Analysis of *listing_Shanghai* from Airbnb
SAS Final Report

11910901 牛圣杰, 11913004 袁灏铖, 11912615 裴彦凯, 11811415 宋晓东

May 29, 2022

Contents

abstract

In this study, we analyze the factors affect the price and review scores on the platform of travel renting, *Airbnb*. First, we conduct the data cleansing on the dataset from *Airbnb*, process the texts into various key words so that we can convert them into dummy variables. Then we extract the features from the cleansed texts, using sentimental analysis and sufficient dimension reduction on the extracted features, in order to reduce the variables and check the NA values. In data analysis, we use linear model to find the variables that determine the *price* of renting houses. And to analysis factors that influence *review score rating* of a house, we conduct a ordinal multinomial logistic regression And obtain some fantasitic conclusions.

Key words: *Feature Engineering, Sentiment Analysis, Sufficient Dimension Reduction, Logistic Regreesion*

1 Introduction

Airbnb was founded in August 2008 and is headquartered in San Francisco, California. It is a platform of travel renting that allows users to publish, search and book rentals online or via a mobile app.

Our dataset contains the basic information of Shanghai Airbnb house, such as scores of various aspects and prices of house. At the same time, our data set includes a lot of text information, such as the self-introduction of the host, the description of the house, the review of the tenant and so on.

Our goal is to conduct the analysis of the price of a house, and factors that affect rentals' scores.

The data is too complex, which make us to do a lot of work in data preprocessing and cleansing. In data preprocessing, the first step is data cleansing. We processed the variables with empty values and unified the format of variables. The second step is characteristic engineering. We classified amenities, bathroom_text and property_type variables. The third step is sentiment analysis, we sorted out the text information in the two variables, description and neighborhood_overview. The fourth step is full dimension reduction. We applied the dimension reduction to the 71 dummy variables derived from the amenities variables.

The analysis process includes data visualization, linear fitting of house prices, and logistic regression of house score.

2 Data Preprocess

2.1 Data Cleansing

The original data has a total of 29159 observations and 59 features. For features with small scale of missing values, we delete the observations that are missing at corresponding these features. After this step, there are 27784 observations left in the dataset.

1	name	1
2	description	1160
3	neighborhood_overview	4509
4	host_since	2
5	host_response_time	5015
6	host_response_rate	5015
7	host_acceptance_rate	3456
8	host_is_superhost	2
9	host_total_listings_count	2
10	host_identity_verified	2
11	bathrooms_text	55
12	bedrooms	1087
13	beds	308
14	first_review	11372
15	last_review	11372
16	review_scores_rating	11372
17	review_scores_accuracy	11880
18	review_scores_cleanliness	11880
19	review_scores_checkin	11882
20	review_scores_communication	11880
21	review_scores_location	11883
22	review_scores_value	11883
23	reviews_per_month	11372

Figure 1. Observe Overall Missing Values

Then we remove observations with price of zero, which also indicates a missing value in price.

	bathrooms_text	bedrooms	beds	amenities	price	minimum_nights	maximum_nights	minimum_minimum_nights
0		NA	NA	["Airport shuttle", "Restaurant", "Long term stays allo...	0.00	1	1095	1
0		NA	NA	["Restaurant", "Long term stays allowed", "Laundry ser...	0.00	1	1095	1
5	3 baths	4	4	["Iron", "Smoke alarm", "Heating", "Breakfast", "Hanger...	0.00	365	1125	365

Figure 2. Price of Zero also Indicates Missing Value

The variable price is currency data, we remove the '$' and ',' symbols in the data, and convert into a numeric value. And variables host_response_rate and host_acceptance_rate are also text data, remove the '$' symbol and convert them to numeric values.

For feature amenities: we remove all symbols like '\u2019s' by regular expression to get all the amenities that have appeared, a total of 255.

```
a full one bedroom flat with living, dinning and balcony, full kitchen of all amenities, and bathroom w/shower. Wash ma
restaurants and nightlife zone within walking distance. New night scene of Yong Kang road, old traditional triangle of D
dropoff allowed"", ""Children\u2019s books and toys"", ""Building staff"", ""Hair dryer"", ""Room-darkening shades"",
d to the room,   it can accommodate 3 people. There is also a fully functional kitchen. The bathroom and the shower ro
our own apartment.. also a lot of restaurants, if u want to taste chinese food u could go around and check which one u
""Hangers"", ""Kitchen"", ""Carbon monoxide alarm"", ""Wifi"", ""Hot water"", ""Smart lock"", ""Elevator"", ""Paid park
ed, the room can accommodate 3 people. There is also a fully functional kitchen. The bathroom and the shower room a
```

Figure 3. Regular Expressions are Noise in Data

We do not select all the features that have appeared, we choose to extract the amenities that appear more than 1000 times in all observations, and convert these amenities into dummy variables, that is, 74 columns of dummy variables are added to the original data set variable.

```
the top 0 is Air conditioning, it appears 27083 times
the top 1 is Wifi, it appears 26824 times
the top 2 is Long term stays allowed, it appears 26444 times
the top 3 is Hair dryer, it appears 25524 times
the top 4 is Shampoo, it appears 25009 times
the top 5 is Essentials, it appears 24271 times
the top 6 is Hangers, it appears 23685 times
the top 7 is Washer, it appears 23204 times
the top 8 is Hot water, it appears 20503 times
the top 9 is Fire extinguisher, it appears 19404 times
the top 10 is TV, it appears 19069 times
...
the top 71 is High chair, it appears 1139 times
the top 72 is Board games, it appears 1083 times
the top 73 is Childrens dinnerware, it appears 1040 times
the top 74 is Waterfront, it appears 989 times
```

Extra pillows and blankets	Iron	Smoke alarm	Oven	Heating	Bed linens	Hangers	Kitchen
1	1	1	1	1	1	1	1
0	1	1	0	0	0	1	1
0	1	1	0	1	0	1	1
0	1	1	0	1	0	1	1
0	1	1	0	1	0	1	1
0	1	1	0	1	0	0	0
1	1	1	0	0	1	1	1
0	0	0	0	1	0	1	1
0	1	0	1	1	1	1	1

We filter for amenities with similar meanings. There are 71 dummy variables left representing different amenities.

