Milestone 03 - Group 09

Daniel Hadley, Kristina Wright

March 14, 2020

Airbnb Listings for Barcelona

Introduction

Airbnb, Inc. is a company founded in 2008 that offers an online marketplace connecting people who offer lodging with people who require accomodations in that locale. The company does not own any of the listed properties and operates as a broker, collecting commissions once a lodging is booked. As a direct competitor to hotels, we are interested in how the users listing properties determine the price they charge.

When accomodations are offered through Airbnb, the person listing the property is called a host, and they must provide a variety of information about the listing including price, neighborhood, type of accommodations offered, and the minimum number of nights a guest must stay if they want to make a booking. In addition to information provided by the host, Airbnb collects and disseminates information about the listing which we use to perform our analysis.

The data is collected using public information compiled from the Airbnb website. Specific collection techniques are not specified, though the Inside Airbnb website states that it uses Open Source technologies such as D3, Boostrap, jQuery, etc. to collect the data and much code was "copied and pasted" from the internet. A major contributer to this code, Tom Slee, described it as a "scrape" of the Airbnb website for each city.

Data Description

The dataset used in this analysis is collected and offered by Inside Airbnb, an independent, non-commercial project started by Murray Cox and John Morris. Their goal is to allow people to see how Airbnb might be affecting the residential housing market. We use the summary data for listings, since it includes the data we are interested in exploring and is more manageable, size-wise, than the detailed listings data.

The data used in this analysis was compiled on November 9, 2019 and includes 20,428 Airbnb listings that travellers see when using the Airbnb website to find accommodations in Barcelona, Spain. The table below describes the available data for each listing in the dataset.

Variable Name	Column Name	Type of Data	Description
Listing ID	`id`	Categorical/Numeric	Numeric identifier unique to each listing
Name	`name`	Character	Short title for the listing provided by the host
Host ID	`host_id`	Categorical/Numeric	Numeric identifier for the host of the listing
Host Name	`host_name`	Categorical/String	Name of the host or hosts of the listing provided by the host(s) to Airbnb

Variable Name	Column Name	Type of Data	Description
Neighbourhood Group	`neighbourhood_group`	Categorical/String	Districts of Barcelona as determined by the coordinates of the listing and the city's definition of its districts; this data is not the data provided by the host
Neighbourhood	`neighbourhood`	Categorical/String	Neighbourhoods of Barcelona are smaller geographical areas than districts and are determined by the coordinates of the listing and compared to the city's boundaries of its neighbourhoods; this data is not the neighbourhood provided by the host
Latitude	`latitude`	Numeric	Latitude coordinates of the listing
Longitude	`longitude`	Numeric	Longitude coordinates of the listing
Type of Accommodation	`room_type`	Categorical/String	Type of accommodations specify whether the listing is for an entire home or apartment, a private room in a shared home or apartment, a hotel room, or a shared room
Price	`price`	Numeric	The price per night, in euros, to book a listing
Minimum Stay	`minimum_nights`	Numeric	The minimum number of nights that a guest must reserve in order to book a listing
Number of Reviews	`number_of_reviews`	Numeric	The number of reviews left by guests after their stay
Last Review	`last_review`	Date	The date of the last review left by a guest
Reviews per Month	`reviews_per_month`	Numeric	The number of reviews left by guests of a listing divided by the number of months the listing has been active
Number of Listings by Host	`calculated_host_listings_count`	Numeric	A count of the number of listings under the same Host Name

Variable Name	Column Name	Type of Data	Description
Availability	`availability_365`	Numeric	The number of days over the next 365 days that the listing can be booked by guests; calculated as 365 minus booked days minus days listing is unavailable as per the host

Exploring the Dataset

Remove Unwanted Data

In this section, we remove columns from the dataset that should have no fundamental influence on listing price. This includes the short title of the listing (`name`), the name of the host(s) (`host_name`), and the availability of the listing over the next 365 days (`availability_365`). While there might end up being a relation between availability and price, since cheap listings for a desirable neighbourhood are likely to be booked, this relationship is backwards; we want to find factors that affect the listing price, not factors affected by the listing price.

