University of California, Los Angeles

Cluster Analysis: Revealing the Highest Revenue of an Airbnb unit in Los Angeles

Price Prediction Ranging from August to December, 2019

Bowen Liu, Zhaojiang He

Statistics 199 - Directed Research in Statistics

Professor Michael Tsiang

13 September 2019

Summary

This research focused on the rental price analysis of different kinds of house layouts as well as locations in the Greater Los Angeles Area. By dividing this area into five sub-areas with respect to their geolocation proximity, this research investigated what type of rental units created the greatest amount of revenue from April 2018 to July 2019, in terms of different layouts and locations. On the other hand, the rental price prediction of each of the best unit identified in each cluster covered from August to December, 2019. With the cleansed data, this research also developed an application that provides future visitors recommendations to find the best living neighborhood in Los Angeles according to their budget.

Introduction

As an important industry of Los Angeles, more precisely speaking, for most large cities all over the world, tourism has a significant effect on both residential income and government tax. According to the "Los Angeles Times," the Los Angeles county had around 50 millions of vistior in 2018, represented a 3.8% increase compared to 2017 and $36.6 billion income. Therefore, providing rental services to visitors is a good choice for making money. However, there does exist a difference in the profit that a landlord can make with respect to different types of houses. In this case, this mainly includes layouts, property types and location, thus choosing the right house at the right place is indispensable for future investors or property owners to maximize their profit.

This research is based on 16 monthly datasets, April 2018 to July 2019, from Airbnb which is the biggest online house rental company in the world. In general, this research has two parts for the investors and visitors individually. The first part has six sections: data cleaning, model selection, data clustering, rental price analysis, conclusion and limitation for the investors to get a better idea about where and what kind of house earn the most profit; the second part includes an application for future visitors to find the best neighborhood to live according to their budgets.

Part One

Data cleaning

For each of the 16 monthly data, there are over 100 variables and more than 46,000 observations. In this case, the research focused on the numeric and logical variables which have a more obvious relationship with our topic, type of houses with highest income, rather than categorical variables except "property type" and "neighborhood" which would be used later. Finally, there are 50 available numeric variables in spite of ID for the further process.

For the numeric variables, there are four steps taken:

1. "$" in some numeric variables such as "price" and "cleaning fee", are removed and transformed into numeric values;
2. The logical variables ("True" and "False") "host_is_superhost" and "requires_license" are transformed into 1s and 0s.
3. For the variables such as "bathroom" and "bedroom", the observations with 0s are removed since they make no sense for people to live in; in addition, since both "bathroom" and "bedroom" are the variables chosen for the model, which will be explained in the next section, we only keep the number of bathroom within 99 percentile.

Moreover, under the table of “bedroom”, the numbers of bedroom 1, 2, and 3, all of which individually is greater than the sum of the rest bedroom number greater or equal to 4, occupy around 93% of the data, thus we create a label of “4” for the number of bedroom greater or equal to 4. For the interpretation illustrated later in the report, bedroom 4 means having 4 or more bedrooms.

4. For location related variables, such as “zipcode”, any value represents outside of Great LA area are removed because this research is limited in the Los Angeles area. This is because some typos exists in the zipcode column.

For categorical variables “property type” and “neighborhood”:

1. For property_type, we looked into the frequency of each property type that exist in the data set and decided to investigate the unit type that are over 5% of the total frequency. As a result, four property types: “Apartment”, “House”, “Condominium” and “Guesthouse” are taken into consideration. They take up around 82% of all the data.
2. For neighborhood, it is the same as all other variables, the rows with missing values are dropped.

Variable Selection

For maximizing the revenue of hosting with Airbnb, price is one of the most important factors without question. What comes after the price variable is that is occupancy rate, but such a variable is not included in this dataset, thus model selection will depend on the correlation between price and other 49 numeric variables by pair.

As a result, 37 out of 50 numeric variables are statistically significant to the variable price using Pearson’s Correlation Test. To find the layouts and location of houses with a high return potential along with determining significant numeric variables, this research first selected categorical variables “property type” and “neighborhood”, followed by “bedroom”, “bathroom”, “longitude”, “latitude” ,“zipcode”, “square feet”, as the statistically significant physical conditions that cannot be altered easily, while other significant variables such as “cleaning fee” that can be changed at owner’s willing would not be considered in this model. However, the recommendations for future host or investor would be based on the variables found to be statistically significant and other variables related to those important factors would be included at the end of this report. On the other hand, since the number of missing data of “square feet” is over 95%, it is dropped. In case there is a collinearity issue, we also checked the VIF value of each pairs, finding all of them having a value below 5. There is no collinearity problem here.

Finally, as mentioned above, the price and occupancy rate are the two most important components of income, whereas, there is no variable here recorded occupancy rate. Therefore, this research used another variable named “review per month” as a standard to measure how often the target house was booked every month.

Data clustering

From the Inside Airbnb website, we are able to collect the datasets from April 2018 to July 2019, and most of our data set consists of more than 40,000 rows and 106 columns. To begin with, we performed a geospatial analysis based on the longitude and latitude variables of the latest dataset on July 2019. Since longitude and latitude are two columns of numeric variables, we decided to perform K-Means Clustering using euclidean distance as a measure of geolocation proximity. By grouping numeric input together based on the euclidean distance, the K-means algorithm is inevitably affected by the “set.seed()” function. However, the K-kmeans algorithm returns a different output when the same code run on different computers. To have the consistent output, we chose to use “RNGkind(sample.kind = "Rounding")”, which allows us to get a consistent cluster and perform similar analysis across different computers.

Elbow method formula:

$$minimize\left(\sum_{k=1}^{k} W(C_k)\right)$$

Where Ck is the kth cluster and W(Ck)W(Ck) is the within-cluster variation. The total within-cluster sum of squares (wss) measures the compactness of the clustering and we want it to be as small as possible.

With the aid of the elbow method, implementing the K-means clustering algorithm allows us to get the optimal value of k (number of clusters) for our dataset. The way to choose the optimal k is looking k values that minimize the total within-cluster sum of square (wss). To find the wss, we computed the k-means clustering for different values of k value from 1-10. As shown in Figure 1, we chose 5 clusters so that adding another cluster did not give as much drop in SSE as it did in the first 5 k values.

