HongKongbnb

From Dim Sum to Dream Stays

Wan Chai Neighborhood

Team :
Asmae, Estela, Trista,
Salman, Vedant

Agenda

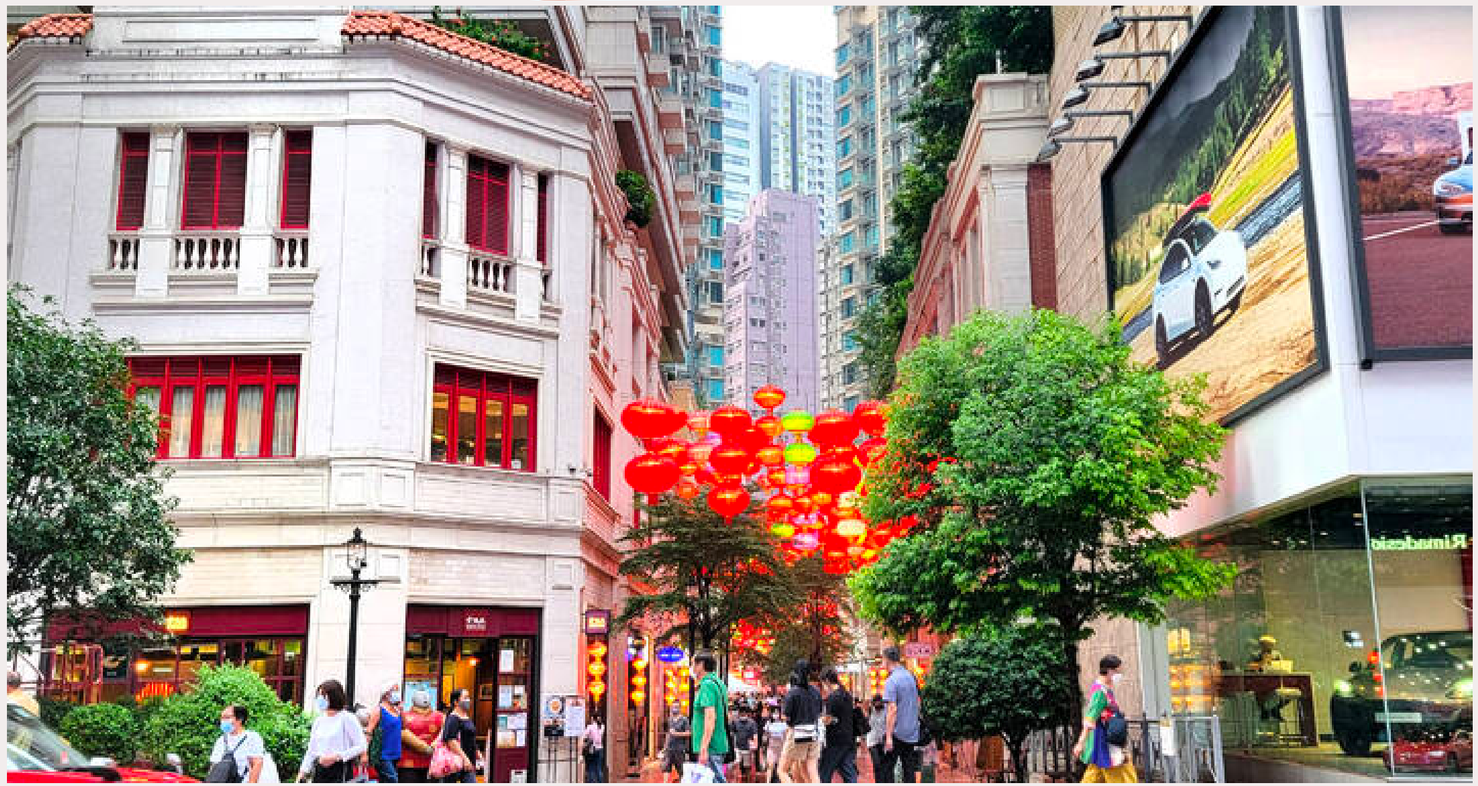
Data Preparation & Exploration

Prediction Model

Classification Model

Clustering Model

Conclusion



STEP 1

Data Preparation
& Exploration

Choosing Relevant Variables

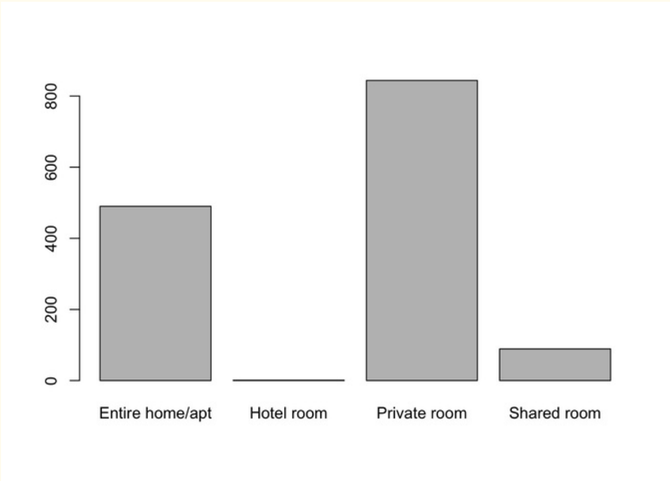
Criteria

- Interpretability
- Amount of information contained
- Relation to other variables
- Adapted to future models
- Amount of missing values

Dealing with Missing Values

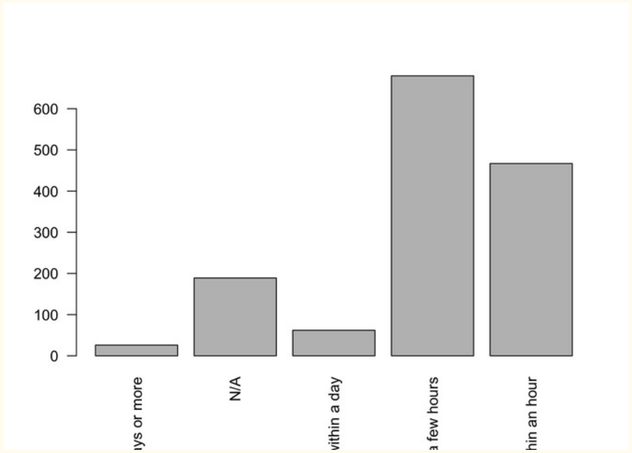
Room Type

Low amount of NAs
Replaced by Private room



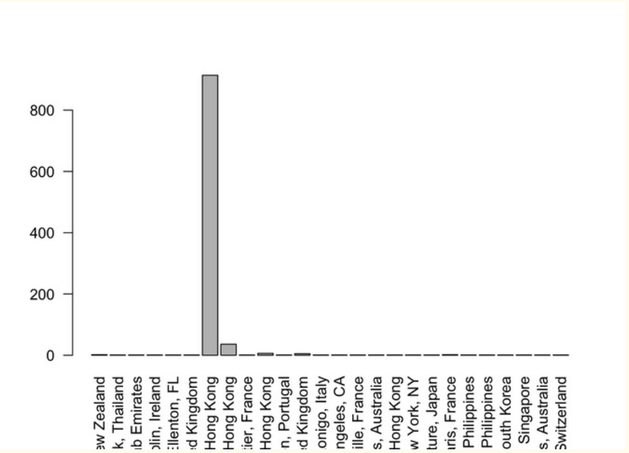
Response rate

Replaced by "within a few hours"

