Airbnb_Paris_France*

Geunchul Shin

Introduction

In this paper, we will be looking at Airbnb listings in Paris, France (12th, December 2023). We will seek the distribution and properties of individual variables of the data. Also we will find possible relationships between variables.

Data

The dataset is from Inside Airbnb which we saved as a local copy. The name if the file is 'airbnb_data.csv'. This file will be added to gitignore as we do not want this to be pushed to GitHub due to its massive size. Speaking of the size, now we will create a parquet file with our favoured variables. We selected important aspects of Airbnbs; such variables such as price, superhost, number of reviews and etc.

After selecting our favoured variables, we cleaned the data. One cleaning process involved with removing dollar sign ($) from the price as we want to keep the variable to be numeric.

[*] Code and data are available at: https://github.com/geunchulshin/Airbnb_paris_france

Distribution and properties of individual variables

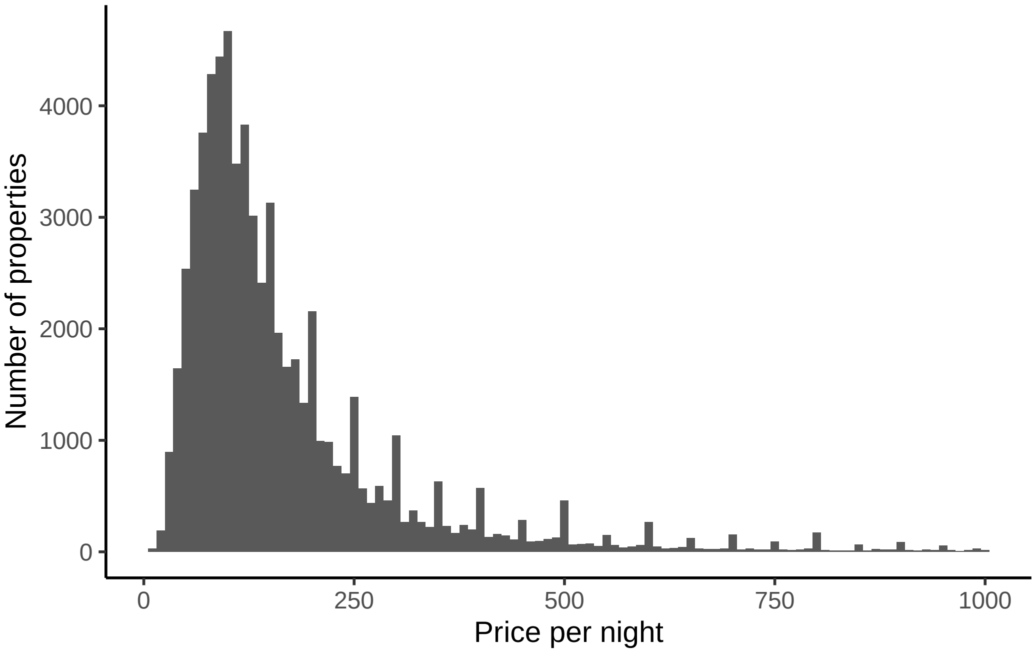

Figure 1: Distribution of prices (less than $1,000) for Airbnb listings in Paris in December 2023