Figure 4. Selection on amenities

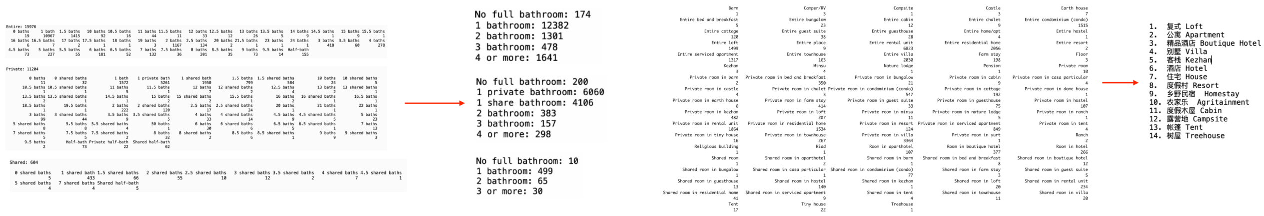
We filiter for amenities with similar meanings. There are 71 dummy variables left representing different amenities.

Considering data such as review_scores, we find that the missing values of total score and first_review and last_review data are actually the same. In other words, the house missing all three data sets is the house that has never been rented out! Therefore, these data can be deleted during review analysis. In addition, when review_scores_rating is 0, all subsequent ratings are missing. Therefore, we believe that Airbnb will give a total rating of 0 to a house that is not highly rated by its guests, which leads to the existence of missing values in the later ratings.

2.2 Feature Extraction

In this section, we conduct the feature engineering on three groups of three variables.

The first one is the bathroom_text. First, we have many complicated labels under different room_type, so we classify these labels, namely, the houses without complete toilet and that with different number of toilets. And when room_type is private_room, there are private toilets and shared toilets.



(a) Bathroom (b) Property

Figure 5. Text Cleansing

The second classified feature is property_type. It is also very complicated to list all the labels, but according to the investigation on the website of Airbnb, all the labels can be divided into 14 building types, which is listed on the right. We classify the labels from the original data into these 14 new levels.

Third is about whether can be scheduled several variables, or say has availability. These variables can be replaced with a single variable, which includes most information. The new variable has five levels, representing the house can be scheduled within 30 days, cannot be scheduled within 30 days but can be scheduled within 60 days, and then is 90 days, 365 days and the last one indicates that reservations are not available.

has_availability	availability_30	availability_60	availability_90	availability_365
t	0	21	51	326
t	0	25	55	330
t	6	6	6	275
t	0	0	0	236
t	0	0	0	49
t	15	45	75	350
t	24	54	84	359
t	28	58	88	363
t	30	60	90	365
t	29	59	89	364

Feature: Variable
Available in 30 days
Available in 60 days
Available in 90 days
Available in 365 days
Not Available

Figure 6. Availability

2.3 Sentiment Analysis

We worked with two complex text variables, description (with 1099 NA values) and neighborhood_overview (with 4368 NA values). And in both variables, compared to whether the description and the neighborhood overview specifically mention some things, what can best reflect the difference between different listings is the activeness of the host's introduction to the listing.

We hope that the text can be converted into a value between 0 and 1 through sentiment analysis to represent the host's positivity for this listing.

The description of house and neighborhood can be classified into Chinese and non-Chinese, in which the sentiment analysis is carried out by SnowNLP and TextBob respectively. Noticeably, the SnowNLP returns a range of polarity values [0,1], which indicates the probability of positive sentiment that hide behind the sentence, which is quite in line with our expectations. But the polarity of the sentiment TextBob returned varied in a range of [-1,1], with -1 representing completely negative and 1 representing completely positive. So we project the sentimental polarity returned by TextBob between [0,1] through a function mapping. In addition, we set the missing value in these two variables to 0, which implies that if the host does not describe the property or the surrounding situation, we will choose to give him the lowest sentimental rating.

has_availability	availability_30	availability_60	availability_90	availability_365
t	0	21	51	326
t	0	25	55	330
t	6	6	6	275
t	0	0	0	236
t	0	0	0	49
t	15	45	75	350
t	24	54	84	359
t	28	58	88	363
t	30	60	90	365
t	29	59	89	364

```
Feature: Variable
Available in 30 days
Available in 60 days
Available in 90 days
Available in 365 days
Not Available
```

Figure 7. Sentiment Analysis

2.4 SDR on Amenity Features

2.4.1 Methodology Statement

Let $\boldsymbol{X}:\Omega\rightarrow\mathbb{R}^p$ a r.v. measurable w.r.t. $\mathscr{R}^p$ the Borel σ-field in $\mathbb{R}^p$, and $Y:\Omega\rightarrow\mathbb{R}$ a r.v. measurable w.r.t. $\mathscr{R}$. **Sufficient Dimension Reduction** (**SDR**) is concerned with the situations where the distribution of Y given $\boldsymbol{X}$ depends on $\boldsymbol{X}$ only through a set of linear combinations of $\boldsymbol{X}$. That is

$$\exists\,\boldsymbol{\beta}\in\mathbb{R}^{p\times r},\ r\leq p,\ \text{s.t.} Y\perp\!\!\!\perp\boldsymbol{X}|\boldsymbol{\beta}'\boldsymbol{X}$$

This relation is **unchanged** if repalce $\boldsymbol{\beta}$ by $\boldsymbol{\beta A}$, $\forall\boldsymbol{A}\in\mathbb{R}^{r\times r}$ nonsingular, that is

$$Y\perp\!\!\!\perp\boldsymbol{X}|\boldsymbol{\beta}'\boldsymbol{X}\Rightarrow Y\perp\!\!\!\perp\boldsymbol{X}|\left(\boldsymbol{\beta A}\right)'\boldsymbol{X},\ \Leftrightarrow Y=h(\boldsymbol{\beta}'\boldsymbol{X},\boldsymbol{\epsilon})$$

If the conditional independence condition above is satisfied for $\boldsymbol{\beta}$, then we call $\mathscr{S}=\operatorname{span}(\boldsymbol{\beta})$ a **Sufficient Reduction Subspace**, or **SDR subspace**. Obiviously, SDR subspace always exists because $Y\perp\!\!\!\perp\boldsymbol{X}|\boldsymbol{X}$ always holds. Furthermore, the SDR subspace is not unique, consider the following fact.

