Exploratory data analysis - Airbnb Paris*

Yichen Ji

2024-03-05

In this essay, we will do the exploratory data analysis on the data of Airbnb in Paris collected at 12th December in 2023.At first, we need to download Paris's data from Airbnb official website, and store it on parquet form. After that, we can start to clean data. The character represents money and comma between numbers are deleted.

Variables

There are 12 variables in our cleaned data.

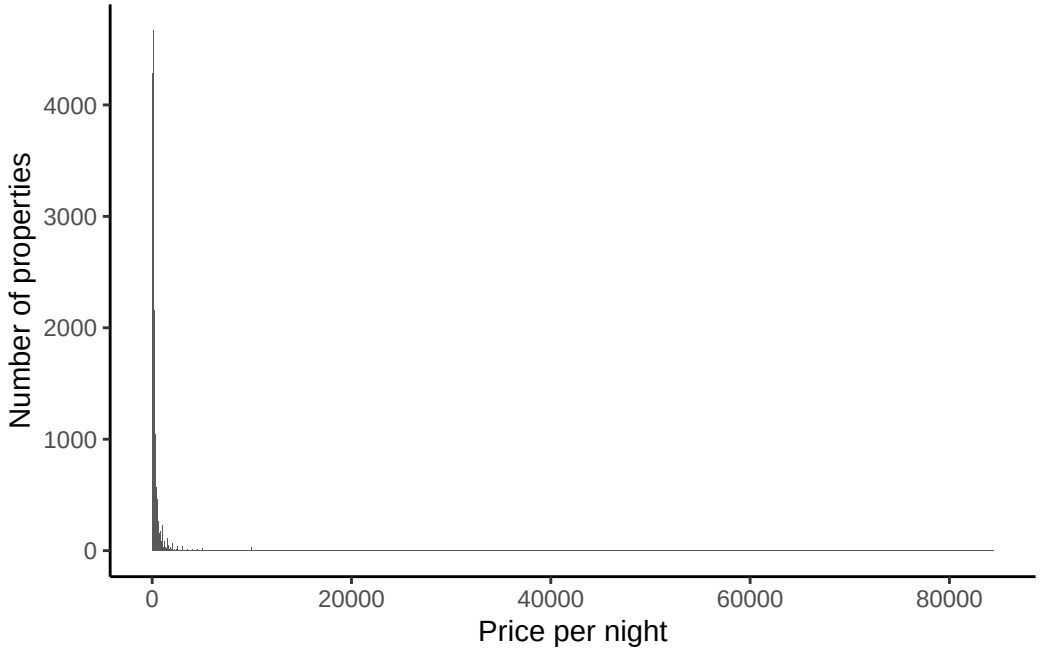
- host_id
- host_response_time
- host_is_superhost
- host_total_listings_count
- neighbourhood_cleansed
- bathrooms
- bedrooms
- price
- number_of_reviews
- review_scores_rating
- review_scores_accuracy

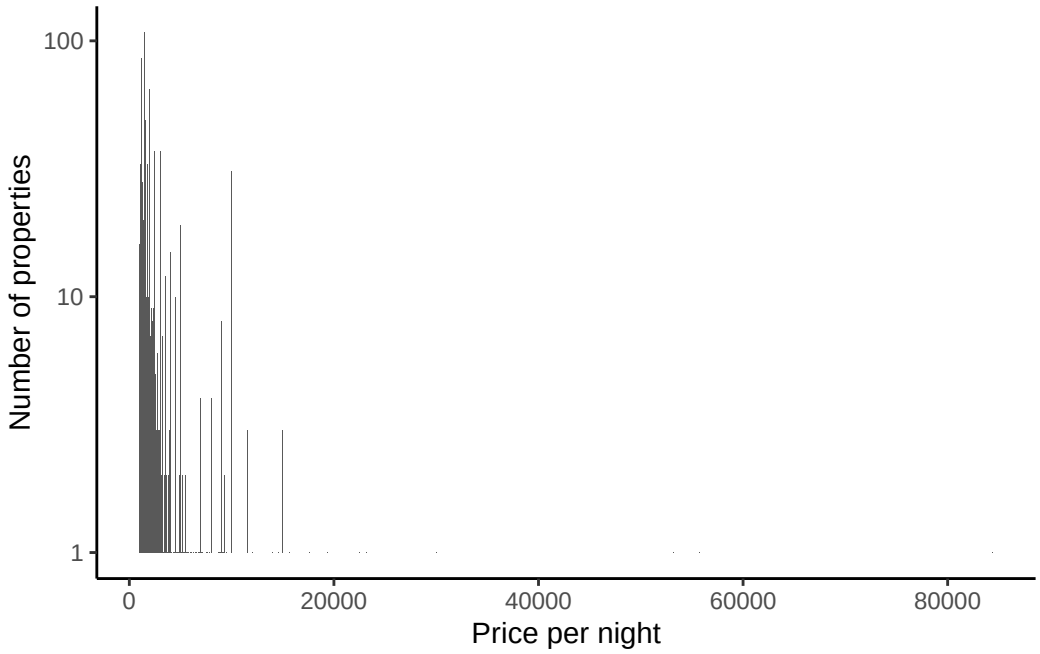
*Code and data are available at: https://github.com/Selinayichenji/small_tasks.git. Please check the folder named mini-essay_8.

