Data Mining for Airbnb Data in Monte Verde, Rome, Italy

Haowen Wang

Agenda

- Data preparation & Exploration
- Prediction
- Classification
- Clustering
- Conclusion

Data preparation & Exploration

- Missing Values on Selected Variables
- Summary Statistics on Price
- Data Visualization on Review_scores_rating
- Mapping
- Wordcloud

Narrow down the variables that we focus on

```
"id"                    "description"              "neighborhood_overview"   "host_id"
"host_name"             "host_since"               "host_response_time"      "host_response_rate"
"host_acceptance_rate"  "host_total_listings_count" "host_has_profile_pic"   "host_identity_verified"
"property_type"         "accommodates"             "amenities"               "price"
"minimum_nights"        "maximum_nights"           "has_availability"        "number_of_reviews"
"review_scores_rating"  "instant_bookable"
```

Variables with missing values:

	Variable	Missing_Count
description	description	29
neighborhood_overview	neighborhood_overview	525
review_scores_rating	review_scores_rating	201

Solution for missing values

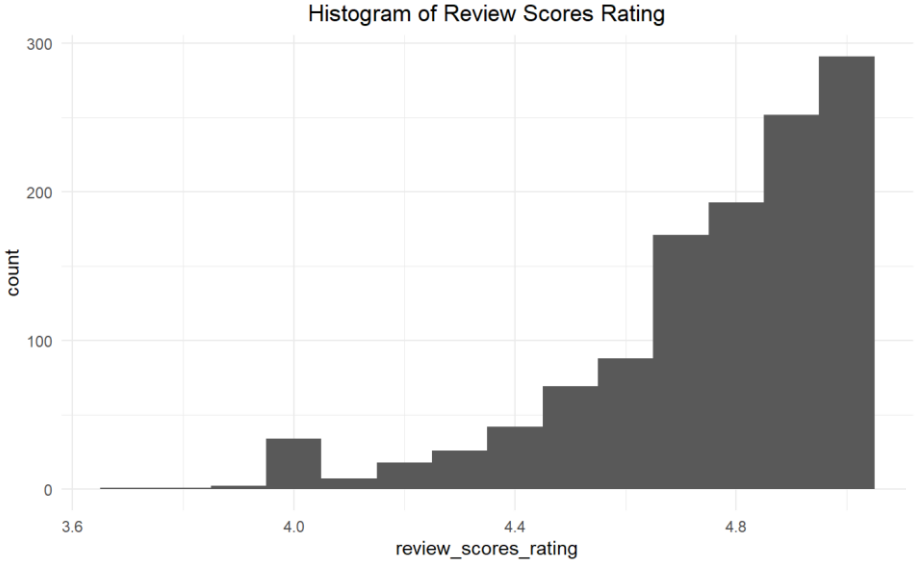
- Replace na with median value for review_scores_rating
  - Too many rows to be just taken out (about 14% of data)
  - Outliers exist which affect the “true” mean value

