Airbnb: Do megahosts improve guest experience? - A Regression Analysis

Nguyen Anh Duong

2025-09-22

Import dataset

```
library("readr")
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

airbnb <- read_csv('airbnb.csv')

## Rows: 19671 Columns: 79

## ── Column specification
────────────────────────────────────────────────────────

## Delimiter: ","
## chr (29): listing_url, last_scraped, source, name, description,
neighborhood...
## dbl (42): id, scrape_id, host_id, host_listings_count,
host_total_listings_c...
## lgl  (8): host_is_superhost, host_has_profile_pic, host_identity_verified,
n...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

Create categorical variable host_type and report the number of listings and proportion of total listings for each of the 3 categories in host_type.

```
airbnb <- airbnb %>%
  mutate(
    host_type = case_when(
      host_total_listings_count >= 21 ~ "megahost",
      host_total_listings_count >=  2 ~ "boutique",
```

```
      host_total_listings_count ==  1 ~ "individual",
      TRUE ~ NA_character_)
  )

q1 <- airbnb %>%
  filter(!is.na(host_type)) %>%
  count(host_type, name = "n_listings") %>%
  mutate(prop = n_listings / sum(n_listings))

q1
```

```
## # A tibble: 3 × 3
##   host_type  n_listings  prop
##   <chr>           <int> <dbl>
## 1 boutique        10844 0.551
## 2 individual       4037 0.205
## 3 megahost         4790 0.244
```

Number of listings:

boutique: 10844

individual: 4037

megahost: 4790

Proportion of total listings:

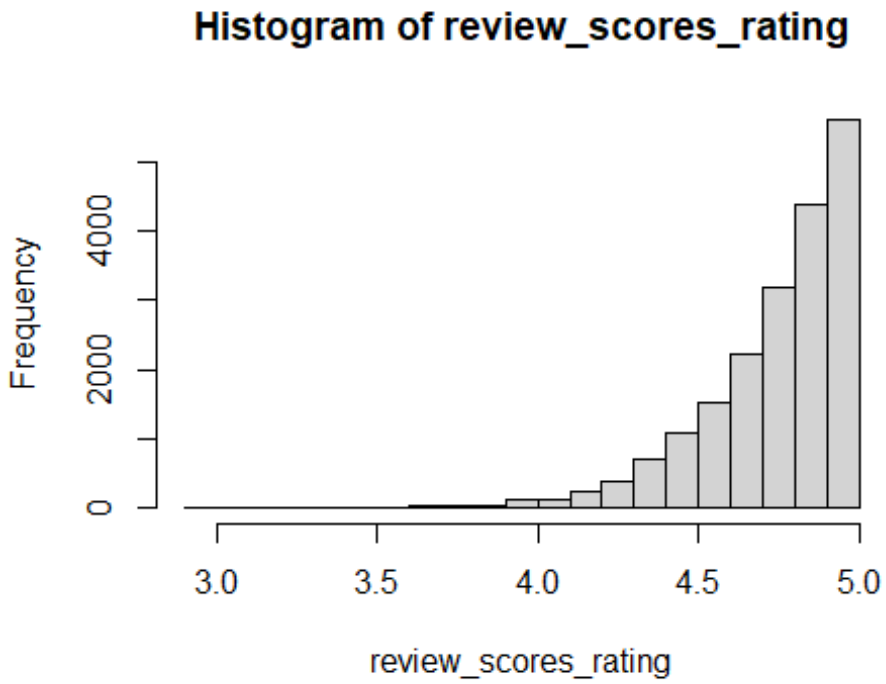
boutique: 55.13% of total listings

individual: 20.52% of total listings

megahost: 24.35% of total listings

Plot a histogram of variable review_scores_rating and calculate proportion of listings that have overall ratings of above 4.

```
hist(airbnb$review_scores_rating, breaks = 20, xlab = "review_scores_rating",
main = "Histogram of review_scores_rating")
```




```
prop_above4 <- mean(airbnb$review_scores_rating > 4, na.rm = TRUE)
prop_above4
```

```
## [1] 0.9900361
```

From the result, we can see that about 99% of listings have overall ratings of above 4.

Run simple regression review_scores_rating = β0 + β1megahost_i + β2boutique_i + u_i

```
# Set the levels of host_type
airbnb$host_type <- factor(airbnb$host_type, levels =
c("individual","megahost","boutique"))

