Northeastern University

Final Project: Final Project: Analyzing Factors Affecting Airbnb Listings in Hawaii

By
Nithin Reddy Penta Reddy

ALY 6010: Probability Theory and Introductory Statistics
Under the guidance of Yvonne Leung
Submission Date:12-16-2023

Table of Contents

INTRODUCTION

The dataset chosen for our final project revolves around Airbnb data sourced from Inside Airbnb: Get the Data [Ref. 2]. This platform provides diverse datasets containing information on Airbnb listings from around the globe. For our specific project needs, we sought a dataset rich in data that could potentially reveal insightful patterns. After scouring through various options, we settled on Hawaii as our focus due to its popularity as a holiday destination, hosting a vast array of Airbnb listings. Our selected file, "listings.csv," encompasses a comprehensive summary of information and key metrics for these listings. In this project we mainly focus on the relations between prices of listing and different factors. To provide more suggestions for travellers who share interests in Hawaii we propose three questions of interests and approach the answers via conducting different types of t-tests and different linear regression analysis. The report would follow structure below: Introduction, Data cleaning, Data Exploration, Questions of interests, and Conclusion.

DATA CLEANING

We took a preliminary look at the dataset, we thoroughly examined the dataset's columns, we decided to trim away certain columns that we anticipated wouldn't significantly contribute to our analysis.
As depicted in figure 1, the remaining columns will be our focus moving forward. These specific columns are the ones we'll delve into and analyze for insights, as they offer the most pertinent and valuable information for our project.

As we delved deeper into data cleaning, an intriguing discovery emerged from the 'name' column. Beyond simply housing the Airbnb names, this column also encompassed additional information. Alongside the names, it featured a rating denoted by a character star and a comprehensive description of the listing, detailing specifics such as the number of bedrooms, beds, and bathrooms (Fig. 2). This amalgamation of details within the 'name' column revealed a rich blend of not just names but also qualitative and quantitative attributes associated with each Airbnb listing.

Discovering the ratings embedded within the string-type 'name' column sparked an idea. Recognizing that not all listings included a rating, we opted to designate 'NA' for those listings lacking this information. Our initial approach to extract this data involved leveraging regular expressions, considering the string nature of the *'name'* column. The use of regular expressions was deemed suitable for this scenario, given its effectiveness in parsing and isolating specific patterns within textual data.
The regular expression (Fig. 3) used is designed to extract a specific pattern (Ratings value) from the strings in the name column of the *airbnb* dataset.

- ★ : Matches the star symbol.
- [0-9]+ : Matches one or more digits (0-9).
- \\. : Matches the dot (escaped with a backslash to treat it as a literal dot).
- [0-9]+ : Matches one or more digits (0-9) after the dot.

Eventually, this extracted and processed rating data was consolidated into a new column within the *'airbnb'* dataframe, called *rating*. This comprehensive approach ensured the ratings were standardized as numerical values, permitting seamless utilization and analysis within the dataset.

Data Exploration & Visualizations

Descriptive Statistics

We found the descriptive statistics for the interesting columns in our dataframe as seen in figure 4. It helps us to understand how data is represented in the excel, it gives information about basic statistics such as mean, median, max, mix, var and sd.

Identifying which neighbourhood has more listings and creating a bar plot. Figure 5 shows the top neighbourhoods as a table and figure 6 shows a Bar chart for how the Airbnb listings are distributed based on the number of listings.
Analysing the table and the graph we can clearly observe that Primary Urban Center has most listings followed by Lahaina, Makena, North Kona and North Shore Kauai.

To better apply better approaches, we decided to remove the outlier based on the descriptive statistics. We have done the following:

- Calculating IQR (Interquartile Range)
- Defining Upper and Lower Limits:
- Filtering Outliers

Data visualizations & Explanations

1. The box plot for Neighbourhood Group vs Price (Figure 8) is closely packed and we do not have any outliers falling outside the whisker. The median for all the neighbourhoods is on the lower side. Maui neighbourhood as symmetric distribution in comparison to others.
2. The box plot of ratings column (Figure 9) shows the distribution of ratings that are all above 4.6, with the majority of them being between approximately 4.8 and 5.0. The y-axis is labelled “Rating” and ranges from about 4.4 to 5.0. A blue rectangular box represents the interquartile range (IQR) of ratings, indicating that most ratings fall between roughly 4.8 and 5.0. There are no visible outliers or extreme values in this dataset as per the plot.
3. We choose our neighbourhood as Hawaii and did not include any NA values in it. Figure 10 shows the descriptive statistics for Ratings column.
4. The density plot for Neighbourhood group and Price (figure 11) shows the distribution of prices in different neighbourhoods in Hawaii. From the plot, we can interpret that:
    - Hawaii has a wider range of prices with a peak at the lower end, indicating that there are more options available at a lower price.
    - Honolulu has a narrower distribution and peaks around mid-range prices.

