Airbnb Price Prediction Analysis

Problem Solving

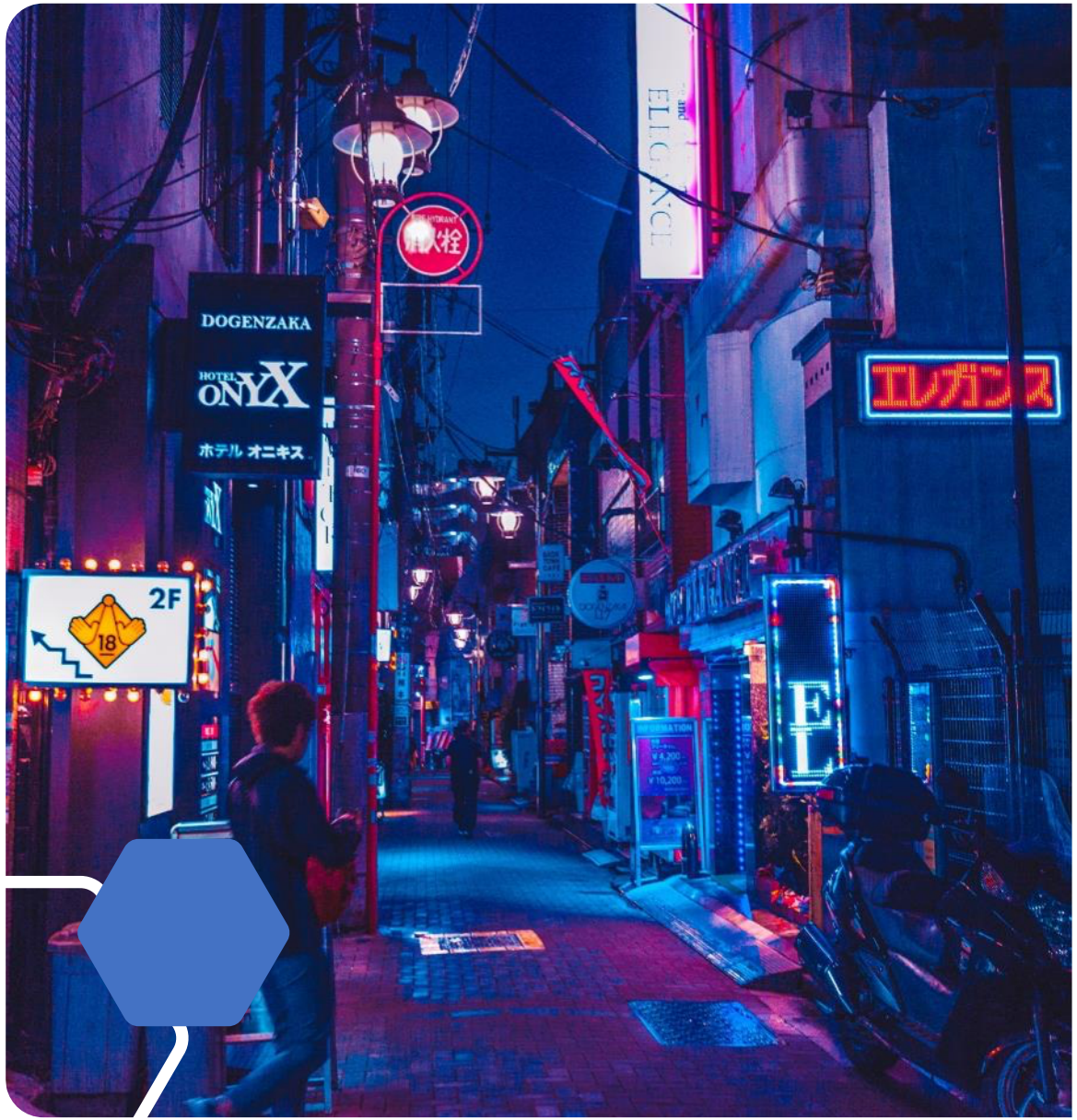
- Problems are at the center of everything you do
- Problem solving is the skill of solving complex problems
- It involves creativity, decision making and evaluating alternatives