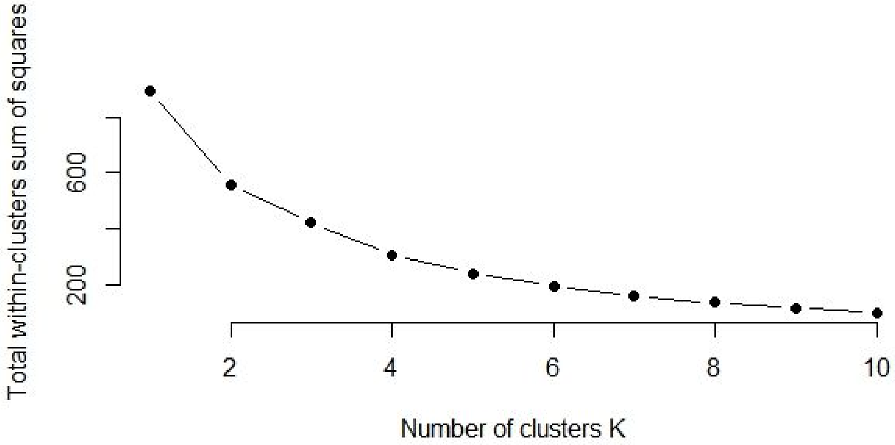


Figure 1: The graph of implementing Elbow Method to find the optimal value of k

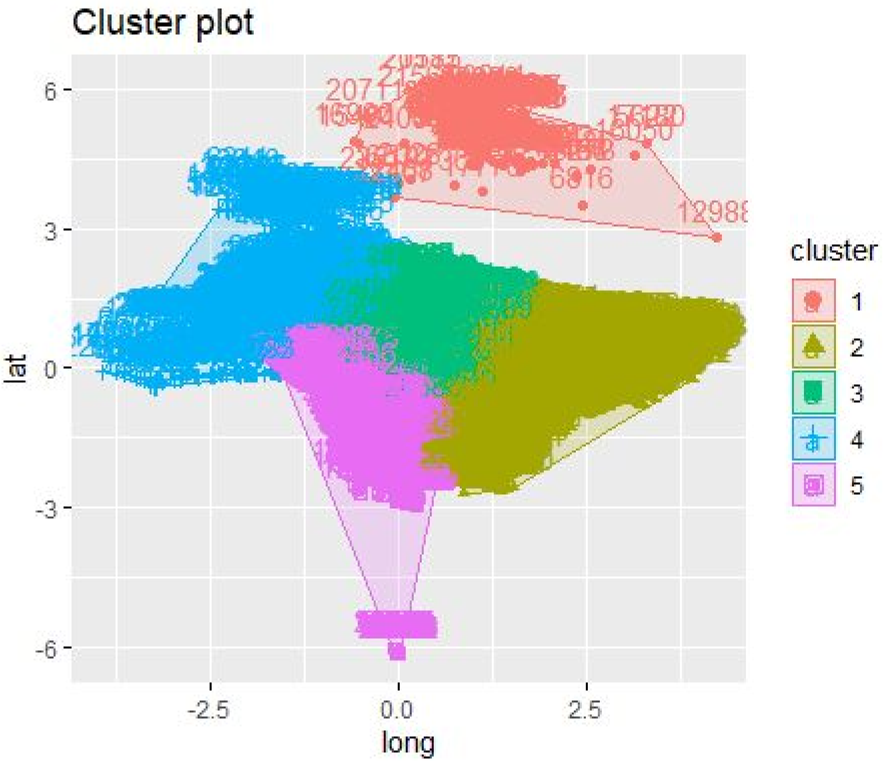


Figure 2: The graph of visualizing five clusters with "factoextra" package in R

Implementing clustering based on the longitude and latitude is more accurate and reasonable, but the issue is that some observations on the boundaries of different cluster region may share the same zip code because clustering longitude and latitude does not take the distribution of different zip code zone into consideration. Moreover, we clustered based on the 2019 July dataset, the K-means might cluster the areas differently in other months with different longitude and latitude. Therefore, in order to keep the regional clustering same for all the months, we performed some careful manual manipulations to reassign the zip codes on the boundary of two groups to only one group. We compared the number of unique zip codes in each cluster and found that group 3 has the greatest number of zip codes, then group 2, 4, 1 and 5 with the smallest number of zip codes. In this process, since losing several zip codes does not have a big effect on the groups with large number of zip codes, we started from group 5 to take the zip code shared with the others, then group 1, 4, and 2. After reallocating those common zip codes, every group has its own and unique zip code, thus rather than running K-means each time with new data from a month, we clustered the new data using the zip codes in different groups from July 2019.

Group Analysis

Since we have implemented K-mean clustering based on the longitude and latitude of Airbnb listings from April 2018 to July 2019 and reassigned some zip codes, we had our data sets divided into five groups. We would look into the similarities shared within a group and the difference between each group of the data. We would reveal what neighbors and zip codes belong to each group and the best financial return correspondingly. By using the 'ggmap' package in R along with the Google Static Map API, we were able to create a visualization of Airbnb listings on top of a map layer of the Greater Los Angeles Area.

Moreover, there is one more step we do before performing the analysis: using the “Moving Average Method” to simulate the data from January 2018 to March 2018, which are missing for all groups, as well as imputing any missing monthly data between April 2018 - July 2019. In total, we have 19-month-long data.

Last but not least, since we use the weighted mean and mean of review per month to calculate the total income, it might cause a “jump” in the price if any host in that type of houses just starts or stops to be an Airbnb host. Therefore, for the price prediction of the best income house of each group in the next five months, there is an assumption that no current host will drop or join to the current market. In this case, the trend of price will tend to be stable because most rental price of a specific house does not change much.

Group 1

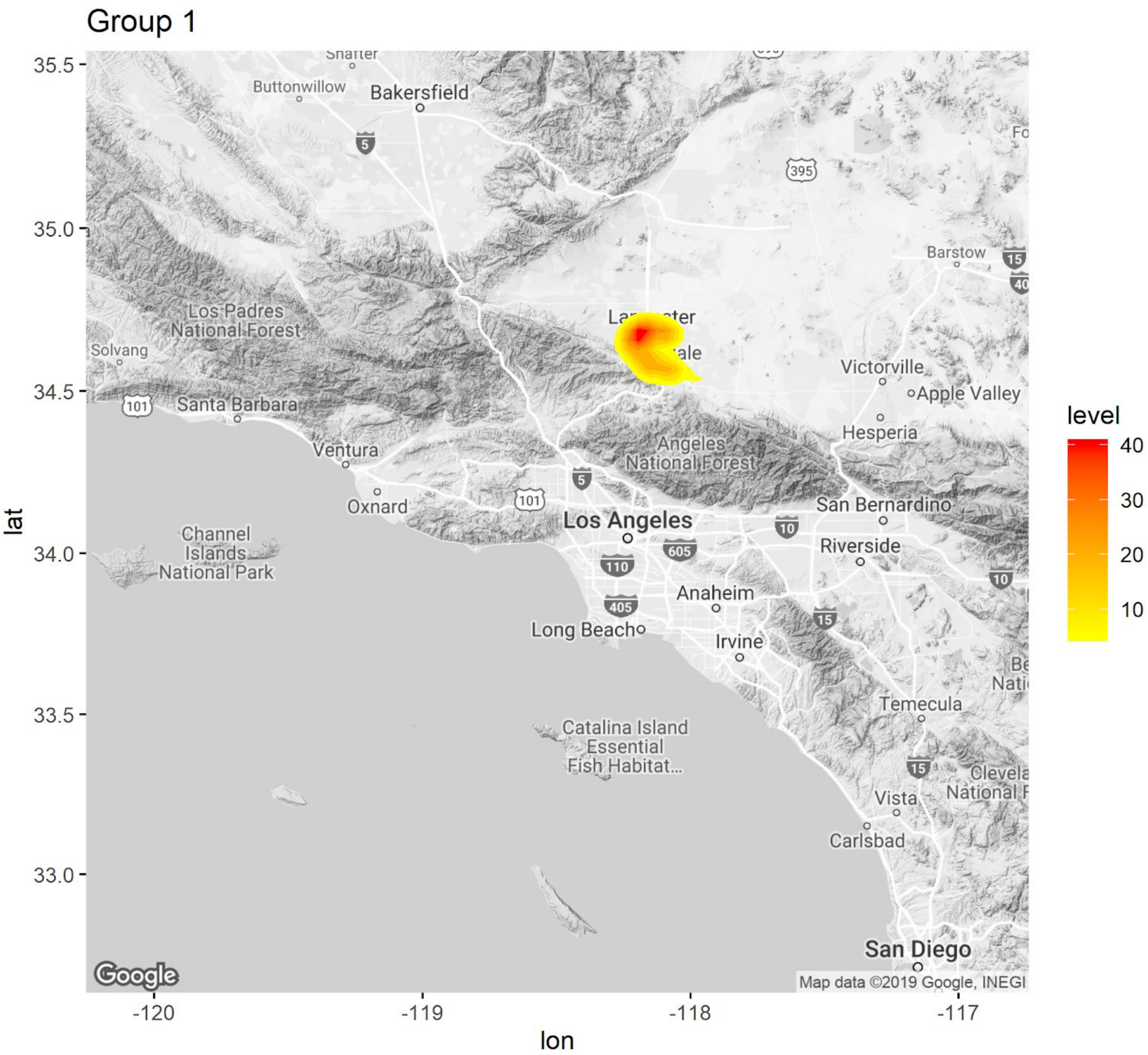


Figure 3: Group 1 visualization of Airbnb listings in the Great Los Angeles Area

```
 [1] Sun Village              Lancaster
 [3] Agua Dulce               Palmdale
 [5] Lake Los Angeles         Southeast Antelope Valley
 [7] Acton                    Northwest Palmdale
 [9] Green Valley             Quartz Hill
[11] Northwest Antelope Valley Desert View Highlands
[13] Leona Valley             Castaic Canyons
```

Figure 4: Neighborhoods in group 1

Based on the total revenue of each type of places with different layouts and locations, we found the property type "house" with 4 or more bedrooms, 3 bathroom, and zip code at 93551 had the greatest income for being a host on the Airbnb platform. Even though this type of house only appears 7 months, its total income is almost 1.5 times as much as the second position in group 1, and its rental price changes are shown below with only one outlier:

