Airbnb-To understand how the different features of a listing in LA can influence the price

Oindrila Sen

November,20, 2017

Table of Contents

Prologue 1

Problem Setup 2

Potential client 2

The Data 2

Data Wrangling and Cleaning 8

Adding New Features 18

Exploratory Data Analysis 18

Create a Model 37

Conclusion 44

Future Scope 46

# Prologue

*"Airbnb happened because two guys could not pay their rent, but did have some space."* Yes, that's how Air-bed and breakfast(Airbnb) started when Joe and Martin from San Francisco rented out their three airbeds to make some cash.Since then, guests and hosts have used Airbnb to travel in a more unique and personalized way.The story of Airbnb's foundation is fascinating and is here: [airbnb](http://www.airbnb.com/home/story). Airbnb is a hospitality service which enables people to lease or rent a short-term accommodation.The company does not own any lodging. It is merely an agent.It receives a percentage service fees (commissions) from both guests and hosts for every booking.  
 A Host can create a listing by selecting the "Host" menu after logging in.The Price for the listing is decided by the host himself. The owner can charge different prices for nightly, weekly, and monthly stays. Hosts then add a description of the residence, amenities, available dates, cancellation policies, and any house rules.Potential guests are required to message the host directly through Airbnb to ask questions regarding the property.After the reservation, the hosts coordinate meeting times and contact information with the guests. Airbnb guests can leave a review and rate a listing after their stay.This rating can be used as an indicator of the experience they had during their stay. Airbnb was started small and soon the concept turned out to be popular among those people who were on the verge of losing their homes due to financial difficulties and also the travelers who are looking for a home away from home. [Wiki](https://en.wikipedia.org/wiki/Airbnb) is always there for more information.

# Problem Setup

In this project, I am working on the Los Angeles based Airbnb listings.  
My objective is to understand how the different features can influence the price for a listing in LA.  
I would provide a tool to estimate the price of a listing in LA based on its features.  
 Thus, to summarize, this Capstone Project will answer the below questions:

1. How the different attributes/features can influence the price for a Airbnb listing in LA?  
2. What price would be suitable for a new listing in LA?  
3. How the price of a listing is affecting the Review Rating?

# Potential client

1. A new host who is looking for some recommendation for the price he should list for his property
2. Or, Someone who is willing to list his property in Airbnb site in Los Angeles and gathering information on what are the factors that would affect the price for a listing
3. Or, an existing host who already has an Airbnb listing in LA and looking for a revision in price to get more customers and better reviews.

# The Data

I am going to use the datasets loaded in [insideairbnb](http://insideairbnb.com/get-the-data.html) for Los Angeles, California. The latest data sets are added on May,2017. The below datasets are included:

1. Detail\_Listings.csv  
2. Summary\_Listings.csv  
3. Detail\_Reviews.csv  
4. Summary\_Reviews.csv  
5. Calendar.csv  
6. Neighbourhoods.csv  
 For this project, I will use the "Detail\_Listings.csv" file alone.This Data file has 31253 records with 1 record for each individual listing.There are 95 columns representing different features for each listing.

Below is a glimpse of the "Detail\_listings.csv" data file:

glimpse(Detail\_listing)