# Run the simple regression
model <- lm(review_scores_rating ~ host_type, data = airbnb)
summary(model)
```

```
##
## Call:
## lm(formula = review_scores_rating ~ host_type, data = airbnb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86472 -0.09472  0.05442  0.15442  0.35442
##
## Coefficients:
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.843683   0.003324 1457.03   <2e-16 ***
## host_typemegahost  -0.198103   0.004513  -43.90   <2e-16 ***
## host_typeboutique  -0.078960   0.003894  -20.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2112 on 19668 degrees of freedom
## Multiple R-squared:  0.09297,    Adjusted R-squared:  0.09288
## F-statistic:  1008 on 2 and 19668 DF,  p-value: < 2.2e-16
```

Interpretation:

Estimated coefficient ˆβ1: Compared to individual hosts, megahost listings have predicted review scores about 0.198 units lower, on average, ceteris paribus.

Estimated coefficient ˆβ2: Compared to individual hosts, boutique listings have predicted review scores about 0.079 units lower, on average, ceteris paribus.

Significance:

Since ˆβ1 has an absolute t value of 43.90 > 2, the estimates ˆβ1 is statistically significant and different from zero.

Since ˆβ2 has an absolute t value of 20.28 > 2, the estimates ˆβ2 is statistically significant and different from zero.

Find the expected review score if the host is an individual and check that this equals the sample mean of individual listings' review scores in the data. Also check if the estimated coefficient ˆβ1 equals the difference in sample means between megahost and individual listings' review scores in the data.

```
# Sample mean by type
means_by_type <- airbnb %>%
  group_by(host_type) %>%
  summarise(mean_rating = mean(review_scores_rating, na.rm = TRUE), .groups =
"drop")

mean_individual <- means_by_type %>% filter(host_type == "individual") %>%
pull(mean_rating)
mean_megahost <- means_by_type %>% filter(host_type == "megahost") %>%
pull(mean_rating)
mean_boutique <- means_by_type %>% filter(host_type == "boutique") %>%
pull(mean_rating)

# Compare
coef(model)[1]

## (Intercept)
##    4.843683
```

```
mean_individual
```

```
## [1] 4.843683
```

```
coef(model)[2]
```

```
## host_typemegahost
##        -0.1981031
```

```
mean_megahost - mean_individual
```

```
## [1] -0.1981031
```

From the simple regression, the expected review score if the host is an individual is the value of coefficient β0 = 4.844

As shown by the 2 comparisons,

The expected review score if the host is an individual (β0 = 4.844) equals the sample mean of individual listings' review scores in the data (mean_individual = 4.844).

Similarly, the estimated coefficient ˆβ1 = -0.198 equals the difference in sample means between megahost and individual listings' review scores in the data (mean_megahost - mean_individual = -0.198)

Potential confounding variable 1: room_type

```
distinct_room_types <- airbnb %>% distinct(room_type) %>% arrange(room_type)
distinct_room_types
```

```
## # A tibble: 2 × 1
##   room_type
##   <chr>
## 1 Entire home/apt
## 2 Private room
```

With variable room_type (1 = Entire home/apt; 0 = Private room), we have:

Effect of room_type on host_type: Megahosts more often operate entire homes -> ˆρxw > 0

Effect of room_type on review_scores_rating: Entire homes usually score higher due to more privacy/space -> β2_notes > 0

Since ˆρxw * β2_notes > 0 -> positive bias on review_scores_rating. If room_type is omitted, the estimated slope (ˆβ1) on review_scores_rating in the simple regression is biased upward

Decision: should control for room_type in the multiple regression

Potential confounding variable 2: minimum_nights

Effect of minimum_nights on host_type: Megahosts often impose higher minimum stays for easier scheduling/utilization -> ˆ ρxw > 0

Effect of minimum_nights on review_scores_rating: the minimum-night rule should not affect the actual stay quality; the reviewer already accepted it. No direct effect on the rating components. -> β2_notes = 0

Since β2_notes = 0, controlling for minimum_nights does not help in reducing bias, but will lead to higher variances in ˆ β1 since host_type is correlated with minimum_nights.

Decision: do not control for minimum_nights in the multiple regression

Potential confounding variable 3: host_identity_verified

Effect of host_identity_verified on host_type: professional/large hosts are more likely to complete verification -> ˆ ρxw > 0

Effect of host_identity_verified on review_scores_rating: whether the platform verified the host's ID should not change the guest experiences. After staying, ratings should not systematically depend on this -> β2_notes = 0

Since β2_notes = 0, controlling for host_identity_verified does not help in reducing bias, but will lead to higher variances in ˆ β1 since host_type is correlated with host_identity_verified.

Decision: do not control for host_identity_verified in the multiple regression

Run multiple regression controlling for 1 additional variable: room_type