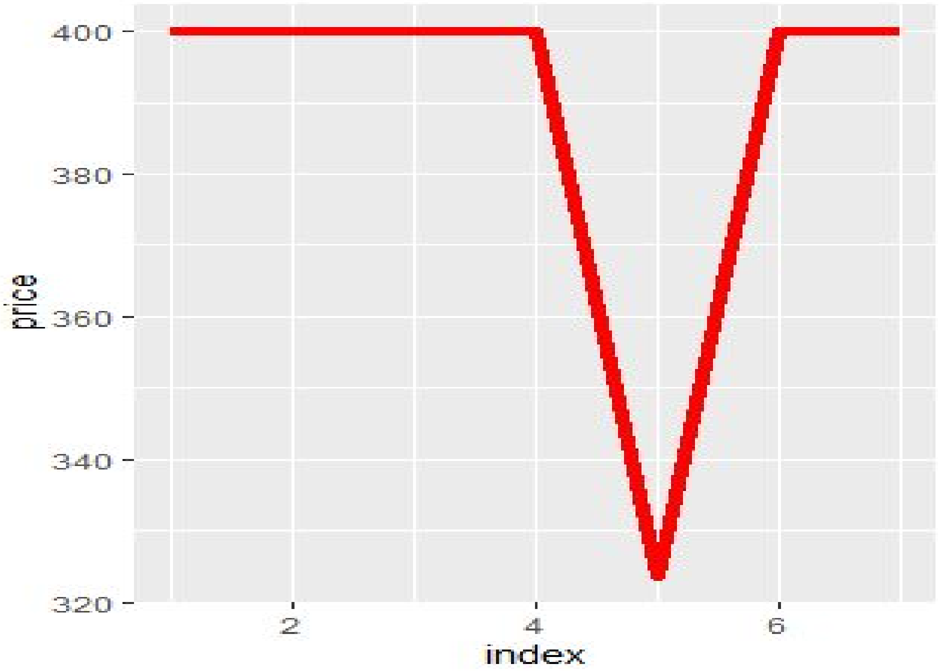


Figure 5: Houses with 4 or more bedrooms, 3 bathrooms, zip code at 93551

As Figure 5 shows, the price variable is a fairly constant value, the prediction for the price in the next five months will be 400 dollars. This is a reasonable prediction because there is only one outlier out of many data points obtained from April 2018 - July 2019.

Group 2

The second group also captures the similarity of geolocation of Airbnb listings of data from April 2018 to July 2019. Interestingly, group 2 consists of the 3 areas in the east and southern part of the Great Los Angeles Area. The Southern subgroup takes the peak level of the density of Airbnb listing within this cluster since the dark red indicates the highest density of Airbnb listings in that region.As shown in Figure 6, the density level reaches its peak at the location near Long Beach. That subgroup covered some popular beaches and attractions for tourists and local people, including The Queen Mary, Aquarium of the Pacific, Long Beach Museum of Art, and the Museum of Latin American Art (MOLAA).

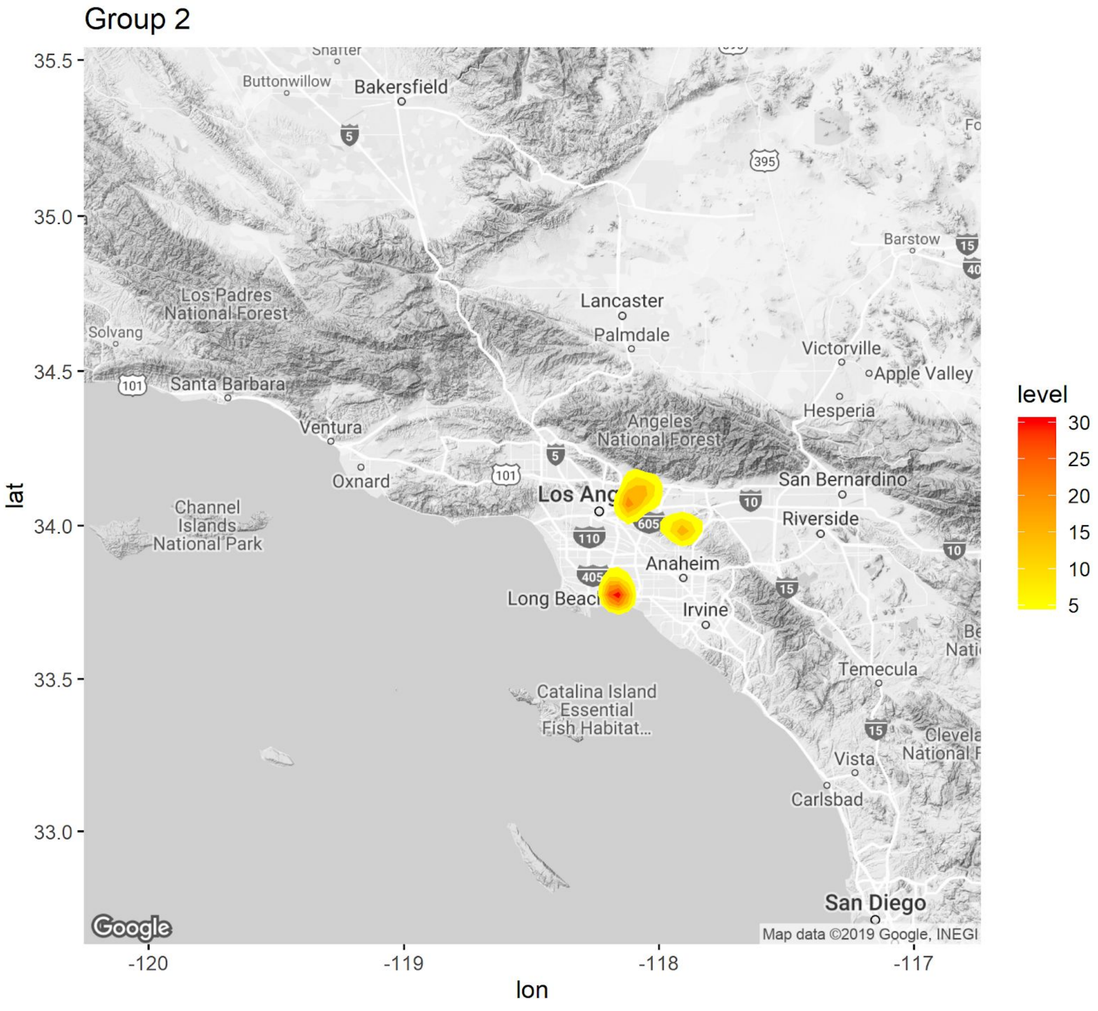
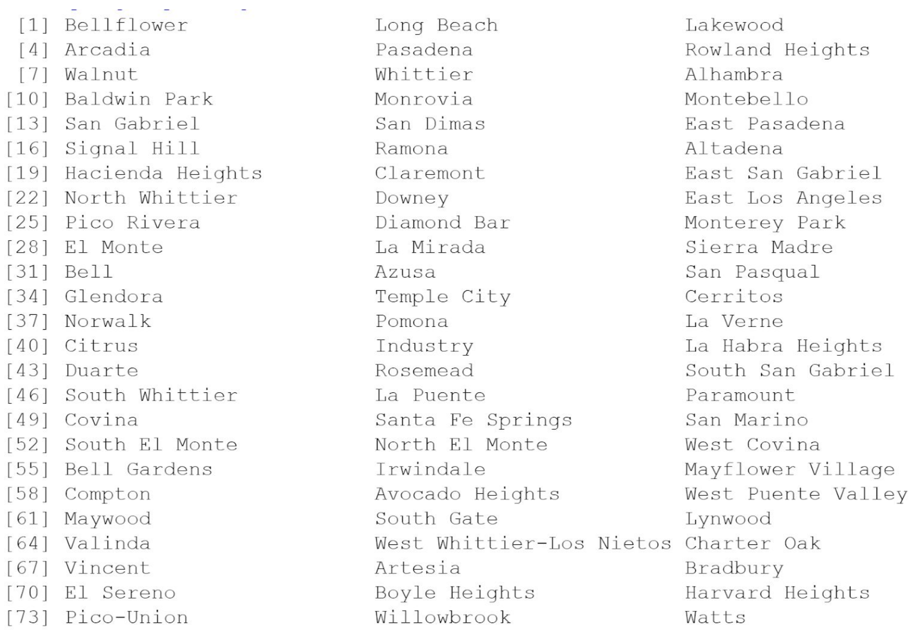