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.00    4.65    4.83    4.71    4.95    5.00     201
```

- Replace na with empty space for description and neighborhood_overview:
  - Empty space simply means no description/words written, instead of purely nothing
  - Does not change their original meanings or affect text mining process.

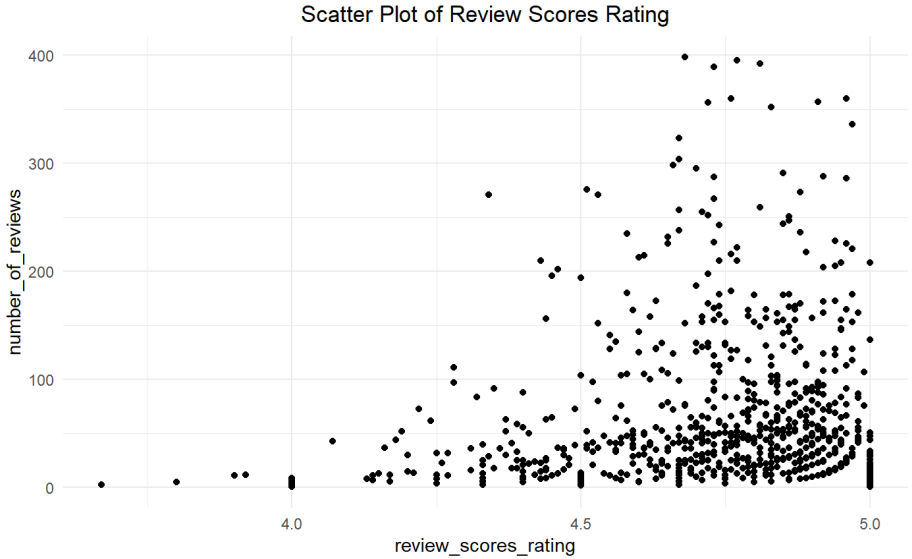
- Summary stats table for price on the right
  - Is 9999 a mistake?
  - High standard deviation and mean as a result from 9999
  - Median price at 110 seems reasonable

	metric	value
1	mean	146.6413
2	median	110.0000
3	minimum	15.0000
4	maximum	9999.0000
5	standard deviation	340.7359

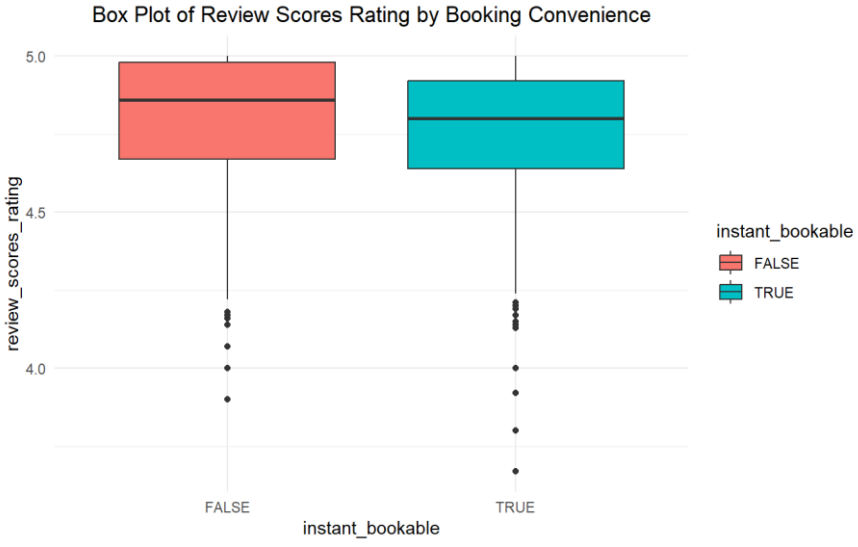


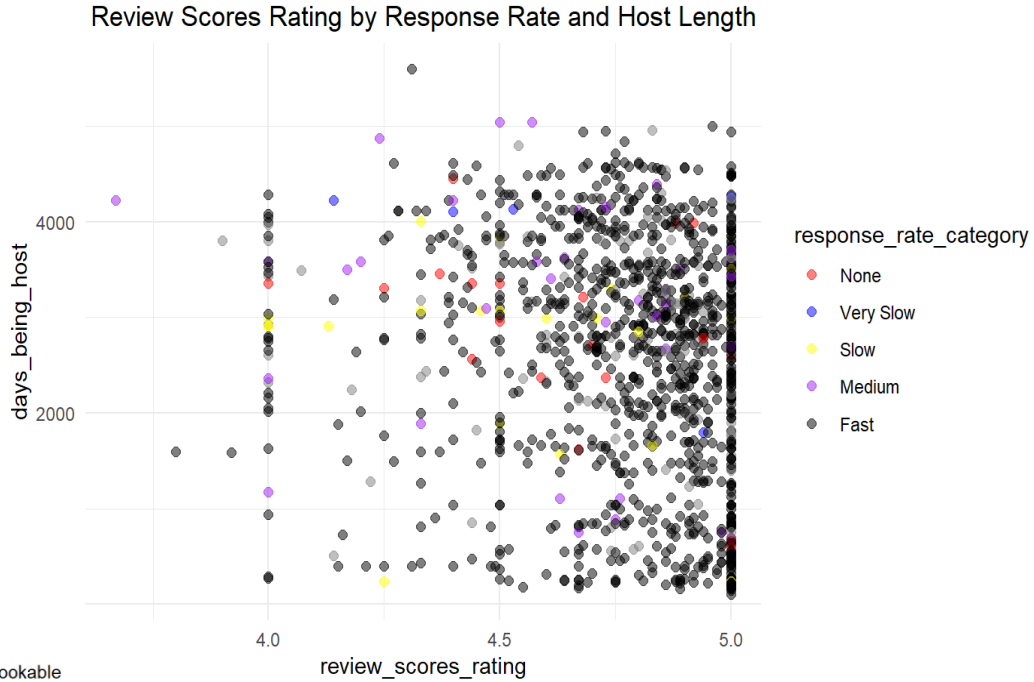
Visualizations on review_scores_rating

- Filter out scores less than 3.5 to make graphs more readable
- Most Airbnb have relatively high scores
- Most have under 100 reviews
- More reviews not equal to higher scores

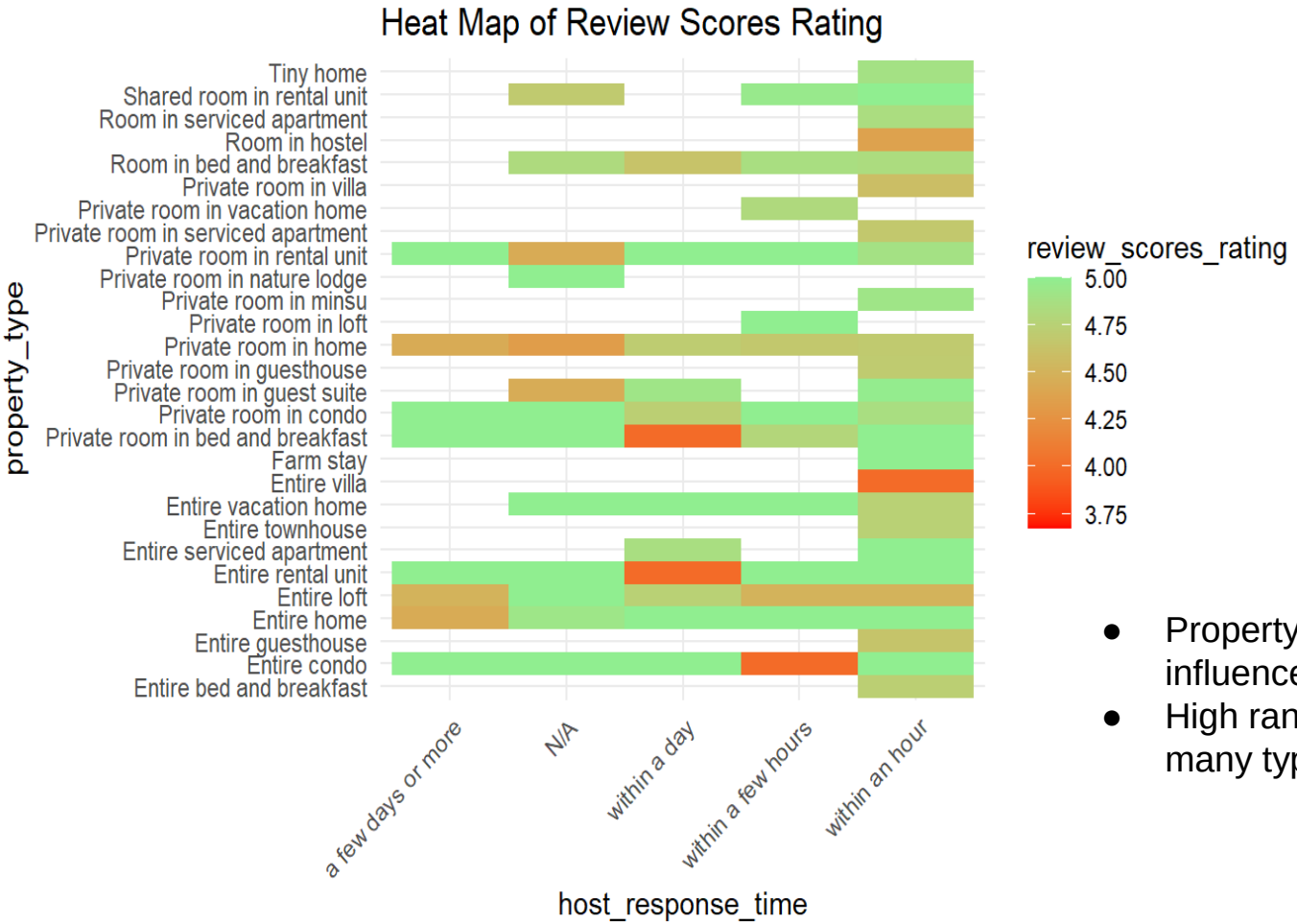


- Review scores rating are not affected by how long the person has been a host