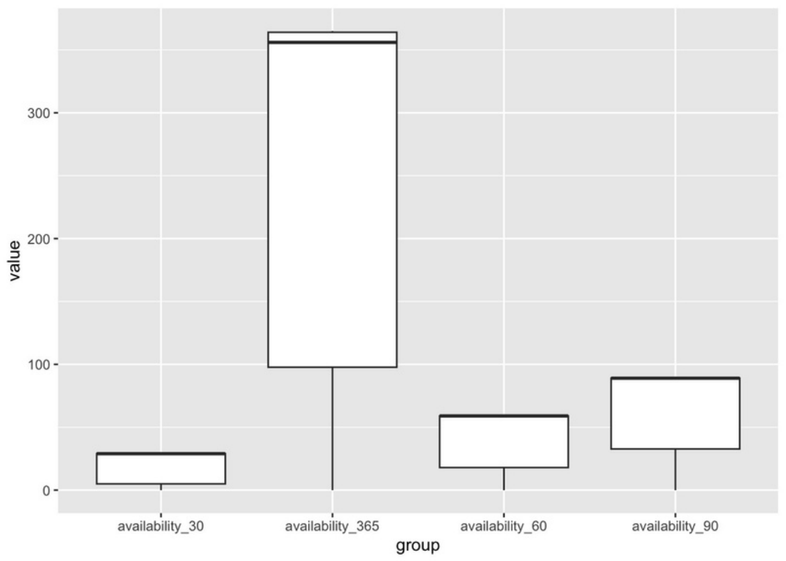


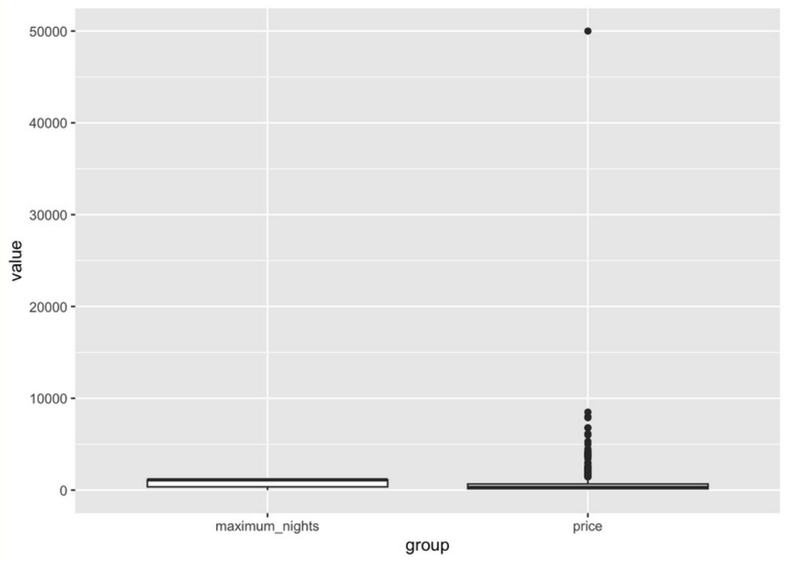
Host location

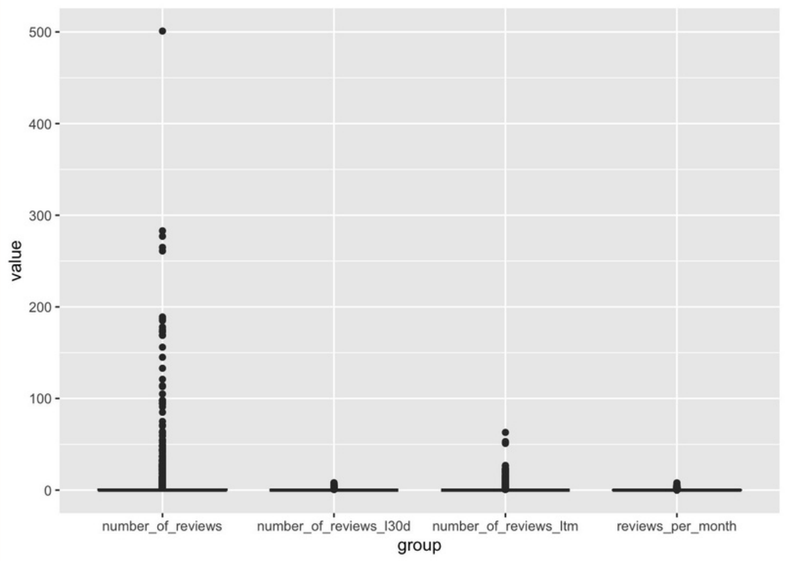
Replaced with Hong Kong



Identifying Outliers







Availability

Price and maximum nights

Reviews

Summary Statistics

Variable analyzed: host_total_listings_count

Five Summaries

```
# five summaries of host_total_listings_count
df.summary <- summary(df$host_total_listings_count)
df.summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    11.0   182.0   235.4   396.0   846.0
```

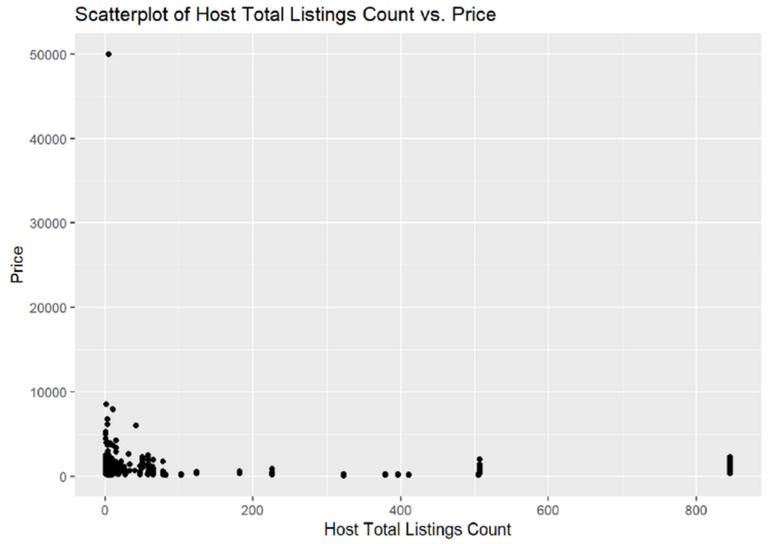
- a wide-ranging distribution of listings: diversity of accommodations offering: varying levels of investment and commitment
- slightly right-skewed distribution indicated by the gap between mean and median value: a few hosts present a notably higher number of listings

Grouped Summary - per Room Type

```
## # A tibble: 4 × 6
##   room_type       mean_count median_count min_count max_count sd_count
##   <chr>                <dbl>        <dbl>     <dbl>     <dbl>    <dbl>
## 1 Entire home/apt       159.         13.5         1       846     276.
## 2 Hotel room             65           65          65        65      NA
## 3 Private room          289.         379           1       846     192.
## 4 Shared room           153.          65           1       505     172.
```

- “Hotel room”: only one dataset - exclusive offer
- “Private room”: highest mean value, point toward a diverse array of private room selections
- “Entire home/apt”: the highest standard deviation & huge gap between mean and median value - notable degree of variability, a significant proportion of hosts within this category have low number of listings