Big Data

- Large amounts of data available for analysis to reveal patterns, trends and associations
- All kinds of data consumer, healthcare, industrial, commercial

Problem Solving Using Big Data

- Data can solve many problems
- Take for instance the issue of pricing for Airbnb
- Price automation can solve two major problems for Airbnb:
  1. Help hosts to decide an optimal price for their property based on recommendations
  2. Increase revenue generation for Airbnb by eliminating the price differences that arise from under-pricing of properties
- The issue of price automation can be solved with a very large dataset on past Airbnb property listings



Focus:

Analyse the Airbnb listings dataset

Identify significant predictors

Build a model for future price automation

Factor Analysis

Features inside data:

- Property Type, Room Type, Bedrooms, Accommodation, Bathrooms, City, Amenities, Reviews, etc.

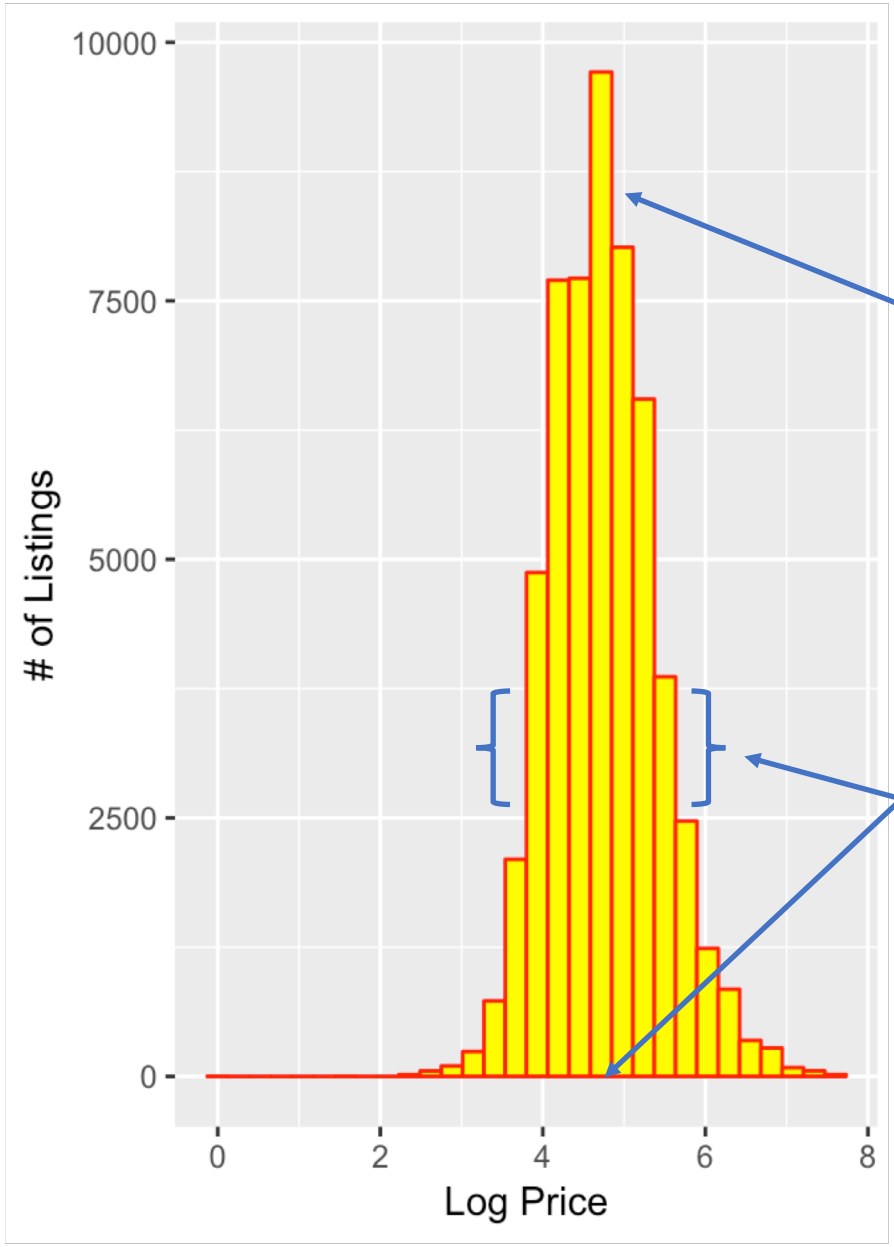
What are we predicting?