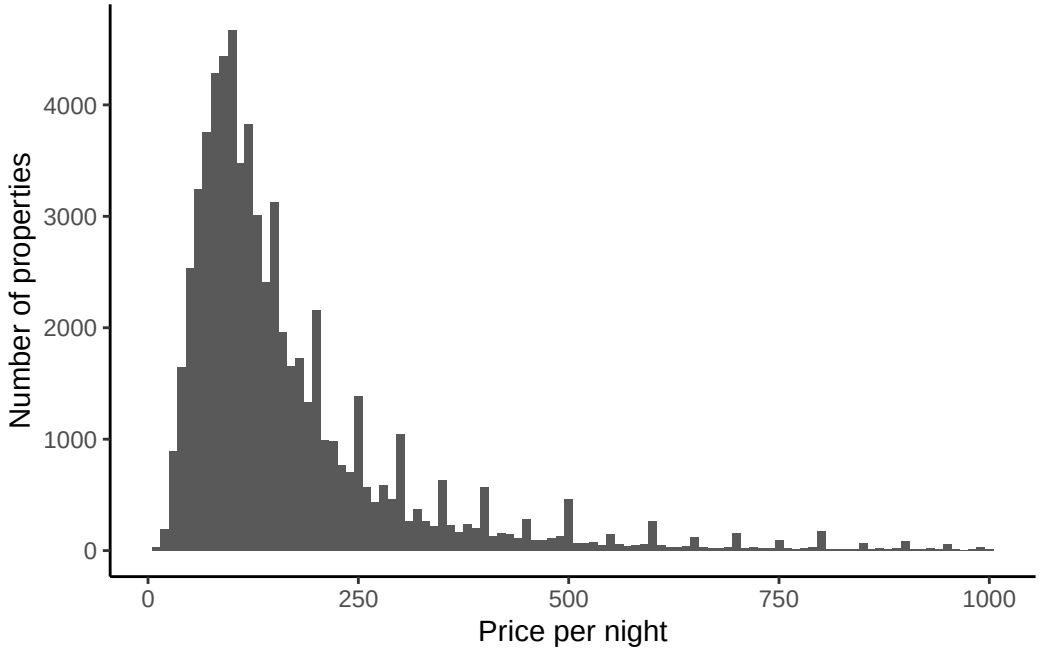
- review_scores_value

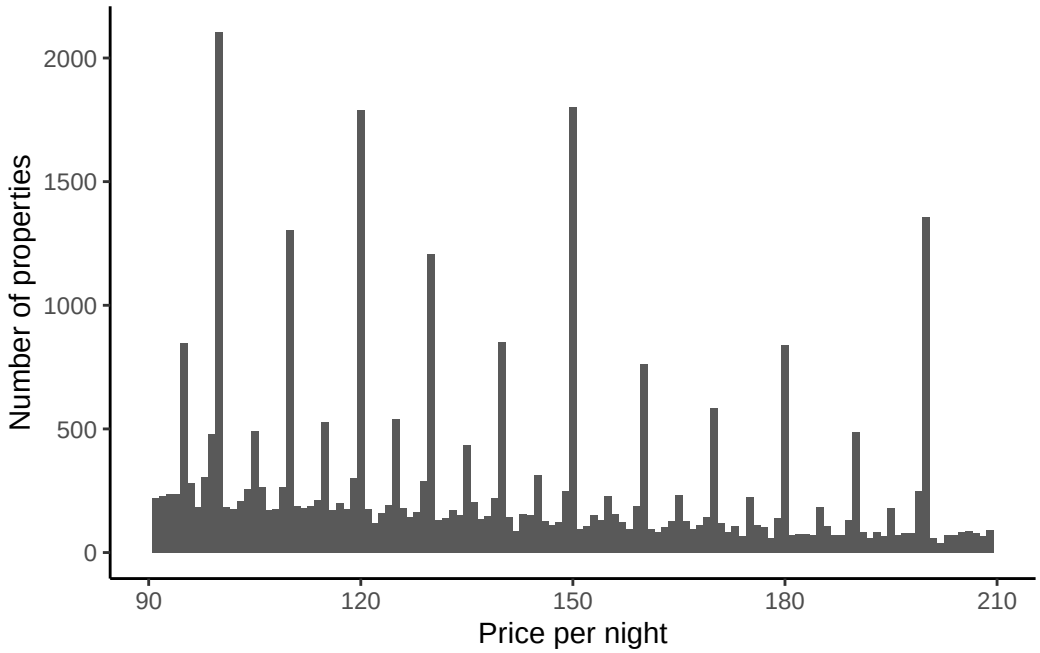
  The host_response_time variable includes 3 choices: “within a few hours”, “within an hour” and “within a day”.
  The host_is_superhost variable column contains only TRUE or FALSE.
  The host_total_listings_count variable shows how many houses the host is operating in the Airbnb.
  The neighbourhood_cleansed variable shows the name of the neighbourhood where the house in. e.g: “Observatoire”.

After a glance of all variables, let’s check distribution and properties of individual variables.




100
10
1