## Observations: 31,253  
## Variables: 95  
## $ id <int> 16228948, 8909380, 14078522, ...  
## $ listing\_url <chr> "https://www.airbnb.com/rooms...  
## $ scrape\_id <dbl> 2.01705e+13, 2.01705e+13, 2.0...  
## $ last\_scraped <chr> "2017-05-03", "2017-05-03", "...  
## $ name <chr> "Robert Ranch in Los Angeles ...  
## $ summary <chr> "Stunning Ranch Estate situat...  
## $ space <chr> NA, "Very private and quite ....  
## $ description <chr> "Stunning Ranch Estate situat...  
## $ experiences\_offered <chr> "none", "none", "none", "none...  
## $ neighborhood\_overview <chr> NA, "I'm in the country what'...  
## $ notes <chr> NA, NA, "Overnight stalls for...  
## $ transit <chr> NA, "The train station is onl...  
## $ access <chr> NA, "Whole house", "Private B...  
## $ interaction <chr> NA, "As much or as little as ...  
## $ house\_rules <chr> "- No noise/music outside all...  
## $ thumbnail\_url <chr> NA, "https://a0.muscache.com/...  
## $ medium\_url <chr> NA, "https://a0.muscache.com/...  
## $ picture\_url <chr> "https://a0.muscache.com/im/p...  
## $ xl\_picture\_url <chr> NA, "https://a0.muscache.com/...  
## $ host\_id <int> 54345288, 44800067, 84615808,...  
## $ host\_url <chr> "https://www.airbnb.com/users...  
## $ host\_name <chr> "Robert", "Kevin", "Kirsten",...  
## $ host\_since <chr> "2016-01-12", "2015-09-22", "...  
## $ host\_location <chr> "Los Angeles, California, Uni...  
## $ host\_about <chr> "Thank You for considering my...  
## $ host\_response\_time <chr> "a few days or more", "within...  
## $ host\_response\_rate <chr> "25%", "100%", "100%", "100%"...  
## $ host\_acceptance\_rate <chr> "N/A", "N/A", "N/A", "N/A", "...  
## $ host\_is\_superhost <chr> "f", "t", "t", "t", "f", "f",...  
## $ host\_thumbnail\_url <chr> "https://a0.muscache.com/im/p...  
## $ host\_picture\_url <chr> "https://a0.muscache.com/im/p...  
## $ host\_neighbourhood <chr> NA, NA, NA, NA, "West Adams",...  
## $ host\_listings\_count <int> 1, 1, 1, 1, 1, 1, 2, 23, 1, 2...  
## $ host\_total\_listings\_count <int> 1, 1, 1, 1, 1, 1, 2, 23, 1, 2...  
## $ host\_verifications <chr> "['email', 'phone', 'reviews'...  
## $ host\_has\_profile\_pic <chr> "t", "t", "t", "t", "t", "t",...  
## $ host\_identity\_verified <chr> "f", "f", "t", "f", "t", "f",...  
## $ street <chr> "Acton, CA 93510, United Stat...  
## $ neighbourhood <chr> NA, NA, NA, NA, "West Adams",...  
## $ neighbourhood\_cleansed <chr> "Acton", "Acton", "Acton", "A...  
## $ neighbourhood\_group\_cleansed <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ city <chr> "Acton", "Palmdale", "Acton",...  
## $ state <chr> "CA", "CA", "CA", "CA", "CA",...  
## $ zipcode <chr> "93510", "93550", "93510", "9...  
## $ market <chr> "Other (Domestic)", "Other (D...  
## $ smart\_location <chr> "Acton, CA", "Palmdale, CA", ...  
## $ country\_code <chr> "US", "US", "US", "US", "US",...  
## $ country <chr> "United States", "United Stat...  
## $ latitude <dbl> 34.53035, 34.48543, 34.51013,...  
## $ longitude <dbl> -118.2271, -118.1254, -118.21...  
## $ is\_location\_exact <chr> "t", "f", "t", "t", "t", "t",...  
## $ property\_type <chr> "House", "House", "House", "O...  
## $ room\_type <chr> "Entire home/apt", "Private r...  
## $ accommodates <int> 10, 2, 2, 6, 1, 1, 2, 2, 2, 2...  
## $ bathrooms <dbl> 7.0, 1.0, 1.0, 1.0, 1.0, 1.0,...  
## $ bedrooms <int> 5, 1, 1, 1, 1, 1, 1, 1, 2, 1,...  
## $ beds <int> 5, 1, 1, 3, 1, 1, 1, 1, 1, 2,...  
## $ bed\_type <chr> "Real Bed", "Real Bed", "Real...  
## $ amenities <chr> "{\"Wireless Internet\",\"Air...  
## $ square\_feet <int> NA, NA, NA, NA, NA, NA, NA, N...  
## $ price <chr> "$3,000.00", "$50.00", "$55.0...  
## $ weekly\_price <chr> NA, NA, NA, NA, NA, NA, NA, "...  
## $ monthly\_price <chr> NA, NA, NA, NA, NA, NA, NA, "...  
## $ security\_deposit <chr> "$2,000.00", NA, NA, NA, NA, ...  
## $ cleaning\_fee <chr> "$200.00", NA, NA, "$35.00", ...  
## $ guests\_included <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## $ extra\_people <chr> "$0.00", "$0.00", "$10.00", "...  
## $ minimum\_nights <int> 1, 1, 1, 1, 1, 2, 1, 1, 1, 1,...  
## $ maximum\_nights <int> 1125, 1125, 1125, 2, 90, 1125...  
## $ calendar\_updated <chr> "5 months ago", "2 months ago...  
## $ has\_availability <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ availability\_30 <int> 27, 29, 18, 3, 0, 4, 6, 22, 2...  
## $ availability\_60 <int> 53, 45, 43, 17, 0, 17, 21, 52...  
## $ availability\_90 <int> 82, 75, 73, 30, 0, 32, 42, 82...  
## $ availability\_365 <int> 352, 350, 348, 87, 0, 32, 42,...  
## $ calendar\_last\_scraped <chr> "2017-05-03", "2017-05-03", "...  
## $ number\_of\_reviews <int> 0, 33, 14, 22, 3, 0, 38, 23, ...  
## $ first\_review <chr> NA, "2015-12-04", "2016-09-02...  
## $ last\_review <chr> NA, "2017-04-22", "2017-04-29...  
## $ review\_scores\_rating <int> NA, 93, 100, 100, 93, NA, 98,...  
## $ review\_scores\_accuracy <int> NA, 10, 10, 10, 10, NA, 10, 9...  
## $ review\_scores\_cleanliness <int> NA, 10, 10, 10, 9, NA, 10, 8,...  
## $ review\_scores\_checkin <int> NA, 10, 10, 10, 10, NA, 10, 9...  
## $ review\_scores\_communication <int> NA, 10, 10, 10, 10, NA, 10, 9...  
## $ review\_scores\_location <int> NA, 9, 10, 10, 9, NA, 10, 8, ...  
## $ review\_scores\_value <int> NA, 9, 10, 10, 10, NA, 10, 9,...  
## $ requires\_license <chr> "f", "f", "f", "f", "f", "f",...  
## $ license <chr> NA, NA, NA, NA, NA, NA, NA, N...  
## $ jurisdiction\_names <chr> NA, NA, NA, NA, "City of Los ...  
## $ instant\_bookable <chr> "t", "f", "f", "t", "f", "f",...  
## $ cancellation\_policy <chr> "strict", "flexible", "flexib...  
## $ require\_guest\_profile\_picture <chr> "f", "f", "f", "f", "f", "f",...  
## $ require\_guest\_phone\_verification <chr> "f", "f", "f", "f", "f", "f",...  
## $ calculated\_host\_listings\_count <int> 1, 1, 1, 1, 1, 1, 2, 19, 1, 2...  
## $ reviews\_per\_month <dbl> NA, 1.91, 1.72, 2.12, 0.18, N...

dim(Detail\_listing)

## [1] 31253 95

# Data Wrangling and Cleaning

"Detail\_Listings.csv" Dataset contains 95 columns. I will remove some of the columns that won't be useful based on the goals I have for this project. I would be using the below columns from the Dataset:

glimpse(clean\_listing)

## Observations: 31,253  
## Variables: 23  
## $ id <int> 16228948, 8909380, 14078522, 13006928, ...  
## $ accommodates <int> 10, 2, 2, 6, 1, 1, 2, 2, 2, 2, 2, 3, 1,...  
## $ host\_response\_rate <chr> "25%", "100%", "100%", "100%", "N/A", "...  
## $ host\_is\_superhost <chr> "f", "t", "t", "t", "f", "f", "t", "f",...  
## $ host\_has\_profile\_pic <chr> "t", "t", "t", "t", "t", "t", "t", "t",...  
## $ host\_identity\_verified <chr> "f", "f", "t", "f", "t", "f", "t", "t",...  
## $ host\_listings\_count <int> 1, 1, 1, 1, 1, 1, 2, 23, 1, 2, 1, 2, 23...  
## $ instant\_bookable <chr> "t", "f", "f", "t", "f", "f", "f", "f",...  
## $ property\_type <chr> "House", "House", "House", "Other", "Ap...  
## $ room\_type <chr> "Entire home/apt", "Private room", "Pri...  
## $ bathrooms <dbl> 7.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, NA, ...  
## $ bedrooms <int> 5, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, ...  
## $ beds <int> 5, 1, 1, 3, 1, 1, 1, 1, 1, 2, 1, 2, 1, ...  
## $ bed\_type <chr> "Real Bed", "Real Bed", "Real Bed", "Re...  
## $ cancellation\_policy <chr> "strict", "flexible", "flexible", "flex...  
## $ availability\_365 <int> 352, 350, 348, 87, 0, 32, 42, 357, 364,...  
## $ price <chr> "$3,000.00", "$50.00", "$55.00", "$150....  
## $ weekly\_price <chr> NA, NA, NA, NA, NA, NA, NA, "$320.00", ...  
## $ monthly\_price <chr> NA, NA, NA, NA, NA, NA, NA, "$999.00", ...  
## $ cleaning\_fee <chr> "$200.00", NA, NA, "$35.00", "$5.00", N...  
## $ security\_deposit <chr> "$2,000.00", NA, NA, NA, NA, NA, NA, NA...  
## $ number\_of\_reviews <int> 0, 33, 14, 22, 3, 0, 38, 23, 0, 42, 15,...  
## $ review\_scores\_rating <int> NA, 93, 100, 100, 93, NA, 98, 89, NA, 9...

