Paris Airbnb Analysis

Ruoxian Wu

2024-03-03

Exploratory data analysis (EDA) is used to predict factors that impact a host's likelihood of achieving superhost status. Based on the findings, we can gain insight into characteristics of superhost listings, significance of guest reviews, and pricing strategies.

Introduction

The Parisian Airbnb market is highly popular among tourists from all over the world as a renowned tourist destination. The wide variety of Airbnb listings available in the city offer valuable insights into trends in urban hospitality, pricing dynamics, and the factors that impact guest satisfaction and host performance.

This study uses data from Inside Airbnb as of 12 December 2023, and uses various data processing and analytical techniques to examine the relationship between listing characteristics, host attributes, and market outcomes. The report uses exploratory data analysis (EDA) to explore the distribution of prices across different neighborhoods, identify the traits of highly rated listings, and examine the correlation between superhost status and review scores. Advanced statistical methods, including logistic regression, are used to predict factors that significantly impact a host's likelihood of achieving superhost status.

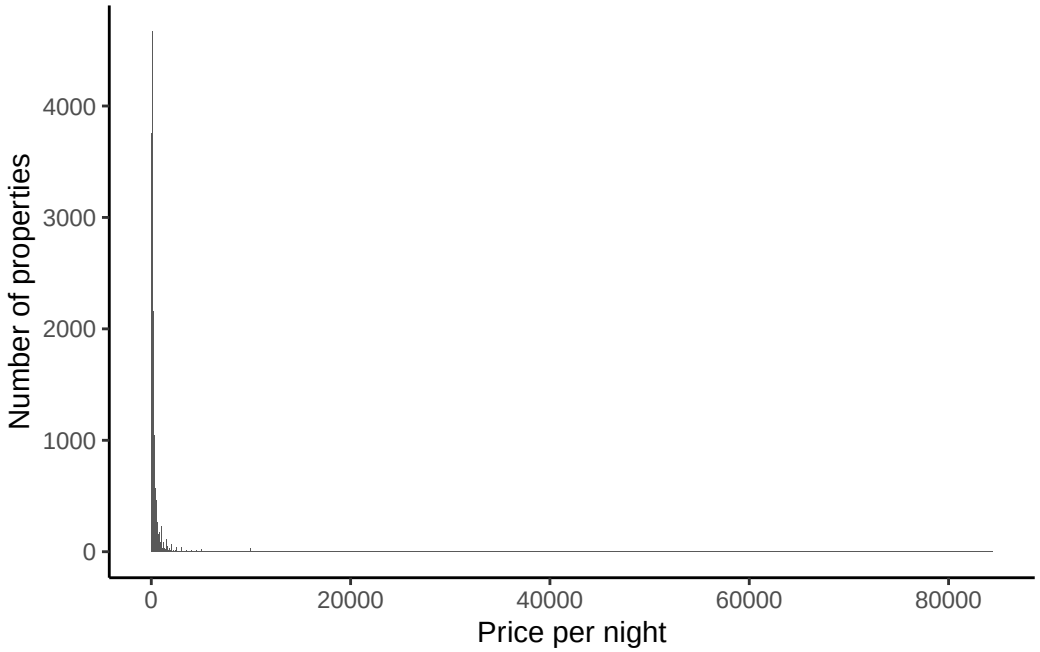
The report aims to provide valuable recommendations for Airbnb hosts looking to optimize their listings for enhanced profitability and guest experience, while also offering guests a lens through which to make informed accommodation choices in Paris.