Relationship with Price

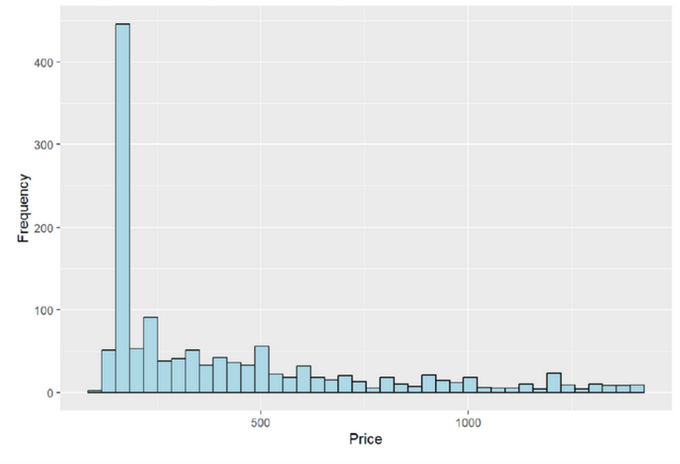


- no clear linear relationship between the listing count and price
- Prices for single listings vary widely, while those with more than 100 listings tend to have less price variation
- hosts to differentiate their offerings based on investment levels
- both basic needs and more extravagant demand can be met

Visualization

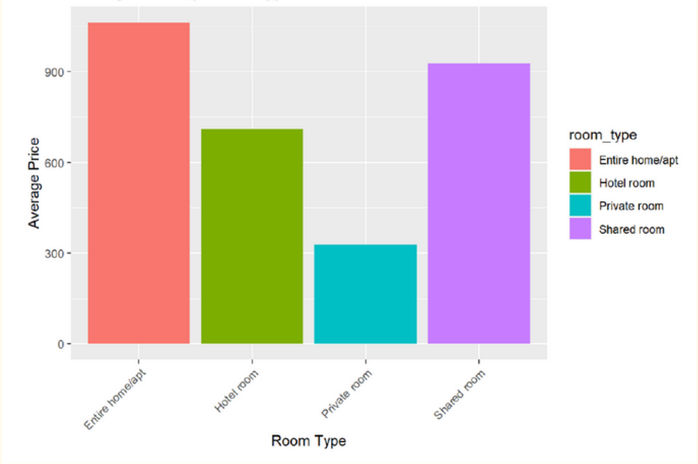
Variable analyzed: Price

Histogram of Price

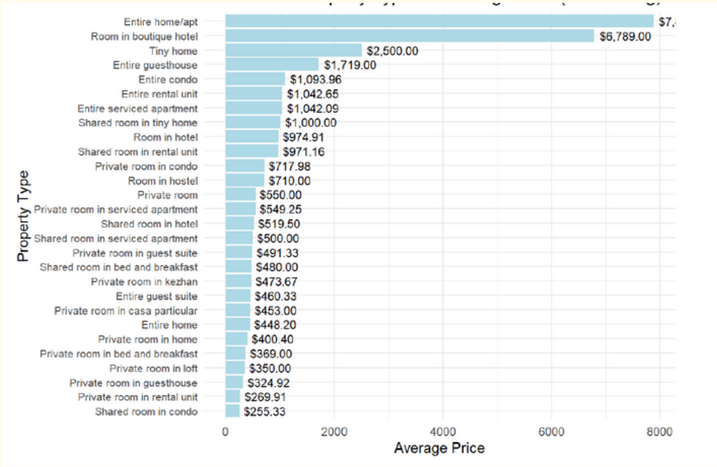


- significant concentration of prices lies below 250, a common price range that attracts a significant portion of potential guests

Average Price by Room Type



Average Price per Property Type

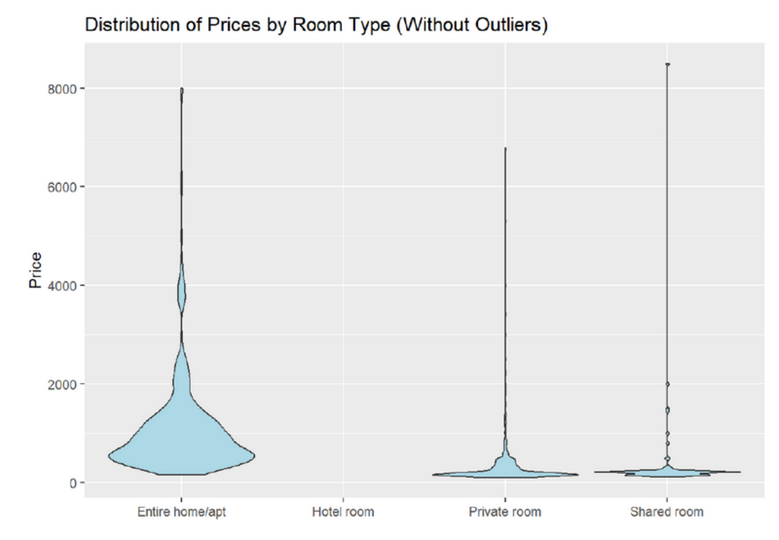


- The detailed categorization of property types indicates the diversity in property types, and the different level of average price highlights the range of options available to travelers, from opulent entire homes to more budget-friendly tiny homes
- The considerable difference between the top two and the subsequent “Tiny home” highlights a substantial pricing variation offer a premium experience and are positioned as a luxurious option.

Visualization

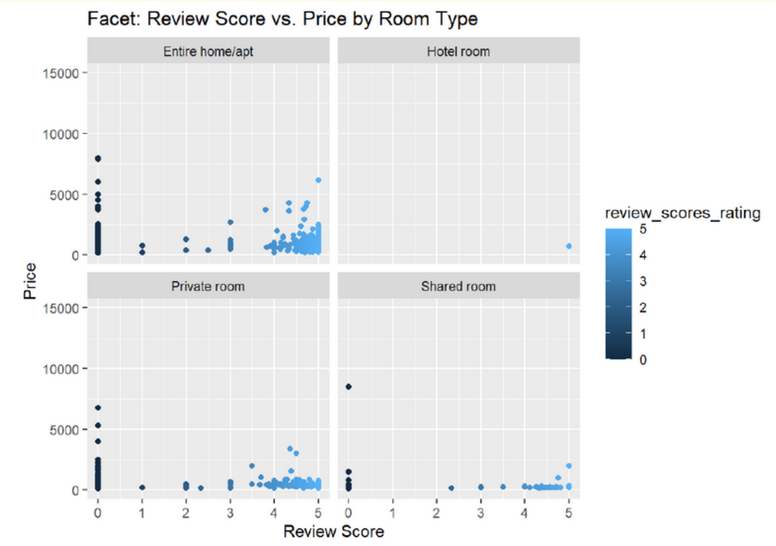
Variable analyzed: Price

Distribution of Prices by Room Type



- "Entire home/apt": both the highest average price and the widest price range. Could be attributed to the uniqueness and luxury accomodates associated
- Private and shared rooms: a similar shape of violin distribution, with shared rooms displaying a more centralized range alongside more extreme outliers

Facet: Review Score vs. Price by Room Type



- review scores do not exhibit a strong correlation with high prices
- suggests that factors beyond pricing significantly influence guest experiences (such as cleanliness, communication, and overall experience)

