Airbnb listings in Paris, France*

Daisy Huo

March 5, 2024

Introduction

In this case study, we will examine Paris, France's Airbnb listings as of March 4, 2024. We will read the dataset, which comes from Inside Airbnb (Cox 2021) and then save a local copy.

Data was collected, cleaned, and analyzed using the statistical programming software R (R Core Team 2023), with additional support from R packages "tidyverse" (Wickham et al. 2019), "modelsummary" (Arel-Bundock 2022), "janitor" (Firke 2023), "knitr" (Xie 2014), "lubridate" (Grolemund and Wickham 2011), "arrow" ("Integration to 'Apache' 'Arrow'," n.d.) and "ggplot2" (Wickham 2016).

As the original dataset is not ours, we will paste the URL copied from Inside Airbnb and download the raw data.

```
# A tibble: 74,329 x 75
      id listing_url             scrape_id last_scraped source name  description
   <dbl> <chr>                       <dbl> <date>       <chr>  <chr> <lgl>
 1  3109 https://www.airbnb.com~   2.02e13 2023-12-12   city ~ Rent~ NA
 2  5396 https://www.airbnb.com~   2.02e13 2023-12-14   city ~ Rent~ NA
 3 81106 https://www.airbnb.com~   2.02e13 2023-12-13   city ~ Rent~ NA
 4  7397 https://www.airbnb.com~   2.02e13 2023-12-13   city ~ Rent~ NA
 5  7964 https://www.airbnb.com~   2.02e13 2023-12-12   city ~ Rent~ NA
 6 81615 https://www.airbnb.com~   2.02e13 2023-12-13   city ~ Rent~ NA
 7  9359 https://www.airbnb.com~   2.02e13 2023-12-13   city ~ Rent~ NA
 8 81870 https://www.airbnb.com~   2.02e13 2023-12-13   city ~ Rent~ NA
 9  9952 https://www.airbnb.com~   2.02e13 2023-12-14   city ~ Rent~ NA
10 86053 https://www.airbnb.com~   2.02e13 2023-12-13   city ~ Rent~ NA
# i 74,319 more rows
```

*Code and data are available at: https://github.com/dai929/Toronto_Homelessness.git

```
# i 68 more variables: neighborhood_overview <chr>, picture_url <chr>,
#   host_id <dbl>, host_url <chr>, host_name <chr>, host_since <date>,
#   host_location <chr>, host_about <chr>, host_response_time <chr>,
#   host_response_rate <chr>, host_acceptance_rate <chr>,
#   host_is_superhost <lgl>, host_thumbnail_url <chr>, host_picture_url <chr>,
#   host_neighbourhood <chr>, host_listings_count <dbl>, ...
```

For exploratory purposes, we will create a parquet file with selected variables.

Distribution and properties of individual variables

The first variable in interest is price. We will need to convert it to a numeric.

```
[1] "$150.00" "$146.00" "$110.00" "$140.00" "$180.00" "$71.00"
```

```
 [1] "$" "1" "5" "0" "." "4" "6" "8" "7" "3" "2" "9" NA  ","
```

```
# A tibble: 1,550 x 1
   price
   <chr>
 1 $1,200.00
 2 $8,000.00
 3 $7,000.00
 4 $1,997.00
 5 $1,000.00
 6 $1,286.00
 7 $2,300.00
 8 $1,500.00
 9 $1,200.00
10 $1,357.00
# i 1,540 more rows
```

Afterthat, we will construct a graph for the distribution of prices (Figure 1) and consider the outliers on the log scale (Figure 2).

