Airbnb data in Paris, France*

Jacob Gilbert

The following is analysis on data provided by data.insideairbnb.com and analysis is done using R, and R packages ...

Firstly lets view some of the pricing data:

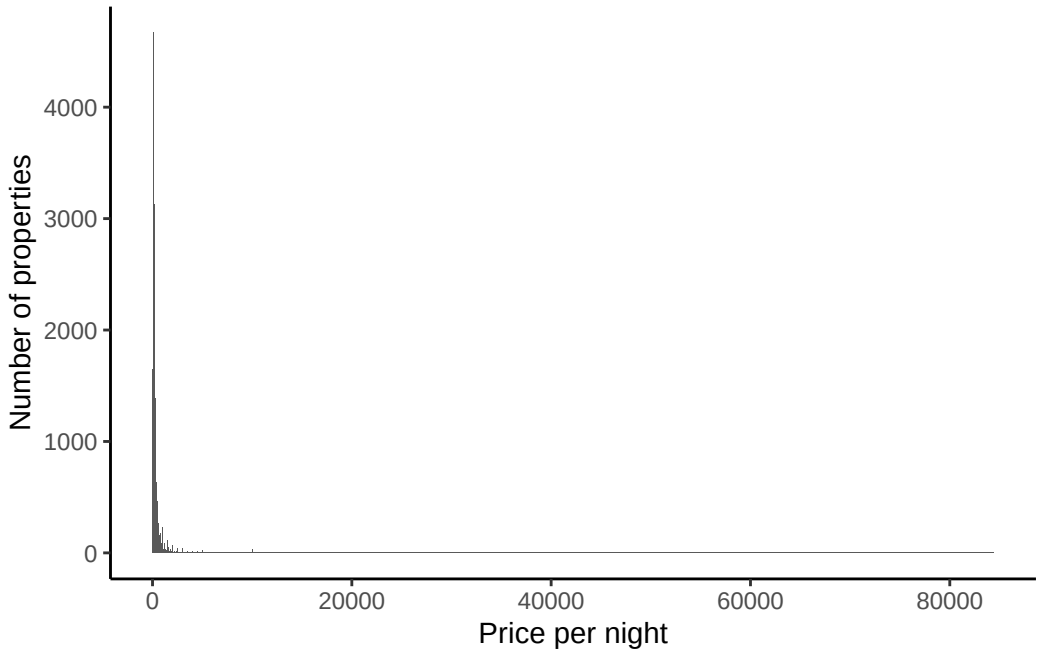
```
[1] "$150.00" "$146.00" "$110.00" "$140.00" "$180.00" "$71.00"
```

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

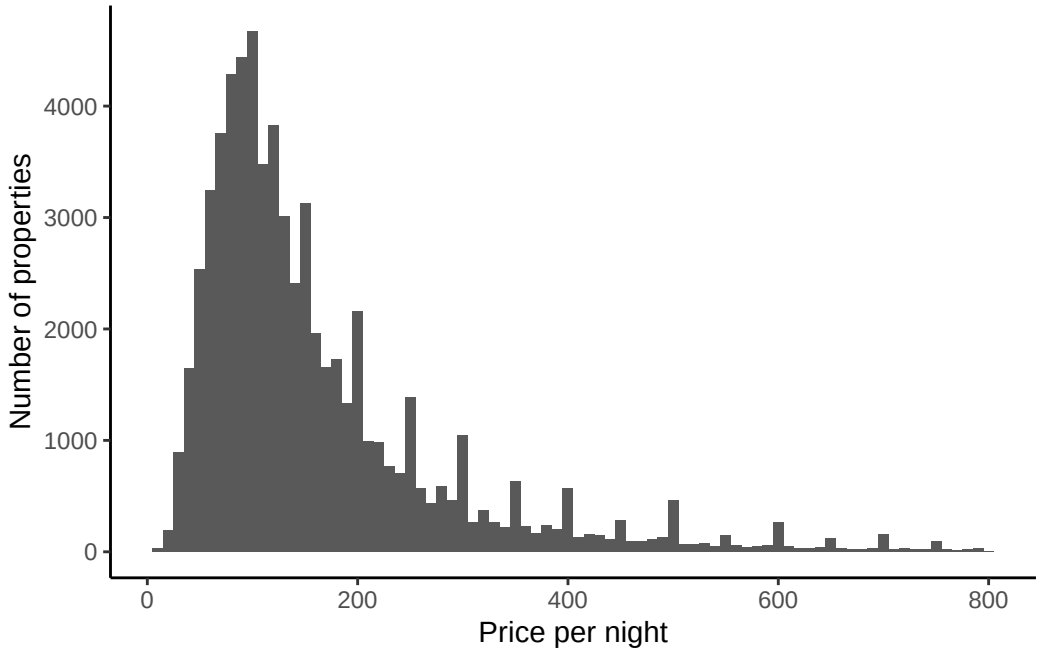
```
# A tibble: 1,550 x 1
   price
   <chr>
 1 $1,200.00
 2 $8,000.00
 3 $7,000.00
 4 $1,997.00
 5 $1,000.00
 6 $1,286.00
 7 $2,300.00
 8 $1,500.00
 9 $1,200.00
10 $1,357.00
# i 1,540 more rows
```

Viewing the distribution of prices we see a poor image of the distibution being caused by few very high prices.