Mapping

Per Room type

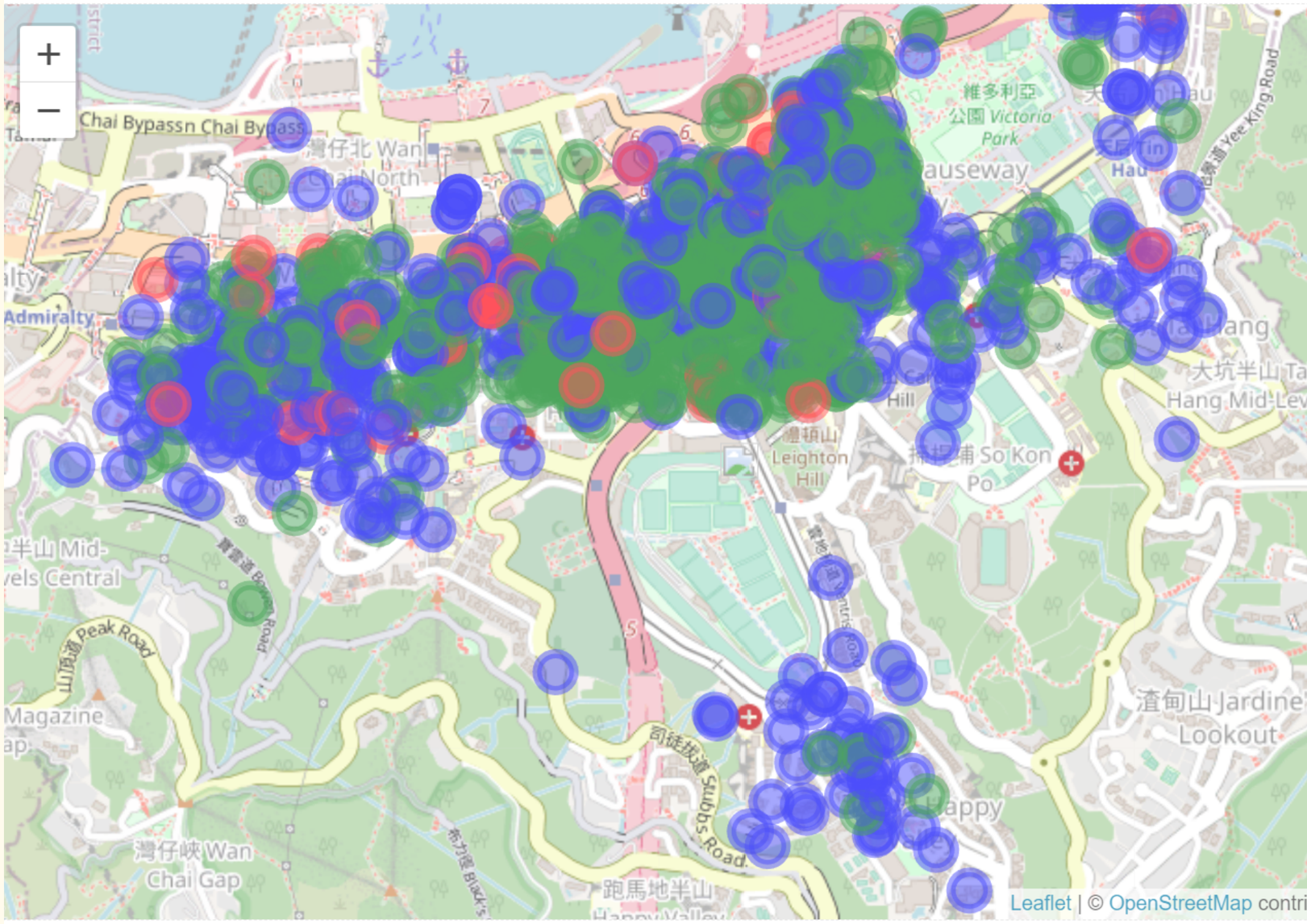
Wan Chai: northern shore of Hong Kong Island

- commercial, residential, and entertainment zones
- accommodations predominantly clustered in specific areas

- **Private rooms**: the northern expanse of Wan Chai & near the HK Cricket Club
  - convenient access to both the northern and southern parts of the district, including local attractions and activities

- **Entire house/apartment**: positioned more towards the outskirts of Wan Chai

- **Shared room:** disperse throughout the area.

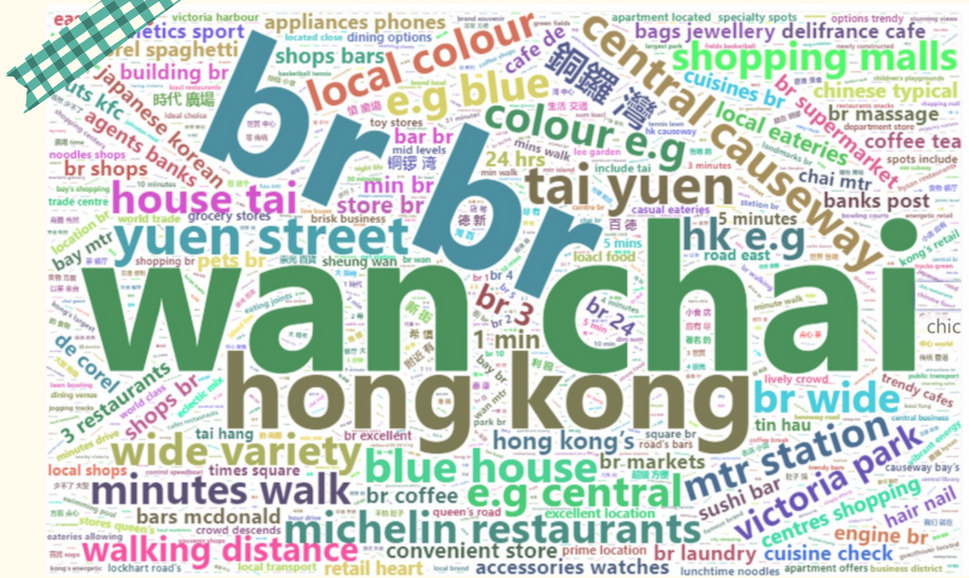


Wordcloud

Variable analyzed: neighborhood_overview

“br”, “Chai”, “Wan,” “restaurants,” “shopping”“centre”,“Causeway,” and “international”

- underscore the commercial dimension of Wan Chai, a diverse range of dining and shopping options



“wan chai”, “hong kong,” “br br”, references to attractions in Chinese, and specific street names like “Yuen Street” or station names

- cultural and tourist highlights
- the convenience and accessibility of Wan Chai - a well-connected district with a focus on local attractions and its close proximity to transportation hubs

