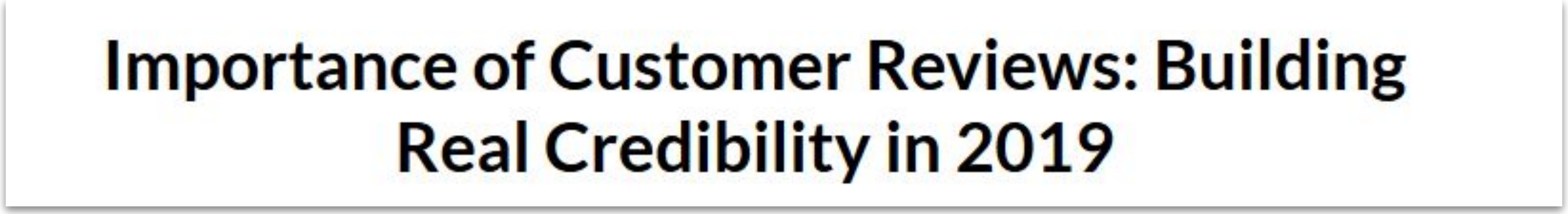
Finding the Impact of Reviews and Predicting Prices on Airbnb

Team A15
: Yehudi Baptiste, Rata Kiewkarnkha, Xinyue Li, Zhipeng Wang, Hirak Bhayani

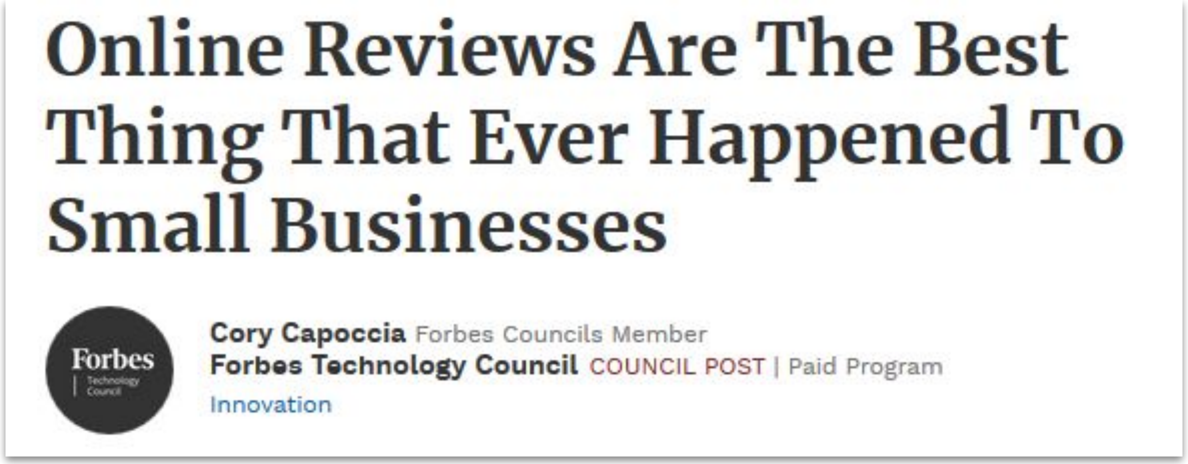
Agenda

1. Business Question(s): Are reviews really beneficials?
   How can we predict prices?
2. Findings
3. The Data
4. Exploratory Data Analysis
5. Regression Analysis
6. Conclusion

There is an underlying assumption that reviews are beneficial



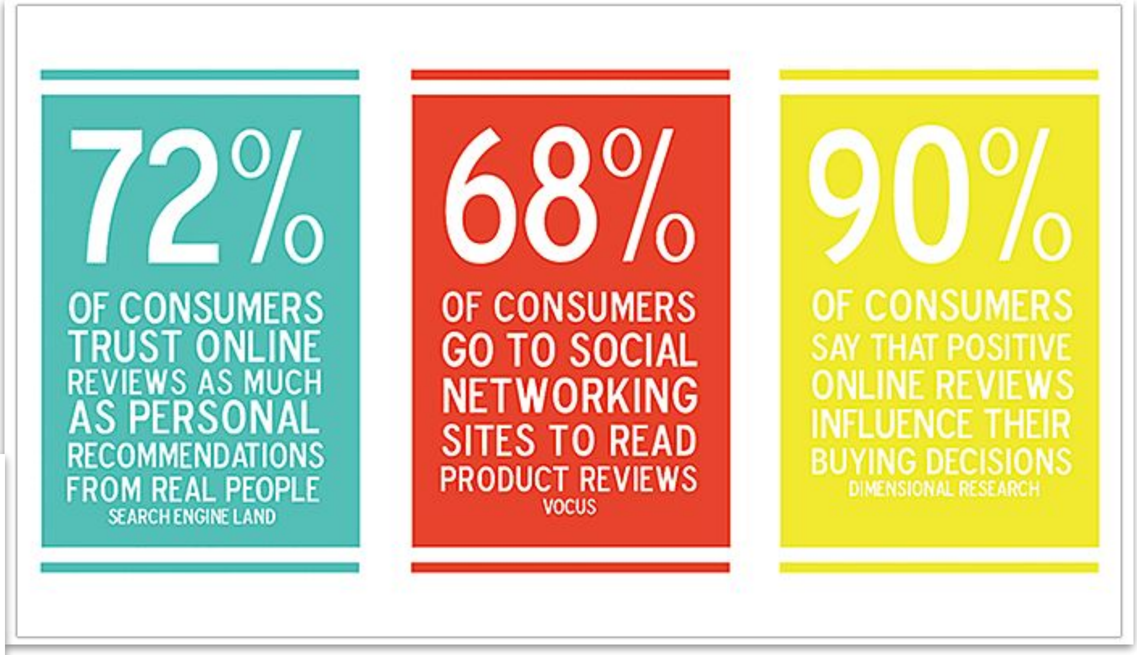
Importance of Customer Reviews: Building Real Credibility in 2019



Online Reviews Are The Best Thing That Ever Happened To Small Businesses

Forbes Technology Council

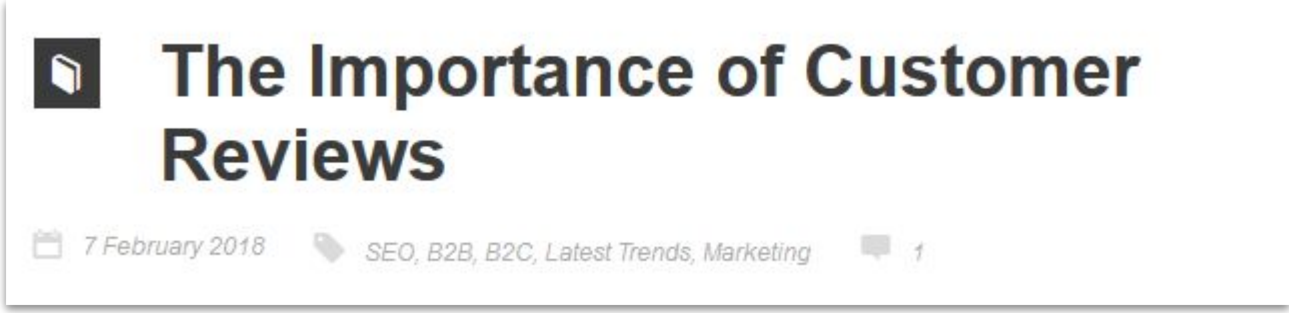
Cory Capoccia Forbes Councils Member
Forbes Technology Council COUNCIL POST | Paid Program
Innovation



72% OF CONSUMERS TRUST ONLINE REVIEWS AS MUCH AS PERSONAL RECOMMENDATIONS FROM REAL PEOPLE
SEARCH ENGINE LAND

68% OF CONSUMERS GO TO SOCIAL NETWORKING SITES TO READ PRODUCT REVIEWS
VOCUS

90% OF CONSUMERS SAY THAT POSITIVE ONLINE REVIEWS INFLUENCE THEIR BUYING DECISIONS
DIMENSIONAL RESEARCH



The Importance of Customer Reviews

7 February 2018 SEO, B2B, B2C, Latest Trends, Marketing 1

Business Questions

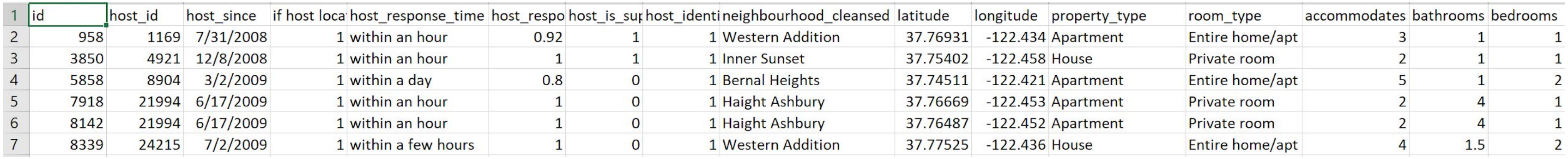
Are higher reviews associated with higher prices?