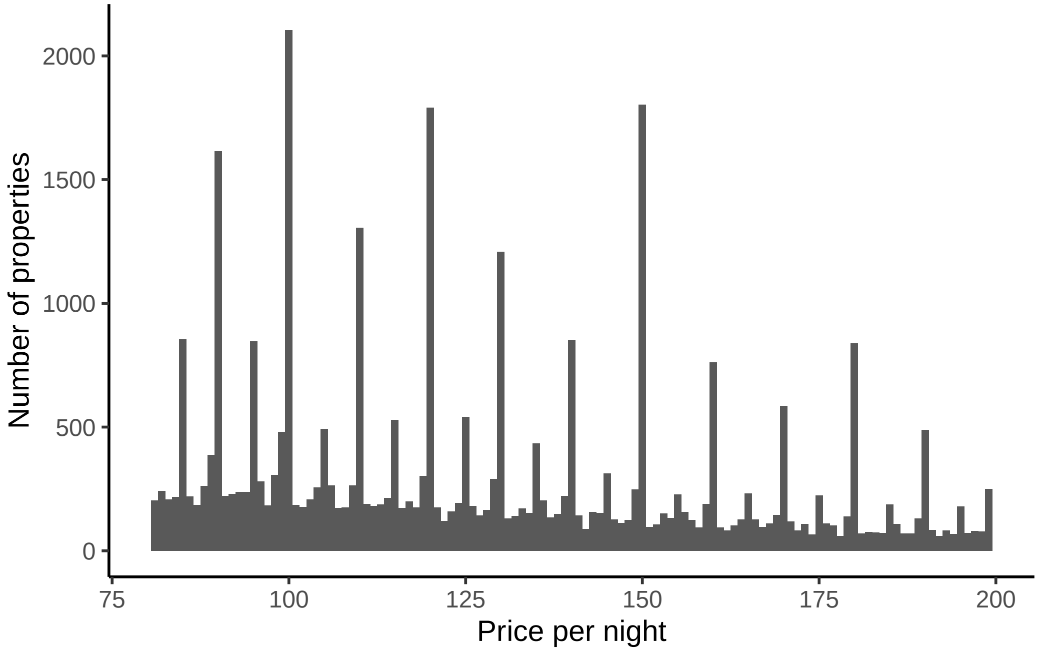

Figure 2: Distribution of prices (between $80 and $200) for Airbnb listings in Paris in December 2023

Let's focus on Airbnbs with *prices* less than 1000. Loooking at Figure 1, we see there are bunching in some prices. If we look Figure 2, in between $80 and $200, it is very clear that there indeed are bunchings on some prices like $120 and $150. From now, we won't bother counting Airbnbs with over $1000 as we consider them as outliers.

Similar to prices, *superhosts* are an important variable when looking at Airbnbs. We got remove of NA values of 'host_is_superhost' variable for clarity. Also, we created a binary variable for efficiency for future usage.

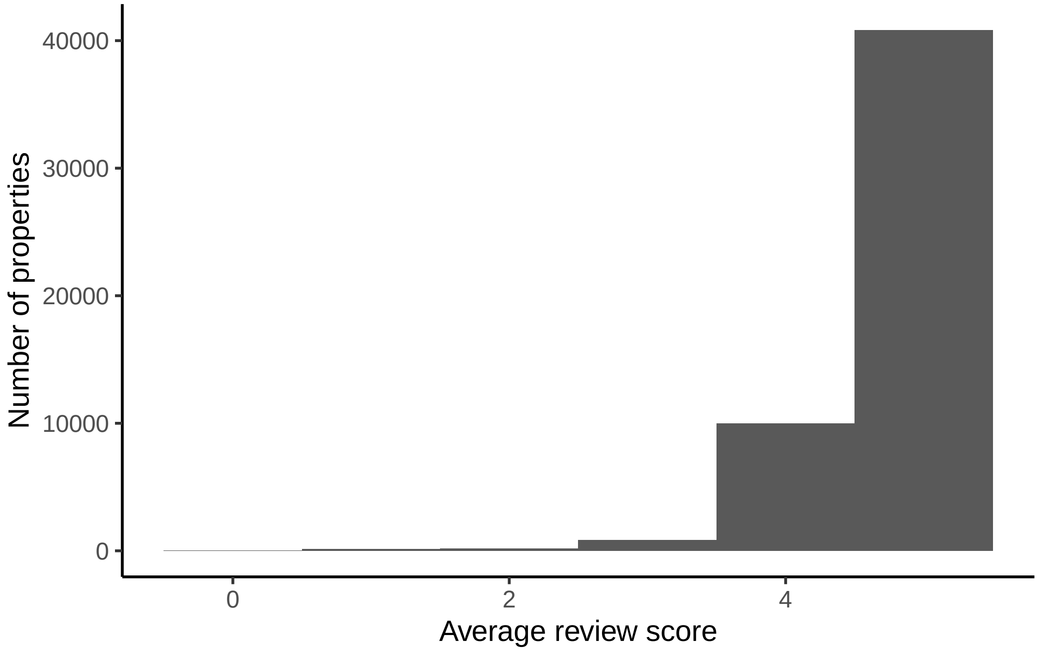

Figure 3: Distribution of review scores for Paris Airbnb rentals in December 2023

Also very similar to superhost property, *review scores* take a big role. We see the distribution of review scores in Figure 3. We also have to take account that there are NAs in this variables. After focusing on non-missing values of the variable, we get Figure 3. The average score seem to be very high with the score of 5.