- The answer is simple, it’s PRICE!

What features does the price depend on?

- That’s what we will figure out!

Histogram for Price Distribution



What does this represent?

-Price with the most listings

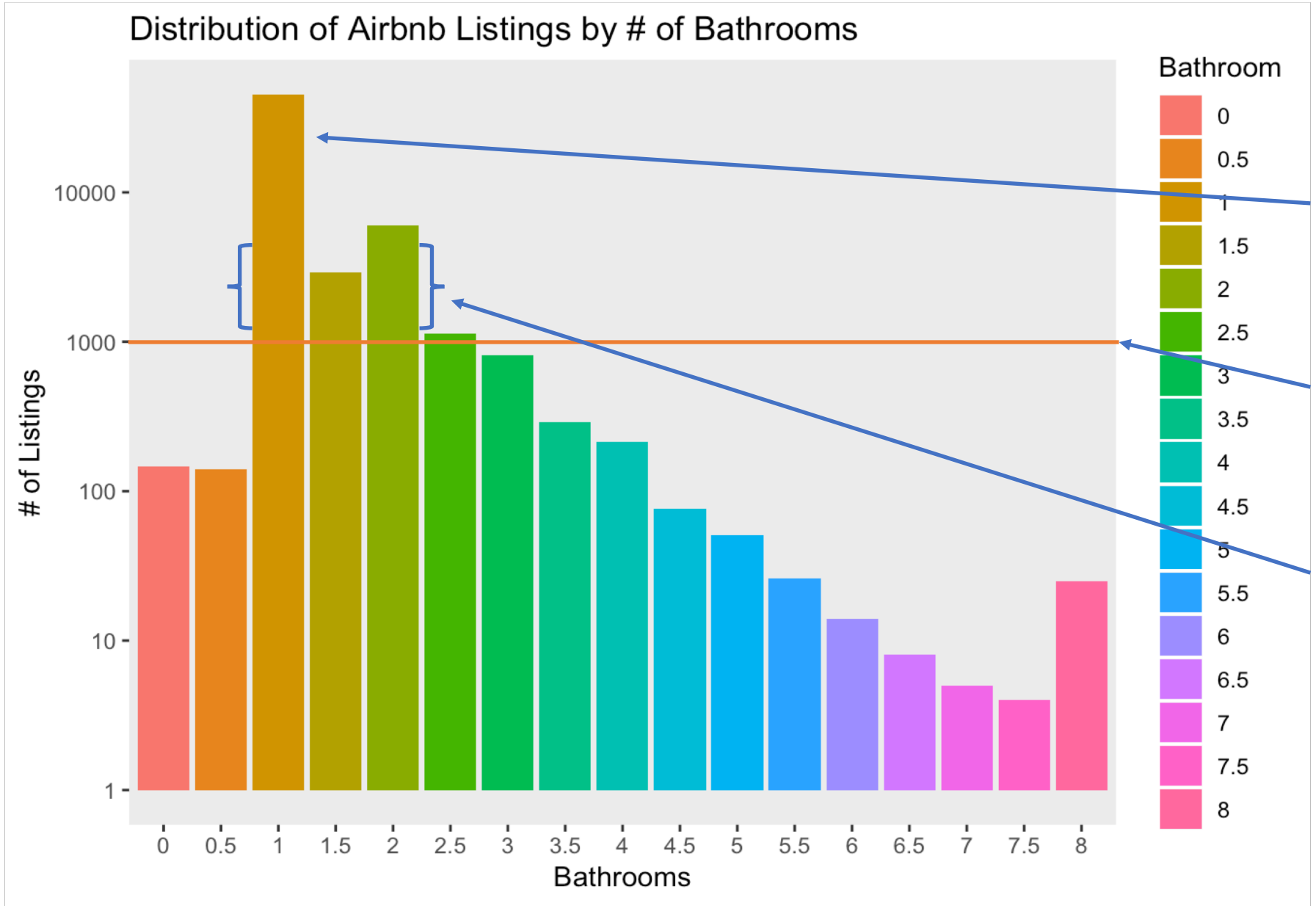
-Bell shaped curve means the mean price is approximately in the middle

