Project Part 2

Parham

5/28/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

```
Airbnb_raw <- fread(paste0("C:/Users/Parham/Desktop/listings.csv"))
Airbnb_DT <- Airbnb_raw %>%
  select(id, name, neighbourhood, last_scraped,
         host_name, host_location, host_id, host_since, host_listings_count, host_is_superhost,
         host_response_time, host_response_rate, host_acceptance_rate, host_verifications,
         latitude, longitude, property_type, room_type, accommodates, bedrooms, beds,price,
         bathrooms_text, amenities,number_of_reviews, first_review, last_review, review_scores_rating,
         review_scores_cleanliness, review_scores_location, review_scores_value,
         review_scores_communication, instant_bookable)
rm(Airbnb_raw)
glimpse(Airbnb_DT)
```

```
## Rows: 4,531
## Columns: 33
## $ id                           <int> 10080, 13188, 13358, 13490, 14267, 16254, ~
## $ name                         <chr> "D1 -  Million Dollar View 2 BR", "Garden ~
## $ neighbourhood                <chr> "", "Vancouver, British Columbia, Canada",~
## $ last_scraped                 <date> 2021-12-11, 2021-12-11, 2021-12-11, 2021-~
## $ host_name                    <chr> "Rami", "Family Guns", "Lynn", "Iris", "Pe~
## $ host_location                <chr> "Vancouver, British Columbia, Canada", "Va~
## $ host_id                      <int> 30899, 51466, 52116, 52467, 56030, 63238, ~
## $ host_since                   <date> 2009-08-10, 2009-11-04, 2009-11-07, 2009-~
## $ host_listings_count          <int> 49, 2, 1, 4, 1, 1, 5, 1, 1, 2, 1, 1, 3, 1,~
## $ host_is_superhost            <chr> "f", "f", "f", "t", "f", "f", "f", "f", "f~
## $ host_response_time           <chr> "within a day", "within an hour", "within ~
## $ host_response_rate           <chr> "90%", "100%", "89%", "100%", "67%", "N/A"~
## $ host_acceptance_rate         <chr> "70%", "100%", "60%", "78%", "64%", "100%"~
## $ host_verifications           <chr> "['email', 'phone', 'reviews', 'jumio']", ~
## $ latitude                     <dbl> 49.28872, 49.24773, 49.28201, 49.25622, 49~
## $ longitude                    <dbl> -123.1205, -123.1051, -123.1267, -123.0661~
## $ property_type                <chr> "Entire condominium (condo)", "Entire rent~
## $ room_type                    <chr> "Entire home/apt", "Entire home/apt", "Ent~
## $ accommodates                 <int> 5, 4, 2, 2, 4, 4, 6, 7, 2, 2, 2, 2, 2, 2, ~
## $ bedrooms                     <int> 2, NA, 1, 1, 1, 2, 3, 4, 1, 1, 1, NA, 1, N~
## $ beds                         <int> 2, 2, 1, 1, 2, 3, 4, 5, 1, 1, 1, 1, 1, 1, ~
## $ price                        <chr> "$158.00", "$150.00", "$85.00", "$150.00",~
## $ bathrooms_text               <chr> "2 baths", "1 bath", "1 bath", "1 bath", "~
## $ amenities                    <chr> "[\"\"Bed linens\"\", \"\"Hot water\"\", \~
## $ number_of_reviews            <int> 16, 254, 431, 91, 34, 5, 3, 7, 203, 118, 3~
## $ first_review                 <date> 2011-11-15, 2010-02-21, 2010-06-22, 2011-~
## $ last_review                  <date> 2017-02-26, 2021-11-26, 2021-11-15, 2021-~
## $ review_scores_rating         <dbl> 4.64, 4.83, 4.67, 4.92, 4.76, 4.80, 4.00, ~
## $ review_scores_cleanliness    <dbl> 4.71, 4.84, 4.80, 4.96, 4.68, 4.80, 3.00, ~
## $ review_scores_location       <dbl> 4.86, 4.88, 4.92, 4.77, 4.77, 5.00, 5.00, ~
## $ review_scores_value          <dbl> 4.50, 4.80, 4.67, 4.89, 4.71, 4.60, 3.67, ~
## $ review_scores_communication <dbl> 4.71, 4.91, 4.80, 4.97, 4.68, 5.00, 4.33, ~
## $ instant_bookable             <chr> "f", "f", "f", "f", "t", "t", "f", "f", "f~
```

```
str(Airbnb_DT$price)
```

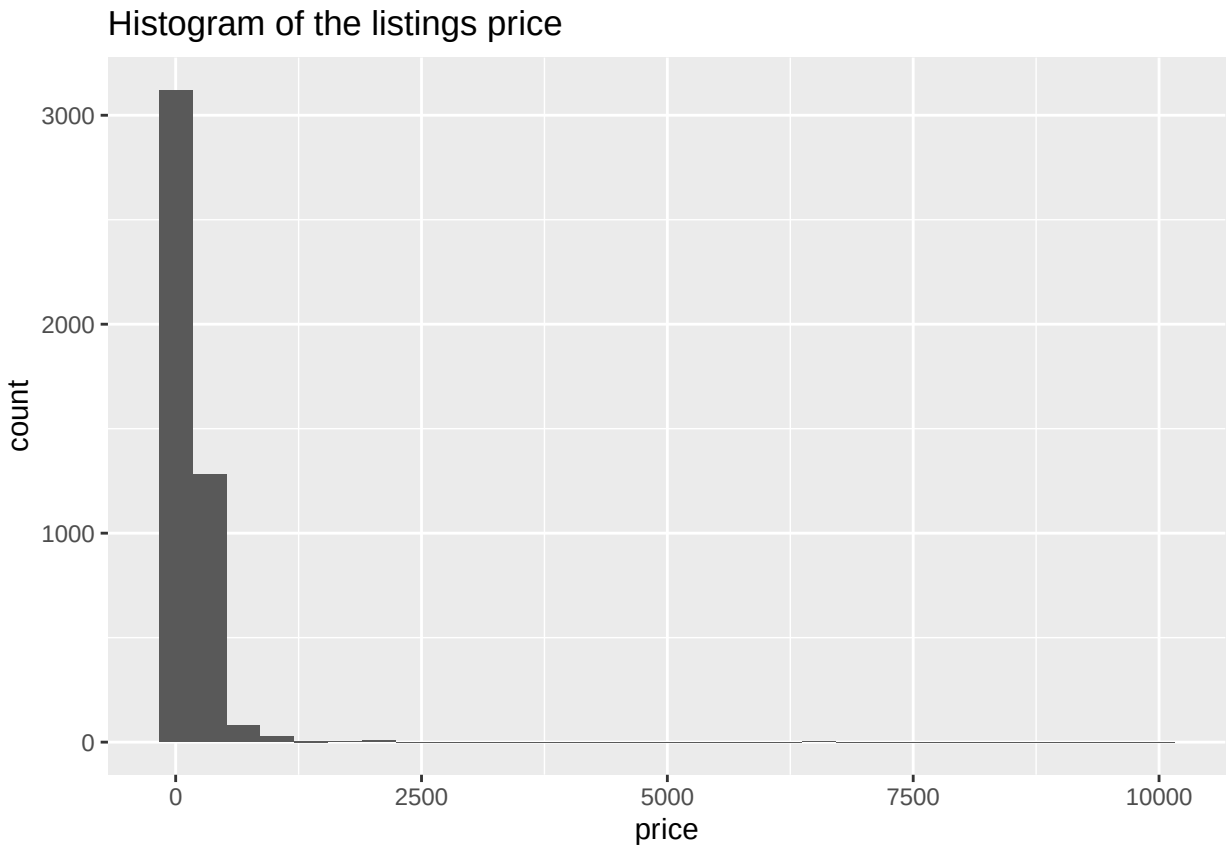
```
##  chr [1:4531] "$158.00" "$150.00" "$85.00" "$150.00" "$150.00" "$350.00" ...
```

```
Airbnb_DT <- Airbnb_DT %>%
  mutate(price = as.numeric(gsub("[\\$,]", "", price)) ,
         host_acceptance_rate = as.numeric(gsub("[\\%,]", "", host_acceptance_rate)),
         host_response_rate = as.numeric(gsub("[\\%,]", "", host_response_rate))
         )
```