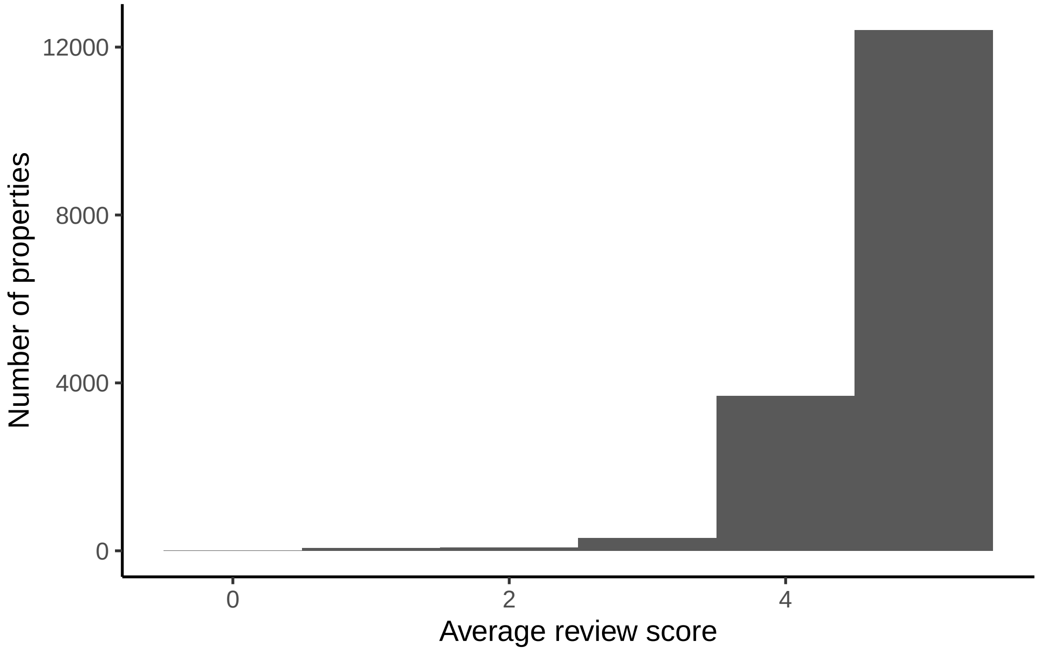

Figure 4: Distribution of review scores for properties with NA response time, for Paris Airbnb rentals in December 2023

A variable we can also consider is Airbnb hosts' response time. But, we found that there are too many NAs in the variable. However, instead of getting rid of the variable, we seeked if there is any possible relation with another variable. As the "NAs" in "host_response_time" were not written properly, we have rewritten it. Then, we checked if there is a relationship with the review score. Here in Figure 4, we constructed distribution of review scores for properties with NA response time.

