AirBnB

Linear regression

Saahil Shroff

2022-11-26

Predicting AirBnB Price

Loading the data and basic sanity checks.

```
## There are 4448 rows and 74 columns.
```

```
## Any NULLS present? TRUE
```

```
## Total number of NULLS: 19868
```

Data cleaning

```
## Selected columns are: name price bedrooms beds room_type accommodates
```

```
## Dimension of the dataframe with selected columns: 4448 6
```

```
##                                                name price bedrooms beds
## 1                       D1 -  Million Dollar View 2 BR   158        2    2
## 2                    Garden level studio in ideal loc.   150       NA    2
## 3 Monthly (or Longer ) Designer One Bedroom Downtown    85        1    1
## 4                          Vancouver's best kept secret   149        1    1
## 5                                    EcoLoft Vancouver   150        1    2
## 6      Close to PNE/Hastings Park Garden level suite   350        2    3
##         room_type accommodates
## 1 Entire home/apt            5
## 2 Entire home/apt            4
## 3 Entire home/apt            2
## 4 Entire home/apt            2
## 5 Entire home/apt            4
## 6 Entire home/apt            4
```

NULL checks

```
## Number of NULLs in price: 0
```

```
## Number of NULLs in bedrooms: 312
```

```
## Number of NULLs in beds: 150
```

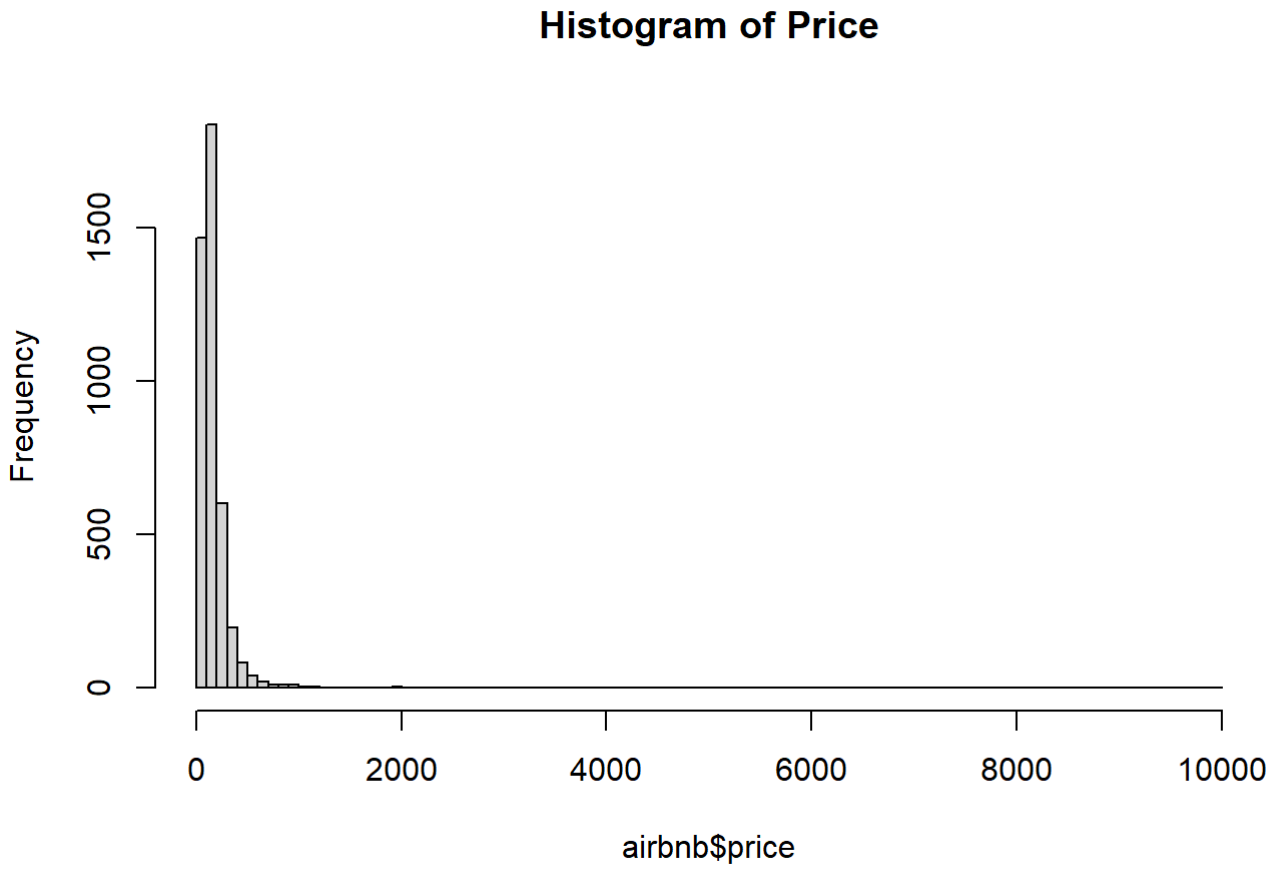
Dropping NULLs from beds-variable & cleaning the bedrooms variable:

```
## Dimension after dropping NULLS from bed-variable: 4298 6
```

