Airbnb listings in Paris, France*

Mary Cheng

March 5, 2024

Introduction

This report uses statistically programming language R (R Core Team 2020) to analyze the data on Paris Airbnb situations. The report will look at the distribution and properties of individual variables first, then look at the relationships between the variables to see if there is any correlation.

The report is going to get the dataset from Inside Airbnb (*Get the Data* 2024), then explore the data features and plot the graphs to illustrate the dataset.

Distribution and properties of individual variables

After cleaning the dataset, the distribution of prices is looked at first. I plotted Figure 1 and Figure 2 to illustrate the distribution. Plotting on regular scale shows multiple outliers, such as the ones with very high prices. So a log scale is used to plot the distribution again to make it more clear. We can see that prices are more on the lower end, number of properties decreases overall as prices increase.

After focusing on prices that are less than $1000, the distribution is more clear. Figure 3 shows that most properties are less than $250, so I decide to analyze more in the range of $100 to $250. Figure 4 shows a overall gradual decrease in number of properties as the price increases. Also, it can be clearly seen that some prices have very high number of properties. It might be due to the fact that landlords tend to choose easy number because those prices are round numbers, such as $100, $120, or $150.

Figure 5 shows the distribution of guest ratings for each properties that they stayed at after cleaning the dataset. It can be seen that majority of the ratings are 5-star; 1-star and 2-star ratings are very rare.

