MGR. In *Figure 6*, you see all 5 clusters superimposed on the map with different colors. The black circle shows the workplace neighborhood, Long Beach.

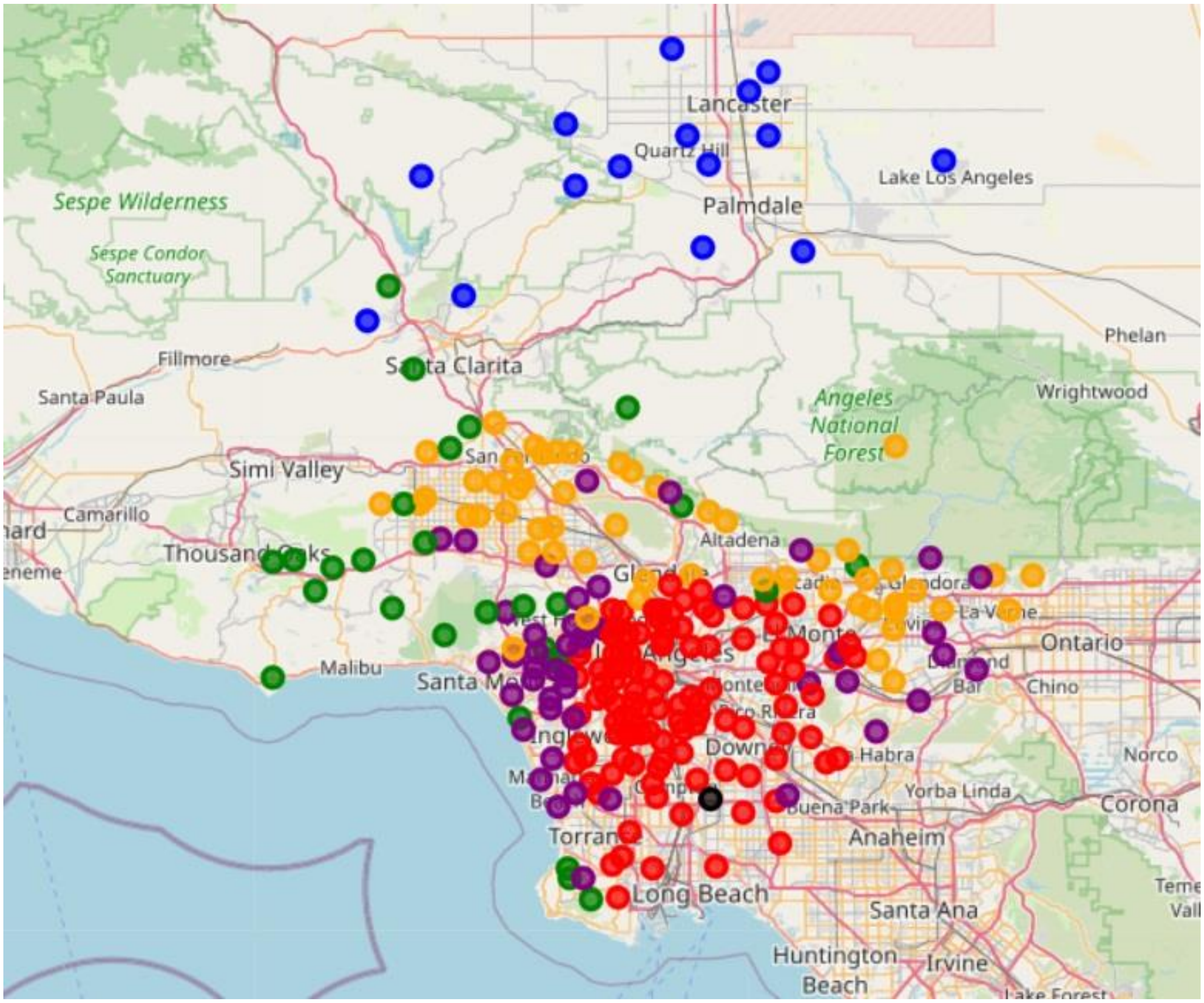
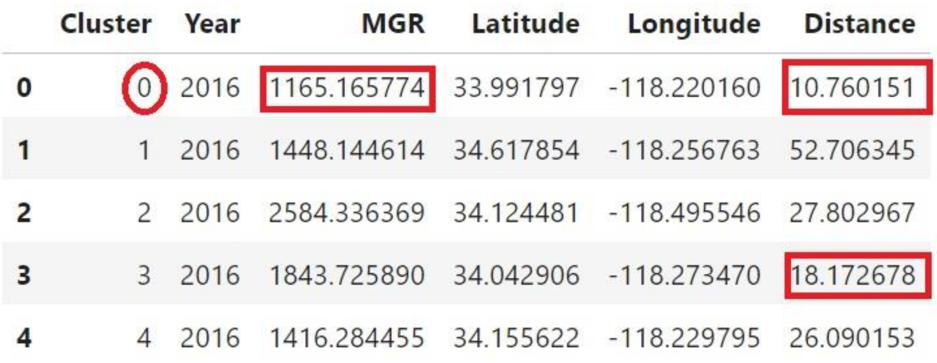


