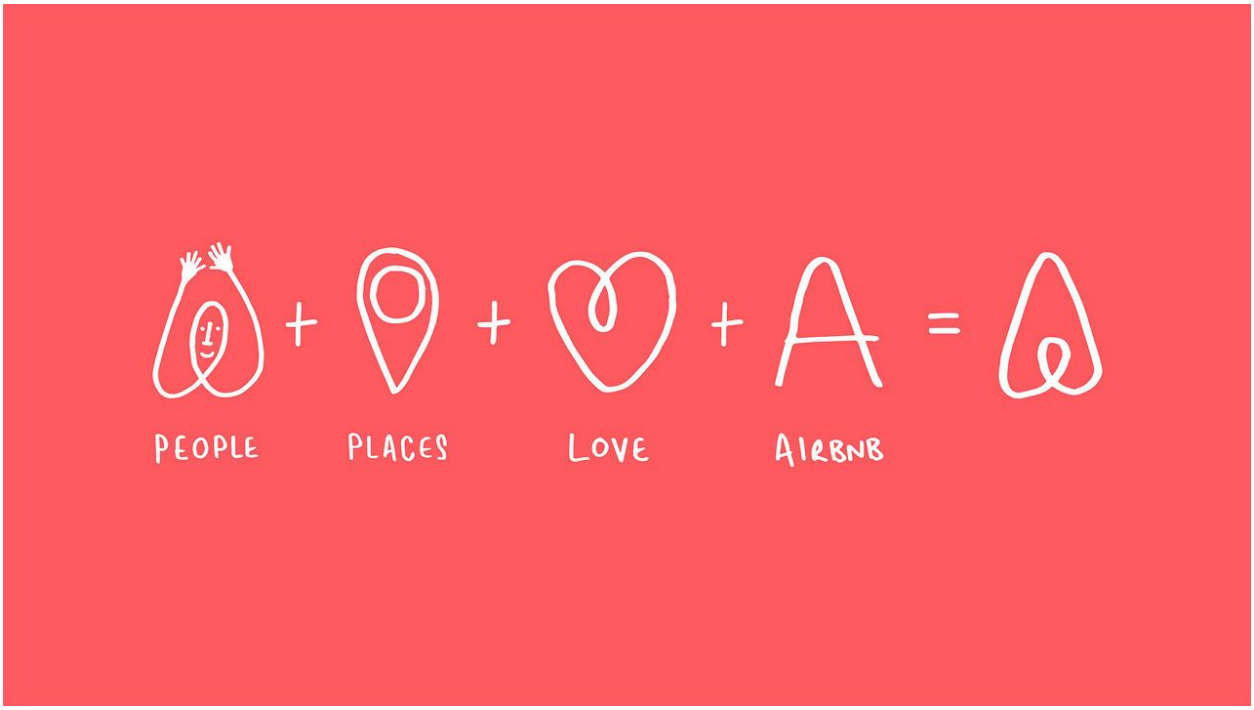
How can I maximize my Airbnb listing earnings in San Francisco?

Exploratory Data Analysis and Predictive Modeling



By: Jessie Huang

Springboard Foundations of Data Science

Mentor: Shmuel Naaman

September 12, 2016

Table of Contents

Introduction

Motivation

There are many lodging options for travelers, such as hotels, bed & breakfast, hostels. Airbnb is a popular bed & breakfast option where homeowners (Hosts) can rent out their entire property or certain rooms (Listing) for travelers. With the increasing popularity of Airbnb coupled with dramatically increased rent prices in San Francisco, I'd like to better understand how San Francisco hosts use Airbnb and what hosts can do to maximize their listing price without sacrificing booking potential. After this analysis, it would be interesting to compare the findings to understand the legal battles between Airbnb and the City of San Francisco. More details on the legal battle are available here.

Objective

The study's goal is to analyze what features affect the listing prices in San Francisco and predict a listing's maximum pricing potential by selecting meaningful features from the dataset.

Data Source

The data source is from Inside Airbnb, a website that compiles the available Airbnb data at the time of scraping by Murray Cox. Inside Airbnb is not associated or affiliated with Airbnb or its competitors.

I will be using the San Francisco listings.csv.gz file, which includes detailed listings data for the active listings as of July 2, 2016.

