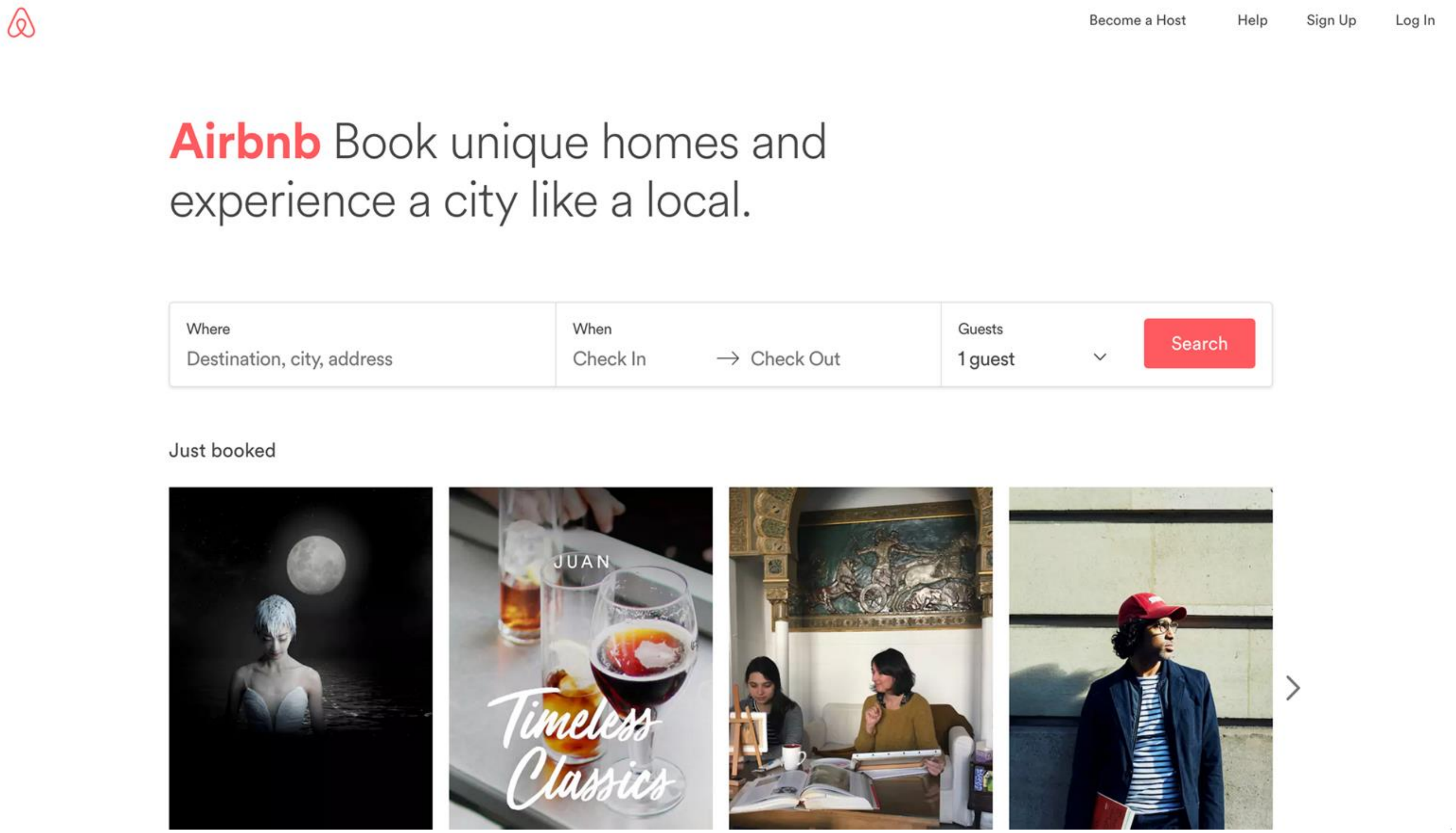
HOW MUCH WOULD I PAY?

Team Members:
Xinyi Gong, Lanyu Shang, Yang Sun

OVERVIEW

Overview

OVERVIEW

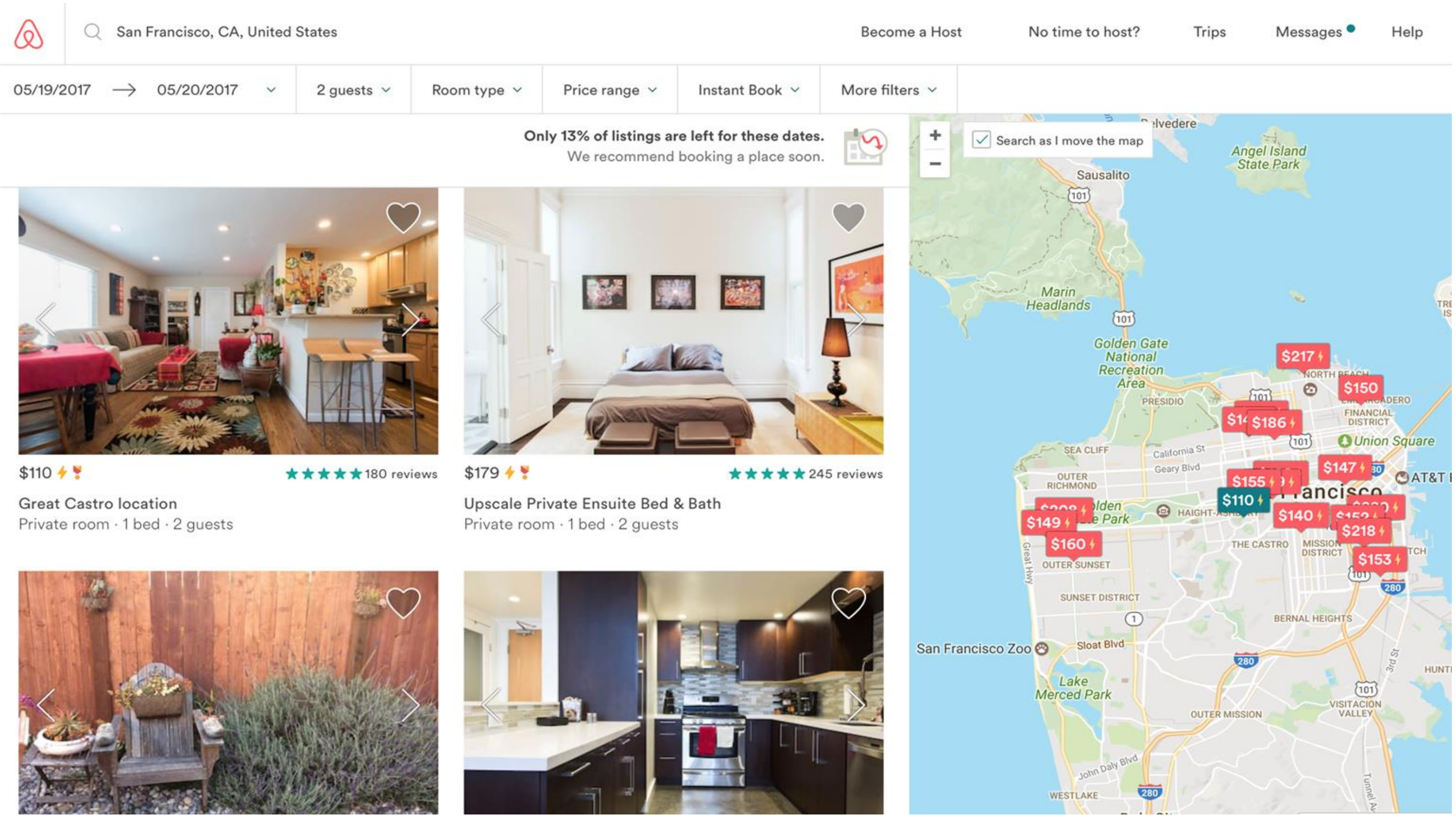

Become a Host
Help
Sign Up
Log In
Airbnb Book unique homes and experience a city like a local.
Where
Destination, city, address
When
Check In
Check Out
Guests
1 guest
Search
Just booked
JUAN
Timeless Classics

STATISTICS

- 100 Million Users (7/2016)
- 0.64 Million Hosts (11/2014)
- 2.3 Million Listing (7/2016)
- 191 Countries (6/2016)
- **San Francisco**: 25.1 million visitors in 2016
- **San Francisco Airbnb guest**: visits for 5.5 days and spends $1,045.

Source: http://expandedramblings.com/index.php/airbnb-statistics/
http://www.sftravel.com
http://blog.airbnb.com/economic-impact-airbnb/#san-francisco

SAMPLE SEARCH



ROLE & PROBLEM

ROLE & PROBLEM:

- Role: Guest
- Problem: Predicting expected price range for a stay given preference.
- Location: San Francisco

8 DATA

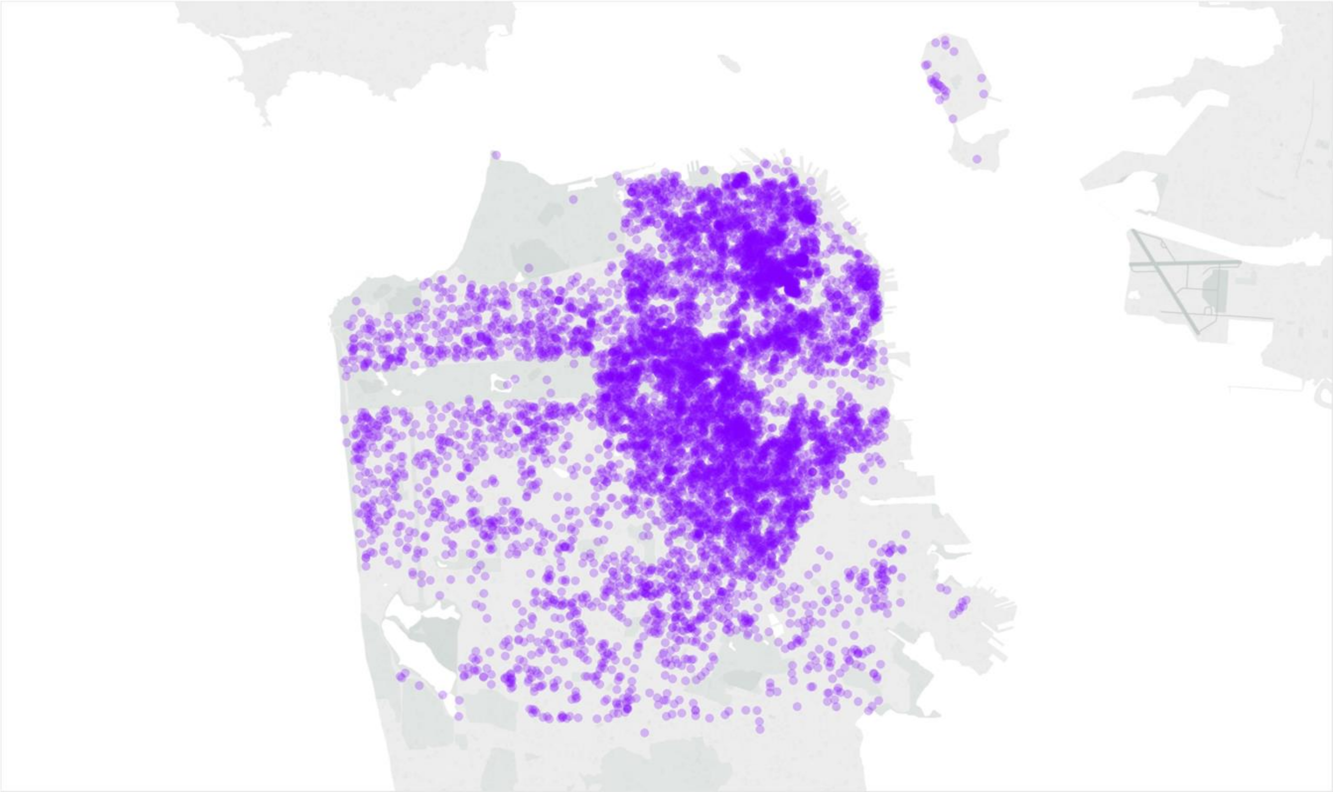
DATA OVERVIEW

- "listing.csv": 8720 listings and 95 variables.
- Target Variable: "*price*"
- Intuitively, 33 predicting variables are selected.
  - **Nominal:** "*id*", "*host_id*", "*host_response_time*", "*property_type*", "*room_type*", "*bed_type*", "*cancellation_policy*"
  - **Binary:** "*host_is_superhost*", "*host_has_profile_pic*", "*host_identity_verified*" "*is_location_exact*", "*instant_bookable*"
  - **Continuous:** "*bathrooms*", "*bedrooms*", "*beds*", "*host_response_rate*", "*host_listings_count*", "*latitude*", "*longitude*", "*security_deposit*", "*guests_included*", "*extra_people*", "*minimum_nights*", "*availability_365*", "*review_scores_rating*", "*review_scores_accuracy*", "*review_scores_cleanliness*", "*review_scores_checkin*", "*review_scores_communication*", "*review_scores_location*", "*review_scores_value*", "*calculated_host_listings_count*", "*reviews_per_month*"