Data Analysis

This study involves several stages of data processing. We start the process by acquiring the data, followed by a detailed cleaning and formatting step to ensure that the results are accurate and reliable, including removing non-numeric characters from prices and converting them

to integers to make sure they are accurate. The study makes use of various R packages. Dplyr, ggplot2, and other tidyverse packages support a wide range of data manipulation, exploration, and visualization tasks for exploratory data analysis (EDA). This EDA focuses on identifying patterns in pricing, the distribution of listings across neighborhoods, and the correlation between review scores and host attributes. The arrow package is utilized for reading and writing data in the Parquet format, enabling efficient data storage and access. The janitor package assists in data cleaning tasks, such as removing duplicate entries and tidying variable names. The knitr package is used for dynamic report generation, allowing for seamless integration of R code and its output into documents. The modelsummary package facilitates the creation of elegant tables summarizing statistical model results. The naniar package offers specialized functions for handling missing data, providing insights into the pattern of missingness.

```
Warning: Removed 7221 rows containing non-finite outside the scale range
(`stat_bin()`).
```

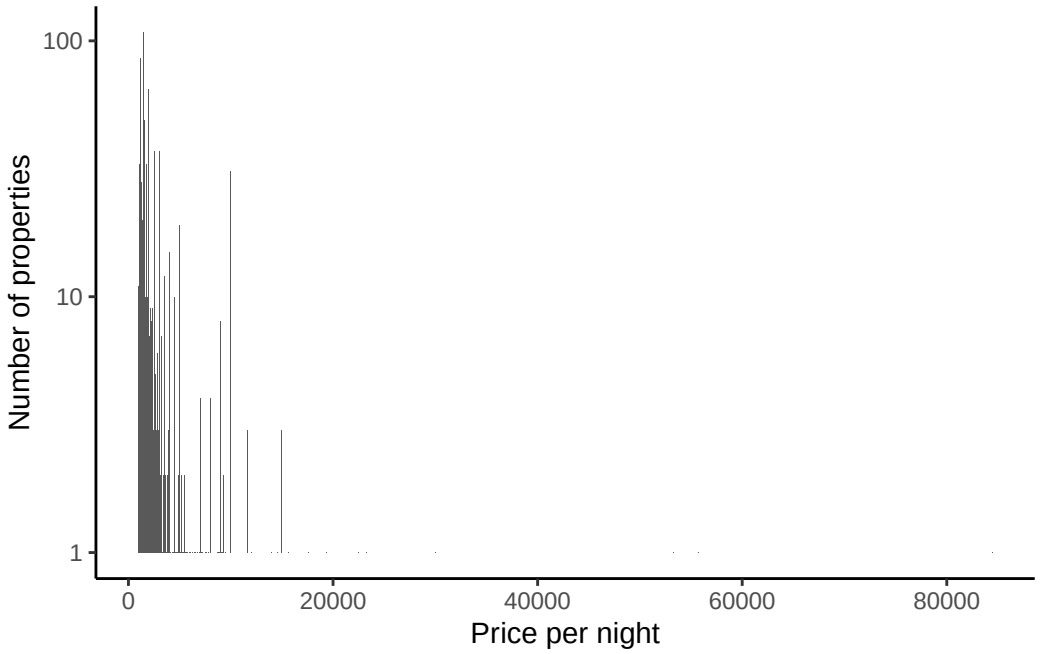


(a) Distribution of price

Figure 1: Distribution of prices

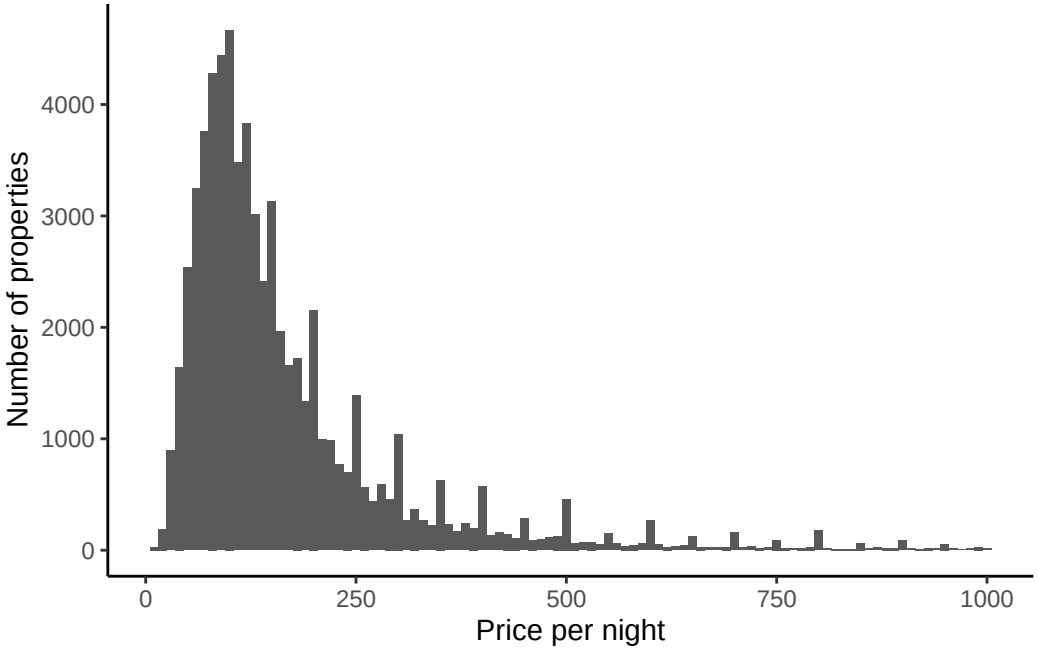
```
Warning in scale_y_log10(): log-10 transformation introduced infinite values.
```

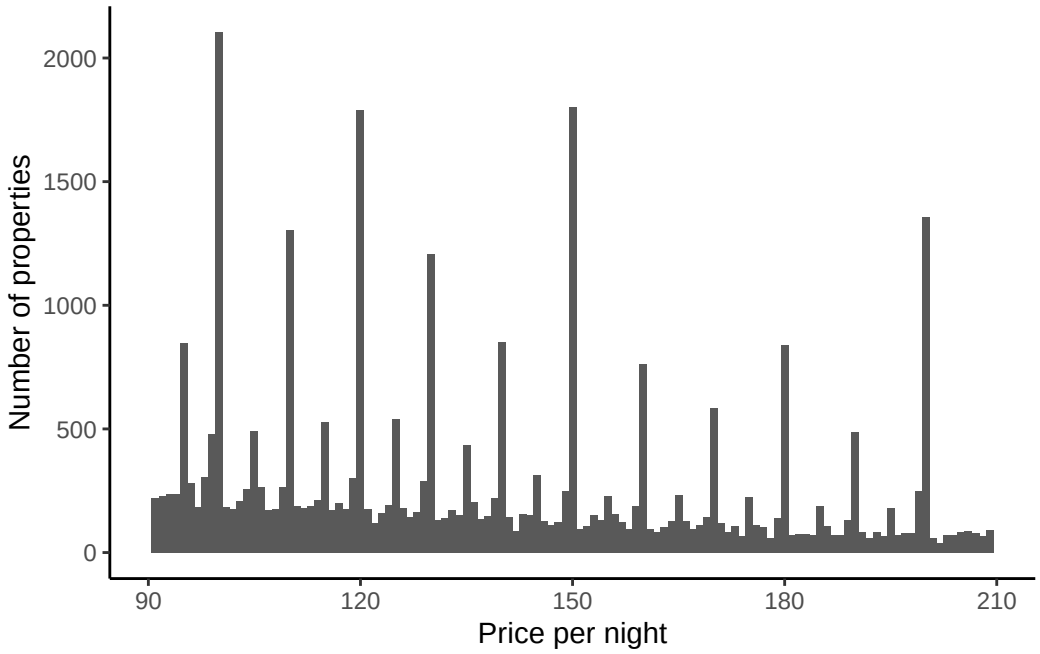
```
Warning: Removed 8085 rows containing missing values or values outside the scale range
(`geom_bar()`).
```