Histogram of price

```
Airbnb_DT %>%
  ggplot(aes(x = price)) +
  geom_histogram() +
  labs(title = "Histogram of the listings price")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Remove outliers to see price distribution better

```
summary(Airbnb_DT$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    13.0    92.5   129.0   179.0   199.0  9999.0
```

```
n = dim(Airbnb_DT)[1]
sum(Airbnb_DT$price > 1250)
```

```
## [1] 22
```

```
indx <- which(Airbnb_DT$price > 1250)
indx
```

```
##  [1]   26  328  603  708 1109 1770 1913 2079 2119 2446 3135 3308 3487 3676 3773
## [16] 3842 3891 4067 4084 4127 4420 4491
```

```
Airbnb_DT[-indx,] %>%
  ggplot(aes(x = price)) +
    geom_histogram() +
    labs(title = "Histogram of the listings price")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

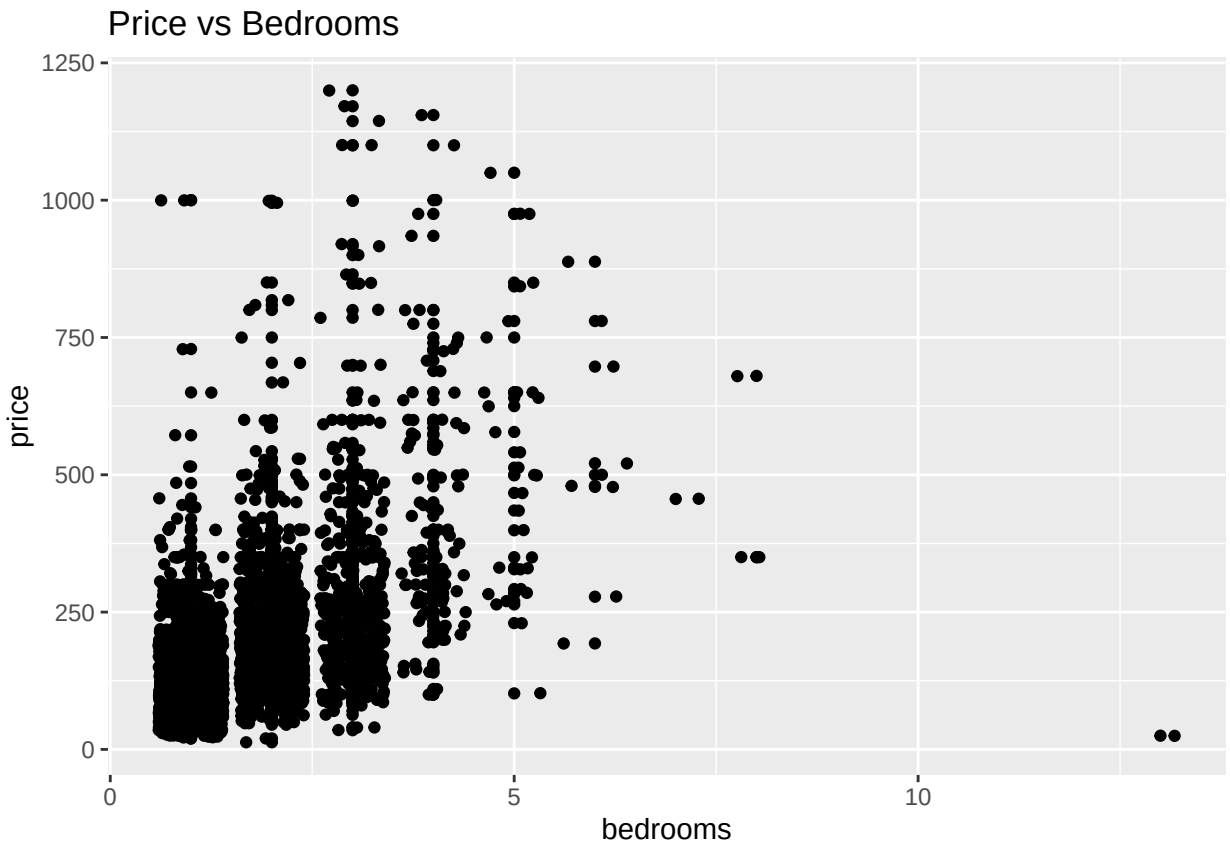
Checking predictors

Beds Or Bedrooms

```
cor(Airbnb_DT$beds,Airbnb_DT$bedrooms,use = "na.or.complete")
```

```
## [1] 0.831141
```

```
Airbnb_DT[-indx,] %>%
  ggplot(aes(x = bedrooms , y = price)) +
  geom_point() +
  geom_jitter()+
labs(title = "Price vs Bedrooms")
```