*Github link: https://github.com/JfpGilbert0/STA302/tree/main/winter/w8



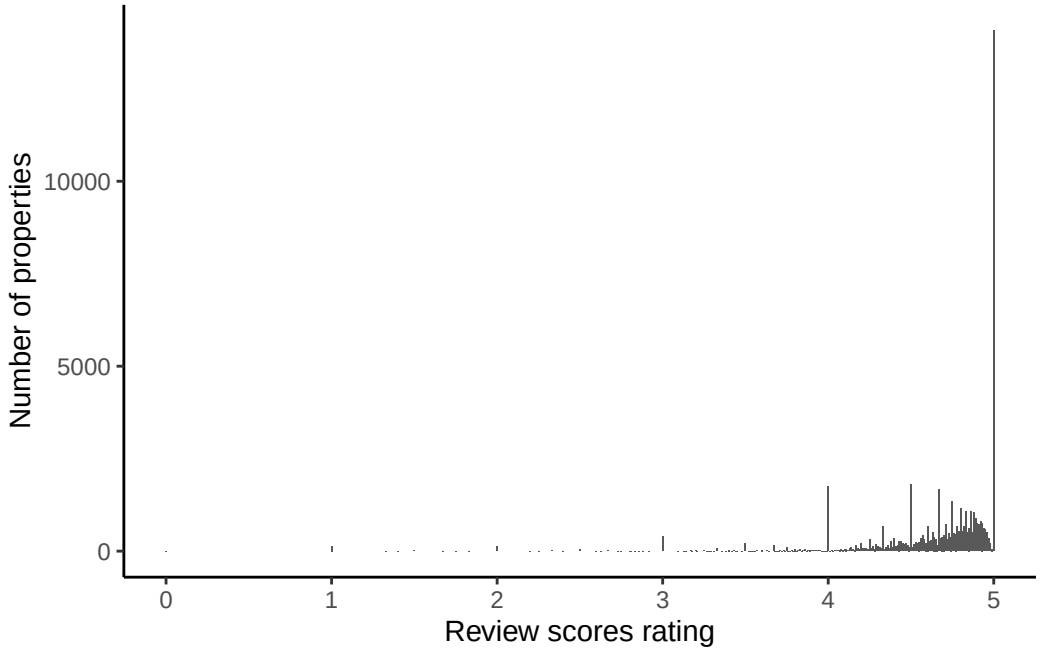
Removing those prices over 800 we get a better image of the bulk of the data.




```
# 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>
```

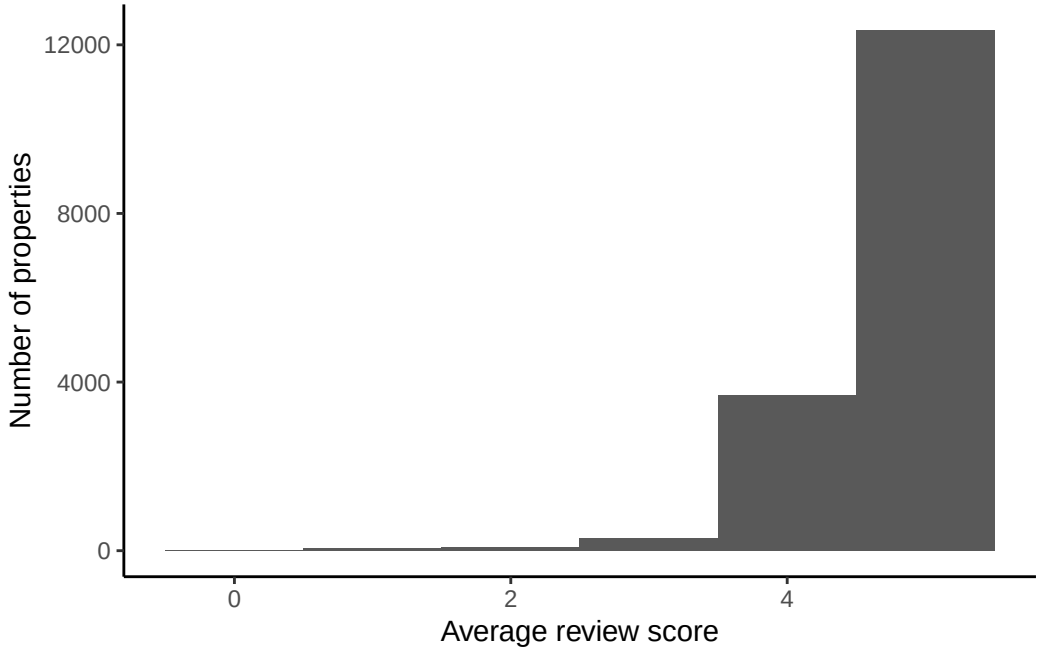
Afteer cleaning some of the data around null superhost entries we can observe a better understanding of the ratings distribution

