Explore Airbnb listings in Paris, France, as at 04 March 2024*

Yimiao Yuan

March 5, 2024

1 Introduction

This report analyzes the Airbnb listings in Paris, France, as at 04 March 2024 using R (R Core Team (2022)). The dataset is read from the Inside Airbnb (Cox (2021)) website, and cleaned and explored using `tidyverse` (Wickham et al. (2019)), `janitor` (Firke (2023)), `knitr` (Xie (2023)), `lubridate` (Grolemund and Wickham (2011)), `mice` (van Buuren and Groothuis-Oudshoorn (2011)), `modelsummary` (Arel-Bundock (2022)), `ggplot2` (Wickham (2016)) and `naniar` (Tierney and Cook (2023)). This report will explore distribution and properties of different variables and analyze the relationship between them.

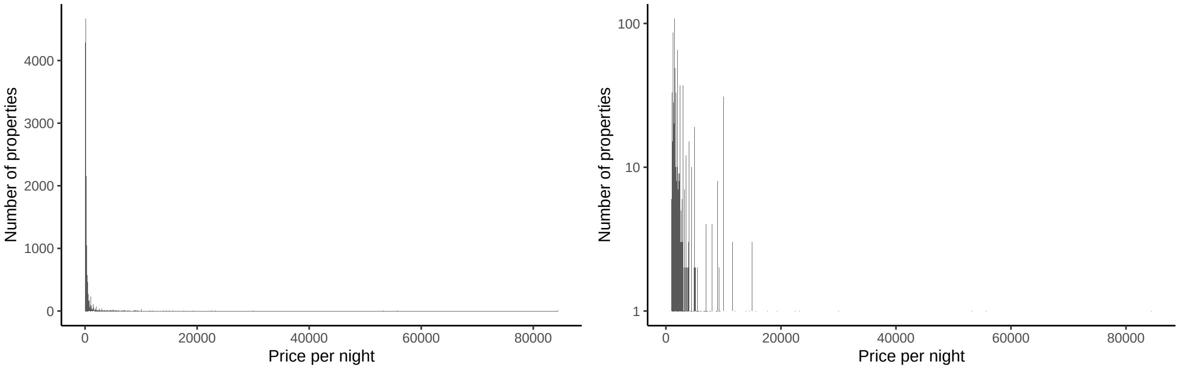
2 Distribution and Properties

The raw dataset is cleaned and used to create a parquet file with selected variables for exploratory purposes. I first look at the price of the Airbnb and plot the distribution in Figure 1. There are outliers on regular scale, so I also use a log scale to plot the distribution. From the graph, we can see that most properties have have low prices, number of properties decreases as the price increases.

Then I turned my focus on price that less than $1000, and the distribution contains more detailed. Figure 2a shows that there are some bunching in the price, so I choose to take a further look at prices between $90 and $250. Figure 2b also shows some bunching in the graph, which might due to the reason that prices around numbers ending in zero or nine.

Next I look at the superhosts in the Airbnb. After removing the NA value in superhost, I plot the distribution of review scores for all superhosts. Figure 3 shows that most of the superhosts' properties have 5 stars, they rarely get 1 or 2 stars.

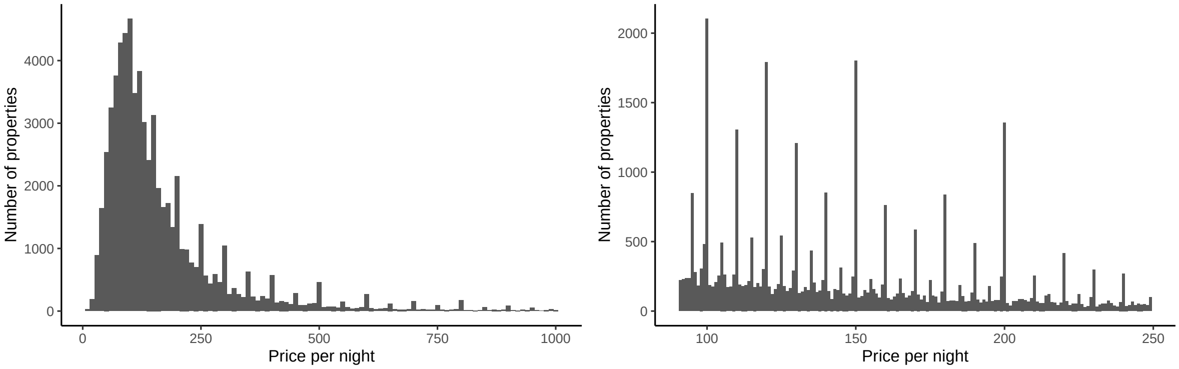
*Code and data are available at: https://github.com/YimiaoYuan09/Airbnb_EDA_Paris