Now, my current Dataset has 23 variables for 31,253 observations. I need to clean the data for further analysis.  
 1. The Price columns are of "Char" data type and contains a Dollar sign and a comma. I need to remove the "'$'" and "," and then convert them into a Numeric variable for my analysis. Below are the price columns that I would be using: \* price  
\* weekly\_price  
\* monthly\_price  
\* cleaning\_fee  
\* security\_deposit

#Remove "$" and "," from the price columns  
# Transforming prices to number  
clean\_listing$price <-  
 gsub("\\$", replacement = "", clean\_listing$price)  
clean\_listing$price <-  
 gsub(",", replacement = "", clean\_listing$price)  
clean\_listing$price <- as.numeric(clean\_listing$price)  
  
clean\_listing$security\_deposit <-  
 gsub("\\$", replacement = "", clean\_listing$security\_deposit)  
clean\_listing$security\_deposit <-  
 gsub(",", replacement = "", clean\_listing$security\_deposit)  
clean\_listing$security\_deposit <-  
 as.numeric(clean\_listing$security\_deposit)  
  
clean\_listing$cleaning\_fee <-  
 gsub("\\$", replacement = "", clean\_listing$cleaning\_fee)  
clean\_listing$cleaning\_fee <-  
 gsub(",", replacement = "", clean\_listing$cleaning\_fee)  
clean\_listing$cleaning\_fee <- as.numeric(clean\_listing$cleaning\_fee)  
  
clean\_listing$weekly\_price <-  
 gsub("\\$", replacement = "", clean\_listing$weekly\_price)  
clean\_listing$weekly\_price <-  
 gsub(",", replacement = "", clean\_listing$weekly\_price)  
clean\_listing$weekly\_price <- as.numeric(clean\_listing$weekly\_price)  
  
clean\_listing$monthly\_price <-  
 gsub("\\$", replacement = "", clean\_listing$monthly\_price)  
clean\_listing$monthly\_price <-  
 gsub(",", replacement = "", clean\_listing$monthly\_price)  
clean\_listing$monthly\_price <-  
 as.numeric(clean\_listing$monthly\_price)

2. The Categorical Variables need to be converted into Factors. Below are the Variables that needs to be converted to Factors:  
\* host\_response\_rate  
\* host\_is\_superhost  
\* host\_identity\_verified \* host\_has\_profile\_pic  
\* instant\_bookable  
\* property\_type  
\* room\_type  
\* bed\_type  
\* cancellation\_policy

# Transforming to factors:  
to\_factor <- c(  
 'host\_has\_profile\_pic',  
 'host\_response\_rate',  
 'host\_is\_superhost',  
 'instant\_bookable',  
 'host\_identity\_verified',  
 'property\_type',  
 'room\_type',  
 'bed\_type',  
 'cancellation\_policy'  
)  
for (col in to\_factor) {  
 clean\_listing[[col]] <- factor(clean\_listing[[col]])  
}

3. Before proceeding further, let's deal with the 'NA' Values. First, let's find out how many 'NA' values we have for each variable and below is the result:

sapply(clean\_listing, function (x) sum(is.na(x)))

## id accommodates host\_response\_rate   
## 0 0 14   
## host\_is\_superhost host\_has\_profile\_pic host\_identity\_verified   
## 14 14 14   
## host\_listings\_count instant\_bookable property\_type   
## 14 0 0   
## room\_type bathrooms bedrooms   
## 0 97 35   
## beds bed\_type cancellation\_policy   
## 59 0 0   
## availability\_365 price weekly\_price   
## 0 0 23823   
## monthly\_price cleaning\_fee security\_deposit   
## 24620 6974 14397   
## number\_of\_reviews review\_scores\_rating   
## 0 7527

* host\_is\_superhost:  
  This Variable has categorical values TRUE and FALSE. So, let's assume the NA records are FALSE i.e. The records for which a value is not there is not a Super Host.

## Assuming that NAs in the host\_is\_superhost columns are False  
clean\_listing$host\_is\_superhost[is.na(clean\_listing$host\_is\_superhost)] <- "f"

* Bedrooms,Bathrooms & Beds:  
  For Bedrooms, Bathrooms and Beds variables, there are a coulple of NA records. These variable are numerical and let's replace the NA value with the most common value for that variable.

## Replacing the NAs with the MEAN value of that column  
table(clean\_listing$bedrooms)

##   
## 0 1 2 3 4 5 6 7 8 9 10   
## 2937 19937 4868 2197 812 297 96 45 9 12 8

clean\_listing$bedrooms[is.na(clean\_listing$bedrooms)] <- 1  
  
table(clean\_listing$bathrooms)

##   
## 0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5 5.5   
## 138 162 21969 1530 4601 866 893 311 250 115 76 68   
## 6 6.5 7 7.5 8   
## 35 18 27 20 77

clean\_listing$bathrooms[is.na(clean\_listing$bathrooms)] <-1  
  
table(clean\_listing$beds)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12   
## 17822 6856 3116 1677 730 444 151 162 58 86 17 19   
## 13 14 15 16   
## 4 5 3 44

clean\_listing$beds[is.na(clean\_listing$beds)] <- 1

* Price columns:  
  Let's assume that the NAs are actually zeros for all the price columns and replace all NA values with 0.

## Assuming that NAs in the price columns are 0  
clean\_listing$cleaning\_fee[is.na(clean\_listing$cleaning\_fee)] <- 0  
clean\_listing$security\_deposit[is.na(clean\_listing$security\_deposit)] <-  
 0  
clean\_listing$weekly\_price[is.na(clean\_listing$weekly\_price)] <- 0  
clean\_listing$monthly\_price[is.na(clean\_listing$monthly\_price)] <- 0

* Review\_scores\_rating:  
  There are 7527 observations where the review\_scores\_rating is NA.Let's Replace the NA values with the median value of the ratings.

## Replacing the NAs 0 % rating  
clean\_listing$review\_scores\_rating[is.na(clean\_listing$review\_scores\_rating)] <- median(clean\_listing$review\_scores\_rating, na.rm = TRUE)

* host\_response\_rate:  
  Host\_reasponse\_rate column has both "N/A" and "NA" values. First, let's convert the records with "N/A" values to "NA" and then convert the NA records to "0%". It would be helpful if this variable is numeric. To convert it to a Number variable, we need to remove the "%" sign first.

# convert the records with "N/A" values to "NA"  
levels(clean\_listing$host\_response\_rate)[levels(clean\_listing$host\_response\_rate) == "N/A"] <- NA  
# convert the NA records to "0%"  
clean\_listing$host\_response\_rate[is.na(clean\_listing$host\_response\_rate)] <- "0%"  
# Remove the % sign  
clean\_listing$host\_response\_rate <-  
 gsub("\\%", replacement = "", clean\_listing$host\_response\_rate)  
# Convert the variable to nemeric  
clean\_listing$host\_response\_rate <-as.numeric(clean\_listing$host\_response\_rate)

* host\_identity\_verified:  
  This Variable has categorical values of TRUE and FALSE. So, let's assume the NA records are FALSE i.e. the records which do not have a value for this variable, the host identity is not verified.

clean\_listing$host\_identity\_verified[is.na(clean\_listing$host\_identity\_verified)] <- "f"

* host\_has\_profile\_pic: This Variable has categorical values TRUE and FALSE. So, let's assume the NA records are FALSE i.e. the records which do not have a value for this variable, the host has not uploaded a profile picture.