Another easier understanding expression is similar to the orthogonal factor model of covariates conditional on responser

$$\boldsymbol{X}_y=\boldsymbol{\mu}+\boldsymbol{A}\boldsymbol{f}_y+\boldsymbol{\epsilon}$$

where $\boldsymbol{X}_y=\boldsymbol{X}|y$, $\boldsymbol{A}$ is the loading matrix fixed and to be estimated, $\boldsymbol{f}_y=\boldsymbol{f}|y$ is the factor conditional on y, $\boldsymbol{X}$'s are independent and $\boldsymbol{X}\perp\!\!\!\perp\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}\sim(\boldsymbol{0},\boldsymbol{\Delta})$.

Here we use the **Sliced Inverse Regreesion** in our study.
Under linearity assumption above, suppose $\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}>0$, then for $\boldsymbol{P_B}(\boldsymbol{\Sigma})=\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B}\right)^{-1}\boldsymbol{B}'\boldsymbol{\Sigma}$

$$E\left(X|\boldsymbol{\beta}'\boldsymbol{X}\right)-E(\boldsymbol{X})=\boldsymbol{P}'_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\left(\boldsymbol{X}-E(\boldsymbol{X})\right)$$

Suppose $\boldsymbol{X}$ is square-integrable and $\boldsymbol{\Sigma}=\operatorname{Var}(\boldsymbol{X})$ is non-singular. Then, under the linearity assumption,

$$\boldsymbol{\Sigma}^{-1}\left(E(\boldsymbol{X}|Y)-E(\boldsymbol{X})\right)\in\mathscr{S}_{Y|\boldsymbol{X}}$$

2.4.2 Extraction Result

To describe the feature of amenity in a hotel, we use 71 dummy variables. This is too tedious for data analysis to use 71 dummy variables. So we want to reduce the dimensions of these variables while avoid the loss of information about *price*. According to the survey on airbnb's website, these amenities can be roughly divided into five aspects: with respect to five areas: basic amenity, security amenity, check-in amenity, kitchen ware and bathroom amenity. To reduce the dimension of amenity variates, we need to score on the amenity conditions.

In traditional processing ways, the sum of dummy variables in each observations would be used as a score. But such a score does not include the information of responser *price*.

In this study, we use a score of the amenities adding the information about *price* by SDR, which can be viewed as a conditional version of factor analysis.

We extract one direction (factor) from each components of amenity variables and obtain the score on these components for each observations and normalize these scores.

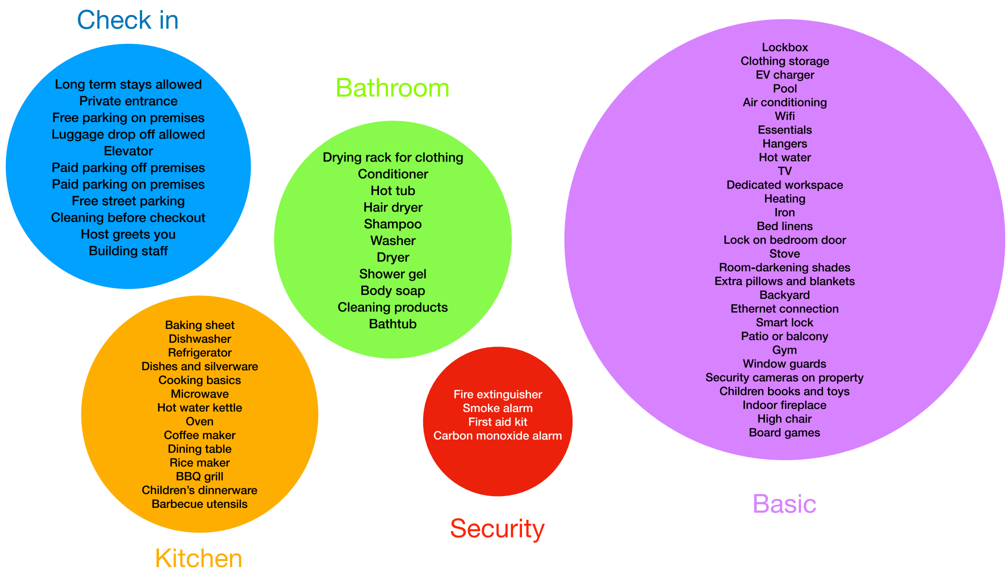
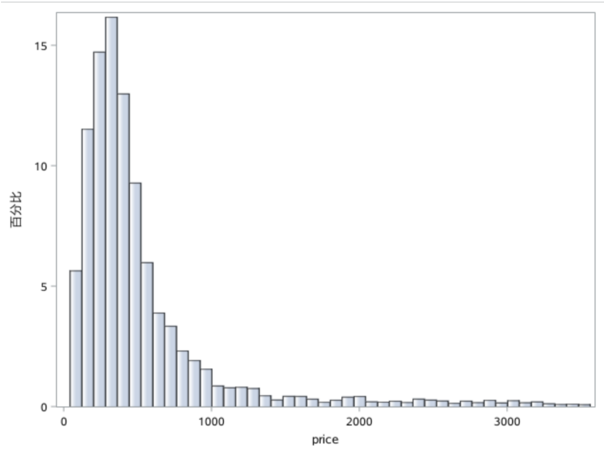