- Kauai has its highest density at the lower price range but also extends to higher prices.
- Maui shows a bimodal distribution with peaks at both lower and higher prices.

Additional feature:

We wrote code that generates an interactive map using Leaflet in R, displaying Airbnb listings' locations. It adds markers on the map for each listing, showing the room type and price when clicked. The markers represent different listings, providing an overview of their geographical distribution.
See figure 12 for a sample output.

Questions of Interest:

Question 1: Do Airbnb listings in specific neighbourhoods differ significantly in their average price?

Hypotheses:
Null Hypothesis (H0): There is no significant difference in the average prices between neighbourhoods North Kona and Primary Urban Centre.
Alternate Hypothesis (H1): There is a significant difference in the average prices between neighbourhoods North Kona and Primary Urban Centre.

Significant level: $\alpha = 0.05$.

In purpose of analysing price differences between specific neighbourhoods, 'North Kona' and 'Primary Urban Centre', within the Airbnb dataset. It executes the following steps:

- Calculating Median Price
- Creating Price-Based Subsets
- Combining Subsets
- Apply two samples t-test:

This test specifically assesses whether there's a significant discrepancy in prices between the neighbourhoods 'North Kona' and 'Primary Urban Centre'. This analysis aims to ascertain if there's a noteworthy variance in Airbnb listing prices between the designated neighbourhoods, providing valuable insights into potential price distinctions across these specific geographical areas.

Interpretation of the results of the t-test (Fig. 13):
The Two Sample t-test was conducted to examine the difference in average prices between two distinct neighbourhoods, namely 'North Kona' and 'Primary Urban Centre,' within the Airbnb dataset. The test yielded a statistically significant result ($t = 27.556$, $df = 6754.8$, $p < 0.0001$), indicating a substantial difference in mean prices between the two neighbourhoods. The 95% confidence interval for the difference in means ranged from 42.544 to 49.061. This interval suggests that, with high confidence, the average price in 'North Kona' (mean = $208.77)

is notably higher than in 'Primary Urban Centre' (mean = $162.97). Therefore, based on this analysis, there appears to be a significant disparity in Airbnb listing prices between these two specific neighbourhoods, with 'North Kona' exhibiting notably higher average prices than 'Primary Urban Centre.
The interpretation of the output aligns with the rejection of the null hypothesis. The obtained p-value ($p < 0.0001$) in the Welch Two Sample t-test provides compelling evidence against the null hypothesis. Hence, we reject the idea that there's no significant difference in average prices between 'North Kona' and 'Primary Urban Centre.'

Instead, the statistical analysis strongly suggests that there is indeed a notable difference in the average prices of Airbnb listings between these two neighbourhoods. The calculated confidence interval (42.544 to 49.061) further solidifies this conclusion, indicating with 95% confidence that 'North Kona' has substantially higher average prices compared to 'Primary Urban Centre.' Therefore, the results support the alternate hypothesis, confirming a significant disparity in Airbnb listing prices between these specific neighbourhoods. Figure 14 shows the box plot for Price for the two neighbourhoods.

Question 2: Is there a significant difference in average price between listings with high and low ratings?

Hypotheses:
Null Hypothesis (H0): The average price of listings with high and low ratings is the same.
Alternate Hypothesis (H1): The average price of listings with high and low ratings differs significantly.

Significant level: $\alpha = 0.05$.

We have written performs several analyses and visualizations; it executes the following steps:

- Separates the combined dataset into two groups.
    a. high rating: listings with ratings greater than or equal to 4.5
    b. low rating: includes listings with ratings equal to or below 3.5.
- Conducts a two-sample t-test
- Regression Analysis
- Visualization

The analysis aims to explore the impact of ratings and neighbourhood on pricing within the Airbnb dataset. The t-test assesses the difference in average prices between high and low-rated listings, while the regression analysis examines how ratings and neighbourhood variables collectively contribute to pricing. Additionally, the scatter plot (Figure 15) visually represents the relationship between ratings and prices, allowing for a better understanding of their potential correlation. A regression line is drawn to show the trend, which appears to be relatively flat, indicating a little weak correlation between the two variables.

Additional Interpretation of High vs. Low Rating Comparison:

Two Sample t-test was conducted to compare the prices between high and low-rated Airbnb listings:

- Test Outcome: The test resulted in a t-value of 2.4461 with a corresponding degree of freedom (df) of approximately 92.172. The calculated p-value is 0.01634, suggesting statistical significance.
- Null Hypothesis Rejection: With a p-value less than the significance level ($\alpha = 0.05$), we reject the null hypothesis. This indicates a statistically significant difference in the average prices between high-rated ('mean of x = 204.85') and low-rated ('mean of y = 183.40') Airbnb listings.
- Confidence Interval: The 95% confidence interval for the difference in means ranges from 4.032 to 38.853. This interval suggests that there's a high likelihood (95% confidence) that the true difference in mean prices between these groups falls within this range.