(a) Distribution of Prices Using the Log Scale for Prices More Than $1,000

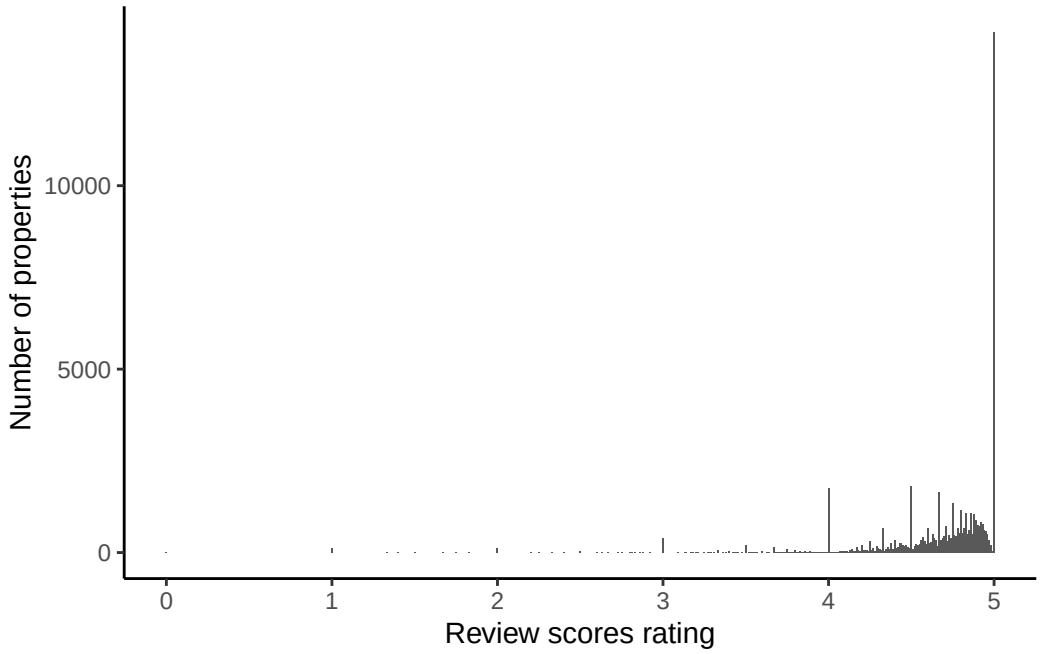
Figure 2: Distribution of prices of Paris Airbnb






```
# 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>
```

```
Warning: Removed 13497 rows containing non-finite outside the scale range
(`stat_count()`).
```