Figure 6. Clusters on the map. The black circle shows the workplace neighborhood.

To select one cluster out of all 5, first the two clusters with lowest average distances to workplace are chosen and then the cluster with lower GMR is marked as the cluster of choice (*Table 4*). As you see, Cluster 0 is the cluster of choice.

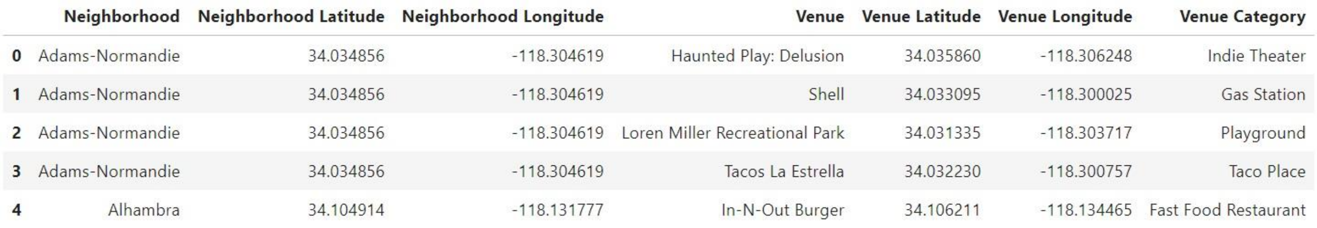


	Cluster	Year	MGR	Latitude	Longitude	Distance
0	0	2016	1165.165774	33.991797	-118.220160	10.760151
1	1	2016	1448.144614	34.617854	-118.256763	52.706345
2	2	2016	2584.336369	34.124481	-118.495546	27.802967
3	3	2016	1843.725890	34.042906	-118.273470	18.172678
4	4	2016	1416.284455	34.155622	-118.229795	26.090153

Table 4. This data frame shows the mean MGR and Distance for each cluster. First the two clusters with lowest Distance are highlighted and then the one with lower GMR is selected as the cluster of choice.

4.4 Filtering the Cluster of Choice

Using Foursquare API, the information about venues around each neighborhood is gathered. These venues are in a radius of 500 meter around each neighborhood. Then, the categories of each venues are extracted and added to the data frame (*Table 5*).



