Explore Airbnb listings in Paris, France, as at 04 March 2024*

Yimiao Yuan

March 4, 2024

1 Download and Save Data

```
# exploratory: create a parquet file with selected variables
airbnb_select <-
  airbnb_raw |>
  select(
    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,
    review_scores_value
  )

rm(airbnb_raw)
```

```
# save the parquet file
write_parquet(
```

*Code and data are available at: https://github.com/YimiaoYuan09/Airbnb_EDA_Paris

```
  x = airbnb_select,
  sink =
    "../../inputs/data/2024-03-04-paris-airbnblistings-select_variables.parquet"
  )
```

2 Distribution and properties of Price

```
airbnb_select$price |>
  head()
```

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

```
# get character from price
airbnb_select$price |>
  str_split("") |>
  unlist() |>
  unique()
```

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

```
# , value
airbnb_select |>
  select(price) |>
  filter(str_detect(price, ","))
```

```
# 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
```

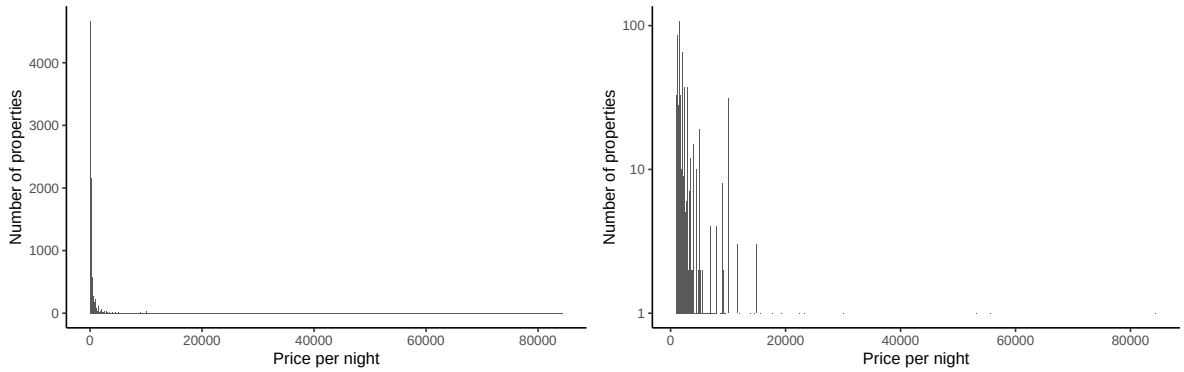
```
10 $1,357.00
# i 1,540 more rows
```

```r
# remove $
airbnb_select <-
  airbnb_select |>
  mutate(
    price = str_remove_all(price, "[\\$,]"),
    price = as.integer(price)
  )
```

```r
# distribution of price
airbnb_select |>
  ggplot(aes(x = price)) +
  geom_histogram(binwidth = 10) +
  theme_classic() +
  labs(
    x = "Price per night",
    y = "Number of properties"
  )

# distribution of price on log scale
airbnb_select |>
  filter(price > 1000) |>
  ggplot(aes(x = price)) +
  geom_histogram(binwidth = 10) +
  theme_classic() +
  labs(
    x = "Price per night",
    y = "Number of properties"
  ) +
  scale_y_log10()
```

```r
# focus on price < 1000
airbnb_select |>
  filter(price < 1000) |>
  ggplot(aes(x = price)) +
  geom_histogram(binwidth = 10) +
  theme_classic() +
  labs(
    x = "Price per night",
```