Figure 8.The Components of amenities Variables

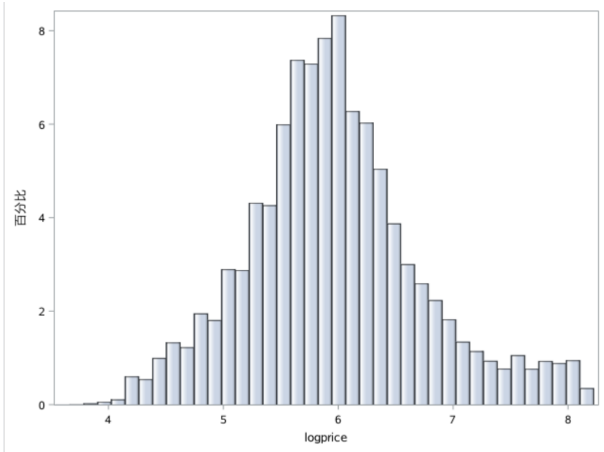
To verify the performance of scoring, we compare the R^2_{adj} of linear model using the 71 amenity variables, the traditional scoring scores and the SDR scores, which are 0.2643, 0.0374 and 0.2169, respectively. So the SDR score preserve the information of the amenities features with respect to *price* when conduct dimension reduction.

3 Exploratory Data Analysis

First of all, regarding the *price* , we drew a histogram for it with the quantile below 95%, that is, 3527. We find that the distribution of *price* was obviously right-skewed even if we excluded the situation that *price* above tens of thousands. Therefore, we think it is necessary to take the logarithm on *price* so that make it normal.



(a) *price* (skewed)



(b) log(*price*) (approximately normal)

Figure 9. Histogram of *price* and log(*price*)

We draw boxplots of some classified variables in the dataset, and it could be seen that most of these variables had some influence on price, such as accommodates, neighbourhood room_type, etc. Of course, there were also some variables that didn't seem to have any influence. Such as host_response_time.

Here we gives the *accommodates*-log(*price*) boxplot, we find that they have a truncated and approximate linearity. Other boxplots will be given in the Appendix.

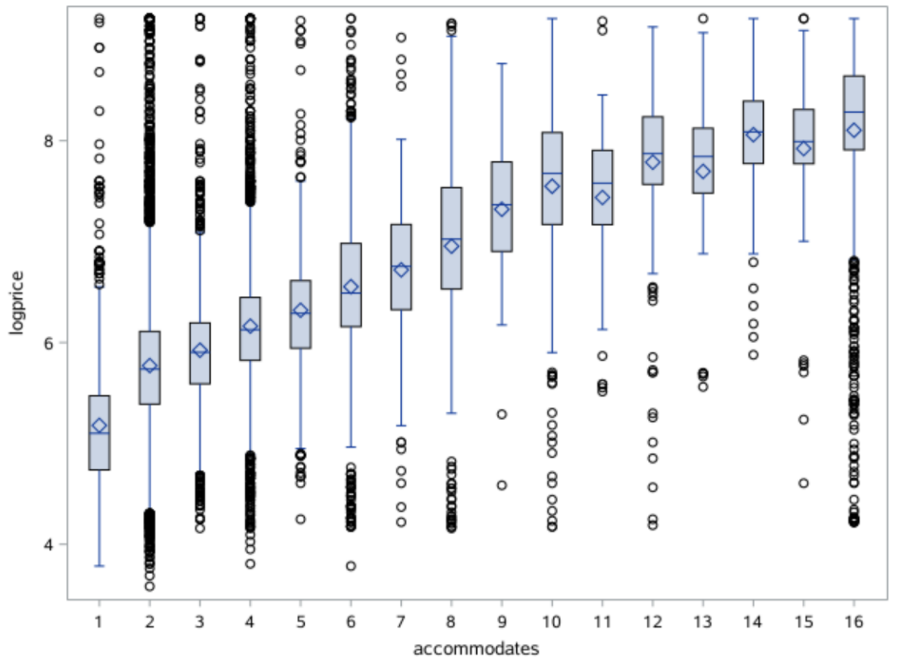


Figure 10. Accommodates-log(*price*)

4 Data Analysis

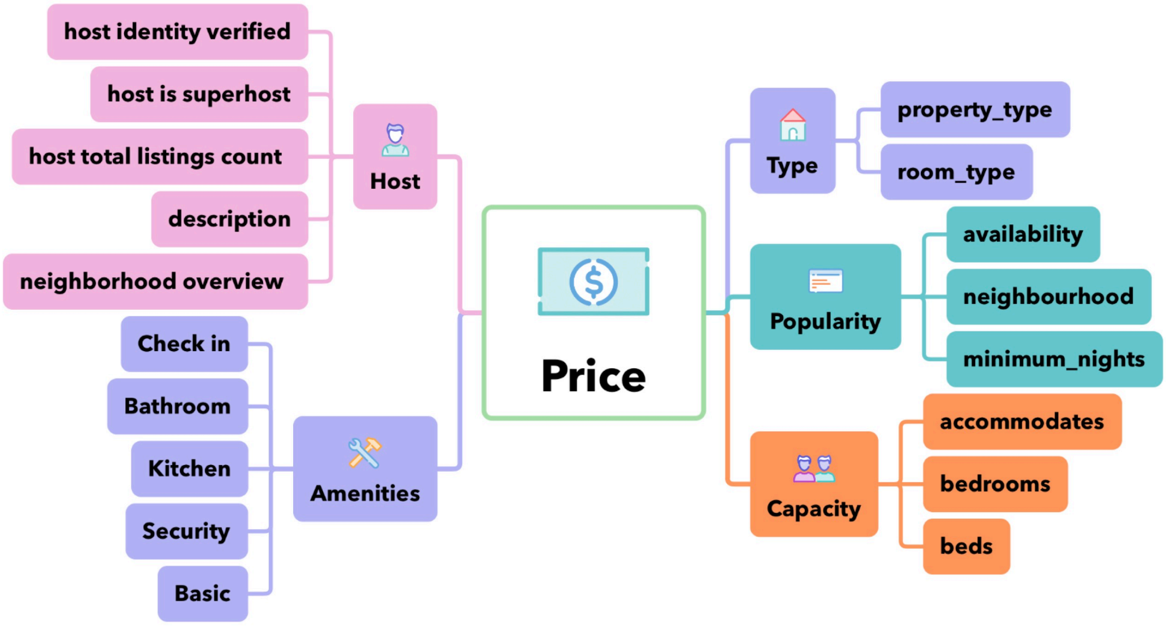


Figure 11. Choice of Explantory Variables

4.1 Analysis of Price

Since the prices of house renting on Airbnb are determined by the owner and have nothing to do with reviews and ratings, we didn't use relevant features when analysis the price. We can divide explanatory variables into five categories.