PRE-PROCESSING

- Remove 11 empty entries.
- Replace NA:
  - With Mean: “*host_response_rate*”, 7 review score related variables
  - With Zero: "*bathrooms*", "*bedrooms*", "*beds*", "*reviews_per_month*", "*security_deposit*"
  - With a New Category: “*host_response_time*”

VISUALIZATION OF LISTING LOCATIONS

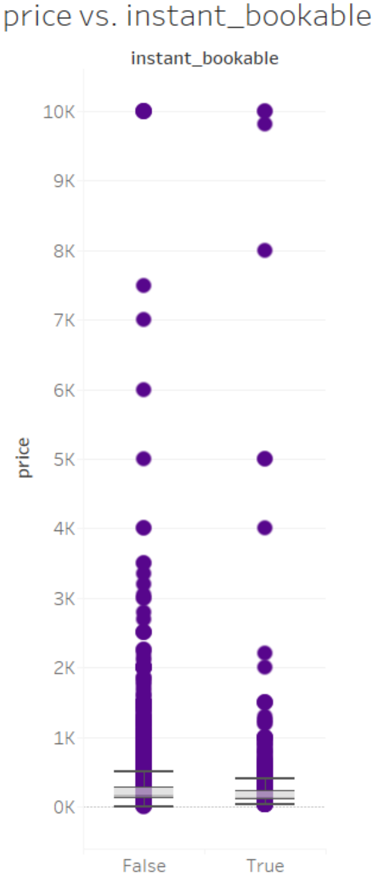


BIVARIATE ANALYSIS

ABOUT BOOKING

- *"instant_bookable"*: whether the listing can be booked immediately without the approval of hosts.
- *"cancellation_policy"*: what will happen if guests want to cancel the booking
- *"security_deposit"*: deposit required to secure the stay (e.g., property damage)
- Intuition: listings with higher price may have higher quality and more regulations: not instantly bookable, strict cancellation policy or greater security deposit.

ABOUT BOOKING (CONT.)

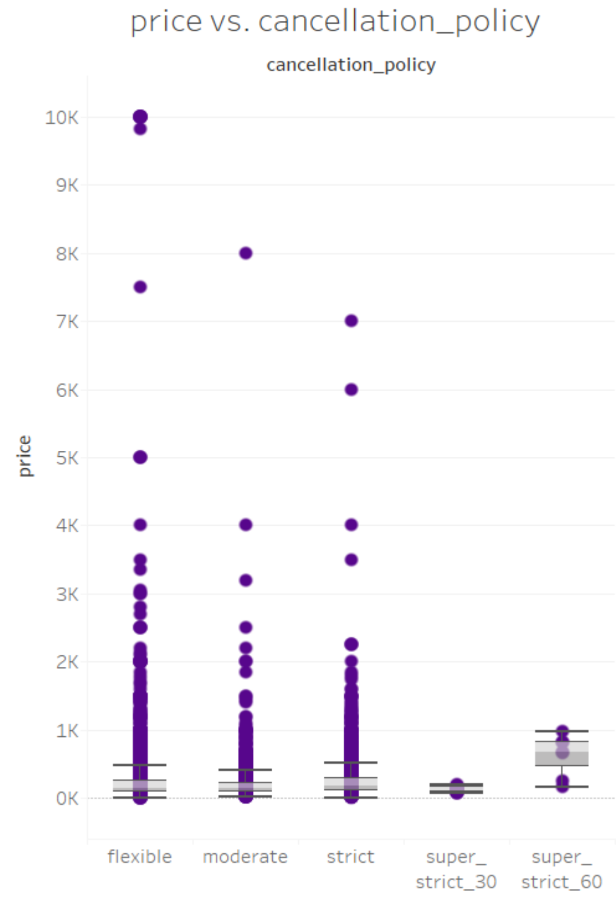


INSTANT BOOKABLE?

```
> t.test(airbnb$price ~ airbnb$instant_bookable)

	Welch Two Sample t-test

data:  airbnb$price by airbnb$instant_bookable
t = -2.4985, df = 2631.1, p-value = 0.01253
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.217682  -2.196351
sample estimates:
mean in group f mean in group t
       219.8149        230.0220
```

CANCELLATION POLICY

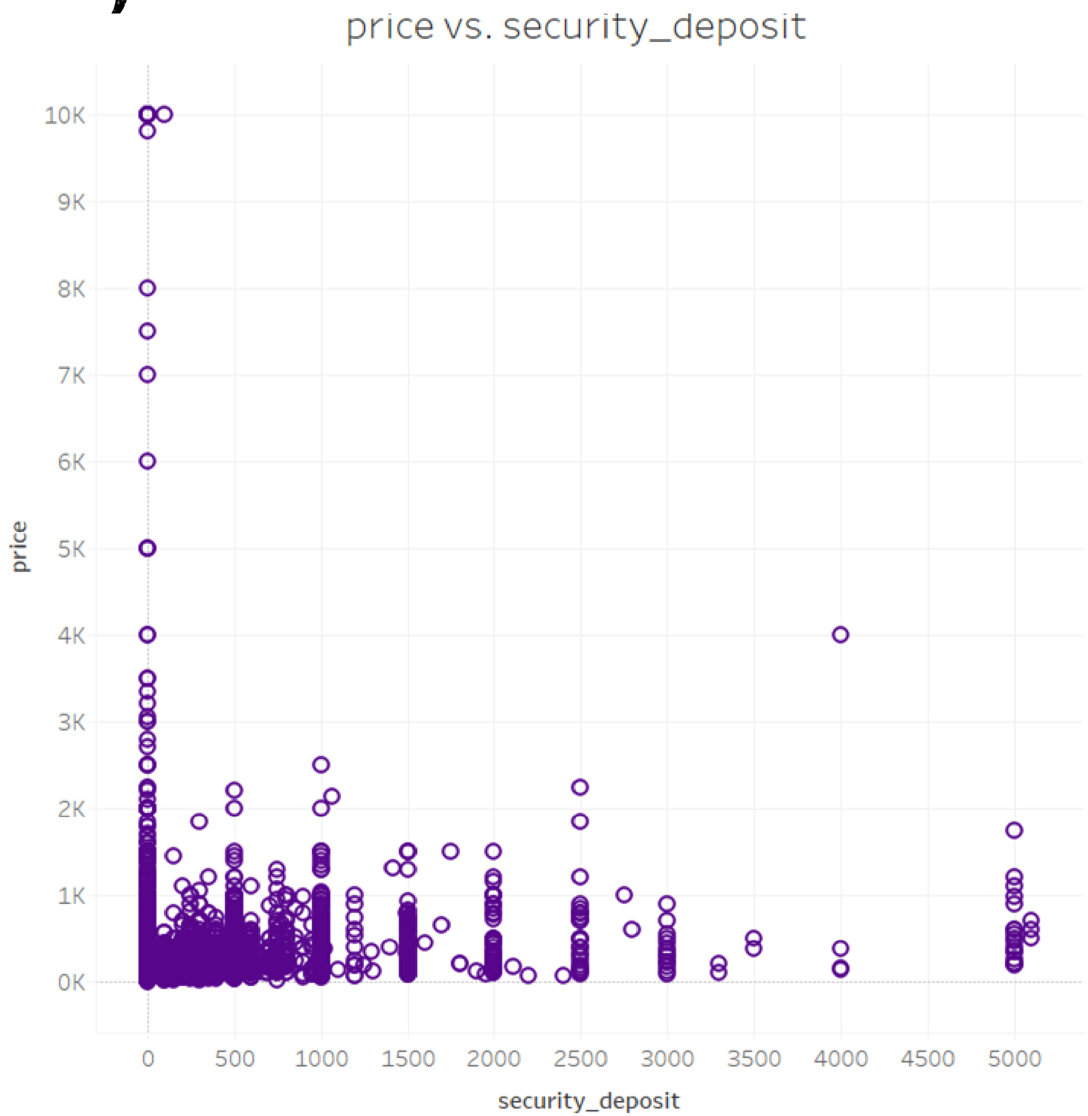
```
> summary(aov(airbnb$price ~ airbnb$cancellation_policy))
                             Df    Sum Sq Mean Sq F value   Pr(>F)
airbnb$cancellation_policy    4 1.198e+07 2994529    14.6 6.94e-12
Residuals                  8704 1.786e+09  205156
```



ABOUT BOOKING (CONT.)

SECURITY DEPOSIT

```
> summary(lm(airbnb$price ~ airbnb$security_deposit))
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              227.0489     5.4044  42.012   <2e-16
airbnb$security_deposit    0.1039     0.0110   9.443   <2e-16

> cor(airbnb$price, airbnb$security_deposit)
[1] 0.100689
```



AVAILABILITY

- *"availability_365"*: how many days in a year will the listing ready to be booked
- Intuition: listings with low availability (either booked or occupied) reflect a high demand and is expected to have higher price.

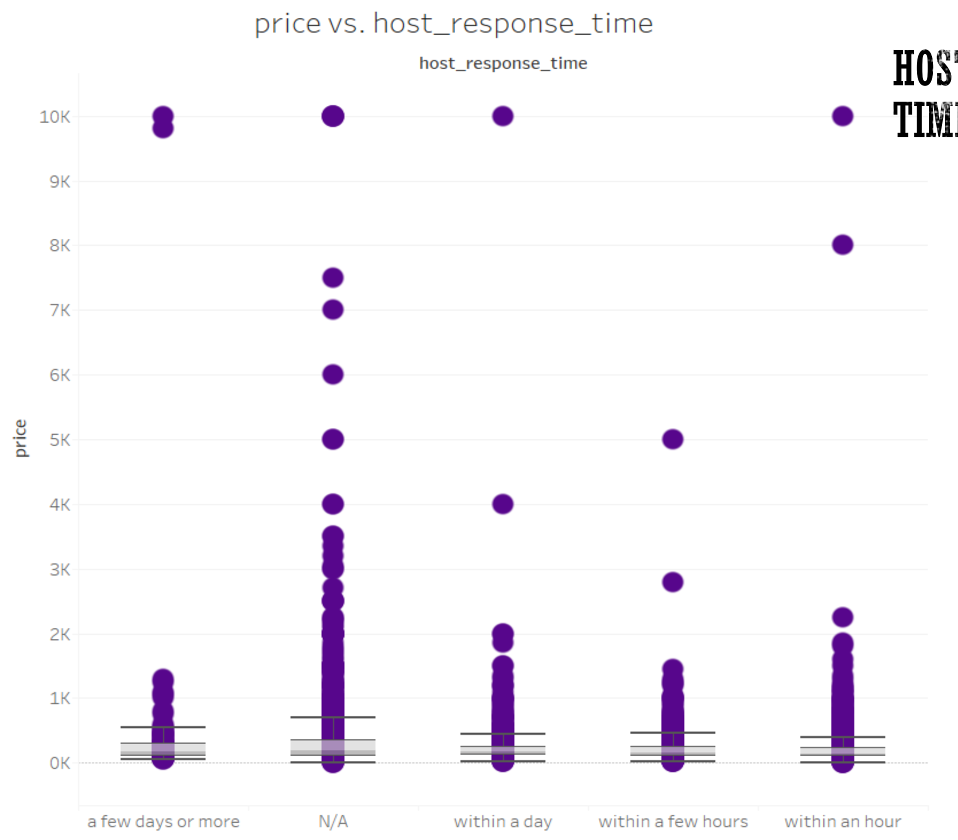

```
> summary(lm(airbnb$price ~ airbnb$availability_365))
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              266.45930    6.63180  40.179  < 2e-16
airbnb$availability_365   -0.13038    0.03501  -3.724 0.000197
> cor(airbnb$price, airbnb$availability_365)
[1] -0.03987573
```

ABOUT HOSTS

- *"host_response_time"*: how long will the host reply the message.
- *"host_is_superhost"*: whether the host is qualified as a superhost*.
- *"host_identity_verified"*: whether the identity of hosts is verified.
- *"calculated_host_listing_count"*: the number of listings of a host.
- Intuition: well-qualified hosts are expected to provide great service which may further affect the price, for example, a super-host is expected to know the market better and thus the price should be more reasonable.

***Superhosts:** complete at least 10 trips in their listings in a year;
respond to guests quickly and maintain a 90% response rate or higher;
provide listings that inspire enthusiastic reviews (At least 80% of their reviews need to be 5 stars);
rarely cancel the reservation.