*Code and data are avaliable at:https://github.com/marycx/Paris_Airbnb.git

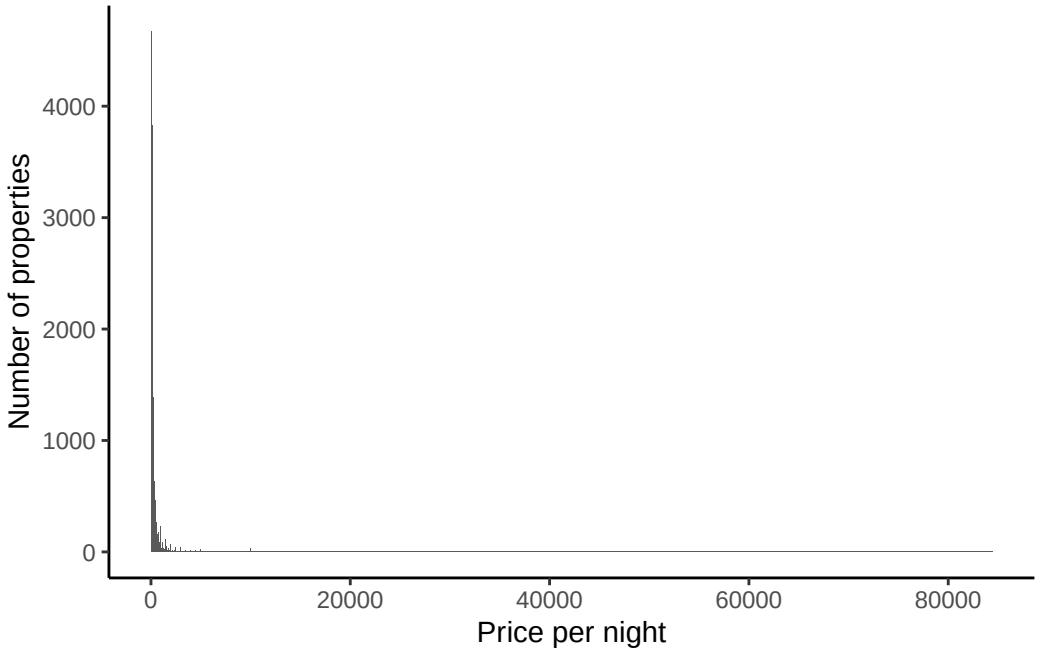

Figure 1: Distribution of prices

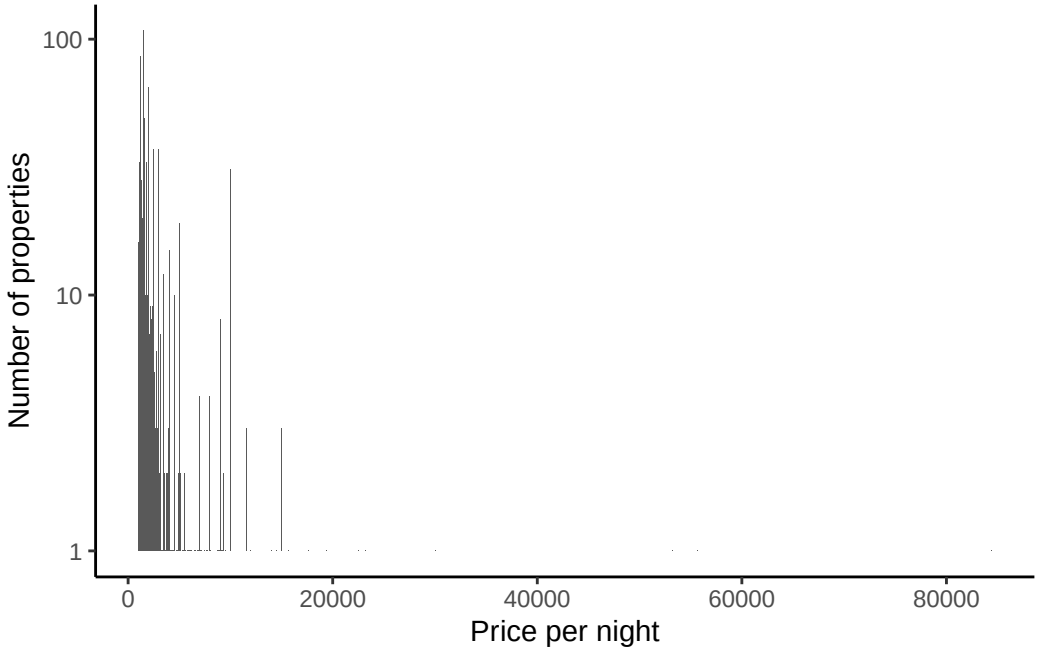

Figure 2: Distribution of prices

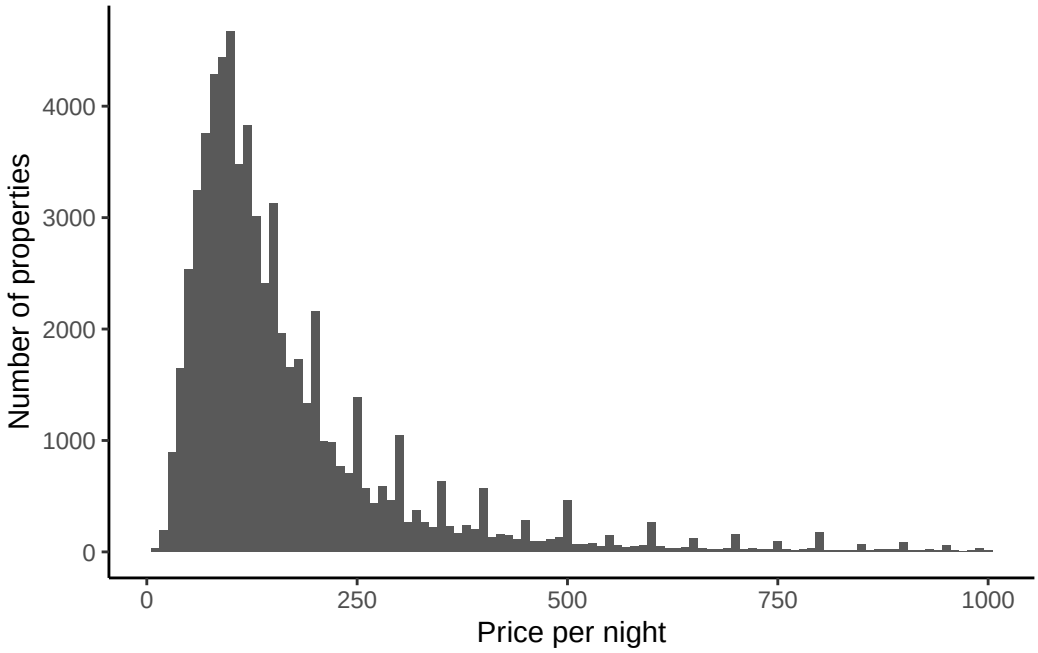

Figure 3: Distribution of prices less than $1000

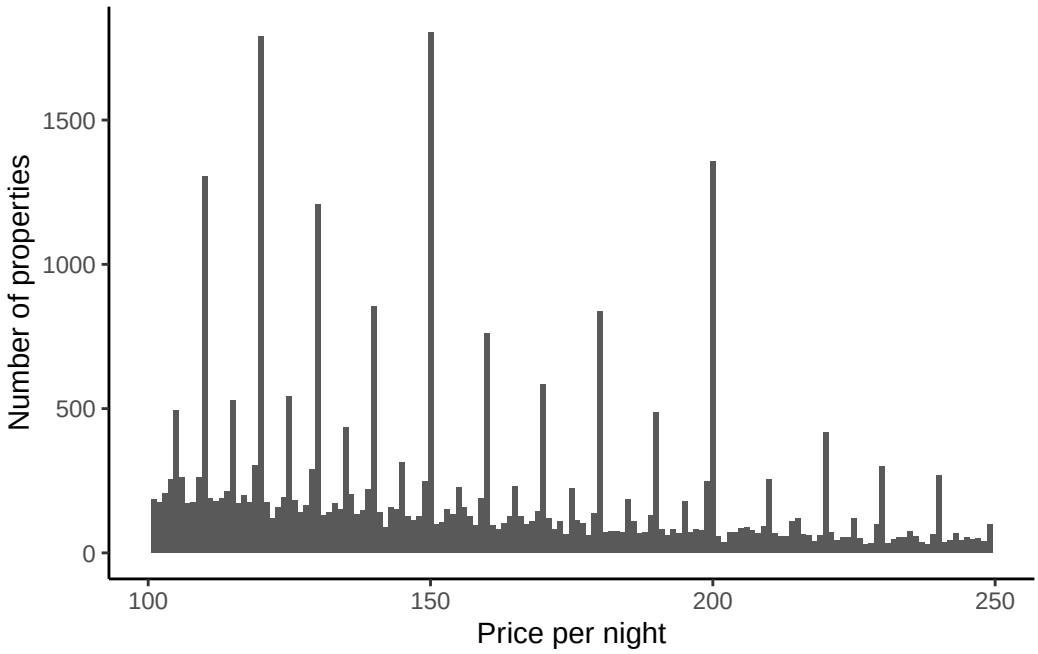

Figure 4: Distribution of prices in range of $250 to $500

I noticed that there are a lot of missing values in this dataset. Using geom_miss_point() from naniar, I plotted Figure 6 to include missing values by host response time.