Rename Columns

Some of the column names are a little long, so we perform the following renamings:

- `neighbourhood_group` is renamed to `district`
- `minimum_nights` is renamed to `min_stay`
- `number_of_reviews` is renamed to `reviews`
- `calculated_host_listings_count` is renamed to `host_listings`

Filter Data

A few extreme outliers skew the density of the price of listings to the right. As a result, we exclude the top 2.5% of listings. Then, we exclude listings with a minimum stay over 5 nights. This should help to limit listings that are catered to tourists by eliminating listings that are better classified as short-term rentals.

Price Density

A kernel density plot is presented for listing prices. An interesting observation from the price density is the tendency for people to price their listings in increments of 50 Euros. For example, the Density Plot, we see multi-modes, where each mode after the largest mode occurs at every 50 Euro increment along the x-axis

Correlogram

Based on the correlogram shown below there is little correlation between the 6 numerical variables presented. All positive correlations are in blue, and all negative correlations are in red.

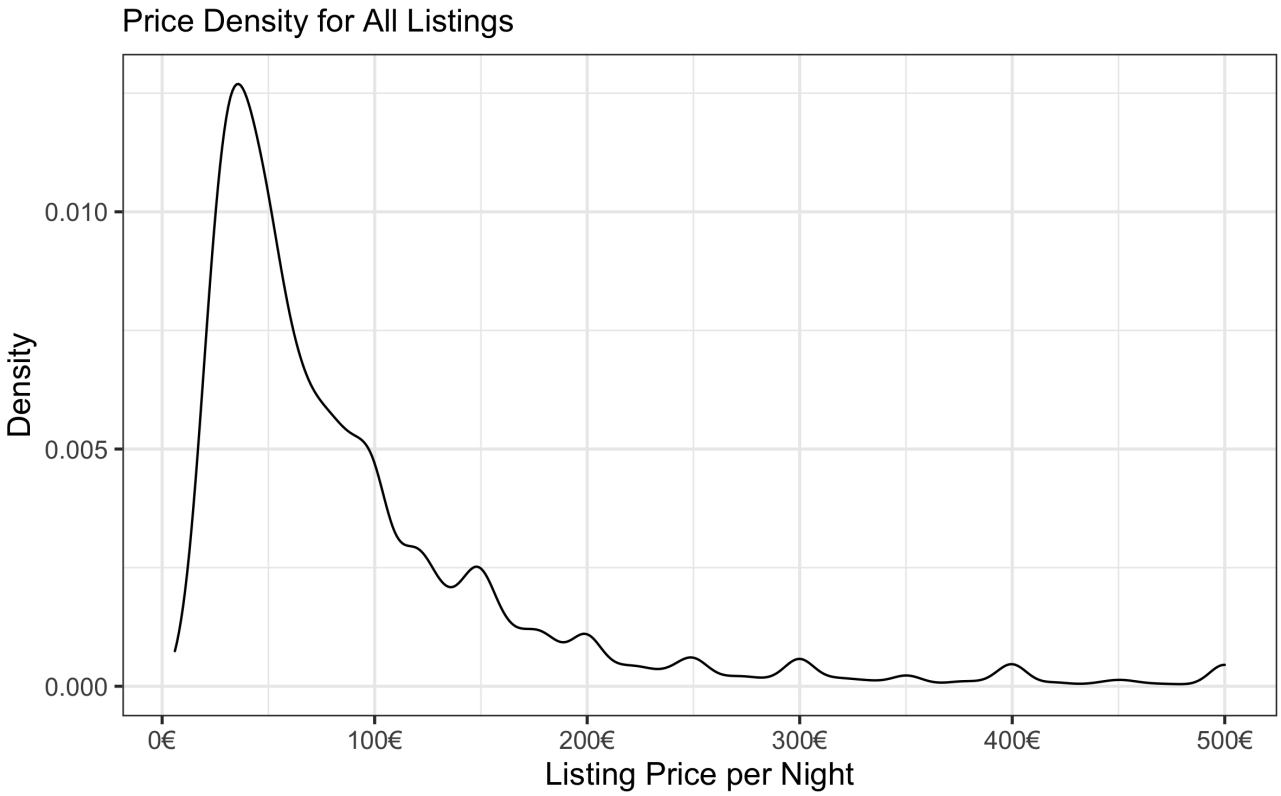

Figure 1: Density plot for prices of Airbnb listings in Barcelona, Spain: Listings have an extreme positive skew and appear to be multimodal. The modality reflects the tendency for listings to be priced in increments of 50 Euros

Figure 2: Correlations for potentially significant variables in the explaining the price of Airbnb listings

Violin Plot

The violin plot below shows the distribution of price in log10 scale for each district in descending order of average price. Based on the plot, Example has the highest priced and Nou Barris has the lowest priced listings.