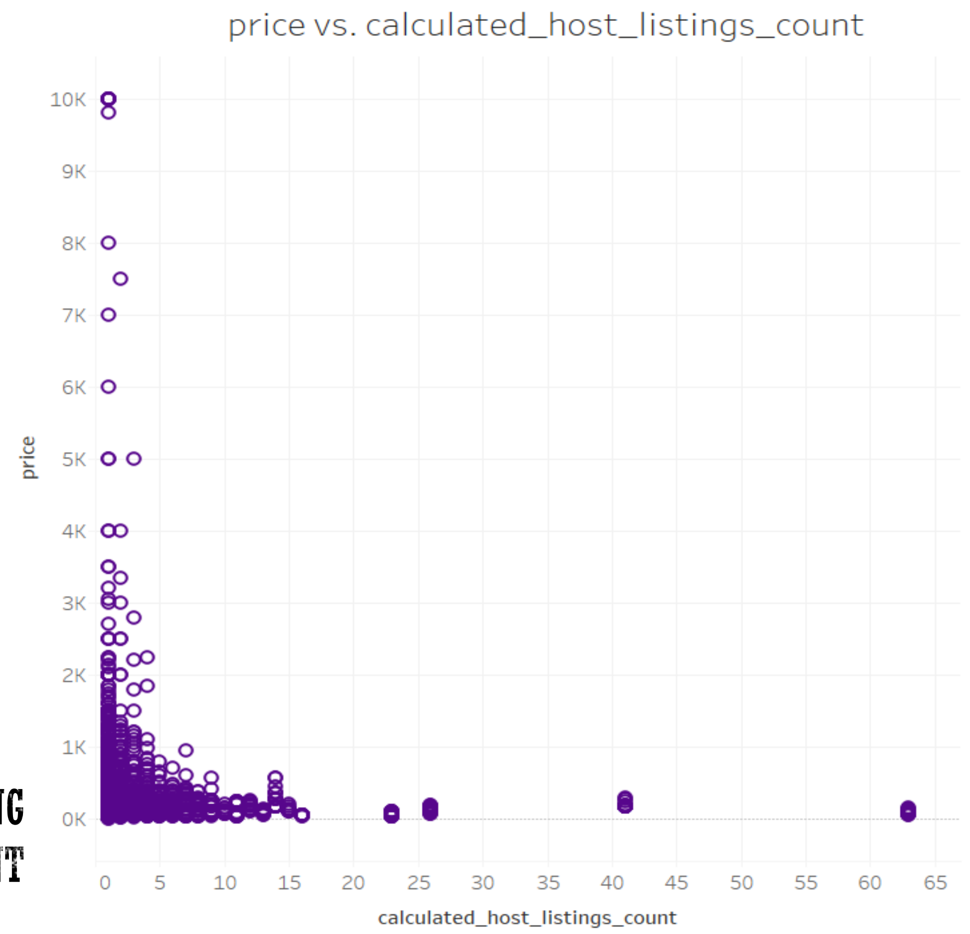
ABOUT HOSTS (CONT.)



HOST RESPONSE TIME

```
> summary(aov(airbnb$price ~ airbnb$host_response_time))
                            Df    Sum Sq Mean Sq F value Pr(>F)
airbnb$host_response_time    4 3.919e+07 9797714    48.5 <2e-16
Residuals                 8704 1.758e+09  202030
```

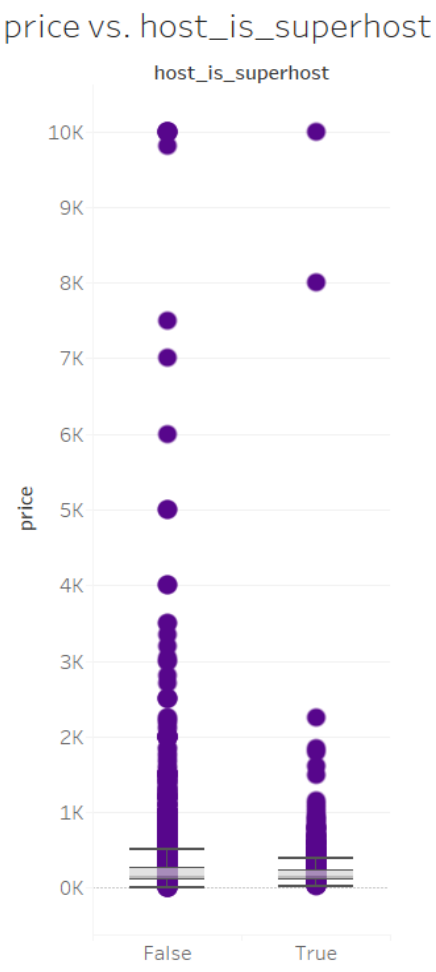
Bivariate Analysis



HOST LISTING COUNT

```
> summary(lm(airbnb$price ~ airbnb$calculated_host_listings_count))
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           259.7076     5.2356  49.604  < 2e-16
airbnb$calculated_host_listings_count  -3.9510     0.7655  -5.161 2.51e-07
> cor(airbnb$price, airbnb$calculated_host_listings_count)
[1] -0.05522791
```

ABOUT HOSTS (CONT.)

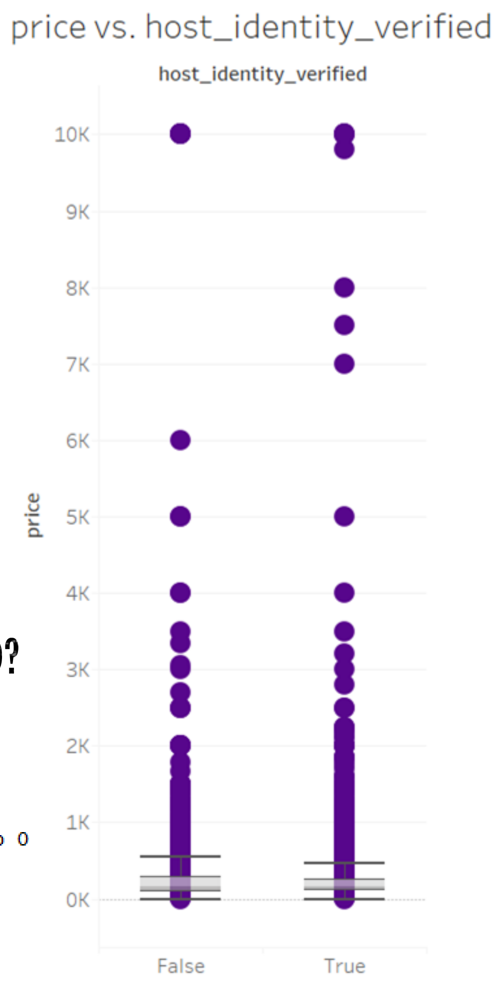


SUPERHOST?

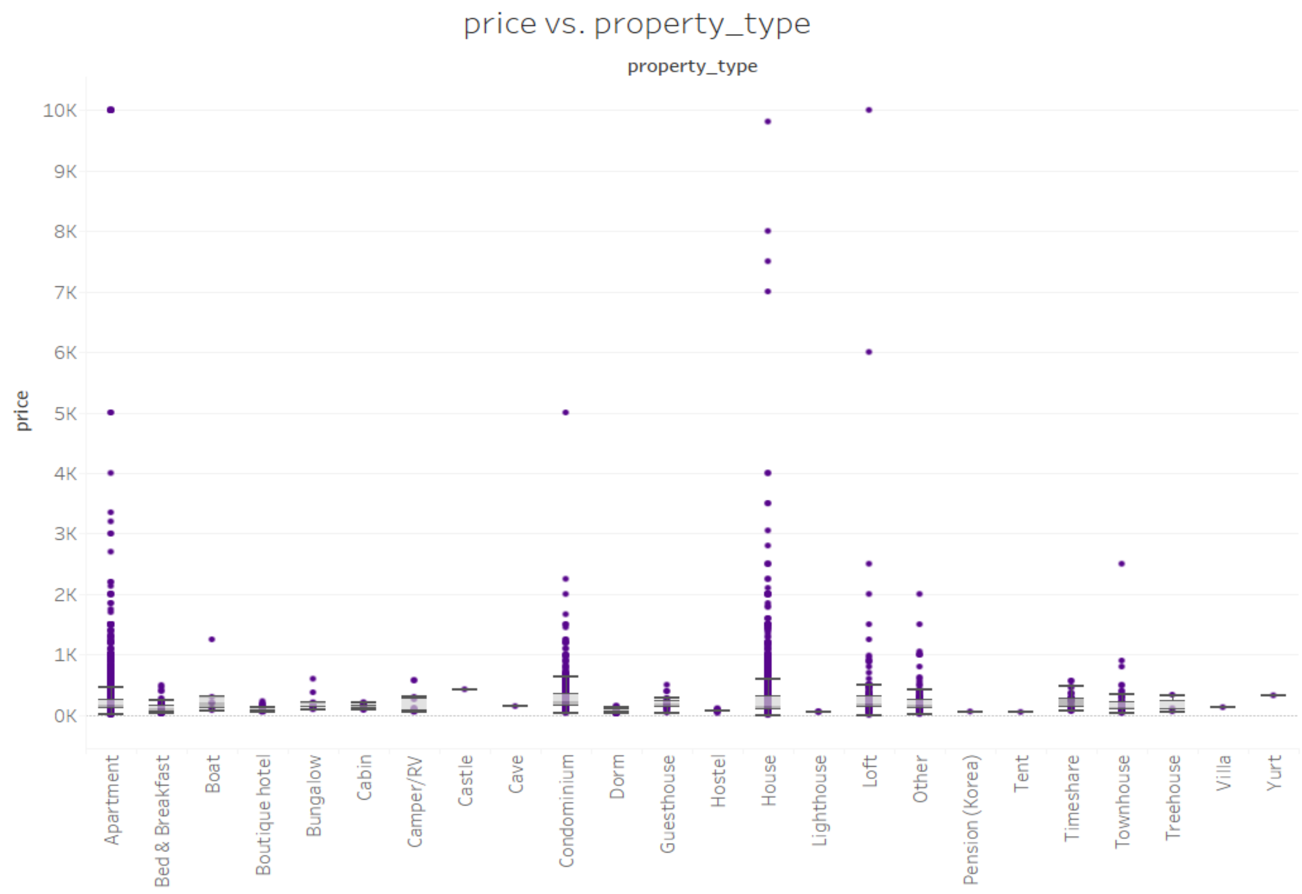
```
> t.test(airbnb$price ~ airbnb$host_is_superhost)

        Welch Two Sample t-test

data:  airbnb$price by airbnb$host_is_superhost
t = 3.9101, df = 2845.4, p-value = 9.443e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 20.85235 62.80382
sample estimates:
mean in group f mean in group t
       257.1402        215.3121
```

IDENTITY VERIFIED?

```
> t.test(airbnb$price ~ airbnb$host_identity_verified)

        Welch Two Sample t-test

data:  airbnb$price by airbnb$host_identity_verified
t = 3.2689, df = 3014.1, p-value = 0.001092
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 17.39904 69.55510
sample estimates:
mean in group f mean in group t
       281.8712        238.3942
```



PROPERTY TYPE



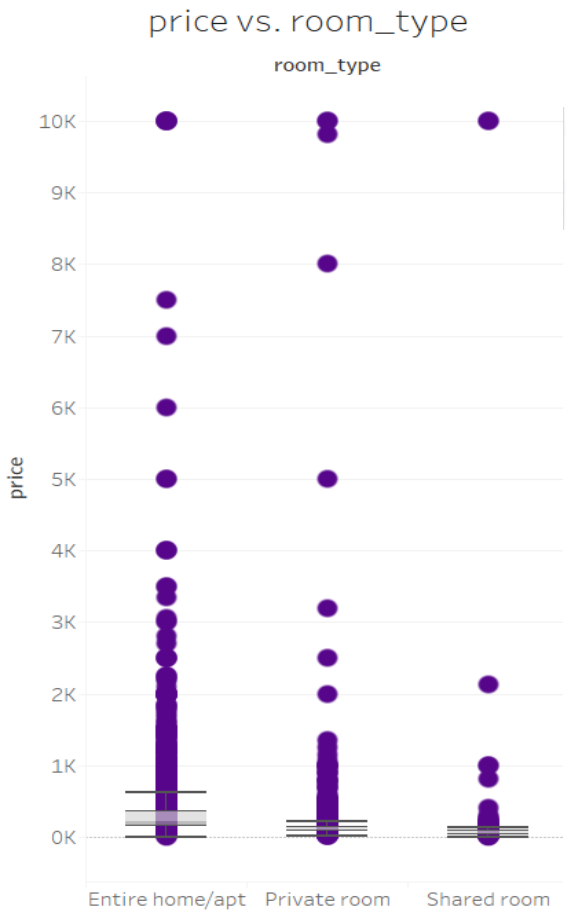
- *"property_type"*: type of the listing
- Intuition: different types can affect the price, e.g., castle is more expensive than tent.

```
> summary(aov(airbnb$price ~ airbnb$property_type))
                       Df    Sum Sq Mean Sq F value   Pr(>F)
airbnb$property_type   23 1.185e+07  515253   2.506 8.67e-05
Residuals            8685 1.786e+09  205620
---
```

ROOM PROPERTIES

- *"room_type"*: type of rooms
- *"bathrooms"*: number of bathrooms in the listing
- *"bedrooms"*: number of bedrooms in the listing
- *"beds"*: number of beds in the listing
- *"bed_type"*: type of beds in the listing
- *"guests_included"*: number of guests the listing supposes to hold
- Intuition: all above variables reflect the listing size and condition, which can affect the price. The bigger and better-condition listings are tended to have higher prices than the smaller ones.

ROOM PROPERTIES (CONT.)