The host category represents the information of the host, including whether he is verified, whether he is a super host, the number of houses he owns, and the score of positivity processed by sentiment analysis before. Amenity category represents the score of various amenties of the house.

These five variables are based on the full dimension reduction method mentioned before to give the score of amenities in different aspects.

The third is the house source type, which can be divided into two types: one is the type of room, the other is the type of building.

The fourth category is what we call the popularity of listings, which can be divided into three categories: availability, location, and minimum number of nights.

The final category is the capacity of the house, which is divided into the number of bedrooms, the number of beds and the maximum capacity of the house.

4.1.1 Variable Correlation Detection

Then we detect the correlations of the variables. It can be found that features under the house capacity category have a high correlation, so we need to test their VIF values (shown in the last column in **Fig 13.**) and find that there is no multicollinearity between variables, that is to say, we can consider a simple linear model for regression.

Pearson 相关系数, N = 27783								
description	description 1.00000	neiover 0.12206	logprice 0.04495	host_total_listings_count -0.03637	Kitchen -0.01570	Bathroom -0.01108	bedrooms 0.01104	Security 0.01035
neiover	neiover 1.00000	description 0.12206	Security 0.06526	minimum_nights -0.05897	Basic -0.05861	Bathroom -0.05559	Checkin 0.04826	host_total_listings_count -0.03817
Checkin	Checkin 1.00000	host_total_listings_count -0.30086	Security 0.29371	Bathroom -0.26136	logprice 0.26132	Kitchen 0.22349	accommodates 0.19739	Basic -0.18575
Basic	Basic 1.00000	logprice -0.41192	accommodates -0.36488	Kitchen -0.32877	bedrooms -0.31241	Bathroom 0.29277	beds -0.28707	Checkin -0.18575
Bathroom	Bathroom 1.00000	Basic 0.29277	Checkin -0.26136	logprice -0.24528	minimum_nights 0.22040	Security -0.18893	Kitchen -0.15525	host_total_listings_count 0.12841
Kitchen	Kitchen 1.00000	Basic -0.32877	accommodates 0.26411	beds 0.23319	bedrooms 0.23089	Checkin 0.22349	logprice 0.22328	Security 0.19619
Security	Security 1.00000	Checkin 0.29371	Kitchen 0.19619	Bathroom -0.18893	Basic -0.15465	beds 0.13049	accommodates 0.12469	bedrooms 0.11819
minimum_nights	minimum_nights 1.00000	Bathroom 0.22040	Checkin -0.07416	logprice -0.06711	accommodates -0.06200	neiover -0.05897	Basic 0.05315	beds -0.05156
accommodates	accommodates 1.00000	beds 0.84171	bedrooms 0.82898	logprice 0.67405	Basic -0.36488	Kitchen 0.26411	Checkin 0.19739	Security 0.12469
bedrooms	bedrooms 1.00000	beds 0.85028	accommodates 0.82898	logprice 0.61064	Basic -0.31241	Kitchen 0.23089	Checkin 0.16490	Security 0.11819
beds	beds 1.00000	bedrooms 0.85028	accommodates 0.84171	logprice 0.55764	Basic -0.28707	Kitchen 0.23319	Checkin 0.15770	Security 0.13049
logprice	logprice 1.00000	accommodates 0.67405	bedrooms 0.61064	beds 0.55764	Basic -0.41192	Checkin 0.26132	Bathroom -0.24528	Kitchen 0.22328
host_total_listings_count	host_total_listings_count 1.00000	Checkin -0.30086	Bathroom 0.12841	logprice -0.11686	Kitchen -0.09179	Security -0.09137	Basic 0.04794	bedrooms -0.04017

Figure 12. Variable Correlation

4.1.2 Regression Model

Since the list of facotrs are too long, here we only gives the estimated β for numeric covariates

