Analysis of Airbnb in Paris, France*

Geunchul Shin

March 3, 2024

1 Introduction

In this paper, we will be looking at Airbnb listings in Paris, France (12th, December 2023). We will seek the distribution and properties of individual variables of the data. Also we will find possible relationships between variables.

2 Data

Data was collected and analyzed with the utilization of the programming software, R Studio (R Core Team 2022). Along with that, the following packages are also used, ggplot2 for plotting (Wickham 2016), tidyverse for data frame manipulation (Wickham et al. 2019), knitr for pdf rendering (Xie 2014), naniar for summarizing (Tierney and Cook 2023), janitor for cleaning data (Firke 2023), modelsummary for creating tables (Arel-Bundock 2022) and arrow (Richardson et al. 2024).

The dataset is from Inside Airbnb which we saved as a local copy. The name of the file is 'airbnb_data.csv'. This file will be added to gitignore as we do not want this to be pushed to GitHub due to its massive size. Speaking of the size, now we will create a parquet file with our favoured variables. We selected important aspects of Airbnbs; such variables such as price, superhost, number of reviews and etc.

After selecting our favoured variables, we cleaned the data. One cleaning process involved with removing dollar sign ($) from the price as we want to keep the variable to be numeric.

*Code and data are available at: https://github.com/geunchulshin/Airbnb_paris_france

3 Distribution and properties of individual variables

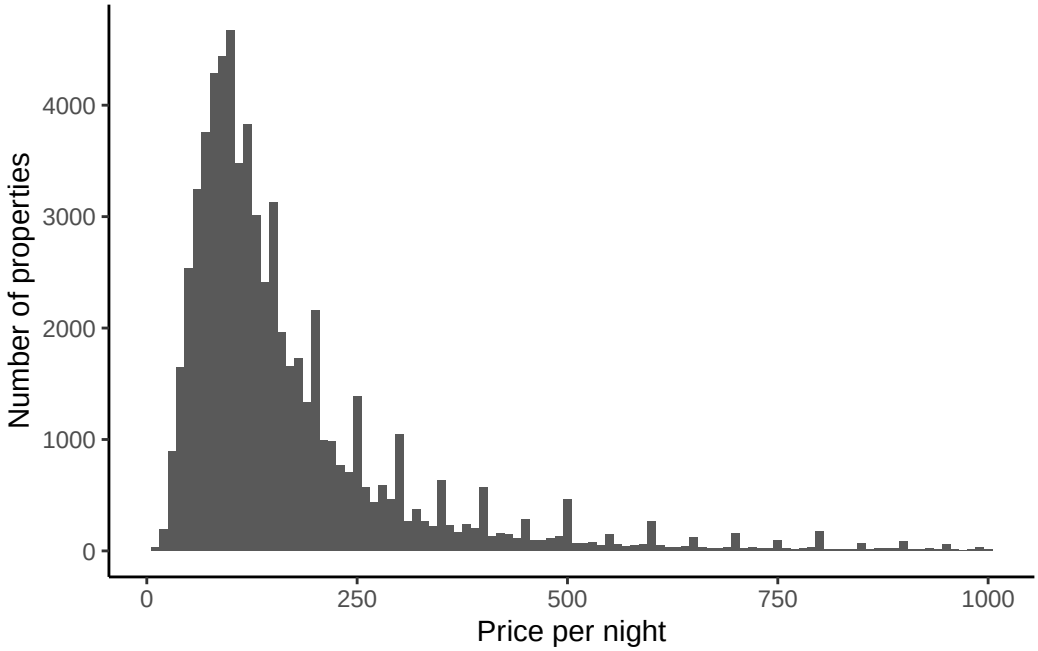

Figure 1: Distribution of prices (less than $1,000) for Airbnb listings in Paris in December 2023