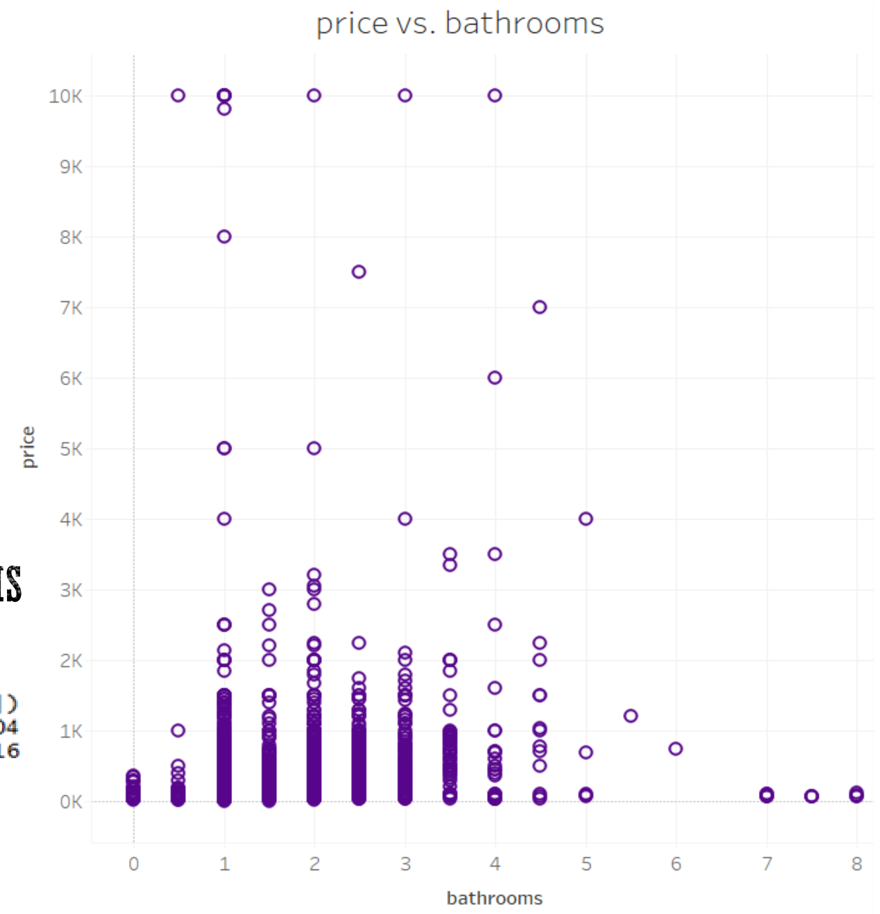


ROOM TYPE

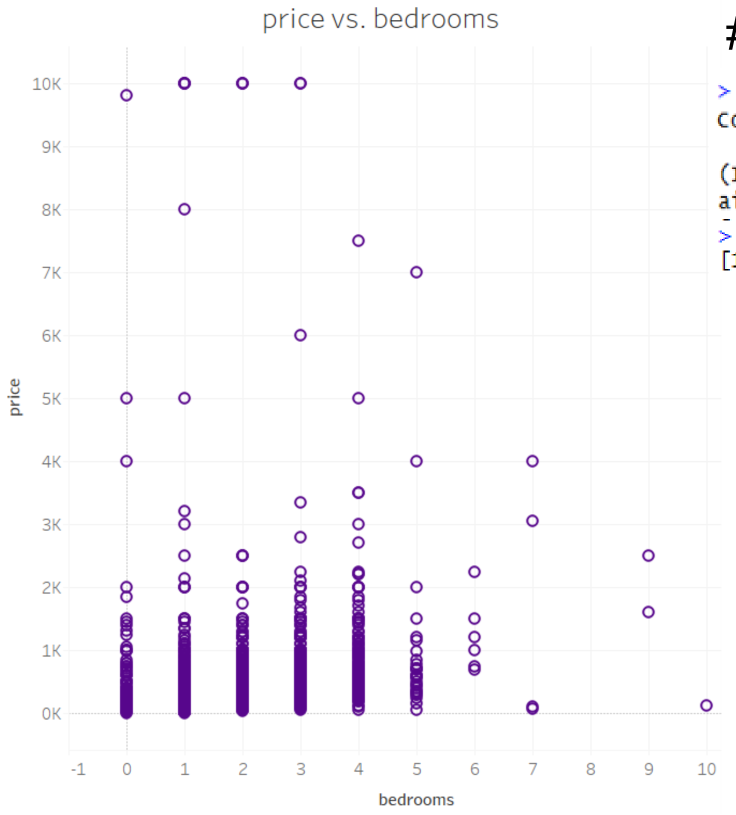
```
> summary(aov(airbnb$price ~ airbnb$room_type, data=air
                   Df    Sum Sq Mean Sq F value Pr(>F)
airbnb$room_type    2   7383314 3691657   175.4 <2e-16
Residuals        8706 183237124   21047
```

OF BATHROOMS

```
> summary(lm(airbnb$price ~ airbnb$bathrooms))
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        38.298     10.815   3.541    4e-04
airbnb$bathrooms  167.583      7.706  21.747   <2e-16
> cor(airbnb$price, airbnb$bathrooms)
[1] 0.2269761
```



ROOM PROPERTIES (CONT.)

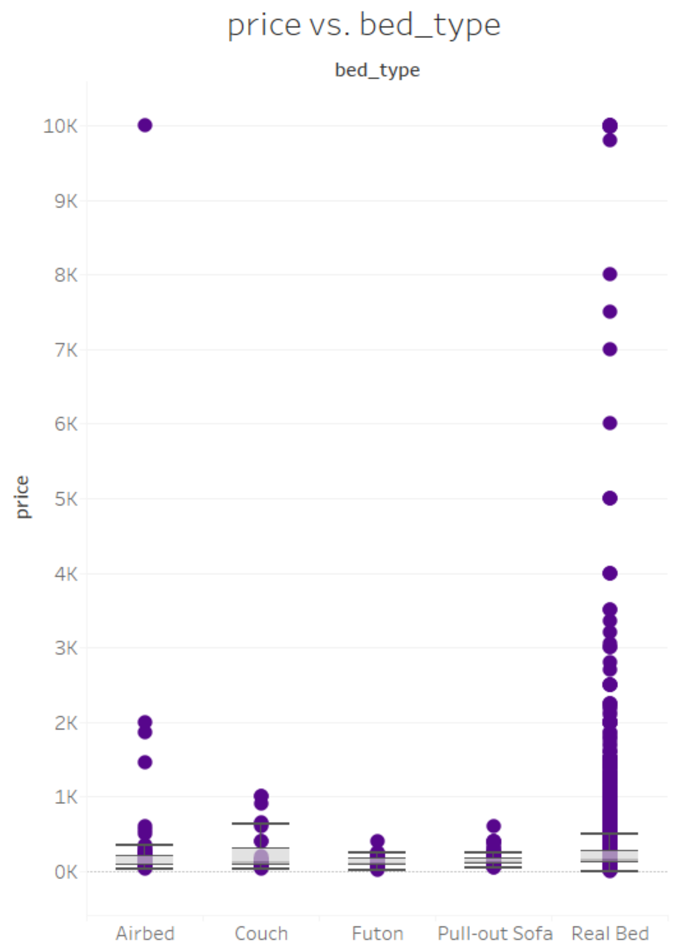


OF BEDROOMS

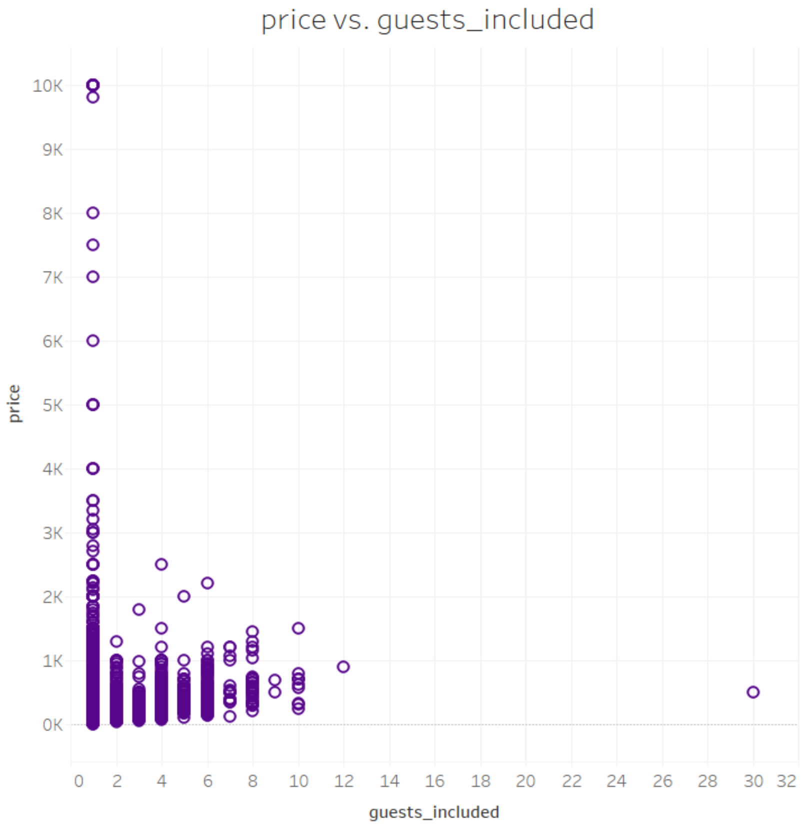
```
> summary(lm(airbnb$price ~ airbnb$bedrooms))
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       46.201      8.377   5.515 3.59e-08
airbnb$bedrooms  152.018      5.208  29.191  < 2e-16
> cor(airbnb$price , airbnb$bedrooms)
[1] 0.2985625
```

BED TYPES

```
> summary(aov(airbnb$price ~ airbnb$bed_type))
                  Df    Sum Sq Mean Sq F value  Pr(>F)
airbnb$bed_type    4 3.177e+06  794249   3.852 0.00394
Residuals       8704 1.794e+09  206168
```

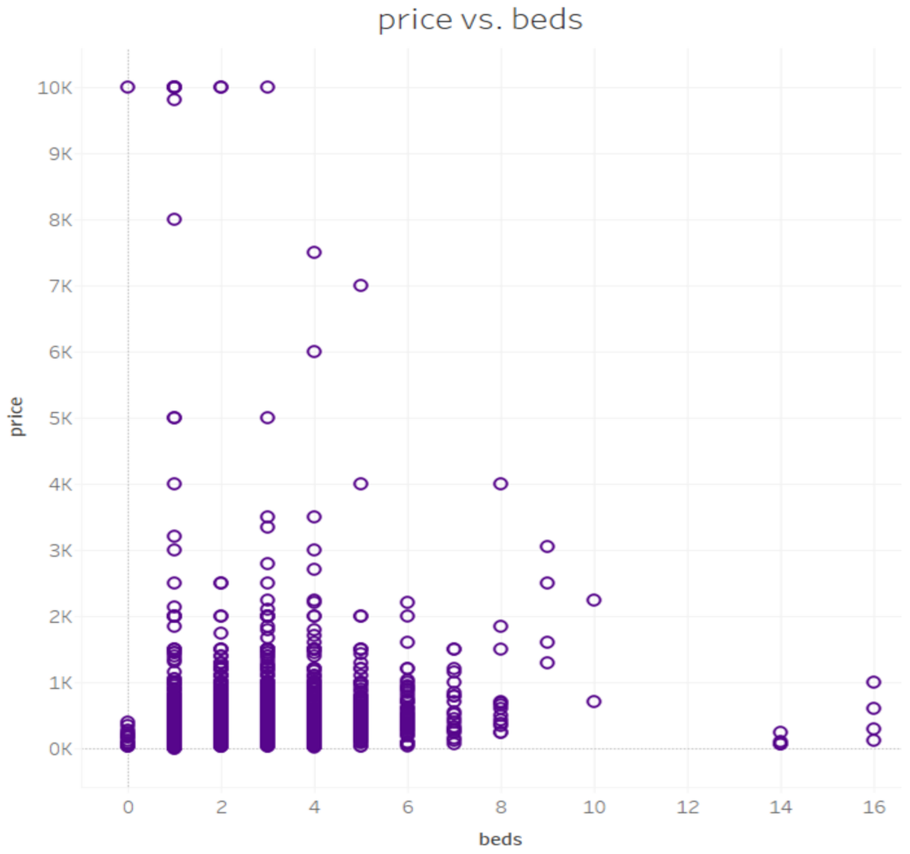


ROOM PROPERTIES (CONT.)