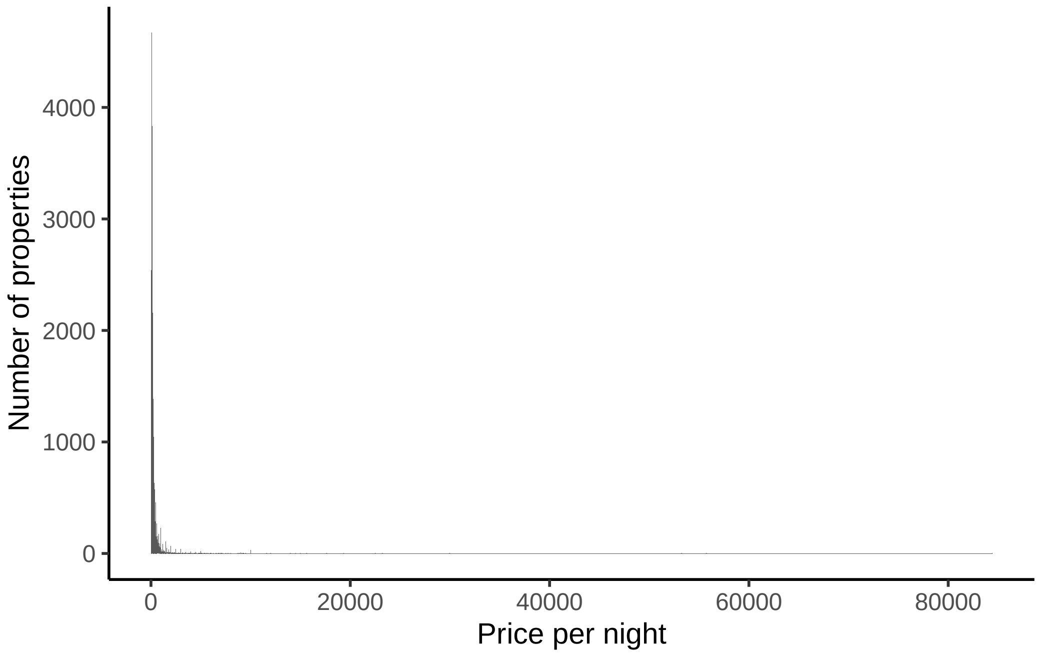

Figure 1: Distribution of prices

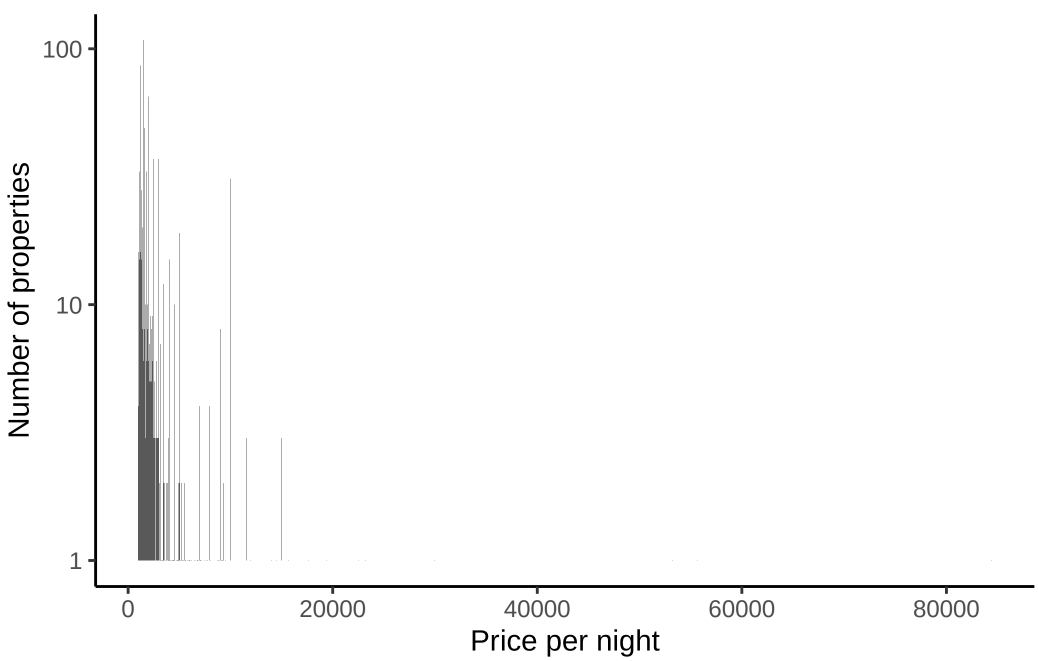

Figure 2: Using the log scale for outliers more than $1000

However, right now we will focus on prices that are less than $1000. Notice that there is some bunching of prices, so we will zoom in by changing to bins to be smaller.