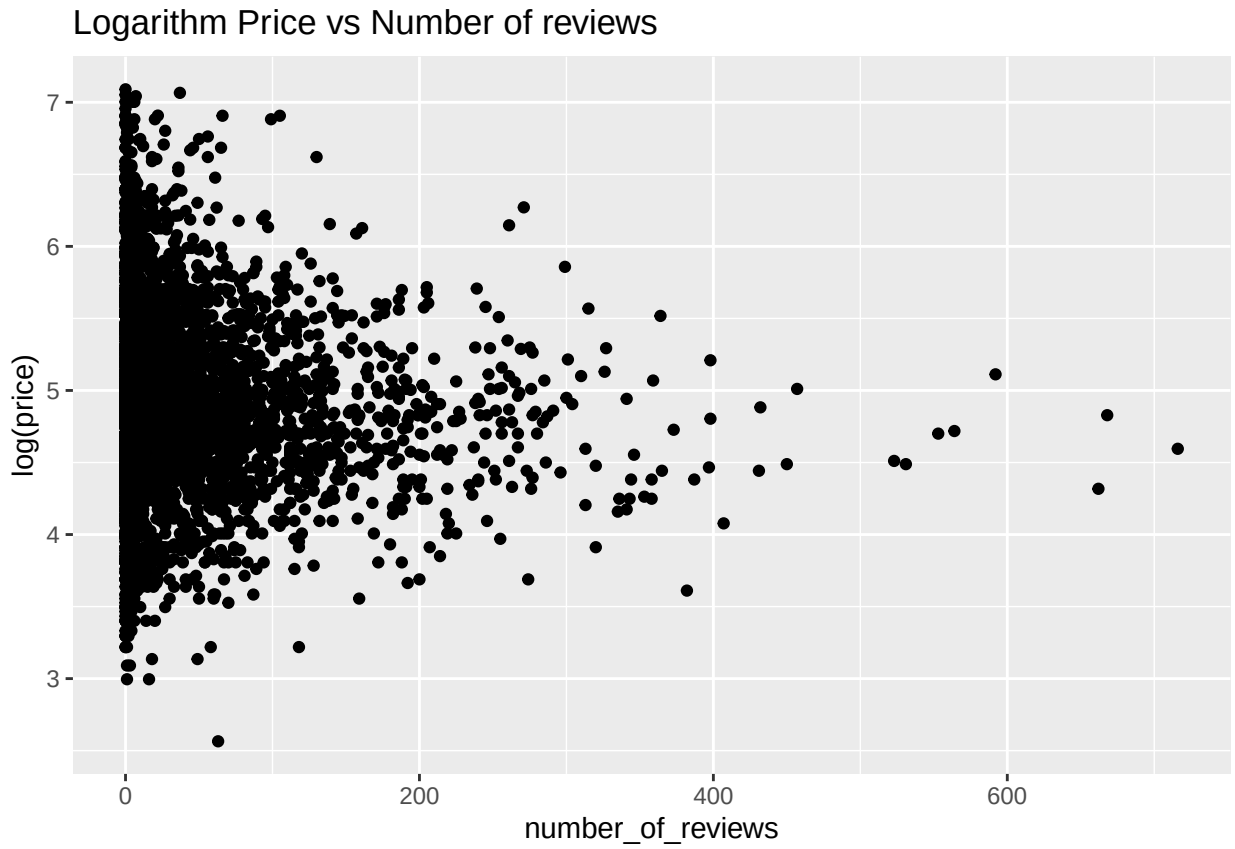

```
Airbnb_DT[-indx,] %>%
  ggplot(aes(x = bedrooms , y = log(price))) +
  geom_point() +
  geom_jitter()+
labs(title = "Logarithm Price vs Bedrooms")
```


```
fit <- lm(log(price) ~ beds ,data = Airbnb_DT[-indx,])
summary(fit)
```

```
##
## Call:
## lm(formula = log(price) ~ beds, data = Airbnb_DT[-indx, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7664 -0.3263 -0.0108  0.3231  2.2430
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.388035   0.015346  285.95   <2e-16 ***
## beds        0.276709   0.006964   39.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5258 on 4349 degrees of freedom
##   (158 observations deleted due to missingness)
## Multiple R-squared:  0.2664, Adjusted R-squared:  0.2662
## F-statistic:  1579 on 1 and 4349 DF,  p-value: < 2.2e-16
```

Number of reviews

```
Airbnb_DT[-indx,] %>%
  ggplot(aes(x = number_of_reviews , y = log(price))) +
  geom_point()+
labs(title = "Logarithm Price vs Number of reviews")
```



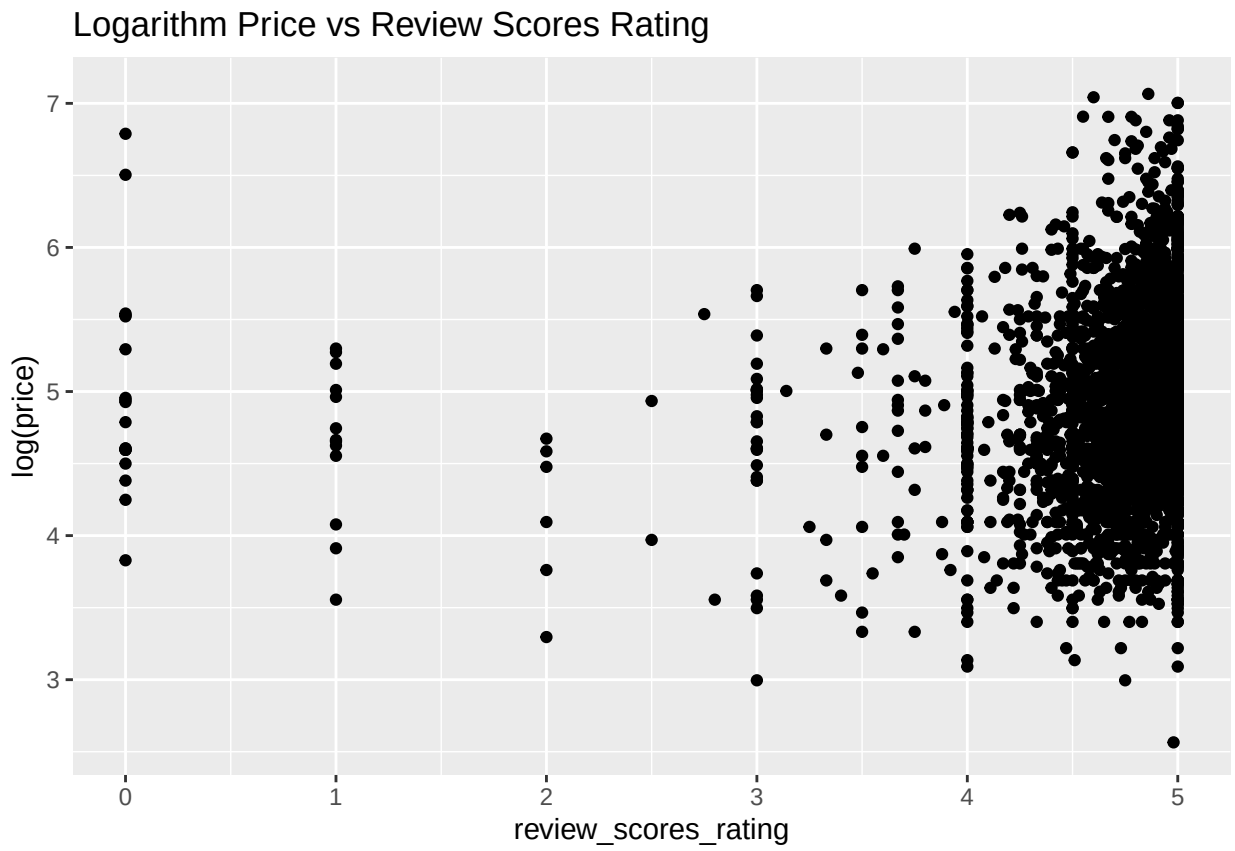

```
fit <- lm(log(price) ~ number_of_reviews ,data = Airbnb_DT[-indx,])
summary(fit)
```

```
##
## Call:
## lm(formula = log(price) ~ number_of_reviews, data = Airbnb_DT[-indx,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32095 -0.36599 -0.04195  0.37775  2.17021
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.9198655  0.0105833 464.872  < 2e-16 ***
## number_of_reviews -0.0005392  0.0001460  -3.694 0.000224 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6158 on 4507 degrees of freedom
```

```
## Multiple R-squared:  0.003018,   Adjusted R-squared:  0.002797
## F-statistic: 13.64 on 1 and 4507 DF,  p-value: 0.0002236
```

Review Scores Rating

```
Airbnb_DT[-indx,] %>%
  ggplot(aes(x = review_scores_rating , y = log(price))) +
  geom_point()+
labs(title = "Logarithm Price vs Review Scores Rating")
```