-Prices for majority listings are between 4 and 6

Actual Statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	4.304	4.700	4.749	5.165	7.600

Bathrooms



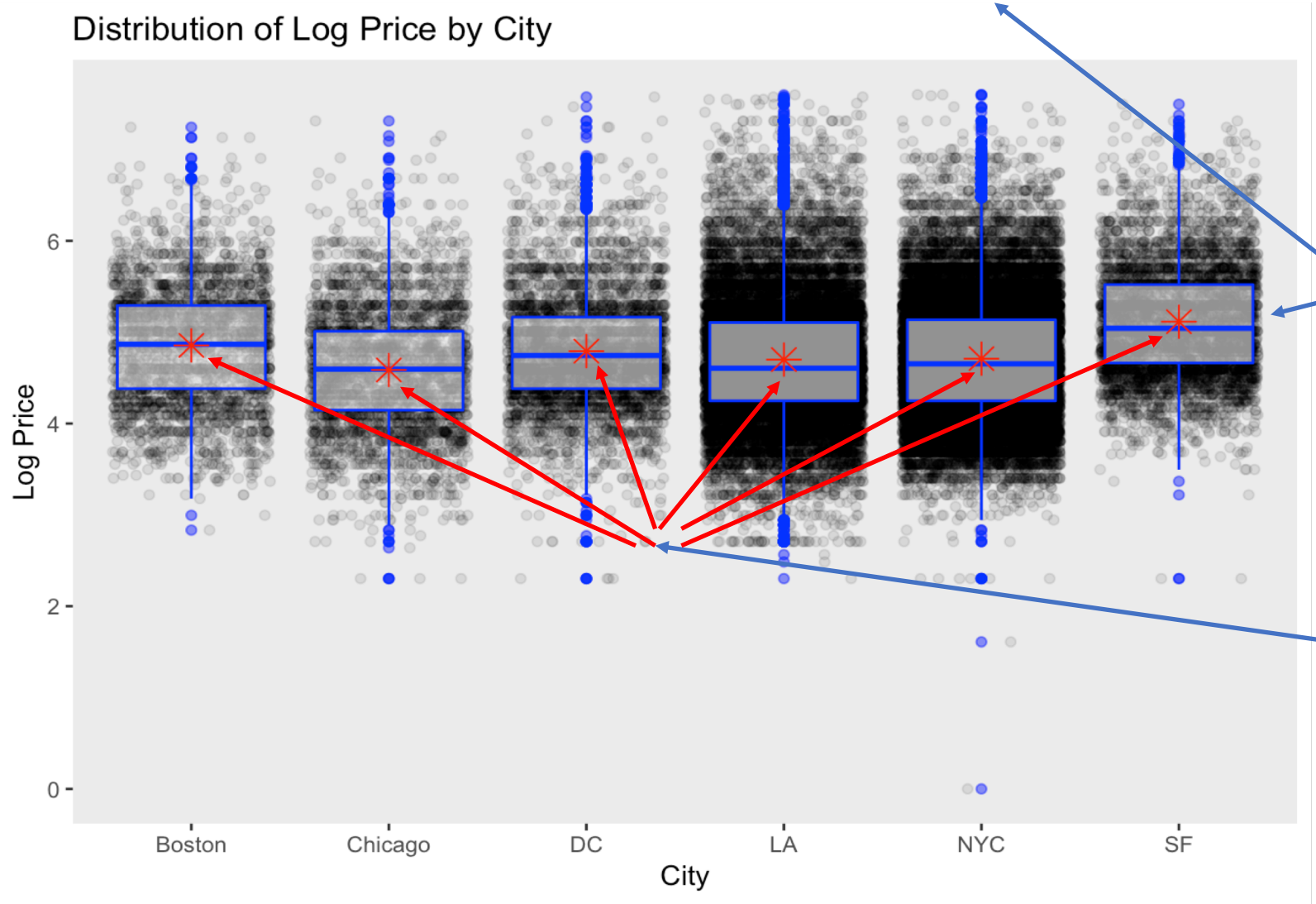
What does this represent?