&

What factors or a listing can help predict its price?

The Data

- insideairbnb.com
- Data for **San Francisco**: 8th of July, 2019
- 7,738 observations and 81 variables
- Removed irrelevant and redundant columns
- Filtered out inactive listings



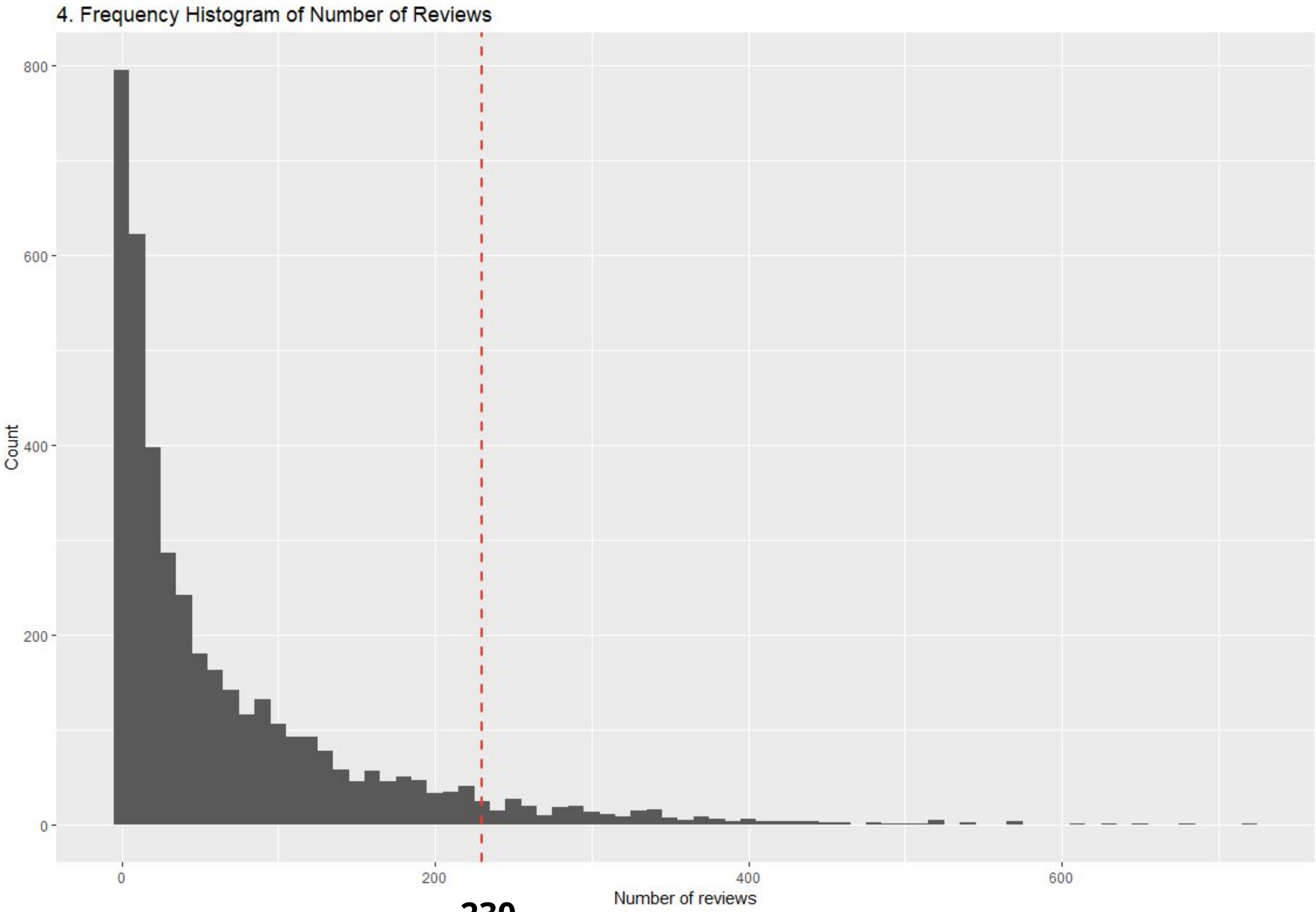
	id	host_id	host_since	if host loca	host_response_time	host_respo	host_is_sup	host_identi	neighbourhood_cleansed	latitude	longitude	property_type	room_type	accommodates	bathrooms	bedrooms
2	958	1169	7/31/2008	1	within an hour	0.92	1	1	Western Addition	37.76931	-122.434	Apartment	Entire home/apt	3	1	1
3	3850	4921	12/8/2008	1	within an hour	1	1	1	Inner Sunset	37.75402	-122.458	House	Private room	2	1	1
4	5858	8904	3/2/2009	1	within a day	0.8	0	1	Bernal Heights	37.74511	-122.421	Apartment	Entire home/apt	5	1	2
5	7918	21994	6/17/2009	1	within an hour	1	0	1	Haight Ashbury	37.76669	-122.453	Apartment	Private room	2	4	1
6	8142	21994	6/17/2009	1	within an hour	1	0	1	Haight Ashbury	37.76487	-122.452	Apartment	Private room	2	4	1
7	8339	24215	7/2/2009	1	within a few hours	1	0	1	Western Addition	37.77525	-122.436	House	Entire home/apt	4	1.5	2