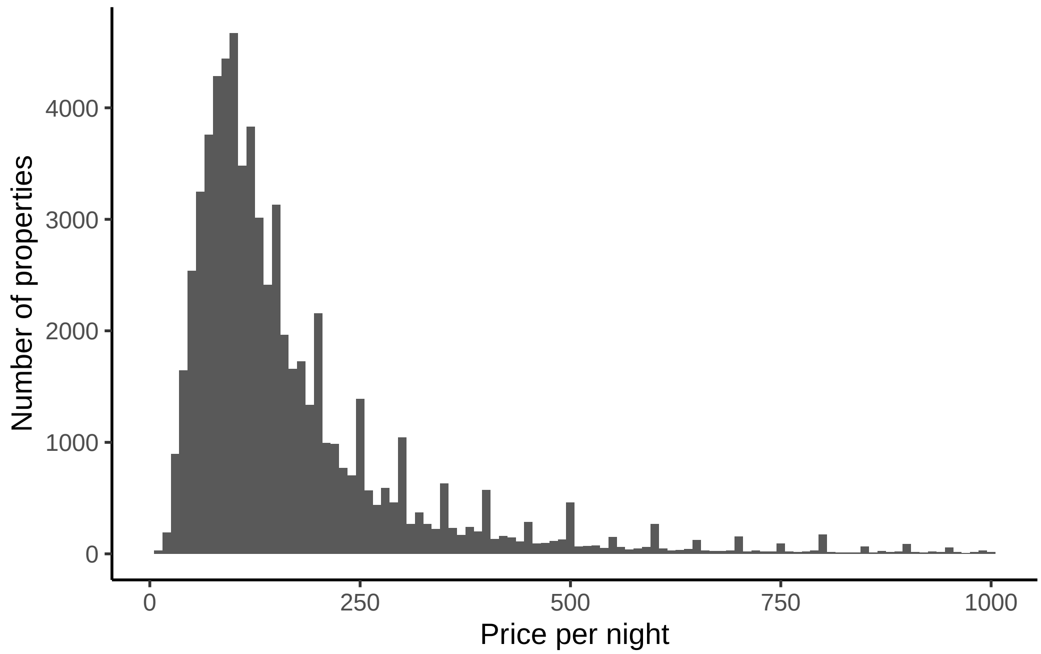

Figure 3: Bunching of prices under $1000

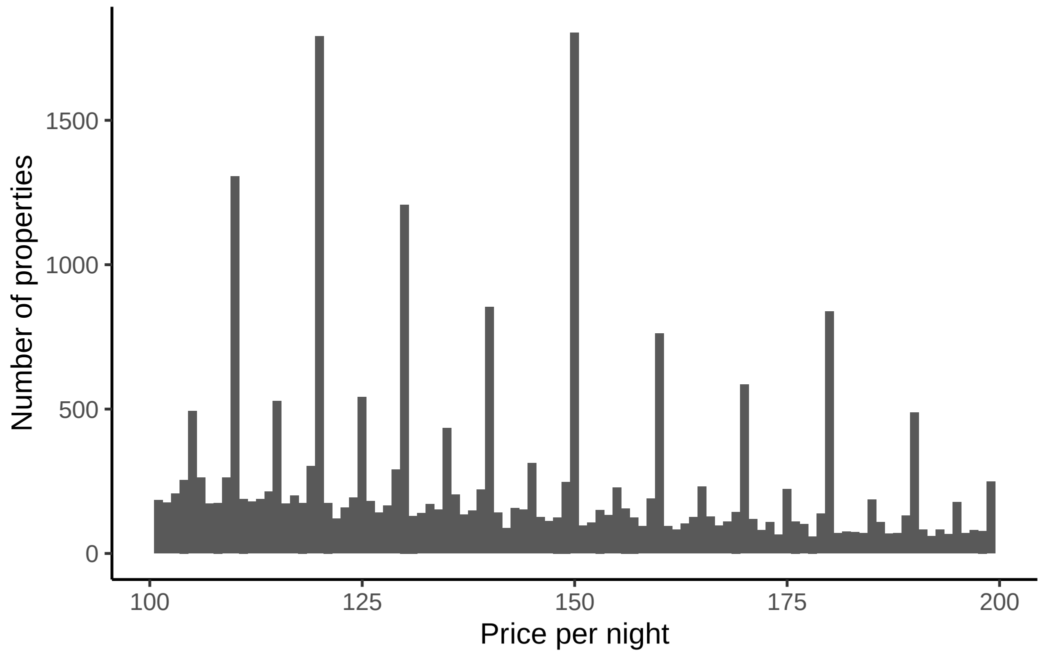

Figure 4: Illustration of bunching of prices between \$100 and \$200

From now on, we will remove all prices above \$999.

We will then turn our attention to superhosts, who are one of the most experienced Airbnb hosts. By creating a binary variable for this group, we can remove anyone else with a NA. Then we will construct a graph for reviews in the dataset, which is a one to five star ratings across multiple aspects.

```
# A tibble: 83 x 12
    host_id host_response_time host_is_superhost host_total_listings_count
      <dbl> <chr>              <lgl>                                  <dbl>
 1 29138344 within an hour     NA                                         3
 2  5869840 within a few hours NA                                         7
 3 35125972 within an hour     NA                                         3
 4 13827149 within a few hours NA                                         3
 5 62919059 within a few hours NA                                         3
 6 22167607 N/A                NA                                         2
 7 10259782 N/A                NA                                         2
 8 62919059 within a few hours NA                                         3
 9 20056470 N/A                NA                                         4
10 20056470 N/A                NA                                         4
# i 73 more rows
# i 8 more variables: neighbourhood_cleansed <chr>, bathrooms <lgl>,
```

```
#   bedrooms <dbl>, price <int>, number_of_reviews <dbl>,
#   review_scores_rating <dbl>, review_scores_accuracy <dbl>,
#   review_scores_value <dbl>
```