Number of properties
0
20000
40000
60000
80000
Price per night


4000
3000
2000
1000
0
Number of properties
0
250
500
750
1000
Price per night



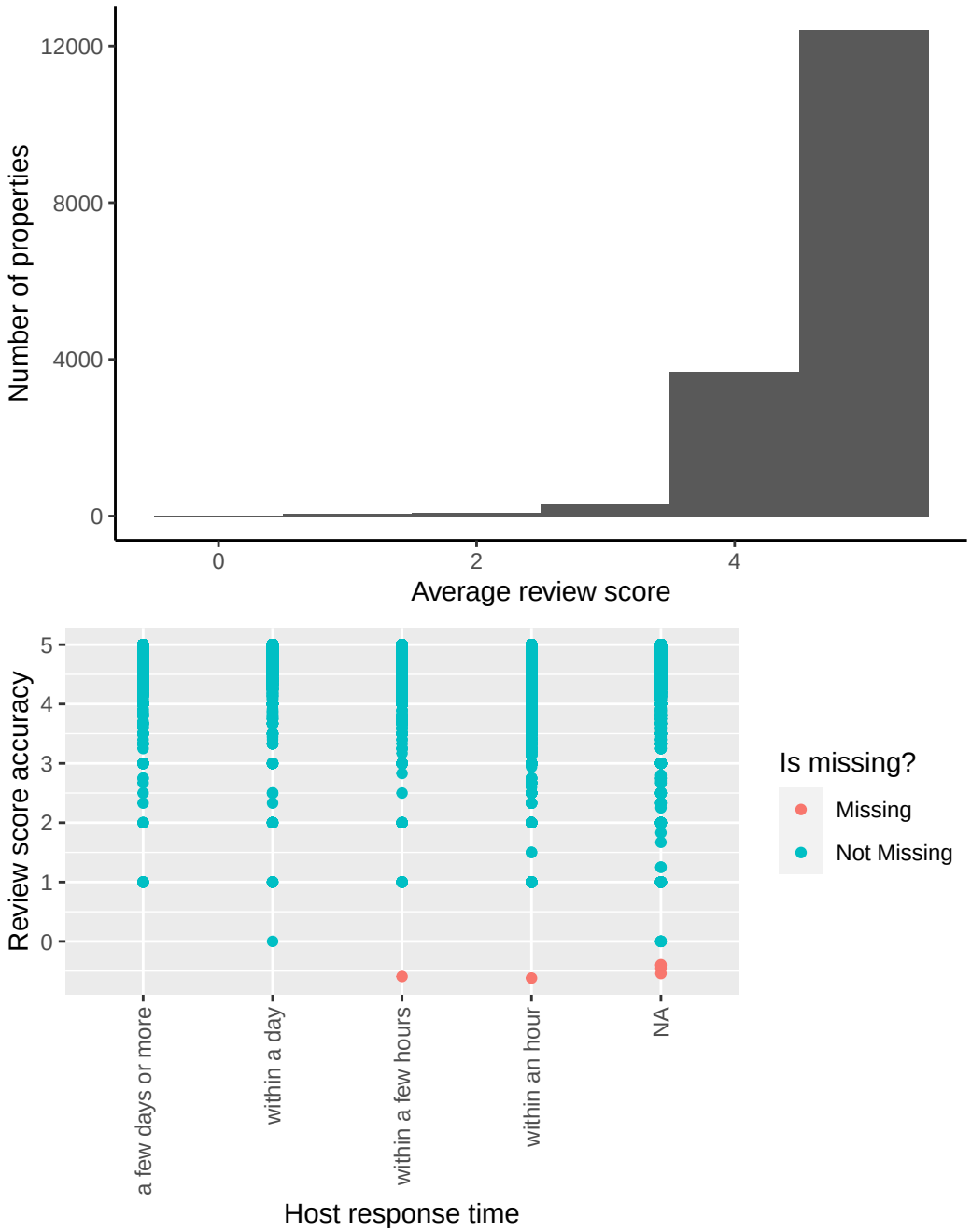
It is obviously to see that lines are closely clustered on the far left of the axis. Most prices per night of Airbnb houses are set inside the range of 250 dollars.The price of Paris houses are seemly normally distributed inside the range. By shortening the range of price per night, we finally stop at the range from 90 dollars to 210 dollars. Interestingly, seen from the figure, hosts prefer to set rental prices in whole hundreds.