Regression Analysis

Linear regression model is constructed to analysis the relations between prices of listing and the rating of the Airbnb in different neighbourhood. From the regression table (Figure 18) we can see the intercept is equal to 20.271 which is considered as the basic price of listings, and the slope of the independent vitiable (rating) is equal to 24.349 shows a positive relation between price and rating. The prices of listing would generally increase 24.349 in USD as each additional increase of the rating. The R square is equal to 0.2826, which means 28.36% of the variation of prices of listing is explained by rating and neighbourhood, the adjust-R squire is slightly below.

However, the good performance of the model could because of the large number of independent variables. Maybe we could introduce model selections in the future discussion.

Overall Insights:

The t-test between high and low-rated listings reveals a significant difference in their average prices. This signifies that there's a discernible pricing distinction between listings based on their ratings. Consequently, ratings might indeed play a notable role in influencing the pricing of Airbnb listings, contributing to the understanding of how ratings impact pricing dynamics within the dataset.

Therefore, there is sufficient evidence to reject the null hypothesis (H0). The data suggests a statistically significant difference in average prices between listings with high ratings (mean: 204.85) and low ratings (mean: 183.40).

Question 3: Does shorter distance between properties and centre of Hawaii have any effect on pricing?

Hypotheses:

Null Hypothesis (H0): There is no significant difference in mean pricing between properties within 30 km and those beyond.
Alternate Hypothesis (H1): There is a significant difference in mean pricing between properties within 30 km and those beyond.

Significant level: $\alpha = 0.05$.

To analysis this question we executes the following steps:

- Using Haversine formula to transfer the coordinates information into the distance.
- Subset the data into two groups.
- Conduct a two samples t-test.
- Regression Analysis

This analysis aims to discover the insight of the relationship between prices of listing and its corresponding distances. It puts distance under considerations when traveler choosing their accommodations in Hawaii.

Additional Interpretation of Comparison:

Since two sample t-test was conducted to compare the prices of listing between two different groups, referring to figure 16 more details of the analysis was given as below:

- Test Outcome: The t-value of the test equal to 9.435 with degree of freedom equal to 1548.9. The calculated p-value is less than 0.0001.
- Null Hypothesis Rejection: Since the p-value is far less than the chosen significant level ($\alpha = 0.05$), we reject the null hypothesis and have strong evidence to conclude that there is a significant difference in the prices of listing between properties within 30 km of the centre of Hawaii and those beyond.
- Confidence Interval: The 95% confidence interval for the difference in means ranges from 18.088 to 27.583. This interval suggests that there's a high likelihood (95% confidence) that the true difference in mean prices between these groups falls within this range.

Regression Analysis

Linear regression model is constructed to analysis the relations between prices of listing and the distance away from the center of big island Hawaii. From the result we can see the intercept is equal to 287.26 which is considered as the basic price of listings, and the slope of the independent vitiable (distance) is equal to -1.463 indicates a negative relation between price and distance. The prices of listing would generally drop 1.463 in USD as each additional kilometer away from the center of big island Hawaii. Also, the value of R square demonstrates that around 11.6% of the variation of prices of listing is explained by the independent variable (distance).

Overall Insights:

This question brings the comparison of prices of listing between two groups with different distances from the center of big island of Hawaii, which allows travellers to make better decisions on their trip to Hawaii. The two-sample t test compares the difference between price and lead to a conclusion that distance indeed a factor affect the prices of listing. The behaviour of differences between pieces of listing implies the well-development of the center area of big island of Hawaii. Also, it somehow shows the popular areas that travellers tend to be and it is clears that compare with other three islands, big island of Hawaii is more preferred.

CONCLUSION

In this project we focus on the data of Airbnb in Hawaii and rise analysis in prices of listing for travellers who shares interests in Hawaii. We have followed the structures of exploratory data analysis. We focus on the relations between prices and different factors, for example, different neighborhood, ratings, and distances to the center of big island of Hawaii. Different challenges come with the analysis. Data extraction as example, to process the analysis of rating we spend large amount of time to extract the rating information from the description of each Airbnb.

Three investigations are rase and focus on the influences on price from neighborhoods, rating of Airbnb, and distances of Airbnb to the center of big island Hawaii. Based on the results of the different types of t-tests, the main takes away is that comparing with the North Kona area the primary urban area has a higher price of listing, which provides valuable insights into potential price distinctions across these specific geographical areas. Also, second investigation reflects how the price of listing is affected by people's comments. The last investigation has moved further into the location information. Comparing with the first investigation, we narrowed down the physical location from neighborhood into circle area of 30 km from the center of big island Hawaii and concluded that there is a significant difference in price of listing between different distances ranges.