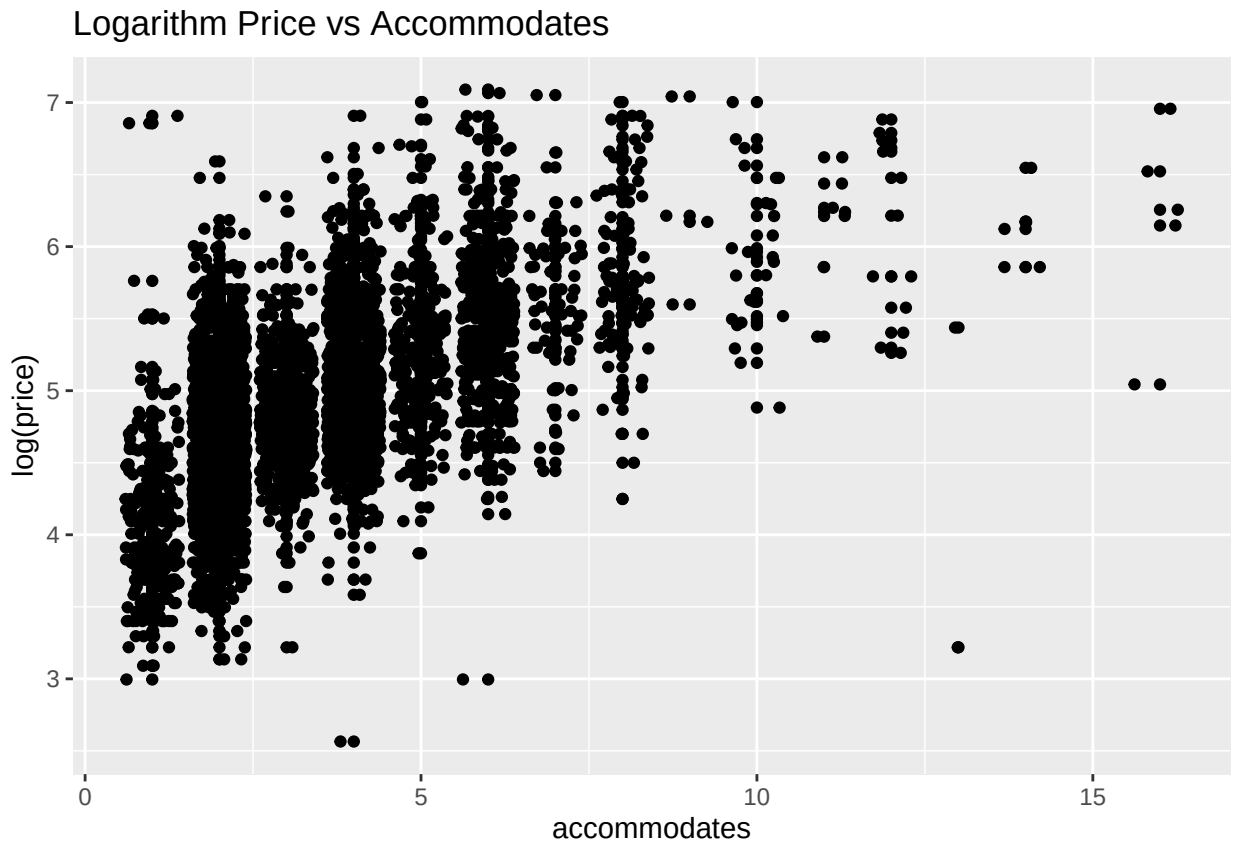

```
fit <- lm(log(price) ~ review_scores_rating ,data = Airbnb_DT[-indx,])
summary(fit)
```

```
##
## Call:
## lm(formula = log(price) ~ review_scores_rating, data = Airbnb_DT[-indx,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36145 -0.37152 -0.04314  0.36964  2.42991
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.35906    0.08919   48.87  < 2e-16 ***
```

```
## review_scores_rating  0.11392    0.01874    6.08 1.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5975 on 3772 degrees of freedom
##   (735 observations deleted due to missingness)
## Multiple R-squared:  0.009705,   Adjusted R-squared:  0.009443
## F-statistic: 36.97 on 1 and 3772 DF,  p-value: 1.321e-09
```

Accommodates

```
Airbnb_DT[-indx,] %>%
  ggplot(aes(x = accommodates , y = log(price))) +
  geom_point()+
  geom_jitter()+
labs(title = "Logarithm Price vs Accommodates")
```



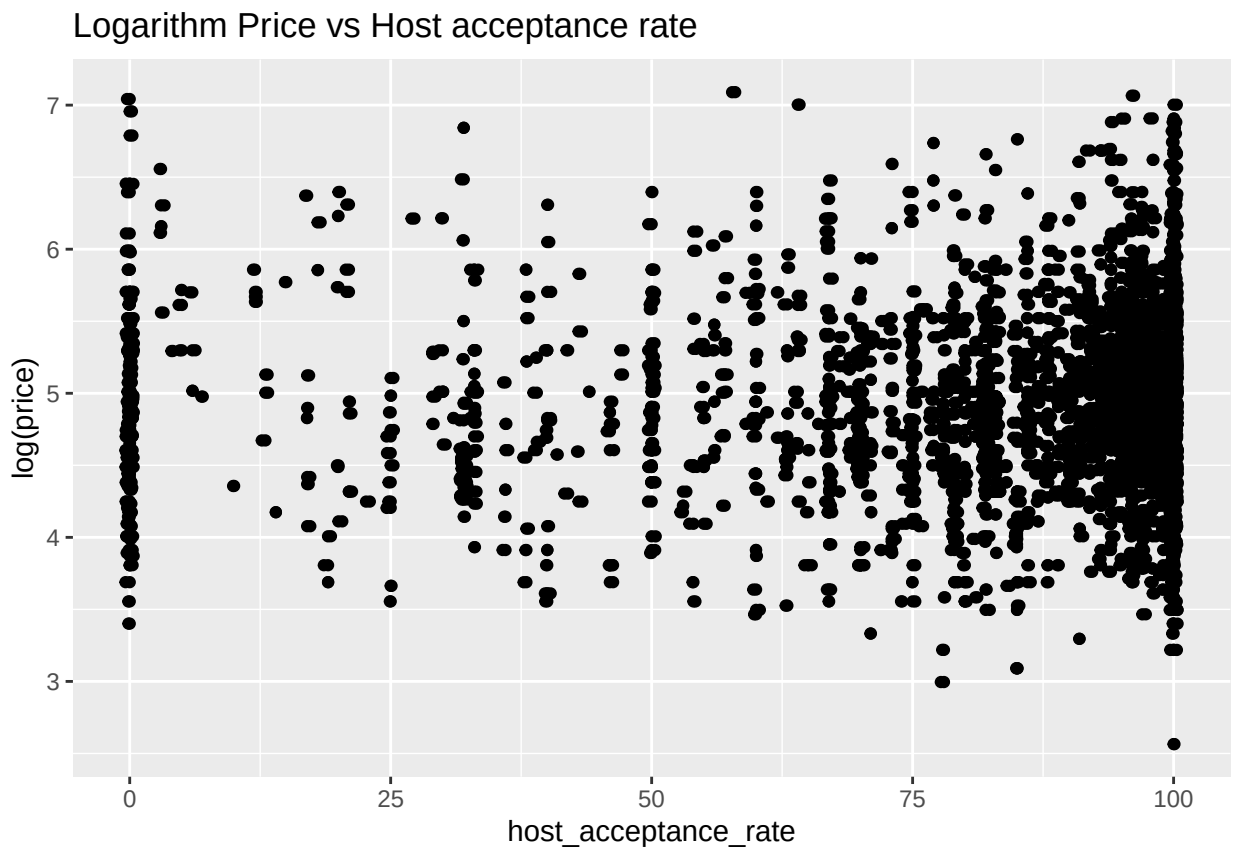

```
fit <- lm(log(price) ~ accommodates ,data = Airbnb_DT[-indx,])
summary(fit)
```

```
##
## Call:
## lm(formula = log(price) ~ accommodates, data = Airbnb_DT[-indx,
##     ])
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5652 -0.3059 -0.0162  0.2913  2.4830
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.22818    0.01519  278.38   <2e-16 ***
## accommodates  0.19660    0.00389   50.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4927 on 4507 degrees of freedom
## Multiple R-squared:  0.3618, Adjusted R-squared:  0.3616
## F-statistic:  2555 on 1 and 4507 DF,  p-value: < 2.2e-16
```