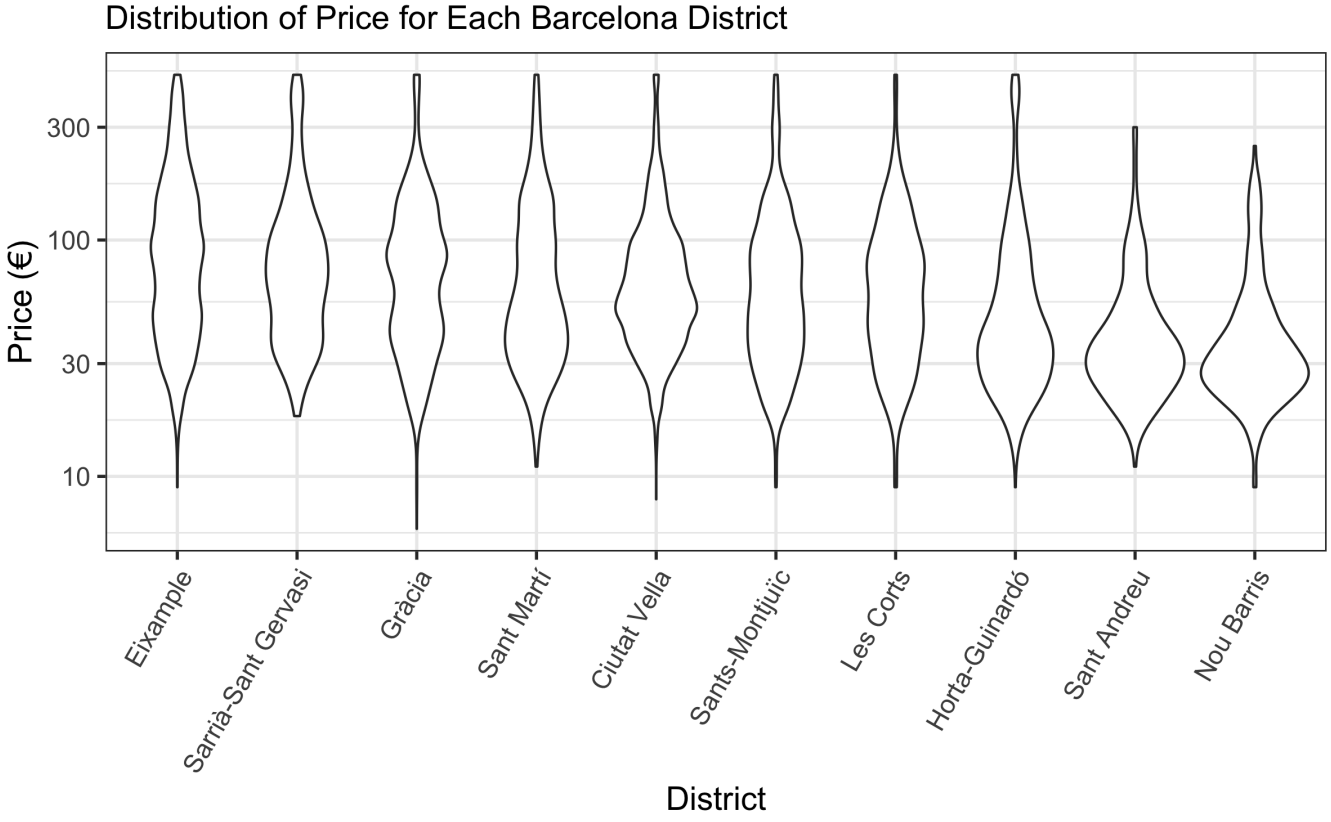

Figure 3: Plot shows the distribution of listing prices for Airbnb listings in Barcelona, Spain by city district

Analysis Methods

We perform linear regression on Airbnb listing prices to the listing's district, type of room, reviews left per month, and distance from city center to see how these variables might be affecting the Airbnb listing prices in Barcelona. First, we run

```
lm(price ~ district + room_type + distance + reviews_per_month, data=df)
```

where our data is housed in the dataframe, `df` and look at the QQ-Plots of the standardized residuals. This plot is given as the left plot below. It is obvious that normality assumptions are not appropriate, so we then run

```
lm(log(price) ~ district + room_type + distance + reviews_per_month, data=df)
```

and look at the QQ-plot for the log transform of price. This QQ-plot is given below on the right and is much closer to the normality assumption even though there is evidence of heavier tails.