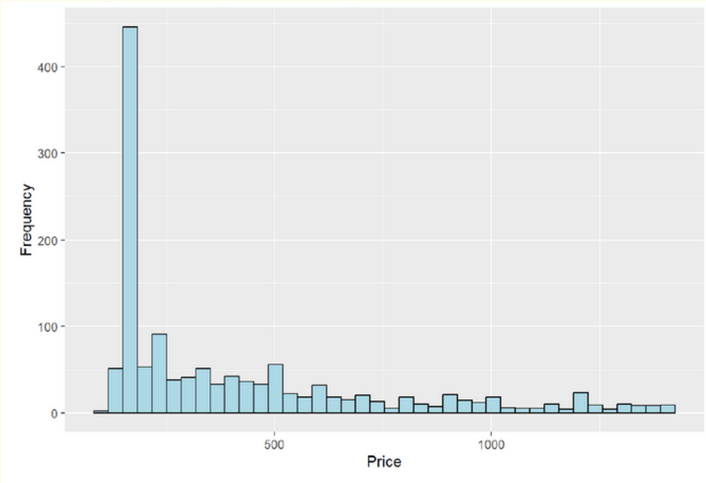
STEP 2

Prediction

Multiple Regression

We will predict the price of an Airbnb rental in HK$

Reminder: price particularities



- Right-skewed
- Presence of significant outliers

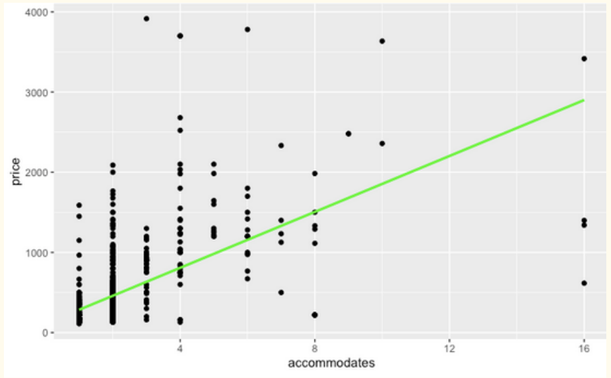
Models building process

1. Data partitioning & variable adjustments

2. Variable elimination:
- with unique values (categorical)
- overly specific
- highly correlated & redundant

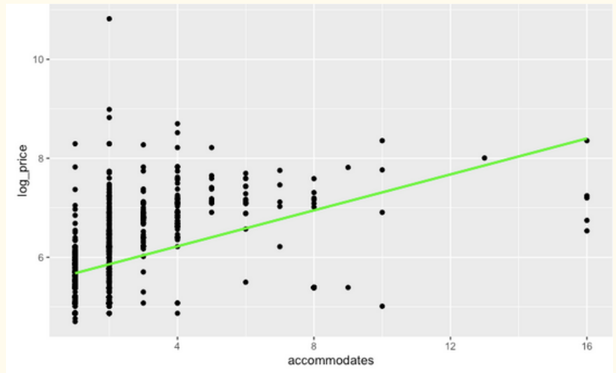
3. Conversion of categorical variables into factors

4. Models creation using backward elimination.



price

Vs.



log(price)

- Relationship is more linear
- Mitigates impact of outliers

2 models to compare:

Response variable
- price
- log(price)

Assessing the models

Model predicting price

Multiple R-squared: 0.5896

F-statistic: 76.36

Validation R-squared: 0.5595

Vs.

Model predicting log(price)

Multiple R-squared: 0.7875

F-statistic: 212.4

Validation R-squared: 0.74656

Validation R-squared + 0.198

Better predictive performance

→ The model doesn't seem overfit.

The significant input variables

- host_response_time
- host_acceptance_rate
- host_listings_count
- host_identity_verified
- room_type
- accommodates
- minimum_nights
- instant_bookable

But...

- RMSE: 388.1555
- Min. residual: -2000.209
- Max. residual: 2027.676

Vs.

- Sd of log(price) in training set 533.3033

Model's prediction errors < natural variability

→ The model makes a mistake of HK$388 on average.
The error margins are not negligible since most prices are ⩽ HK$250.

STEP 3

Classification

K-Nearest Neighbors

We will predict whether a rental will have a washer

The input variables

minimum_nights
maximum_nights

Determines wether a customer would need a washer or not

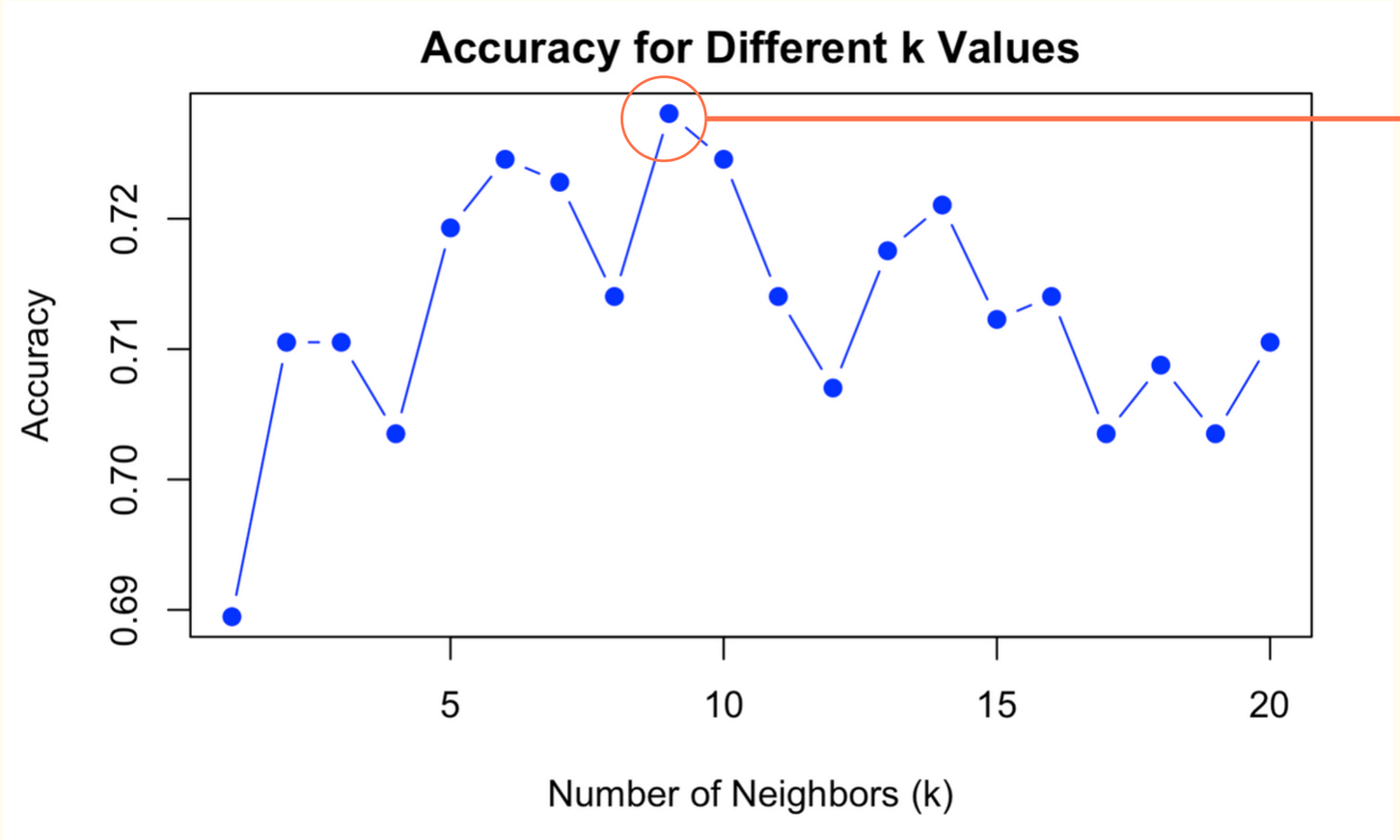
price

If a rental is more expensive, it may include a washer

host_listings_count

If the host has more listings it may be a professional company renting apartments where washers are included in the building.