```
# Set the levels of room_type
airbnb$room_type <- factor(airbnb$room_type, levels = c("Private room",
"Entire home/apt"))

# Run the regression
model_multiple <- lm(review_scores_rating ~ host_type + room_type,
                     data = airbnb)

summary(model_multiple)

##
## Call:
## lm(formula = review_scores_rating ~ host_type + room_type, data = airbnb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85857 -0.09566  0.05143  0.14694  0.35694
##
## Coefficients:
```

```
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  4.858296   0.004216 1152.382  < 2e-16 ***
## host_typemegahost           -0.195544   0.004532  -43.146  < 2e-16 ***
## host_typeboutique           -0.080030   0.003896  -20.542  < 2e-16 ***
## room_typeEntire home/apt -0.019697   0.003499   -5.629 1.84e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2111 on 19667 degrees of freedom
## Multiple R-squared:  0.09443,    Adjusted R-squared:  0.09429
## F-statistic: 683.6 on 3 and 19667 DF,  p-value: < 2.2e-16
```

Interpretation:

Estimated coefficient ˆβ1: Compared to individual hosts, megahost listings have predicted review scores about 0.196 units lower, on average, ceteris paribus.

Estimated coefficient ˆβ2: Compared to individual hosts, boutique listings have predicted review scores about 0.080 units lower, on average, ceteris paribus.

Significance:

Since ˆβ1 has an absolute t value of 43.146 > 2, the estimates ˆβ1 is statistically significant and different from zero.

Since ˆβ2 has an absolute t value of 20.542 > 2, the estimates ˆβ2 is statistically significant and different from zero.

Comments: The simple regression earlier gave roughly: ˆβ1 = −0.198, and ˆβ2 = −0.079.

Controlling for room_type, ˆβ1 moved upwards (less negative) to −0.196 (change ≈ +0.002), and ˆβ2 moved downwards (more negative) to −0.080 (change ≈ -0.001).

The direction of impact of megahost status on review rating has not changed: it remains negative.

Run multiple regression controlling for additional confounding variables

In this multiple regression, I choose to control for 1 additional variable: price

I found price to be a potential (highly plausible) confounding variable, since megahosts tend to charge higher (positive correlation between host_type and price), and higher price tends to come with tougher reviews (positive correlation between price and review_scores_rating).So I expect that omitting price can introduce a megative bias.

```r
# Price is originally a char value, containing "$". So we convert price into
numeric values
airbnb <- airbnb %>%
  mutate(
    price = parse_number(price)
  )
```

```r
# Run the regression
model_multiple_2 <- lm(review_scores_rating ~ host_type + room_type + price,
data = airbnb)

summary(model_multiple_2)
```

```
##
## Call:
## lm(formula = review_scores_rating ~ host_type + room_type + price,
##     data = airbnb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85124 -0.09369  0.05018  0.14297  0.40879
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.840e+00  4.218e-03 1147.30   <2e-16 ***
## host_typemegahost         -2.124e-01  4.514e-03  -47.05   <2e-16 ***
## host_typeboutique         -8.707e-02  3.847e-03  -22.64   <2e-16 ***
## room_typeEntire home/apt -5.657e-02  3.751e-03  -15.08   <2e-16 ***
## price                      2.968e-04  1.194e-05   24.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2078 on 19666 degrees of freedom
## Multiple R-squared:  0.122,  Adjusted R-squared:  0.1219
## F-statistic: 683.4 on 4 and 19666 DF,  p-value: < 2.2e-16
```

Interpretation:

Estimated coefficient $\hat{\beta}1$: Compared to individual hosts, megahost listings have predicted review scores about 0.212 units lower, on average, ceteris paribus.

Estimated coefficient $\hat{\beta}2$: Compared to individual hosts, boutique listings have predicted review scores about 0.087 units lower, on average, ceteris paribus.

Significance:

Since $\hat{\beta}1$ has an absolute t value of 47.05 > 2, the estimates $\hat{\beta}1$ is statistically significant and different from zero.

Since $\hat{\beta}2$ has an absolute t value of 22.64 > 2, the estimates $\hat{\beta}2$ is statistically significant and different from zero.