Host acceptance rate & Host response rate

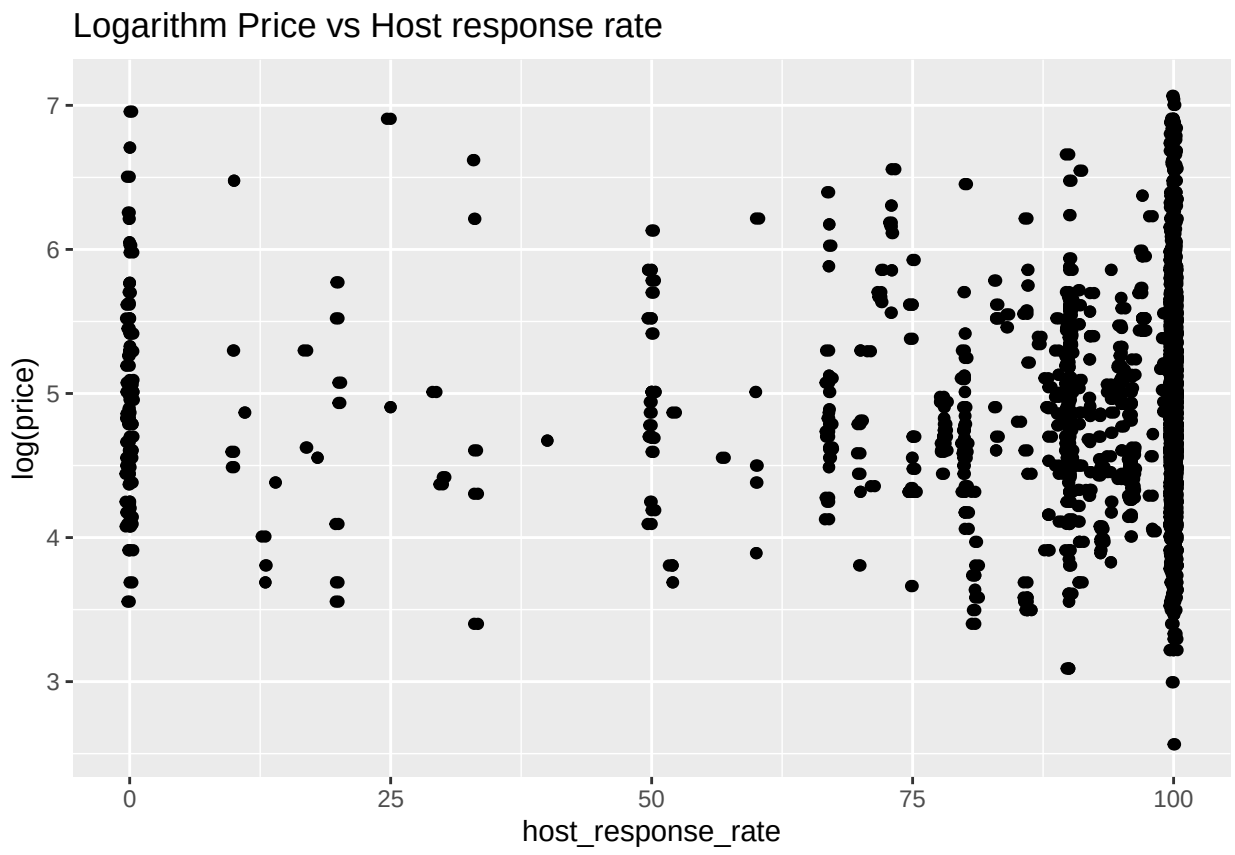
```
Airbnb_DT[-indx,] %>%
  ggplot(aes(x = host_acceptance_rate , y = log(price))) +
  geom_point()+
  geom_jitter()+
  labs(title = "Logarithm Price vs Host acceptance rate")
```




```
fit <- lm(log(price) ~ host_acceptance_rate ,data = Airbnb_DT[-indx,])
summary(fit)
```

```
##
## Call:
## lm(formula = log(price) ~ host_acceptance_rate, data = Airbnb_DT[-indx,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3504 -0.3615 -0.0464  0.3678  2.1949
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.8671948  0.0345270  140.97   <2e-16 ***
## host_acceptance_rate 0.0004818  0.0003981    1.21    0.226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6019 on 3317 degrees of freedom
##   (1190 observations deleted due to missingness)
## Multiple R-squared:  0.0004414,  Adjusted R-squared:  0.00014
## F-statistic: 1.465 on 1 and 3317 DF,  p-value: 0.2263
```

```
Airbnb_DT[-indx,] %>%
  ggplot(aes(x = host_response_rate , y = log(price))) +
  geom_point()+
  geom_jitter()+
  labs(title = "Logarithm Price vs Host response rate")
```