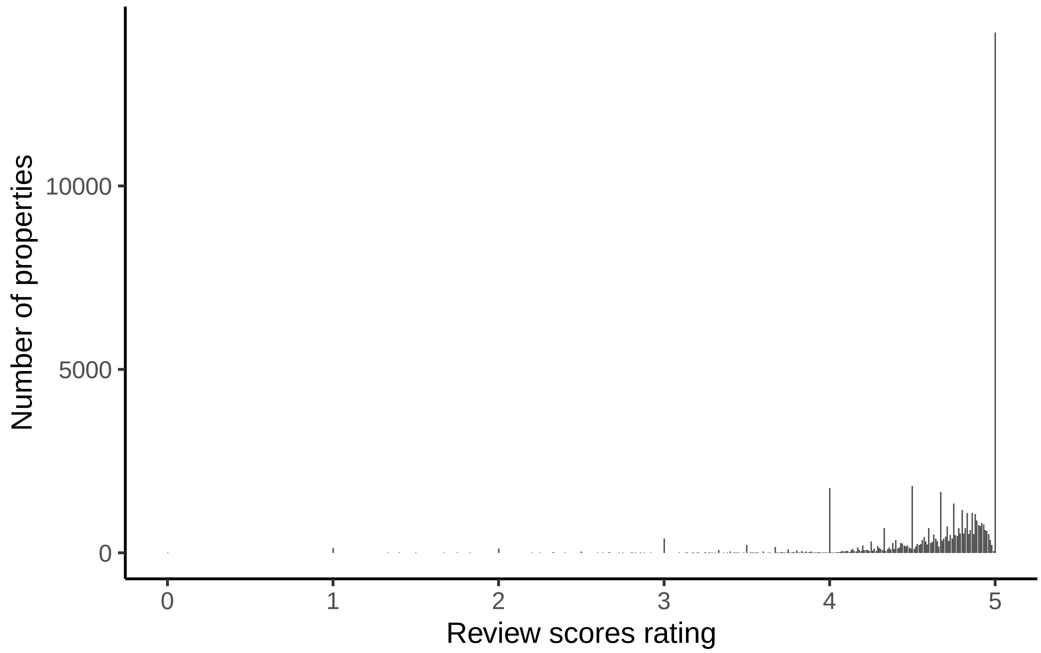

Figure 5: Distribution of reviews for Paris Airbnb in March, 2024

The NAs in the reviews are complicated to deal with. We could just focus on the main review scores and remove anyone with an NA, which is a large proportion of the entire observations. From figure 6, we can tell that guests mostly reviewed five-star for their experiences in Paris Airbnb.

```
[1] 13497
```

```
number_of_reviews
    0
13497
```

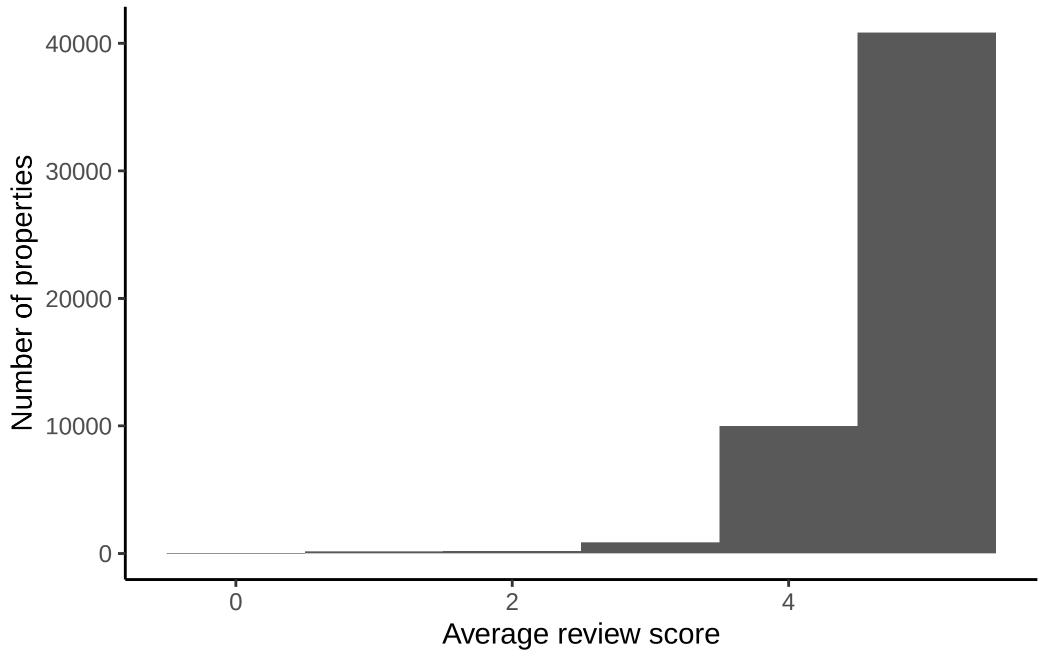

Figure 6: Distribution of main reviews for Paris Airbnb in March 2024

Another factor we will take into account is the hosts' response time. Again, people with NAs for this variable also created an issue, as there are a large number of them. We will construct a graph to see if there is any relationship with the reviews for NA response time.

```
# A tibble: 6 x 2
  host_response_time     n
  <chr>              <int>
1 N/A                16531
2 a few days or more  1243
3 within a day        5297
4 within a few hours  6811
5 within an hour     22094
6 <NA>                   2
```