## Assuming that NAs in the host\_has\_profile\_pic columns are False  
clean\_listing$host\_has\_profile\_pic[is.na(clean\_listing$host\_has\_profile\_pic)] <- "f"

* host\_listings\_count: This Variable has numerical value.So, let's assume the NA records are equal 1 since at least one listing is there for the host.

## Replacing the NAs with the 1  
clean\_listing$host\_listings\_count[is.na(clean\_listing$host\_listings\_count)] <- 1

Let's check again if any column has a NA value:

sapply(clean\_listing, function (x) sum(is.na(x)))

## id accommodates host\_response\_rate   
## 0 0 0   
## host\_is\_superhost host\_has\_profile\_pic host\_identity\_verified   
## 0 0 0   
## host\_listings\_count instant\_bookable property\_type   
## 0 0 0   
## room\_type bathrooms bedrooms   
## 0 0 0   
## beds bed\_type cancellation\_policy   
## 0 0 0   
## availability\_365 price weekly\_price   
## 0 0 0   
## monthly\_price cleaning\_fee security\_deposit   
## 0 0 0   
## number\_of\_reviews review\_scores\_rating   
## 0 0

So, now we do not have any NA records.

4. For the Property\_Type Variable, most of the listings are either an Apartment or an independent House. There are a few listings for the rest of the categories.

table(clean\_listing$property\_type)

##   
## Apartment Bed & Breakfast Boat   
## 13801 328 44   
## Boutique hotel Bungalow Cabin   
## 23 455 77   
## Camper/RV Casa particular Castle   
## 109 1 10   
## Cave Chalet Condominium   
## 1 4 917   
## Dorm Earth House Guest suite   
## 128 4 25   
## Guesthouse Hostel House   
## 560 66 12736   
## Hut In-law Island   
## 8 3 1   
## Lighthouse Loft Other   
## 1 563 318   
## Parking Space Plane Serviced apartment   
## 1 1 14   
## Tent Tipi Townhouse   
## 18 4 769   
## Train Treehouse Vacation home   
## 1 9 9   
## Villa Yurt   
## 234 10

So, let's keep only 3 categories for the property\_type variable as below for simplicity:  
1. Apartment  
2. House  
3. Others

5. Let's check the price variable.

summary(clean\_listing$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.0 69.0 100.0 180.1 170.0 10000.0

quantile(clean\_listing$price, probs = seq(0.1, 0.9, 0.1))

## 10% 20% 30% 40% 50% 60% 70% 80% 90%   
## 46 60 75 90 100 125 150 199 299

Let's restrict our analysis to those listings whose price is less or equal to $500/per night. The goal of this project is to understand the prices for the vast majority of the listings. We would not like our analysis to be influenced by some unusually expensive properties.

# take a subset of the records with price <= 500  
clean\_listing <-  
 clean\_listing %>%  
 filter(price <= 500)  
dim(clean\_listing)

## [1] 29982 23

So, now our current dataset has **29,982** observations which is 95% of the original dataset with **31,253** observations. 6. Let's check the cancellation\_policy for the listings:

table(clean\_listing$cancellation\_policy)

##   
## flexible moderate strict super\_strict\_30   
## 9280 7638 13054 1   
## super\_strict\_60   
## 9

Since there are only a few observations under the categories "super\_strict\_30" and "super\_strict\_60", let's change the cancellation\_policy to "strict" for these observations.

# Change the few observations that have cancelation policy of super\_strict\_30 and super\_strict\_60 to strict  
levels(clean\_listing$cancellation\_policy)[levels(clean\_listing$cancellation\_policy) =="super\_strict\_30"] <- "strict"  
levels(clean\_listing$cancellation\_policy)[levels(clean\_listing$cancellation\_policy) =="super\_strict\_60"] <- "strict"

# Adding New Features

**1. has\_weekly\_discount & has\_monthly\_discount:** There is a weekly\_price variable and a monthly\_price variable. For, some of the listings, there are values for these two variables. Let's add two new features based on these two variable to see if there is a discount if someone is booking the listing weekly basis or a monthly basis.

# Add 2 new columns has\_weekly\_discount and has\_monthly\_discount  
clean\_listing$has\_weekly\_discount <-   
ifelse(((clean\_listing$price \* 7) - clean\_listing$weekly\_price) > 0 & clean\_listing$weekly\_price!= 0, "Yes", "No" )  
clean\_listing$has\_monthly\_discount <- ifelse(((clean\_listing$price \* 30) - clean\_listing$monthly\_price) > 0 & clean\_listing$monthly\_price!= 0, "Yes", "No" )

**2. good\_host\_response:** Based on the existing variable "host\_response\_rate", let's add a new feature "good\_host\_response". Considering, if a host is responding 60% of the time, then the host\_response\_rate is good.

clean\_listing$good\_host\_response <- ifelse(clean\_listing$host\_response\_rate > 60, "Yes", "No" )

**Now, the data is ready for analysis.**

# Exploratory Data Analysis

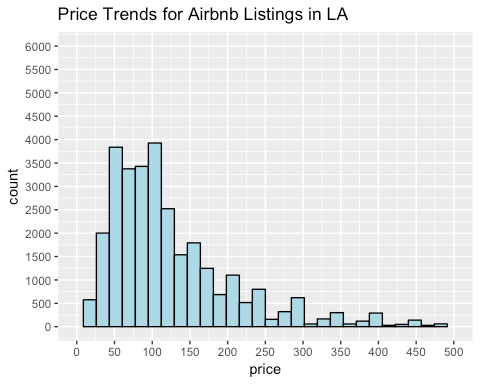
Let's summarize the variables with some visuals.Here, my goal is as below:  
 1. Understand the variables individually  
 2. Spot relationships between all the explanatory variables and the price

**1. Understand the variables individually**

**1. Price:**  
Since we are working on understanding the Price for the Airbnb listings in LA, let's first check what's the price range for the listings in LA.

ggplot(clean\_listing, aes(x = price)) +  
 geom\_histogram(color = 'black',  
 fill = 'light blue',  
 na.rm = TRUE) +  
 scale\_x\_continuous(limits = c(0, 500), breaks = seq(0, 500, 50)) +  
 scale\_y\_continuous(limits = c(0, 6000), breaks = seq(0, 6000, 500))+  
 ggtitle("Price Trends for Airbnb Listings in LA")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 **Conclusion:** It looks like most of the Airbnb properties in LA are in the range of 50 - 120 dollars per night. There are only a few listing in higher price range.