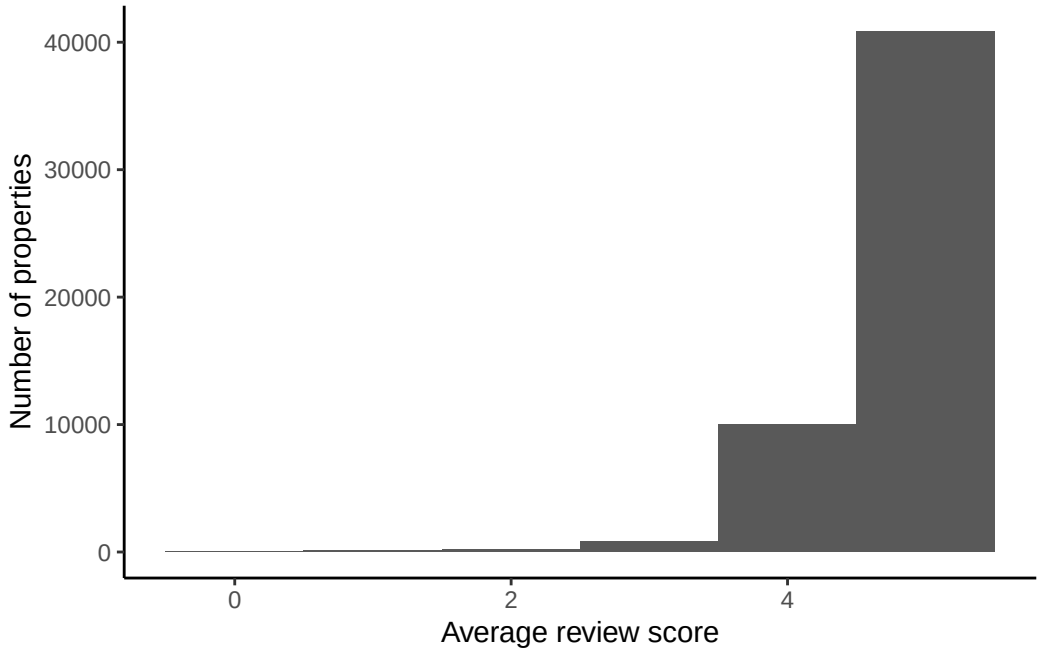



```
[1] 13497
```

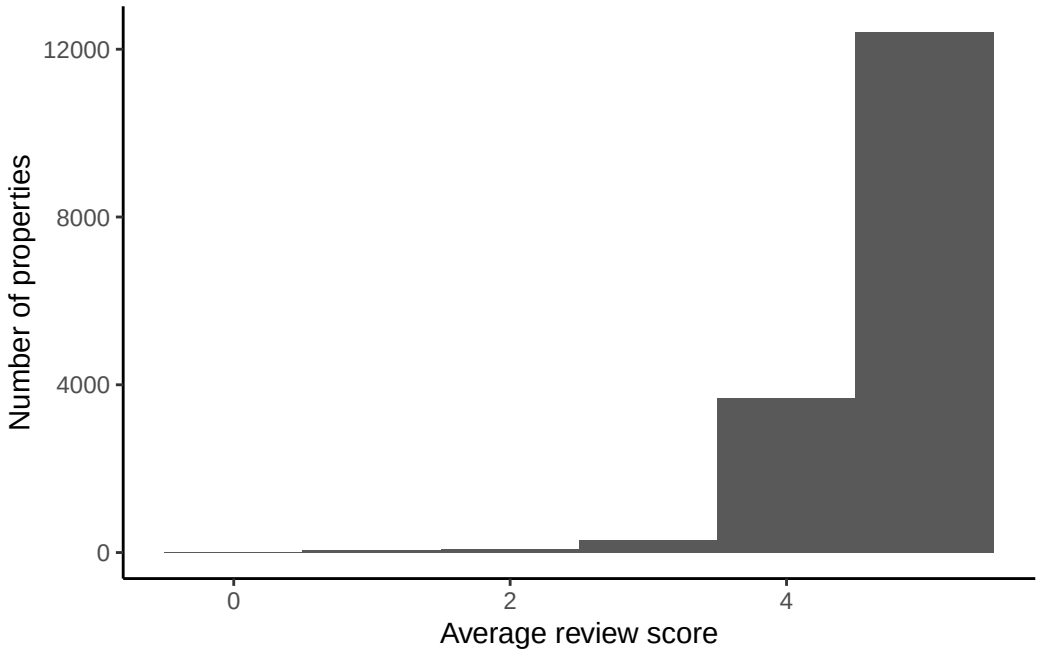
```
number_of_reviews
    0
13497
```

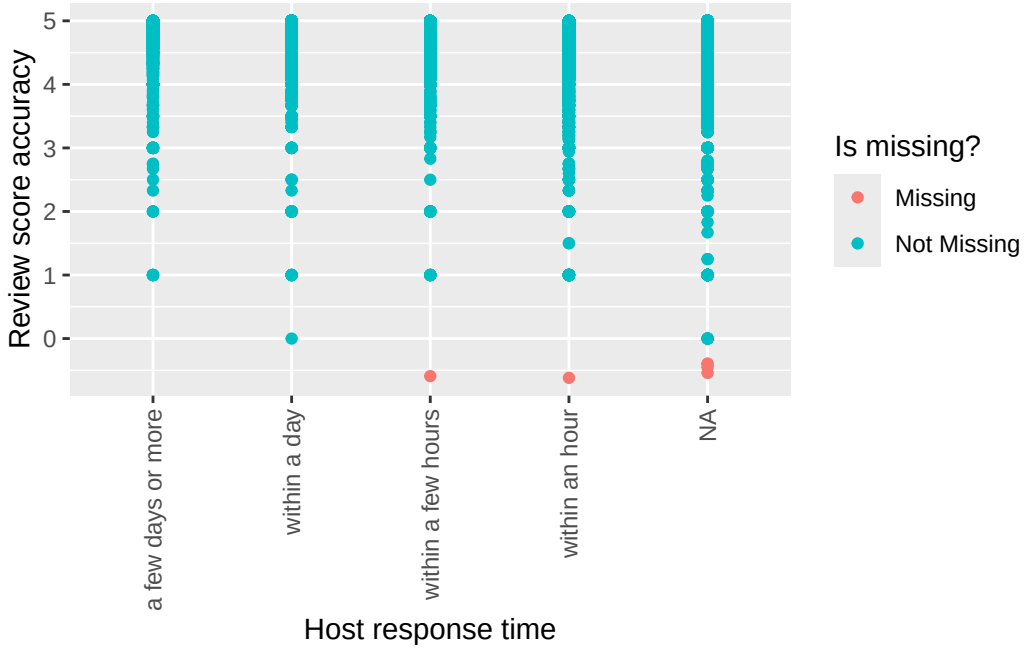
```
# A tibble: 65,475 x 13
   host_id host_response_time host_is_superhost host_total_listings_count
     <dbl> <chr>              <lgl>                                 <dbl>
 1    3631 within a few hours FALSE                                     2
 2    7903 within an hour     FALSE                                     3
 3  439130 within a few hours FALSE                                     1
 4    2626 within an hour     TRUE                                      9
 5   22155 N/A                FALSE                                     1