Exploratory Data Analysis

3 Variables

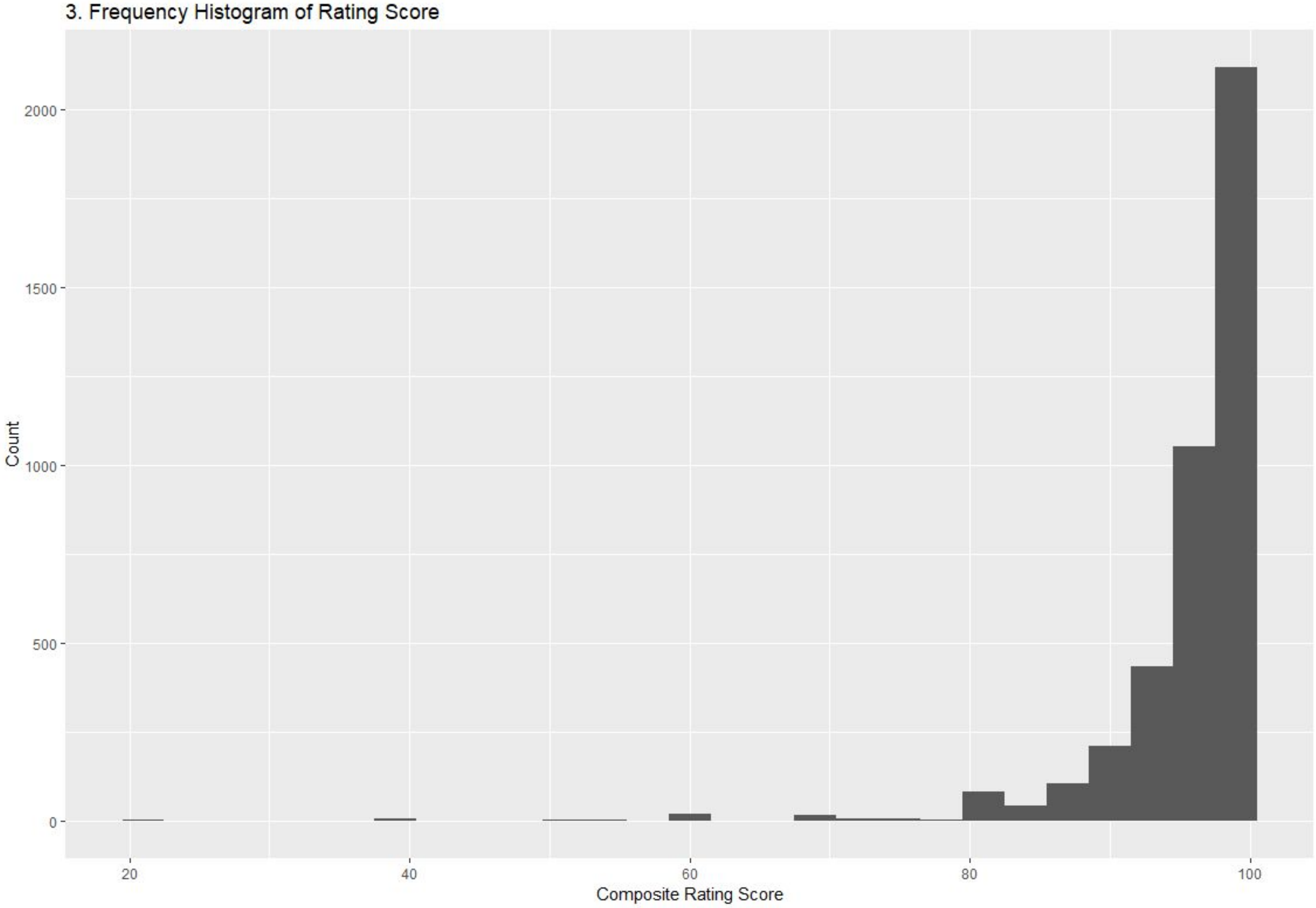
1. Reviews
2. Ratings
3. Price

1. Reviews: Skewed to the Right



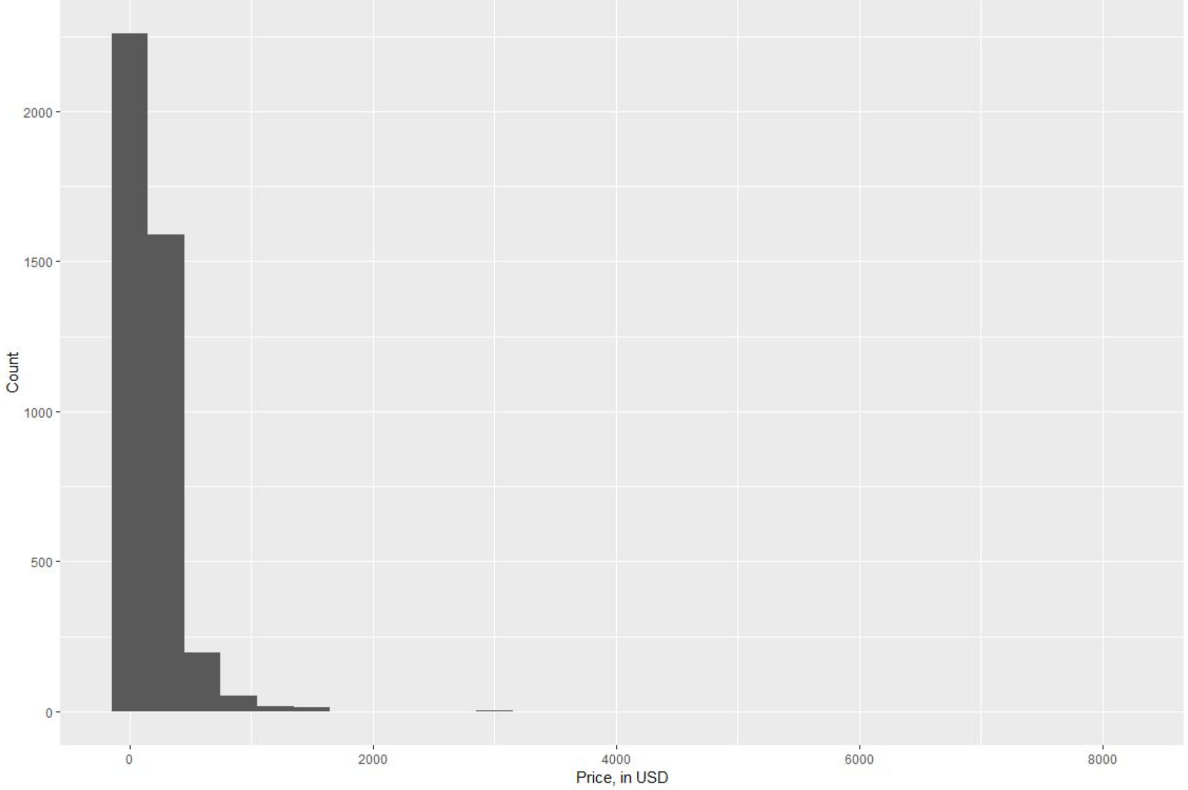
230

2. Ratings: Skewed to the Left

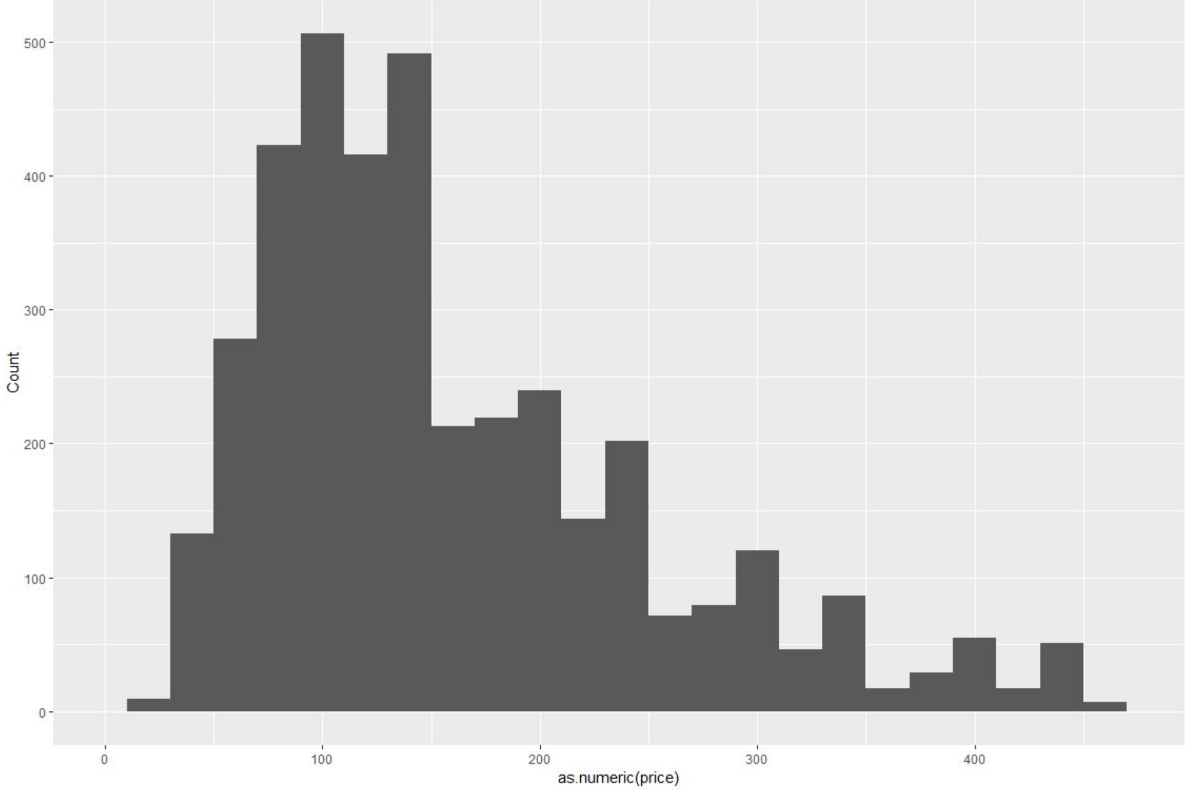


3. Price: Skewed to the Right; Has Outliers

1. Frequency Histogram of Prices



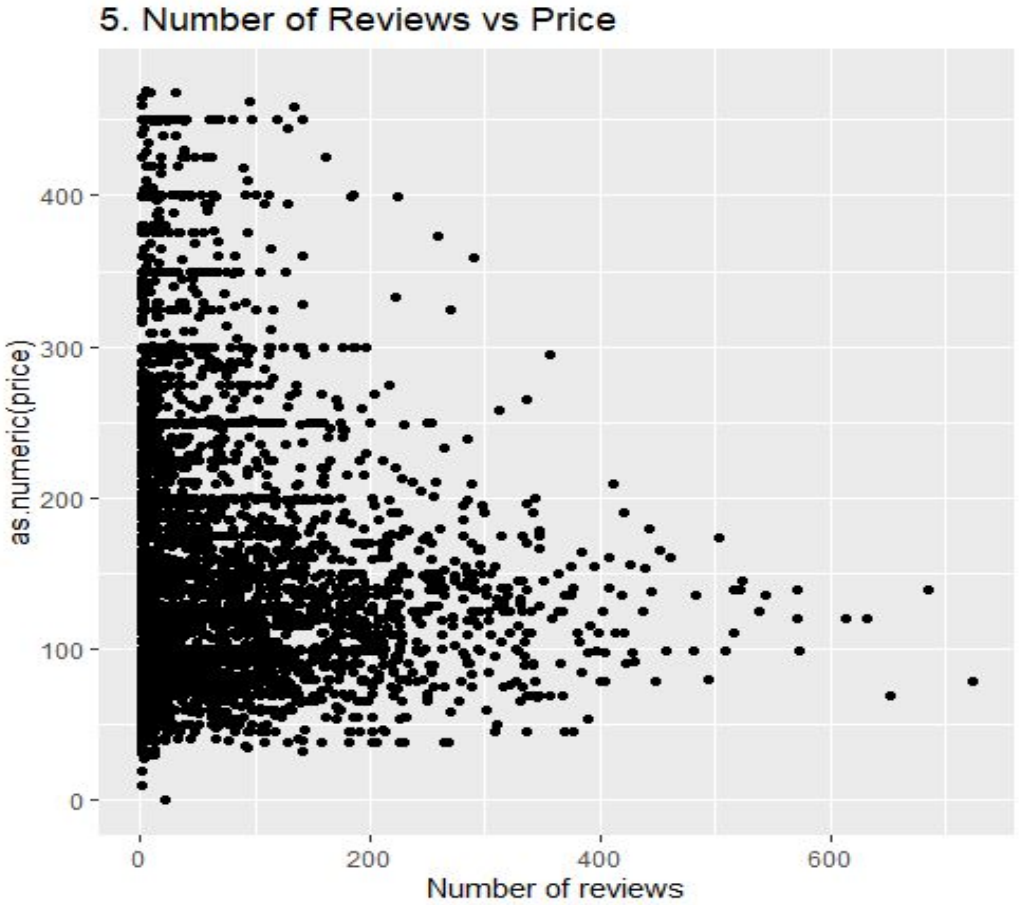
2. Frequency Histogram of Prices (no outliers)