Moreover, regression analysis is employed to three investigations. We first consider the price of the listing as the dependent variable and the ratings; neighborhood are the independent variables. About 28% variation of the price are explained by the independent variables. Second regression model is constructed to analysis the relationship between price of listing and distance of Airbnb from the center of big island Hawaii. The regression model shows a negative relation between price and distance which answering the question that the shorter the distance from the center of big island of Hawaii the higher the price of listing would have.

Lastly, this project mainly focusses on the relations between prices of listing and different factors. Different investigation has been approached via conducting different types of t-tests and linear regression to provide information for traveller who considers Hawaii as their destination from three different aspects.

REFERENCES

1. *Mike Yi: A Complete Guide to Box Plots; [online]:* https://chartio.com/learn/charts/box-plot-complete-guide/
2. *Github: Making Maps in R; [online]:* https://cengel.github.io/R-spatial/mapping.html
3. *Dataworkie (8, October, 2022): Scaling Density Plots;* [online]: https://datawookie.dev/blog/2022/10/scaling-density-plots/
4. *Statistic Solutions;* [online]: https://www.statisticssolutions.com/independent-sample-t-test/
5. *Leung, Y. (2023, December 10). Aly 6010 module 1-5. ALY 6010, Slide:104;* [online]:https://docs.google.com/presentation/d/1qX9-qxRoRI87VxBYHvN2Lv_fkXD4LF3a4aJ9ULQP8_8/edit
6. *BY ZACH (2020, December, 23): How to Plot Multiple Linear Regression Results in R;* [online]: https://www.statology.org/plot-multiple-linear-regression-in-r/
7. *Geeks Community: Multiple linear regression using ggplot2 in R;* [online]: https://www.geeksforgeeks.org/multiple-linear-regression-using-ggplot2-in-r/
8. *Keon-Woong Moon (2020, October,6): Visualize multiple regression model;* [online]: https://cran.r-project.org/web/packages/ggiraphExtra/vignettes/ggPredict.html
9. *Rpubs: Multiple Linear Regression:* [online]: https://rpubs.com/bensonsyd/385183
10. *Yvonne Leung, PhD (30, October 2023): ALY 6010 Module 1-6, Slide 153- Slide 156; [online]:* https://docs.google.com/presentation/d/1qX9-qxRoRI87VxBYHvN2Lv_fkXD4LF3a4aJ9ULQP8_8/edit#slide=id.g2a5b5e3d203_0_51

APPENDIX

```
> colnames(airbnb)
 [1] "name"                "neighbourhood_group" "neighbourhood"       "latitude"
 [5] "longitude"           "room_type"           "price"               "minimum_nights"
 [9] "number_of_reviews"   "availability_365"    "rating"
```

Figure 1: Column names from the main data frame

	name
1	Resort in Lahaina · 2 bedrooms · 3 beds · 2 baths
2	Rental unit in Honolulu · ★4.57 · 1 bedroom · 1 bed · 1 share...
3	Condo in Honolulu · ★4.72 · 1 bedroom · 1 bed · 1 bath
4	Condo in Honolulu · ★4.71 · Studio · 1 bed · 1 bath
5	Rental unit in Kamuela · ★4.67 · 1 bedroom · 1 bed · 1 bath
6	Condo in Princeville · ★4.54 · 1 bedroom · 1 bed · 1 bath
7	Rental unit in Captain Cook · ★4.65 · Studio · 2 beds · 1 bath

Figure 2: Data in the name column

```
#The name column contains the name of the Airbnb along with a rating
# Extract ratings using a regular expression
ratings <- str_extract(airbnb$name, "★[0-9]+\\.[0-9]+")

# Convert the extracted ratings to numeric
ratings_numeric <- as.numeric(sub("★", "", ratings))

# Replace rows with no valid rating with NA
ratings_numeric[is.na(ratings_numeric)] <- NA

# Add the ratings to your data frame as a new column
airbnb$rating <- ratings_numeric
```

Figure 3: Code for extracting ratings using regular expressions

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Rating	1	21099	4.77696	0.2578115	4.85	4.821639	0.177912	0	5	5	-3.070009	22.61001	0.00177489
Price	1	33339	410.93026	1133.5688822	245.00	284.765081	155.673000	0	99110	99110	31.915360	1953.28051	6.20828479

Figure 4 Descriptive statistics for columns 'Rating' and 'Price'

	neighbourhood	listing_count
1	Primary Urban Center	6830
2	Lahaina	4807
3	Kihei-Makena	4572
4	North Kona	3928
5	North Shore Kauai	2185