**2. Property\_Type & Room\_Type:**  
Let's see what type of property are usually rented out as part of an Airbnb in LA.

table(clean\_listing$property\_type, clean\_listing$room\_type)

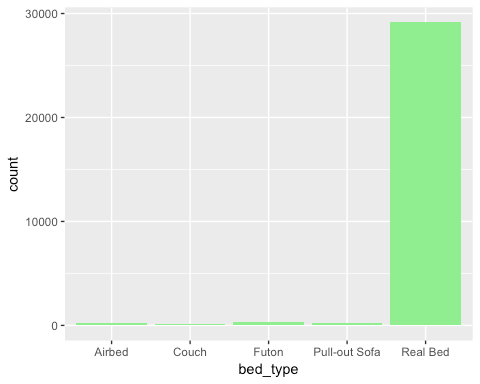
##   
## Entire home/apt Private room Shared room  
## Apartment 9015 4009 689  
## Others 2531 1684 306  
## House 5240 5878 630

ggplot(clean\_listing, aes(x=property\_type,fill=room\_type)) +  
 geom\_bar(na.rm = TRUE)

 **Conclusion:** For Airbnb, people usually rent out their entire home or entire apartment or even a private room in their house.Also, it is observed here that entire home or a private room is more in demand than a shared accommodation which is really understandable.

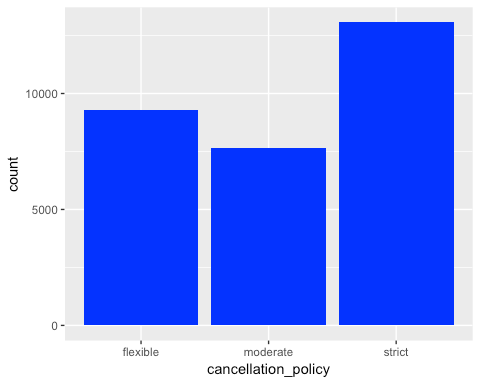
**3. Bed\_Type:**  
Let's see what types of beds are mostly used in a Airbnb.

ggplot(clean\_listing, aes(x = bed\_type)) +  
 geom\_bar(fill = 'light green', na.rm = TRUE)

 **Conclusion:** Though the concept of Airbnb started with Airbeds, it looks like Real Reds are much more popular nowadays.

**4.Cancellation\_policy:**  
Let's see what cancellation policy are more appropriate for an Airbnb in LA.

ggplot(clean\_listing, aes(x = cancellation\_policy)) +  
 geom\_bar(fill = 'blue', na.rm = TRUE)

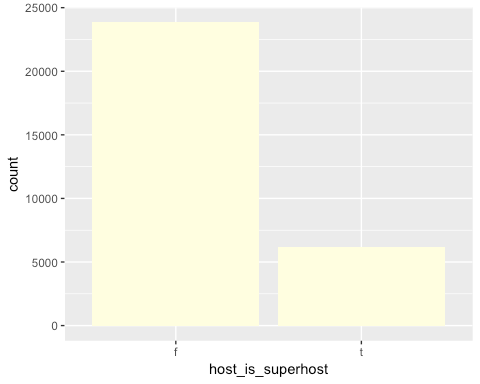
 **Conclusion:** Hosts are usually strict regarding any cancellation which is quite natural because a cancellation means some loss in business.

**5. Host\_is\_superhost:**  
Let's see the ratio of Super Hosts vs Normal Host

prop.table(table(clean\_listing$host\_is\_superhost))

##   
## f t   
## 0.795077 0.204923

ggplot(clean\_listing, aes(x = host\_is\_superhost)) +  
 geom\_bar(fill = 'light yellow')

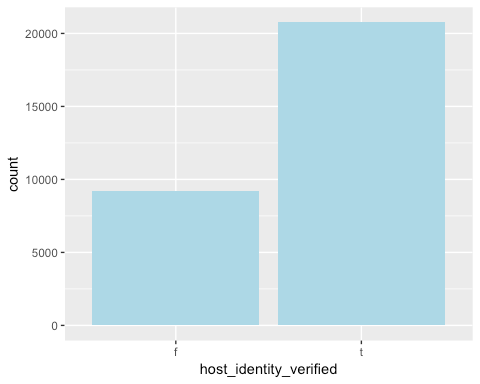
 **Conclusion:** Only 20% of the hosts in LA are identified as a "Super" host.

**6. Host\_identity\_verified:**  
Let's see if most of the host's identity is verified in LA Airbnb

prop.table(table(clean\_listing$host\_identity\_verified))

##   
## f t   
## 0.3078514 0.6921486

ggplot(clean\_listing, aes(x = host\_identity\_verified)) +  
 geom\_bar(fill = 'light blue')

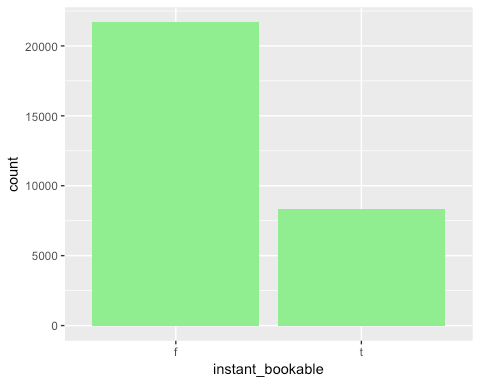
 **Conclusion:** 70% of the host's identity is verified in LA and still there are 30% of the hosts whose identity is not yet verified and that's alarming.

**7. Instant\_bookable:**  
Instant Book listings don't require an approval from the host before they can be booked. Instead, guests can just choose their travel dates, book, and discuss check-in plans with the host.

prop.table(table(clean\_listing$instant\_bookable))

##   
## f t   
## 0.7230005 0.2769995

ggplot(clean\_listing, aes(x = instant\_bookable)) +  
 geom\_bar(fill = 'light green')

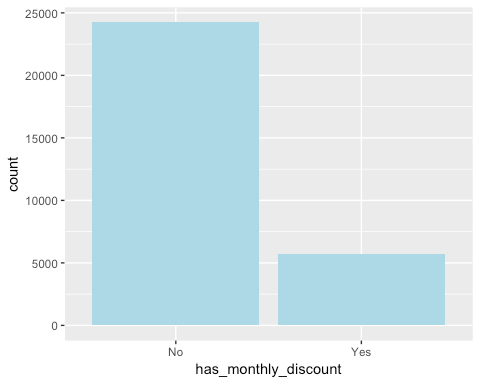
 **Conclusion:** Instant\_booking is NOT opted by most of the hosts yet.But this feature can attract more guests since it can save some time.

**8. has\_monthly\_discount:**

prop.table(table(clean\_listing$has\_monthly\_discount))

##   
## No Yes   
## 0.8082516 0.1917484

ggplot(clean\_listing, aes(x = has\_monthly\_discount)) +  
 geom\_bar(fill = 'light blue')

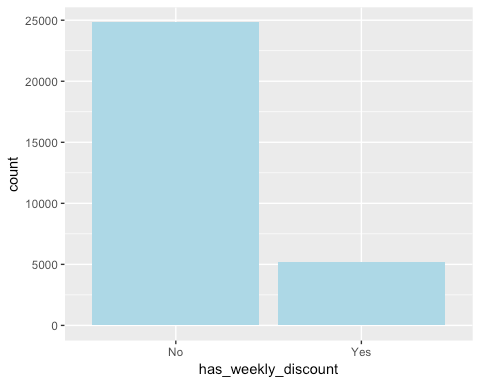
 Most of the host is not providing any monthly discount and it's unlikely to book a listing for a month.