- Majority of 5s are fast responders
- The slow and very slow response tend to get lower ratings, but not by much



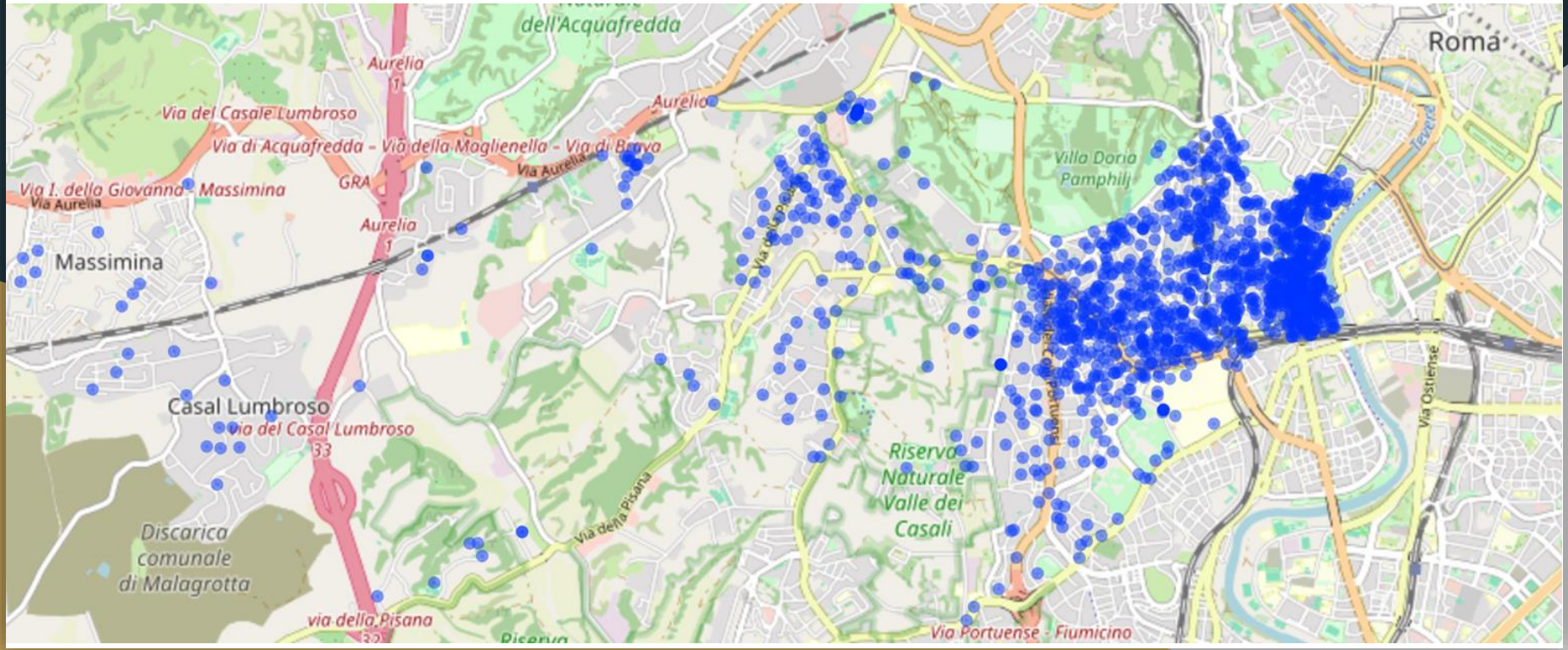


- Properties that are not instant bookable have slightly lower scores



- Property type does not influence ratings that much
- High range of scores across many types

Property location map



Neighborhood overview wordcloud

- Getting rid of stop words in both English and Italian
- ‘Br’ was originally the most common word, but means line break
- ‘Roma’/‘Rome’ is where Monteverde belongs to
- ‘trastevere’ is considered the heart of Rome
- ‘Quartiere’ means an area in a town/city in Italian

Multiply Linear Regression

- Data wrangling
- Model selection
- Fitting the model
- Multiple linear regression & improvement
- Performance Measurement

Data wrangling

- Replace NA with median value (Review_scores_rating, Bedrooms, beds, Host_is_superhost)

- Extract numeric information from price variable
- Log transform price variable
- Standarize availability_365

$1,000.00
$1,000.00
$1,000.00
$1,000.00
$1,129.00

- Dummify variables— room_type