GUESTS INCLUDED

```
> summary(lm(airbnb$price ~ airbnb$guests_included))
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             204.711      7.835  26.127  < 2e-16
airbnb$guests_included   28.189      3.856   7.311 2.89e-13
> cor(airbnb$price, airbnb$guests_included)
[1] 0.07811042
```



OF BEDS

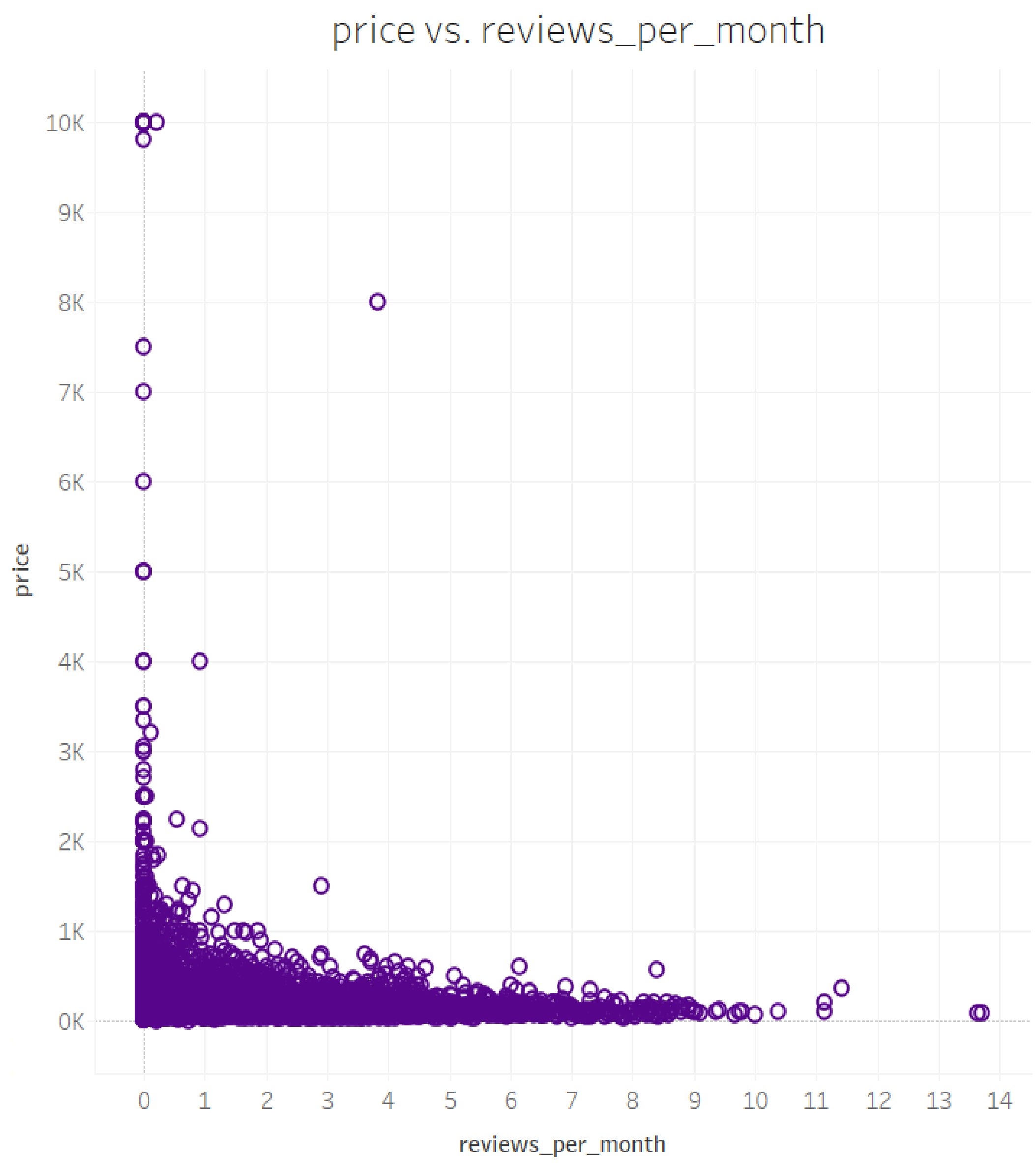
```
> summary(lm(airbnb$price ~ airbnb$beds))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   92.656      8.313   11.14   <2e-16
airbnb$beds   92.019      4.007   22.96   <2e-16
> cor(airbnb$price, airbnb$beds)
[1] 0.2389547
```

REVIEWS/MONTH

- *"reviews_per_month"*: number of reviews received per month

- Intuition: a reasonably priced listing is expected to be more popular and thus to receive more reviews per month.

```
> summary(lm(airbnb$price ~ airbnb$reviews_per_month))
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               294.787      6.056   48.68   <2e-16
airbnb$reviews_per_month  -35.534      2.881  -12.33   <2e-16
> cor(airbnb$price, airbnb$calculated_host_listings_count)
[1] -0.05522791
```

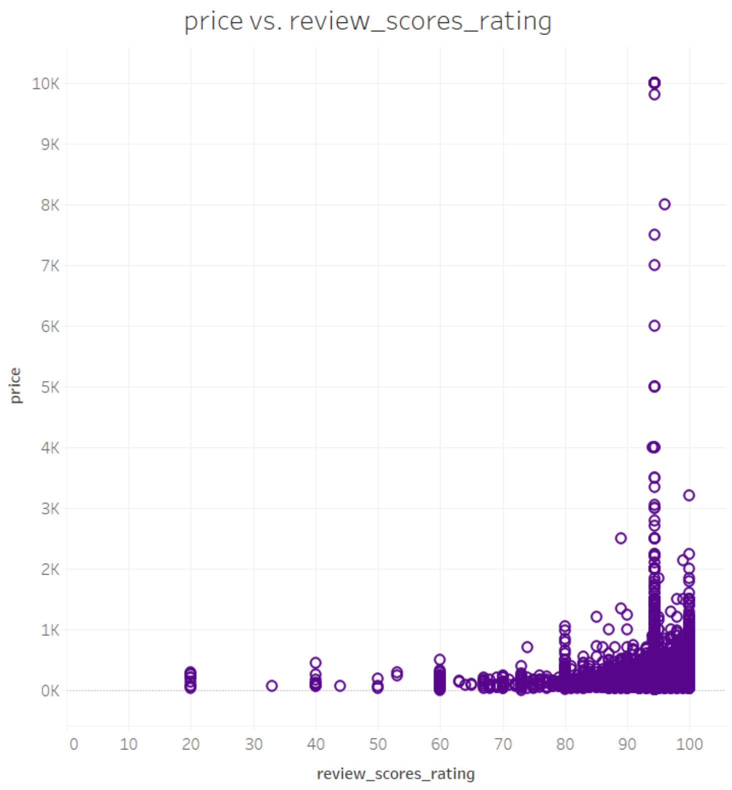


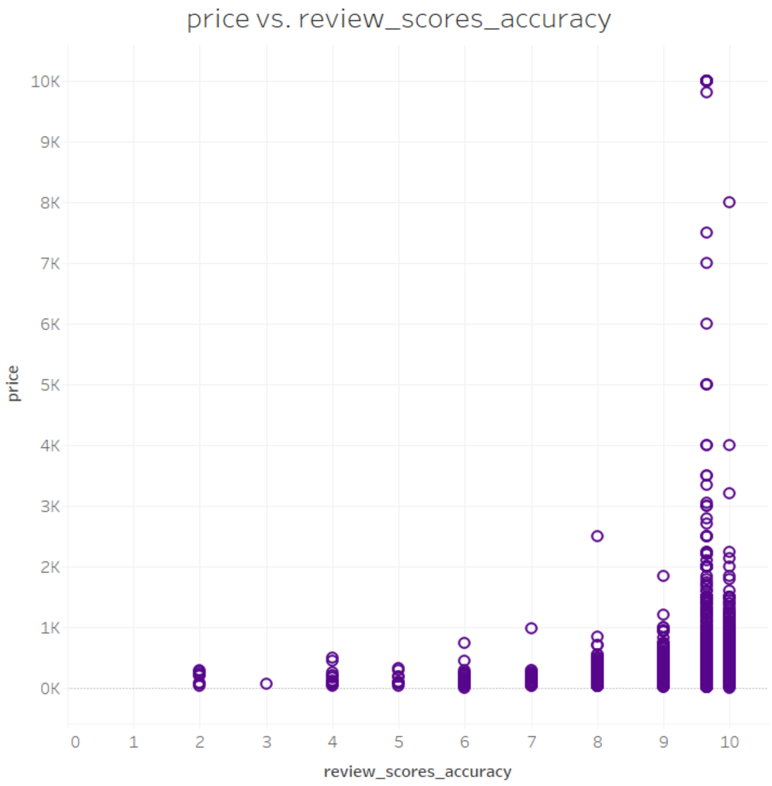
REVIEW SCORES

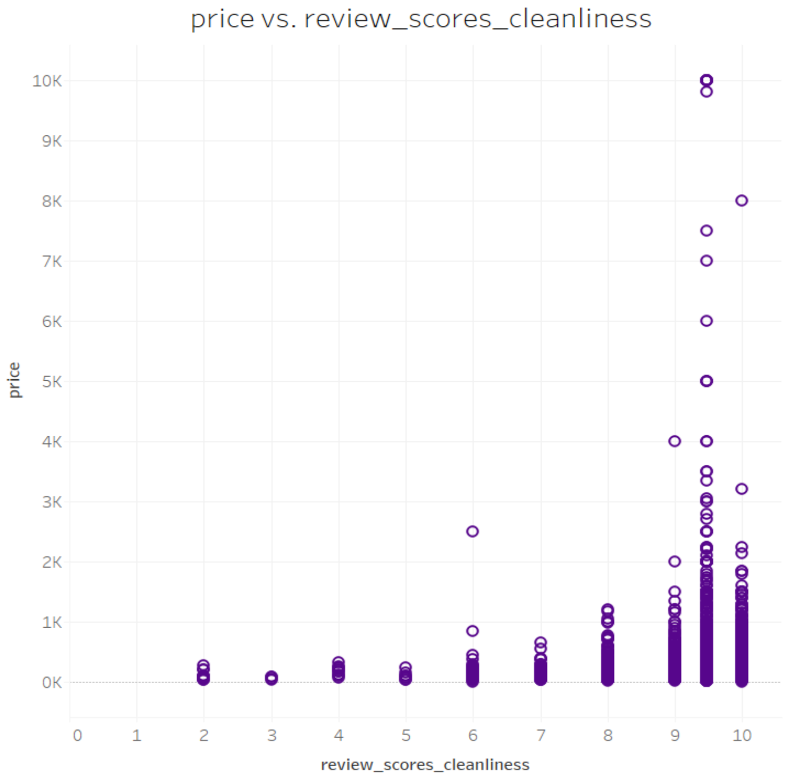
Variable Name	Description	p-value of Linear Model	Correlation
"review_scores_rating"	overall rating of stay	1.797e-05	0.04593894
"review_scores_accuracy"	the accuracy of listing description	0.006464	0.02917963
"review_scores_cleanliness"	the cleanliness of listing	0.0005223	0.03716541
"review_scores_checkin"	the check-in process	0.04402	0.0215808
"review_scores_communication"	the communication with hosts	0.02993	0.02326415
"review_scores_location"	the location of listing	0.0008218	0.03584101

- Intuition: a listing with higher review is expected to have higher price as it reflects the good quality and service.

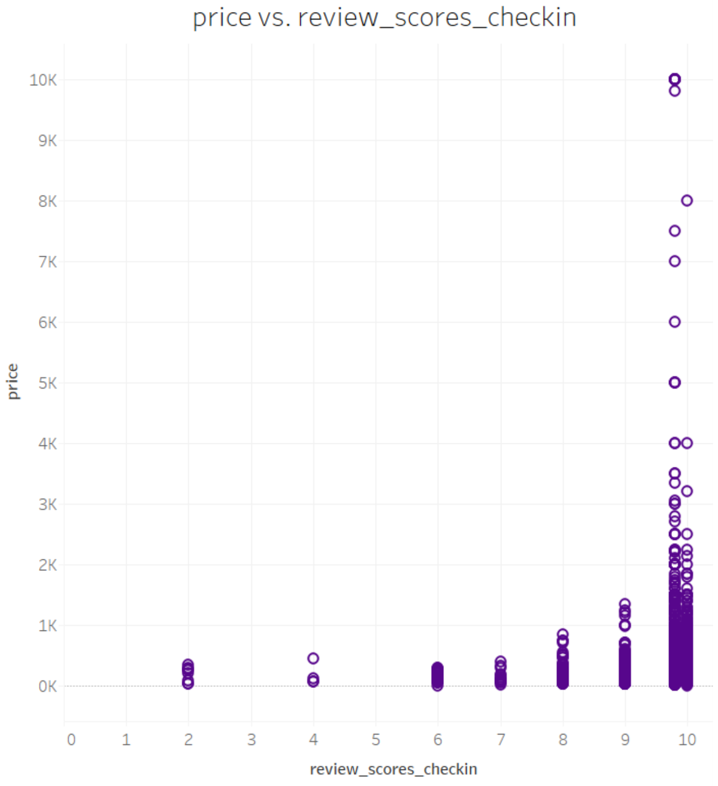
REVIEW SCORES (CONT.)