The we start to investigate the variable host_is_superhost. For further analysis, we need clean all the rows whose super_host column do not contain information (NA). Then for the convenience, we transfer results from TRUE/FALSE to binary number 1 and 0.

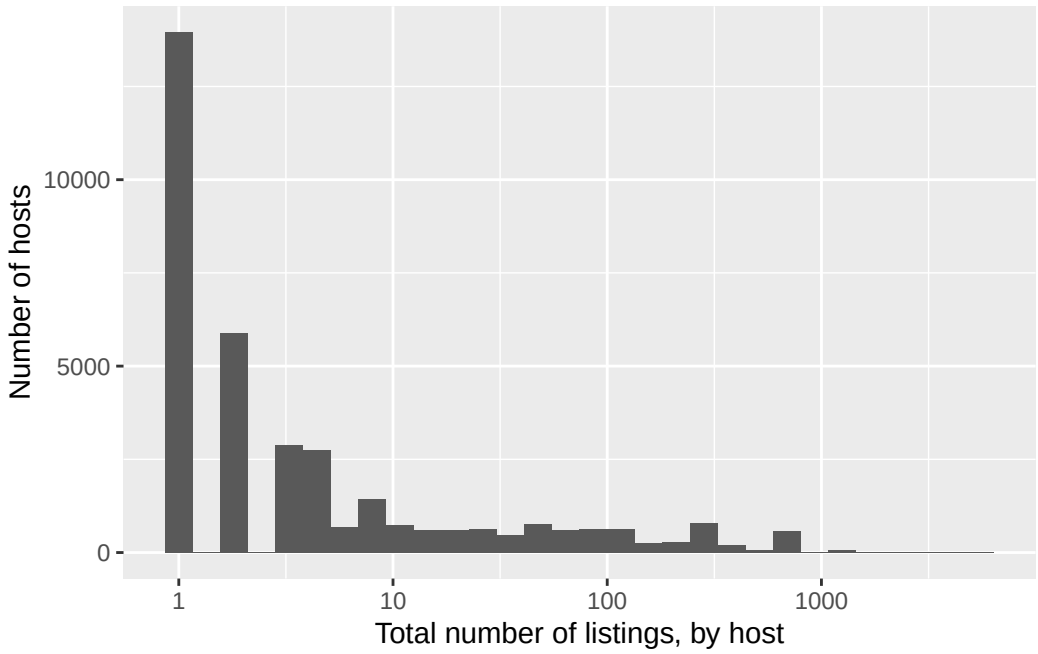
After counting number of all reposnses in Average review score column, we can see it contains lots of NA.

```
# 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
```

The average review scores after cleaning NA is shown below, most comments concentrate on high scores.