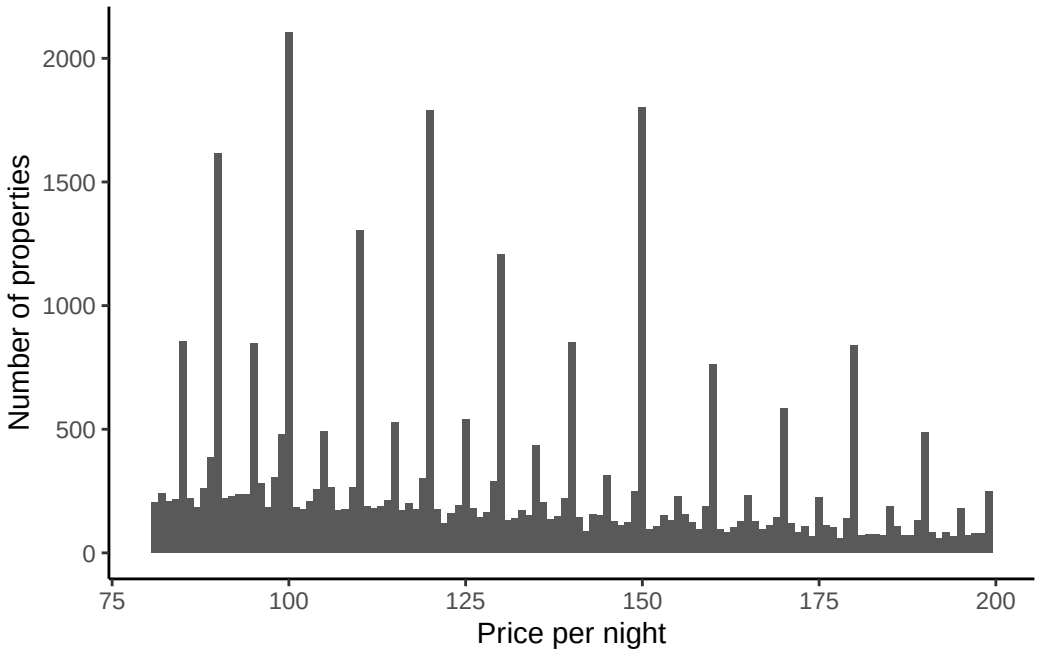

Figure 2: Distribution of prices (between $80 and $200) for Airbnb listings in Paris in December 2023

Let's focus on Airbnbs with *prices* less than 1000. Loooking at Figure 1, we see there are bunching in some prices. If we look Figure 2, in between $80 and $200, it is very clear that there indeed are bunchings on some prices like $120 and $150. From now, we won't bother counting Airbnbs with over $1000 as we consider them as outliers.

Similar to prices, *superhosts* are an important variable when looking at Airbnbs. We got remove of NA values of 'host_is_superhost' variable for clarity. Also, we created a binary variable for efficiency for future usage.