```
room_typeEntire home/apt
room_typeHotel room
room_typePrivate room
```

- Transfer boolean values to numeric(0/1)

host_is_superhost
TRUE
TRUE
TRUE

Model selection— minimize collinearity

High Correlating variables: greater than 0.6

Group1:

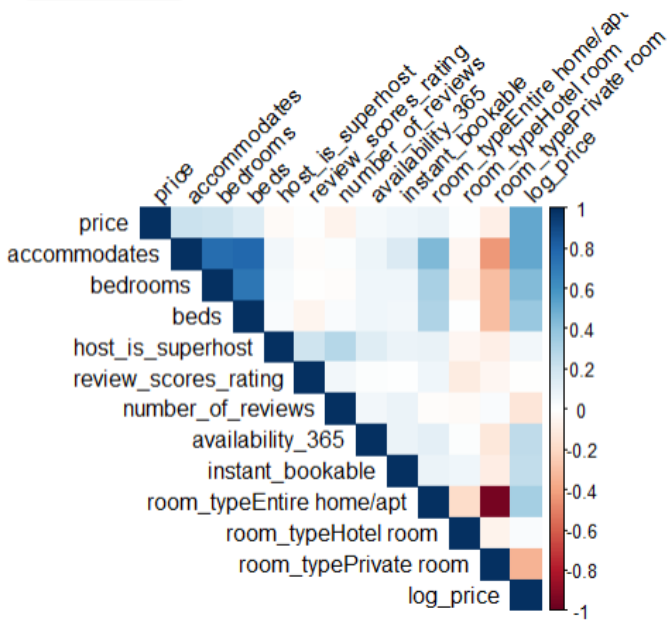
Accommodates vs bedrooms

Accommodates vs beds

Group2:

Room_type/entire apartment

vs Room_type/private room



Fitting the model

- Split dataset into training dataset & validating dataset
- Portion 0.6/0.4

- Predictors: bedrooms beds host_is_superhost review_scores_rating number_of_reviews availability_365 instant_bookable
- room_typeEntire home/apt room_typeHotel room

- Response variable: log price

Multiple linear regression

Variable with P value > 0.05:

- Host is superhost
- review_score _rating
- room type/hotel room

R-squared :0.371