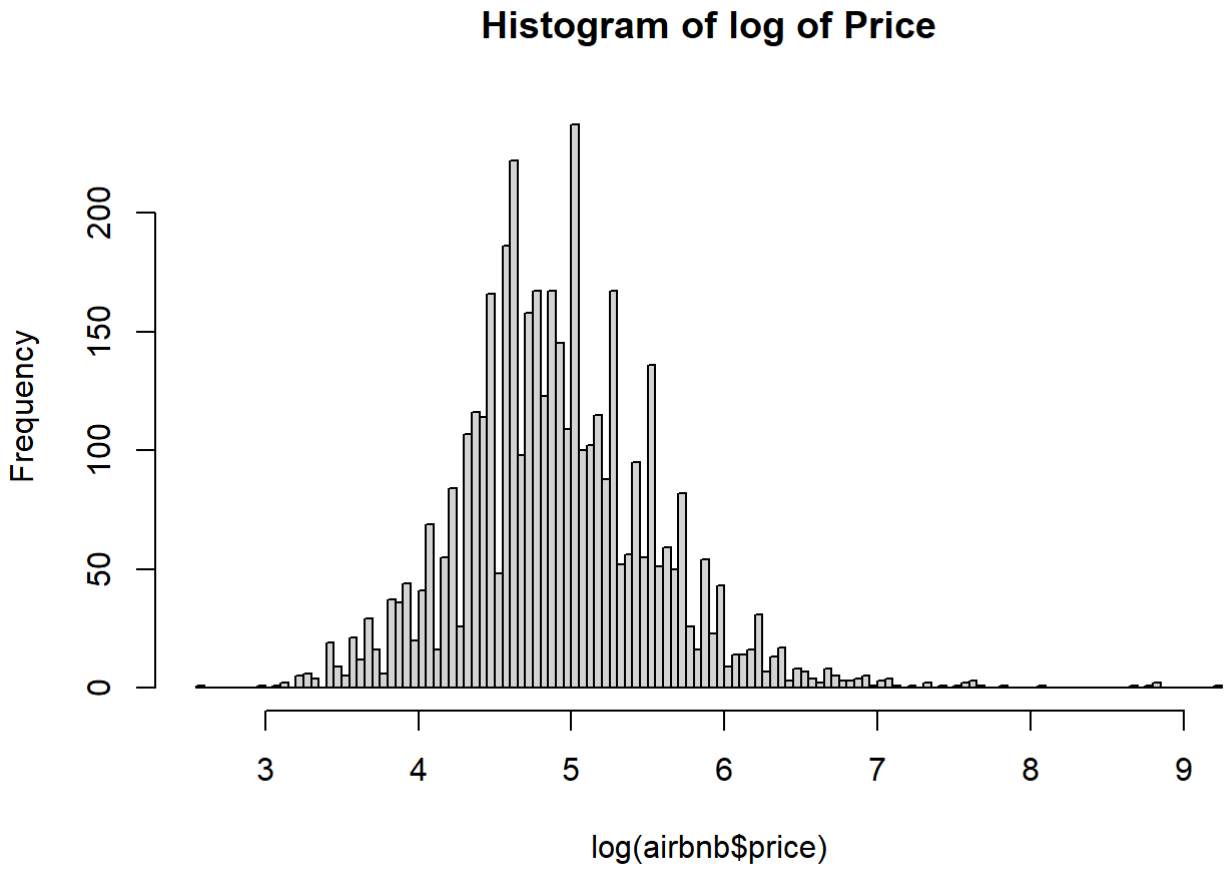
```
##                                                 name price bedrooms beds
## 1                       D1 -  Million Dollar View 2 BR   158        2    2
## 2                    Garden level studio in ideal loc.   150        2    2
## 3 Monthly (or Longer ) Designer One Bedroom Downtown     85        1    1
## 4                          Vancouver's best kept secret   149        1    1
## 5                                    EcoLoft Vancouver   150        1    2
## 6      Close to PNE/Hastings Park Garden level suite   350        2    3
##          room_type accommodates
## 1 Entire home/apt            5
## 2 Entire home/apt            4
## 3 Entire home/apt            2
## 4 Entire home/apt            2
## 5 Entire home/apt            4
## 6 Entire home/apt            4
```