(a) Distribution of prices (b) Using the log scale for prices more than $1,000

Figure 1: Distribution of prices of Paris Airbnb rentals in March 2024

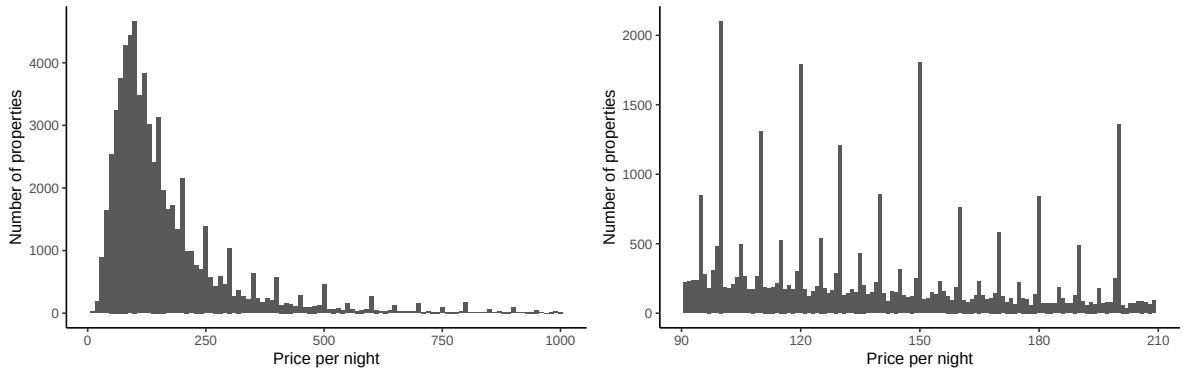
```
    y = "Number of properties"
  )

airbnb_select |>
  filter(price > 90) |>
  filter(price < 210) |>
  ggplot(aes(x = price)) +
  geom_histogram(binwidth = 1) +
  theme_classic() +
  labs(
    x = "Price per night",
    y = "Number of properties"
  )
```

3 Distribution and properties of Superhost

```
# NA value
airbnb_data_less_1000 |>
  filter(is.na(host_is_superhost))
```

```
# 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
```



(a) Prices less than $1,000 suggest some bunching

(b) Prices between $90 and $210 illustrate the bunching more clearly

Figure 2: Distribution of prices of Paris Airbnb rentals in March 2024

```
 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>
```

```
# NA in review scores rating
airbnb_data_no_superhost_na |>
  filter(is.na(review_scores_rating)) |>
  nrow()
```

```
[1] 13497
```

```
airbnb_data_no_superhost_na |>
  filter(is.na(review_scores_rating)) |>
  select(number_of_reviews) |>
```

```
  table()
```

```
number_of_reviews
    0
13497
```

```
# no NA in review scores rating
airbnb_data_no_superhost_na |>
  filter(!is.na(review_scores_rating)) |>
  ggplot(aes(x = review_scores_rating)) +
  geom_histogram(binwidth = 1) +
  theme_classic() +
  labs(
    x = "Average review score",
    y = "Number of properties"
  )
```

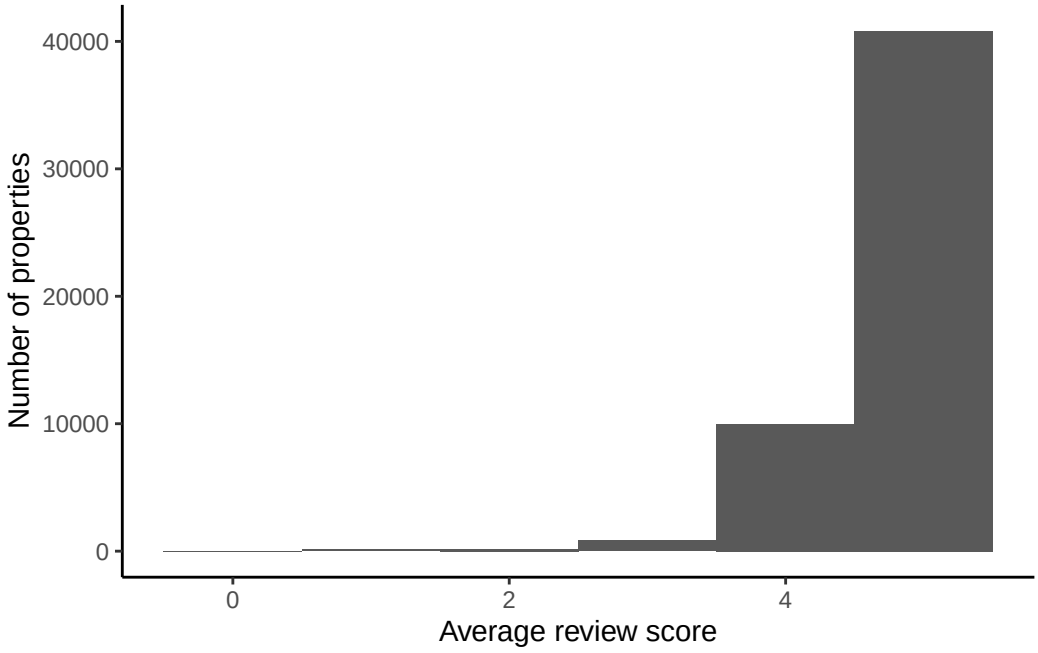


Figure 3: Distribution of review scores for Paris Airbnb rentals in March 2024

```r
# host response time
airbnb_data_has_reviews |>
  count(host_response_time)
```

```
# 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
```

```r
# host response time NA
# relationship with review scores rating
airbnb_data_has_reviews |>
  filter(is.na(host_response_time)) |>
  ggplot(aes(x = review_scores_rating)) +
  geom_histogram(binwidth = 1) +
  theme_classic() +
  labs(
    x = "Average review score",
    y = "Number of properties"
  )
```