Figure 5: Top neighbourhoods as a table

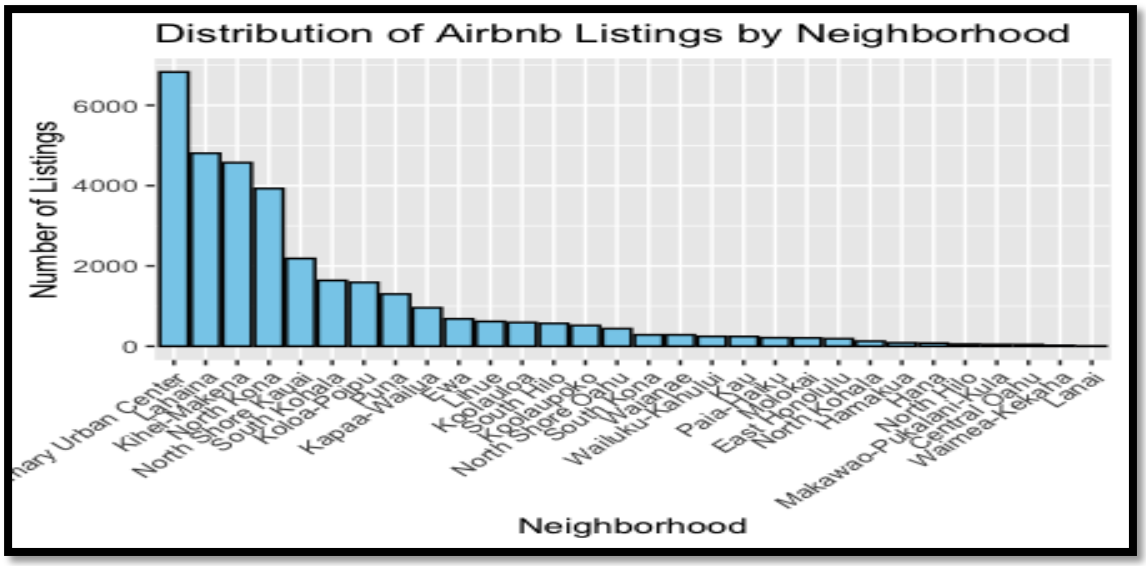

Figure 6: Bar chart for distribution of neighbourhoods based on their listings count