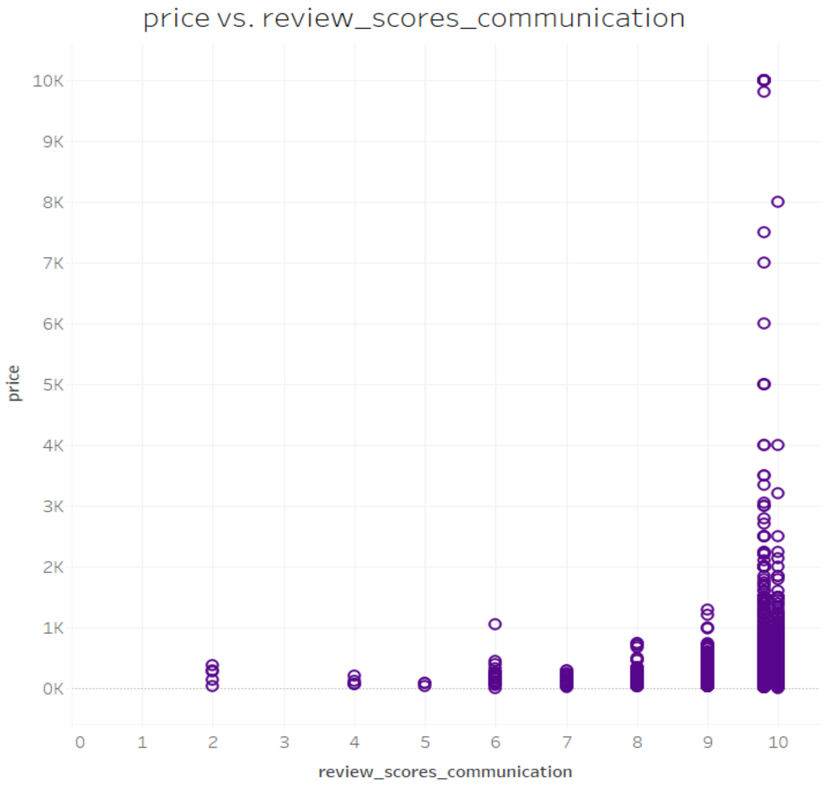


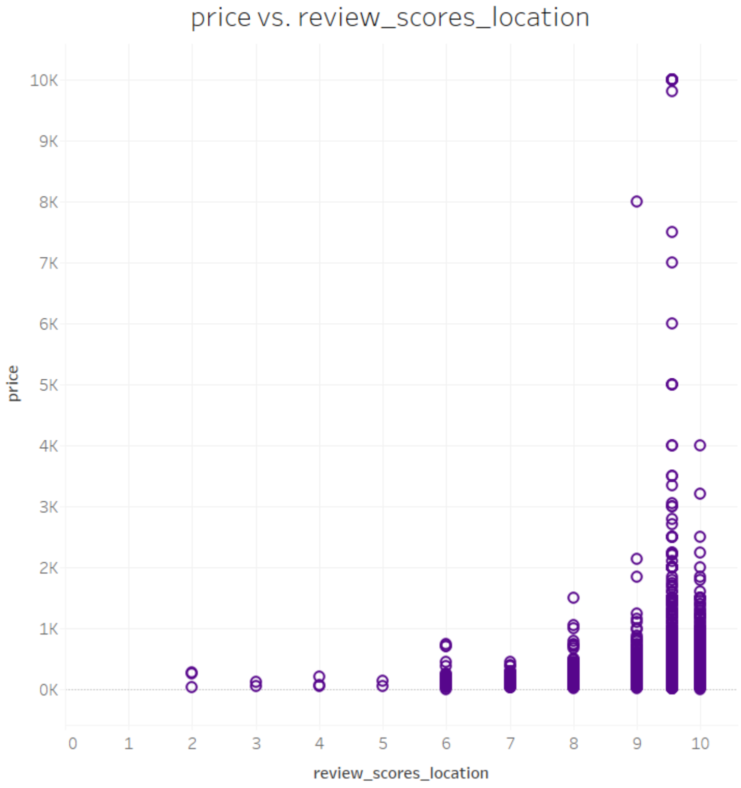




REVIEW SCORES (CONT.)







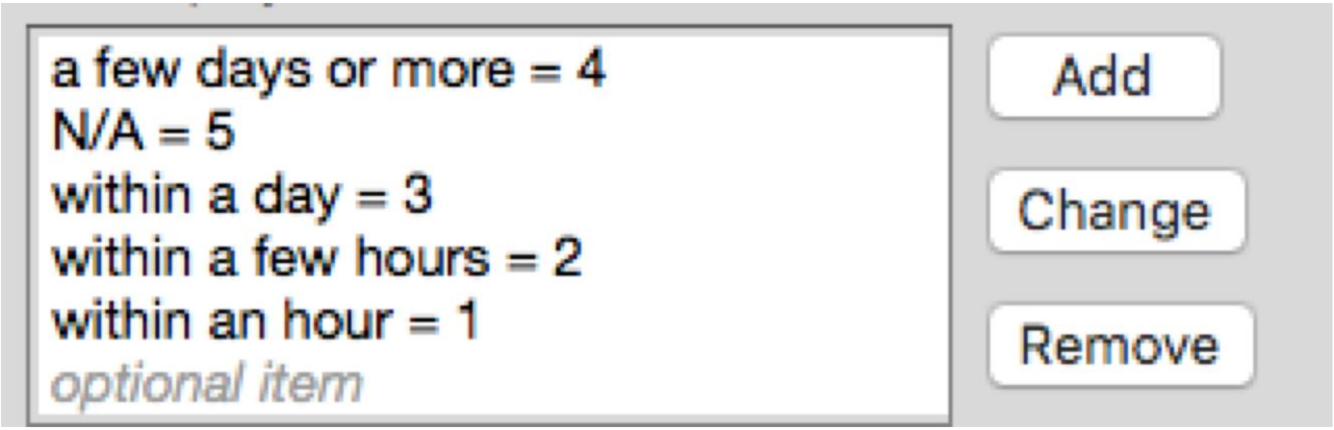
POTENTIAL PREDICTORS FOR *"PRICE"*

1. *room_type*
2. *instant_bookable*
3. *host_response_time*
4. *host_is_superhost*
5. *host_identity_verified*
6. *property_type*
7. *bathrooms*
8. *bedrooms*
9. *beds*
10. *bed_type*
11. *security_deposit*
12. *guests_included*
13. *availability_365*
14. *review_scores_rating*
15. *review_scores_accuracy*
16. *review_scores_cleanliness*
17. *review_scores_checkin*
18. *review_scores_communication*
19. *review_scores_location*
20. *cancellation_policy*
21. *calculated_host_listings_count*
22. *reviews_per_month*

REGRESSION

VALUE LABELS

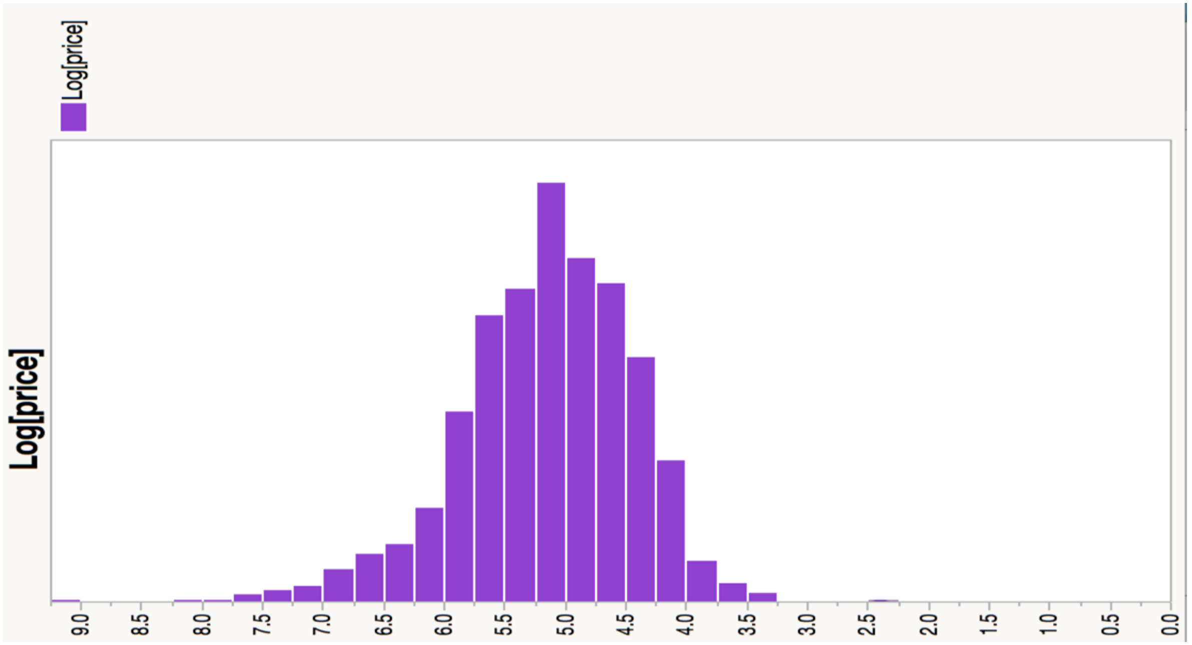
- For example: host_response_time



TARGET VARIABLE

Price

Logged Price



LINEAR VS SEMI-LOG MODEL

Response price

Whole Model

Summary of Fit

RSquare	0.143416
RSquare Adj	0.13837
Root Mean Square Error	421.7498
Mean of Response	249.6765
Observations (or Sum Wgts)	8709

AICc	BIC
130051.2	130425.3

VS

Response Log[price]

Summary of Fit

RSquare	0.539081
RSquare Adj	0.536366
Root Mean Square Error	0.503146
Mean of Response	5.159059
Observations (or Sum Wgts)	8708

AICc	BIC
12804.13	13178.29

MODEL IMPROVEMENT

- Look at Indicator function parametrization, almost all of the "*property_type*" have high p-value:

property_type[Apartment]	0.1145028	0.057641	1.99	0.0470*
property_type[Bed & Breakfast]	0.0648308	0.078476	0.83	0.4088
property_type[Boat]	0.2612633	0.205255	1.27	0.2031
property_type[Boutique hotel]	0.119454	0.102827	1.16	0.2454
property_type[Bungalow]	0.2835966	0.156331	1.81	0.0697
property_type[Cabin]	0.105946	0.170788	0.62	0.5351
property_type[Camper/RV]	-0.282659	0.146026	-1.94	0.0529
property_type[Castle]	0.5522486	0.485688	1.14	0.2556
property_type[Cave]	0.4713284	0.485598	0.97	0.3318
property_type[Condominium]	0.3006949	0.060937	4.93	<.0001*
property_type[Dorm]	-0.158258	0.087766	-1.80	0.0714
property_type[Guesthouse]	0.1386084	0.119821	1.16	0.2474
property_type[Hostel]	-0.370354	0.168353	-2.20	0.0278*
property_type[House]	0.064158	0.058217	1.10	0.2705
property_type[Lighthouse]	-0.738206	0.28506	-2.59	0.0096*
property_type[Loft]	0.3815994	0.070506	5.41	<.0001*
property_type[Other]	0.2812282	0.071905	3.91	<.0001*
property_type[Pension (Korea)]	-0.501882	0.486896	-1.03	0.3027
property_type[Tent]	-1.332774	0.485795	-2.74	0.0061*
property_type[Timeshare]	0.2807826	0.101128	2.78	0.0055*
property_type[Townhouse]	0.0330366	0.084737	0.39	0.6966
property_type[Treehouse]	-0.505936	0.284135	-1.78	0.0750
property_type[Villa]	0.0325353	0.485541	0.07	0.9466

NOMINAL VARIABLES

- Some categorical variables are too many levels, which will bias this predictor's coefficients up.

property_type[Apartment]	0.1145028	0.057641	1.99	0.0470*
property_type[Bed & Breakfast]	0.0648308	0.078476	0.83	0.4088
property_type[Boat]	0.2612633	0.205255	1.27	0.2031
property_type[Boutique hotel]	0.119454	0.102827	1.16	0.2454
property_type[Bungalow]	0.2835966	0.156331	1.81	0.0697
property_type[Cabin]	0.105946	0.170788	0.62	0.5351
property_type[Camper/RV]	-0.282659	0.146026	-1.94	0.0529
property_type[Castle]	0.5522486	0.485688	1.14	0.2556
property_type[Cave]	0.4713284	0.485598	0.97	0.3318
property_type[Condominium]	0.3006949	0.060937	4.93	<.0001*
property_type[Dorm]	-0.158258	0.087766	-1.80	0.0714
property_type[Guesthouse]	0.1386084	0.119821	1.16	0.2474
property_type[Hostel]	-0.370354	0.168353	-2.20	0.0278*
property_type[House]	0.064158	0.058217	1.10	0.2705
property_type[Lighthouse]	-0.738206	0.28506	-2.59	0.0096*
property_type[Loft]	0.3815994	0.070506	5.41	<.0001*
property_type[Other]	0.2812282	0.071905	3.91	<.0001*
property_type[Pension (Korea)]	-0.501882	0.486896	-1.03	0.3027
property_type[Tent]	-1.332774	0.485795	-2.74	0.0061*
property_type[Timeshare]	0.2807826	0.101128	2.78	0.0055*
property_type[Townhouse]	0.0330366	0.084737	0.39	0.6966
property_type[Treehouse]	-0.505936	0.284135	-1.78	0.0750
property_type[Villa]	0.0325353	0.485541	0.07	0.9466

FEATURE DELETION

- Trade-off: fewer features might decrease model accuracy, but will also decrease model complexity.