Figure 6: Group 2 visualization of Airbnb listings in the Greater Los Angeles Area



```
 [1] Bellflower            Long Beach                Lakewood
 [4] Arcadia               Pasadena                  Rowland Heights
 [7] Walnut                Whittier                  Alhambra
[10] Baldwin Park          Monrovia                  Montebello
[13] San Gabriel           San Dimas                 East Pasadena
[16] Signal Hill           Ramona                    Altadena
[19] Hacienda Heights      Claremont                 East San Gabriel
[22] North Whittier        Downey                    East Los Angeles
[25] Pico Rivera           Diamond Bar               Monterey Park
[28] El Monte              La Mirada                 Sierra Madre
[31] Bell                  Azusa                     San Pasqual
[34] Glendora              Temple City               Cerritos
[37] Norwalk               Pomona                    La Verne
[40] Citrus                Industry                  La Habra Heights
[43] Duarte                Rosemead                  South San Gabriel
[46] South Whittier        La Puente                 Paramount
[49] Covina                Santa Fe Springs          San Marino
[52] South El Monte        North El Monte            West Covina
[55] Bell Gardens          Irwindale                 Mayflower Village
[58] Compton               Avocado Heights           West Puente Valley
[61] Maywood               South Gate                Lynwood
[64] Valinda               West Whittier-Los Nietos Charter Oak
[67] Vincent               Artesia                   Bradbury
[70] El Sereno             Boyle Heights             Harvard Heights
[73] Pico-Union            Willowbrook               Watts
```

Figure 7: Neighborhoods in Group 2

Based on the total income of each type of place with different layouts and location, we found the property type “apartment” with 2 bedrooms, 1 bathroom and zip code 91107 made the most income in the past 19 months. As shown in Figure 8, the trend of price is very stable at the beginning, followed by a few jumps.

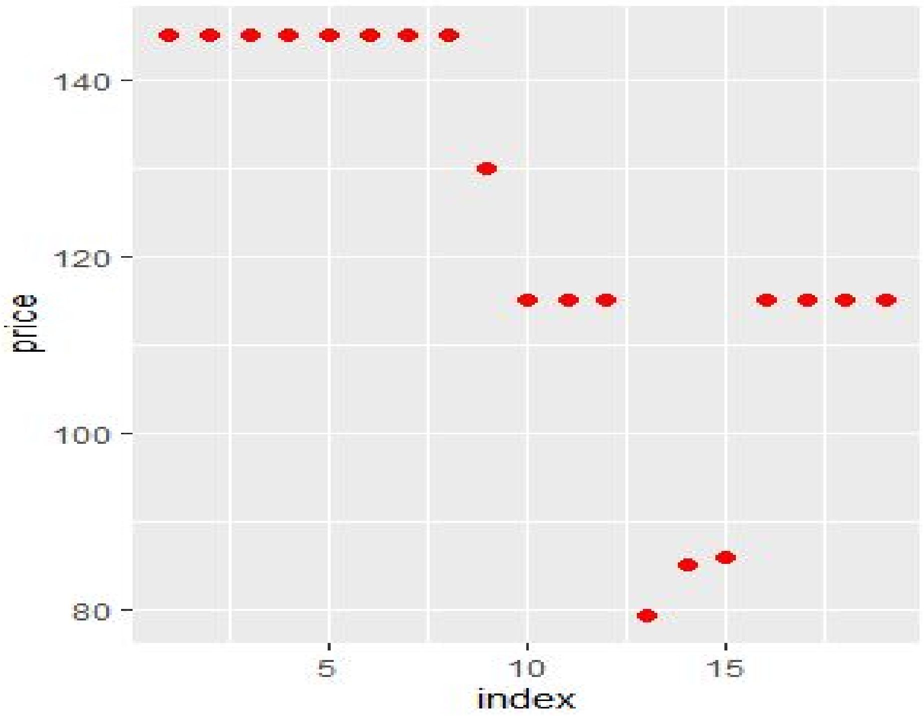


Figure 8: Airbnb listings of apartments with 2 bedrooms, 1 bathroom, and zip code 91107

By doing the log transformation to the price and splitting this dataset into 5 groups (months 1-8, 9, 10-12, 13-15, 16-19), we find the variance is nearly constant as 0. Therefore, assuming the price will not change much in the next five months, the prediction price will be 115.

Group 3

There are so many must-see attractions in this area. The mode of the zip code in group 3 is 90028, which covers the heart of the Greater Los Angeles Area, including the West Hollywood area, Walk of Fame, Chinese Theatre. As shown in Figure 9, it is very close to Universal City in which Universal Studios Hollywood, a film studio and theme park is located. Similarly, Warner Brother Studio l, Walt Disney Studios, and Marvel Studio are all located in

Burbank. What's more, the Hollywood sign and Griffith Observatory are fairly close. Tourists can take a short uber ride to different attractions easily. The greatest density being over 150 is reasonable and significantly greater than other groups derived from our clustering algorithms.

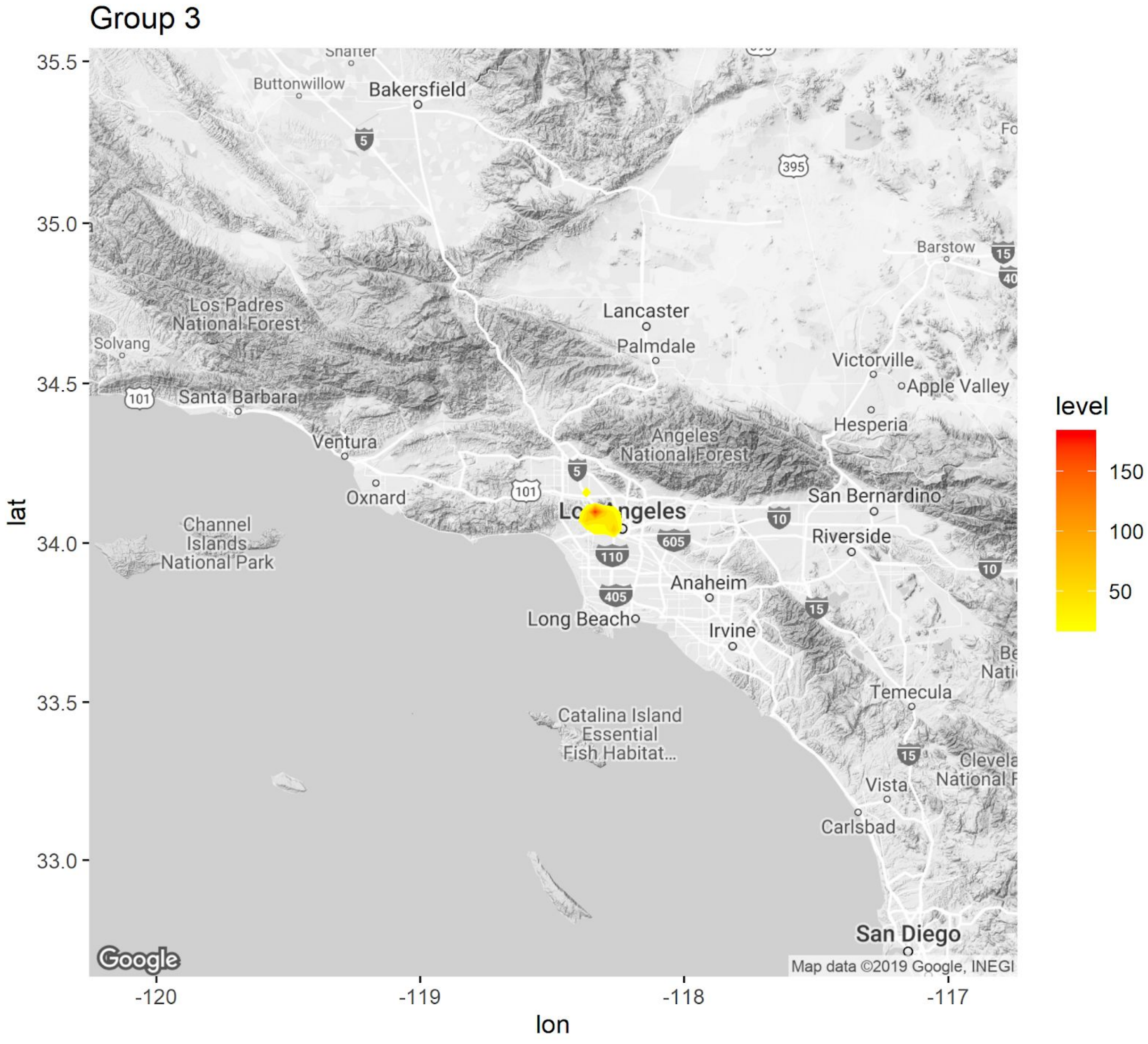
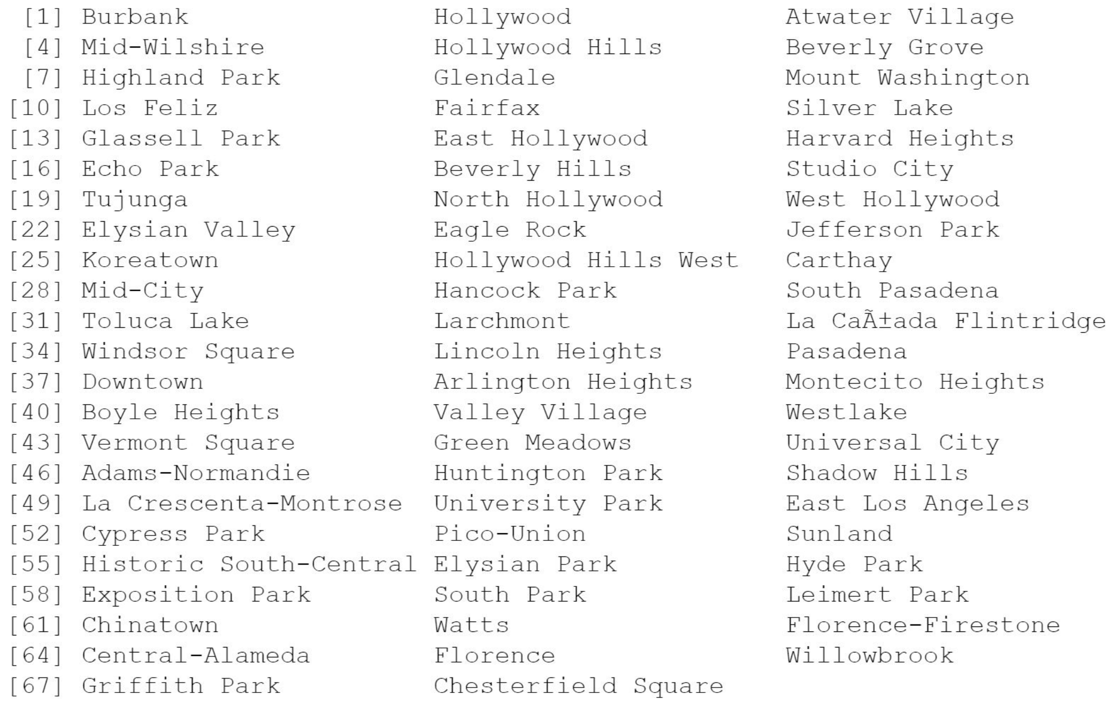