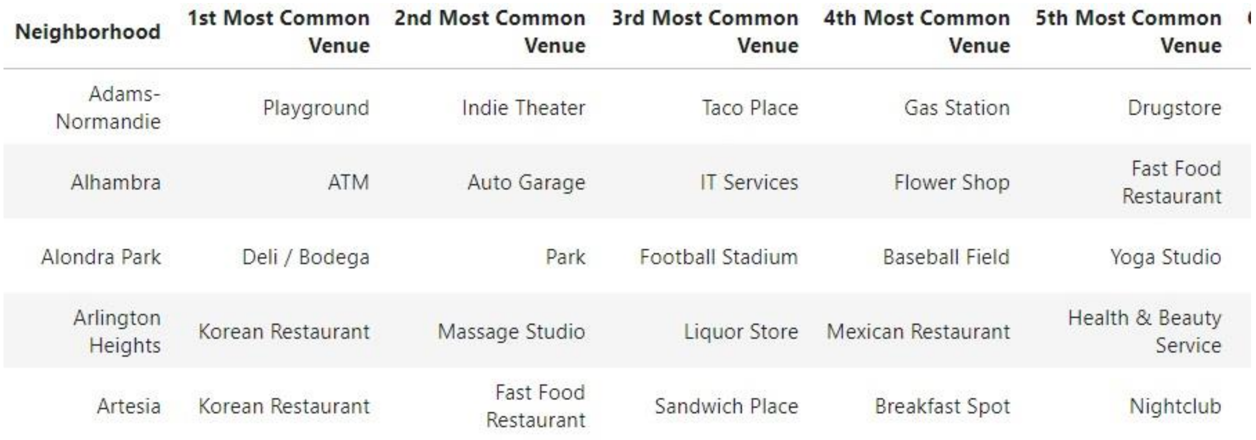
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adams-Normandie	34.034856	-118.304619	Haunted Play: Delusion	34.035860	-118.306248	Indie Theater
1	Adams-Normandie	34.034856	-118.304619	Shell	34.033095	-118.300025	Gas Station
2	Adams-Normandie	34.034856	-118.304619	Loren Miller Recreational Park	34.031335	-118.303717	Playground
3	Adams-Normandie	34.034856	-118.304619	Tacos La Estrella	34.032230	-118.300757	Taco Place
4	Alhambra	34.104914	-118.131777	In-N-Out Burger	34.106211	-118.134465	Fast Food Restaurant

Table 5. Venues and their categories

Then the frequency of each category in each neighborhood has been calculated and 10 most frequent category per each neighborhood is defined in a new data frame (*Table 6*). Then the user selects three venue categories that are important to them for relocation:

Park, Mexican Restaurant and Gas Station.

Based on these three categories, each neighborhood gets a score between 0 to 30. A higher score means that those select categories are more common in that neighborhood. The loop examines all neighborhood and finds the Willowbrook with the highest score of 27 and recommend it for relocation to the user. *Figure 7* shows both workplace and recommend neighborhoods on the map.



Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Adams-Normandie	Playground	Indie Theater	Taco Place	Gas Station	Drugstore
Alhambra	ATM	Auto Garage	IT Services	Flower Shop	Fast Food Restaurant
Alondra Park	Deli / Bodega	Park	Football Stadium	Baseball Field	Yoga Studio
Arlington Heights	Korean Restaurant	Massage Studio	Liquor Store	Mexican Restaurant	Health & Beauty Service
Artesia	Korean Restaurant	Fast Food Restaurant	Sandwich Place	Breakfast Spot	Nightclub

Table 6. Ten most frequent venue categories per each neighborhood. Only 5 first categories are shown.