Summary of Fit

RSquare	0.527881
RSquare Adj	0.526412
Root Mean Square Error	0.508518
Mean of Response	5.159059
Observations (or Sum Wgts)	8708

AICc	BIC
12964.75	13169.64

CONTINUOUS VARIABLES

- Some features have similar meanings, and they may also decrease model quality.
- e.g.,

review_scores_rating	0.0050408	0.001631	3.09	0.0020*
review_scores_accuracy	0.0063552	0.012709	0.50	0.6170
review_scores_cleanliness	0.0683882	0.010255	6.67	<.0001*
review_scores_checkin	-0.001803	0.013844	-0.13	0.8964
review_scores_communication	-0.010447	0.014582	-0.72	0.4738
review_scores_location	0.0963082	0.009648	9.98	<.0001*
review_scores_value	-0.074286	0.011735	-6.33	<.0001*

PRINCIPAL COMPONENT ANALYSIS

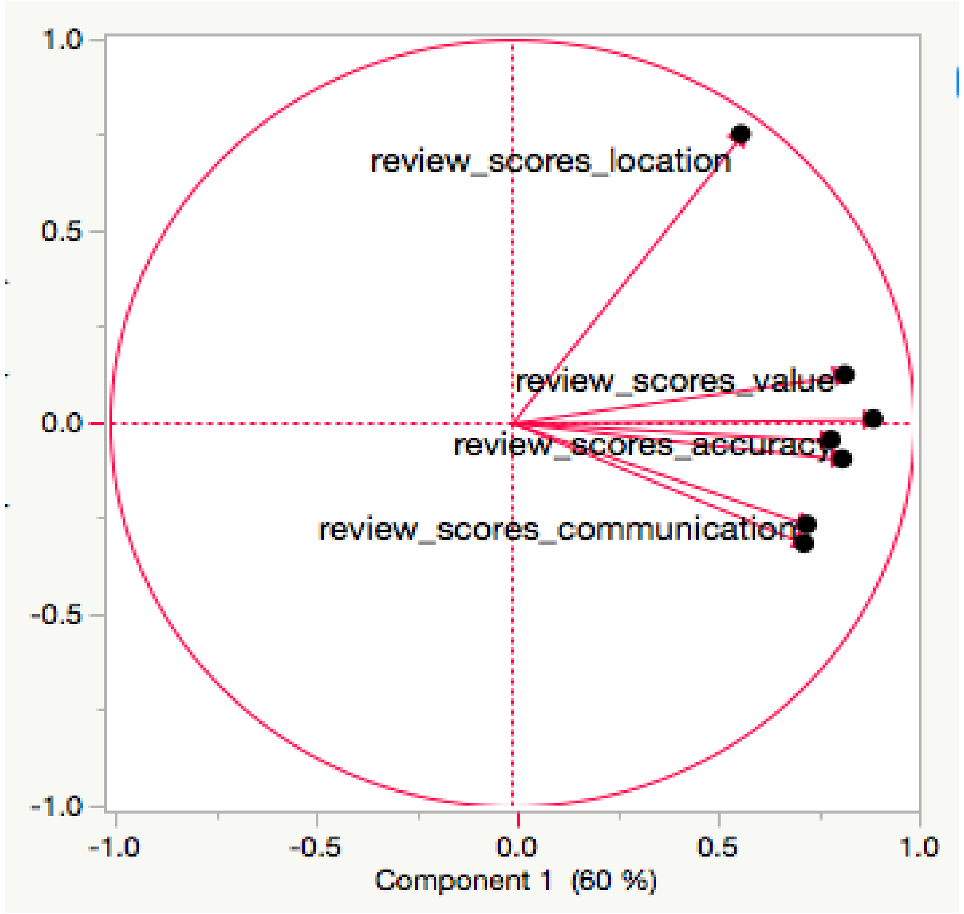
FACTOR ANALYSIS

- Several predictors have similar meanings.
- e.g., 7 “review_scores” , they complicate the model.
- We can do PCA to see if we will have a simplified model and a similar model predictive power.

PCA ON 7 REVIEW SCORES

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	4.2032	60.046		60.046
2	0.7630	10.900		70.946
3	0.6990	9.985		80.932
4	0.4033	5.762		86.693
5	0.3798	5.426		92.120
6	0.3444	4.921		97.040
7	0.2072	2.960		100.000



RULE OF THUMB: KEEP FACTOR THAT HAS AN E.V > 1

- The regression result is

Summary of Fit

RSquare	0.518577
RSquare Adj	0.517413
Root Mean Square Error	0.513327
Mean of Response	5.159059
Observations (or Sum Wgts)	8708

AICc	BIC
13122.61	13285.14

- We can see that R-square dropped and AIC/BIC increased, so we should keep some review scores that have low p-values.

BACK TO ORIGINAL SEMI-LOG MODEL

- We do see some review scores have high p-value. Let's see the regression result after dropping them:

review_scores_rating	1	1	3.69381	14.2873	0.0002*
review_scores_cleanliness	1	1	12.82526	49.6068	<.0001*
review_scores_location	1	1	29.12589	112.6560	<.0001*
review_scores_value	1	1	12.37603	47.8692	<.0001*

Summary of Fit

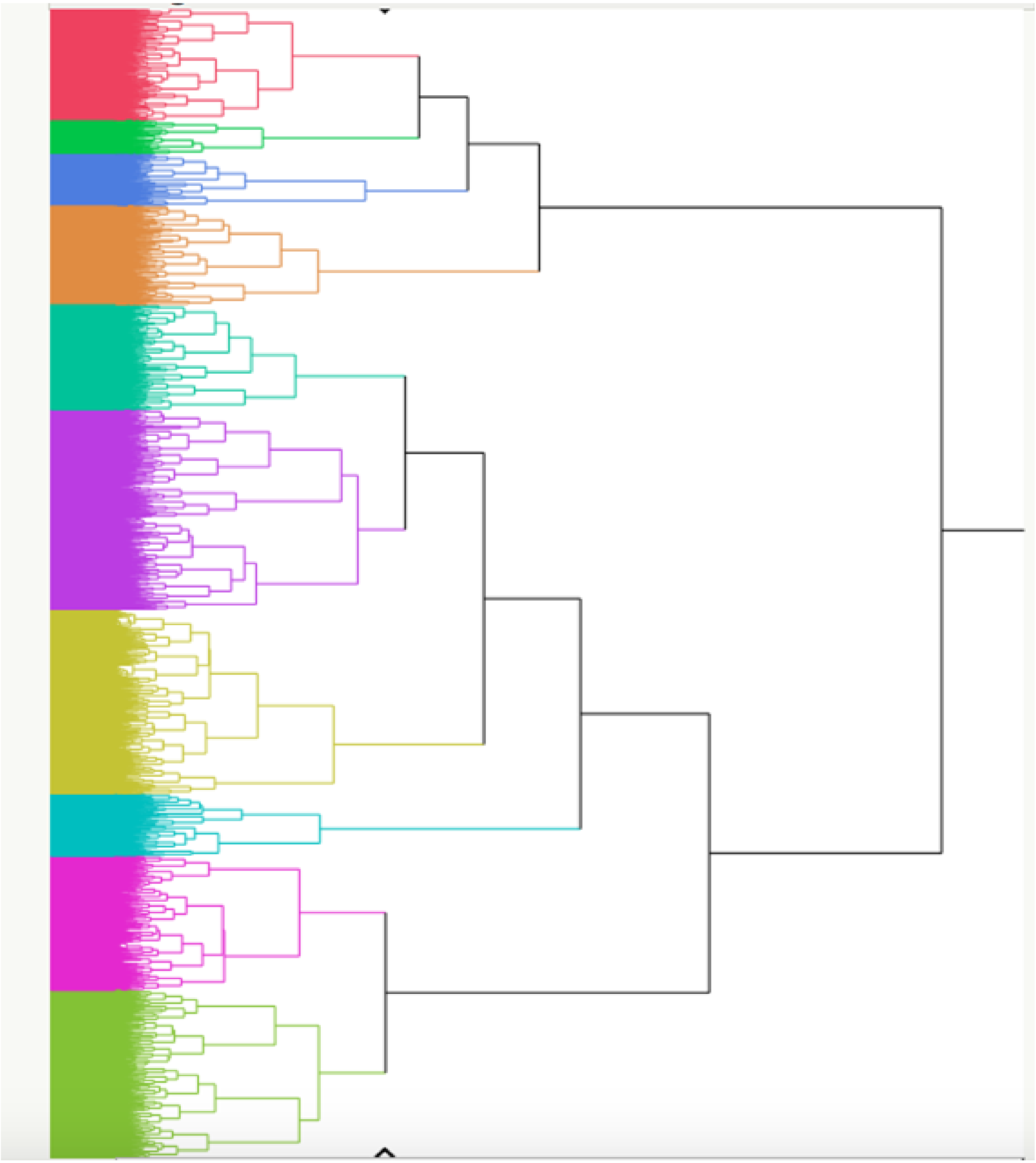
RSquare	0.527814
RSquare Adj	0.526509
Root Mean Square Error	0.508467
Mean of Response	5.159059
Observations (or Sum Wgts)	8708

