# Paris Airbnb EDA*

Yuchen Chen

2024-03-05

## 1 Introduction

In this paper we have considered exploratory data analysis (EDA) by using tool (R Core Team 2022) to reformat (Wickham et al. 2019) the dataset of Airbnb listings in Paris on December 2023. We focused on missing data, the distributions of variables, and the relationships between variables.

## 2 Distribution and properties of individual variables

```
airbnb_data_selected <-
  airbnb_data_selected |>
  mutate(
    price = str_remove_all(price, "[\\$,]"),
    price = as.integer(price)
  )
```

Initially, our focus might be on the price, which is currently stored as a text. This is a frequent issue, and we must ensure that it doesn't simply result in a conversion to missing values (NAs). Merely coercing the price variable into a numeric format could lead to NA values, because numerous text elements, like the dollar sign "$", don't have a direct numeric equivalent. Therefore, we need to strip away these characters before proceeding with the conversion.

*Code and data are available at: https://github.com/Victor1114/Paris-Airbnb-EDA.git

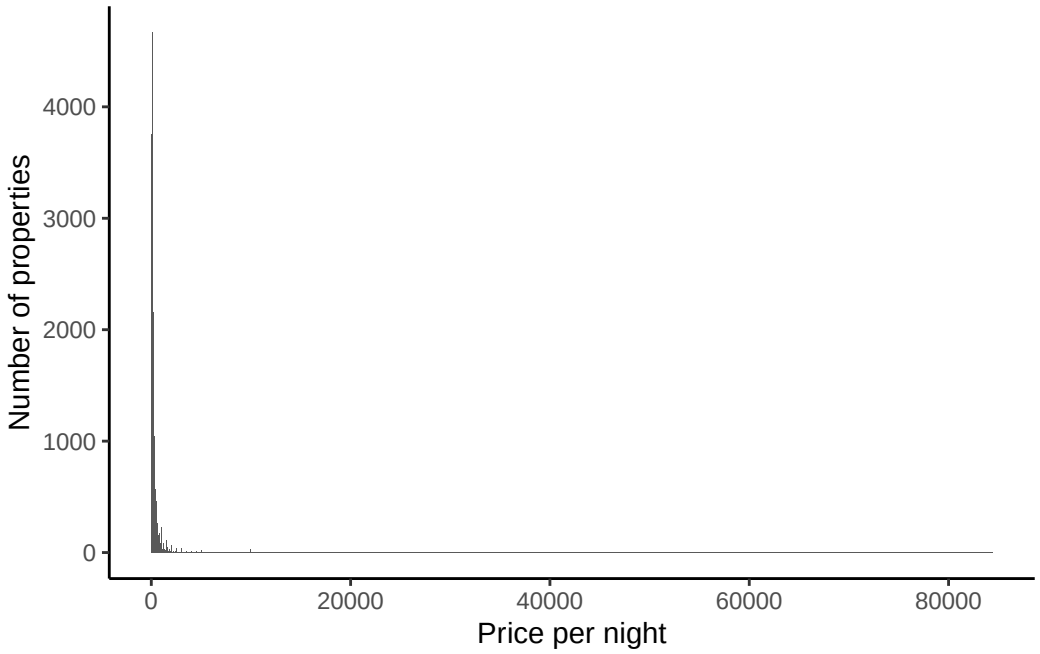


Figure 1: Distribution of prices

## 3 Price per night

Since Figure 1 have many outliers, so clean the dataset by only consider the price per night under $900.

```
airbnb_data_less_900 <-
  airbnb_data_selected |>
  filter(price < 900)

airbnb_data_no_superhost_nas <-
  airbnb_data_less_900 |>
  filter(!is.na(host_is_superhost)) |>
  mutate(
    host_is_superhost_binary =
      as.numeric(host_is_superhost)
  )
```

Then removing all prices that more than $899 and anyone with a NA for whether they are a superhost.