```
Call:
lm(formula = log_price ~ . - price, data = train.df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76320 -0.30066  0.00095  0.26029  2.91210

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.77929    0.20153  18.753  < 2e-16 ***
bedrooms                  0.27113    0.03327   8.149 1.38e-15 ***
beds                      0.03728    0.01801   2.070   0.0388 *
host_is_superhost         0.01825    0.03928   0.465   0.6423
review_scores_rating      0.03091    0.04183   0.739   0.4601
number_of_reviews        -0.07905    0.01777  -4.449 9.83e-06 ***
availability_365          0.12548    0.01700   7.382 3.84e-13 ***
instant_bookable          0.17953    0.03464   5.183 2.76e-07 ***
`room_typeEntire home/apt` 0.26160    0.04268   6.130 1.37e-09 ***
`room_typeHotel room`     0.29225    0.17498   1.670   0.0953 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4828 on 816 degrees of freedom
Multiple R-squared:  0.371,     Adjusted R-squared:  0.3641
F-statistic: 53.48 on 9 and 816 DF,  p-value: < 2.2e-16
```

Backward elimination

- Using the highest p value to remove
- Mitigation of Multicollinearity
- Improved Generalization: training set

```
lm(formula = log_price ~ bedrooms + beds + number_of_reviews +
    availability_365 + instant_bookable + `room_typeEntire home/apt` +
    `room_typeHotel room`, data = train.df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76003 -0.29673  0.00272  0.26068  2.91064