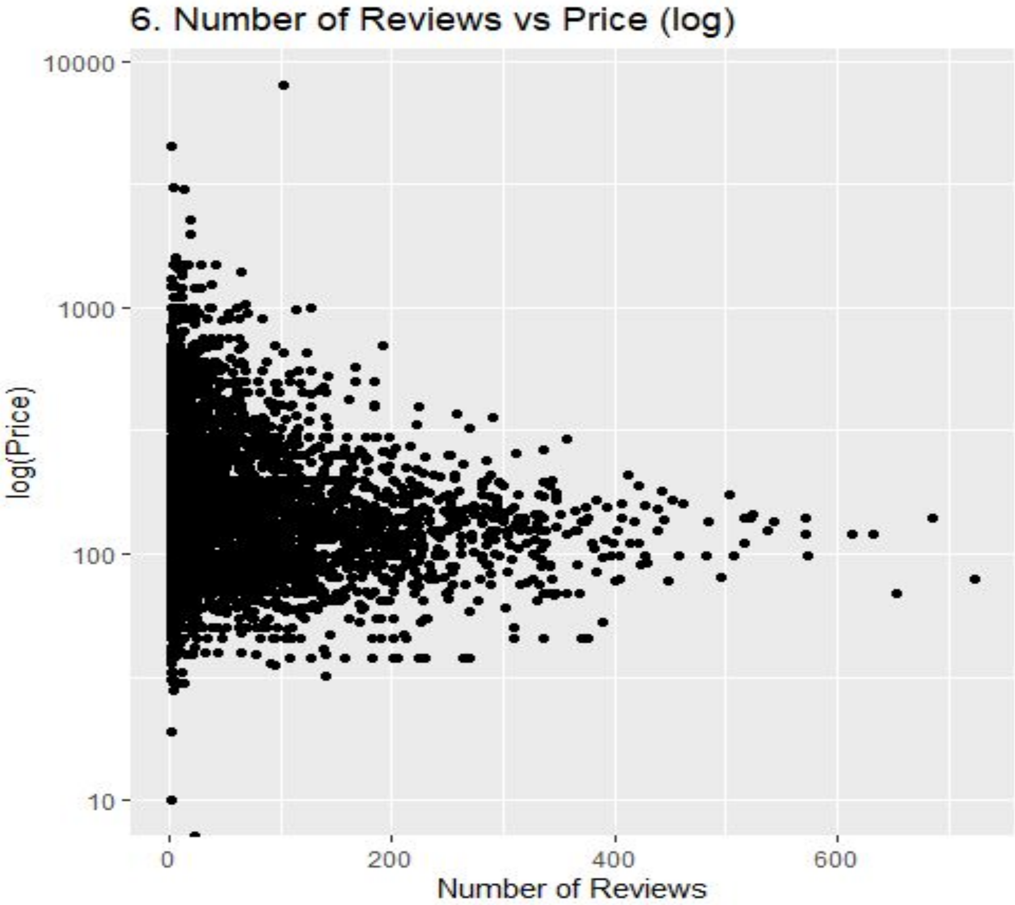
Additional Analysis

- Reviews vs. Price
- Highly Reviewed Listings
- Superhosts

Reviews vs. Price: No Clear Relationship

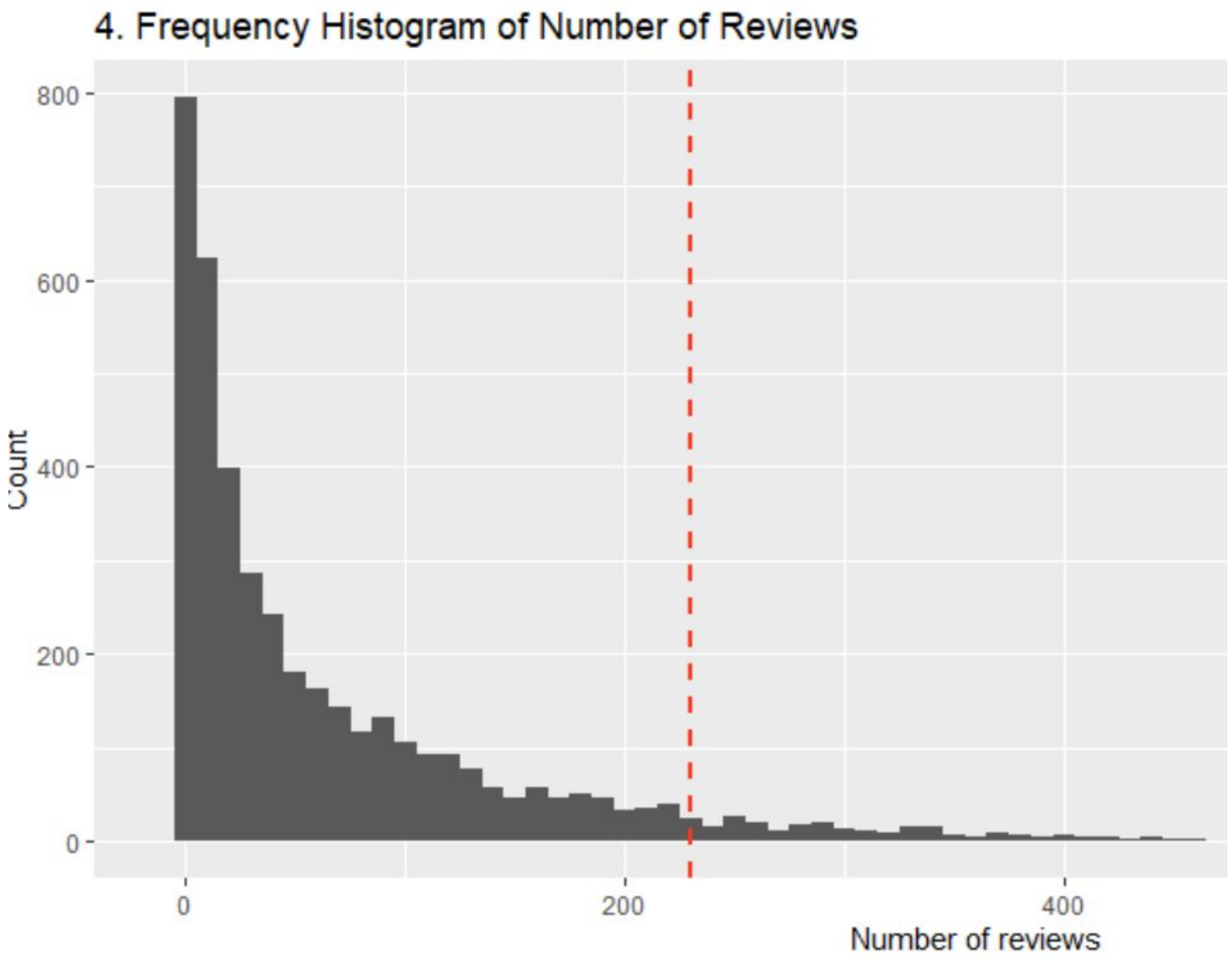


5. Number of Reviews vs Price



