Airbnb - Review Score Analysis

Kitu Komya

December 20, 2017

Introduction

In this short group project, my part of the project was to analyze how a host's amenities and attributes affect her review score rating. In this report are the following sections: R Code, Method, Analysis Findings, and Conclusions. I hope you will enjoy reading through this report as much as I did doing it!

R Code

```r
# load packages
library(readr)
library(stringi)
library(dplyr)
library(tidyr)
library(plyr)
library(ggplot2)
library(knitr)

# load in data
listings.gz <- read_csv(file = "listings.csv.gz")

# look at raw data
head(listings.gz)
summary(listings.gz)


# ------ clean dataframe ------

# clean listings.gz to use only relevant variables
list2 <- listings.gz[ , c(49:50, 52:59, 80)]

# clean amenities variable
list2$amenities <- stri_replace_all_regex(str = list2$amenities, pattern = "\\{", replacement = "")
list2$amenities <- stri_replace_all_regex(str = list2$amenities, pattern = "\\}", replacement = "")
list2$amenities <- stri_replace_all_regex(str = list2$amenities, pattern = "\"", replacement = "")

# use complete case
list2$amenities <- ifelse(list2$amenities == "", NA, list2$amenities)
list3 <- list2[complete.cases(list2), ]

# look at cleaned and transformed dataset
head(list3)
summary(list3)
```

```
# ------ analyze attributes ------

# linear regression on attributes
fit <- lm(review_scores_rating ~ property_type + room_type + accommodates +
            bathrooms + bedrooms + beds + bed_type, data = list3)
summary(fit)

# dataframes of significant attributes
attributes <- c("property type: Bungalow", "property type: Condominium", "property type: Guesthouse",
                "property type: House", "property type: Loft", "property type: Townhouse",
                "room type: Private Room", "room type: Shared Room", "accommodates", "bedrooms",
                "beds", "bed type: Futon", "bed type: Real Bed")
estimates <- c(2.33567, 2.11278, 3.35298, 1.50701, 2.65329, 1.30589, -0.56145, -4.06204, -0.33131,
               0.60600, -0.22101, -1.67170, -2.06934)

# create one dataframe of significant attributes
df <- as.data.frame(attributes, stringsAsFactors = FALSE)
df2 <- as.data.frame(estimates)
df <- cbind(df, df2)

# one for positive, one for negative
df_pos <- df[df$estimates > 0, ]
df_neg <- df[df$estimates < 0, ]

# make bar plot of positive attributes
a <- ggplot(data = df_pos, aes(x = reorder(attributes, estimates), y = estimates)) + geom_bar(stat = "i
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) + ylab("Change in Review Score") +
  xlab("Host's Attributes") + ggtitle("Attributes of Hosts that Significantly and Positively \nAffect R
  theme(plot.title = element_text(hjust = 0.5))

# make bar plot of negative attributes
b <- ggplot(data = df_neg, aes(x = reorder(attributes, -estimates), y = estimates)) + geom_bar(stat = "
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) + ylab("Change in Review Score") +
  xlab("Host's Attributes") + ggtitle("Attributes of Hosts that Significantly and Negatively \nAffect R
  theme(plot.title = element_text(hjust = 0.5))


# ------ analyze amenities ------

# give df IDs
list3$id <- 1:nrow(list3)

# split amenities' list into multiple variables
dat <- with(list3, strsplit(amenities, ','))
df2 <- data.frame(id = factor(rep(list3$id, times = lengths(dat)),
                              levels = list3$id), amenities = unlist(dat))
df3 <- as.data.frame(cbind(id = list3$id, table(df2$id, df2$amenities)))

# get new df of just reviews
reviews <- as.data.frame(list3$review_scores_rating)

# merge amenities with reviews
amenities <- cbind(df3, reviews)
```

```r
# clean amenities
names(amenities)[94] <- "review_scores_rating"
amenities <- amenities[ , -1]

# fit linear regression on amenities
fit2 <- lm(review_scores_rating ~ ., data = amenities)
summary(fit2)

# dataframes of significant amenities
amen <- c("24-hour check-in", "Cable TV", "Children's books and toys", "Dryer", "Elevator in building",
          "First aid kit", "Free parking on premises", "Game console", "Hair dryer",
          "Hangers", "Heating", "Hot tub", "Indoor fireplace", "Laptop friendly workspace",
          "Pets allowed", "Pets live on this property", "Pool", "Private entrance", "Safety card",
          "Shampoo", "Smoking allowed", "TV", "Washer")
values <- c(-0.56896, 0.47176, 2.00865, 1.20878, -0.39856, 0.80373, 0.62664, -2.29282, 0.54309,
            0.46016, 0.89403, -0.68816, 0.54685, 0.75075, -0.65192, 1.56824, 0.34171, 0.50053,
            -0.70855, 0.29488, -0.82392, 0.57804, -0.74120)

# create only 1 df of significant amenities
df_amen <- as.data.frame(amen, stringsAsFactors = FALSE)
df2_amen <- as.data.frame(values)
df_amen <- cbind(df_amen, df2_amen)

# create pos and neg df
amen_pos <- df_amen[df_amen$values > 0, ]
amen_neg <- df_amen[df_amen$values < 0, ]

# make bar plot of positive amenities
c <- ggplot(data = amen_pos, aes(x = reorder(amen, values), y = values)) + geom_bar(stat = "identity", 
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) + ylab("Change in Review Score") +
  xlab("Host's Amenities") + ggtitle("Amenities of Hosts that Significantly and Positively \nAffect Rev
  theme(plot.title = element_text(hjust = 0.5))

# make bar plot of negative amenities
d <- ggplot(data = amen_neg, aes(x = reorder(amen, -values), y = values)) + geom_bar(stat = "identity",
  theme(axis.text.x = element_text(angle = 30, hjust = 1)) + ylab("Change in Review Score") +
  xlab("Host's Amenities") + ggtitle("Amenities of Hosts that Significantly and Negatively \nAffect Rev
  theme(plot.title = element_text(hjust = 0.5))
```