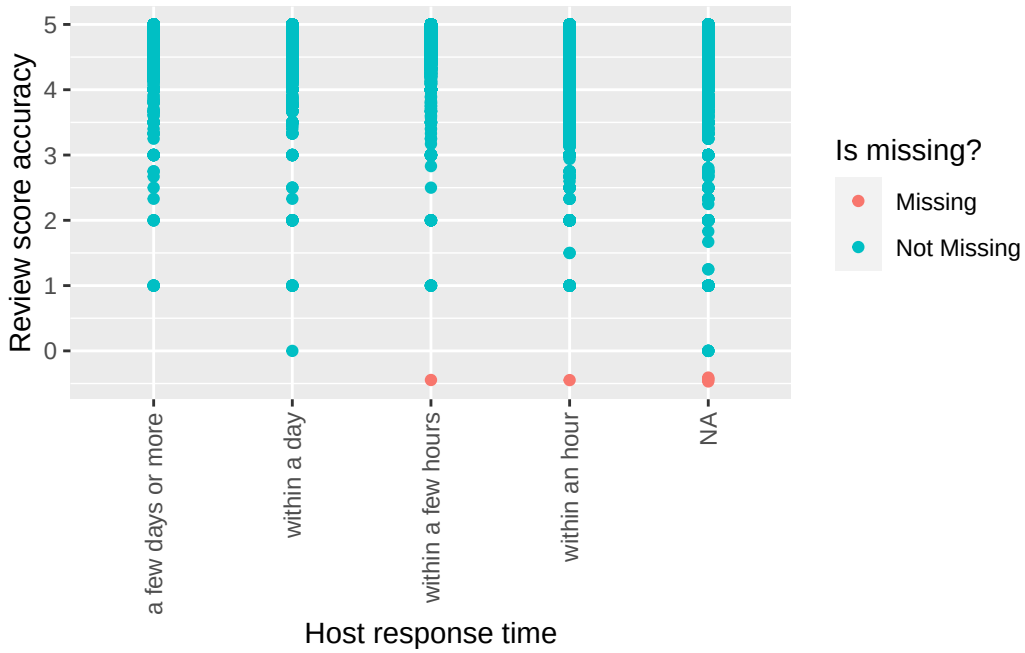



```
# A tibble: 6 x 2
  host_response_time     n
  <chr>              <int>
```

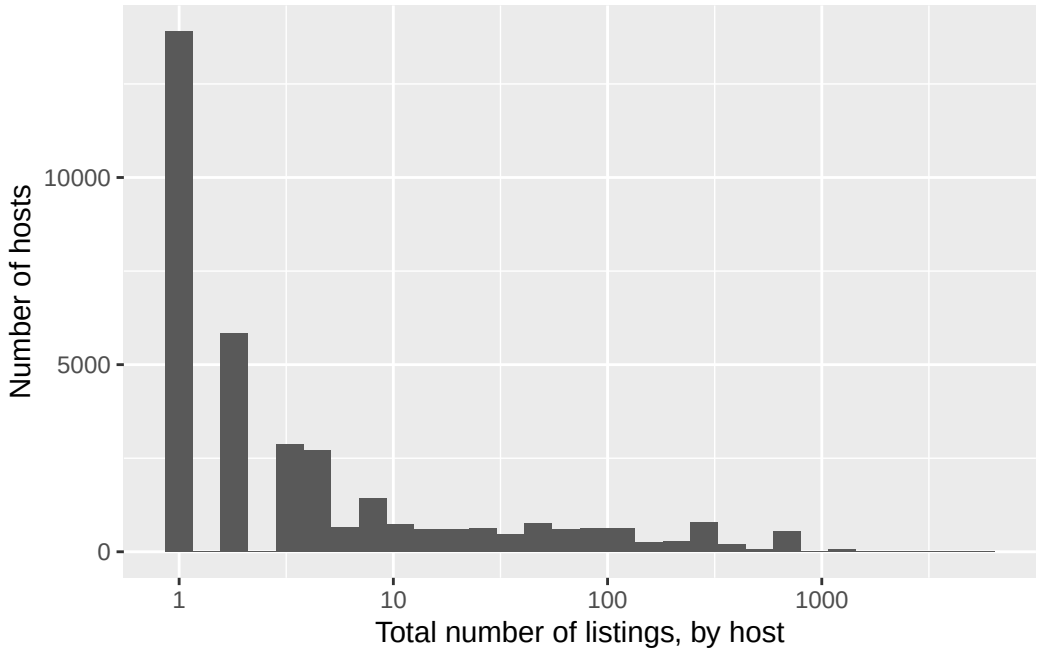
```
1 N/A                16462
2 a few days or more  1240
3 within a day        5263
4 within a few hours  6768
5 within an hour     21967
6 <NA>                   2
```