Deep Dive into Dataset

Overview of Data

In taking a look at the summary of the dataset, the information can be categorized into the following areas:

1. **Subjective summary of Listing** by the Host, such as:
    a. Listing name
    b. Listing summary
    c. Summary of the listing space
    d. Description of the listing
    e. Neighborhood description
    f. Host notes
    g. Transit options
2. **Objective Host information** determined by Airbnb's criteria, such as:
    a. Host id
    b. Date when they became a Host
    c. Host response rate & response time
    d. Total number of listings the Host has
    e. Host verifications
3. **Objective and Subjective Listing information:**
    a. Address
    b. Neighborhood
    c. Property type (house, apartment, condominium, boat, etc.) selected by the Host
    d. Room type (entire space, private room, or shared room)
    e. Number of rooms, bathrooms, and beds
4. **Cost of Listing**
    a. Cost per night, week, and month
    b. Cleaning fees
    c. Security deposit
    d. Cost of extra guests
    e. Minimum and maximum nights of stay
5. **Reviews**
    a. First and most recent dates of reviews
    b. Average review score
    c. Number of review per month

Data Wrangling

The dataset is relatively clean because it was compiled through regular scrapings of the Airbnb website.

However, there are many features provided that aren't relevant for the study, such as whether the host has a profile picture, redundant features like where the jurisdiction of the Listing, the exact coordinates of the Listing by latitude and longitude, the host and listing urls (I can easily obtain this using the listing and host id instead), the time scraped and scrape id, to name a few. These items were removed to help decrease the size of the dataset.

There were also some outliers prices of $1 and $10,000 per night. Some of these turned out to be test accounts by Airbnb interns or spam accounts, so these observations were removed from the dataset. An interesting learning from this exercise was that even though these were outliers, removing them did not change the overall data or plots very much. This is because the number of observations that were removed were too few to influence the overall data.

The dataset also includes a number of repetitive columns due to the information that is available on the Airbnb website but also with how Inside Airbnb has scraped and cleaned the data already. By using the one-way analysis of variance (ANOVA) test, some features were removed due to lack of statistical significance.

Limitations of the Dataset

Because booking data is not available, it's not possible to plot booking rates over a calendar year to understand the demand. Therefore, there will not be an analysis of how different events in San Francisco influence the listing price.

Exploring and Analyzing the Dataset

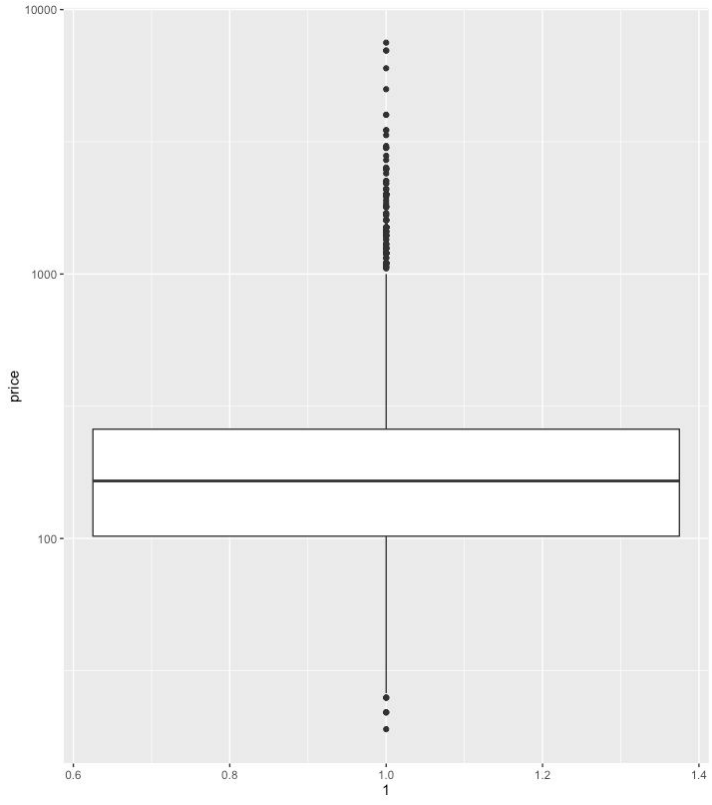
Some of the predictable significant features to determining the listing price were confirmed using the ANOVA test:

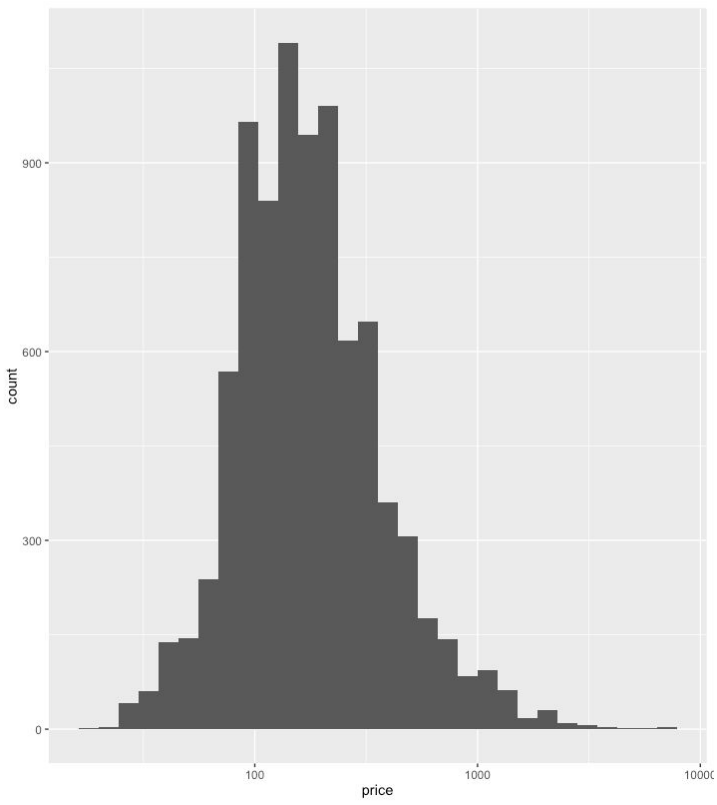
- Location of the listing
- Type of room (entire apartment, private room, shared room)
- Type of property (condominium, house, apartment, loft, etc)
- Accommodation size (the number of people that the listing allows for)
- Listing ratings and reviews
- Host identity verified

After data wrangling and decreasing the dataset, the initial plots showed that:

Distribution of Price

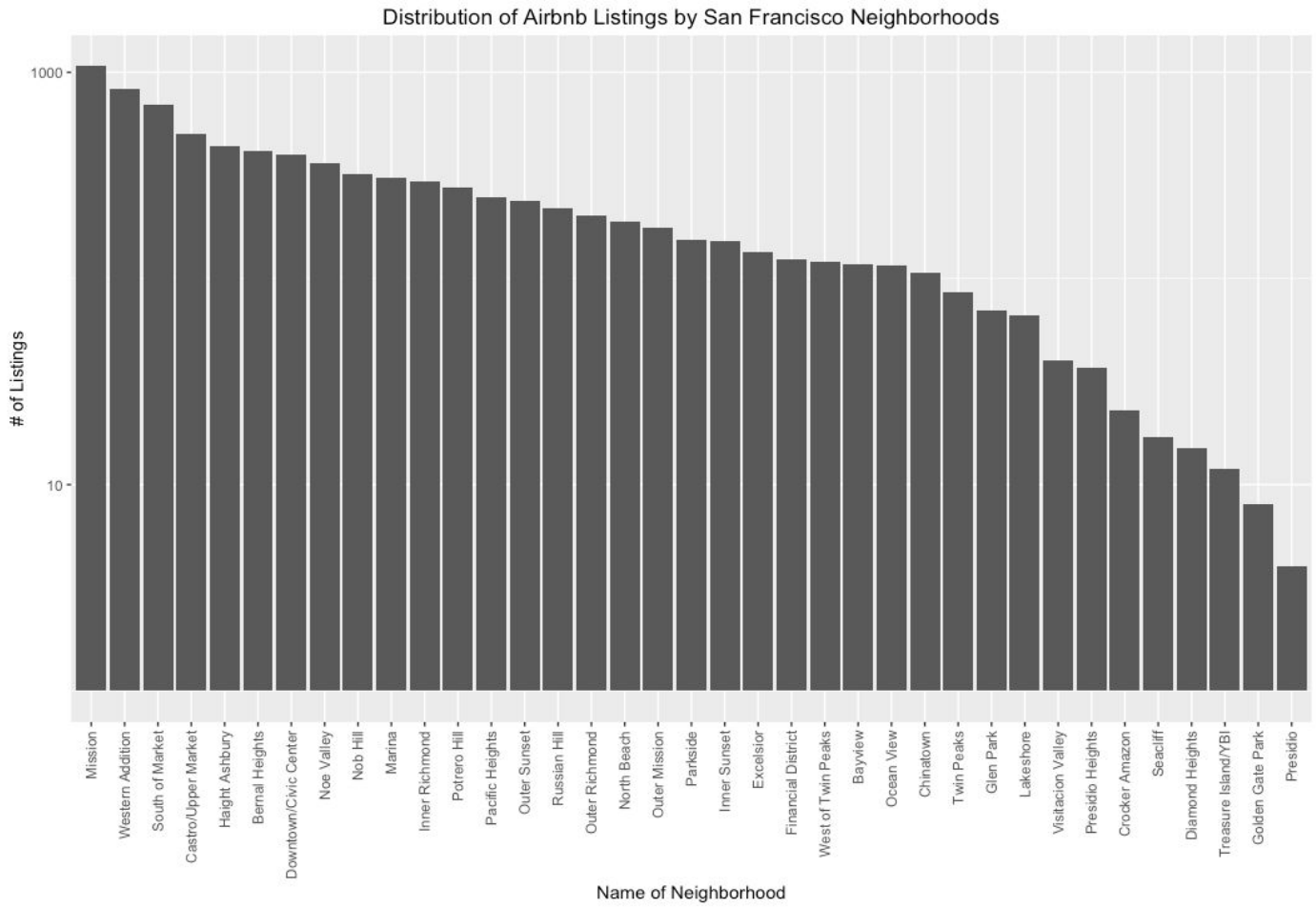
- 25% to 75% of the listings have a listing price of $103 to $260 per night
- The mean price is $245
- The median price is $165