Here, I have substituted the value of beds into the bedrooms-variable wherever it was NULL. I did so because when I went through the dataset before making this modification, it seemed like the beds and bedrooms were usually the same in most cases.

Analyzing the distribution of price.



The above plot looks very similar to the Pareto distribution.



Taking the log of the first curve, we can see that the plot resembles a normal-distribution.

Converting the number of bedrooms into another variable with a limited number of categories only, such as 0, 1, 2, 3+ to use these categories in the models below

Adding a new variable - bedrooms_sel:

```
##                                                name price bedrooms beds
## 1                      D1 -  Million Dollar View 2 BR   158        2    2
## 2                    Garden level studio in ideal loc.   150        2    2
## 3 Monthly (or Longer ) Designer One Bedroom Downtown    85        1    1
## 4                          Vancouver's best kept secret   149        1    1
## 5                                    EcoLoft Vancouver   150        1    2
## 6      Close to PNE/Hastings Park Garden level suite   350        2    3
##         room_type accommodates bedrooms_sel
## 1 Entire home/apt            5            1
## 2 Entire home/apt            4            1
## 3 Entire home/apt            2            0
## 4 Entire home/apt            2            0
## 5 Entire home/apt            4            0
## 6 Entire home/apt            4            1
```

Estimating a linear regression model where you explain log price

with number of BR-s (the BR categories done above).

Modelling regression model of price & BRs

```
##
## Call:
## lm(formula = price ~ bedrooms_sel, data = airbnb_final)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -519.1  -62.5  -27.5   22.5 9871.5
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     127.523      5.601  22.769  < 2e-16 ***
## bedrooms_sel1    64.698      9.786   6.611 4.27e-11 ***
## bedrooms_sel2   163.579     14.722  11.111  < 2e-16 ***
## bedrooms_sel3+  416.584     21.732  19.169  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 279.4 on 4294 degrees of freedom
## Multiple R-squared:  0.09625,    Adjusted R-squared:  0.09562
## F-statistic: 152.4 on 3 and 4294 DF,  p-value: < 2.2e-16
```

Modelling regression model of log(price) & BRs

```
##
## Call:
## lm(formula = log(price) ~ bedrooms_sel, data = airbnb_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6739 -0.3228 -0.0285  0.3369  4.5567
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.65352    0.01095  425.04   <2e-16 ***
## bedrooms_sel1   0.48108    0.01913   25.15   <2e-16 ***
## bedrooms_sel2   0.78720    0.02878   27.35   <2e-16 ***
## bedrooms_sel3+  1.23922    0.04248   29.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5461 on 4294 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.286
## F-statistic: 574.6 on 3 and 4294 DF,  p-value: < 2.2e-16
```

Looking at both the models, we can see that the model on log(price) has higher R-squared value than the model that just uses price. This clearly shows that the log(price) vs bedrooms are more interdependent as compared to price vs bedrooms.