Determining the ideal value of K



Assessing the model

The accuracy of the k-NN model is calculated as the proportion of correct predictions

		Actual	
		False	**True**
Predicted	**False**	**171**	75
	True	82	**242**

Naive rule

0.553

Our model is 23.6% more accurate

(TP + TN) / (TP + TN + FP + FN)
(242+171) / (242+171+82+75) =
413/570=

Accuracy: 0.724

Naive Bayes

The output variable:
review scores

The input variables

We selected variables we thought have more influence on ratings

Amenities

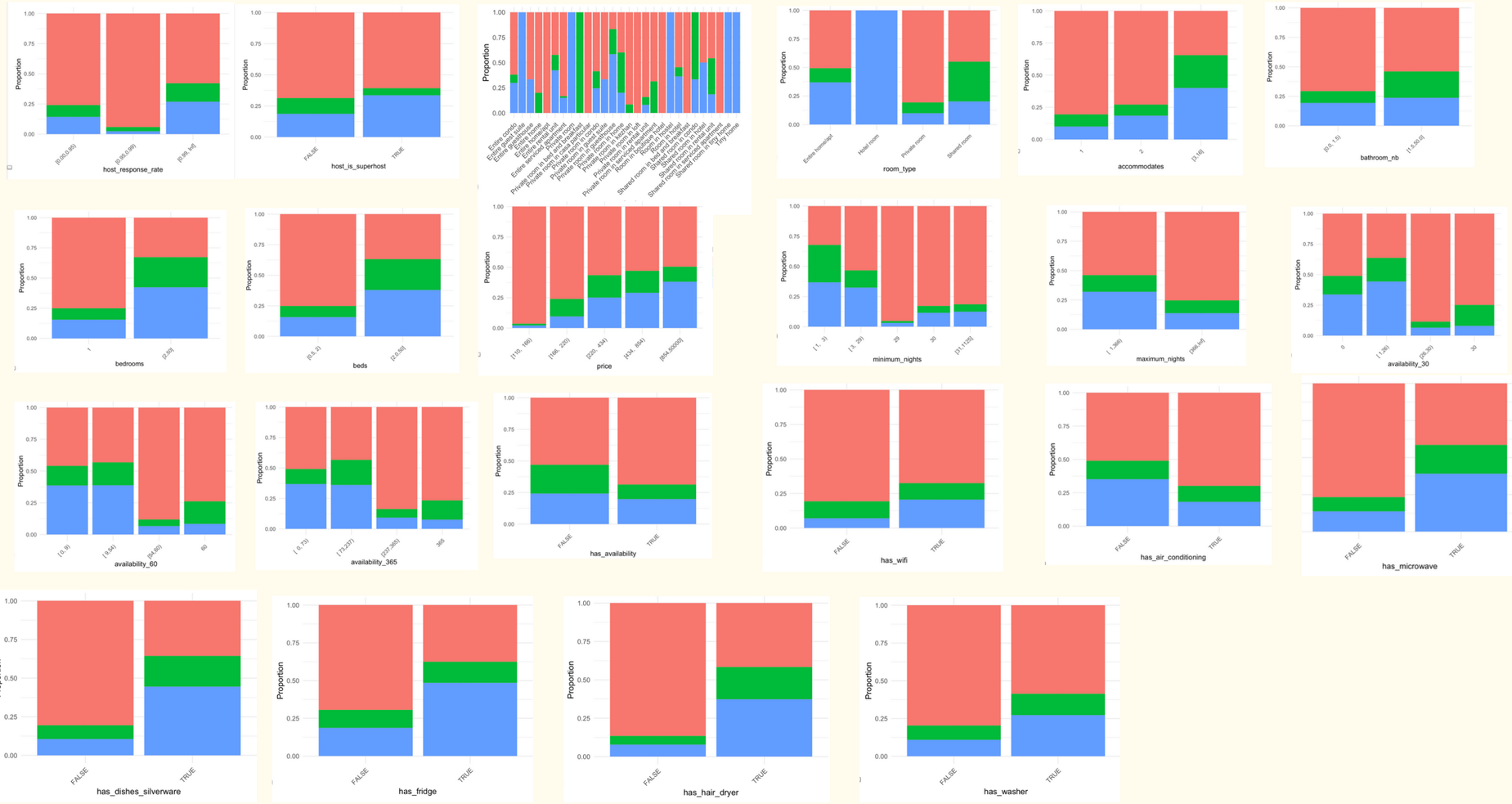
air conditioning
washer
wifi
microwave
dishes and silverware
mini fridge
hair dryer

Other variables

host_response_rate
host_is_superhost
property_type
room_type
accommodates
bathroom_nb
bedrooms
beds
price
minimum_nights
maximum_nights
availability_30
availability_60
availability_365
has_availability

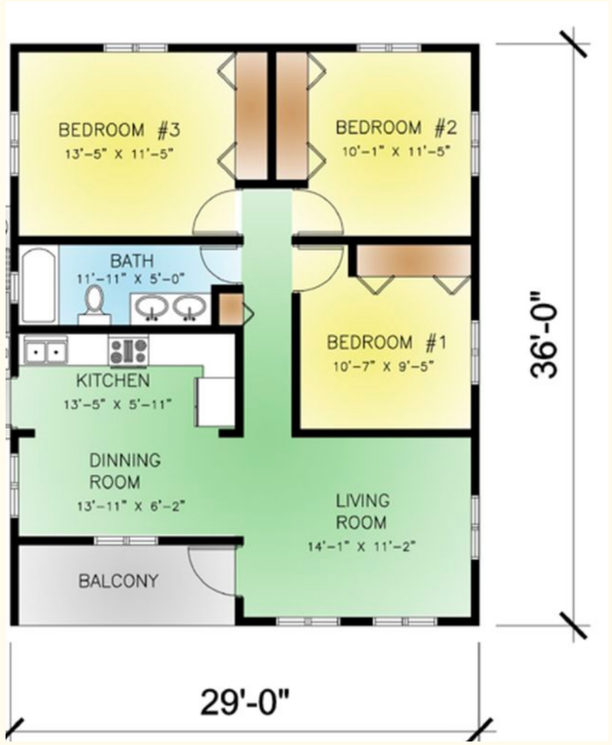
We reviewed their predictive power using proportional bar plots:

All variables differentiate the outcome



Fictional property prediction

We used the model to predict the ratings for a fictional rental.



3BD, 1 Bath, no washer, with wifi +fridge, super host

The result:

Bin 2: 1.0 - 4.6

We decided to keep all the variables because they can all differentiate the outcome

Accuracy on the training set: 0.7143

Accuracy on the validation set: 0.7474.

Confusion Matrix and Statistics

	Reference		
Prediction	0.0	[1.0,4.6)	[4.6,5.0]
0.0	441	29	31
[1.0,4.6)	17	40	19
[4.6,5.0]	103	45	129

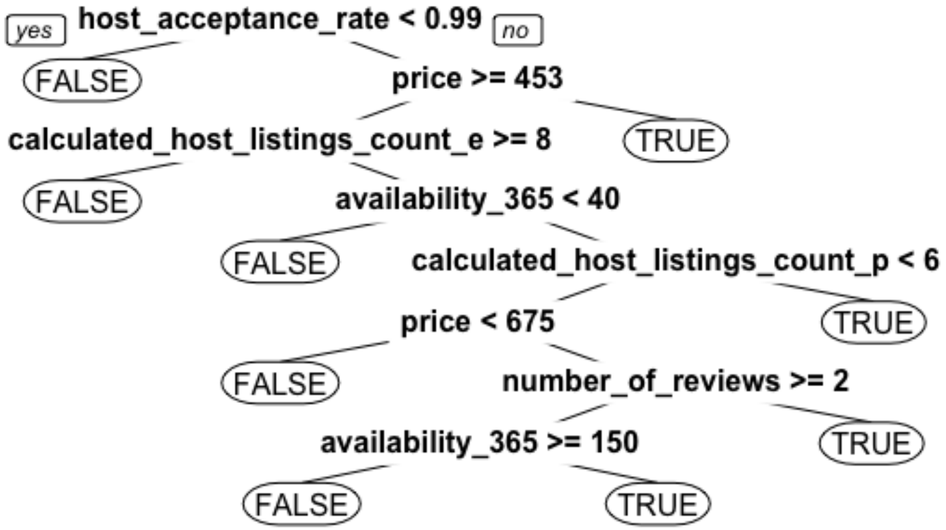
	Reference		
Prediction	0.0	[1.0,4.6)	[4.6,5.0]
0.0	326	13	13
[1.0,4.6)	17	16	8
[4.6,5.0]	64	29	84

This means the model has a high predictive power. It is not overfitting to the training data and is making accurate predictions on unseen data.