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 3.92973    0.04376  89.812  < 2e-16 ***
bedrooms                    0.27114    0.03325   8.154 1.32e-15 ***
beds                        0.03656    0.01798   2.033   0.0424 *
number_of_reviews          -0.07600    0.01706  -4.454 9.61e-06 ***
availability_365            0.12676    0.01686   7.516 1.48e-13 ***
instant_bookable            0.18018    0.03451   5.222 2.25e-07 ***
`room_typeEntire home/apt`  0.26559    0.04235   6.272 5.77e-10 ***
`room_typeHotel room`       0.28318    0.17457   1.622   0.1052
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4825 on 818 degrees of freedom
Multiple R-squared:  0.3703,    Adjusted R-squared:  0.3649
F-statistic: 68.71 on 7 and 818 DF,  p-value: < 2.2e-16
```

Performance Measurement

- Limitations in accurately predicting prices, especially underestimating them.
- The higher RMSE and MAE on the test set compared to the training set may indicate some degree of overfitting or lack of generalization to new data.
- Further model refinement and tuning may be necessary to improve predictive accuracy, especially considering the negative trends in ME, MPE, and potentially high MAPE.

	ME	RMSE	MAE	MPE	MAPE
Test set	21.73021	251.6526	55.08459	-11.5602	38.57608

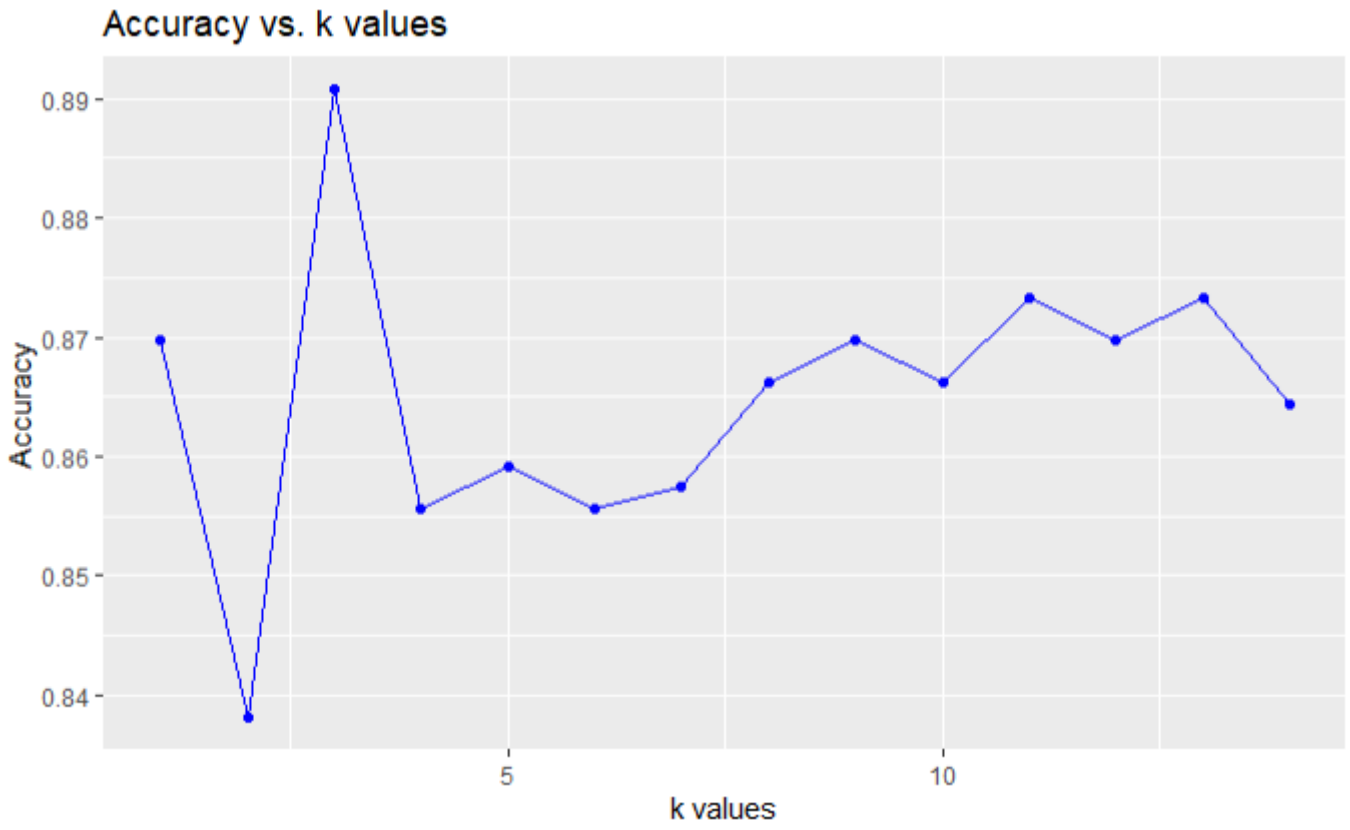
	ME	RMSE	MAE	MPE	MAPE
Test set	43.47765	427.852	71.64218	-4.864977	37.1256

Classification

- K-NN
- Naive Bayes
- Classification tree

K-NN Model

- Target variable: Amenities (Kitchen)
- Kitchen is the classic label of large and modern apartments.
- Generated two new columns called number of bedrooms and number of beds by extracting information from description column
- Input variables: price, accommodates, number of bedrooms, number of beds , calculated private rooms, calculated shared rooms
- Two sample t-test for feature selection with threshold of 0.1.
- Using all six features to find and plot the best k parameter.
- Confusion Matrix


Accuracy vs. k values
0.89
0.88
0.87
0.86
0.85
0.84
Accuracy
5
10
k values

Summary Statistics for K-NN Model

	Actual 0	Actual 1
Predicted 0	29	20