Price versus neighborhood

There are 37 total neighborhoods in San Francisco, with more than 35% of listings are in the Mission, Western Addition, South of Market (SoMa), and Castro neighborhoods.



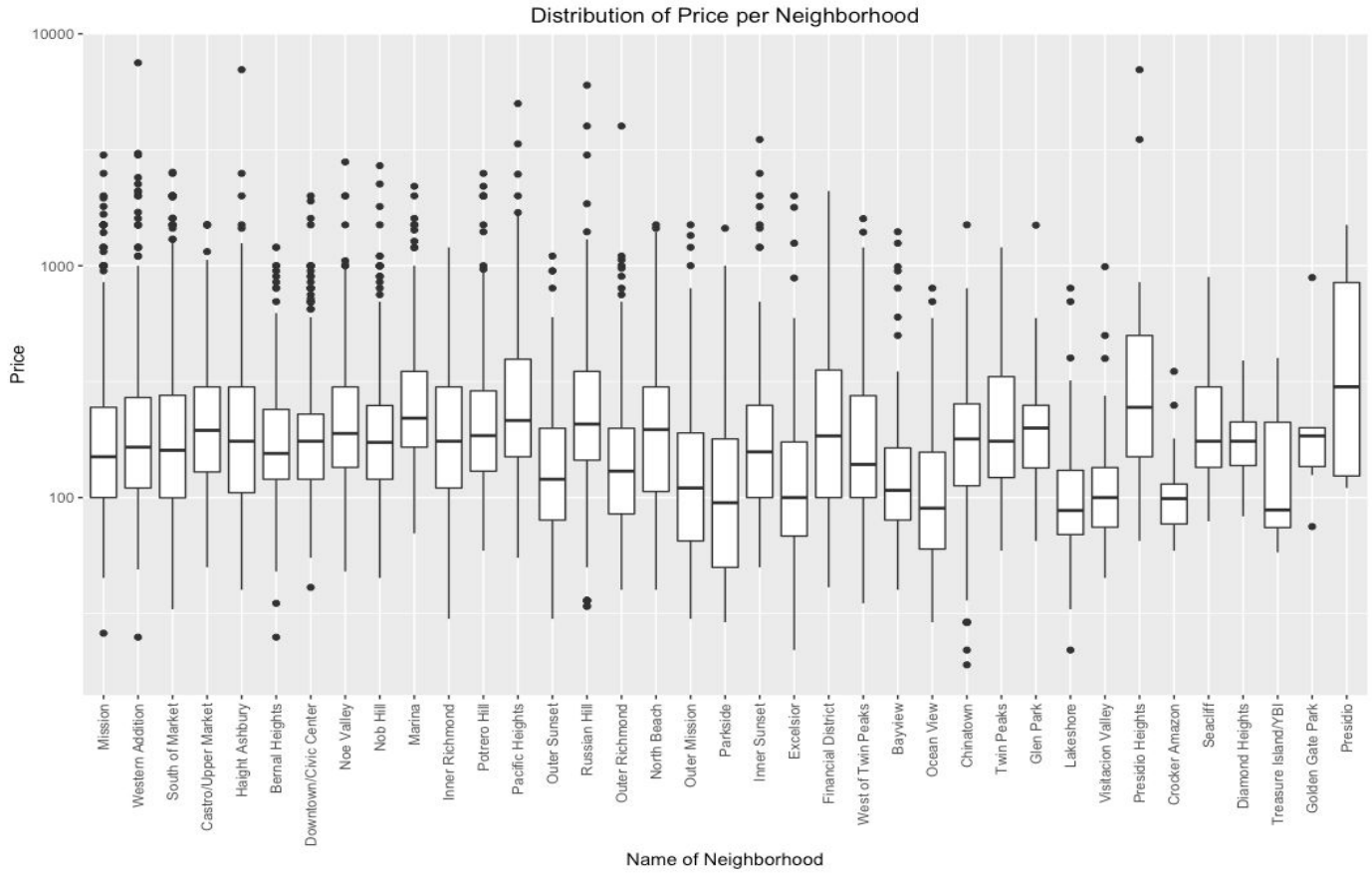
Please note: Because the majority of the plots were not a normal distribution, the log scale is used to allow for a more normal distribution.