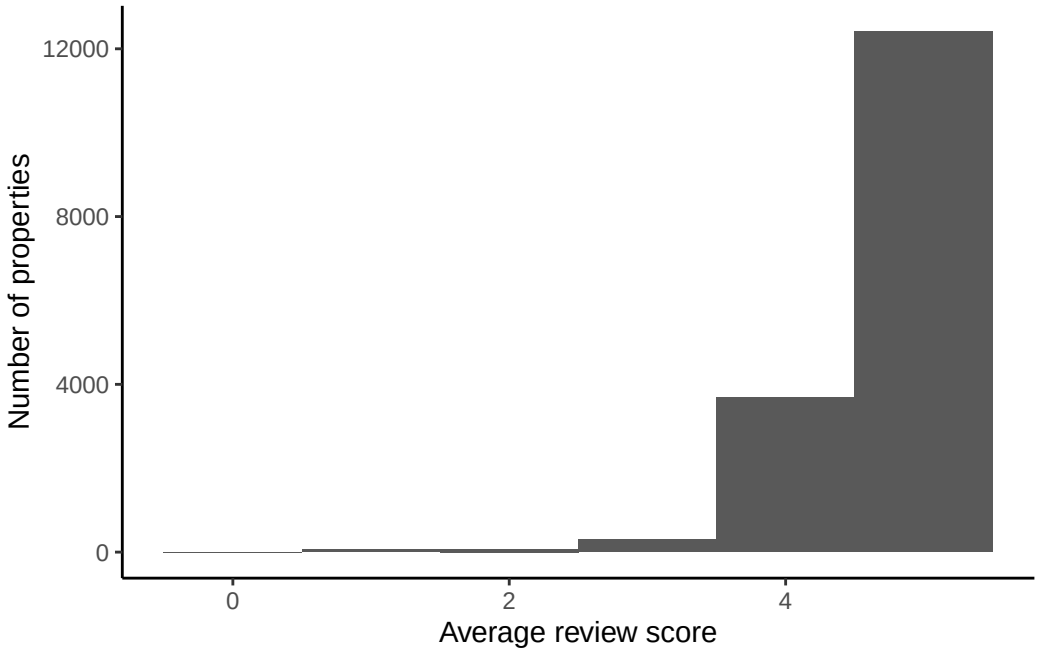

Figure 5: Distribution of review scores rating

We can also look at how many properties a host has on Airbnb in Figure 7 . It can be observed that most hosts have 1 property. Majority of the hosts have less than 10 properties, which makes sense. Few of them have the number of properties around 100 - 1000, which sounds a bit strange, but might be the case if the host is in real estate business.