6. Number of Reviews vs Price (log)

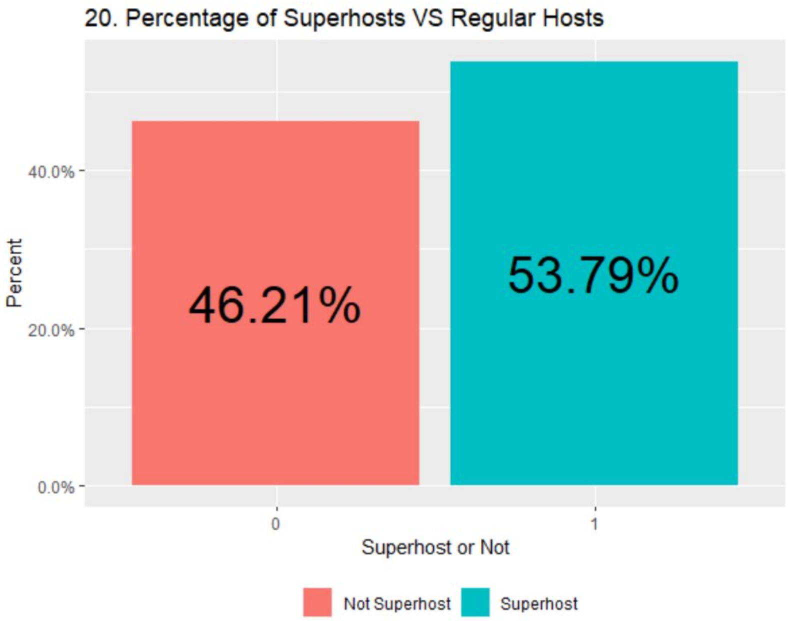
Highly Reviewed Listings: Cut off Point at 230

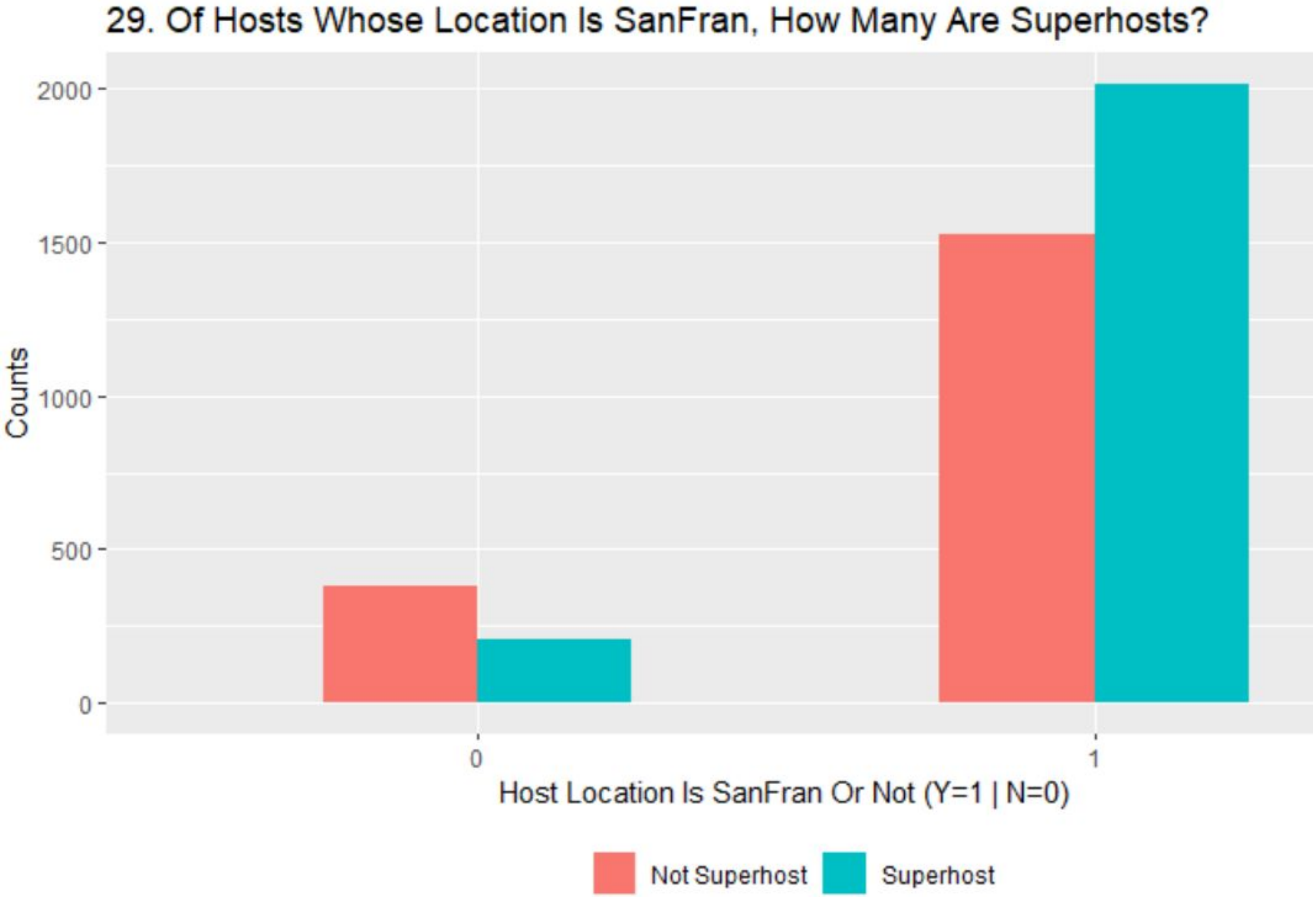


230

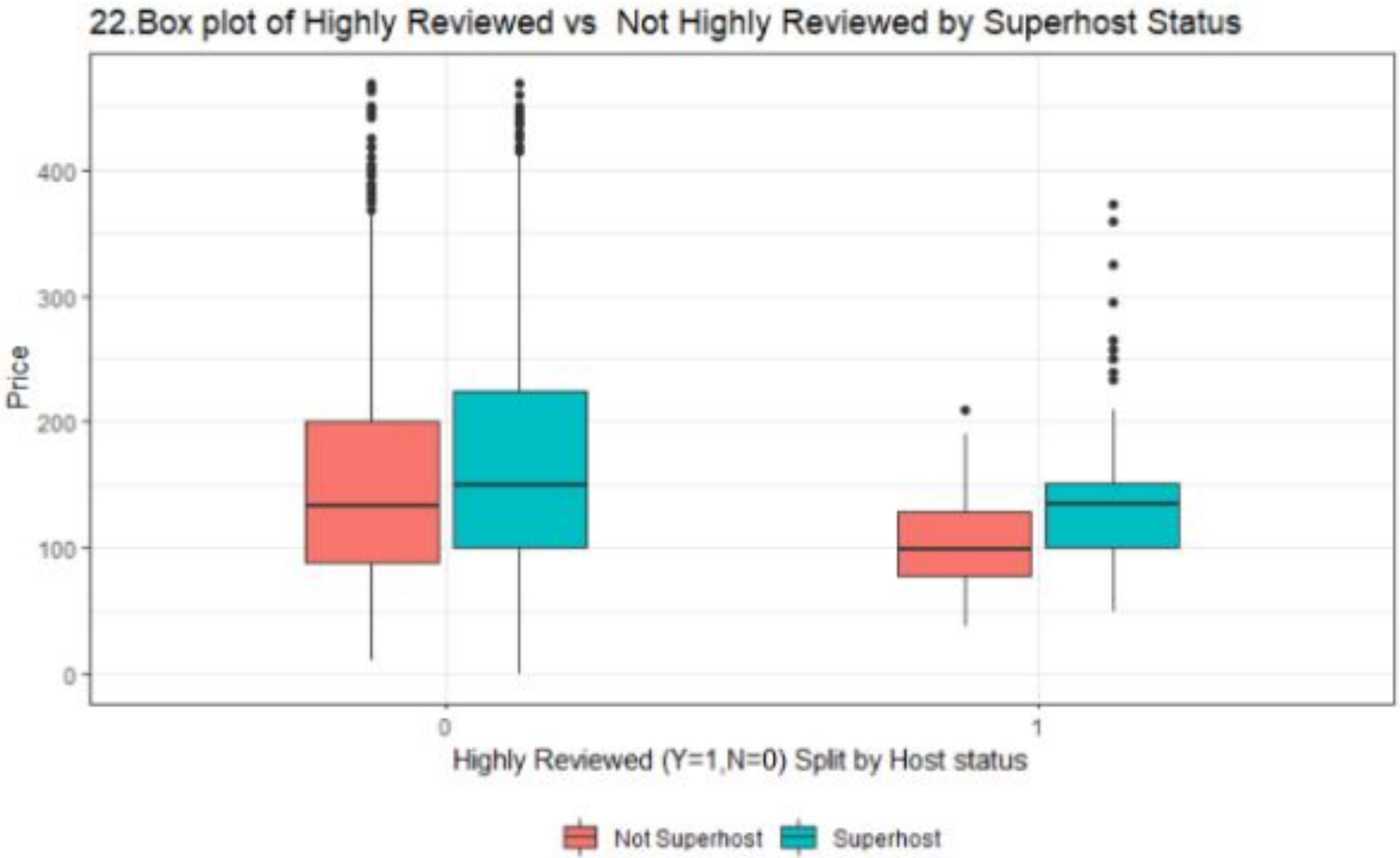
Superhosts: 54% are superhosts, mostly based in San Francisco