**9. has\_weekly\_discount:**

prop.table(table(clean\_listing$has\_weekly\_discount))

##   
## No Yes   
## 0.8276633 0.1723367

ggplot(clean\_listing, aes(x = has\_weekly\_discount)) +  
 geom\_bar(fill = 'light blue')



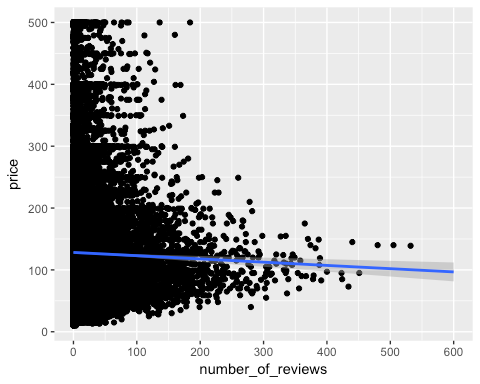
Again, weekly discount is also not available for most of the listings.

**2. Spot relationships between the explanatory variables and price**  
 1. Plot price vs number\_of\_reviews

ggplot(clean\_listing, aes(x = number\_of\_reviews, y = price)) +  
 geom\_point() +  
 geom\_smooth(method = lm, fullrange=TRUE)+  
 scale\_x\_continuous(limits = c(0, 600), breaks = seq(0, 600, 100))

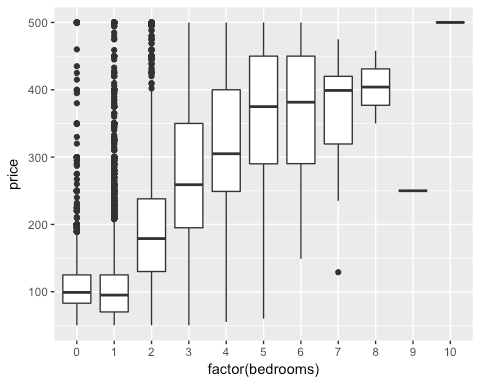
## Warning: Removed 1 rows containing non-finite values (stat\_smooth).

## Warning: Removed 1 rows containing missing values (geom\_point).

 It looks like usually the less price listings has more reviews which is expected since people usually look for some budget friendly properties.

2. Plot price vs bedrooms

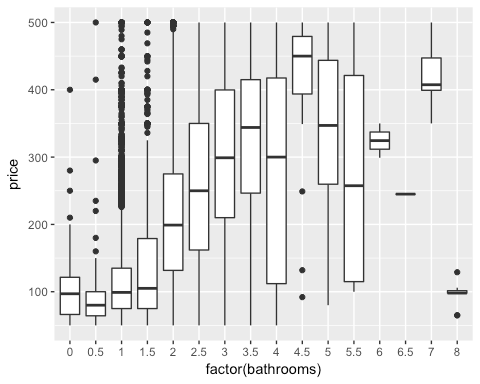
ggplot(clean\_listing, aes(x = factor(bedrooms), y = price)) +  
 geom\_boxplot(na.rm = TRUE) +  
 ylim(50, 500)



As per the above plot results, the price for the listings with more bedrooms are usually more. That means a 1 bedroom apartment is less price than a 3 bedroom apartment and the result is as expected.

3. Plot price vs bathrooms

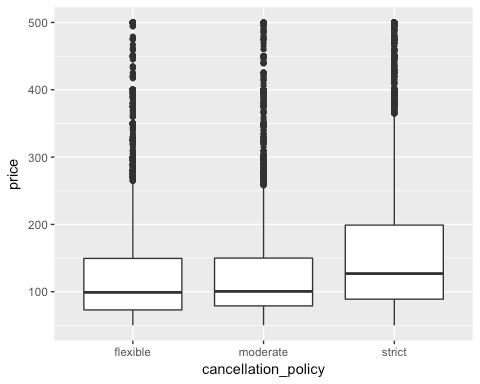
ggplot(clean\_listing, aes(x = factor(bathrooms), y = price)) +  
 geom\_boxplot(na.rm = TRUE) +  
 ylim(50, 500)



Again, as per the above plot results, the price for the listings with more bathrooms are usually more. That means a 2 bedroom apartment with 1 bath is less price than a 2 bedroom apartment with 2 baths and the result is as expected.

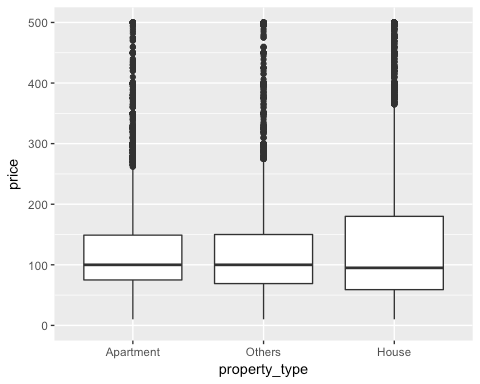
4. Plot price vs cancellation\_policy

ggplot(clean\_listing, aes(x = cancellation\_policy, y = price)) +  
 geom\_boxplot(na.rm = TRUE) +  
 ylim(50, 500)

 Cancellation Policy does not influence the price much.So, a host can select a strict cancellation policy and choose the same price as a listing with a flexible cancellation policy.

5. Plot price vs property\_type

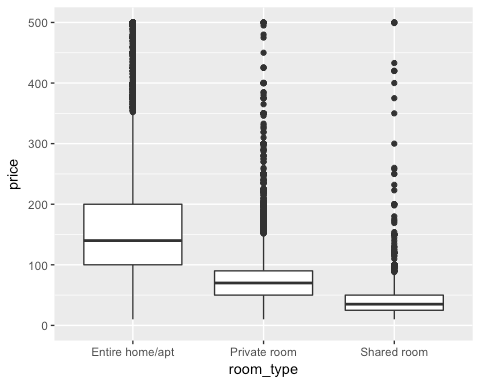
ggplot(clean\_listing, aes(x = property\_type, y = price)) +  
 geom\_boxplot(na.rm = TRUE) +  
 ylim(00, 500)



As per the plot, the Mean price for each property type is at the range of $100 per night and the property type does not have a great influence on price.