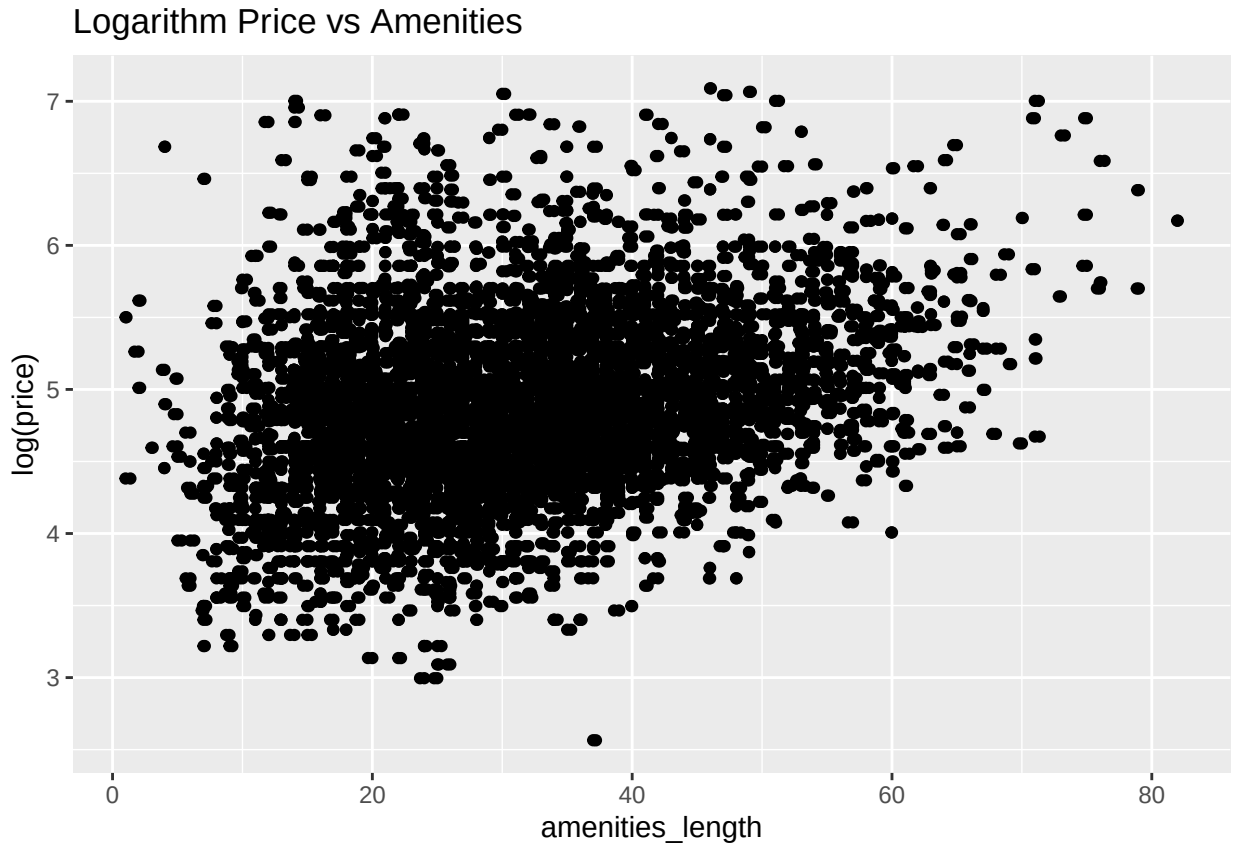


```
fit <- lm(log(price) ~ host_response_rate ,data = Airbnb_DT[-indx,])
summary(fit)
```

```
##
## Call:
## lm(formula = log(price) ~ host_response_rate, data = Airbnb_DT[-indx,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34528 -0.35635 -0.04269  0.36277  2.15538
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.8709727  0.0551987  88.244   <2e-16 ***
## host_response_rate 0.0003926  0.0005761   0.681    0.496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6008 on 3122 degrees of freedom
##   (1385 observations deleted due to missingness)
## Multiple R-squared:  0.0001487,  Adjusted R-squared:  -0.0001716
## F-statistic: 0.4643 on 1 and 3122 DF,  p-value: 0.4957
```

Amenities

```
p <- ""
k <- c()
for(j in 1:nrow(Airbnb_DT))
{
  x<-gsub(" ", "", Airbnb_DT$amenities[j])
  x<-strsplit(x, ",")[[1]]
  z<-c()
  for(i in x)
  {
    #print(gsub("[^0-9A-Za-z///' ]","'" , i))
    p<-gsub("[^0-9A-Za-z///' ]","'" , i)
    p<-gsub("'","",p)
    z = c(z,p)
  }
  k<-c(k,length(z))
}
Airbnb_DT <- Airbnb_DT%>%
  mutate(amenities_length = k)
Airbnb_DT[-indx,] %>%
  ggplot(aes(x = amenities_length , y = log(price))) +
  geom_point()+
  geom_jitter()+
  labs(title = "Logarithm Price vs Amenities")
```




```
fit<-lm(log(price) ~ amenities_length ,data = Airbnb_DT[-indx,])
summary(fit)
```

```
##
## Call:
## lm(formula = log(price) ~ amenities_length, data = Airbnb_DT[-indx,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.40879 -0.37800 -0.04961  0.33948  2.33191
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.4869735  0.0230520  194.65   <2e-16 ***
## amenities_length 0.0131560  0.0006778   19.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5925 on 4507 degrees of freedom
## Multiple R-squared:  0.07715,    Adjusted R-squared:  0.07694
## F-statistic: 376.8 on 1 and 4507 DF,  p-value: < 2.2e-16
```

We can see correlation matrix

```
attach(Airbnb_DT)
data = cbind(price,host_response_rate,host_acceptance_rate,bedrooms,beds,host_listings_count,accommodat
cor(data,use = "na.or.complete")[1,]
```

```
##                price   host_response_rate host_acceptance_rate
##           1.00000000           0.02661787           0.01375713
##             bedrooms                 beds  host_listings_count
##           0.62026421           0.56674160          -0.04426302
##         accommodates review_scores_rating    number_of_reviews
##           0.61236746           0.07692298          -0.07123096
##            longitude             latitude     amenities_length
##          -0.17485875           0.09872218           0.25214777
```