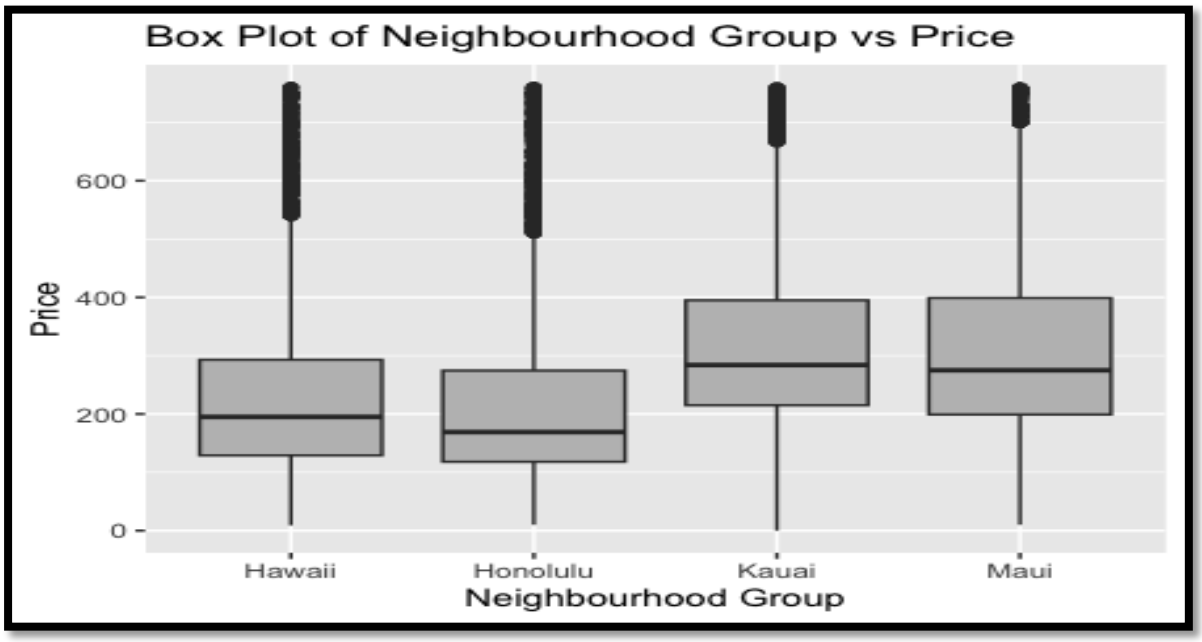

Figure 7: Box plot – Neighbourhood group vs Price

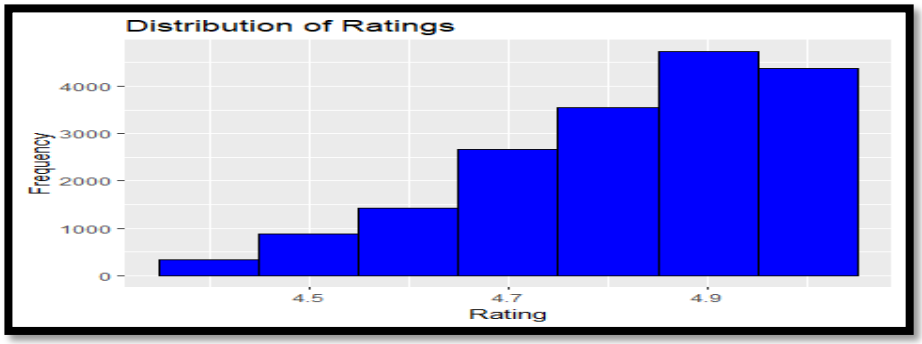

Figure 8 Histogram for distribution of Ratings

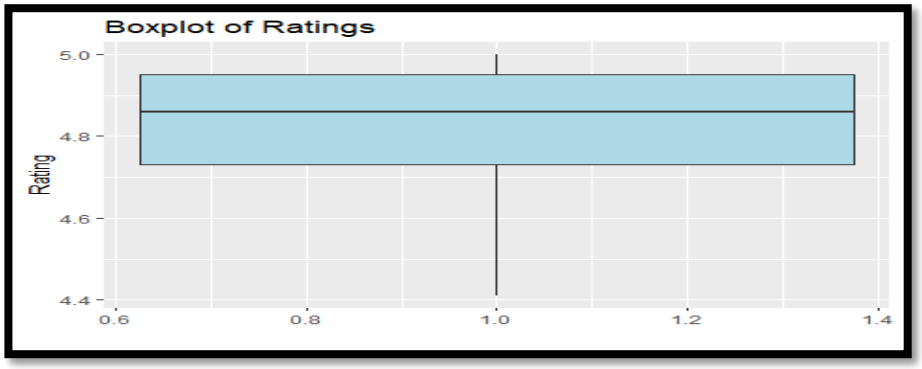

Figure 9 Box plot for Ratings

```
> airbnb_neigh_group <- airbnb[airbnb$neighbourhood_group == "Hawaii" & !is.na(airbnb$rating),
]
> summary(airbnb_neigh_group$rating)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   4.730   4.880   4.805   4.970   5.000
> mean(airbnb_neigh_group$rating)
[1] 4.805285
> median(airbnb_neigh_group$rating)
[1] 4.88
> sd(airbnb_neigh_group$rating)
[1] 0.2419333
> quantile(airbnb_neigh_group$rating)
  0%  25%  50%  75% 100%
2.00 4.73 4.88 4.97 5.00
```