(a) Distribution of prices

(b) Using the log scale for prices more than $1,000

Figure 1: Distribution of prices of Paris Airbnb rentals in March 2024



(a) Prices less than $1,000 suggest some bunching

(b) Prices between $90 and $250 illustrate the bunching more clearly

Figure 2: Distribution of prices of Paris Airbnb rentals in March 2024

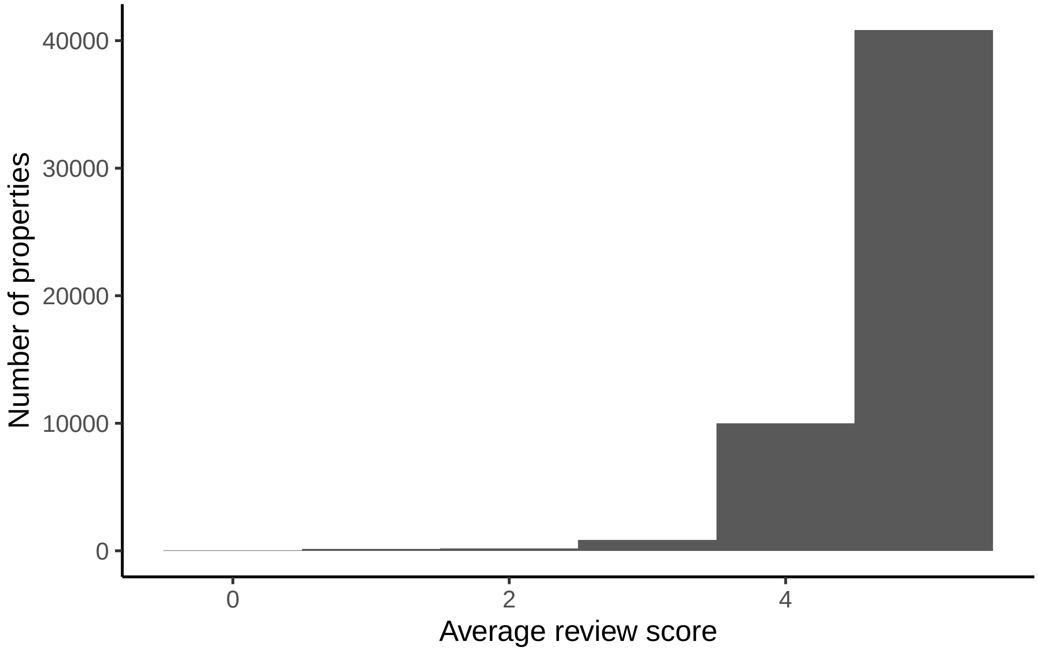


Figure 3: Distribution of review scores for Paris Airbnb rentals in March 2024

There are a lot of NA values in the review scores rating since they do not have enough reviews. For this report, I just remove NA values in review scores rating.

I also interested in the host response time, which is how quickly a host responds to an inquiry. There are also a lot of NAs in this variable, so I want to see that if there is a relationship with the review score. Figure 4 shows that most of properties get a rate over 4 stars. However, `ggplot2` drops a lot of missing values, so I use geom_miss_point() from `naniar` to include them in the graph (Figure 5).

I also interested in how many properties a host has on Airbnb. From Figure 6, we can see that most hosts have 1 properties, some hosts have less than 10 properties. Few of them have more than 100 properties, which is a little bit strange.