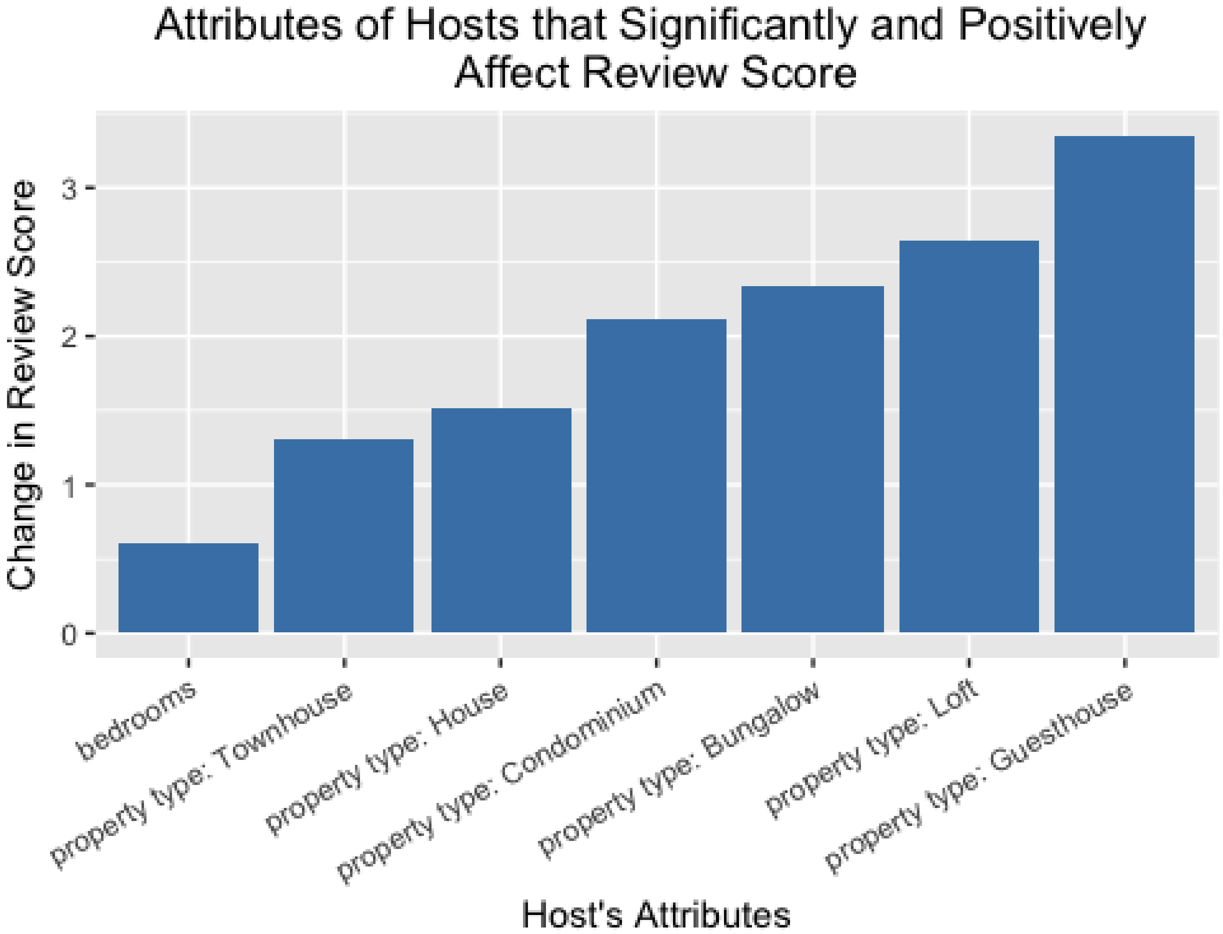
Method

I wanted to explore the question: how are ratings affected by what a host can offer? This question directly translates to: what should a host do to improve her ratings?

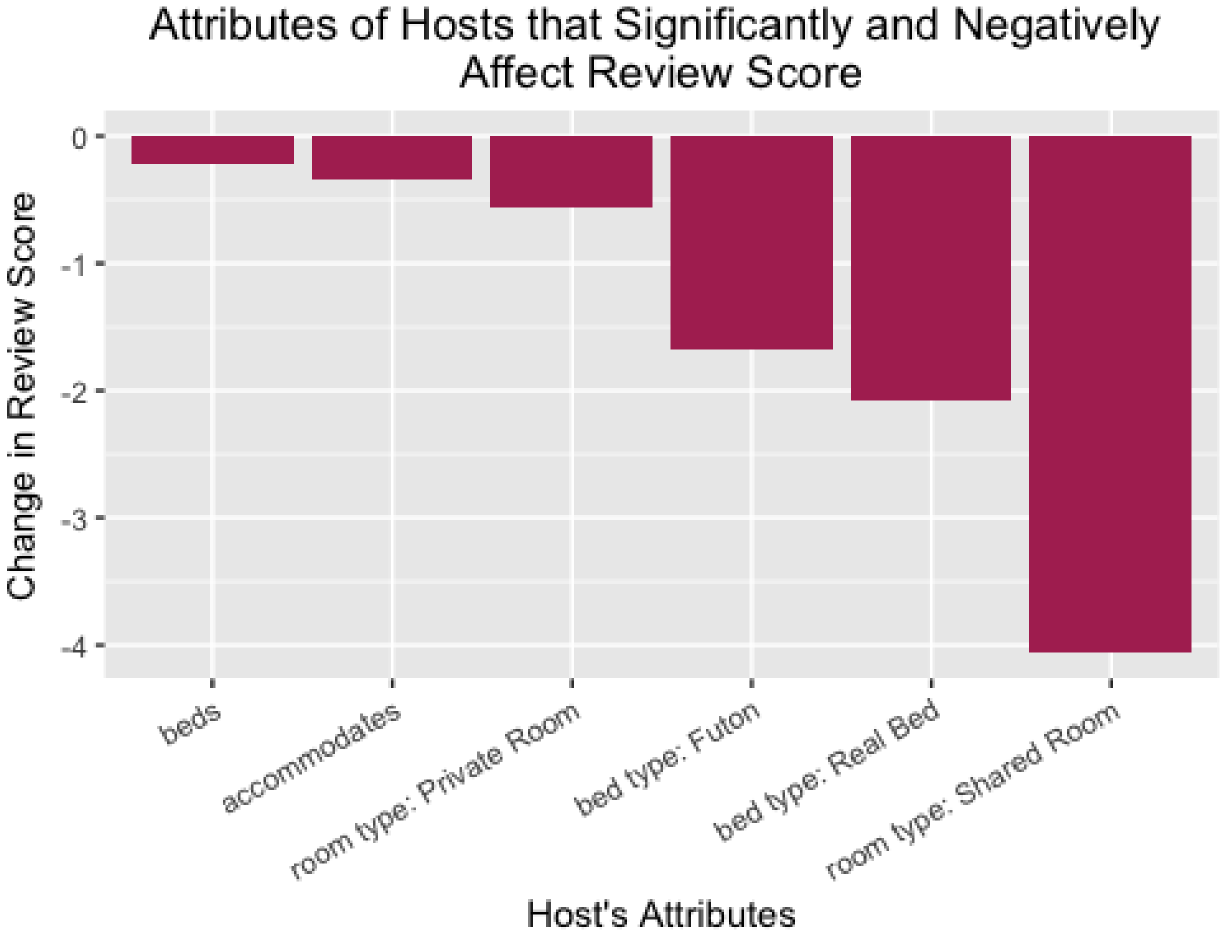
Thus, I looked into attributes and amenities. Attributes were all in separate variables and was thus easy to explore since I could conduct linear regression on all of the variables. Amenities, however, was in one single variable. I soon realized that the variable was in list form and held multiple strings of factor type. I then did some data wrangling to separate this one variable into all distinct string type variables present in the amenities, and then used binary 0 or 1 to indicate whether each property contained it. This was the most challenging part of my process - data wrangling into a different data structure.