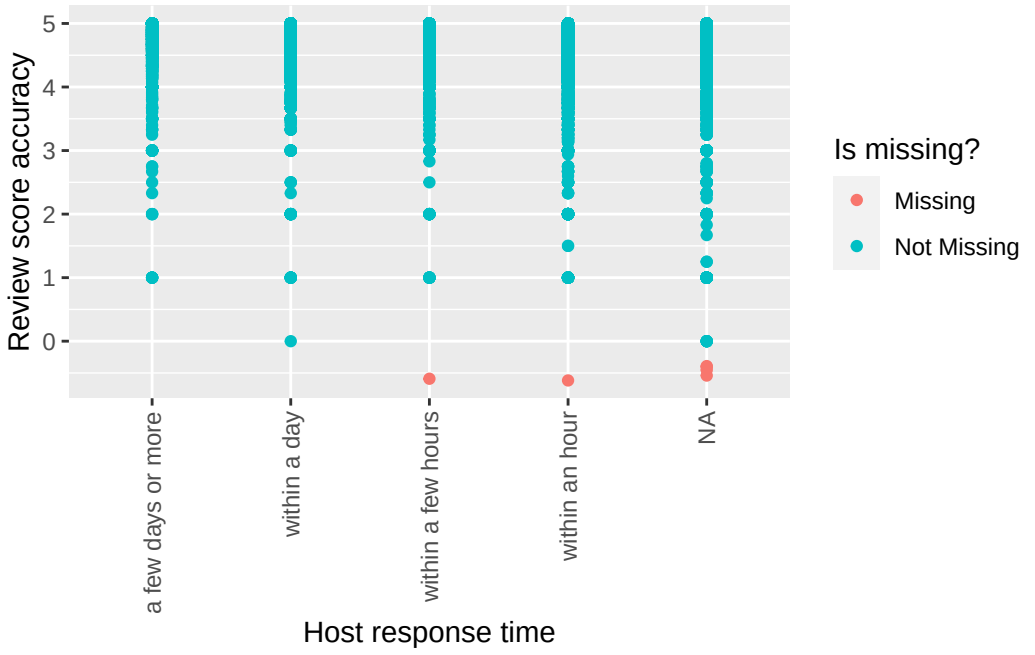

Figure 6: Missing values in Paris Airbnb data, by host response time

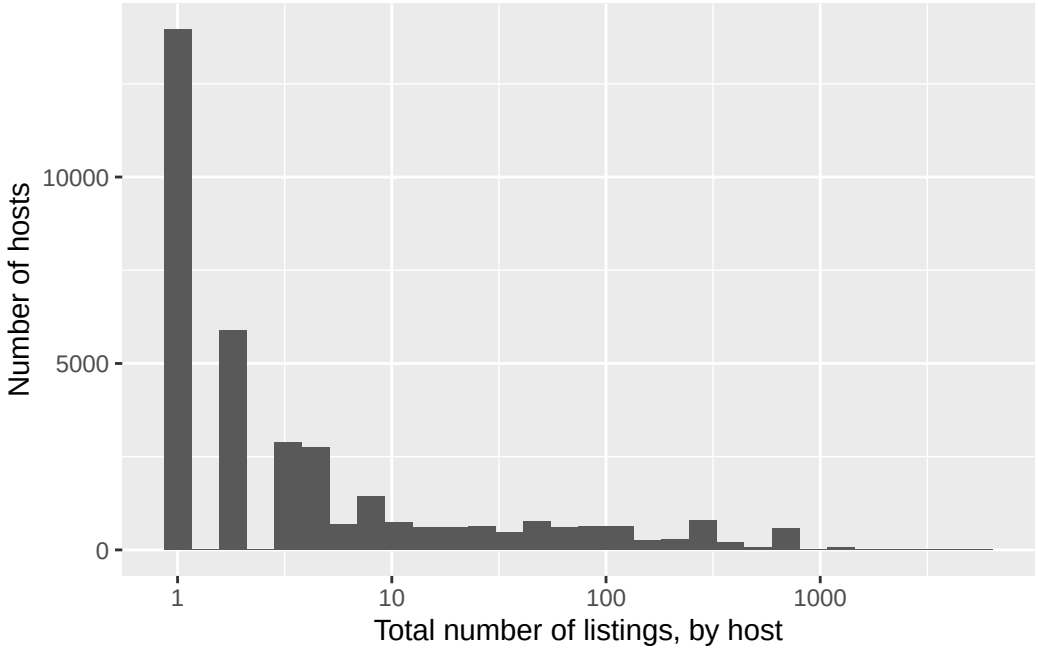

Figure 7: Distribution of the number of properties a host has on Airbnb

Relationships between variables

After looking at individual variables, the relationships between variables will be explored. First the relationship between price and reviews, and whether they are a super-host, for properties with more than 1 review will be analyzed. Figure 8 shows the relationship between these variables. We can see that superhosts generally relate to 5-star ratings and in the price range of 250-500 per night. This makes sense because usually superhosts would receive high ratings as they would provide good housing and services. Also, if the price of the properties are very low, it is less likely that the property has a high standard, thus less likely the host will receive good rating and become a superhost. If the property price is