Values that these two variables (room type and accommodates)

take

Room type:

```
##
## Entire home/apt      Hotel room    Private room     Shared room
##            3484               4             802               8
```

Accomodates:

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
##  260 1615  449 1091  246  406   61  100    8   27    5   14    3    7    1    5
```

Converting the room type into 3 categories: Entire home/apt, Private room, Other; and recode accommodates into 3 categories: “1”, “2”, “3 or more”.

```
## types_of_room
## Entire home/apt           Other    Private room
##            3484              12             802
```

```
##                                             name price bedrooms beds
## 1                     D1 -  Million Dollar View 2 BR   158        2    2
## 2                   Garden level studio in ideal loc.   150        2    2
## 3 Monthly (or Longer ) Designer One Bedroom Downtown    85        1    1
## 4                        Vancouver's best kept secret   149        1    1
## 5                                    EcoLoft Vancouver   150        1    2
## 6      Close to PNE/Hastings Park Garden level suite   350        2    3
##         room_type accommodates bedrooms_sel   types_of_room new_accomodates
## 1 Entire home/apt            5            1 Entire home/apt            more
## 2 Entire home/apt            4            1 Entire home/apt            more
## 3 Entire home/apt            2            0 Entire home/apt               2
## 4 Entire home/apt            2            0 Entire home/apt               2
## 5 Entire home/apt            4            0 Entire home/apt            more
## 6 Entire home/apt            4            1 Entire home/apt            more
```

Adding new variables (new_accomodates + types_of_room) to

the previous prediction model and interpreting the model.

```
##
## Call:
## lm(formula = log(price) ~ bedrooms_sel + new_accomodates + types_of_room,
##     data = airbnb_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7219 -0.3097 -0.0370  0.2714  4.8558
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.44538    0.03730 119.175  < 2e-16 ***
## bedrooms_sel1               0.16338    0.02230   7.328 2.78e-13 ***
## bedrooms_sel2               0.43896    0.03051  14.388  < 2e-16 ***
## bedrooms_sel3+              0.90329    0.04135  21.844  < 2e-16 ***
## new_accomodates2            0.28314    0.03597   7.871 4.41e-15 ***
## new_accomodates3            0.35809    0.04332   8.265  < 2e-16 ***
## new_accomodatesmore         0.59213    0.04071  14.544  < 2e-16 ***
## types_of_roomOther         -0.10393    0.14478  -0.718    0.473
## types_of_roomPrivate room -0.37407    0.02335 -16.017  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4958 on 4289 degrees of freedom
## Multiple R-squared:  0.4125, Adjusted R-squared:  0.4114
## F-statistic: 376.4 on 8 and 4289 DF,  p-value: < 2.2e-16
```

For each category, a reference is set and other categorical data is computed with respect to it. For e.g.: In variable “bedrooms_sel” - bedrooms_sel1, bedrooms_sel2, and bedrooms_sel3+ are calculated in reference to bedrooms_sel0. Additionally, this model has better R-sqaured value as compared to the previous model. In my view, it is due to the extra variables that we have taken to predict the price; these extra factors influence the price of the airbnb as compared to just # of bedrooms, and aids us in predicting the price of the airbnb unit.

types_of_roomOther is not statiscally significant, i.e., this categorical data has no influence in predicting the price of the airbnb. It may be because the data present in this category is too small as compared to other categories in the same variable to make an effect on the price of the airbnb unit.

Using the model above to predict (log) price for each listing in the data

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.071   4.729   4.884   4.917   5.201   5.941
```

Root-mean-squared-error (RMSE) of the predictions.

```
## [1] 0.495303
```

Using the model to predict log price for a 2-bedroom apartment that accommodates 4 (i.e., a full 2BR apartment).

```
## Log price of a 2-Bedroom apartment that accomodates 4: 5.476477
```