 6  429406 within a day       FALSE                                     5
 7   28422 N/A                FALSE                                     4
 8  152242 within an hour     FALSE                                   245
 9   33534 within a few hours TRUE                                      1
10  296615 within a few hours TRUE                                      1
# i 65,465 more rows
# i 9 more variables: neighbourhood_cleansed <chr>, bathrooms <lgl>,
#   bedrooms <dbl>, price <int>, number_of_reviews <dbl>,
#   review_scores_rating <dbl>, review_scores_accuracy <dbl>,
```

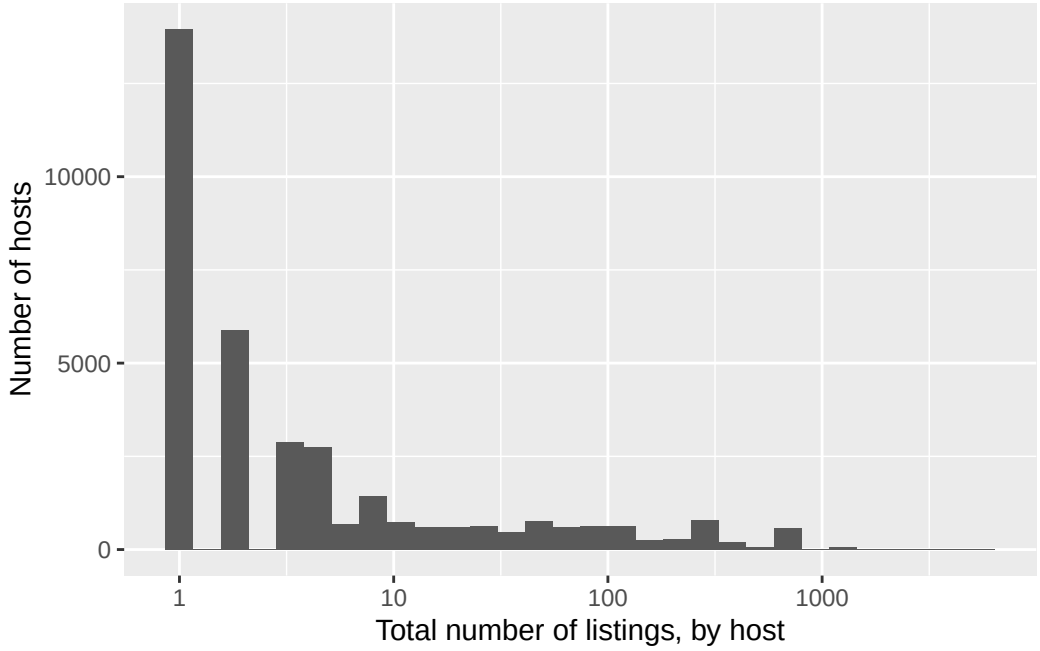
```
#   review_scores_value <dbl>, host_is_superhost_binary <dbl>
```




```
# 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
```