-Properties with a single bathroom have more than 10,000 listings

-Bathrooms below the line have less than a 1000 listings

-Majority listings have between 1 and 2 bathrooms

Cities and Price



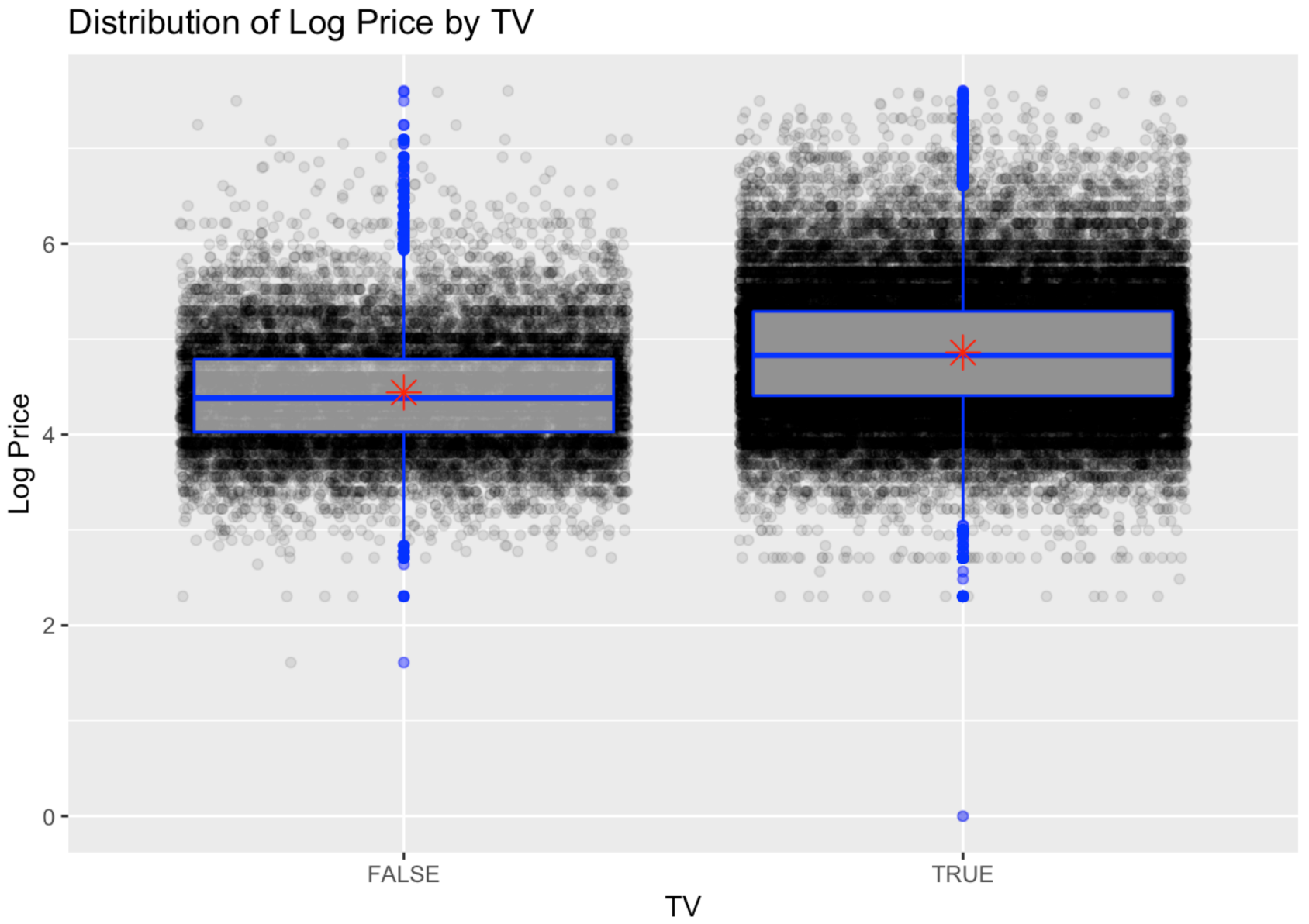
What does this represent?

-Which city has the highest mean price?

-Which cities generally attract the most number of guests

-Do different cities have different prices on average?