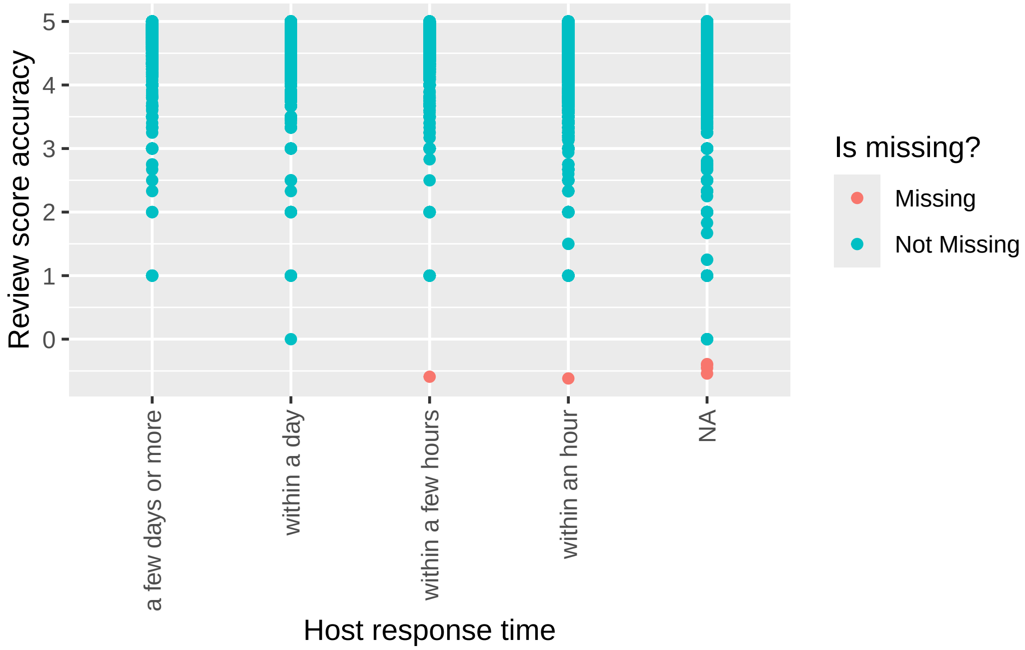

Figure 5: Missing values in Paris Airbnb data, by host response time

It is also interesting how many missing values are being dropped by looking at Figure 5. We removed hosts with NA as their response time and this almost removed 19 percent of the observations.

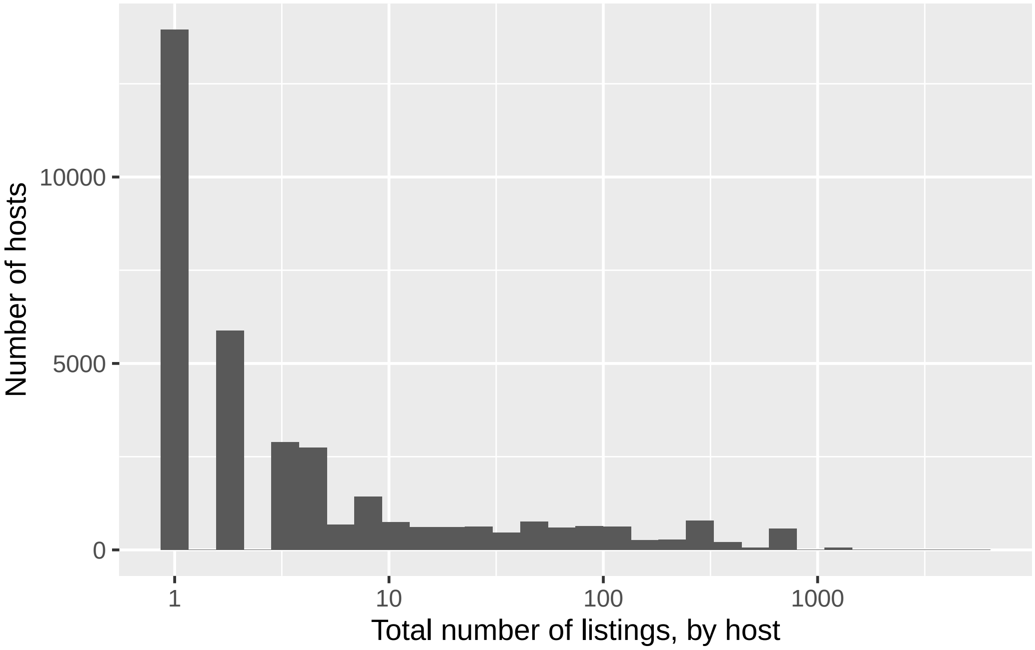

Figure 6: Distribution of the number of properties a host has on Airbnb, for Paris Airbnb rentals in December 2023

We also wondered how many properties a host might have on Airbnb. Here in Figure 6, we see how many properties a host has on Airbnb. It seems that majority of the hosts owned a single property. Therefore, we will put our focus more on these.

Relationship between variables