Use a stepwise algorithm

```
fit <- lm(log(price) ~ host_response_rate+host_acceptance_rate+bedrooms+host_listings_count+accommodates
summary(fit)
```

```
##
## Call:
## lm(formula = log(price) ~ host_response_rate + host_acceptance_rate +
##     bedrooms + host_listings_count + accommodates + review_scores_rating +
##     number_of_reviews + longitude + latitude + amenities_length,
##     data = na.omit(Airbnb_DT))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.06059 -0.25668 -0.01862  0.24079  1.75626
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -7.341e+02  3.240e+01 -22.661  < 2e-16 ***
## host_response_rate    2.350e-04  6.127e-04   0.384 0.701359
## host_acceptance_rate  9.656e-04  4.228e-04   2.284 0.022480 *
## bedrooms              1.934e-01  1.729e-02  11.187  < 2e-16 ***
## host_listings_count   3.951e-04  3.448e-04   1.146 0.252001
## accommodates          1.186e-01  7.819e-03  15.164  < 2e-16 ***
## review_scores_rating  1.740e-01  2.091e-02   8.323  < 2e-16 ***
## number_of_reviews    -4.539e-04  1.180e-04  -3.846 0.000123 ***
## longitude            -2.826e+00  2.208e-01 -12.798  < 2e-16 ***
## latitude              7.903e+00  4.212e-01  18.763  < 2e-16 ***
## amenities_length      5.889e-03  7.006e-04   8.405  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3923 on 2337 degrees of freedom
## Multiple R-squared:  0.5723, Adjusted R-squared:  0.5704
## F-statistic: 312.7 on 10 and 2337 DF,  p-value: < 2.2e-16
```

```
anova(fit)
```

```
## Analysis of Variance Table
```

```
##
## Response: log(price)
##                         Df Sum Sq Mean Sq   F value    Pr(>F)
## host_response_rate       1   1.68   1.677   10.8960 0.0009782 ***
## host_acceptance_rate     1   1.04   1.045    6.7897 0.0092269 **
## bedrooms                 1 301.01 301.010 1955.8628 < 2.2e-16 ***
## host_listings_count      1   0.30   0.302    1.9643 0.1611866
## accommodates             1  55.56  55.560  361.0113 < 2.2e-16 ***
## review_scores_rating     1  16.01  16.015  104.0592 < 2.2e-16 ***
## number_of_reviews        1   2.39   2.390   15.5299 8.359e-05 ***
## longitude                1  33.20  33.198  215.7104 < 2.2e-16 ***
## latitude                 1  59.11  59.106  384.0537 < 2.2e-16 ***
## amenities_length         1  10.87  10.873   70.6486 < 2.2e-16 ***
## Residuals             2337 359.67   0.154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
step(fit)
```

```
## Start:  AIC=-4383.17
## log(price) ~ host_response_rate + host_acceptance_rate + bedrooms +
##     host_listings_count + accommodates + review_scores_rating +
##     number_of_reviews + longitude + latitude + amenities_length
##
##                         Df Sum of Sq    RSS     AIC
## - host_response_rate     1     0.023 359.69 -4385.0
## - host_listings_count    1     0.202 359.87 -4383.9
## <none>                               359.67 -4383.2
## - host_acceptance_rate   1     0.803 360.47 -4379.9
## - number_of_reviews      1     2.277 361.94 -4370.4
## - review_scores_rating   1    10.662 370.33 -4316.6
## - amenities_length       1    10.873 370.54 -4315.2
## - bedrooms               1    19.261 378.93 -4262.7
## - longitude              1    25.207 384.87 -4226.1
## - accommodates           1    35.390 395.06 -4164.8
## - latitude               1    54.179 413.85 -4055.7
##
## Step:  AIC=-4385.02
## log(price) ~ host_acceptance_rate + bedrooms + host_listings_count +
##     accommodates + review_scores_rating + number_of_reviews +
##     longitude + latitude + amenities_length
##
##                         Df Sum of Sq    RSS     AIC
## - host_listings_count    1     0.212 359.90 -4385.6
## <none>                               359.69 -4385.0
## - host_acceptance_rate   1     1.180 360.87 -4379.3
## - number_of_reviews      1     2.268 361.96 -4372.3
## - review_scores_rating   1    10.724 370.41 -4318.0
## - amenities_length       1    10.993 370.68 -4316.3
## - bedrooms               1    19.271 378.96 -4264.5
## - longitude              1    25.185 384.88 -4228.1
## - accommodates           1    35.373 395.06 -4166.8
## - latitude               1    54.234 413.92 -4057.3
##
```

```
## Step:  AIC=-4385.64
## log(price) ~ host_acceptance_rate + bedrooms + accommodates +
##     review_scores_rating + number_of_reviews + longitude + latitude +
##     amenities_length
##
##                         Df Sum of Sq    RSS     AIC
## <none>                               359.90 -4385.6
## - host_acceptance_rate  1     1.163 361.07 -4380.1
## - number_of_reviews     1     2.463 362.37 -4371.6
## - review_scores_rating  1    10.513 370.42 -4320.0
## - amenities_length      1    10.781 370.68 -4318.3
## - bedrooms              1    19.242 379.15 -4265.3
## - longitude             1    25.495 385.40 -4226.9
## - accommodates          1    35.320 395.22 -4167.8
## - latitude              1    55.972 415.88 -4048.2
```

```
##
## Call:
## lm(formula = log(price) ~ host_acceptance_rate + bedrooms + accommodates +
##     review_scores_rating + number_of_reviews + longitude + latitude +
##     amenities_length, data = na.omit(Airbnb_DT))
##
## Coefficients:
##          (Intercept)  host_acceptance_rate              bedrooms
##           -7.387e+02             1.033e-03             1.933e-01
##         accommodates  review_scores_rating     number_of_reviews
##            1.184e-01             1.707e-01            -4.689e-04
##            longitude              latitude      amenities_length
##           -2.837e+00             7.968e+00             5.797e-03
```

```
fit2 <- lm(formula = log(price) ~ host_acceptance_rate + bedrooms + accommodates +
review_scores_rating + number_of_reviews + longitude + latitude +
amenities_length, data = na.omit(Airbnb_DT))
summary(fit2)
```

```
##
## Call:
## lm(formula = log(price) ~ host_acceptance_rate + bedrooms + accommodates +
##     review_scores_rating + number_of_reviews + longitude + latitude +
##     amenities_length, data = na.omit(Airbnb_DT))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05854 -0.25398 -0.01848  0.24082  1.75845
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -7.387e+02  3.214e+01 -22.981  < 2e-16 ***
## host_acceptance_rate  1.033e-03  3.756e-04   2.750  0.00601 **
## bedrooms              1.933e-01  1.729e-02  11.183  < 2e-16 ***
## accommodates          1.184e-01  7.817e-03  15.151  < 2e-16 ***
## review_scores_rating  1.707e-01  2.065e-02   8.266 2.30e-16 ***
## number_of_reviews    -4.689e-04  1.172e-04  -4.001 6.51e-05 ***
```

```
## longitude            -2.837e+00  2.204e-01 -12.872  < 2e-16 ***
## latitude              7.968e+00  4.177e-01  19.073  < 2e-16 ***
## amenities_length      5.797e-03  6.926e-04   8.371  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3923 on 2339 degrees of freedom
## Multiple R-squared:  0.572,  Adjusted R-squared:  0.5705
## F-statistic: 390.7 on 8 and 2339 DF,  p-value: < 2.2e-16
```

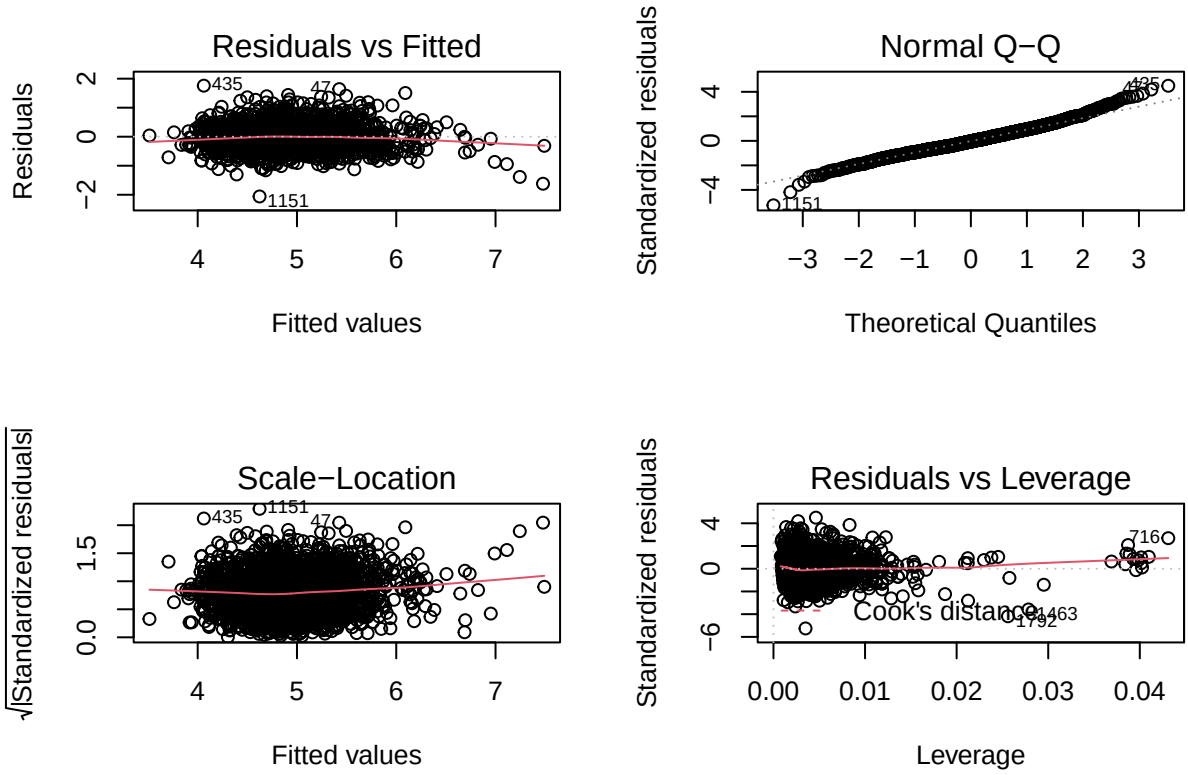
```
AIC(fit2)
```

```
## [1] 2279.696
```