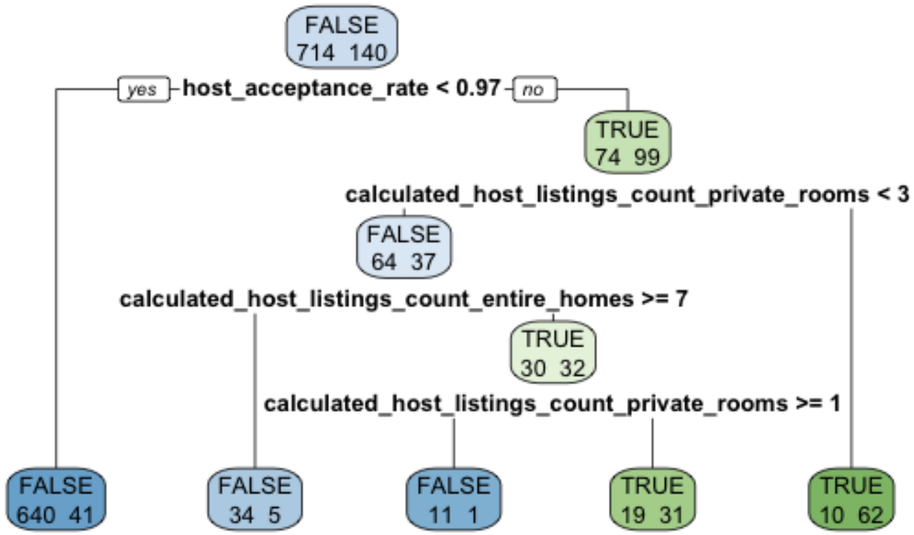
Classification Tree

Predict : Instant Bookable

Tree without bias



Tree with complexity parameter



Evaluation of the Model

- High accuracy of 87%
- Low overfitting
- Both trees have close accuracy
- Better at predicting not instant bookable

STEP 4

Clustering

K-Means

Choice of Variables

Numerical variables

K-means works with numerical variables.

Amount of variables

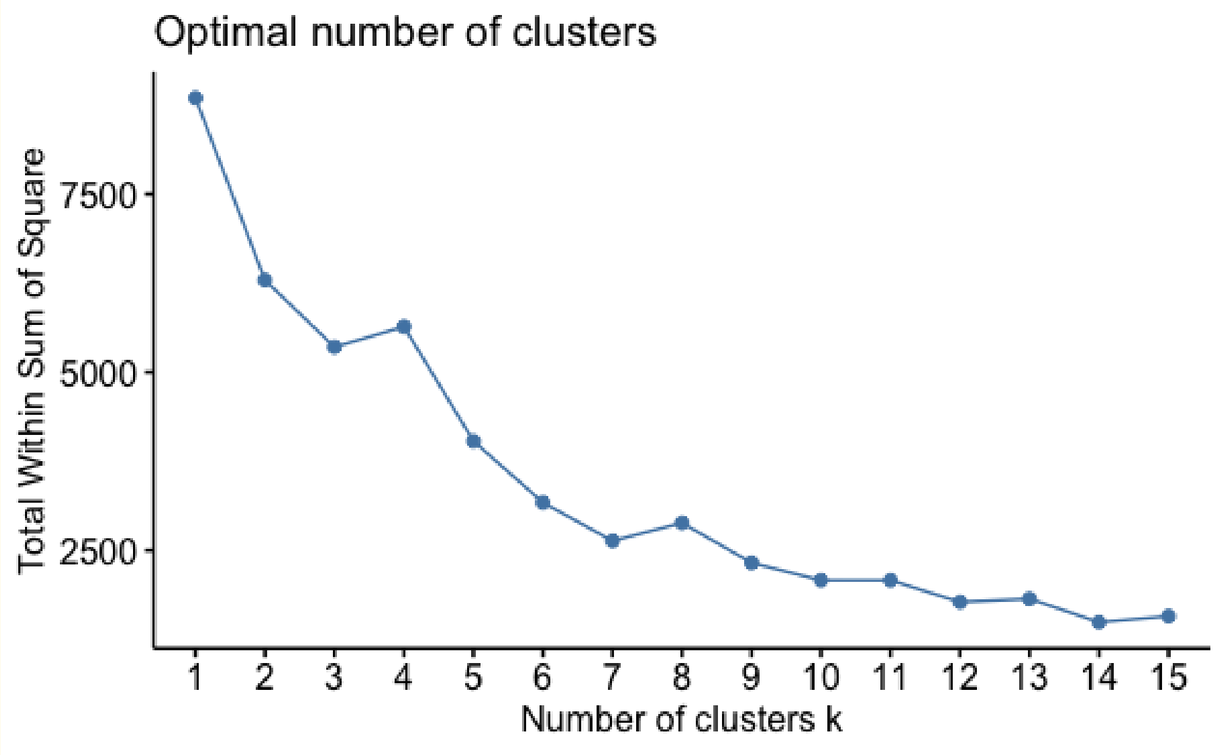
The model has to be easy to interpret with less variables.

Correlation

High correlation exaggerates certain variables.