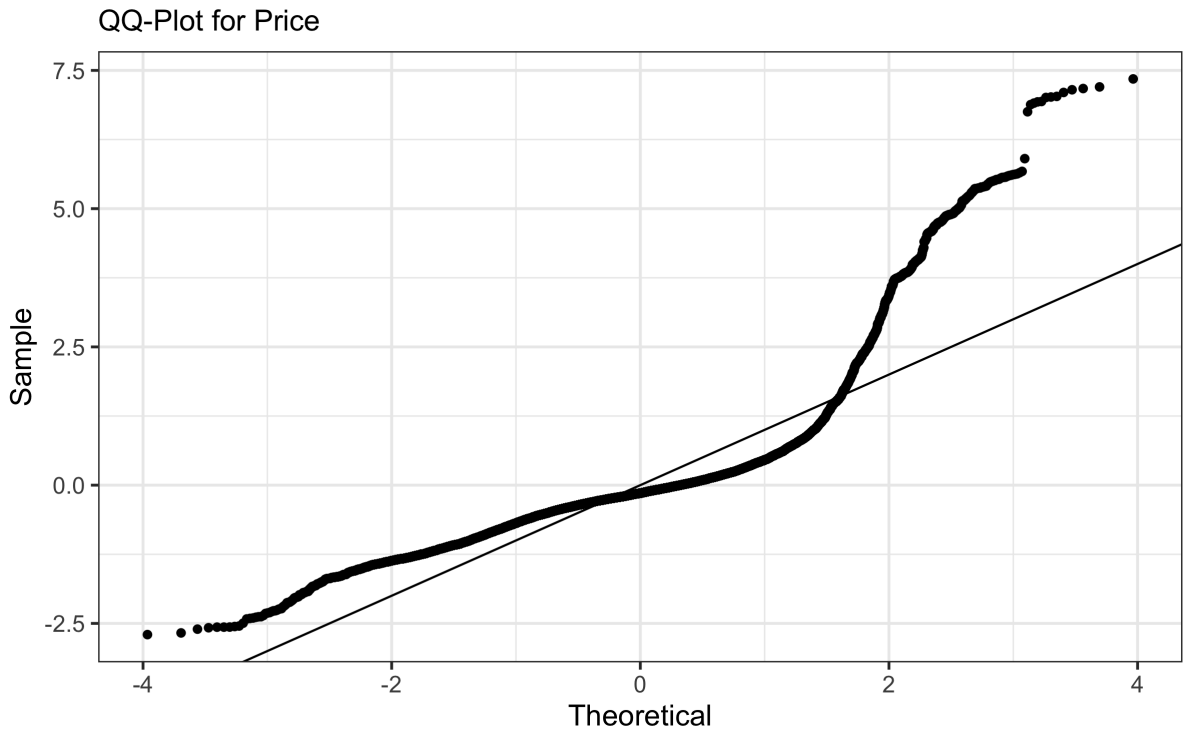

Figure 4: QQ-Plot for standardized residuals of the regression on the raw price data: The residuals show evidence that the errors are not normally distributed