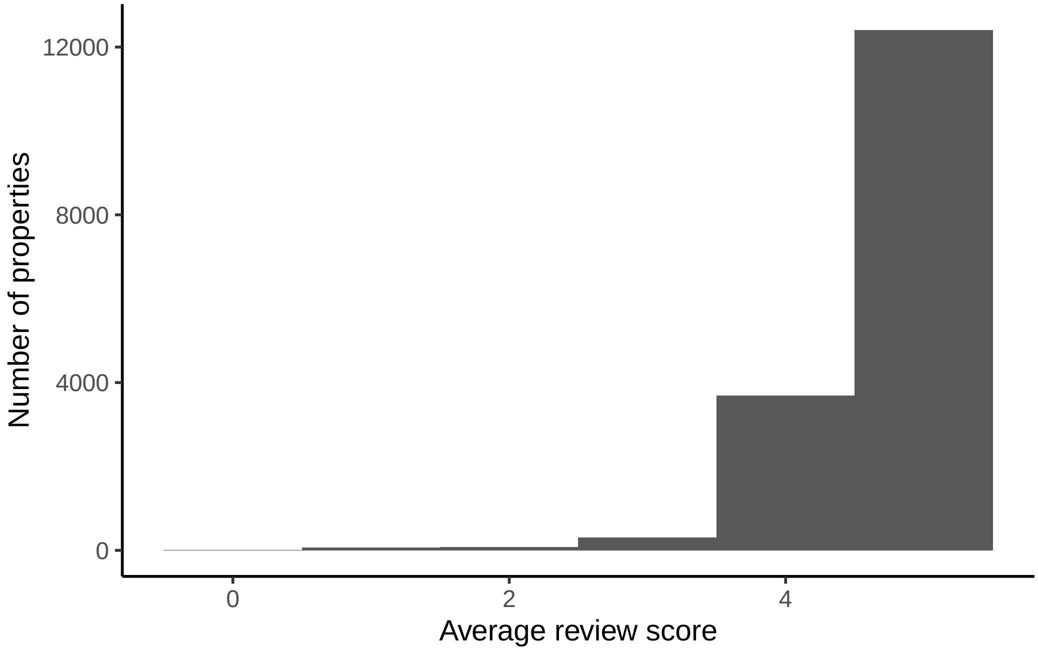


Figure 4: Distribution of review scores for properties with NA response time, for Paris Airbnb rentals in March 2024