Figure 9: Group 3 visualization of Airbnb listings in the Greater Los Angeles Area



```
 [1] Burbank                 Hollywood               Atwater Village
 [4] Mid-Wilshire            Hollywood Hills         Beverly Grove
 [7] Highland Park           Glendale                Mount Washington
[10] Los Feliz               Fairfax                 Silver Lake
[13] Glassell Park           East Hollywood          Harvard Heights
[16] Echo Park               Beverly Hills           Studio City
[19] Tujunga                 North Hollywood         West Hollywood
[22] Elysian Valley          Eagle Rock              Jefferson Park
[25] Koreatown               Hollywood Hills West    Carthay
[28] Mid-City                Hancock Park            South Pasadena
[31] Toluca Lake             Larchmont               La CaÃ±ada Flintridge
[34] Windsor Square          Lincoln Heights         Pasadena
[37] Downtown                Arlington Heights       Montecito Heights
[40] Boyle Heights           Valley Village          Westlake
[43] Vermont Square          Green Meadows           Universal City
[46] Adams-Normandie         Huntington Park         Shadow Hills
[49] La Crescenta-Montrose   University Park         East Los Angeles
[52] Cypress Park            Pico-Union              Sunland
[55] Historic South-Central  Elysian Park            Hyde Park
[58] Exposition Park         South Park              Leimert Park
[61] Chinatown               Watts                   Florence-Firestone
[64] Central-Alameda         Florence                Willowbrook
[67] Griffith Park           Chesterfield Square
```

Figure 10: Neighborhoods in Group 3

Based on the total income of each type of places with different layouts and location, we found the property type “Apartment” with 3 bedrooms, 3 bathroom, and zip code 90028 made the most income in the past 19 months, and the price changes is shown below:

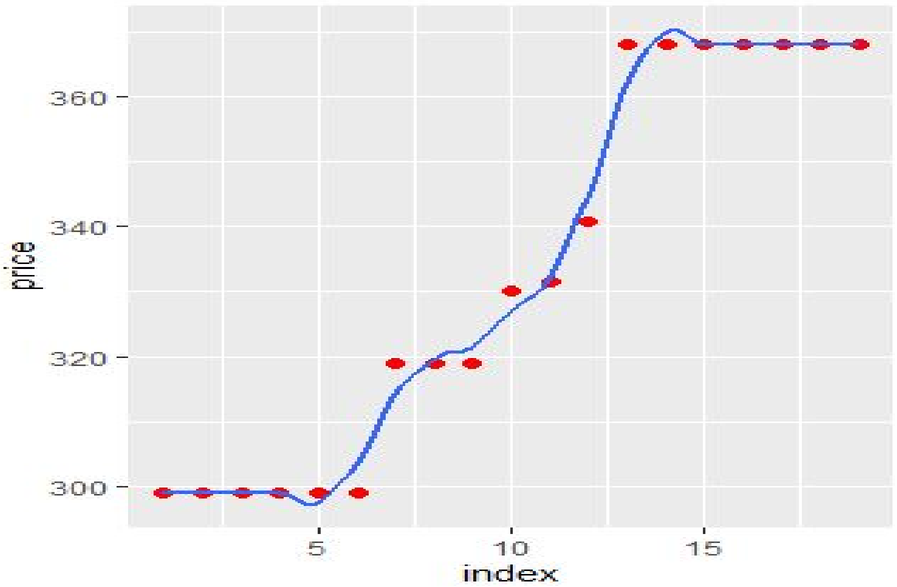


Figure 11: Price of apartments with 3 bedrooms, 3 bathrooms, and zip code 90028

By using the loess regression with an optimum span of 0.37, we predict the price in the next five months is constantly 368. Based on having a flat line in 2019, this prediction is consistent with the pattern we have observed in Figure 11.

Group 4

The mode of the zip code in group 4 is 91604, which corresponds to the North Western Los Angeles Area. The area next to the North Hollywood is Burbank. Different studios mentioned in group 3 analysis also bring a positive influence on the emergence of Airbnb listings in this area. It is likely that tourists would reserve Airbnb in this region because they book it only a few days before arrival or they tend to find a relatively cheaper place to stay while traveling. This region is also close to the Six Flag: Magic Mountain. As shown in Figure 12, this visualization confirms the popularity of some attractions like the Warner Brother Studios Hollywood tour because of the peak of density shown on the graph.