AICc	BIC
12959.95	13143.66

CLUSTERING

DENDROGRAM

- Hierarchical Clustering



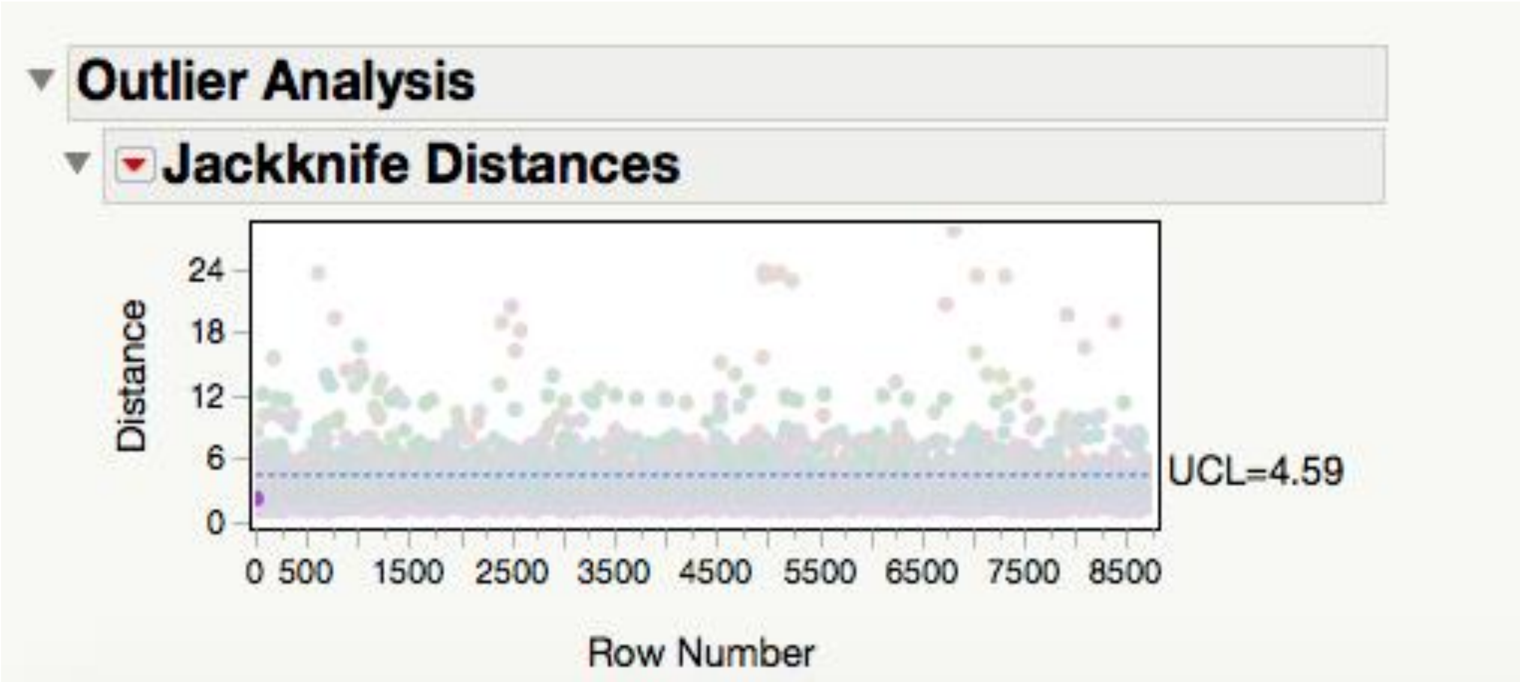
OUTLIERS

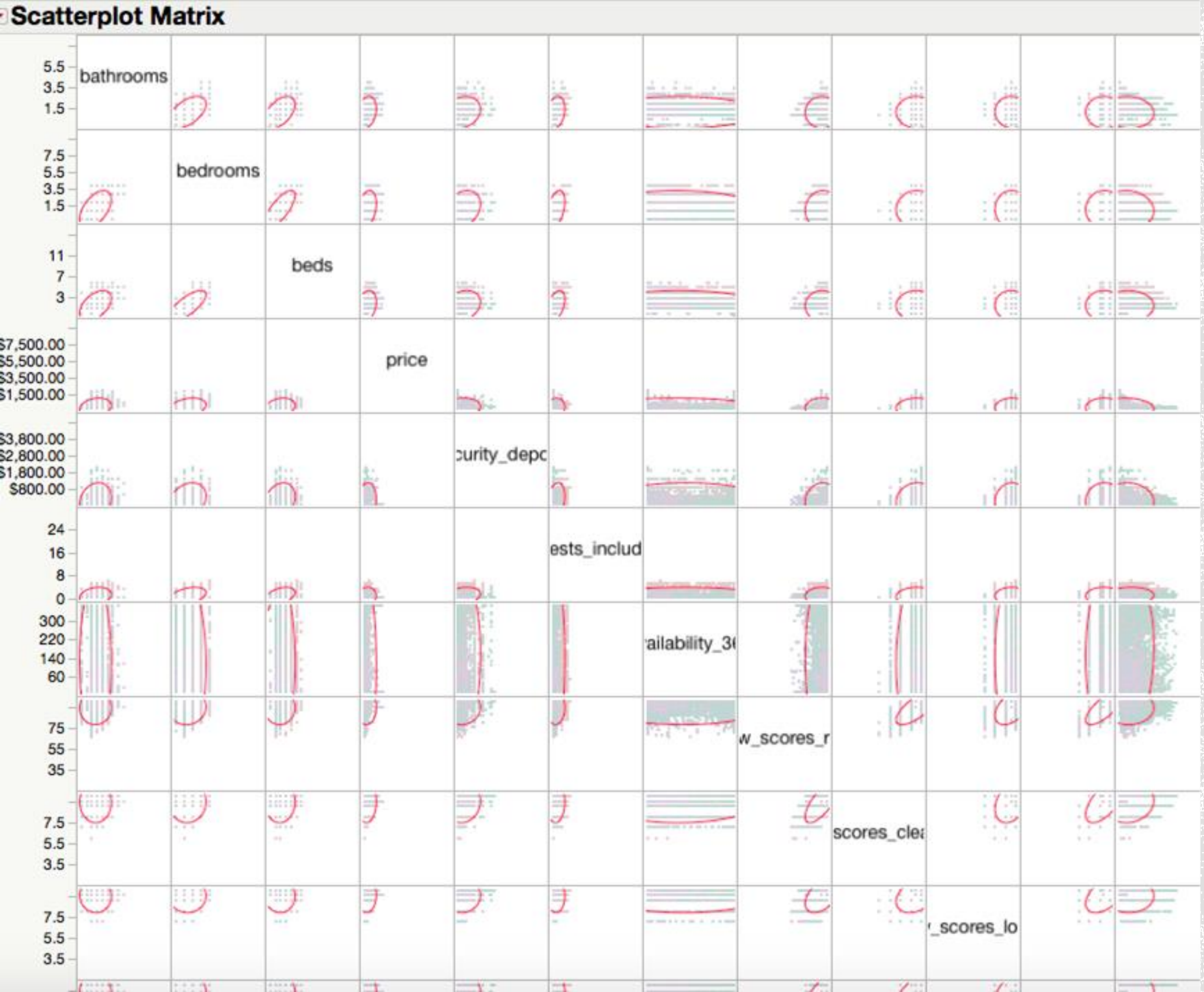
- Even though the dataset has been cleaned, there are still several outliers that may affect accuracy of the model.

- e.g.,

		price	security_deposit
73	3	$215.00	$500.00
74	2	$350.00	0
75	1	$89.00	0
76	2	$157.00	$199.00
77	6	$400.00	$5,000.00

EXCLUDE AND HIDE THE OUTLIERS





MULTIVARIATE ANALYSIS

without outliers

RESULT OF OUTLIERS-EXCLUDED MODEL

- R-square gets improved by 2% and AIC/BICs both decrease by 26%.

Summary of Fit

RSquare	0.538525
RSquare Adj	0.537138
Root Mean Square Error	0.45743
Mean of Response	5.128155
Observations (or Sum Wgts)	8010

AICc	BIC
10228.79	10410.31

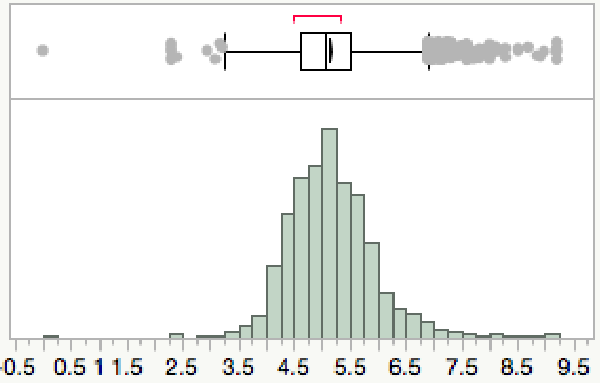
FINAL MODEL FEATURES & COEFFICIENTS

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.6076363	0.234367	11.13	<.0001*
host_response_time[a few days or more]	0.078334	0.041978	1.87	0.0621
host_response_time[N/A]	0.2049233	0.016141	12.70	<.0001*
host_response_time[within a day]	-0.015854	0.018256	-0.87	0.3852
host_response_time[within a few hours]	-0.038932	0.015729	-2.48	0.0133*
host_is_superhost[f]	-0.089616	0.015249	-5.88	<.0001*
host_identity_verified[f]	0.0237336	0.012213	1.94	0.0520
room_type[Entire home/apt]	1.141381	0.037536	30.41	<.0001*
room_type[Private room]	0.5698977	0.037204	15.32	<.0001*
bathrooms	0.155993	0.013081	11.93	<.0001*
bedrooms	0.1845283	0.010899	16.93	<.0001*
beds	0.0788507	0.009406	8.38	<.0001*
security_deposit	1.6821e-5	1.975e-5	0.85	0.3945
guests_included	0.0058403	0.005822	1.00	0.3158
availability_365	0.0005147	4.178e-5	12.32	<.0001*
instant_bookable[f]	-0.002534	0.014027	-0.18	0.8566
reviews_per_month	-0.049811	0.004042	-12.32	<.0001*
cancellation_policy[flexible]	-0.468438	0.188179	-2.49	0.0128*
cancellation_policy[moderate]	-0.552193	0.188131	-2.94	0.0033*
cancellation_policy[strict]	-0.535849	0.187956	-2.85	0.0044*
cancellation_policy[super_strict_30]	-1.202577	0.22172	-5.42	<.0001*
review_scores_rating	0.0050617	0.001842	2.75	0.0060*
review_scores_cleanliness	0.07296	0.011707	6.23	<.0001*
review_scores_location	0.1152075	0.011071	10.41	<.0001*
review_scores_value	-0.072432	0.013074	-5.54	<.0001*

DISTRIBUTION OF LOGGED PRICE WITHOUT OUTLIERS

Distributions

Log[price]



Quantiles

100.0%	maximum	9.210340372
99.5%		7.6009024595
97.5%		6.8023947633
90.0%		5.9914645471
75.0%	quartile	5.5214609179
50.0%	median	5.0751738152
25.0%	quartile	4.605170186
10.0%		4.3174881135
2.5%		3.9120230054
0.5%		3.4965075615
0.0%	minimum	0

Summary Statistics

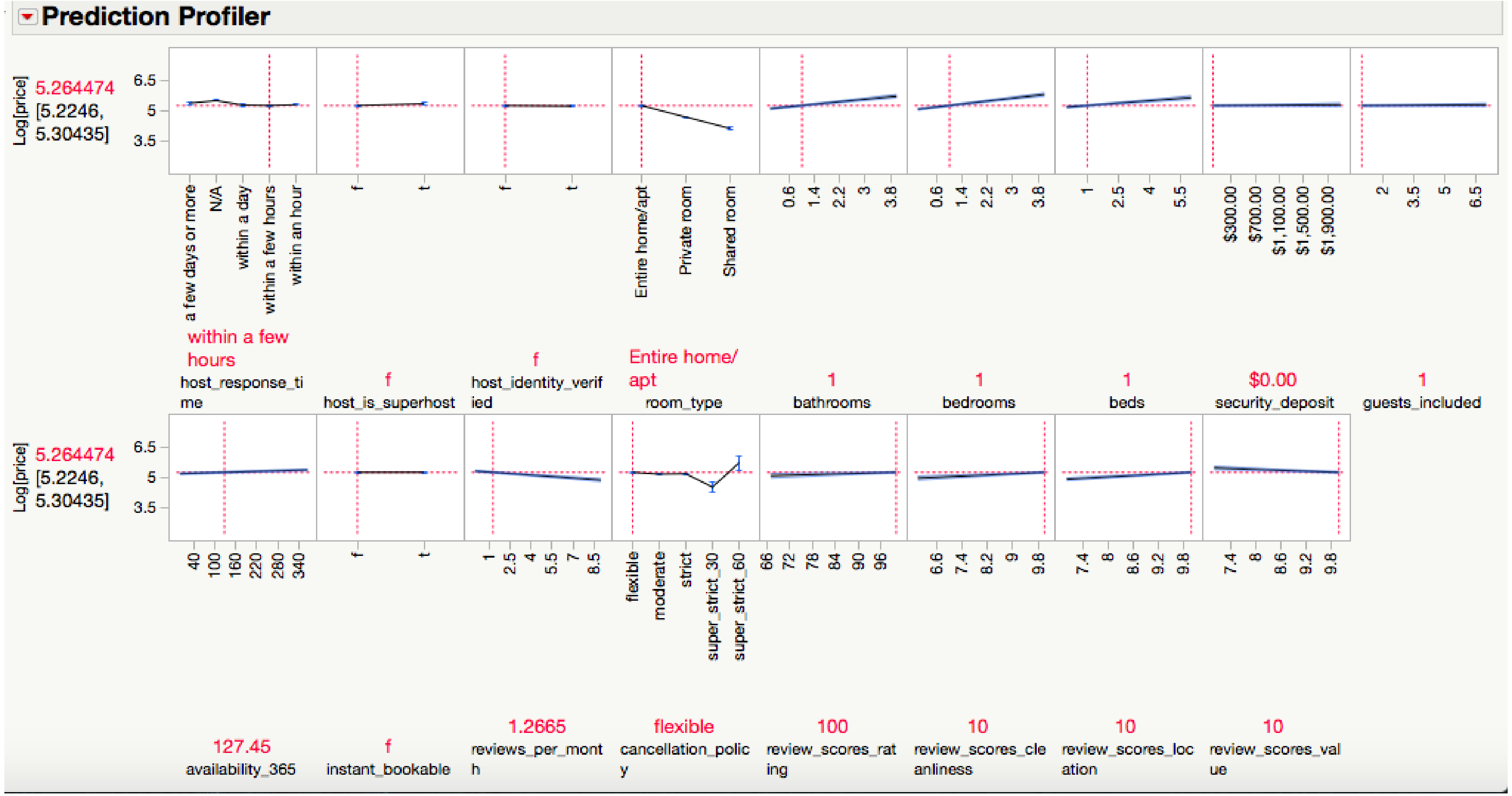
Mean	5.1420181
Std Dev	0.7232371
Std Err Mean	0.0078157
Upper 95% Mean	5.1573387
Lower 95% Mean	5.1266974
N	8563

APPLICATIONS

EXAMPLE

I am a traveler to San Francisco, and I'd like 2 nights' stay at an apartment on Airbnb, with all full score review, 1 bed, 1 bathroom; I do not need the host to be a super host. But I prefer that the host will reply within a few hours and does not require security deposit.

RESULT


Prediction Profiler
Log[price]
5.264474
[5.2246,
5.30435]
6.5
5
3.5
within a few hours
host_response_time
f
host_is_superhost
f
host_identity_verified
Entire home/apt
room_type
1
bathrooms
1
bedrooms
1
beds
$0.00
security_deposit
1
guests_included
127.45
availability_365
f
instant_bookable
1.2665
reviews_per_month
flexible
cancellation_policy
100
review_scores_rating
10
review_scores_cleanliness
10
review_scores_location
10
review_scores_value

EXPECTED PRICE

$$p = 2 * \exp(5.2645) = 2*193 = \$386$$

Note: we use the average value to calculate expected price if the guest does not have a particular requirement for the predictors in the model.

CONCLUSION

As a guest, we normally have an expected price for accommodation. Using this model, we should first fill in our requirements for stay. For example, what is the room type, what are reviews for the host, etc. Then, we can calculate the expected price per night. If the host price is above the model's estimated price, we should pick another host, otherwise, the host is an ideal choice for our trip.

DISCUSSION

- Model complexity vs model accuracy
- Outliers' negative impact on model quality
- Prices' increase rarely linearly (e.g., flight, soup, etc.). Text mining and other machine learning algorithm might help (in the original dataset, there are many variables that are fully composed of text/characters)

Q & A
THANK YOU