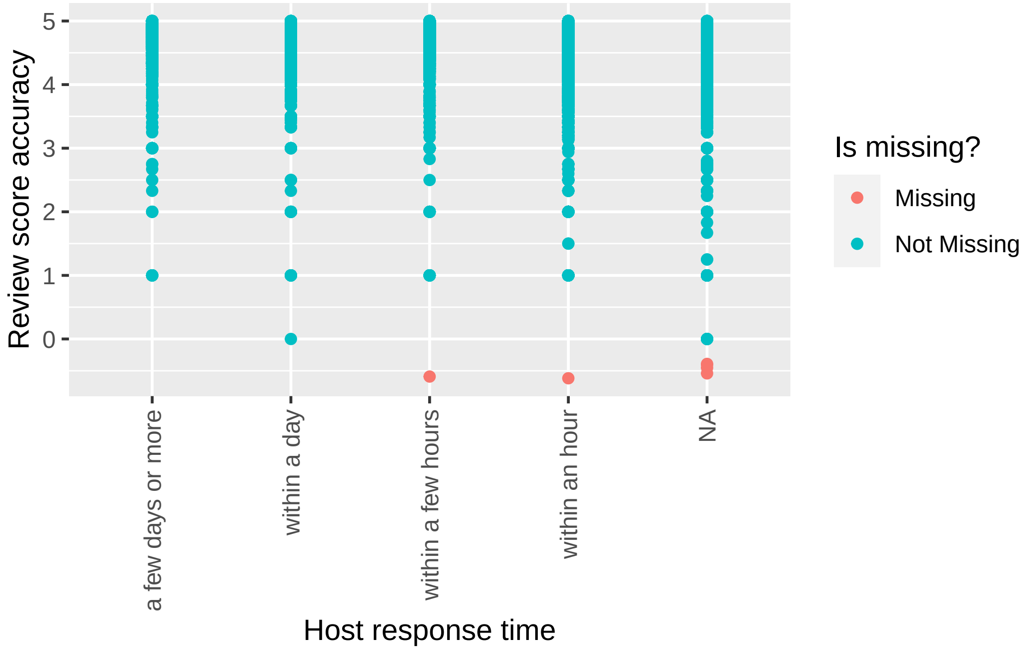


Figure 5: Missing values in Paris Airbnb data, by host response time

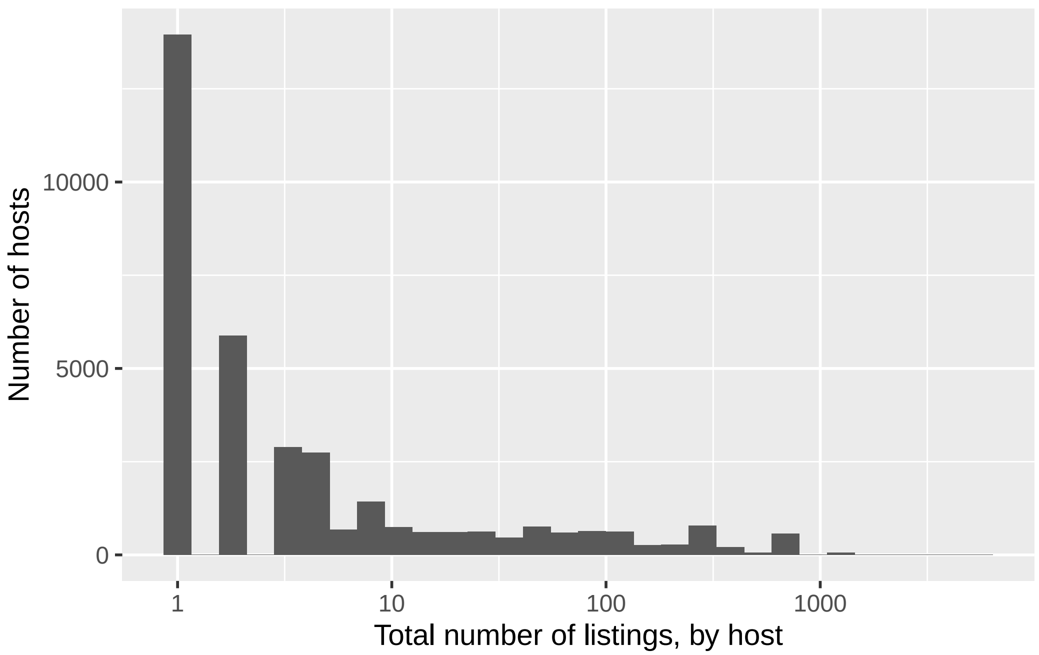


Figure 6: Distribution of the number of properties a host has on Airbnb, for Paris Airbnb rentals in March 2024

3 Relationship between variables

After explore distribution of individual variables, I also interested in relationship between variables. Figure 7 shows the relationship between price and review and whether a host is a superhost, for properties with more than 1 review. From the graph, we can see that most properties have price less than /$250 and gain a 4 or 5 stars review. If the host is a superhost, the properties are more likely have higher price and have more 5 stars review. As the price increases, the average review score also increases.

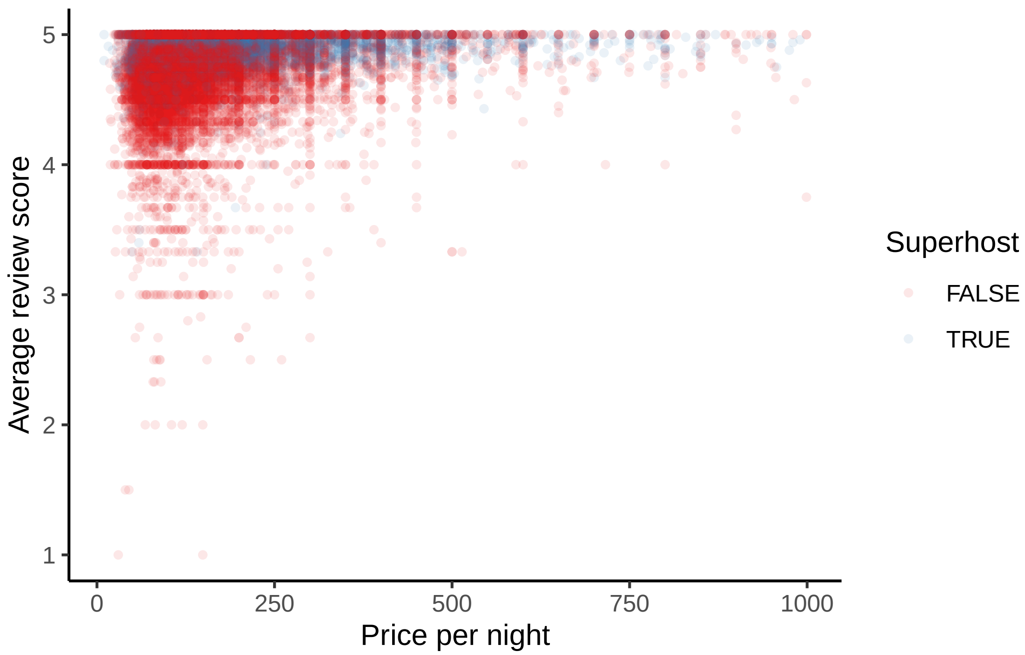


Figure 7: Relationship between price and review and whether a host is a superhost, for Paris Airbnb rentals in March 2024

4 Model

A model is run on the dataset to gain a better understanding of relationships between multiple variables. This model is going to predict whether someone is a superhost, using host response time and review scores rating. Since the output is a binary value, I use the logistic regression for this model. The result shows that each variable is positively associated with the probability of being a superhost.

	(1)
(Intercept)	−16.262
	(0.481)
host_response_timewithin a day	2.019
	(0.211)
host_response_timewithin a few hours	2.695
	(0.210)
host_response_timewithin an hour	2.972
	(0.209)
review_scores_rating	2.624
	(0.089)
Num.Obs.	22 047
AIC	24 165.0
BIC	24 205.0
Log.Lik.	−12 077.507
RMSE	0.43

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.

Cox, Murray. 2021. *Inside Airbnb*. http://insideairbnb.com/.

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

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. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.