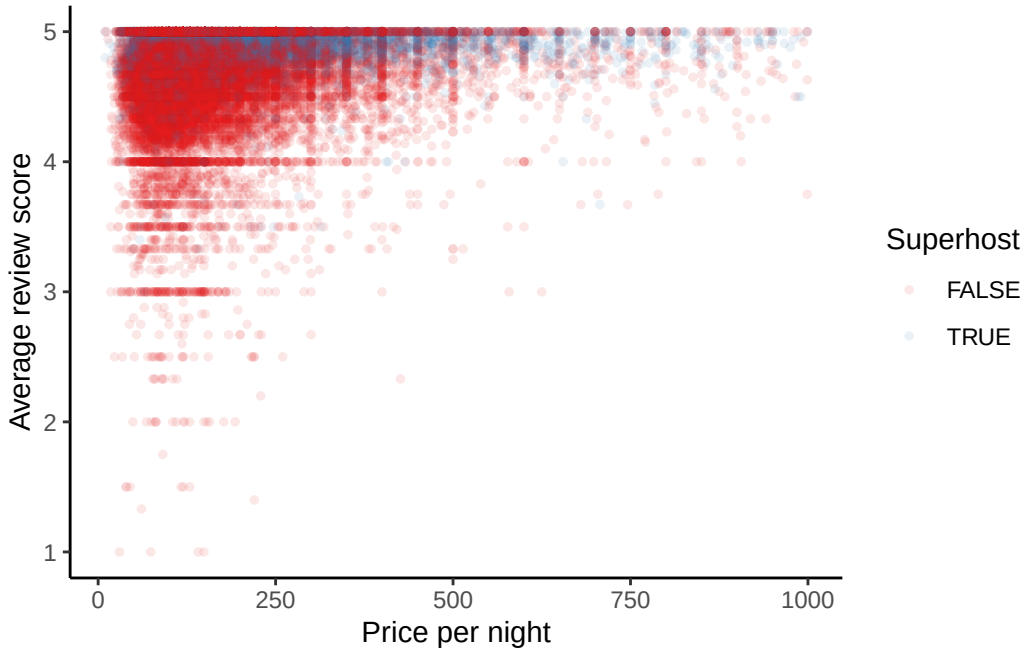


For the column host response time, NA are also exists, but the number of NA is small.



As shown above, most hosts have only one house for renting in Airbnb. The number of hosts who own and rent out that amount of properties begins to plummet after reaching six or seven.

After analysising individual variables, the next step is analysising the relationship between several variables.



As shown on @super_fig_1, as price per night increases, trend of average review scores increases obviously and focus between 4 and 5, the blue points represent super host starts to take a higher percentage than red points (not super host).

Except these 2 variables(price per night and average scores review), host response time also correlated with whether a host is super host. From the data below, it is obvious to see that shorter the response time, more possible the host is a super host.

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)

neighbourhood_cleansed	n	percent
Buttes-Montmartre	3737	10.5%
Popincourt	3076	8.7%
Vaugirard	2587	7.3%
Entrepôt	2552	7.2%
Batignolles-Monceau	2197	6.2%
Buttes-Chaumont	1895	5.3%
Temple	1848	5.2%

Passy	1835	5.2%
Ménilmontant	1834	5.2%
Opéra	1783	5.0%
Bourse	1458	4.1%
Hôtel-de-Ville	1407	4.0%
Reuilly	1281	3.6%
Observatoire	1256	3.5%
Panthéon	1240	3.5%
Élysée	1199	3.4%
Luxembourg	1118	3.2%
Gobelins	1096	3.1%
Palais-Bourbon	1073	3.0%
Louvre	973	2.7%
Total	35445	-

The last variable is neighbourhood, the Buttes-Montmartre is the neighbourhood has the most hosts, 10.5% hosts are in this neighbourhood. Popincourt follows the Buttes-Montmartre, its data is 8.7%. For neighbourhoods Temple and Passy, they have the least percentage of hosts, only 5.2% hosts came from each of them.

Now, we want to predict whether a host can be a super host or not by using the variables we analysis before. As the result of being super host is expressed as binary number, we can use the glm model to approach the goal.

The formula of glm model is:

$$\text{Prob}(\text{Is superhost} = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 \text{Response time} + \beta_2 \text{Reviews} + \epsilon)$$

The coefficiencies shown on the table illustrate that both of reponse time and review are positively correlated with probability of becoming a superhost, and the faster the host response, the more possible he/she will become a super host.

The last step is analysis is rewriting parquet file and store it again.

	(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
F	674.466
RMSE	0.43