"Superhosts are experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests"





Prices for Highly Reviewed Listings: Lower for Both Statuses

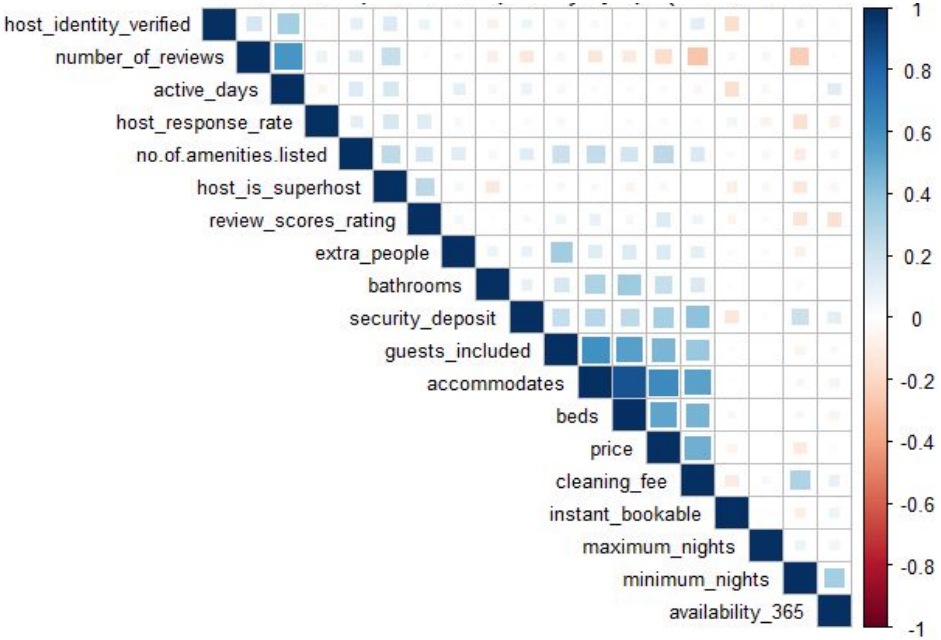


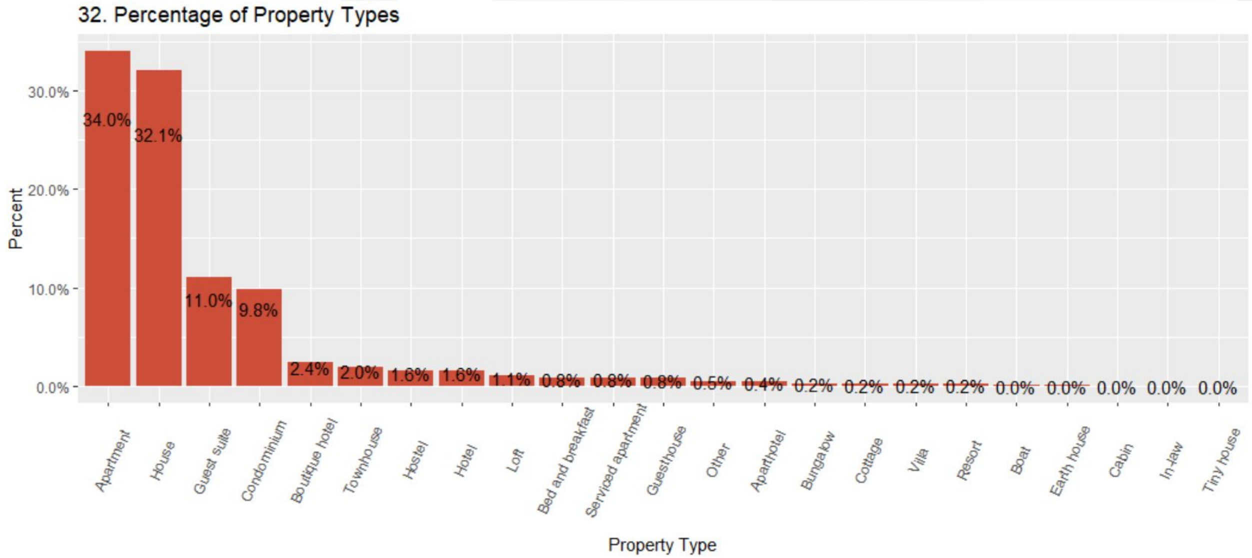
*Cut off Price Outliners

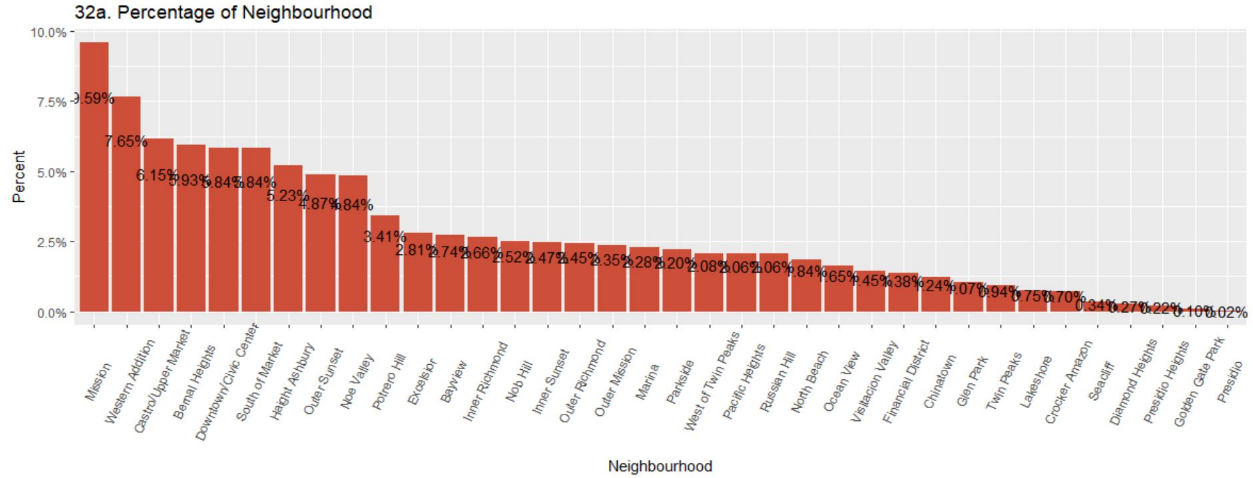
5. Regression Analysis

1. Preparing for regression

- Drop the variables that do not make sense
- Find the **highly correlated** variables
- Reduce the number of **levels for category variables**




32. Percentage of Property Types


32a. Percentage of Neighbourhood

2. Predicting Price

-Automatic selection

```
Model_1=step(fit.nothing,direction='forward',scope=formula('FitAll'))

Model_2=step(FitAll,direction='backward',scope=formula('fit.nothing))

Model_3=step(fit.nothing,direction='both',scope=formula('FitAll'))

Model_4=step(FitAll,direction='both',scope=formula('fit.nothing'))
```