After converting the price to a log scale and applying the ANOVA test, the average price among the different neighborhoods is significantly different for a p-value of <0.05. There appears to be a level of significance in influencing the listing price for 28 of the 37 neighborhoods, except for the neighborhoods below:

1. Crocker Amazon
2. Diamond Heights
3. Excelsior

4. Golden Gate Park
5. Outer Mission
6. Outer Richmond
7. Outer Sunset
8. Treasure Island/YBI
9. Visitacion Valley

While it can be assumed that there may not enough data (and possibly a high variation) to determine the significance level for Crocker Amazon, Diamond Heights, Treasure Island, Golden Gate Park, and Visitacion Valley, the remaining neighborhoods have a very high number of listings, but are not considered significant to affect the listing price. The Presidio neighborhood is an interesting result to explore because it is considered statistically significant even though it has the least amount of data and has the largest range in price.

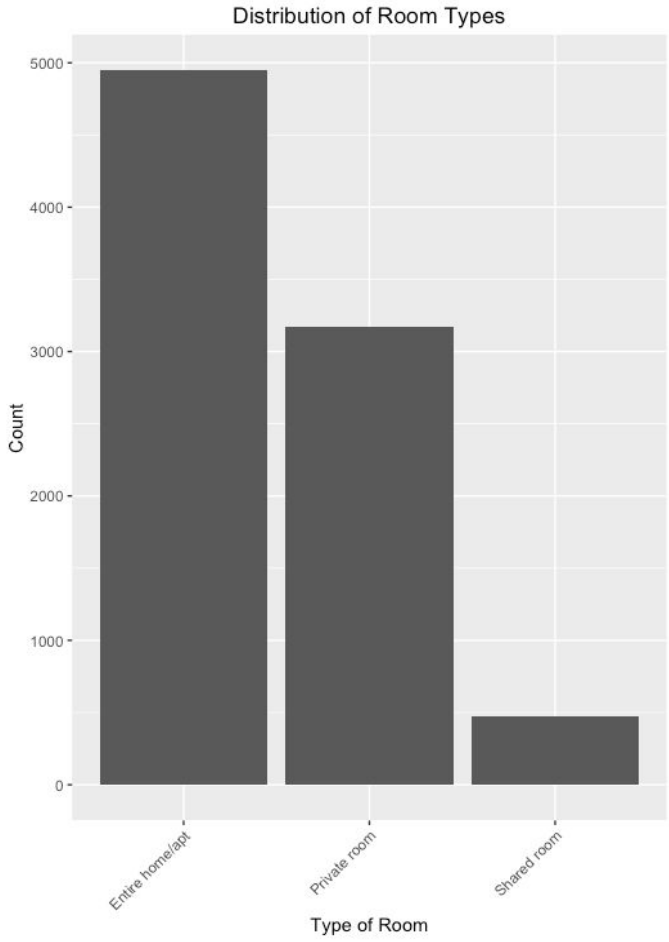


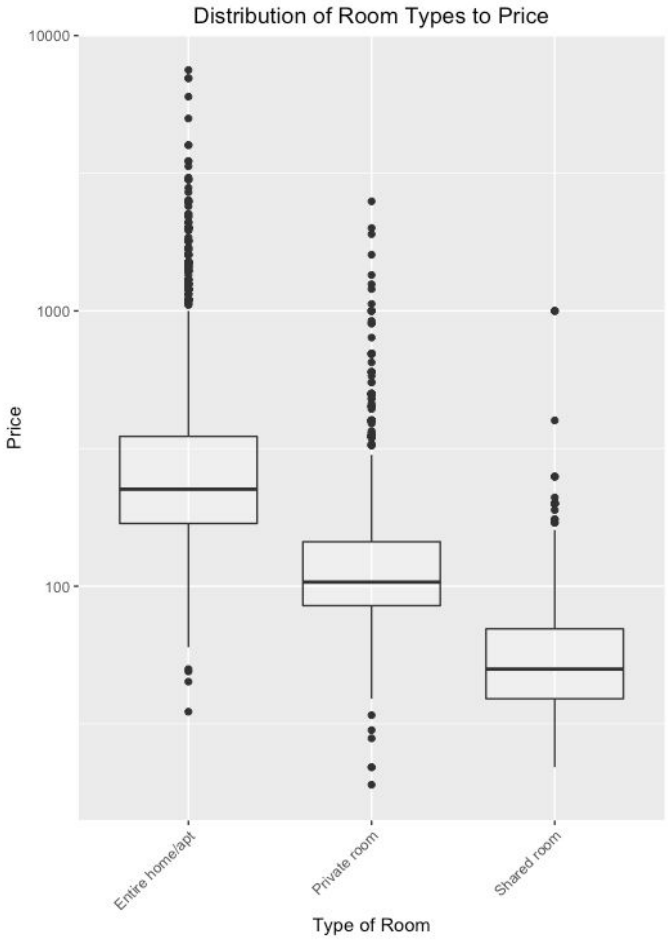
Type of Room versus Price

In taking a further look at the type of room (room_type), the three types of rooms (Entire Home, Private Room, and Shared Room) are statistically significant in influencing the listing price. 58%

of listings are for an Entire Home, compared to the 37% of listings for a Private Room and a mere 5% of listings for a Shared Room.

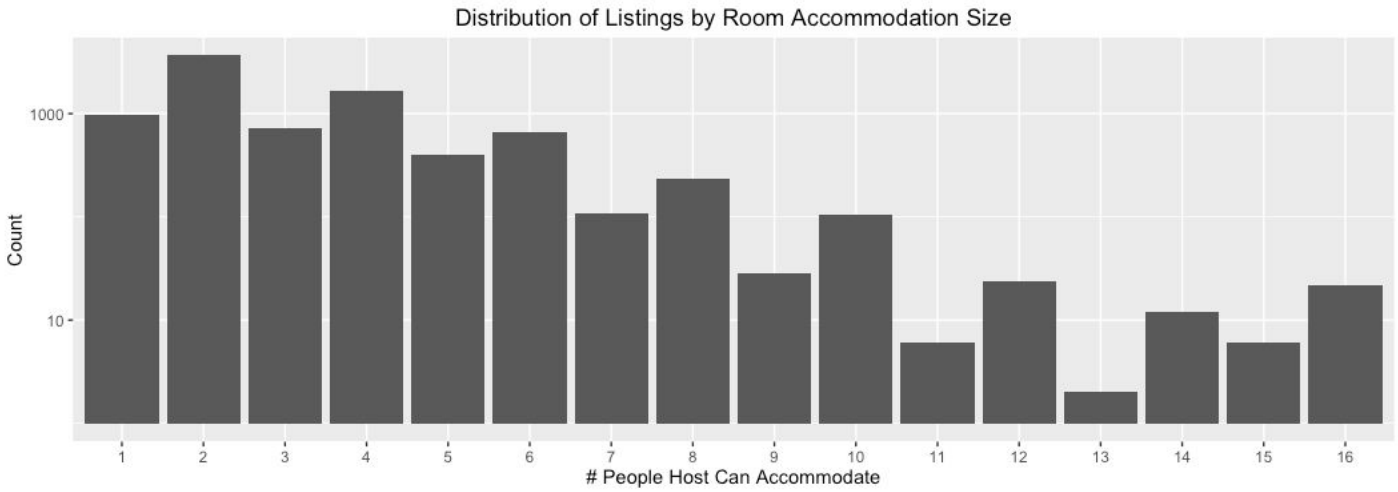
After converting the price to a log scale and applying the ANOVA test, all three room types are significantly significant in affecting the listing price for a p-value of <0.01. As expected, the larger the space that is available for booking, the higher the price the host can list for:

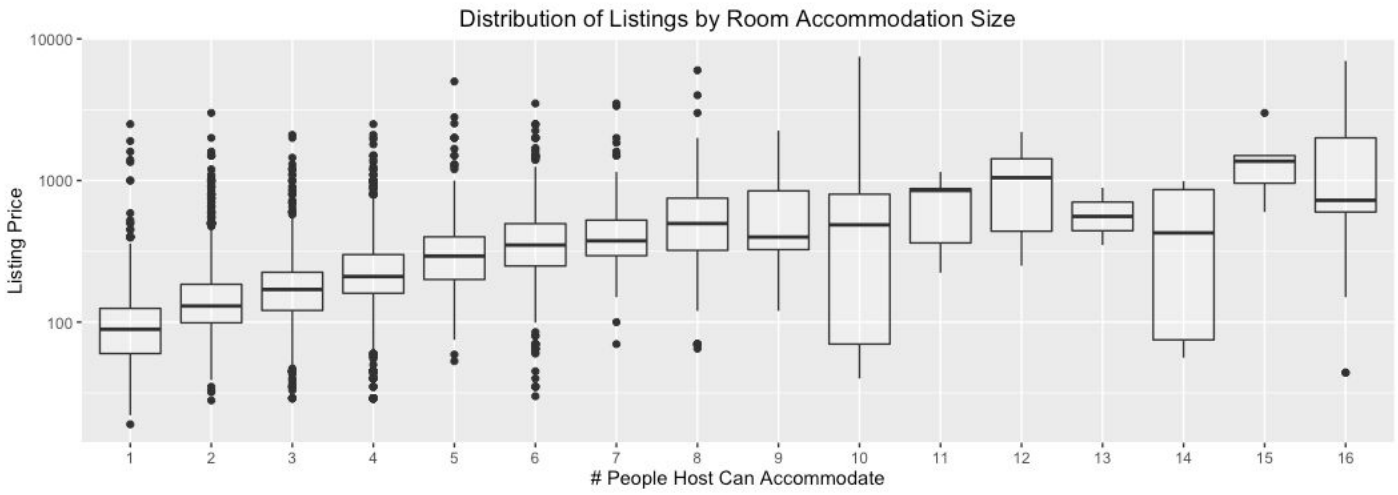




Accommodation Size versus Price

When taking an initial look at how the listing accommodation size influences the price, the price steadily increases from 1 to 8 individuals. 97% of listings have an accommodation size of 1, 2, 3, 4, 5, and 6 people. The two largest accommodation sizes are listings for two individuals (43%) and three individuals (84%).