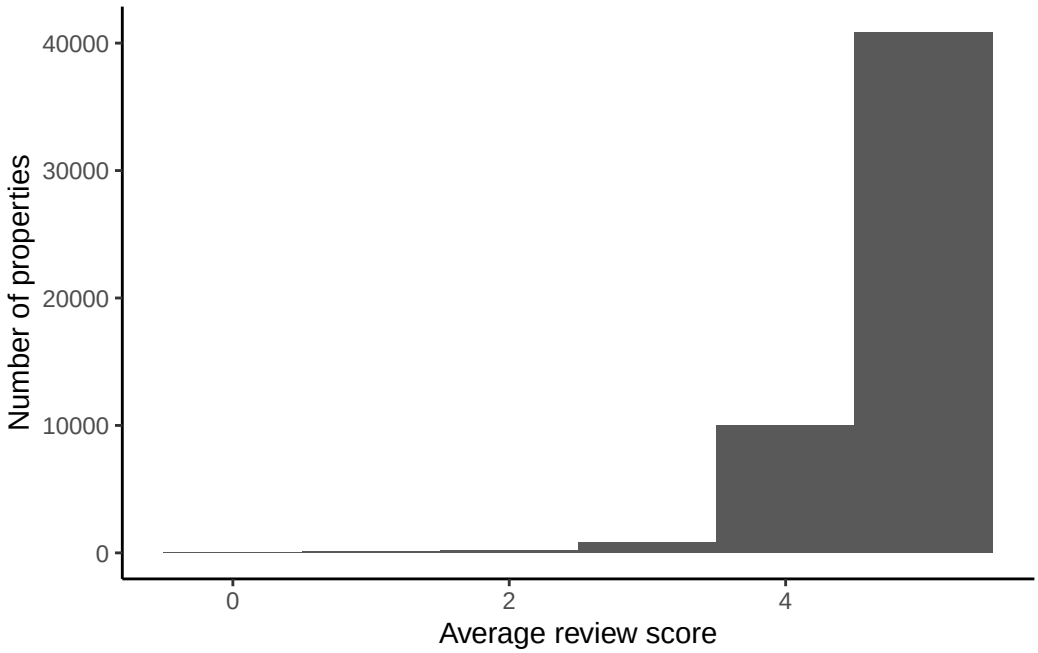

Figure 3: Distribution of review scores for Paris Airbnb rentals in December 2023

Also very similar to superhost property, *review scores* take a big role. We see the distribution of review scores in Figure 3. We also have to take account that there are NAs in this variables. After focusing on non-missing values of the variable, we get Figure 3. The average score seem to be very high with the score of 5.

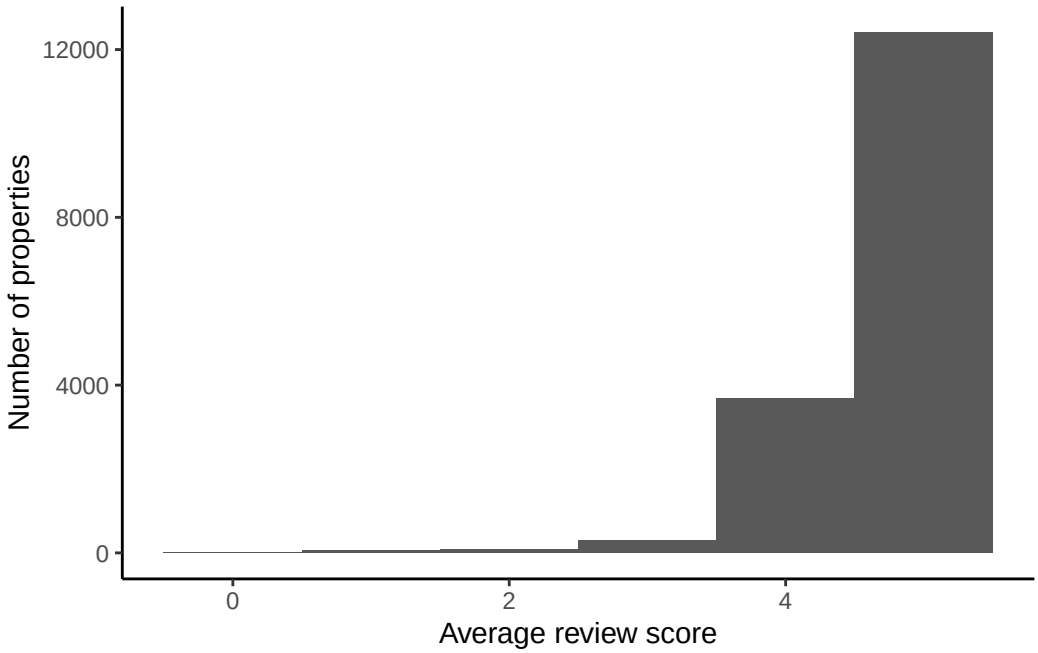

Figure 4: Distribution of review scores for properties with NA response time, for Paris Airbnb rentals in December 2023

A variable we can also consider is Airbnb hosts' response time. But, we found that there are too many NAs in the variable. However, instead of getting rid of the variable, we seeked if there is any possible relation with another variable. As the "NAs" in "host_response_time" were not written properly, we have rewritten it. Then, we checked if there is a relationship with the review score. Here in Figure 4, we constructed distribution of review scores for properties with NA response time.