very high, less people will choose to stay there, since airbnb is a more affordable choice for many people. If less people stay there, the host would receive less reviews in general, making it hard to become a superhost.

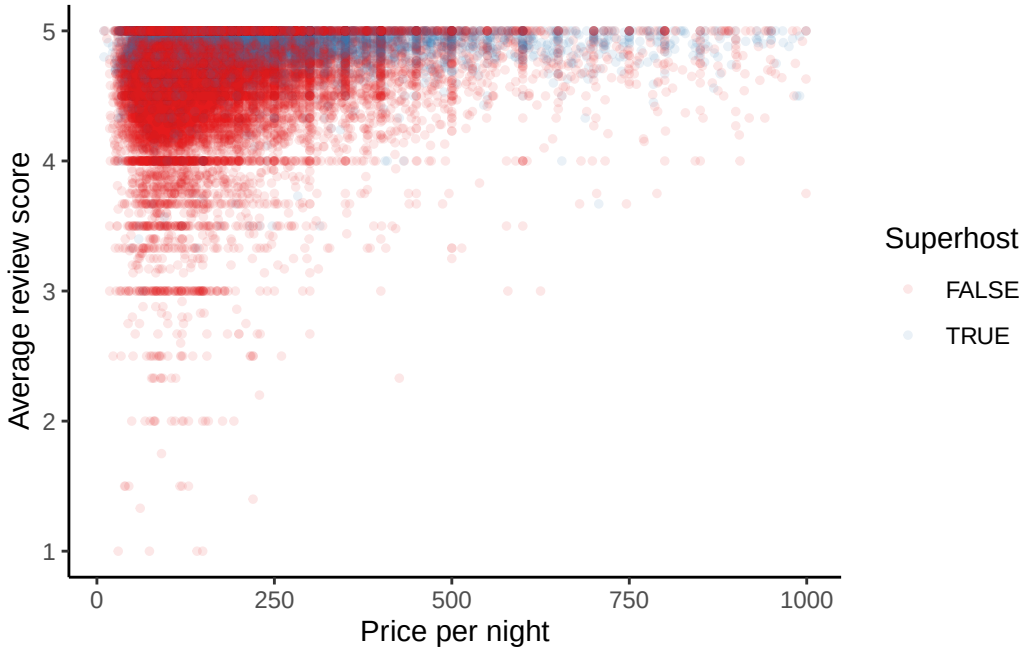

Figure 8: Distribution of the number of properties a host has on Airbnb

A model is run on the dataset to get a better sense of relationships. Here, the model is used to predict whether someone a superhost or not, using variables: "host_response_time" and "review_scores_rating". The output is a binary variable, so we use logistic regression. We can see the model results using modelsummary(). It seems that each variable is positively correlated with the probability of being a superhost, since they all have positive numbers in model results.

	(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

Reference

Get the Data. 2024. http://insideairbnb.com/get-the-data/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.