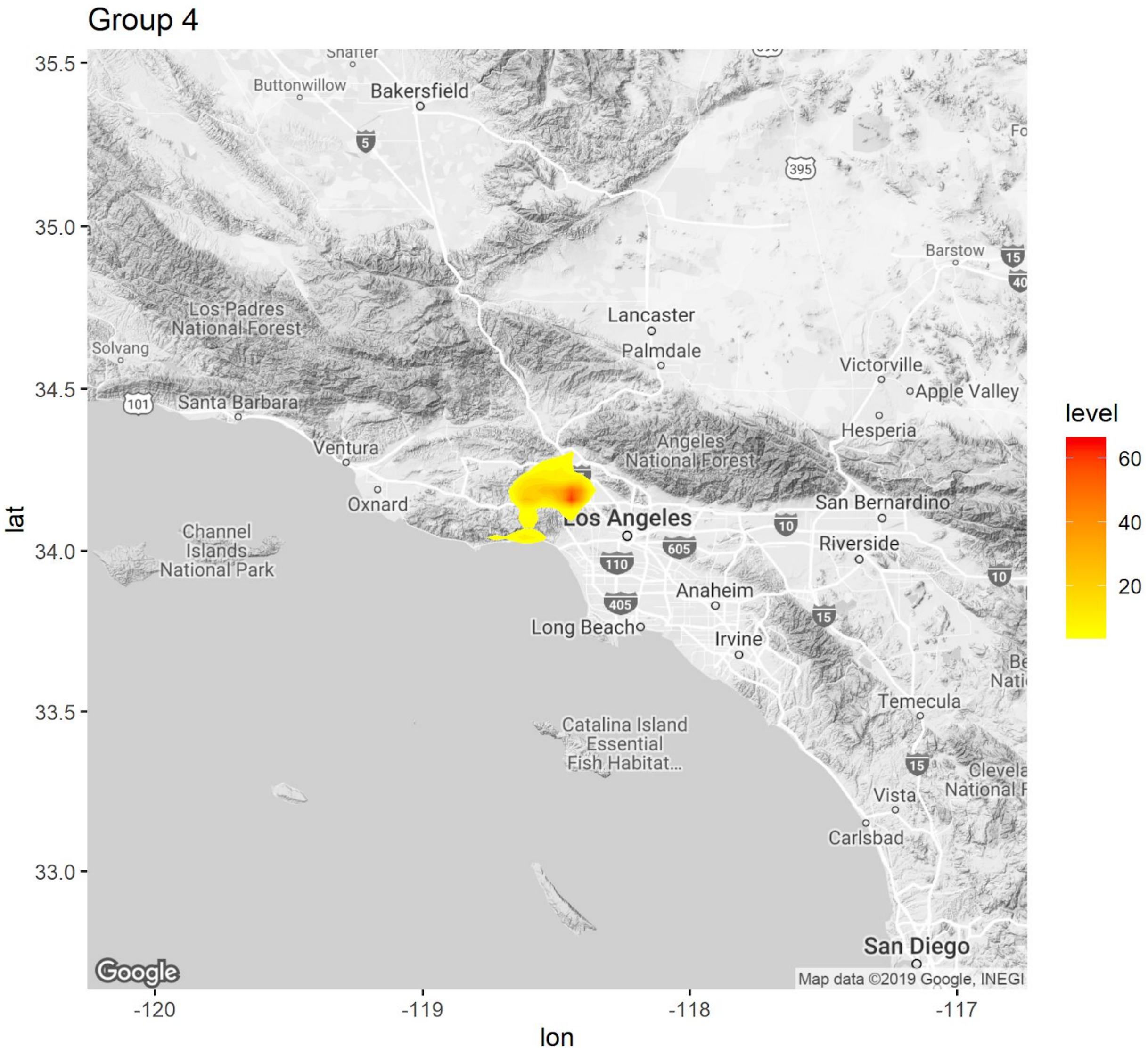
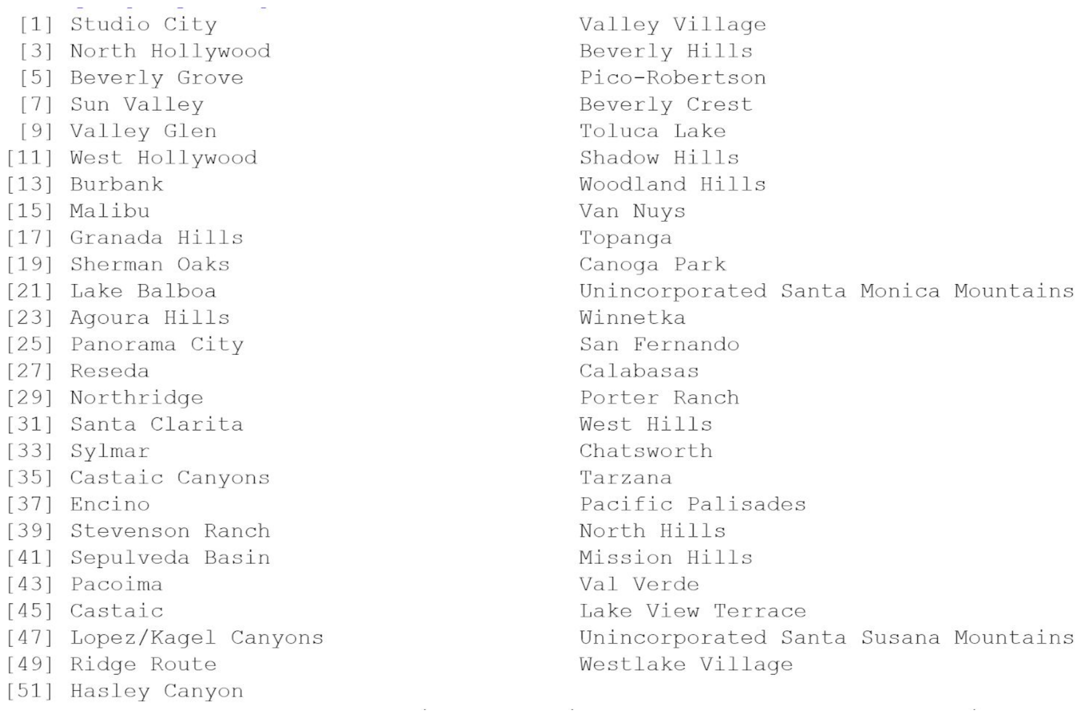


Figure 12: Group 4 visualization of Airbnb listings in the Great Los Angeles Area



```
 [1] Studio City                       Valley Village
 [3] North Hollywood                   Beverly Hills
 [5] Beverly Grove                     Pico-Robertson
 [7] Sun Valley                        Beverly Crest
 [9] Valley Glen                       Toluca Lake
[11] West Hollywood                    Shadow Hills
[13] Burbank                           Woodland Hills
[15] Malibu                            Van Nuys
[17] Granada Hills                     Topanga
[19] Sherman Oaks                      Canoga Park
[21] Lake Balboa                       Unincorporated Santa Monica Mountains
[23] Agoura Hills                      Winnetka
[25] Panorama City                     San Fernando
[27] Reseda                            Calabasas
[29] Northridge                        Porter Ranch
[31] Santa Clarita                     West Hills
[33] Sylmar                            Chatsworth
[35] Castaic Canyons                   Tarzana
[37] Encino                            Pacific Palisades
[39] Stevenson Ranch                   North Hills
[41] Sepulveda Basin                   Mission Hills
[43] Pacoima                           Val Verde
[45] Castaic                           Lake View Terrace
[47] Lopez/Kagel Canyons               Unincorporated Santa Susana Mountains
[49] Ridge Route                       Westlake Village
[51] Hasley Canyon
```

Figure 13: Neighborhood in Group 4

Based on the total income of each type of places with different layouts and location, we found the property type “Apartment” with 3 bedrooms, 3 bathroom and zip code 90265 made the most income in the past 19 months. The trend of the price variables changes is shown in Figure 14.

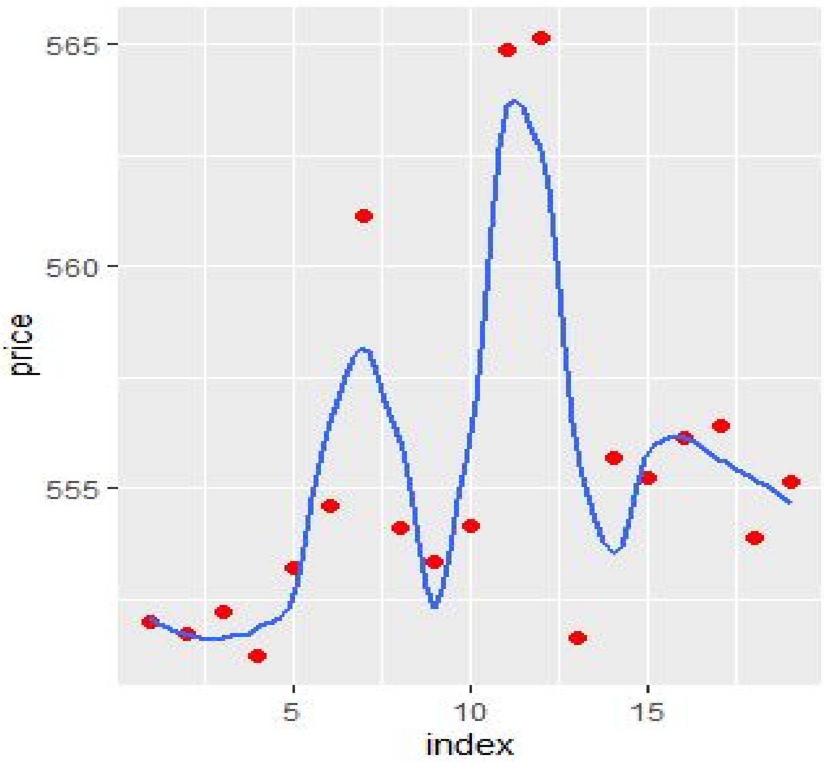


Figure 14: Price of apartments with 3 bedrooms, 3 bathrooms, and zip code 90265

As shown in the blue line in Figure 14, by using the loess regression with an optimum span at 0.42, we predict the price in the next five months is 554.1247 , 553.7799, 553.6798 , 553.8657, 554.3638 and 555.1912. We do not expect a huge change in price in this case.