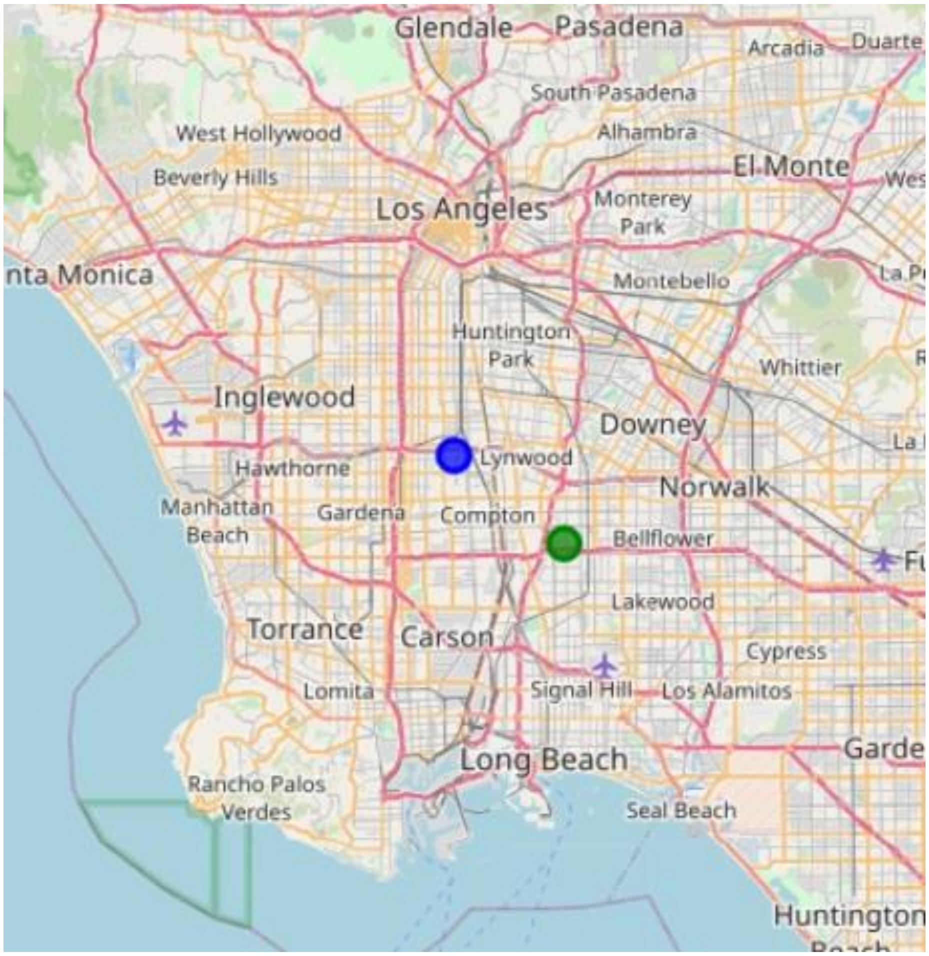


Figure 7. Green circle shows the workplace neighborhood, Long Beach and the blue one shows the recommended neighborhood for relocation, Willowbrook.

5 Discussion

Results showed that Willobrook is the best neighborhood to move in if you get a job in Long Beach. This recommendation is based on the Median Gross Rent of a neighborhood as well as its distance from the workplace. Another factor involved is the venues that are important to the user. Our model showed that clustering the neighborhoods based on two first features will help the user by narrowing down the available choices. Although this model was successful of clustering the neighborhoods, more data seems to be helpful for classifications including but not limited to school districts ranks and crime rate. Also, by evaluating the neighborhood cluster, some outliers are obvious in the data. Removing the outliers before and after clustering will help the model to lead into even better results.

6 Conclusion

This project tried to recommend a neighborhood in Los Angeles county to relocate after getting a new job in the most populous county in the United States with over 10 million residents. This recommendation was based on the MGR, distance to workplace and venues available in proximity of the neighborhood. This model showed that we can achieve this buy having accurate data and comprehensive knowledge of the machine learning. In this model we were limited to Los Angeles county neighborhood and some data were not accurate including some coordinates of neighborhoods. Also, the quantity of venues for some neighborhood was too low and this heavily affected the filtering results. Overall, the model is an example of how machine learning can help humans to make faster decisions which not only save their money but also a lot of their times. This model can be improved by providing it more data as discussed in the discussion section and expand it to the other geographical areas in the Country.