→ **Pick "best of the best" model**

2. Predicting Price

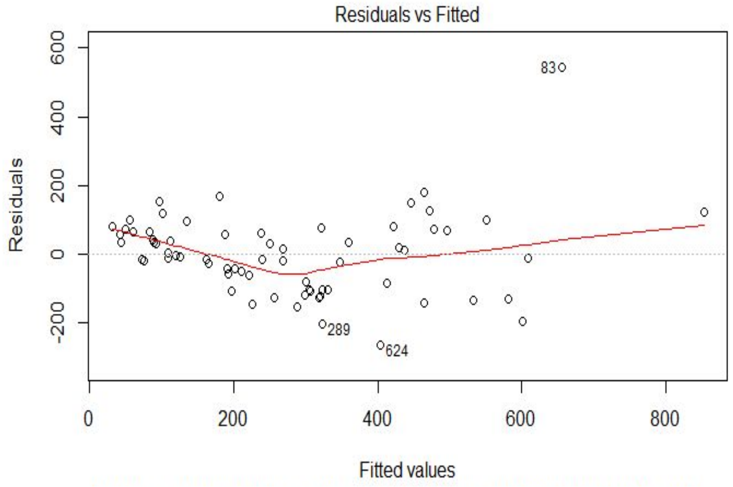
-Homoscedasticity check

price ~ accommodates + extra_people + review_scores_rating + number_of_reviews + minimum_nights + instant_bookable

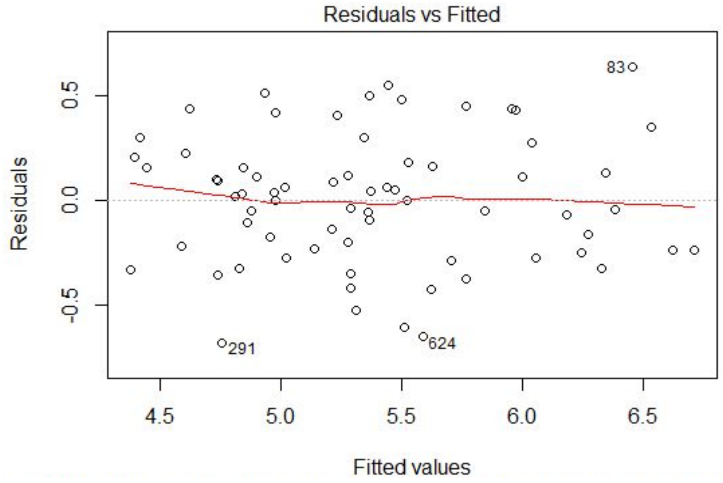
Log Transformation

Log(price) ~ host_response_rate + room_type + accommodates + bathrooms + security_deposit + minimum_nights + number_of_reviews + review_scores_rating + cancellation_policy + active_days + neighbourhood_cleansed

Residuals plots for price regression (Original)



Residuals plots for price regression (Log transformation)



2. Predicting Price

-Realistic consideration & Result

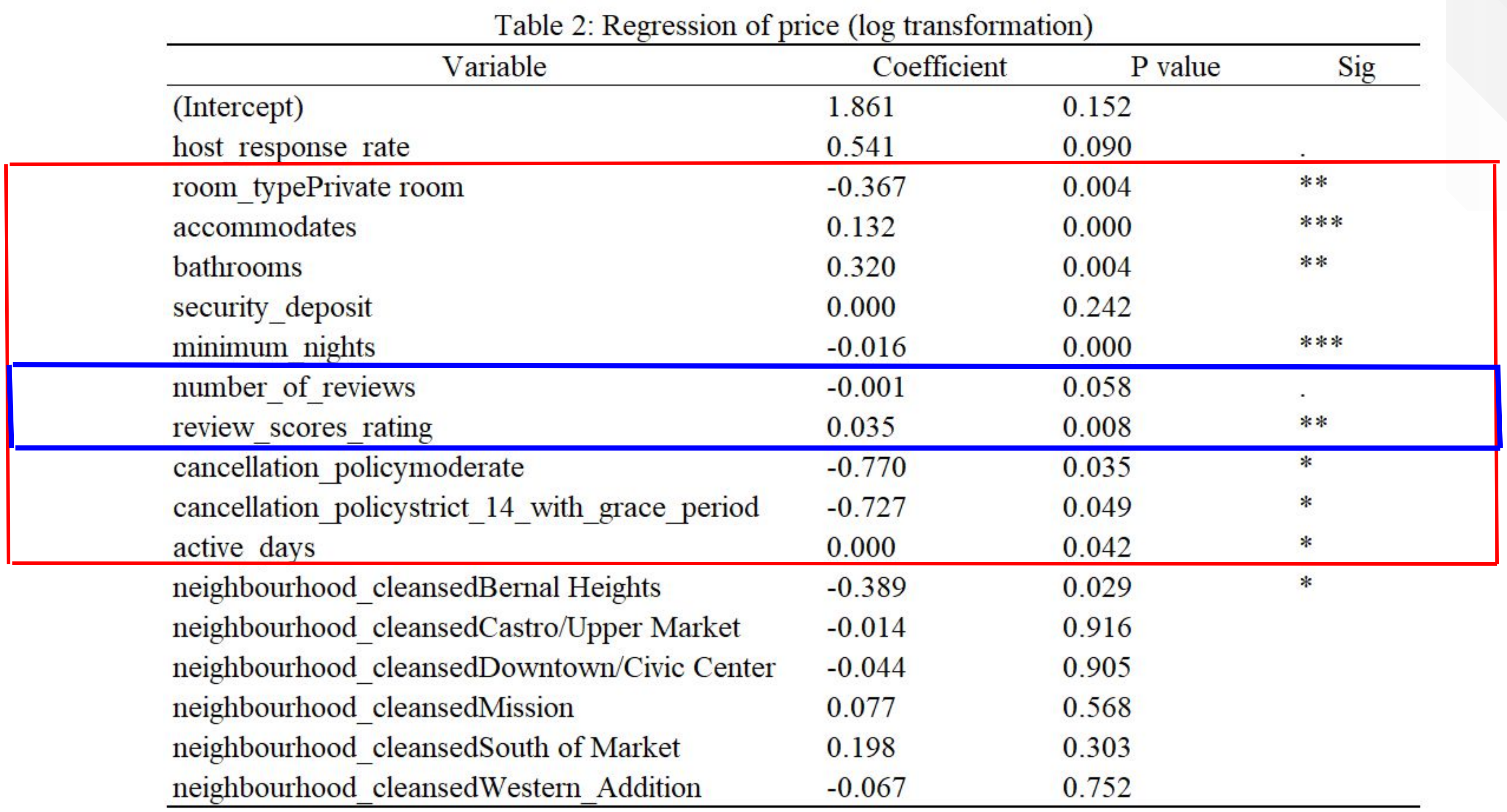


Table 2: Regression of price (log transformation)

Variable	Coefficient	P value	Sig
(Intercept)	1.861	0.152	
host_response_rate	0.541	0.090	.
room_typePrivate room	-0.367	0.004	**
accommodates	0.132	0.000	***
bathrooms	0.320	0.004	**
security_deposit	0.000	0.242	
minimum_nights	-0.016	0.000	***
number_of_reviews	-0.001	0.058	.
review_scores_rating	0.035	0.008	**
cancellation_policymoderate	-0.770	0.035	*
cancellation_policystrict_14_with_grace_period	-0.727	0.049	*
active_days	0.000	0.042	*
neighbourhood_cleansedBernal Heights	-0.389	0.029	*
neighbourhood_cleansedCastro/Upper Market	-0.014	0.916	
neighbourhood_cleansedDowntown/Civic Center	-0.044	0.905	
neighbourhood_cleansedMission	0.077	0.568	
neighbourhood_cleansedSouth of Market	0.198	0.303	
neighbourhood_cleansedWestern_Addition	-0.067	0.752	

3. Predicting Number of Reviews

number_of_reviews ~ bathrooms + active_days + minimum_nights + host_is_superhost + price + accommodates