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: log(price)
##                         Df Sum Sq Mean Sq  F value    Pr(>F)
## host_acceptance_rate     1   2.30   2.302   14.958 0.0001129 ***
## bedrooms                 1 301.00 301.000 1956.194 < 2.2e-16 ***
## accommodates             1  55.31  55.309  359.451 < 2.2e-16 ***
## review_scores_rating     1  14.59  14.593   94.836 < 2.2e-16 ***
## number_of_reviews        1   2.90   2.904   18.875 1.456e-05 ***
## longitude                1  33.96  33.956  220.679 < 2.2e-16 ***
## latitude                 1  60.10  60.098  390.573 < 2.2e-16 ***
## amenities_length         1  10.78  10.781   70.067 < 2.2e-16 ***
## Residuals             2339 359.90   0.154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(2,2))
plot(fit2)
```



Negative Binomial & Gamma Model

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

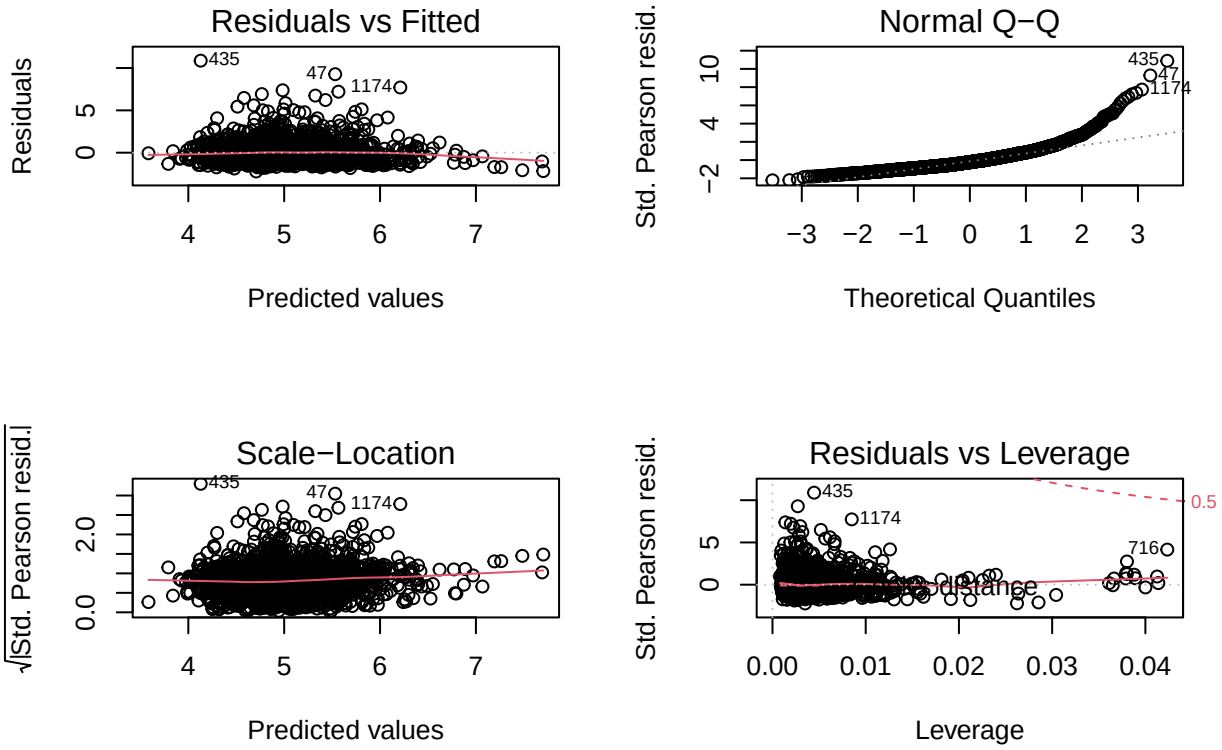
```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
fit3 <- glm.nb(price ~ host_acceptance_rate + bedrooms + accommodates +
review_scores_rating + number_of_reviews + longitude + latitude +
amenities_length, data = na.omit(Airbnb_DT))
summary(fit3)
```

```
##
## Call:
## glm.nb(formula = price ~ host_acceptance_rate + bedrooms + accommodates +
##     review_scores_rating + number_of_reviews + longitude + latitude +
##     amenities_length, data = na.omit(Airbnb_DT), init.theta = 6.680336887,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -3.8232  -0.7996  -0.2390   0.4269   5.8418
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -7.391e+02  3.258e+01 -22.683  < 2e-16 ***
## host_acceptance_rate  7.184e-04  3.794e-04   1.893   0.0583 .
## bedrooms              2.285e-01  1.736e-02  13.159  < 2e-16 ***
## accommodates          1.109e-01  7.865e-03  14.105  < 2e-16 ***
## review_scores_rating  1.725e-01  2.109e-02   8.177 2.92e-16 ***
## number_of_reviews    -5.661e-04  1.191e-04  -4.754 1.99e-06 ***
## longitude            -2.817e+00  2.229e-01 -12.637  < 2e-16 ***
## latitude              8.028e+00  4.238e-01  18.942  < 2e-16 ***
## amenities_length      5.491e-03  6.983e-04   7.863 3.74e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(6.6803) family taken to be 1)
##
##     Null deviance: 5815.5  on 2347  degrees of freedom
## Residual deviance: 2391.3  on 2339  degrees of freedom
## AIC: 25590
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  6.680
##           Std. Err.:  0.199
##
##  2 x log-likelihood:  -25570.278
```

```
par(mfrow = c(2,2))
plot(fit3)
```




```
fit4 <- glm(price ~ host_acceptance_rate + bedrooms + accommodates +
review_scores_rating + number_of_reviews + longitude + latitude +
amenities_length, data = na.omit(Airbnb_DT), family = Gamma("log"))
summary(fit4)
```

```
##
## Call:
## glm(formula = price ~ host_acceptance_rate + bedrooms + accommodates +
##     review_scores_rating + number_of_reviews + longitude + latitude +
##     amenities_length, family = Gamma("log"), data = na.omit(Airbnb_DT))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.58725  -0.31764  -0.09417   0.17066   2.34235
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -7.402e+02  3.659e+01 -20.230  < 2e-16 ***
## host_acceptance_rate  7.312e-04  4.275e-04   1.710   0.0873 .
## bedrooms              2.273e-01  1.968e-02  11.553  < 2e-16 ***
## accommodates          1.120e-01  8.899e-03  12.581  < 2e-16 ***
## review_scores_rating  1.713e-01  2.351e-02   7.288 4.29e-13 ***
## number_of_reviews    -5.601e-04  1.334e-04  -4.199 2.79e-05 ***
## longitude            -2.821e+00  2.509e-01 -11.242  < 2e-16 ***
## latitude              8.042e+00  4.755e-01  16.910  < 2e-16 ***
```

```
## amenities_length      5.500e-03  7.884e-04   6.977 3.92e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1993915)
##
##     Null deviance: 909.68  on 2347  degrees of freedom
## Residual deviance: 375.59  on 2339  degrees of freedom
## AIC: 25576
##
## Number of Fisher Scoring iterations: 5
```

```
par(mfrow = c(2,2))
plot(fit4)
```