Does the Price Depend on TV?



What does this represent?

-Does the property have a TV or not?

-Mean price for listings with and without a TV

-In general, listings with a TV cost more to rent

Linear Regression for Price against Bed Types

```
##
## Call:
## lm(formula = Log_Price ~ Bed_Type, data = new_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7614 -0.4439 -0.0454  0.4148  3.0503
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.22900    0.03600 117.462  < 2e-16 ***
## Bed_TypeCouch         -0.07111    0.06392  -1.112    0.266
## Bed_TypeFuton          0.05259    0.04508   1.167    0.243
## Bed_TypePull-out Sofa  0.20876    0.04684   4.457 8.33e-06 ***
## Bed_TypeReal Bed       0.53235    0.03611  14.741  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6639 on 57001 degrees of freedom
## Multiple R-squared:  0.01312,    Adjusted R-squared:  0.01305
## F-statistic: 189.5 on 4 and 57001 DF,  p-value: < 2.2e-16
```

What does this represent?

-Which bed types have a significantly different price from the intercept (or the first bed type on the x-axis)

-Which bed types affect the price significantly

-Is the way in which each of the bed types influence the price different

Look for difference in coefficient estimates

T-test for if the Listing can be Instantly Booked against Price

```
##
##  Two Sample t-test
##
## data:  Log_Price by Instant_Bookable
## t = 12.011, df = 57004, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06338561 0.08810650
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            4.769023            4.693277
```

Very small p-value → significant

What does this represent?

-Mean price for a listing that can not be booked instantly

-Mean price for a listing that can be booked instantly

-Is there a price difference between the two kinds of listings

Factor Analysis

	Log_Price <dbl>	Property_Type <fctr>	Room_Type <fctr>	City <fctr>	Bed_Type <fctr>	Cancellation_I <fctr>
1	5.010635	Apartment	Entire home/apt	NYC	Real Bed	strict
2	5.129899	Apartment	Entire home/apt	NYC	Real Bed	strict
3	4.976734	Apartment	Entire home/apt	NYC	Real Bed	moderate
4	4.744932	Apartment	Entire home/apt	DC	Real Bed	moderate
5	4.442651	Apartment	Private room	SF	Real Bed	strict
6	4.418841	Apartment	Entire home/apt	LA	Real Bed	moderate

- After analysing all these different features separately and in connection to the price, I found out that some of these features barely have any connection to the price.

- Since price is what we are trying to predict, I will separate these features from the ones that are responsible for the price

Significant Features

## 1	Property Type
## 2	Room Type
## 3	Accommodates
## 4	Bathrooms
## 5	City
## 6	Bed Type
## 7	Cancellation Policy
## 8	Cleaning Fee
## 9	Review Scores Rating
## 10	Beds
## 11	TV
## 12	Internet
## 13	Kitchen

Nonsignificant Features

## 1	Number of Reviews
## 2	Instant Bookable
## 3	Profile Picture
## 4	Bedrooms
## 5	Parking

Model for Price Automation

Model Creation

Now that we know what features are responsible for the price:

1) Use these features to create a mathematical model
2) Divide the data into two parts: Training and Testing
3) Train the model using training data
4) Apply the trained model on the testing data
5) Predict prices on the testing data
6) Measure model performance

Example - R Code for training the Airbnb Model

```
```{r message=FALSE, warning=FALSE}
library(caret)
# ensure results are repeatable
set.seed(123)
inTraining <- createDataPartition(final_data$Log_Price, p = .8, list = FALSE)
training <- final_data[ inTraining,]
testing  <- final_data[-inTraining,]
# Manual Grid Search
control <- trainControl(method="repeatedcv",
                        number=5,
                        repeats=1,
                        search="grid")
tunegrid <- expand.grid(.mtry=8)
model <- train(Log_Price ~.,
               data=training,
               method="rf",
               metric="RMSE",
               tuneGrid=tunegrid,
               trControl=control,
               ntree=200)
```
```

Airbnb can use this model for price automation of future listings. Thus, solving the problem for hosts and the company itself.

Thank You

Aarij Khawaja

aarijak@gmail.com

https://www.linkedin.com/in/aarij-khawaja/