Predicted 1	42	477

- Accuracy rate = 89.08%
- TPR rate = 95.98%
- Positive class : ‘1’ (Those apartments with kitchen)

Naive Bayes Classification

- Goal: Classify and predict review scores rating into 4 categories: Low, Mid, High, NA
- Algorithm: Naive Bayes
- Feature selected:
  - Superhost: T or F
  - Host Response Time: A few days or more, NA, within a day, within a few hour, with an hour
  - Host Acceptance rate: low, high
  - Room type: Entire home/apt, Hotel room, Private room, Share room

Naive Bayes Performance

Train Performance

Prediction	Reference low	mid	high	NA_cate
low	191	20	6	20
mid	84	122	6	29
high	104	92	12	31
NA_cate	75	10	4	34

Accuracy Rate: 0.4274

Validation Performance

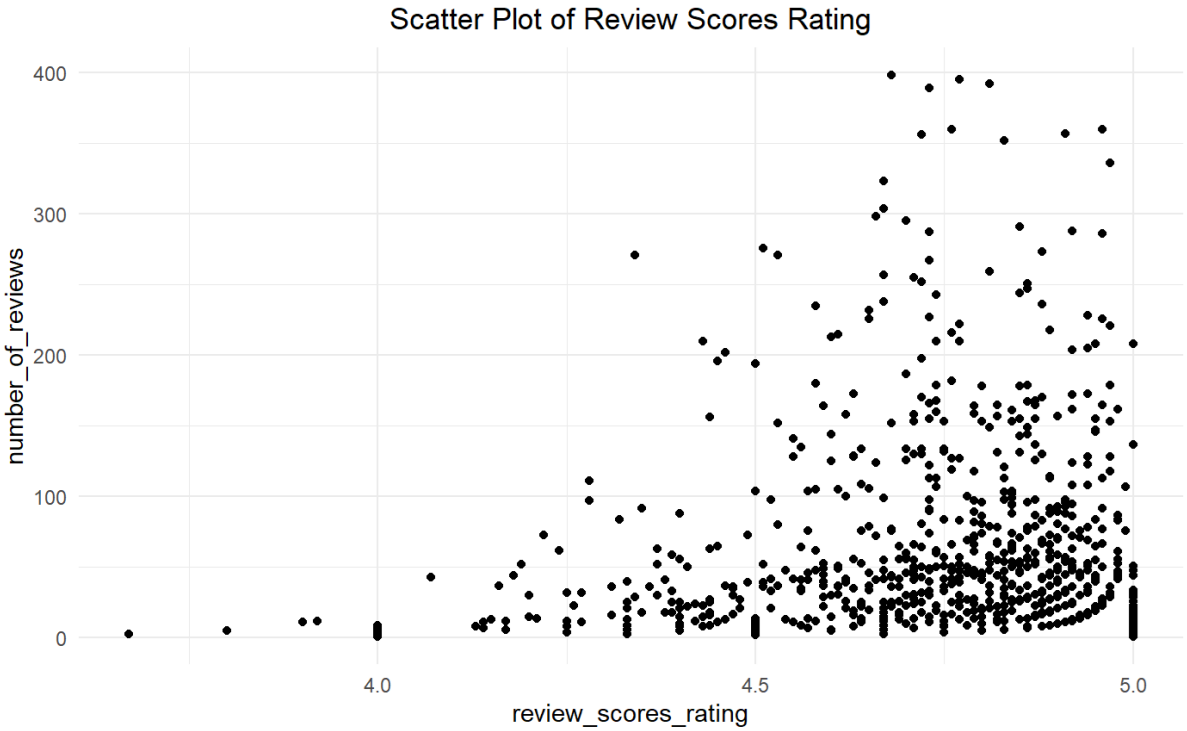
Prediction	Reference low	mid	high	NA_cate
low	125	18	10	16
mid	73	79	4	16
high	48	66	10	18
NA_cate	44	5	3	25

Accuracy Rate: 0.4268

- Evenly distributed outcome
- 42%> 25%, better performance than naive rule
- Similar Accuracy for both Train and Validation, no overfitting

Naive Bayes Test

- Fictional House:
  - Host is a superhost
  - Response within an hour
  - High acceptance rate
  - Private room
- Result: ‘Mid’ level Review Scores Rating


Scatter Plot of Review Scores Rating
number_of_reviews
400
300
200
100
0
4.0
4.5
5.0
review_scores_rating

Classification Tree

- Target variable: Instant bookable
- Input variables: minimum nights, maximum nights, minimum minimum nights, maximum minimum nights, minimum maximum nights, maximum maximum nights, minimum nights average ntm, maximum nights average ntm, availability 30, availability 60, availability 90, availability 365.
- Fit the tree with all inputs
- Cross-Validation
- Generate new trees based on best parameters
- Confusion Matrix

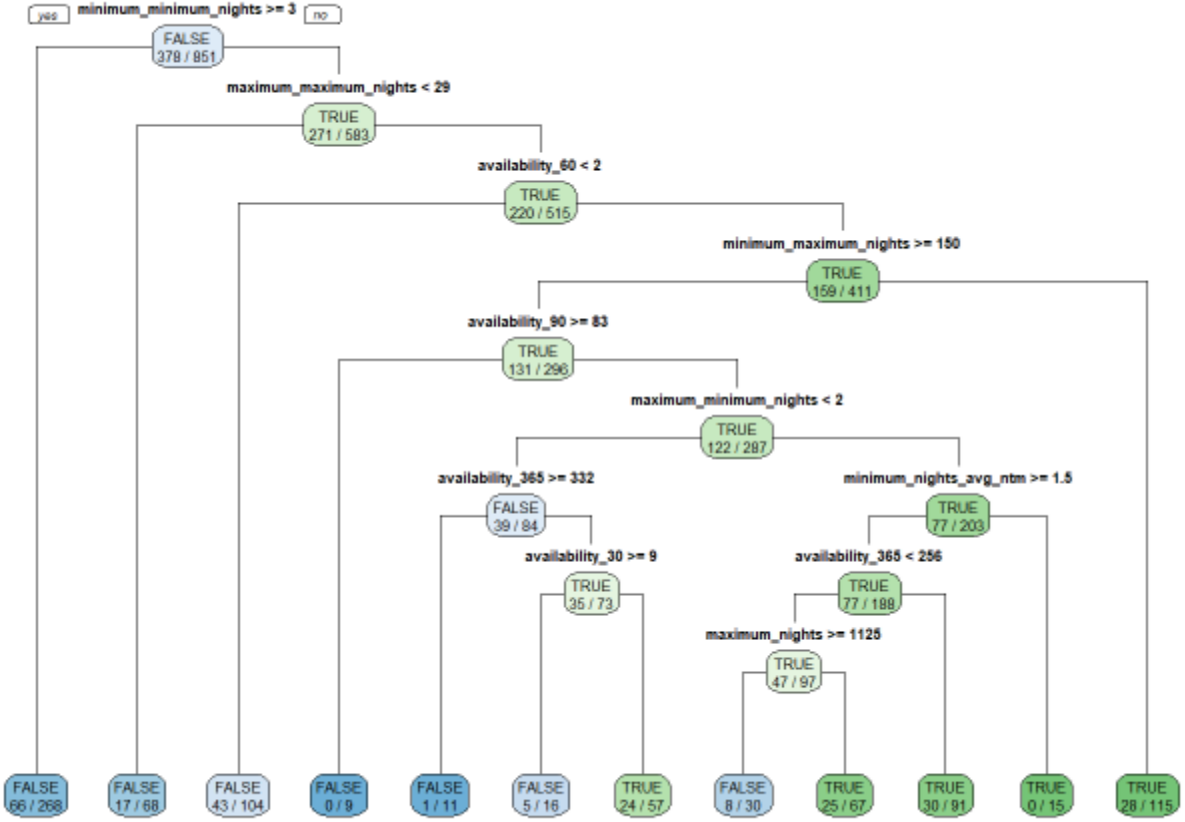
```
Variables actually used in tree construction:
 [1] availability_30        availability_365       availability_60
 [4] availability_90        maximum_maximum_nights maximum_minimum_nights
 [7] maximum_nights         maximum_nights_avg_ntm minimum_maximum_nights
[10] minimum_minimum_nights minimum_nights_avg_ntm