Group 5

This group indicates the density of Airbnb listings in the West Los Angeles region. As for tourism, there are some famous attractions: Santa Monica Beach and Venice Beach. As for transportation, the Los Angeles International Airport is also located on the westside of Los Angeles. Venice Beach is famous for having a lot of surfers, skateboarders, and stars. There are also some museums: the Getty Center, Los Angeles County Museum of Art. the Getty Center is known for its architecture, gardens, and views overlooking Los Angeles. The Los Angeles and the Petersen Automotive Museum are near this region as well.

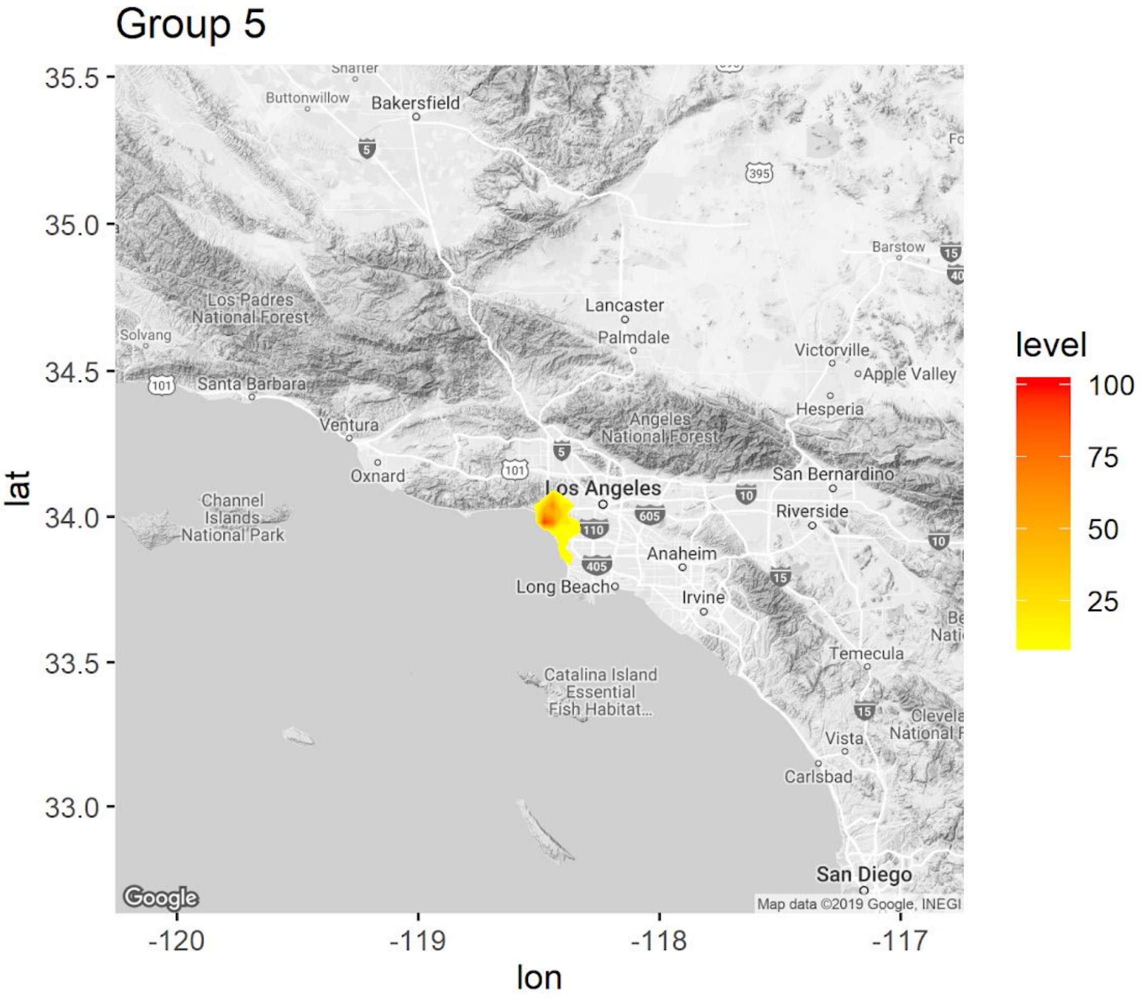
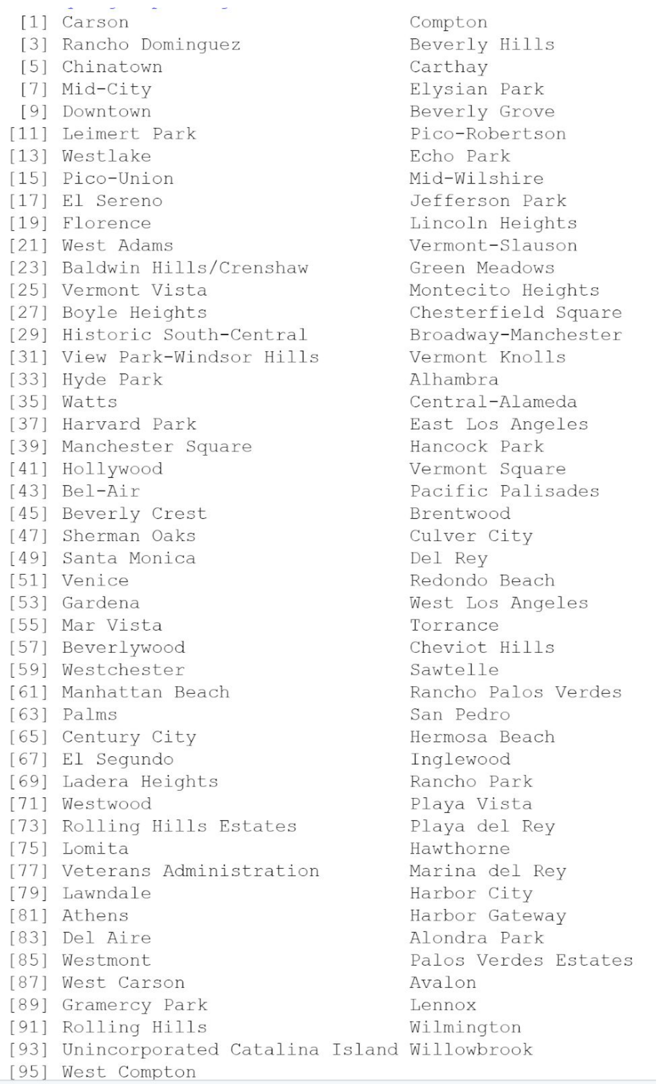