12000
8000
4000
0
Number of properties
0
2
4
Average review score

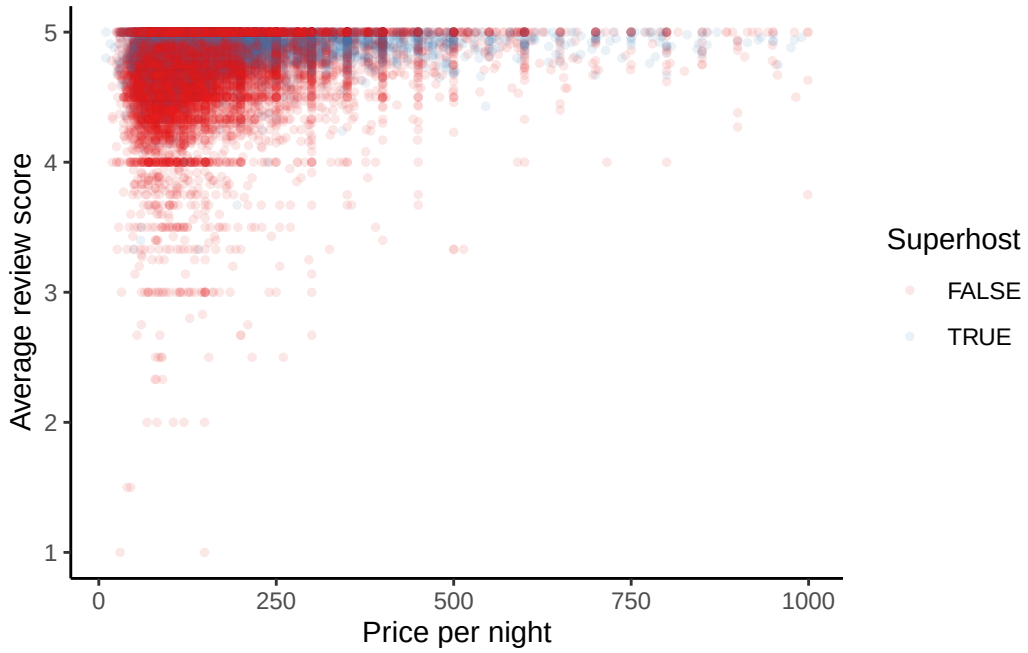

5
4
3
2
1
0
Review score accuracy
a few days or more
within a day
within a few hours
within an hour
NA
Host response time
Is missing?
Missing
Not Missing




```
# A tibble: 6 x 13
   host_id host_response_time host_is_superhost host_total_listings_count
     <dbl> <fct>              <lgl>                                  <dbl>
1 50502817 within an hour     FALSE                                    778
2 50502817 within an hour     FALSE                                    778
3 50502817 within an hour     FALSE                                    778
4 50502817 within an hour     FALSE                                    778
5 50502817 within an hour     FALSE                                    778
6 50502817 within an hour     FALSE                                    778
# i 9 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>, host_is_superhost_binary <dbl>
```

```
# A tibble: 6 x 13
   host_id host_response_time host_is_superhost host_total_listings_count
     <dbl> <fct>              <lgl>                                  <dbl>
1 50502817 within an hour     FALSE                                    778
2 50502817 within an hour     FALSE                                    778
3 50502817 within an hour     FALSE                                    778
4 50502817 within an hour     FALSE                                    778
5 50502817 within an hour     FALSE                                    778
6 50502817 within an hour     FALSE                                    778
```

```
# i 9 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>, host_is_superhost_binary <dbl>
```




```
# 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)

	(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