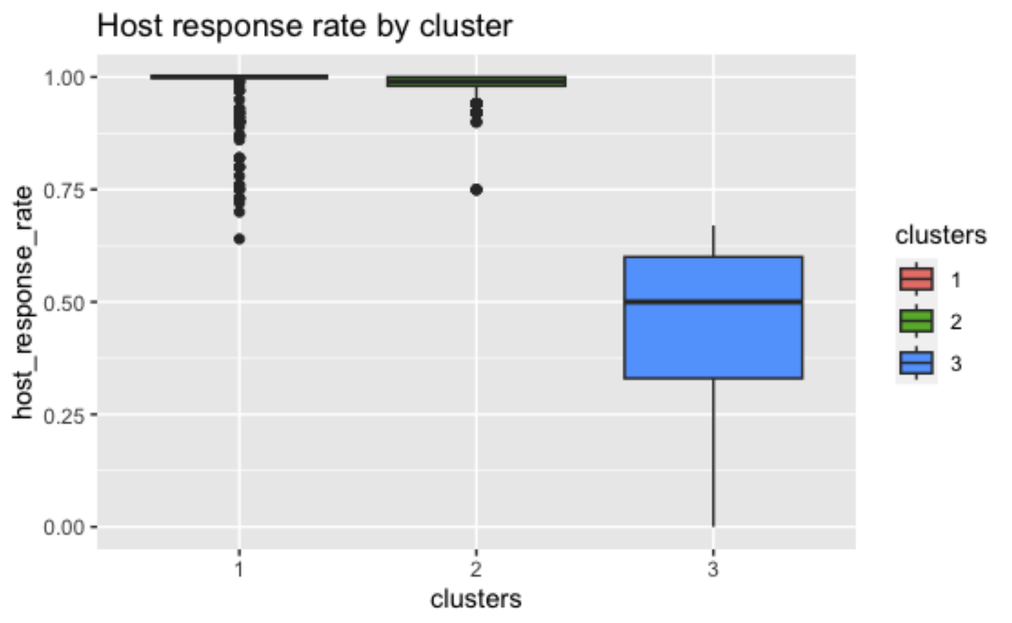
Choosing the # of clusters

Elbow plot

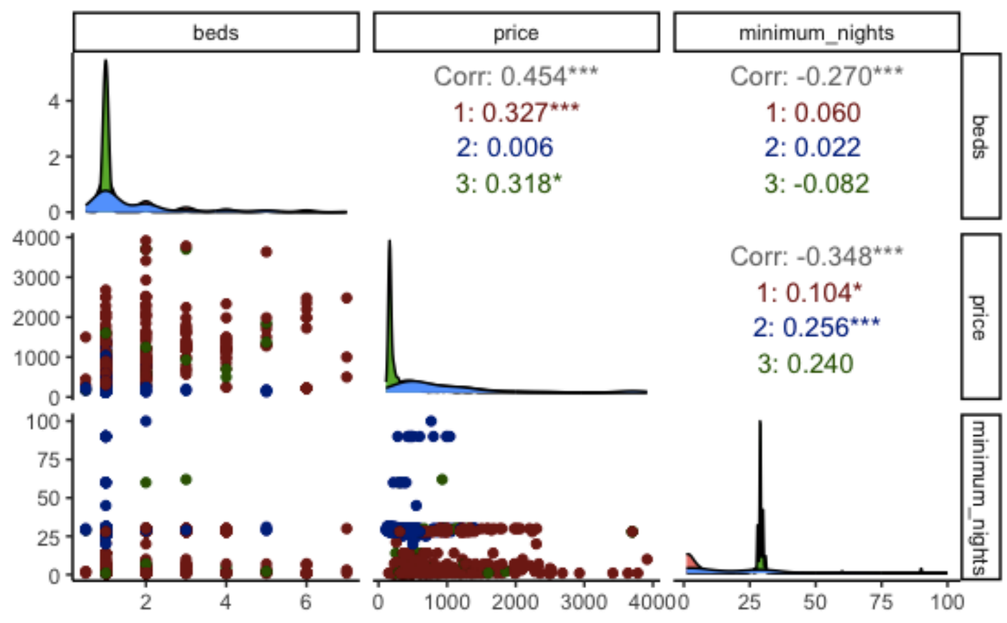


Understanding the clusters

Host response per cluster

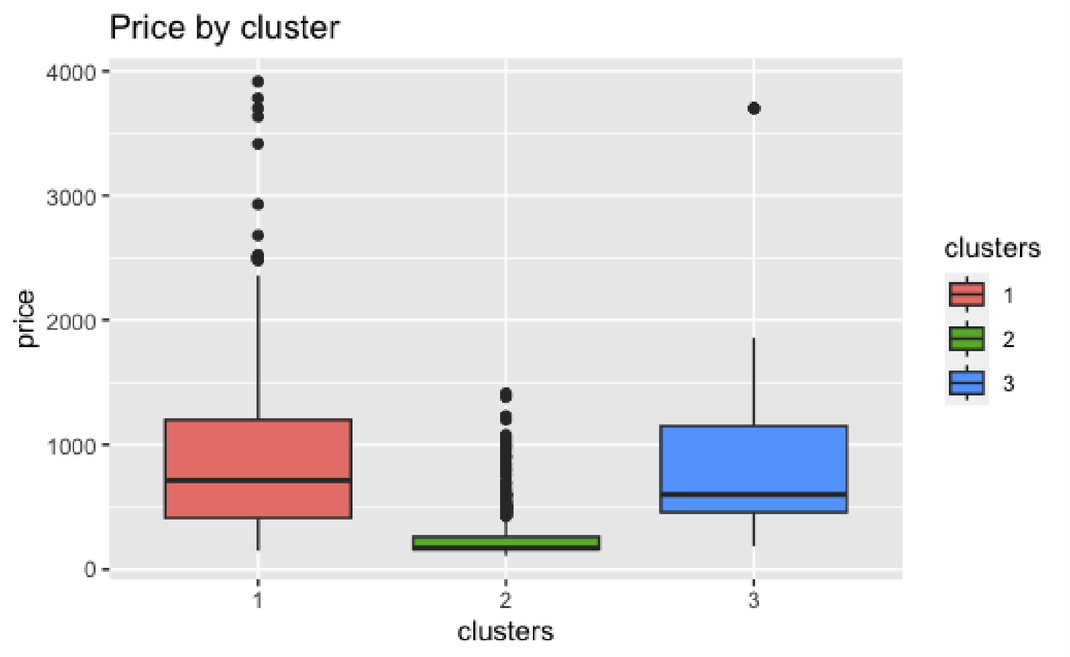


Pair plot for beds, price and minimum nights

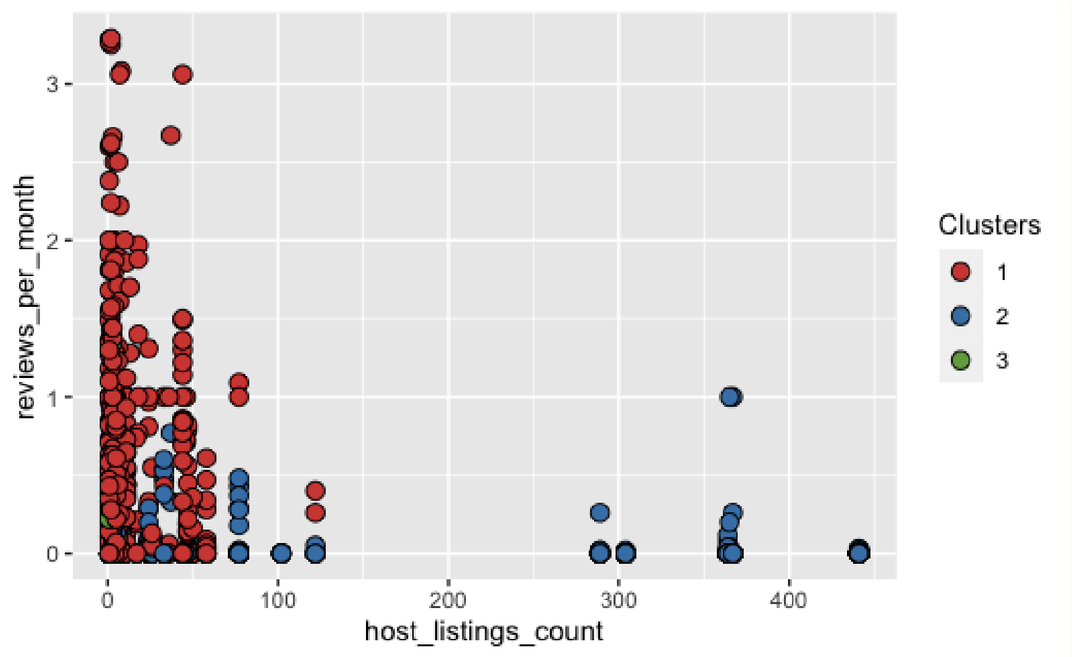


Understanding the clusters

Price per cluster



Number of listings and reviews per month relationship



Meaning of the Clusters

Cluster 1: Popular Short Stays

- high price range
- Low number of minimum nights, indicating a focus on short stays.
- Higher number of reviews high reputation
- Lower host listings count, owners are not corporate.

Cluster 2: Budget Long-Stay

- With low prices,
- high minimum nights,
- low bed counts
- Cheaper long term stays

Cluster 3: Mid-Market Mixed-Stay

- The moderate price range
- average minimum nights
- average bed counts
- suggest this cluster is a mix of both short and long stays at medium price points.

.

Conclusion

Challenges

- Massive data set
- Selecting the right variables for the job
- Fundamental understanding of models

Insights

- Visualizations give direction
- Value of feature reduction
- Dependance of model performance on data format