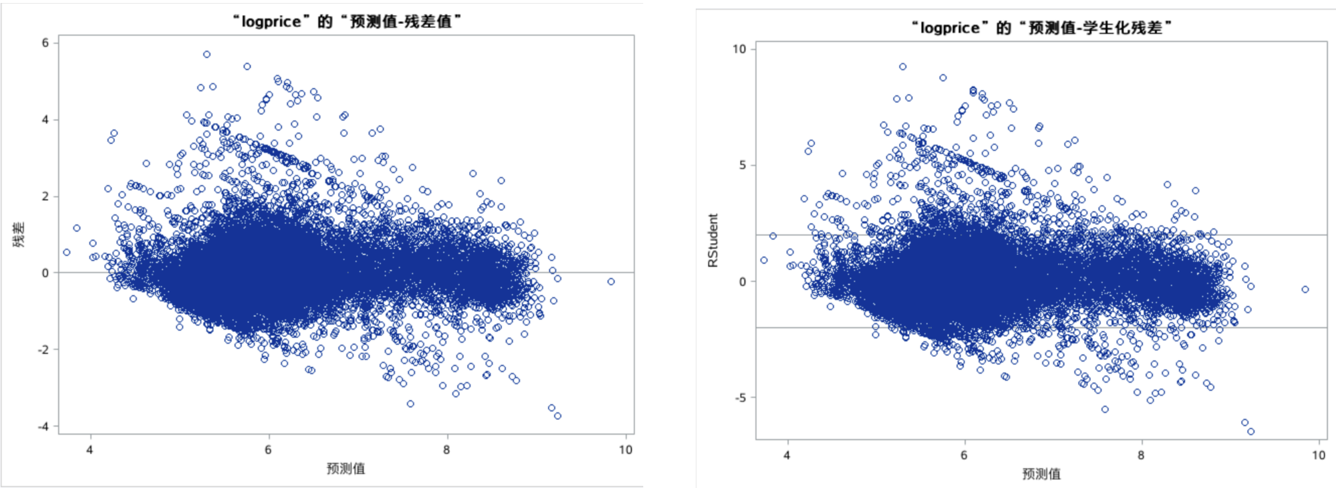
参数估计							
变量	自由度	参数估计	标准误差	t 值	Pr > |t|	容差	方差膨胀
Intercept	1	6.24224	0.03543	176.18	<.0001	.	0
description	1	0.09196	0.00877	10.48	<.0001	0.98188	1.01846
neiover	1	-0.00804	0.00998	-0.81	0.4207	0.97268	1.02808
Checkin	1	0.47872	0.02830	16.92	<.0001	0.77228	1.29486
Basic	1	-1.26634	0.04004	-31.62	<.0001	0.75398	1.32630
Bathroom	1	-0.68211	0.02986	-22.85	<.0001	0.81931	1.22054
Kitchen	1	-0.14590	0.03395	-4.30	<.0001	0.83416	1.19882
Security	1	-0.07928	0.01348	-5.88	<.0001	0.87721	1.13998
minimum_nights	1	0.00001952	0.00012906	0.15	0.8798	0.93966	1.06422
host_total_listings_count	1	-0.00078510	0.00005284	-14.86	<.0001	0.89206	1.12100
accommodates	1	0.14726	0.00238	61.85	<.0001	0.23092	4.33055
bedrooms	1	0.11434	0.00446	25.63	<.0001	0.23185	4.31321
beds	1	-0.04748	0.00310	-15.34	<.0001	0.21522	4.64644

Figure 13. Estimated β for Numeric Covariates

4.1.3 Model Diadnosis

We tested the correlation of the above variables and found that the variables under the maximum capacity class had a high correlation. Therefore, we tested their VIF values and find that the variables did not directly have multicollinearity.

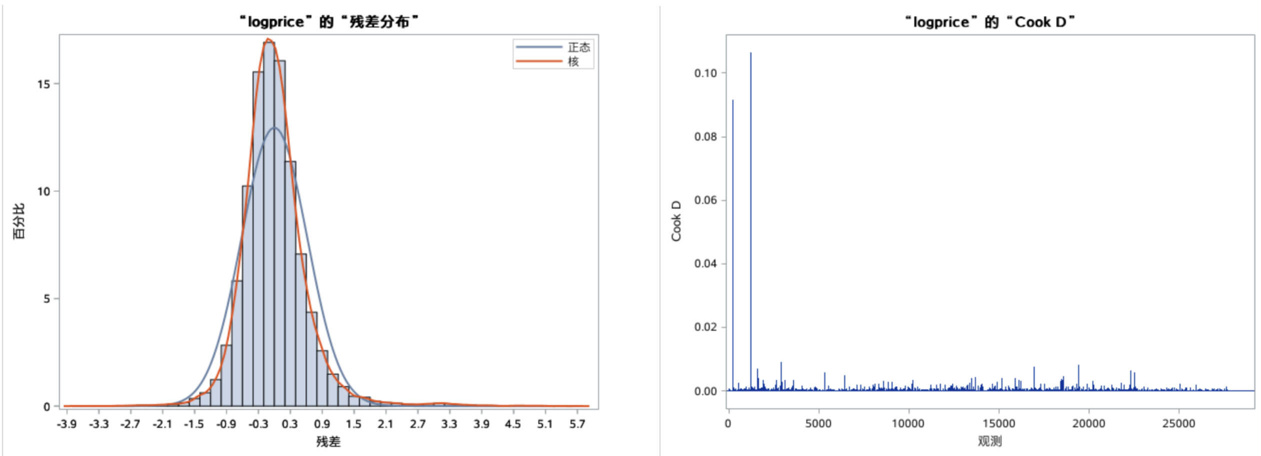
The residual plot of linear model after Box-Cox transfromation is shown as follows



(a) Residuals (b) Studentized residuals

Figure 14. Residual Plot

Except the prediction at a low levels, the distribution of the residuals are almost at random. The distribution of residual is approximately normal.



(a) Distribution of Residual (b) Cook's D

Figure 15. Model Diadnosis

The Cook distance is a common distance used in statistical analysis to diagnose the presence of abnormal data in various regression analyses. A large Cook distance indicates a fundamental change in coefficients after cases are excluded from regression statistics and calculations.

Only two observation points are with large Cook's D, which are also smaller than 0.5. So we can assume that there are few strong points in the data

4.1.4 Variable Selection

Therefore, we select the 18 variables mentioned above to perform stepwise linear regression, and the regression result is shown in the figure. The R^2_{adj} reaches 0.6198, and one variable need to be delete. According to the order of entry of variables, we consider the maximum capacity of the house, the infrastructure of the house, and the location and room type of the house to be the most important characteristics.