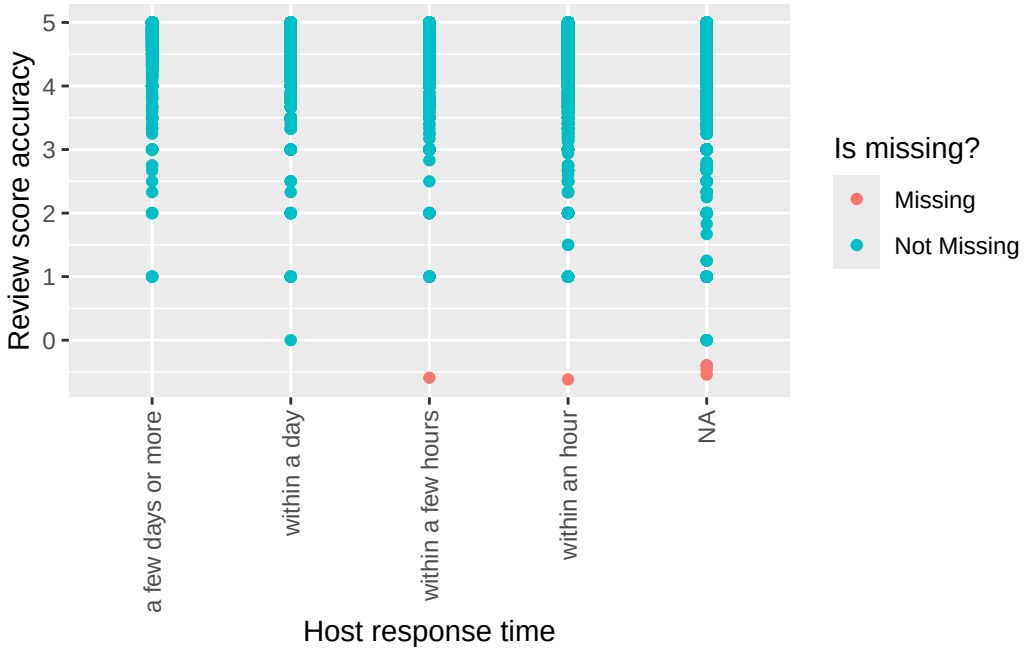

Figure 5: Missing values in Paris Airbnb data, by host response time

It is also interesting how many missing values are being dropped by looking at Figure 5. We removed hosts with NA as their response time and this almost removed 19 percent of the observations.

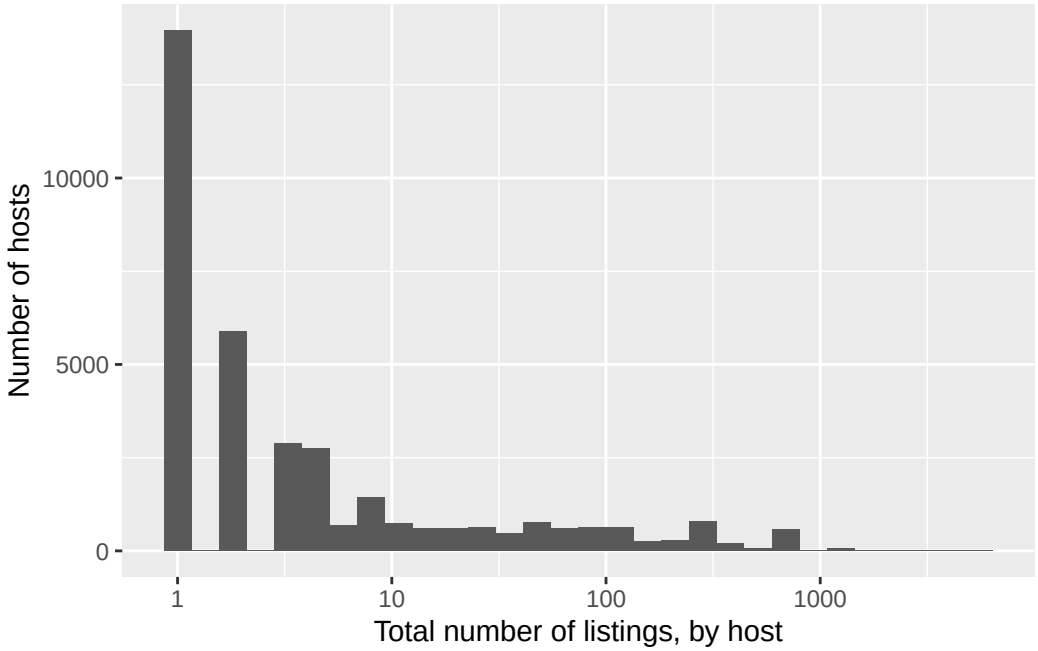

Figure 6: Distribution of the number of properties a host has on Airbnb, for Paris Airbnb rentals in December 2023

We also wondered how many properties a host might have on Airbnb. Here in Figure 6, we see how many properties a host has on Airbnb. It seems that majority of the hosts owned a single property. Therefore, we will put our focus more on these.