I then made two linear regression models: one that tried to explain the relationship between all attributes and review score, and one between amenities and review score. From this, I only chose those attributes and amenities that were significant at a 95% level. Finally, I made 4 barplots of statistically significant and positive as well as statistically significant and negative attributes and amenities in relation to review score.

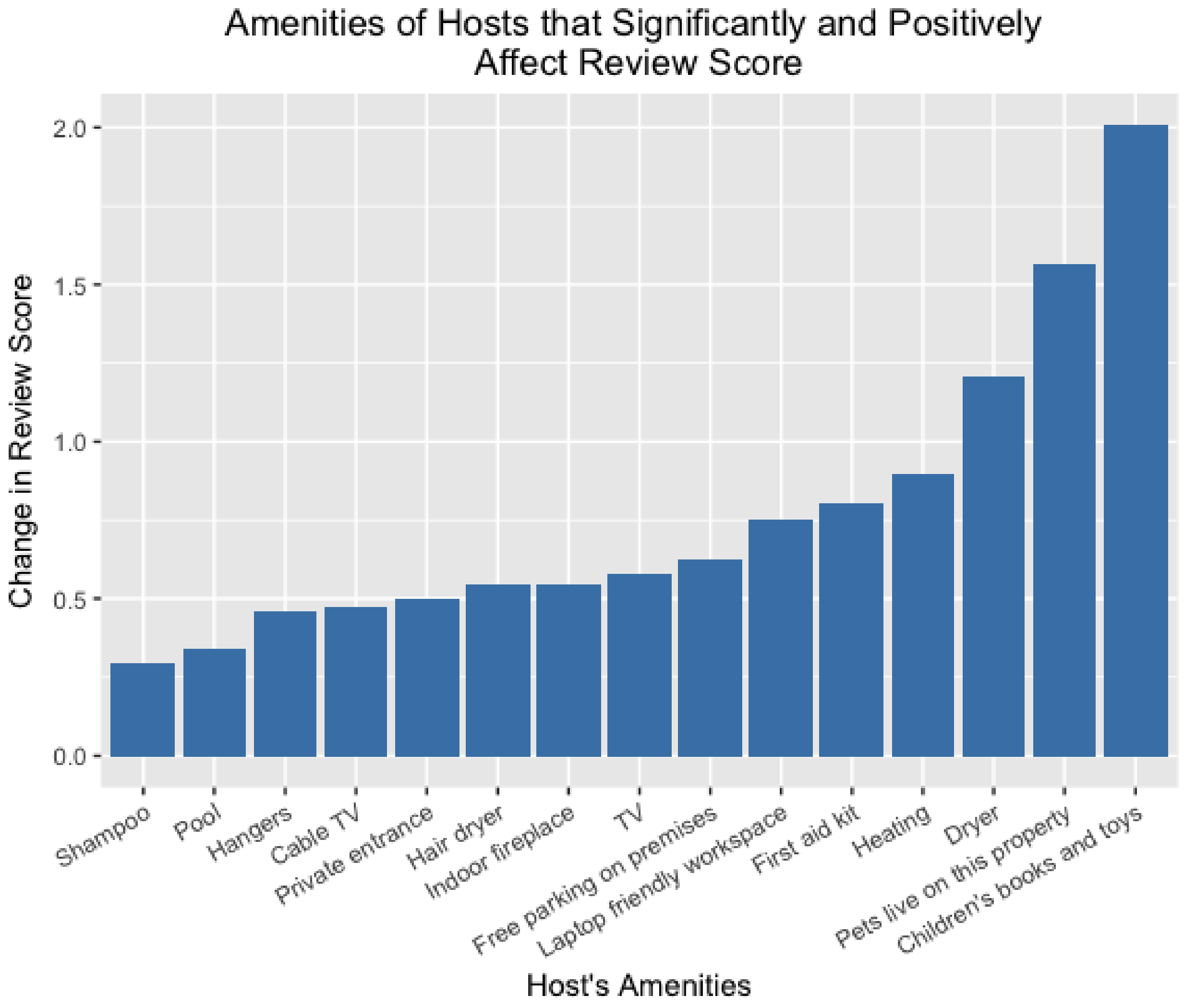
Analysis Findings



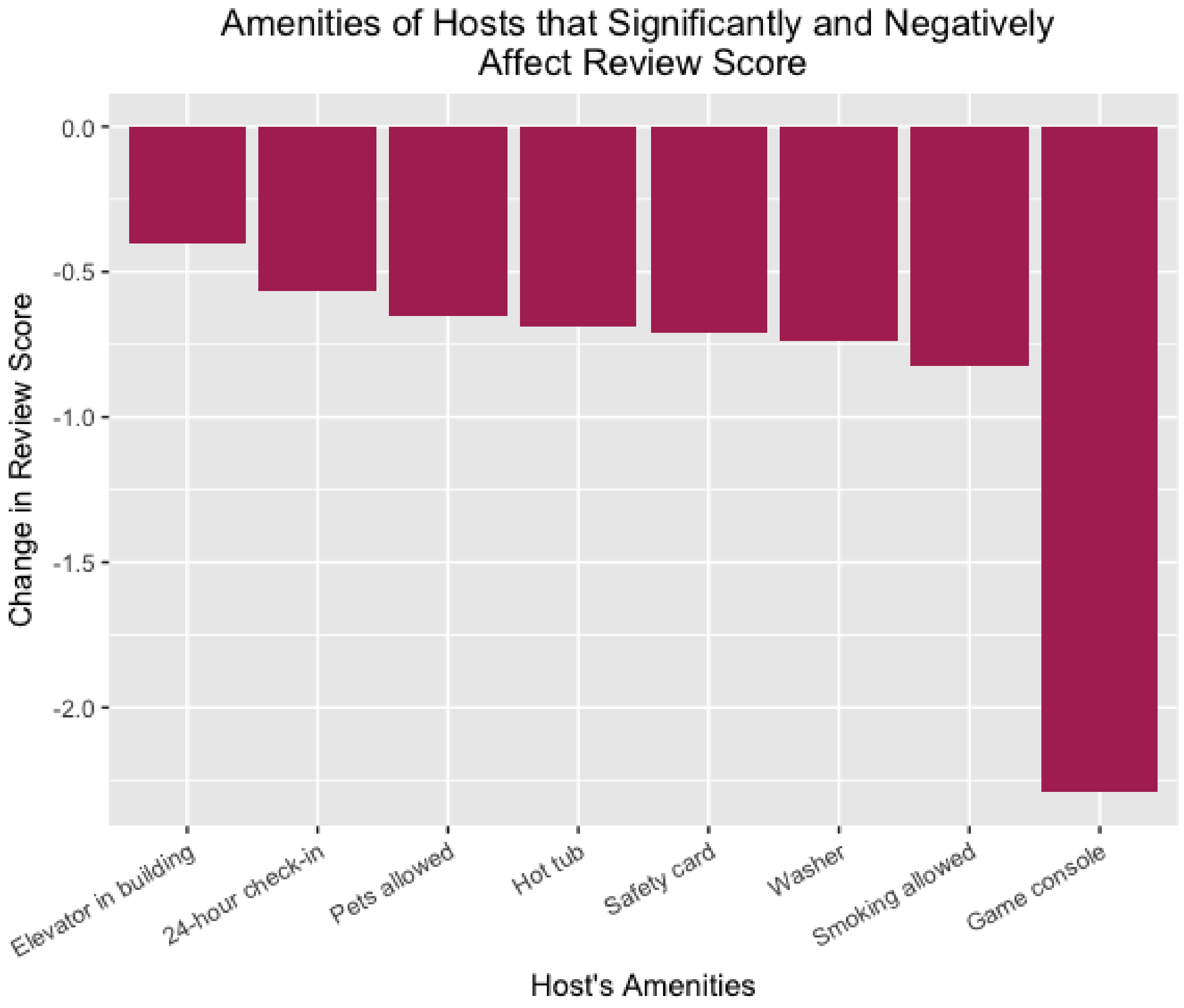
Attributes that are significant and positive mainly relate to property type. More “exotic” or vacation-type or non-conventional property types, such as bungalow, loft, guesthouse, and condo, are in this category. Renters perhaps have a better experience in such properties they usually don’t reside in or experience regularly.



Attributes that are significant and negative include those that relate to living with more people. Such attributes include accommodates, beds, futon, and shared room. Interestingly, both room type: Private room and room type: Shared room are in this category, which I found strange at first glance - both private and shared rooms