6. Plot price vs room\_type

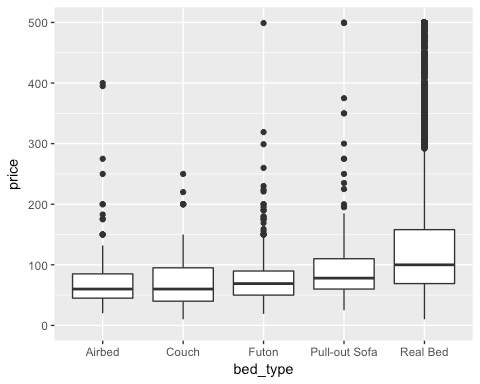
ggplot(clean\_listing, aes(x = room\_type, y = price)) +  
 geom\_boxplot(na.rm = TRUE) +  
 ylim(00, 500)



As per the plot, shared room has the least Mean price. The private rooms have a little higher Mean price than a shared accommodation, but quite less than the entire home. The entire home shows a lot more variation on average than a private room or a shared accommodation.

7. Plot price vs bed\_type

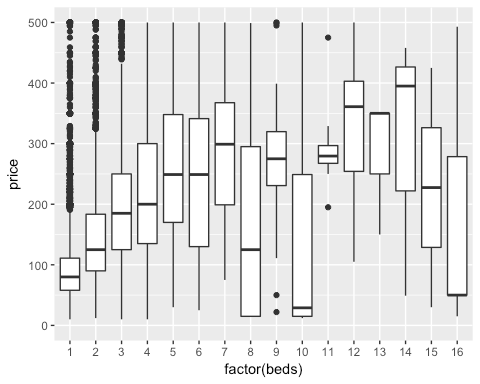
ggplot(clean\_listing, aes(x = bed\_type, y = price)) +  
 geom\_boxplot(na.rm = TRUE) +  
 ylim(00, 500)



The price is not much influenced by what type of Bed is there.Real Beds has a slightly more Mean price than any other types.

8. Plot Price vs No. of Beds

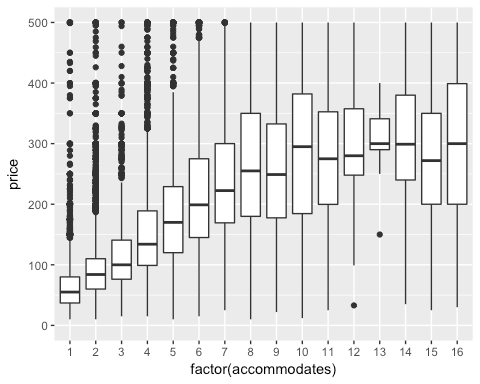
ggplot(clean\_listing, aes(x = factor(beds), y = price)) +  
 geom\_boxplot(na.rm = TRUE) +  
 ylim(0, 500)



As per the plot, it looks like with an increase in number of beds, the price is increasing. But then, there is a little drop in price for beds number equal to 8 or 10 or 16.

9. Plot price vs accommodates

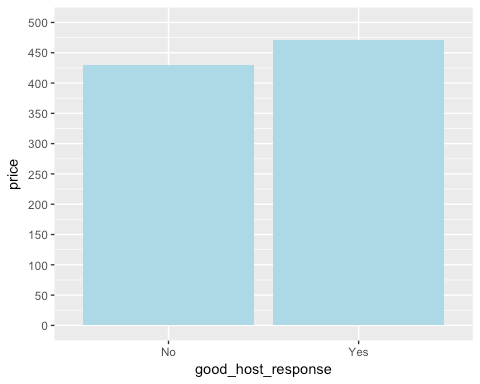
ggplot(clean\_listing, aes(x = factor(accommodates), y = price)) +  
 geom\_boxplot(na.rm = TRUE) +  
 ylim(0, 500)



As per the plot, the price is increasing with an increase in the number of accommodates the host does allow for a listing.

10. Plot price vs host\_response\_rate

ggplot(clean\_listing, aes(x = good\_host\_response, y = price)) +  
 geom\_bar(fill = 'light blue',na.rm = TRUE, stat = "identity") +  
 scale\_y\_continuous(limits = c(0, 500), breaks = seq(0, 500, 50))

 Apparently, host with better response can claim better price for their listing.

# Create a Model

Let's create a Linear Regression Model to predict the price of an Airbnb Listing in LA. First, we will divide the records into training and test dataset. 80% of the Records goes to Training Set and the rest 20% to Test Set.

dim(train\_data)

## [1] 23986 26

dim(test\_data)

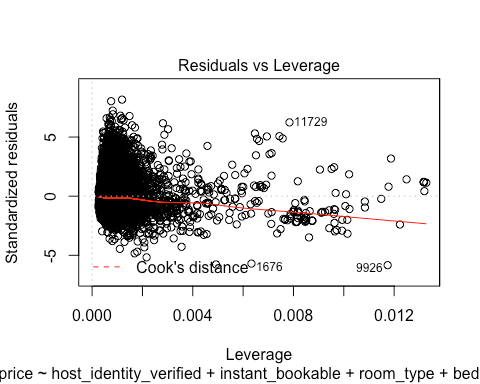
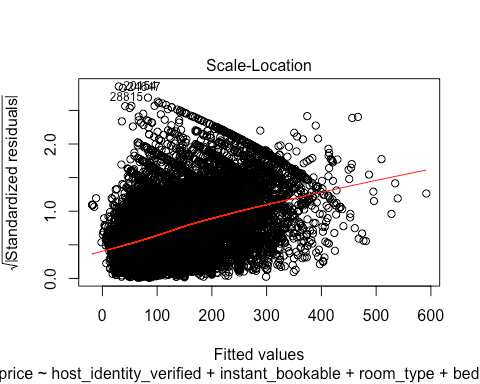
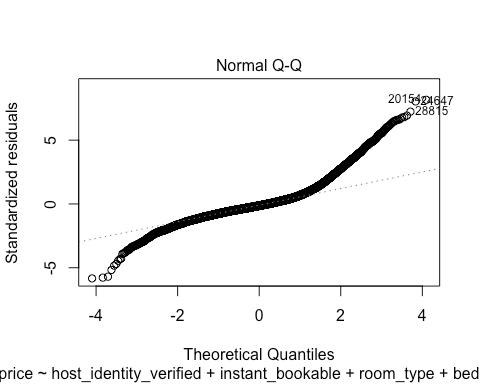
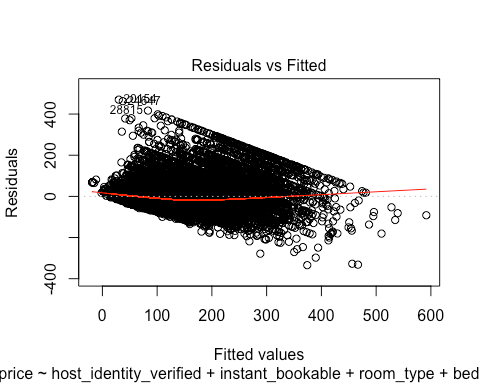
## [1] 5996 26

Let's build our model to predict price for the Budget Listings.