4 Relationship between variables

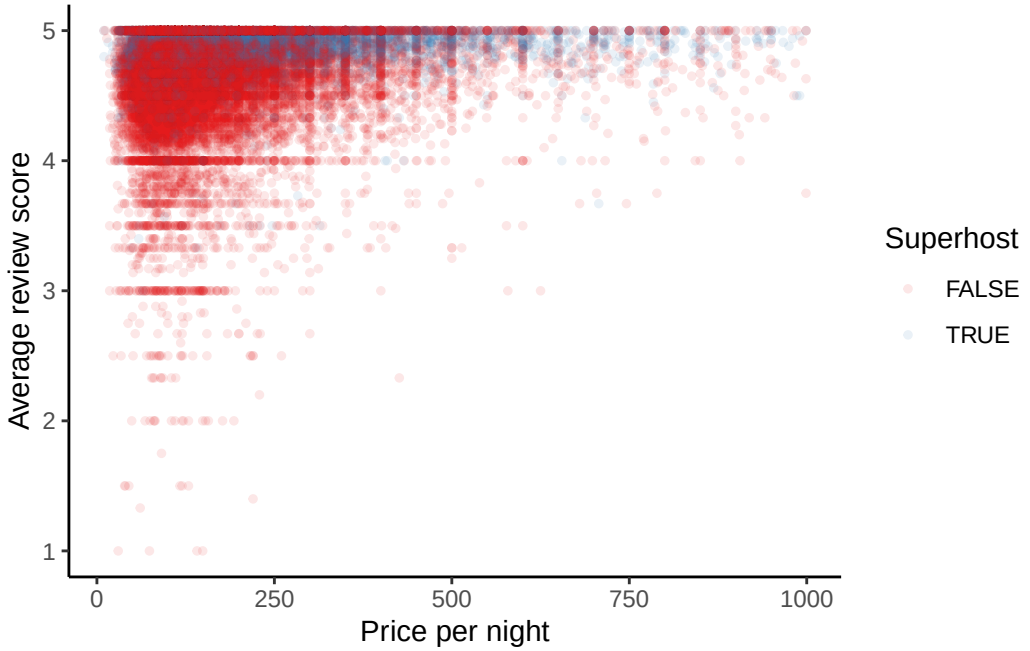

Figure 7: Relationship between price and review and whether a host is a superhost, for Paris Airbnb rentals in December 2023

As seen in Figure 7, we see the relationship between price and review and whether a host is a superhost.

One important aspect of the eligibility of superhost is how fast the host respond. As seen here, we find that more superhosts compare to normal hosts respond within an hour.

Table 1: Table of response time of hosts/superhosts

	host_is_superhost	
host_response_time	FALSE	TRUE
a few days or more	5% (1,219)	0% (24)
within a day	17% (4,326)	10% (971)
within a few hours	18% (4,660)	22% (2,151)
within an hour	60% (15,352)	68% (6,742)

With the findings previously, we use our insight to construct the model that we want to

Table 2: Explaining whether a host is a superhost based on their response time

	(1)
(Intercept)	−18.384
	(0.377)
host_response_timewithin a day	2.283
	(0.210)
host_response_timewithin a few hours	3.015
	(0.209)
host_response_timewithin an hour	3.190
	(0.208)
review_scores_rating	3.021
	(0.065)
Num.Obs.	35 445
AIC	37 601.0
BIC	37 643.4
Log.Lik.	−18 795.504
RMSE	0.43

estimate:

$$Prob(Issuperhost = 1) = logit^{-1}(\beta_0 + \beta_1 Responsetime + \beta_2 Reviews + \varepsilon))))$$

As we see here Table 2, each of the levels have a positive association with the chance of being a superhost.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=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/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

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.

Xie, Yihui. 2014. *Knitr: Elegant Graphics for Data Analysis*. In Victoria Stodden, Friedrich Leisch; Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman; Hall/CRC.