are correlated to more negative reviews? However, upon further exploration of the data, I realized that there is a third room type: Entire house option as well, which makes sense - renters probably have a better experience when they rent out an entire house instead of just one room, whether it be private or shared. Thus, many of the findings need to be complemented by the other data available by Airbnb.



Amenities that are significant and positive include common items you would find in a house. Such amenities include shampoo, hangers, cable TV, first aid kit, private entrance, and indoor fireplace. These amenities lead me to believe that most renters are looking for a more private and homely visit than a public and guest-like one. Cutely, children's books and toys also fall into this category - hotels, for instance, don't carry these, which is probably why renting at a place that also caters to a family's kids is very attractive to renters. Also adorably, "pets live on this property" is significantly positive, but "pets allowed" is significantly negative. This goes to show that existing pets are welcome, but public pets are not.



Amenities that are significant and negative include those that are conventional when staying at a building rather than at a home. Such amenities include elevator, 24-hour check in, and safety card, which are usually in buildings like hotels where there is less of a "home-like" feel and more of a "guest" one. Other environmental amenities like "smoking allowed" and "pets allowed" are in this category. Again, these are more common in public buildings to private ones. Interestingly, game console, washer, and hot tub are also here. Perhaps other renters were using the game console too much which contributed to a negative review score. As for the other two, more research and insight need to be conducted!

Conclusions

1. Renters generally look for a private and "home-like" stay instead of a public and "guest-like" one.
2. A prospective host can use these findings to figure out how to improve her review scores.
3. A prospective renter can use these findings to estimate how her stay will be.

Future Analysis

4. Complement analysis with other data that is available from Airbnb to understand how a renter may have reviewed her stay.
5. Explore psychological insight as to why some amenities and attributes are negative or positive.