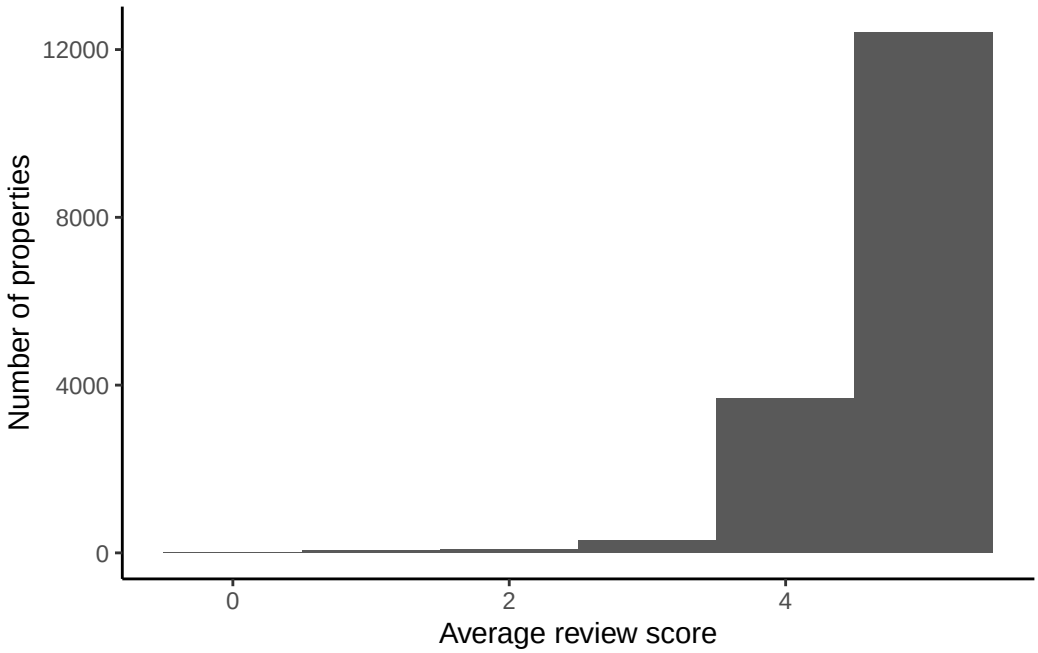


Figure 4: Distribution of review scores for properties with NA response time, for Paris Airbnb rentals in March 2024

```
# include missing data
airbnb_data_has_reviews |>
  ggplot(aes(
    x = host_response_time,
    y = review_scores_accuracy
  )) +
  geom_miss_point() +
  labs(
    x = "Host response time",
    y = "Review score accuracy",
    color = "Is missing?"
  ) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

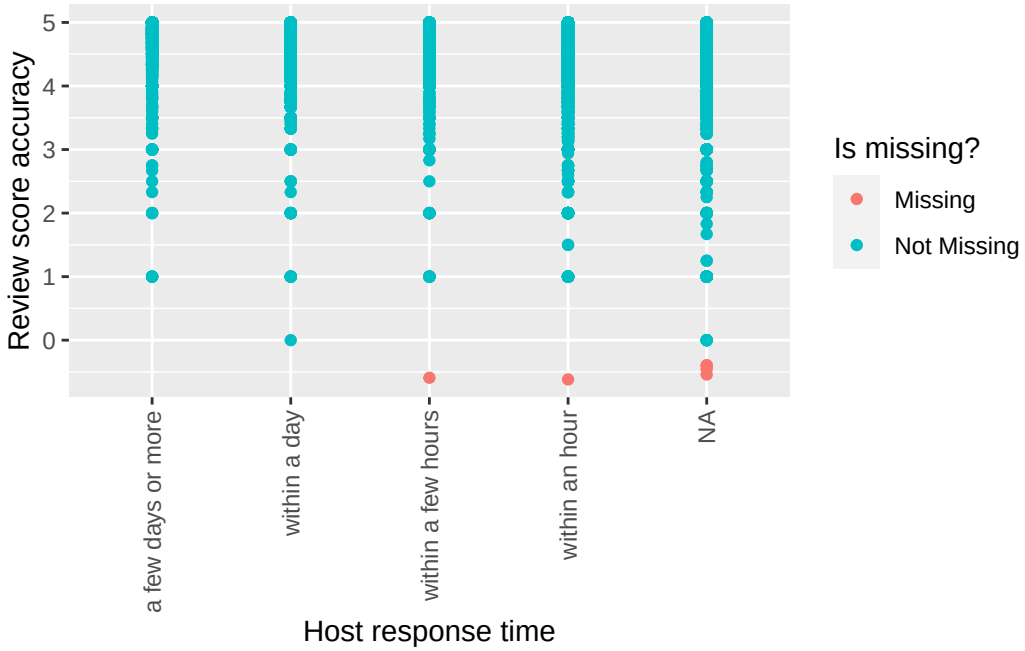


Figure 5: Missing values in Paris Airbnb data, by host response time