Lets look at the relationships between some of these differen characteristics. Stating with Host response time and their rating.





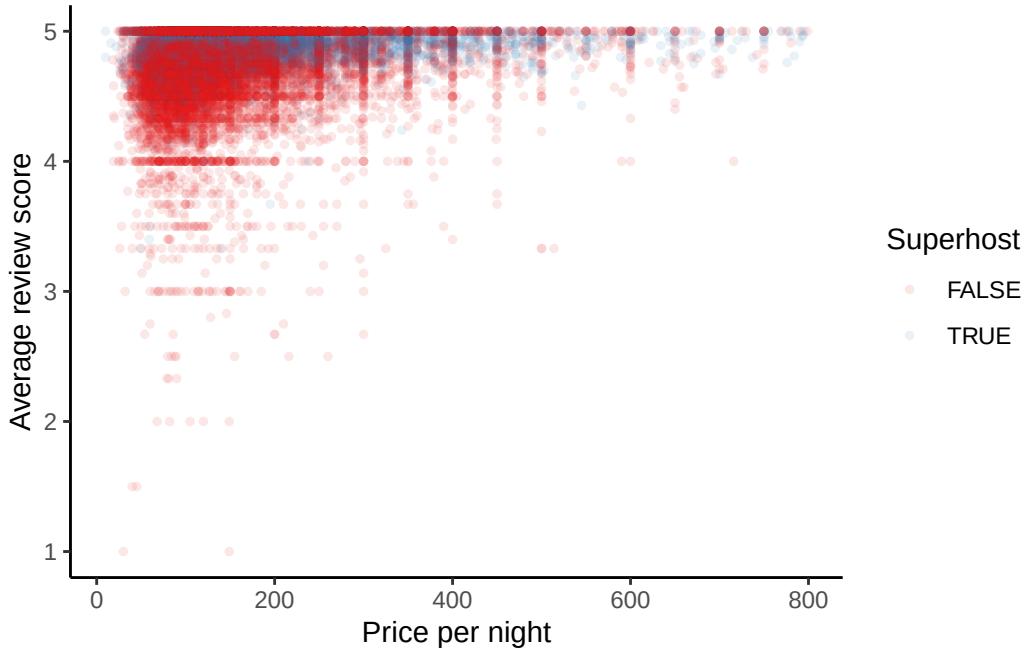
We can also observe how many properties hosts tend to have.



A very small group of hosts has a massive amount of listings. We will focus on those with 1 property.

```
# 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>
```

Here we observe price and rating's relationship aswell as where wihing this relationship we observe the superhosts.



A summary of the host data:

```
# A tibble: 2 x 3
  host_is_superhost     n proportion
  <lgl>             <int>      <dbl>
1 FALSE             15770       0.72
2 TRUE               6207       0.28
```

	(1)
(Intercept)	−16.272
	(0.482)
host_response_timewithin a day	2.015
	(0.211)
host_response_timewithin a few hours	2.696
	(0.210)
host_response_timewithin an hour	2.972
	(0.209)
review_scores_rating	2.626
	(0.089)
Num.Obs.	21 977
AIC	24 081.7
BIC	24 121.6
Log.Lik.	−12 035.825
RMSE	0.43

	host_is_superhost			
host_response_time	FALSE		TRUE	
a few days or more	6%	(952)	0%	(24)
within a day	22%	(3,501)	12%	(765)
within a few hours	24%	(3,786)	26%	(1,610)
within an hour	48%	(7,531)	61%	(3,808)

Neighborhood summary:

neighbourhood_cleansed	n	percent
Buttes-Montmartre	2840	12.9%
Popincourt	2198	10.0%
Entrepôt	1708	7.8%
Vaugirard	1676	7.6%
Ménilmontant	1437	6.5%
Buttes-Chaumont	1432	6.5%

Finally a regression for determiniting chances of becoming a superhost