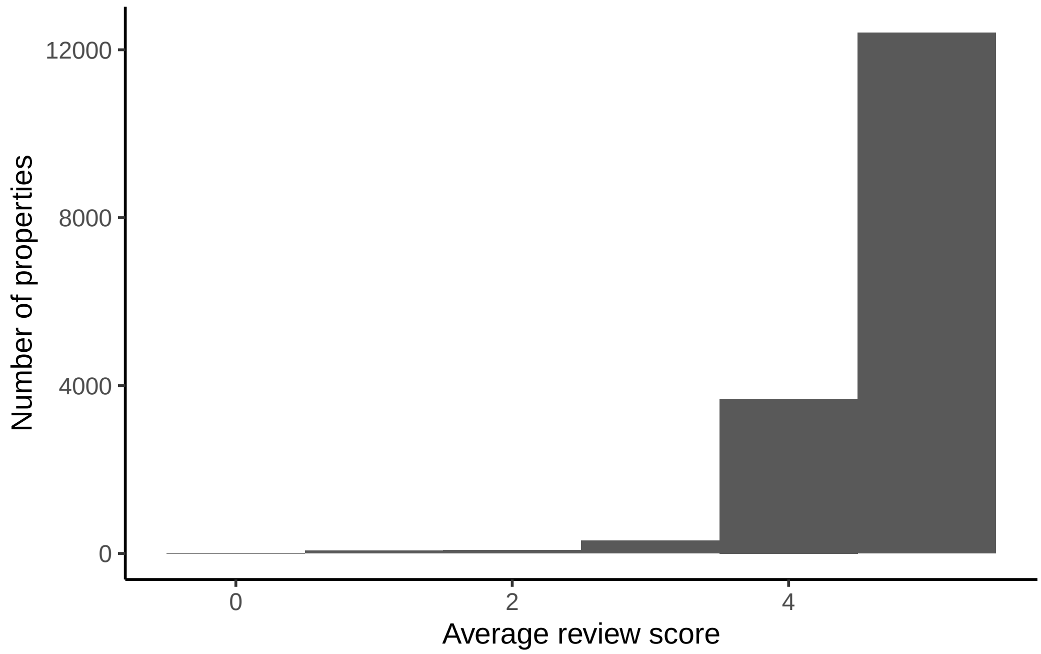

Figure 7: Distrubution of reviews for NA response time for Paris Airbnb in March 2024

From now on, we will remove all people with a NA in response time.

We will construct a graph for distribution of the number of properties a host has. In addition, from now on, we will only deal with the hosts with one property.

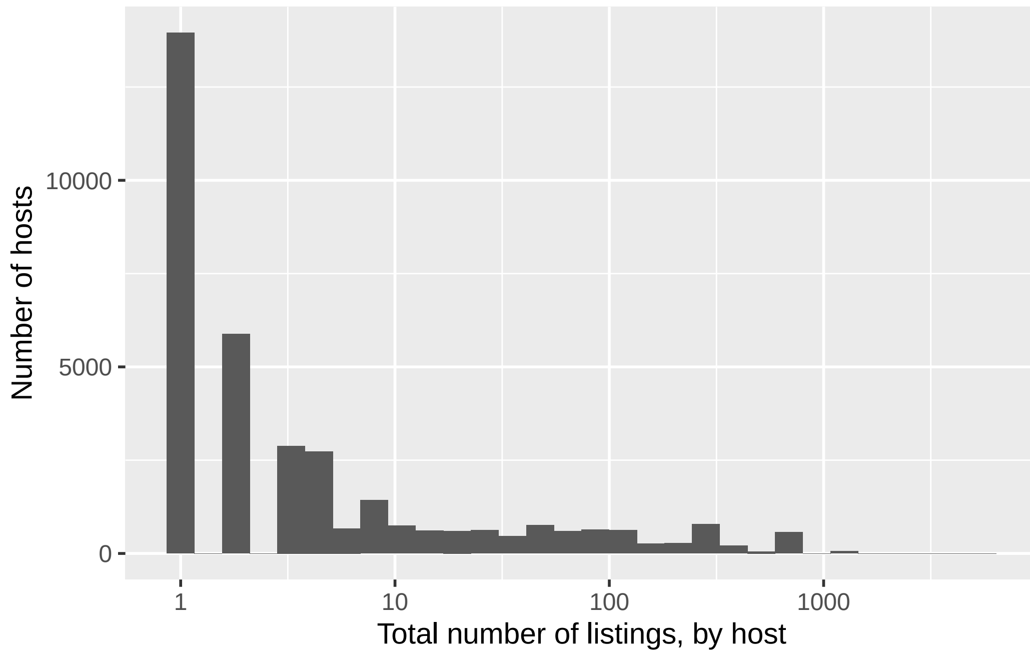

Figure 8: Distrubution of the number of properties a host has for Paris Airbnb in March 2024

Relationships between variables

We will construct a graph to see if there is any relationship between price per night, whether a host is a superhost and average review scores.