```
# remove NA in host_response_time
# superhost, has revies, has response time
airbnb_select <-
  airbnb_data_has_reviews |>
  filter(!is.na(host_response_time))
```

4 Distribution and properties of Host Properties

```
# how many properties a host has on Airbnb
airbnb_select |>
  ggplot(aes(x = host_total_listings_count)) +
  geom_histogram() +
  scale_x_log10() +
  labs(
    x = "Total number of listings, by host",
    y = "Number of hosts"
  )
```

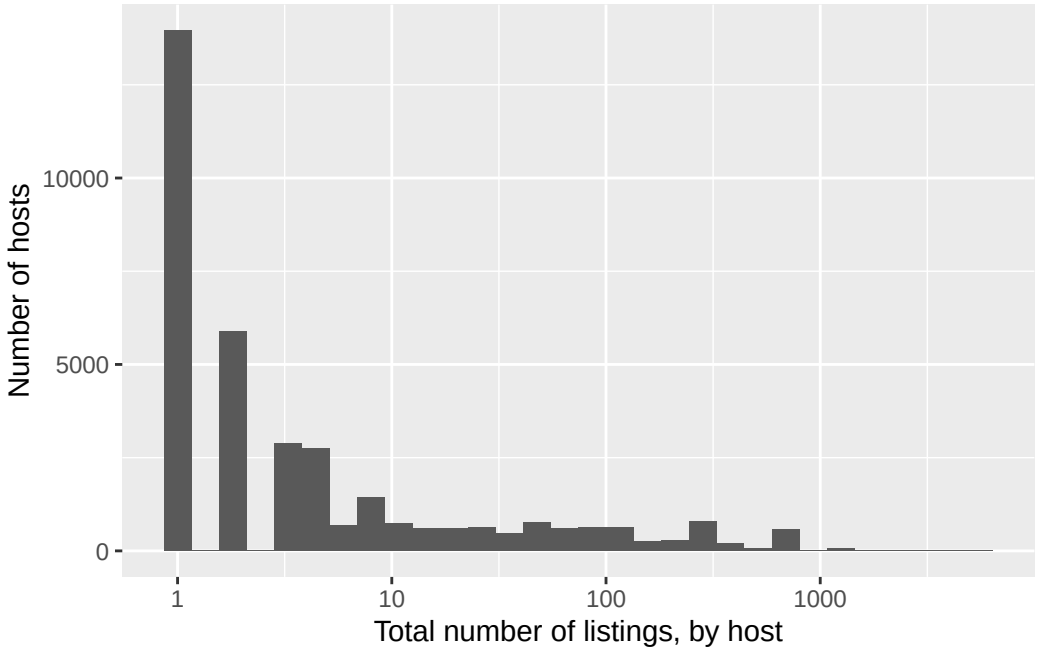


Figure 6: Distribution of the number of properties a host has on Airbnb, for Paris Airbnb rentals in March 2024

```
# host with number of listings > 500
airbnb_select |>
  filter(host_total_listings_count >= 500) |>
  head()
```

```
# 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>
```

```r
# focus on host with only 1 property
airbnb_select <-
  airbnb_select |>
  add_count(host_id) |>
  filter(n == 1) |>
  select(-n)
```

5 Relationship between prices and reviews, superhosts, number of properties, neighborhood