`number_of_reviews`

0
13335

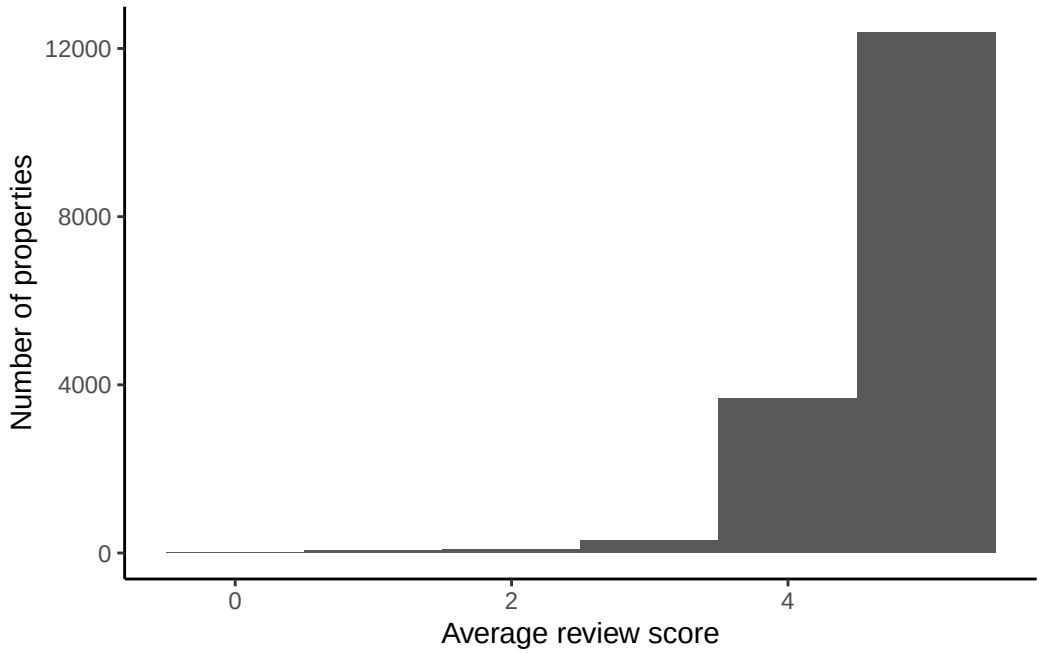


Figure 2: Distribution of review scores for properties with NA response time, for Paris Airbnb rentals in December 2023

Next, we consider the star ratings which five is the highest and zero is lowest. From Figure 2, it is obviously that most of people rate between 4 to 5 which is high evaluate for cleanliness, accuracy, value, and others. During this process, we delet the NAs in "review_scores_rating".

## 4 Host respond time and review score accuracy

Figure 3 was drawing by using (Tierney and Cook 2023) which showing the relationship between the host response time and the review score for listings in the Airbnb dataset. We can find out that most listings have review score around 4 to 5, which suggests a generally high level of accuracy in listings' descriptions. There is a clear relationship of points towards the higher end of the review score accuracy, especially for hosts who respond within an hour or a few hours. Hence, there is a potential correlation which faster response times might make higher review scores accuracy. What's more, people might don't want to give review score if host not respond, which cause many missing in NA of response time.

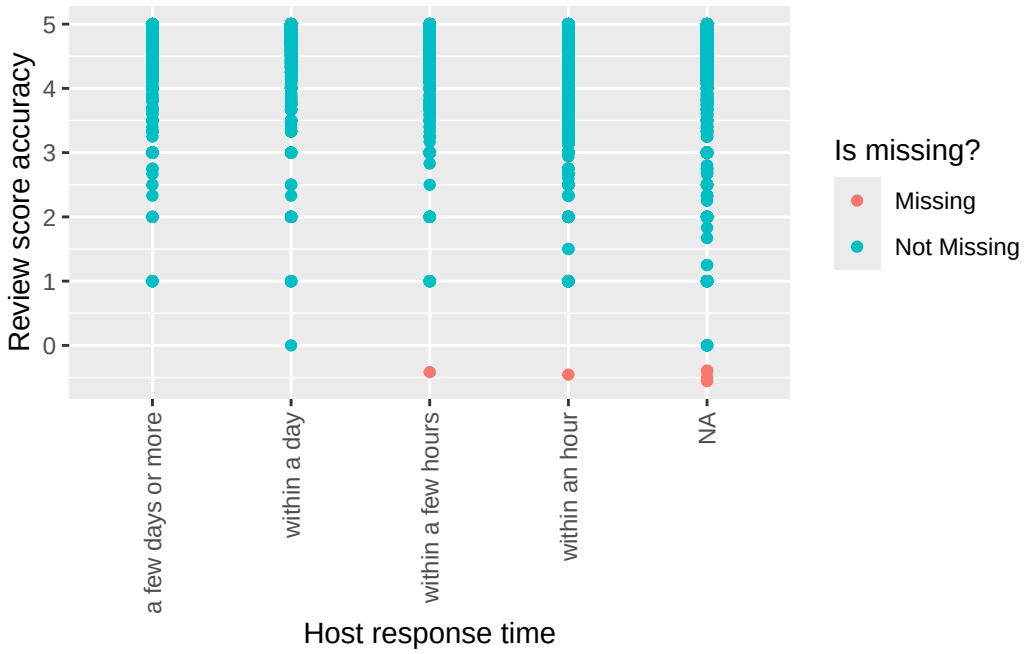


Figure 3

## 5 Relationship

Figure 4 visualizes data for Paris Airbnb rentals in December 2023. First, it displaying most of people choose low price listings with a broad range of average review scores predominantly between 4 and 5, indicative of high guest satisfaction. Because of the price per night is plotted along the x-axis, there is not evident to have correlation between price and guest satisfaction. The superhosts are evenly distributed in different price per night with high average review score, which illustrates the correlation between superhosts and high average review score. Moreover, if the price per night less than $250, people easier to rate the review score under 4 points and with the price increasing, the trend is dcreasing.