Figure 15: Group 5 visualization of Airbnb listings in the Greater Los Angeles Area



```
 [1] Carson                        Compton
 [3] Rancho Dominguez              Beverly Hills
 [5] Chinatown                     Carthay
 [7] Mid-City                      Elysian Park
 [9] Downtown                      Beverly Grove
[11] Leimert Park                  Pico-Robertson
[13] Westlake                      Echo Park
[15] Pico-Union                    Mid-Wilshire
[17] El Sereno                     Jefferson Park
[19] Florence                      Lincoln Heights
[21] West Adams                    Vermont-Slauson
[23] Baldwin Hills/Crenshaw        Green Meadows
[25] Vermont Vista                 Montecito Heights
[27] Boyle Heights                 Chesterfield Square
[29] Historic South-Central        Broadway-Manchester
[31] View Park-Windsor Hills       Vermont Knolls
[33] Hyde Park                     Alhambra
[35] Watts                         Central-Alameda
[37] Harvard Park                  East Los Angeles
[39] Manchester Square             Hancock Park
[41] Hollywood                     Vermont Square
[43] Bel-Air                       Pacific Palisades
[45] Beverly Crest                 Brentwood
[47] Sherman Oaks                  Culver City
[49] Santa Monica                  Del Rey
[51] Venice                        Redondo Beach
[53] Gardena                       West Los Angeles
[55] Mar Vista                     Torrance
[57] Beverlywood                   Cheviot Hills
[59] Westchester                   Sawtelle
[61] Manhattan Beach               Rancho Palos Verdes
[63] Palms                         San Pedro
[65] Century City                  Hermosa Beach
[67] El Segundo                    Inglewood
[69] Ladera Heights                Rancho Park
[71] Westwood                      Playa Vista
[73] Rolling Hills Estates         Playa del Rey
[75] Lomita                        Hawthorne
[77] Veterans Administration       Marina del Rey
[79] Lawndale                      Harbor City
[81] Athens                        Harbor Gateway
[83] Del Aire                      Alondra Park
[85] Westmont                      Palos Verdes Estates
[87] West Carson                   Avalon
[89] Gramercy Park                 Lennox
[91] Rolling Hills                 Wilmington
[93] Unincorporated Catalina Island Willowbrook
[95] West Compton
```

Figure 16: Neighborhood in Group 5

Based on the total income of each type of places with different layouts and location, we found the property type “Apartment” with 3 bedrooms, 3 bathroom and zip code 90015” made the most income in the past 19 months, and the price changes are shown below:

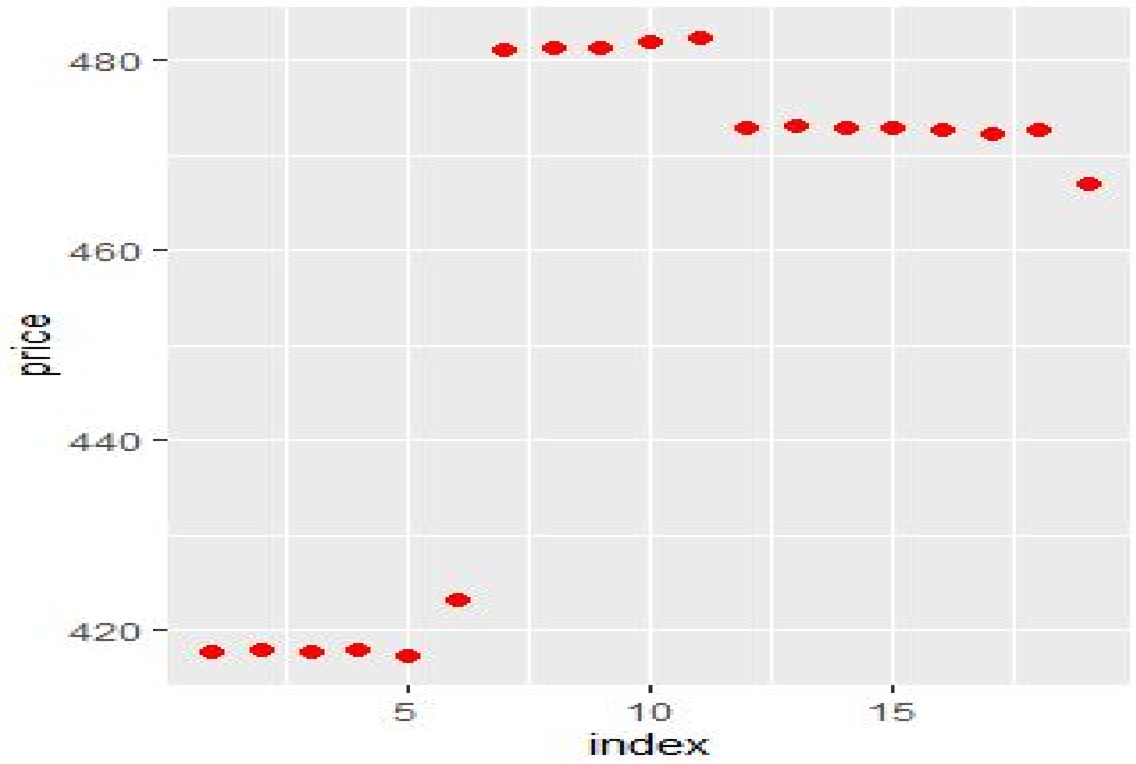


Figure 17: Price of apartment with 3 bedrooms, 3 bathrooms and zip code 90265

By performing the log transformation to the price and splitting this dataset into 5 groups(months 1-6, 9, 7-11, 12- 19), we find the variance is nearly constant as 0. Therefore, assuming the price will not change much in the next five months, the prediction price will be 456.9 constantly in those months.

Suggestions For Future Airbnb Hosts:

As the report mentioned at the variable selection part, there are several statistically significant numeric variables that were not used, which are more likely to be a personal requirement among different owners. After summarizing these variables, we find several features that future Airbnb host should pay attention to:

1. Precise location to the customer
2. Number of people can live in and charge for extra people
3. Amount of service fees such as cleaning & security
4. Availability for being a host
5. Check-in procedure
6. Customer Reviews

Conclusion

Throughout our research about the price analysis of different layouts and geolocation in the Greater Los Angeles Area by implementing the elbow method with K-means clustering provides information and dividing the whole area into 5 groups. After removing extreme cases and reassign zip codes on the group boundaries for a better interpretability, we are able to find the type of houses with the best income, which is calculated by weighted average price and number of review, within each group in terms of "number of beds, number of baths, type of unit, zip code". More so, we also make the price predictions in the next five month assuming that no host drop or add in.

As a conclusion, house with more than 4 bedrooms and 3 bathrooms make the most income as an Airbnb host in group one, which is around upper North Los Angeles; apartment with 2 bedrooms and 1 bathroom make the most income in group 2, which is around East and South Los Angeles; apartment with 3 bedrooms and 3 bathrooms make the most income in group 3, which is around downtown Los Angeles; apartment with 3 bedrooms and 3 bathrooms make the most income in group 4, which is around north western Los Angeles; apartment with 3 bedrooms and 3 bathrooms make the most income in group five, which is around West Los Angeles. Through the rental price changes and price prediction, there does exist a difference among these five groups, and we could say that such a difference is related to the local market prices, which means the areas with higher house prices are more likely to be expensive in their renal house prices.

Limitation:

The biggest limitation of this research is choosing the type of rental houses within each group in spite of the house price. As the conclusion mentioned, expensive places in Great Los Angeles such as Beverly Hills and Malibu are more likely to have higher rental price. Even though this research takes the idea of occupancy rate into account, the "number of reviews per month" is just a standard to compare, rather than a variable to calculate the approximate monthly income due to the topic of this research. Therefore, two more processes could be done if we have more time to make this research a real instruction for the investors: First, find the datasets to prove the assumption above: expensive places are more likely to be expensive in their rental price; then use those house prices combined with the rental house prices and locations we found in this research to calculate the rate of return on investment. In this case, we can provide a more accurate and practical blue book for the future investors.

Part Two

This part is designed for the future visitors to find the best neighborhood to live in based on their requirements such as number of people, number of bedrooms and budget. Since the bathroom can be shared and 0 bathroom is removed, these three variables are the most basic standard to find the ideal house demand.

The steps are shown as below:

1. Enter the number of people to live in
2. Enter the number of bedrooms needed
3. Enter the budget per night

Then, it will output the price difference between the price entered and the median market Price. Moreover, it will also output the neighborhood with the rental price closest to the price entered.

Citation

Inside Airbnb- Adding data to debate: http://insideairbnb.com/get-the-data.html

“Los Angeles Tourism Generated a Record $36.6 Billion for the Region's Economy Last Year.” *Los Angeles Times*, Los Angeles Times, 7 May 2019, https://www.latimes.com/business/la-fi-tourism-impact-record-20190507-story.html.

Prabhakaran, Selva. *Loess Regression and Smoothing With R*, http://r-statistics.co/Loess-Regression-With-R.html.