Figure 10 Descriptive statistics for Ratings column after cleaning the same

Figure 11 Density plot for Price and Neighbourhood group、

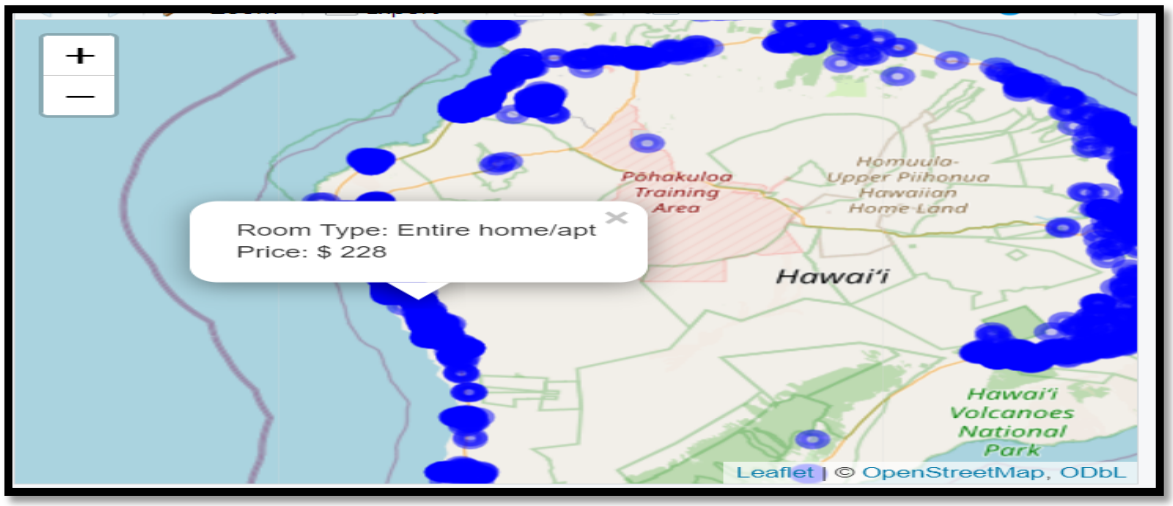

Figure 12 Output for leaflet map feature

```
        Welch Two Sample t-test

data:  price by neighbourhood
t = 27.556, df = 6754.8, p-value < 0.00000000000000022
alternative hypothesis: true difference in means between group North Kona and group Primary Ur
ban Center is not equal to 0
95 percent confidence interval:
 42.54450 49.06134
sample estimates:
          mean in group North Kona mean in group Primary Urban Center
                          208.7698                            162.9669
```