## 6 Respond time and is superhost

Table 1 was drew by using (Firke 2023) to demonstrate that a host does not respond within an hour then it is unlikely that they are a superhost.

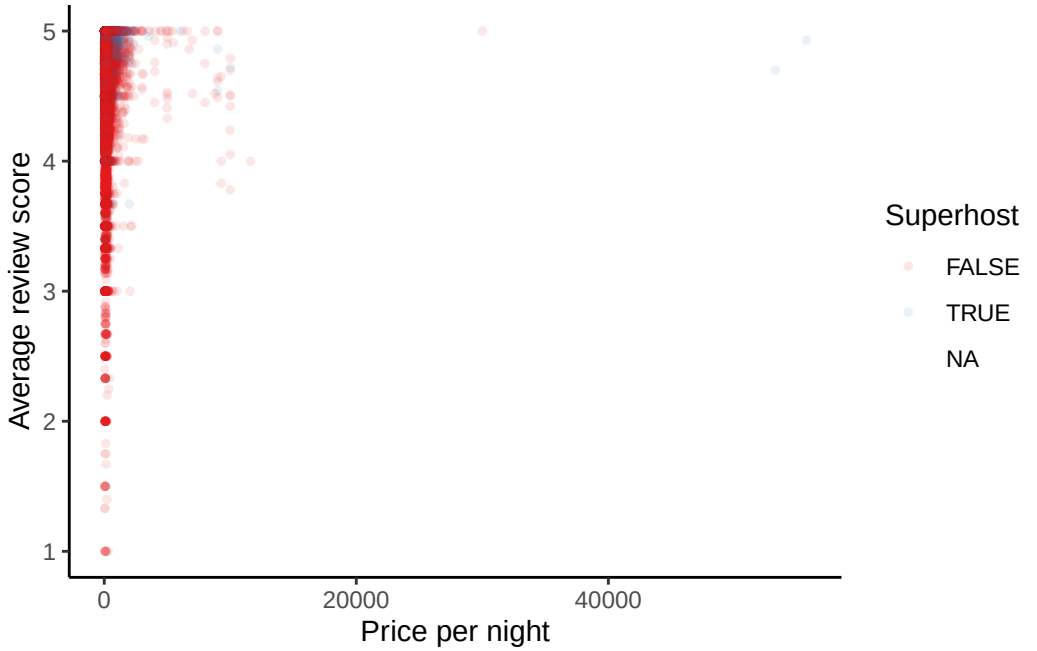


Figure 4: Relationship between price and review and whether a host is a superhost, for Paris Airbnb rentals in December 2023

Table 1

| | host_is_superhost | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| host_response_time | FALSE | | TRUE | | NA_ | | | |
| a few days or more | 3% | (1,963) | 0% | (26) | 0% | (0) | | |
| N/A | 47% | (29,147) | 8% | (935) | 17 | | | (15) |
| within a day | 9% | (5,743) | 9% | (1,029) | 3% | (3) | | |
| within a few hours | 10% | (5,975) | 20% | (2,288) | 16 | | | (14) |
| within an hour | 32% | (19,734) | 63% | (7,390) | 64 | | | (58) |
| <NA> | 0% | (8) | 0% | (1) | 0% | (0) | | |

## 7 Results

-More travelers like listings prices less than $250,but lower price listings may not good as expect. -Review score or guest satisfaction is nothing about price per night for listings. -Hosts' Respond time is less might get higher review score and short respond time have more possibility for host is superhost. -Most of people satisfy about the experience of Airbnb.

## 8 Reference


http://insideairbnb.com/get-the-data

https://r4ds.had.co.nz/exploratory-data-analysis.html

https://medium.com/mlearning-ai/basic-exploratory-data-analysis-template-for-regression-problems-20ca00c58f7d

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Tierney, Nicholas, and Dianne Cook. 2023. "Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations." *Journal of Statistical Software* 105 (7): 1–31. https://doi.org/10.18637/jss.v105.i07.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.