```r
# more then 1 review
airbnb_select |>
  filter(number_of_reviews > 1) |>
  ggplot(aes(x = price, y = review_scores_rating,
             color = host_is_superhost)) +
  geom_point(size = 1, alpha = 0.1) +
  theme_classic() +
  labs(
    x = "Price per night",
    y = "Average review score",
    color = "Superhost"
  ) +
  scale_color_brewer(palette = "Set1")
```

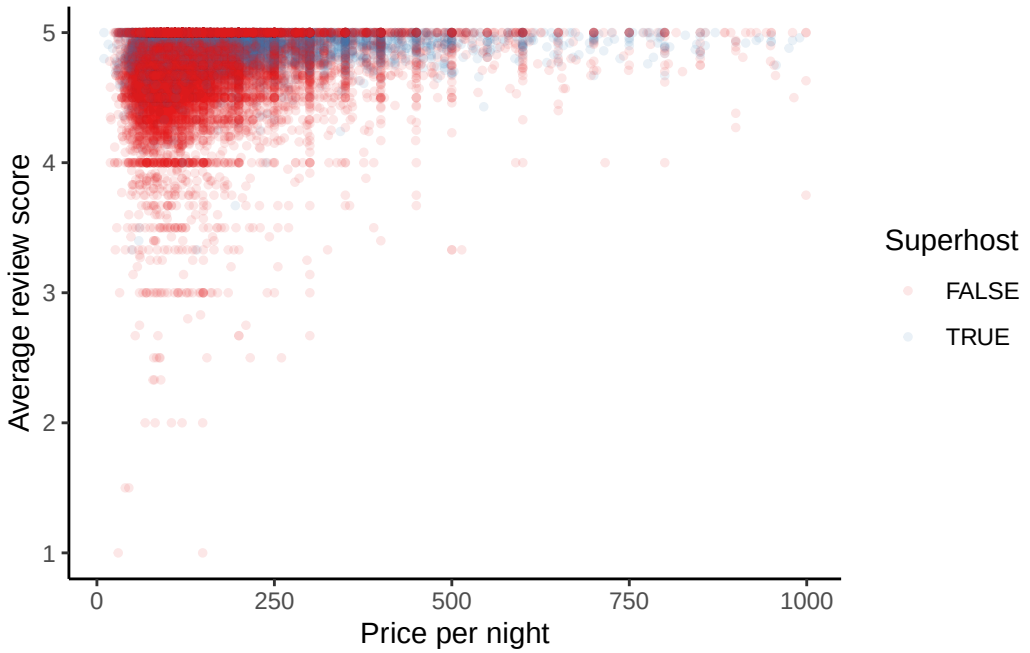


Figure 7: Relationship between price and review and whether a host is a superhost, for Paris Airbnb rentals in March 2024

```
# proportion of superhost
airbnb_select |>
  count(host_is_superhost) |>
  mutate(
    proportion = n / sum(n),
    proportion = round(proportion, digits = 2)
  )
```

```
# A tibble: 2 x 3
  host_is_superhost     n proportion
  <lgl>             <int>      <dbl>
1 FALSE             15820       0.72
2 TRUE               6227       0.28
```

```
# host response time, by superhost
airbnb_select |>
  tabyl(host_response_time, host_is_superhost) |>
  adorn_percentages("col") |>
```

```
  adorn_pct_formatting(digits = 0) |>
  adorn_ns() |>
  adorn_title()
```

```
                   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)
```

```
# neighbourhood
airbnb_select |>
  tabyl(neighbourhood_cleansed) |>
  adorn_pct_formatting() |>
  arrange(-n) |>
  filter(n > 100) |>
  adorn_totals("row") |>
  head()
```

```
neighbourhood_cleansed    n percent
     Buttes-Montmartre 2842   12.9%
            Popincourt 2202   10.0%
              Entrepôt 1713    7.8%
             Vaugirard 1681    7.6%
          Ménilmontant 1438    6.5%
       Buttes-Chaumont 1430    6.5%
```

6 Model

```
# forecast whether someone is a superhost
# logistic regression
# affected by faster responses and better reviews
logistic_reg_superhost_response_review <-
  glm(
    host_is_superhost ~
      host_response_time +
      review_scores_rating,
    data = airbnb_select,
```

	(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

```
    family = binomial
  )
```

```
modelsummary(logistic_reg_superhost_response_review)
```

7 Save Analysis Dataset

```
# save analysis data
write_parquet(
  x = airbnb_select,
  sink = "../../outputs/data/2024-03-04-paris-airbnblistings-analysis_dataset.parquet"
  )
```