Analysis Results

First, let's look at the results of the linear regression for the untransformed price response variable. In the output below, we see that only 2 of the 10 districts, with Ciutat Vella as the base district, are statistically significant at the 95% confidence level. These are Eixample and Sarrià-Sant Gervasi.

```
#Linear Model on Price
lm.1 <- readRDS(file=here::here("data", "lm1_results.rds"))
tidy(lm.1)
```

```
## # A tibble: 15 x 5
##    term                          estimate std.error statistic  p.value
##    <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                     154.        1.76     87.6   0.
##  2 districtEixample                 13.7       1.84      7.44  1.05e-13
##  3 districtGràcia                   -0.959     2.80     -0.342 7.32e- 1
##  4 districtHorta-Guinardó            4.49      3.92      1.14  2.53e- 1
##  5 districtLes Corts                 0.938     5.09      0.184 8.54e- 1
##  6 districtNou Barris                4.50      6.08      0.741 4.59e- 1
##  7 districtSant Andreu              -3.51      5.11     -0.687 4.92e- 1
##  8 districtSant Martí                4.21      2.49      1.69  9.02e- 2
##  9 districtSants-Montjuïc           -2.76      2.67     -1.03  3.02e- 1
## 10 districtSarrià-Sant Gervasi      18.2       4.20      4.33  1.48e- 5
## 11 room_typeHotel room              30.7       3.58      8.58  1.03e-17
## 12 room_typePrivate room           -89.9       1.20    -74.7   0.
## 13 room_typeShared room            -92.5       6.85    -13.5   3.19e-41
```

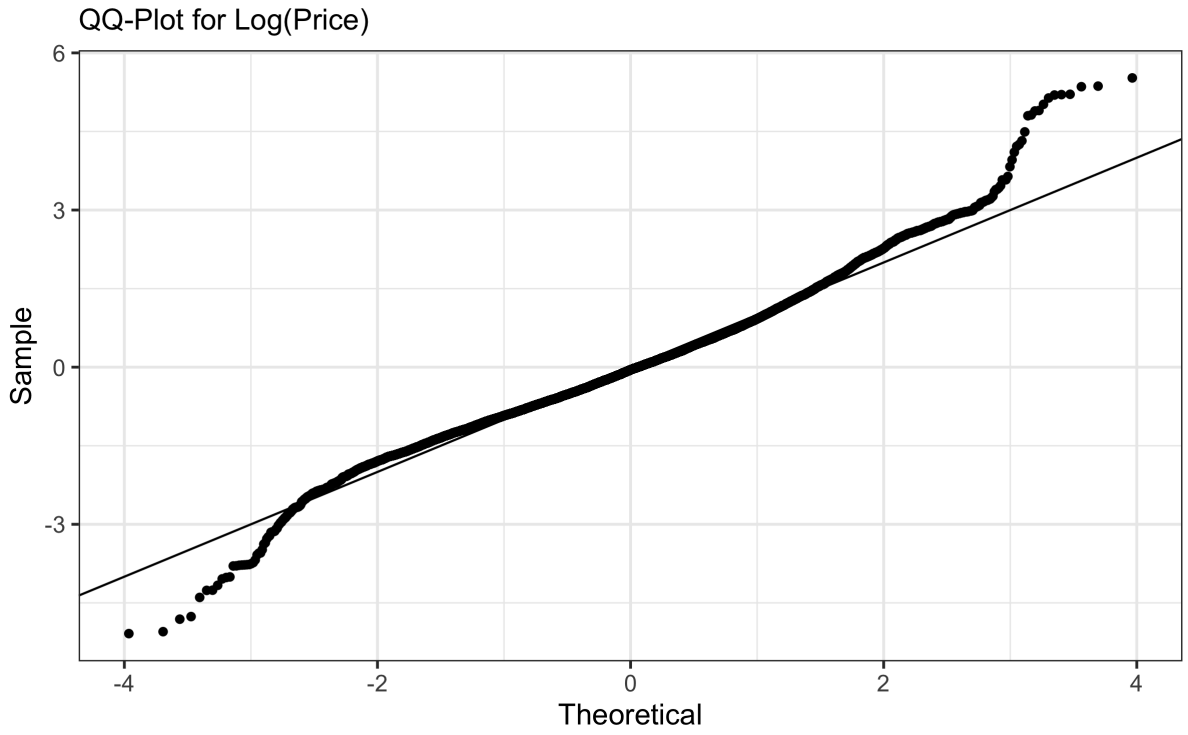

Figure 5: QQ-Plot for standardized residuals of the regression on the log transform of price data: The residuals seem to have heavier tails than the normal distribution, but are better suited for linear regression than the untransformed price data

```
## 14 distance                    -489.       68.2      -7.17  7.80e-13
## 15 reviews_per_month             -3.13      0.320    -9.80  1.40e-22
```

We can interpret the estimate for Eixample, 13.7, as saying that prices of Airbnb listings in the Eixample district differ from listings in the Ciutat Vella district, on average, by 13.7 Euro's per night. All room types, distance, and reviews per month are statistically significant at any reasonable confidence level.