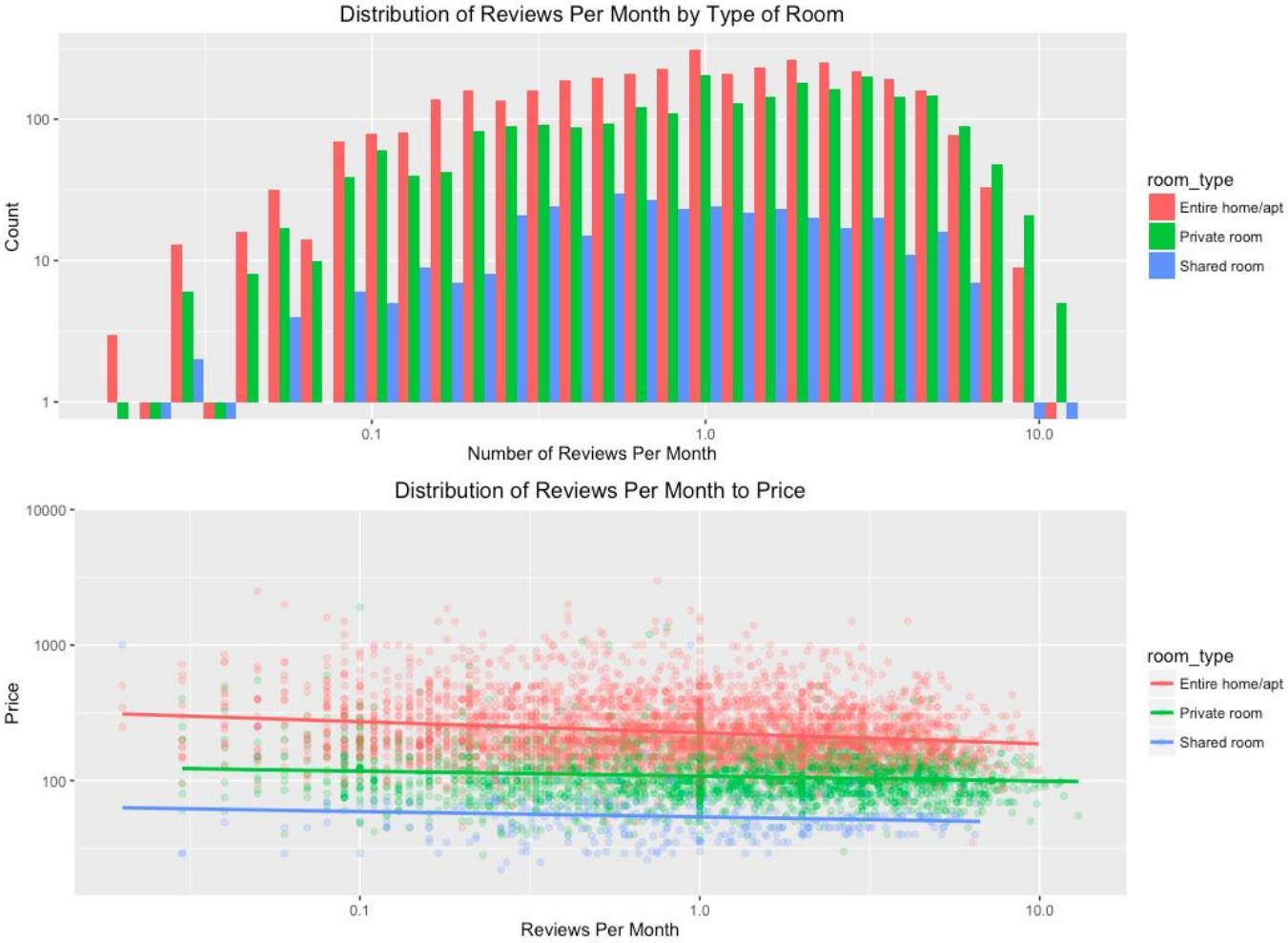




However, after performing the ANOVA statistical test, it appears that all accommodation sizes are statistically significant in influencing the listing price, despite the smaller dataset for some of the sizes (such as accommodation sizes from 11 to 16 people).

Number of Reviews versus Price

The total number of reviews that a listing has received and the average number of reviews per month the listing receives also tells an insightful story. The more reviews a listing receives, the lower the listing price:



Using the Pearson correlation coefficient, the correlation is -0.1547419 with a 95% confidence interval. This means that there is a small but significant negative correlation between the reviews per month and the listing price. Receiving more reviews for a listing actually decreases the amount the host can price their listing at.

Some unexpected features that have statistical significance to influence the price using the ANOVA test are:

- Host behavior, which includes: response time, response rate, acceptance rate
- The minimum number of nights a guest has to book
- Whether the listing can be instantly booked or not
- The total number of Airbnb listings the host has
- How strict the cancellation policy is for the listing

Next Steps

The next steps are to understand how the property type, accommodation size, and the unexpected features that have statistical significance influence the listing price. Once this analysis is done, a predictive model will be created and used to predict the listing price based on the characteristics of a listing.

After analyzing the price, it would also be interesting to explore what it takes to get a listing booked. I will perform an analysis of how hosts are naming their listing and compare the listing price and reviews per month.