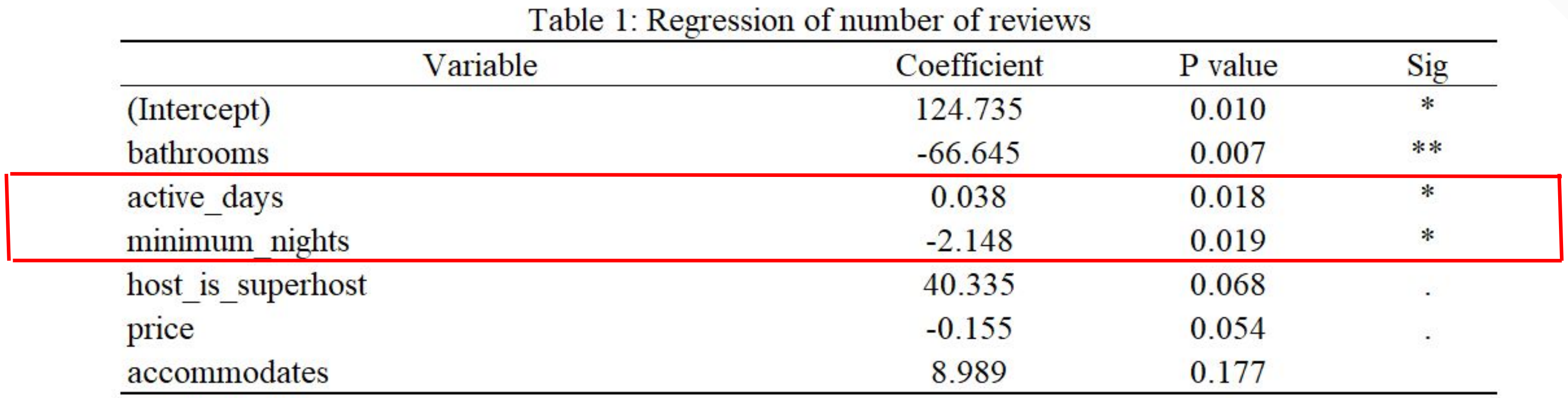


Table 1: Regression of number of reviews

Variable	Coefficient	P value	Sig
(Intercept)	124.735	0.010	*
bathrooms	-66.645	0.007	**
active_days	0.038	0.018	*
minimum_nights	-2.148	0.019	*
host_is_superhost	40.335	0.068	.
price	-0.155	0.054	.
accommodates	8.989	0.177	

4. Predicting Ratings

Review_scores_rating ~ host_response_time + host_response_rate + host_is_superhost + accommodates + bathrooms + price + extra_people + minimum_nights

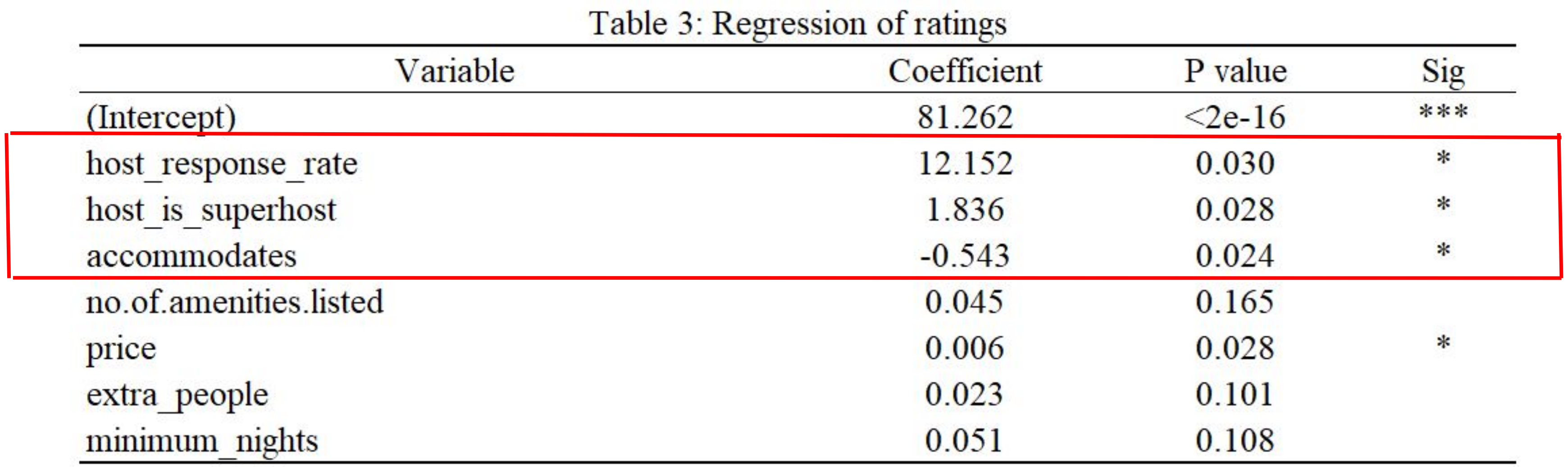


Table 3: Regression of ratings

Variable	Coefficient	P value	Sig
(Intercept)	81.262	<2e-16	***
host_response_rate	12.152	0.030	*
host_is_superhost	1.836	0.028	*
accommodates	-0.543	0.024	*
no.of.amenities.listed	0.045	0.165	
price	0.006	0.028	*
extra_people	0.023	0.101	
minimum_nights	0.051	0.108	

5. Predicting Superhost

host_is_superhost ~ review_scores_rating + extra_people + maximum_nights + number_of_reviews + require_guest_phone_verification + minimum_nights + room_type + security_deposit + active_days

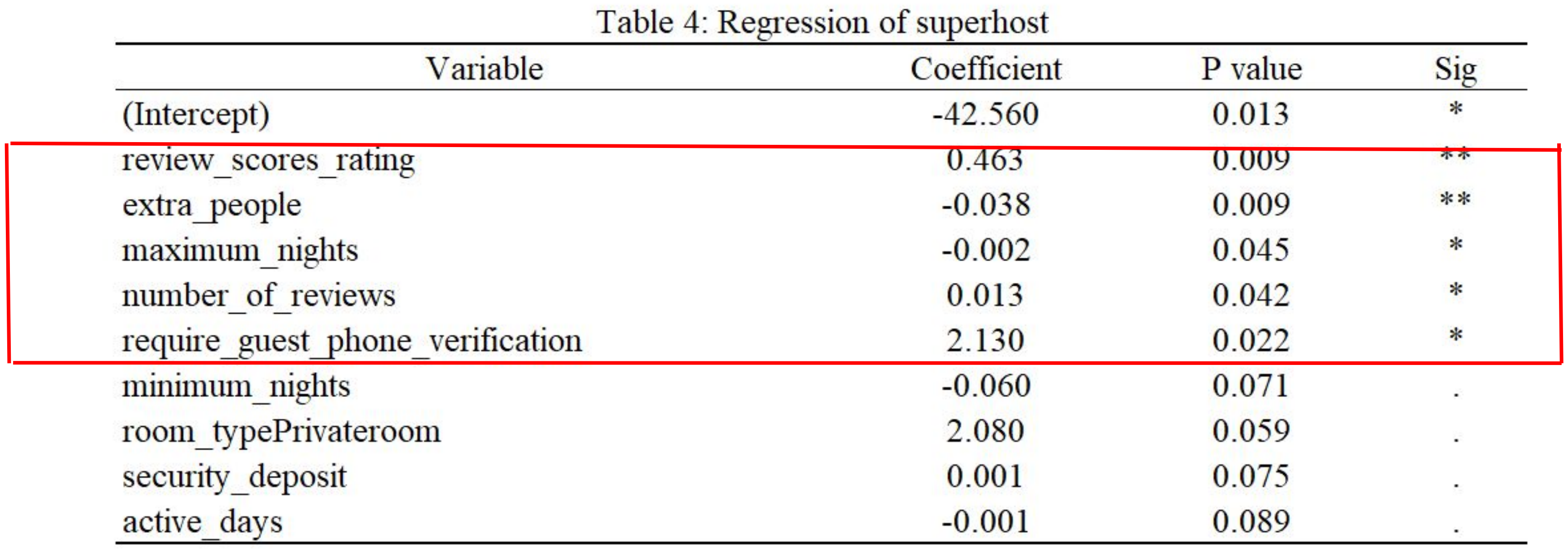


Table 4: Regression of superhost

Variable	Coefficient	P value	Sig
(Intercept)	-42.560	0.013	*
review_scores_rating	0.463	0.009	**
extra_people	-0.038	0.009	**
maximum_nights	-0.002	0.045	*
number_of_reviews	0.013	0.042	*
require_guest_phone_verification	2.130	0.022	*
minimum_nights	-0.060	0.071	.
room_typePrivateroom	2.080	0.059	.
security_deposit	0.001	0.075	.
active_days	-0.001	0.089	.

6. Conclusion

Conclusion

First thought, most interested in the relationship between **listing's characteristics & number of reviews**

Review Model turned out to be the **least interesting/informative with lowest coefficient of determination, producing the least insightful findings**

Model for price produces the most interesting / interpretable results

number of reviews do not have a statistically significant impact on price
but ratings do