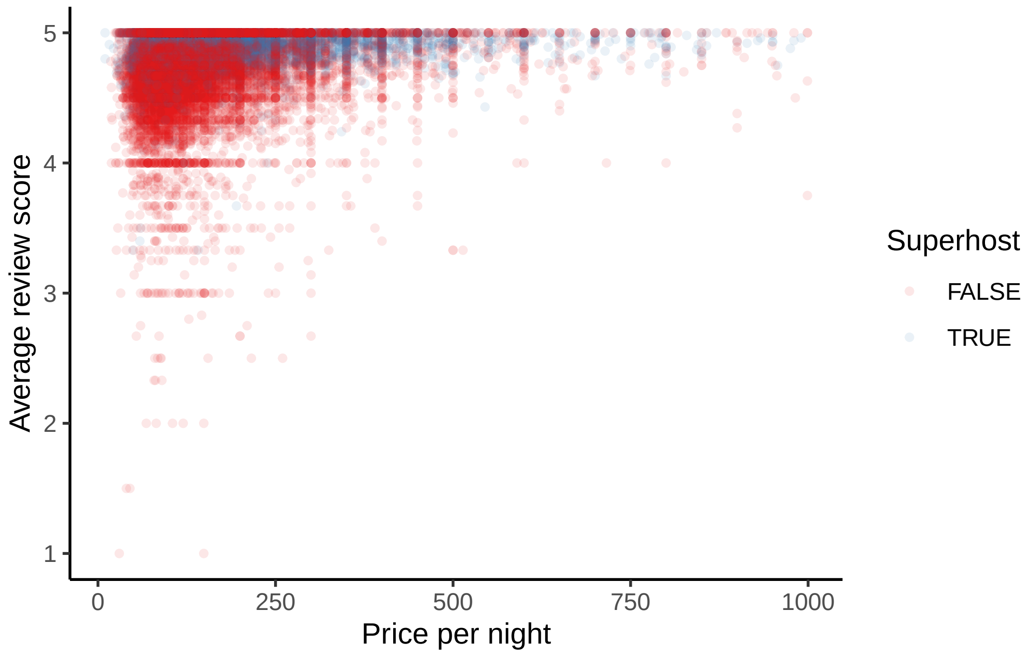

Figure 9: Relationship between price, review and whether a host is a superhost for Paris Airbnb in March 2024

We will then look for possible values of superhost by the response time. It is obvious that hosts with a faster response time, especially within an hour, are more likely to become a superhost. None of the hosts with a response time of a few days or more becomes a superhost.

```
# A tibble: 2 x 3
  host_is_superhost     n proportion
  <lgl>             <int>      <dbl>
1 FALSE             15820       0.72
2 TRUE               6227       0.28
```

```
                    host_is_superhost
 host_response_time            FALSE          TRUE
 a few days or more        6%   (953)  0%     (24)
       within a day       22% (3,511) 12%    (770)
 within a few hours       24% (3,802) 26% (1,614)
     within an hour       48% (7,554) 61% (3,819)
```

Finally, we are able to carry out an Airbnb EDA for Paris. In this case study, we have a hypothesis that superhosts are positively related with faster response time and higher review scores. We estimate the model as follows.

	(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. "Paris." *Insideairbnb.com*. http://insideairbnb.com/paris.

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/.

"Integration to 'Apache' 'Arrow'." n.d. https://github.com/apache/arrow/, https://arrow.apache.org/docs/r/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.