Figure 13 Output for t-test for question 1

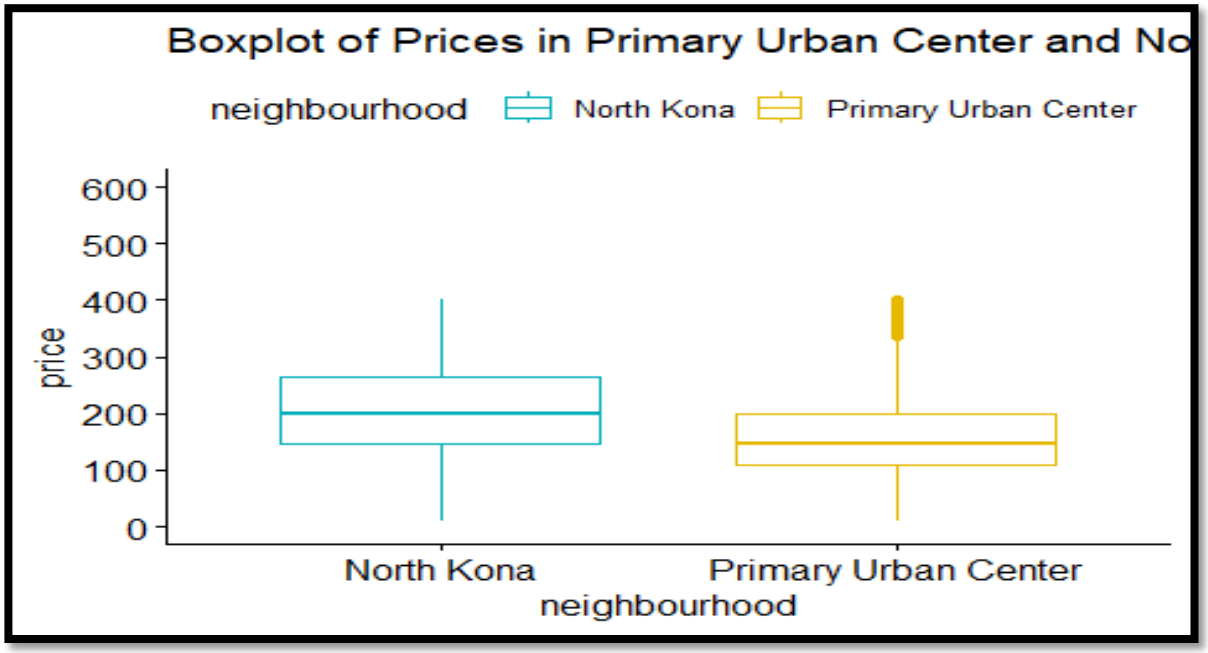

Figure 14 Box plot for visualizing the output of the t-test

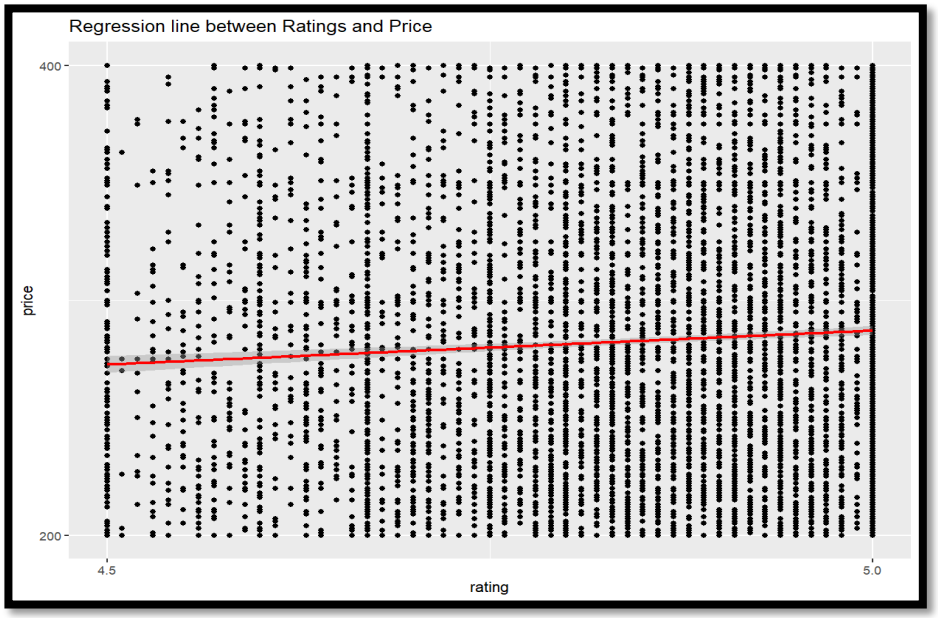

Figure 15 Scatter plot for question 2

```
        Welch Two Sample t-test

data:  subset_within_30km$price and subset_beyond_30km$price
t = 9.4346, df = 1548.8, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 18.08806 27.58341
sample estimates:
mean of x mean of y
 229.6838  206.8481
```

Figure 16 t-test result for question 3

```
Call:
lm(formula = price ~ distance, data = subset_data)

Residuals:
    Min      1Q  Median      3Q     Max
-230.26  -81.44  -24.50   54.55  443.59

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 287.26121    2.42148  118.63 <0.0000000000000002 ***
distance     -1.46271    0.04909  -29.79 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 112.3 on 6784 degrees of freedom
Multiple R-squared:  0.1157,    Adjusted R-squared:  0.1156
F-statistic: 887.7 on 1 and 6784 DF,  p-value: < 0.00000000000000022
```

Figure 17 Regression table for relation between price and distance

```
Call:
lm(formula = price ~ rating + neighbourhood, data = combined_subset)

Residuals:
    Min      1Q  Median      3Q     Max
-243.63  -53.87  -11.31   46.23  265.37

Coefficients:
                                Estimate Std. Error t value             Pr(>|t|)
(Intercept)                       20.271     18.124   1.118              0.263386
rating                            24.349      2.171  11.218 < 0.0000000000000002 ***
neighbourhoodEast Honolulu        34.556     17.232   2.005              0.044935 *
neighbourhoodEwa                  86.759     15.885   5.462  0.00000004785377579 ***
neighbourhoodHamakua               9.375     17.695   0.530              0.596222
neighbourhoodHana                134.992     17.891   7.545  0.00000000000004734 ***
neighbourhoodKapaa-Wailua        101.283     15.209   6.659  0.00000000002831537 ***
neighbourhoodKau                  -7.901     15.819  -0.499              0.617443
neighbourhoodKihei-Makena         97.049     14.860   6.531  0.00000000006720246 ***
neighbourhoodKoloa-Poipu         154.178     15.064  10.235 < 0.0000000000000002 ***
neighbourhoodKoolauloa           120.559     15.321   7.869  0.00000000000000379 ***
neighbourhoodKoolaupoko           51.033     15.540   3.284              0.001025 **
neighbourhoodLahaina             117.682     14.896   7.900  0.00000000000000294 ***
neighbourhoodLanai               113.058     39.828   2.839              0.004536 **
neighbourhoodLihue               106.245     15.708   6.764  0.00000000001385704 ***
neighbourhoodMakawao-Pukalani-Kula 108.946   20.529   5.307  0.00000011279195094 ***
neighbourhoodMolokai              -2.567     15.903  -0.161              0.871764
neighbourhoodNorth Hilo           24.179     19.147   1.263              0.206692
neighbourhoodNorth Kohala         14.923     16.971   0.879              0.379248
neighbourhoodNorth Kona           54.499     14.881   3.662              0.000251 ***
neighbourhoodNorth Shore Kauai   115.571     14.981   7.715  0.00000000000001280 ***
neighbourhoodNorth Shore Oahu     68.297     15.571   4.386  0.00001160922318194 ***
neighbourhoodPaia-Haiku           92.993     15.992   5.815  0.00000000617286427 ***
neighbourhoodPrimary Urban Center 21.754     14.839   1.466              0.142663
neighbourhoodPuna                 -4.351     14.972  -0.291              0.771336
neighbourhoodSouth Hilo           11.360     15.176   0.749              0.454132
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 73.96 on 17477 degrees of freedom
  (7501 observations deleted due to missingness)
Multiple R-squared:  0.2826,    Adjusted R-squared:  0.2814
F-statistic: 229.5 on 30 and 17477 DF,  p-value: < 0.00000000000000022
```

Figure 18 Regression Table for Relation between Price and Rating