Next, we look at the results of the linear regression on the log(price) of Airbnb listings. In the output below, we see many more districts are now statistically significant, but the interpretation of their coefficient estimates are not as straightforward as the results when regressing on listing price.

```
#Linear Model on log(Price)
lm.2 <- readRDS(file=here::here("data", "lm2_results.rds"))
tidy(lm.2)
```

```
## # A tibble: 15 x 5
##    term                         estimate std.error statistic   p.value
##    <chr>                           <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                    4.98     0.0137    363.    0.
##  2 districtEixample               0.0771   0.0144      5.36  8.49e-  8
##  3 districtGràcia                -0.0280   0.0219     -1.28  2.00e-  1
##  4 districtHorta-Guinardó        -0.107    0.0306     -3.48  4.99e-  4
##  5 districtLes Corts              0.0537   0.0398      1.35  1.77e-  1
##  6 districtNou Barris            -0.0810   0.0475     -1.71  8.79e-  2
##  7 districtSant Andreu           -0.181    0.0399     -4.53  5.90e-  6
##  8 districtSant Martí            -0.0238   0.0194     -1.22  2.21e-  1
##  9 districtSants-Montjuïc        -0.0706   0.0209     -3.38  7.26e-  4
## 10 districtSarrià-Sant Gervasi    0.177    0.0328      5.39  7.11e-  8
## 11 room_typeHotel room           -0.0167   0.0280     -0.598 5.50e-  1
## 12 room_typePrivate room         -1.01     0.00940  -108.    0.
## 13 room_typeShared room          -1.27     0.0536    -23.7   9.59e-122
## 14 distance                      -6.25     0.533     -11.7   1.28e- 31
## 15 reviews_per_month             -0.0214   0.00250    -8.57  1.18e- 17
```

Due to the improved QQ-Plot for the log(price) model, we only consider this model's results going forward. We can look at how well this model fits the data by randomly selecting 10 listings and comparing the price to the fitted price.

```
set.seed(100)
augment(lm.2) %>%
  slice(sample(1:nrow(augment(lm.2)), size=10, replace=FALSE)) %>%
  select(.rownames, district, room_type, distance, reviews_per_month, log.price., .fitted) %>%
  rename(price = log.price.,
         fitted = .fitted) %>%
  mutate(price = exp(price),
         fitted = exp(fitted))
```

```
## # A tibble: 10 x 7
##    .rownames district        room_type    distance reviews_per_mon~ price fitted
##    <chr>     <chr>           <chr>           <dbl>            <dbl> <dbl>  <dbl>
##  1 4036      Ciutat Vella    Entire home~   0.0123            0.02 135.   134.
##  2 514       Eixample        Entire home~   0.0135            0.56 149    142.
##  3 3623      Sarrià-Sant Ge~ Private room   0.0404            3.27  84.0   45.5
##  4 3935      Ciutat Vella    Private room   0.0110            0.13  40     49.0
##  5 4375      Sant Andreu     Entire home~   0.0331            1.16  60.0   95.9
##  6 8530      Eixample        Private room   0.0297            0.27  50.0   46.9
##  7 3190      Eixample        Entire home~   0.0226            0.12  85.   136.
```

```
##  8 12181     Sant Martí      Entire home~   0.0284            4.02 180    109.
##  9 8849      Eixample        Private room   0.0173            3.33  55.    47.5
## 10 9365      Sants-Montjuïc  Entire home~   0.0218            0.47 120.   117.
```

In the output above, the `price` columns shows the listing's price from the Airbnb website and `fitted` shows the price given by the linear model. Note that we exponentiated this output so that it is in Euro's and straightforward to compare.

Finally, we can look at the adjusted r-squared to see how much of the variance in listing price is explained by district, type of room, distance of city center, and reviews per month. By using the `glance()` function from the `broom` package, we see that 52.36% of the variance of the log transform of listing is explained by our model.

Discussion

Based on the results of our model using the log transform of listing price, district, type of room, distance from city center, and reviews per month all play a significant role in the price of an Airbnb listing. However, much of the variance is left unexplained in our model so that we may want to look for additional variables.

References

Airbnb dataset