逐步选择汇总										
步	进入的效应	删除的效应	引入效应数	引入参数个数	调整R方	AIC	CP	SBC	F 值	Pr > F
0	Intercept		1	1	0.0000	23568.6943	45047.4865	-4054.079	0.00	1.0000
1	accommodates		2	16	0.4760	5730.6547	10467.3123	-21768.719	1673.82	<.0001
2	Basic		3	17	0.5062	4087.5638	8269.0348	-23403.584	1694.01	<.0001
3	bedrooms		4	39	0.5320	2630.6760	6416.7675	-24679.486	70.01	<.0001
4	neighbourhood		5	54	0.5520	1436.7919	4974.9342	-25749.971	83.26	<.0001
5	room_type		6	56	0.5684	406.7124	3785.1774	-26763.597	525.77	<.0001
6	Bathroom		7	57	0.5810	-409.8877	2873.3516	-27571.970	829.13	<.0001
7	bath		8	63	0.5922	-1155.3701	2063.5353	-28268.093	127.70	<.0001
8	property_type		9	76	0.6006	-1712.6512	1471.8930	-28718.428	45.22	<.0001
9	host_total_listings_		10	77	0.6074	-2188.1410	977.8910	-29185.691	480.30	<.0001
10	Checkin		11	78	0.6109	-2434.1109	725.6691	-29423.434	248.38	<.0001
11	has_availability		12	82	0.6133	-2602.2264	554.4941	-29558.643*	44.04	<.0001
12	minimum_nights		13	148	0.6162	-2740.2786	413.9592	-29153.739	4.09	<.0001
13	beds		14	181	0.6181	-2845.4355	308.6028	-28987.417	5.17	<.0001
14	description		15	182	0.6192	-2928.3359	225.9966	-29062.091	84.47	<.0001
15	host_is_superhost		16	183	0.6196	-2951.9548	202.5175	-29077.483	25.46	<.0001
16	Security		17	184	0.6198	-2965.3674	189.2003	-29082.669	15.31	<.0001
17	Kitchen		18	185	0.6198	-2967.3021	187.2914	-29076.377	3.91	0.0480
18	neiover		19	186	0.6198*	-2968.9452*	185.6726*	-29069.794	3.62	0.0571

Figure 16. Stepwise Selection

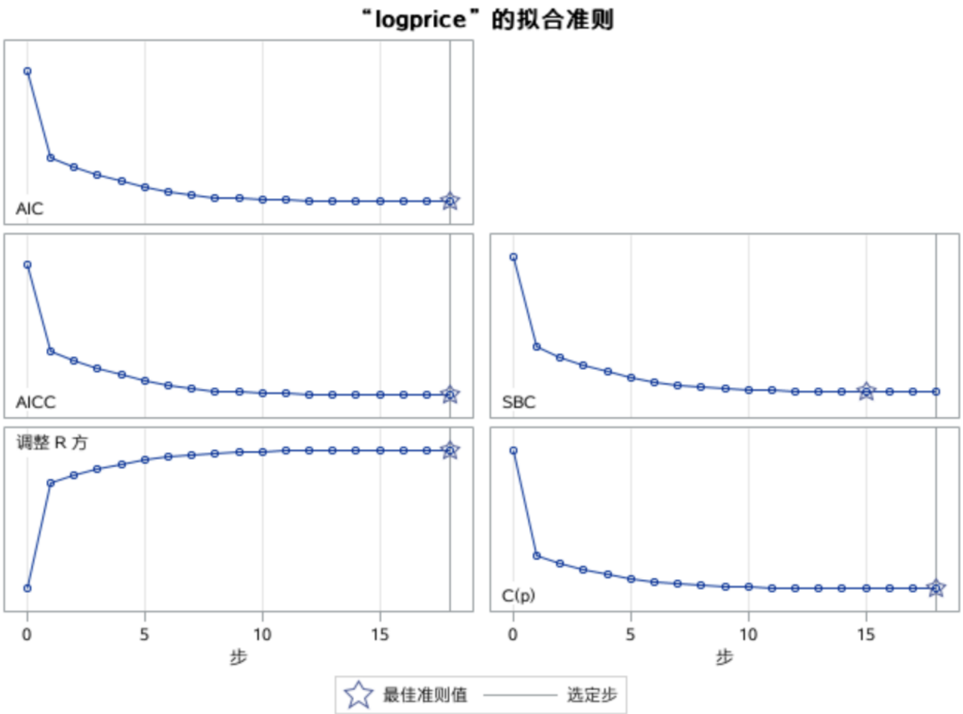


Figure 17. Criterion Statistics

4.2 Analysis of Review Scores Rating

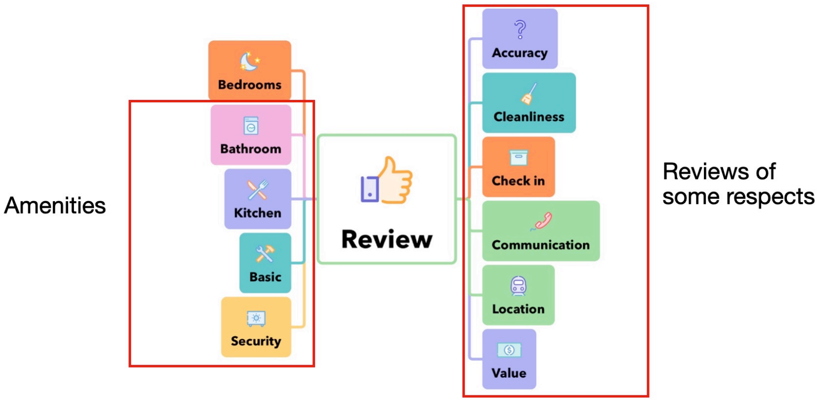


Figure 18. Choice of Explantory Variables

Next, we analyzed the data of the variable review _scores _rating.

By observing the remaining missing values, we find that the number of missing values in first review, last review and review score rating is exactly the same.

This is because these houses have not been successfully rented, so the missing value in this part has no impact on our study of user scores.

Secondly, we also found that the missing value of the remaining missing values was less than the missing value of the other scores. This is because if the tenant did not score, the remaining missing values would be zero and all the scores of these aspects would be missing. These missing values are also not within the scope of our study.

After deletion, 16,758 data were left for analysis.

4.2.1 About Review Score Rating

The house can be divided into three levels according to the review score when screening the house on Airbnb. Therefore, we can also divide the house into four levels: above 4.9, between 4.7 and 4.9, between 4.5 and 4.7 and below 4.5, according to the review score similar to that on Airbnb.