Root node error: 378/851 = 0.44418

n= 851

        CP nsplit rel error  xerror     xstd
1 0.108466      0   1.00000 1.00000 0.038346
2 0.089947      1   0.89153 1.02910 0.038445
3 0.047619      2   0.80159 0.84127 0.037335
4 0.011905      3   0.75397 0.81481 0.037087
5 0.010582     11   0.65344 0.80423 0.036981
6 0.010000     13   0.63228 0.82275 0.037164
```

- Complexity Parameter of 0.010582 with number of split of 11 provides the least x error here.


yes
minimum_minimum_nights >= 3
no
FALSE
379 / 851
maximum_maximum_nights < 29
TRUE
271 / 583
availability_60 < 2
TRUE
220 / 515
minimum_maximum_nights >= 150
TRUE
159 / 411
availability_90 >= 83
TRUE
131 / 296
maximum_minimum_nights < 2
TRUE
122 / 287
availability_365 >= 332
FALSE
39 / 84
availability_30 >= 9
TRUE
35 / 73
minimum_nights_avg_ntm >= 1.5
TRUE
77 / 203
availability_365 < 256
TRUE
77 / 188
maximum_nights >= 1125
TRUE
47 / 97
FALSE
66 / 268
FALSE
17 / 68
FALSE
43 / 104
FALSE
0 / 9
FALSE
1 / 11
FALSE
5 / 16
TRUE
24 / 57
FALSE
8 / 30
TRUE
25 / 67
TRUE
30 / 91
TRUE
0 / 15
TRUE
28 / 115

Summary Statistics

Train.df	Actual False	Actual True
Predicted False	366	140
Predicted True	107	238

Valid.df	Actual False	Actual True
Predicted False	233	103
Predicted True	97	135

	Accuracy	TPR
Train.df	70.98%	62.96%
Valid.df	64.79%	56.72%

- No signs of overfitting here
- The True Positive Rate is a measure of how well a model is at capturing the positive cases

Clustering Analysis

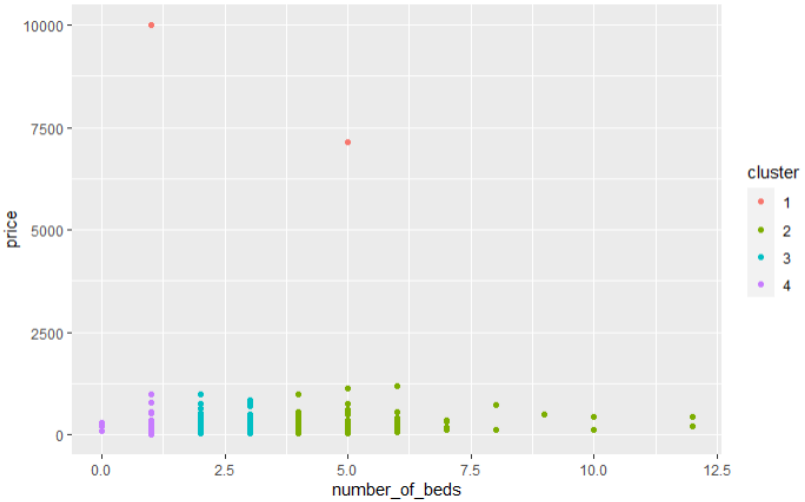
- Clustering Comparison
- Categories Analysis

Cluster Analysis

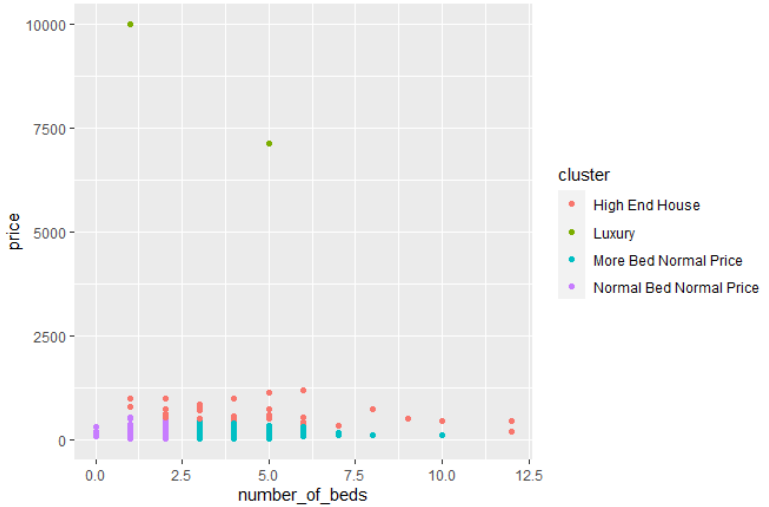
- Goal: Cluster houses into different categories
- Algorithm: K-mean Clustering with k=4
- Feature selected: Scaled Number of beds, Scaled Price
- Weight of each Feature: Number of Bed*100%, Price*200%
- Cluster Label: **Normal Bed Normal Price, More Bed Normal Price, High End House, Luxury**

Clustering Comparison

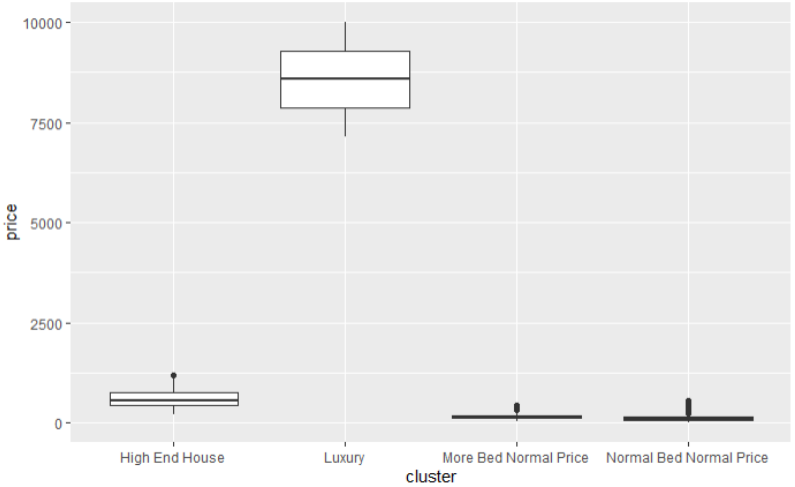
Even Weighted Clustering result

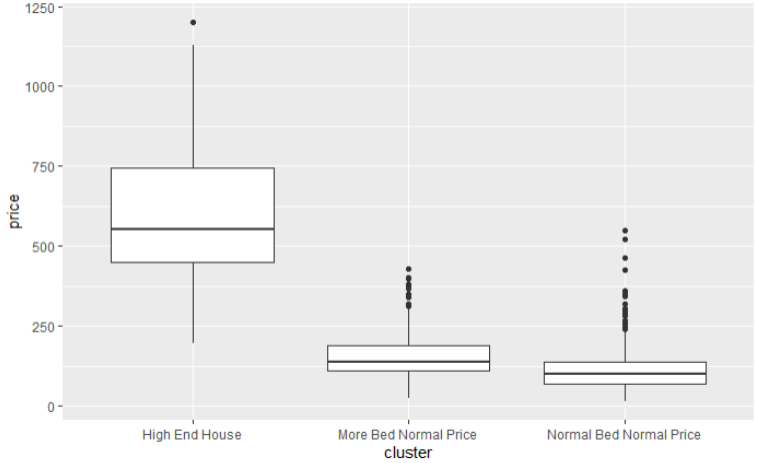


Double Price Weighted Clustering result



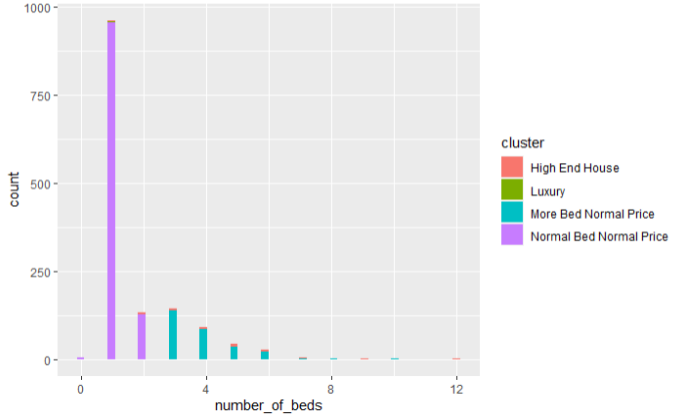
Categories Analysis

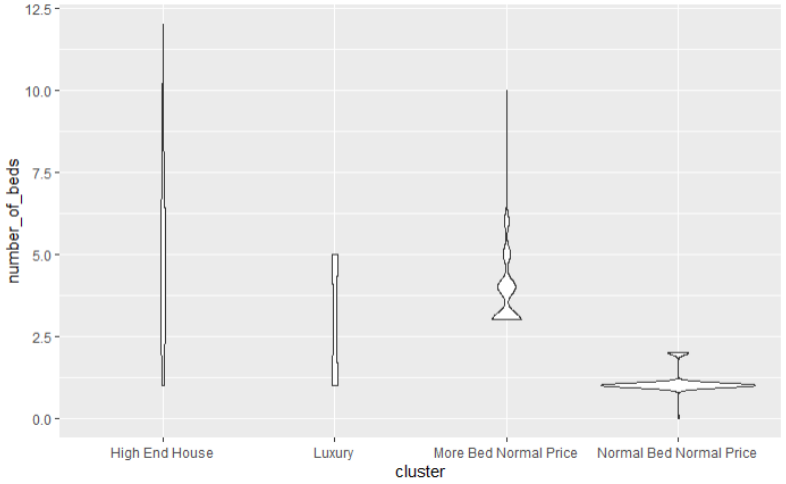




Price Range: Luxury> >High End house> More Bed Normal Price>=Normal Bed Normal Price

Categories Analysis





- Luxury: 1 bed and 5 beds
- Normal Bed Normal Price: 1 to 2 beds
- More bed Normal Price: > 3 beds
- High end house: various range

Conclusion

- Importance of data preparation process and effective ways to track useful data in the future
- Clustering analysis for segmentation(i.e. using surveys to put customers into groups)
- Other ways to deal with missing values