# Create a Linear Regression Model  
lm\_model <-  
 lm(  
 price ~ host\_identity\_verified + instant\_bookable + room\_type + bedrooms + bathrooms +  
 beds + availability\_365 +  
 accommodates + security\_deposit + host\_is\_superhost +cleaning\_fee + property\_type  
 + has\_weekly\_discount + has\_monthly\_discount  
 + good\_host\_response + cancellation\_policy ,  
 data = train\_data  
 )  
summary(lm\_model)

##   
## Call:  
## lm(formula = price ~ host\_identity\_verified + instant\_bookable +   
## room\_type + bedrooms + bathrooms + beds + availability\_365 +   
## accommodates + security\_deposit + host\_is\_superhost + cleaning\_fee +   
## property\_type + has\_weekly\_discount + has\_monthly\_discount +   
## good\_host\_response + cancellation\_policy, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -334.16 -30.82 -7.74 19.80 470.05   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 67.827745 1.447174 46.869 < 2e-16 \*\*\*  
## host\_identity\_verifiedt -2.472554 0.843299 -2.932 0.00337 \*\*   
## instant\_bookablet -1.788636 0.862854 -2.073 0.03819 \*   
## room\_typePrivate room -46.882663 0.987875 -47.458 < 2e-16 \*\*\*  
## room\_typeShared room -76.772182 1.876612 -40.910 < 2e-16 \*\*\*  
## bedrooms 28.083060 0.724519 38.761 < 2e-16 \*\*\*  
## bathrooms 14.257811 0.740938 19.243 < 2e-16 \*\*\*  
## beds -3.230659 0.449636 -7.185 6.91e-13 \*\*\*  
## availability\_365 0.042988 0.002848 15.092 < 2e-16 \*\*\*  
## accommodates 7.188325 0.320357 22.439 < 2e-16 \*\*\*  
## security\_deposit 0.025128 0.001211 20.741 < 2e-16 \*\*\*  
## host\_is\_superhostt 5.031987 0.972737 5.173 2.32e-07 \*\*\*  
## cleaning\_fee 0.322851 0.010006 32.266 < 2e-16 \*\*\*  
## property\_typeOthers 7.947201 1.118353 7.106 1.23e-12 \*\*\*  
## property\_typeHouse 6.636721 0.883311 7.513 5.96e-14 \*\*\*  
## has\_weekly\_discountYes 7.384364 1.352018 5.462 4.76e-08 \*\*\*  
## has\_monthly\_discountYes -6.707560 1.312600 -5.110 3.24e-07 \*\*\*  
## good\_host\_responseYes -21.262682 1.019791 -20.850 < 2e-16 \*\*\*  
## cancellation\_policymoderate -8.684707 1.056477 -8.220 < 2e-16 \*\*\*  
## cancellation\_policystrict -7.304214 0.984920 -7.416 1.25e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 57.61 on 23966 degrees of freedom  
## Multiple R-squared: 0.5914, Adjusted R-squared: 0.591   
## F-statistic: 1825 on 19 and 23966 DF, p-value: < 2.2e-16

plot(lm\_model)

 Let's check the accuracy of this model.

predict.lm\_model <- predict(lm\_model, test\_data)  
head(predict.lm\_model)

## 1539 13343 14238 25361 7203 11368   
## 77.76655 86.08840 88.88619 198.62701 132.27435 40.67321

head(test\_data$price)

## [1] 70 44 99 249 119 38

SSE <- sum((test\_data$price - predict.lm\_model) ^ 2)  
SST <- sum((test\_data$price - mean(test\_data$price)) ^ 2)  
R\_Squared\_Value <- 1 - SSE / SST  
R\_Squared\_Value

## [1] 0.5842782

The Adjusted R-Squared value for the linear regression model to predict the price of an Airbnb listing is **0.591**. For the test data set, the R-Squared value is **0.5883346** which is almost the same as the training set data.

# Conclusion

The scope of this project is to understand how the different features can influence the price and to create a tool to predict the price for an Airbnb listing. Though, the current Linear Regression Model to predict the price is not perfect (R-Squared Value = 0.59), the model still would be helpful for providing a guidance for pricing.  
 Finally, let's summarize our observations for the Airbnb dataset and build our own data story.

1. Price for most of the listings in LA is in the range of $50 to $120 per night.

2. Airbnb started with 3 Airbeds and with time Real Bed has gained more popularity.

3. Most of the Airbnb hosts prefer a strict cancellation policy as a last minute cancellation can incur loss.

4. Renting out an entire home or an entire apartment can be more beneficial to quote a better price than a private room or a shared one.

5. Usually an Apartment or an independent house is more in demand. Maybe that's because many families prefer an Airbnb during a vacation to get a feeling and comfort of their own home. Adventurous options like a Boat or a Treehouse are there but less in number.

Now, let's understand how the different features are related to the Price.

coefficients(lm\_model)

## (Intercept) host\_identity\_verifiedt   
## 67.82774521 -2.47255371   
## instant\_bookablet room\_typePrivate room   
## -1.78863591 -46.88266293   
## room\_typeShared room bedrooms   
## -76.77218225 28.08305973   
## bathrooms beds   
## 14.25781090 -3.23065858   
## availability\_365 accommodates   
## 0.04298835 7.18832513   
## security\_deposit host\_is\_superhostt   
## 0.02512804 5.03198681   
## cleaning\_fee property\_typeOthers   
## 0.32285080 7.94720067   
## property\_typeHouse has\_weekly\_discountYes   
## 6.63672068 7.38436422   
## has\_monthly\_discountYes good\_host\_responseYes   
## -6.70755957 -21.26268160   
## cancellation\_policymoderate cancellation\_policystrict   
## -8.68470732 -7.30421411

The below features have a **Positive** relation with price. i.e. while selecting a price for a new listing in LA, we can conclude that if the host is a "Super Host", he can ask for a better price. Or, a house with more bedrooms and bathrooms (a bigger house) that accommodates more people can get better price. 1. bathrooms  
2. availability\_365  
3. security\_deposit  
4. cleaning\_fee  
5. property\_typeHouse  
6. property\_typeOthers  
7. has\_weekly\_discountYes  
8. bedrooms  
9. accommodates  
10. host\_is\_superhostt  
 The below features have a **Negative** relation with price as per the data we are using for our analysis. i.e. That means the listings with Private rooms and Shared rooms can get less price. If the Host is Strict or Moderate with the cancellation policies, usually they ask for a lesser price. 1. instant\_bookablet  
2. room\_typeShared room  
3. room\_typePrivate room  
4. has\_monthly\_discountYes  
5. cancellation\_policymoderate  
6. cancellation\_policystrict  
7. good\_host\_responseYes  
8. beds  
9. host\_identity\_verifiedt

# Future Scope

1. Exploring and adding some more features could improve the current model for predicting price of an LA based Airbnb Listing.
2. There is a future scope to analyze and understand how the review scores are getting influenced by the different features and build a model to predict the review scores ratings.
3. Using the other available Datasets in Airbnb like the one for Reviews, we can also try to build a Recommendation Model for the Guests.
4. We can try to categorize the Guests' sentiments for each of the listings.  
    --The End--