The distribution of the review score is shown as follows

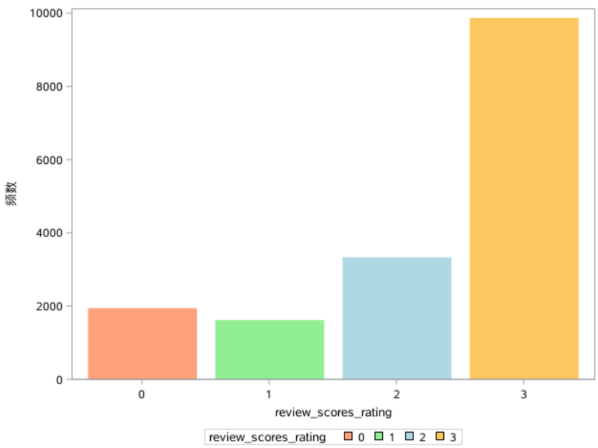


Figure 19. Distribution of the Review Score Rating

It can be found that more than half of tenants score 4.9 or above.

4.2.2 Ordinal Multinomial Logistic Regression

In the study of satisfaction with tenants to check in, we have three categories explained variable, and every aspect of house to other evaluation, including cleanliness, location and price satisfaction, such as the complement of house amenties, here we don't have the check-in amenties score, because tenants to check in score can better show this part information. Finally, there is the number of bedrooms, an explanatory variable used to give information about the size of the house.

We divide the dataset into two subsets, with proportion of 3:1, as the training set and testing set.

The result of logistic regression is shown as follows

最大似然估计分析						
参数		自由度	估计	标准误差	Wald 卡方	Pr > 卡方
Intercept	3	1	-87.4813	1.3626	4121.9900	<.0001
Intercept	2	1	-85.1033	1.3424	4018.8264	<.0001
Intercept	1	1	-82.5187	1.3135	3946.6442	<.0001
review_scores_accura		1	3.0693	0.1593	371.2652	<.0001
review_scores_cleanl		1	3.9044	0.1223	1019.4575	<.0001
review_scores_checki		1	3.3813	0.1953	299.7389	<.0001
review_scores_commun		1	3.0706	0.2079	218.1517	<.0001
review_scores_locati		1	0.6485	0.1192	29.5876	<.0001
review_scores_value		1	3.9052	0.1277	935.5228	<.0001
bath	1 bathroom	1	-0.3025	0.0650	21.6586	<.0001
bath	1 private bathroom	1	0.1340	0.0768	3.0447	0.0810
bath	1 share bathroom	1	0.2299	0.0851	7.2936	0.0069
bath	2 bathroom	1	-0.1037	0.0879	1.3895	0.2385
bath	3 bathroom	1	-0.0700	0.1462	0.2294	0.6320
bath	4 or more	1	0.1100	0.1511	0.5298	0.4667
Basic		1	-0.3611	0.2449	2.1744	0.1403
Bathroom		1	-0.4037	0.2175	3.4435	0.0635
Kitchen		1	0.3138	0.2032	2.3852	0.1225
Security		1	-0.0309	0.0751	0.1692	0.6809
bedrooms		1	0.0851	0.0277	9.4356	0.0021

Figure 20. Logistic Regression

This is the parameter estimation and significance result of each variable. We find that most variables are significant, but the degree of complete safety amenities is not significant. We believe that this may be due to the fact that most houses have complete safety amenities, which leads to the effect similar to Simpson's paradox.

And the performance of logistic model on the training set and testing set are respectively

预测概率和观测响应的关联			
一致部分所占百分比	92.9	Somers D	0.860
不一致部分所占百分比	6.9	Gamma	0.862
结值百分比	0.2	Tau-a	0.509
对	47050692	c	0.930

(a) Train

预测概率和观测响应的关联			
一致部分所占百分比	92.7	Somers D	0.855
不一致部分所占百分比	7.1	Gamma	0.857
结值百分比	0.2	Tau-a	0.503
对	5061843	c	0.928

(b) Test

Figure 21. Performance of Logistic Model on the Training Set and Testing Sets

Our logistic regression has a very good accuracy rate in both the training set and the test set, and there is no over-fitting phenomenon to a large extent, indicating that the variables we selected have a good explanation for the total score.

5 Conclusion

5.1 Analysis of Price

In general, all the host, the degree of complete amenities of the house, the type of house, the popularity of the house and the capacity of the house have a significant impact on the price.

As far as hosts are concerned, verified host and super host usually set a high price, and host who are more active in introducing their listing will also set a relatively higher price. The prices of all kinds of houses with complete amenities will also increase, and the degree of complete basic amenities has the greatest impact on the price of house.

For the type of house, in terms of overall price, entire house is greater than that of private room and greater than the shared room, and among building types, upscale buildings such as campsites and villas are more expensive; shabby buildings such as tents and farmhouses are less expensive. In terms of popularity, high-priced listings are often unavailable within 30 days and available within 60 days; listings in Huangpu and Xuhui districts are overpriced; and listings in Jiading and Fengxian districts are underpriced. In terms of house capacity, the larger the capacity and the more the number of bedrooms, the price shows an increasing trend.

According to stepwise feature entry order, we believe that the maximum capacity of the house, the infrastructure of the house, and the geographical location and room type of the house are the most important features.

5.2 Analysis of Scores

In analyzing the factors that influence the score, we divided the score into four levels. After that, logistic regression was conducted on training set and the model was tested with the testing set. The accuracy of prediction results is over 90%, which is almost the same with that on the training set, indicating no overdispersion.

In general, the scores of various aspects, the degree of completeness of amenities and the number of bedrooms have a significant impact on the overall evaluation of users, and almost all aspects have a positive impact on the satisfaction of tenants.

In terms of completeness of amenities, kitchen amenities are the most one which can enhance the satisfaction of the tenants with the occupancy, while tenants are less concerned about security amenities, because most properties have relatively complete security amenities.

Although the number of bedrooms has a significant effect on occupancy satisfaction, an increase in the number of bedrooms has less improvement in occupancy satisfaction.

A Appendix: Boxplots