Predicting Nightly Price of Airbnb Listings in Toronto, Canada

Milestone Report

George Tang

1 Introduction

Airbnb, Inc. is an online marketplace and hospitality service brokerage company. What began as an idea of putting an air mattress in the living room and turning it into to bed-and-breakfast in 2007 "for a few bucks" has grown into an international business with annual revenue of over $2 billion as of 2017. Members (hosts) uses the company's platform to list their properties to provide accommodation services, in which Airbnb receives commissions from each booking.

One biggest decision that hosts need to make is setting the prices for their listings. Hosts marking their prices too high or too low may risk driving potential customers away or shortchanging themselves. On the other hand, hosts that set prices based on the properties' locations and features along with competitors' prices can fully leverage the properties' true value and maximize their revenues.

While hosts can spend hours searching the Airbnb's website to get a reference rate, it is time-consuming and often it is difficult to identify properties with features similar to the hosts' in their vicinity. In this project, we will analyze the historical listing information with data analytics, from which factors that affect price will be identified. We will then build machine learning model to predict the listing prices based on input such as host information, properties' features and booking policy.

1.1 Objective

The objectives of is project are:

- To explore and analyze Airbnb's listings in Toronto, Canada
- To identify features that affect the prices of a nightly stay
- To develop machine learning models that predict the prices of a nightly stay based on relevant features

In this Milestone Report, we will present a description of the data set (Section 2), the steps for data cleaning and wrangling (Section 3), develop a data story which includes exploratory data analysis (EDA) and statistical analysis (Section 4), and outline the next steps of this project (Chapter 5).

The Python codes used for this report can be found here.

1.2 Significance

Through this project, we will take a deep dive into Toronto's Airbnb listing, and with this effort, we will identify the important features that determine pricing, which the hosts can use as a reference to renovate their properties or booking policy to make them more attractive to customers. We will also develop

machine learning models that can be used by the hosts to set fair and competitive prices and ultimately maximize their revenues.

2 Dataset

2.1 Airbnb Listings Data

The dataset is obtained from the website Inside Airbnb. It is an independent, non-commercial website that allows users to explore how Airbnb is used in cities around the world. The dataset used in this project, referred to as "listings" thereafter, was collected on June 4, 2019.

The listings dataset consists of 20,769 listings (row), and 106 features (columns). Each row consists of a listing in the Greater Toronto area on June 4, 2019.

The features are divided into the following 7 subcategories:

1. Host information
2. Property information
3. Booking information and policy
4. Availability
5. Reviews
6. Airbnb listing information
7. Web scraping information

A list of the features is shown in Appendix I. Most of the feature names are self-explanatory.

2.2 Toronto Geographical Information

Two Wikipedia pages (here and here) provides the information that links the listings' postal codes to their city names. The use of this information will be discussed in Section 3.4.1.

2.3 Mapquest API

As discussed in Section 3.4.1, some listings come with ambiguous geographical information. As such, their geographical information are obtained through Mapquest's API with their latitudes and longitudes as input.

3 Data Cleaning and Wrangling

The purpose the data cleaning and wrangling steps are:

1. To ensure the all features are of the correct data type
2. To ensure missing data are properly imputed
3. To create potentially useful features
4. To prepare the dataset for EDA and statistical analysis

3.1 Data Type Correction

The numerical features price, security deposit, cleaning fee, charge for extra people and host response rate are stored as string in the dataset, and as such, their date types are converted to numeric.

The datetime features last scraped, host since, calendar last scraped, first review and last review are stored as string in the dataset, and as such, their data types are converted to datetime.

3.2 Incorrect Price Data Elimination

The feature price is the focus of this study, and as such, its data integrity is of utmost importance.
We find that there are four (4) listings with price of 0, which indicates data issue. Further investigation of the listings' website indicates that the prices are indeed non-zero; nonetheless, for simplicity, those listings are dropped from the dataset.

3.3 Missing Values Imputation

A list of features with missing values is shown in Appendix II. Overall, 54 of the 106 features consist of missing values, with counts from 1 (0.005%) to 20,769 (100%).

Only features that are considered potentially useful for data analysis will be imputed. For imputation of each feature, a new feature with name [variable]_NA is created to record the listings that originally consist of missing value. It may be useful if the reason for missing is systemic (i.e. non-random).

3.3.1 Numeric Features

The numeric features host listings count, number of bathrooms, host response rate, bedrooms and beds are imputed with their respective medians.

Missing security deposit and cleaning fee are due the hosts' decision to not include one, which is equivalent to a value of 0. As such, the missing values will be imputed with 0.

For review scores, the missing values are likely due to the facts that either the listings are new with few customers, or their customers did not leave a review score. The missing values are imputed with the feature median values.

3.3.2 Categorical Features

The categorical feature host response time is imputed with the feature mode.

3.3.3 Datetime Features

The datetime features host since , first review, and last review are imputed with feature medians.

3.4 New Feature Creation

3.4.1 City Names

The dataset consists of two features, namely neighbourhood and cleansed neighbourhood, that provide the name of the neighhourhood for each listing. There are 140 unique values for each feature, which may be too granular for data analysis. Instead the city name for each listing may be more appropriate. The information is obtained with the feature zipcode, which is the postal code of the listing. Canada's postal code consists of six characters, with the first three characters known as the Forward Sortation Area (FSA). The FSA is than matched with the list of cities as discussed Section 2.2.

For listings with erroneous FSA or unknown city names, their city names are obtained with Mapquests API as discussed in Section 2.3, with the listing's longitude and latitude information as input.

After these steps, there are two listings with missing city information. Additionally, the number of listings for the cities of Thornhill (7), Mississauga (2), Pickering (2), and Markham (1) are exceptionally low. Since all of those are cities of considerable sizes, it is likely that only a small fraction of their listings is included in in this dataset, which make the information non-generalizable. As such, we decided to remove the listings from these cities.

3.4.2 Indicator Variable for Amenities

The feature "amenities" contains a list of attributes provided by the host that the property contains. It is stored as a string enclosed in curly brackets with each attribute separated by comma. To further evaluate those amenities, a dummy variable is created for each amenity, with '1' and '0' indicating the presence and absence of that amenity, respectively.

A list of amenities along with the percentage of listings with each amenity is shown in Appendix III. In total, there are 196 unique amenities. The rarest amenities are tennis court, brick oven, pool toys and hammock, each only available in 1 listing, while the most common amenities are wifi, heating, essentials, and smoke detector. The amenities should be used with caution nonetheless because the information may not be complete. For instance, hot water should be provided in most properties; yet, only 59% of listings are listed as providing hot water. It is possible that this amenity is so trivial that hosts did not bother to include them.

3.4.3 Days since Reference Day

The number of days since the recorded events can be a feature more useful than the dates. A reference date of 2019/6/27, which is the date the listings data was scrapped, is chosen.

4 Data Storytelling

4.1 Price

As summarized in Table 4-1, the prices of a nightly stay show considerable variation. While the least expensive listing is $13, the most expensive listing is $13,422. To further investigate, this expensive offer is an "Art Collector's House" that "will have you living luxuriously just steps from Toronto's most stylish neighbourhood".

Table 4-1 Summary of price and log price distribution

Mean	143.35	4.66
Standard Deviation	234.24	0.72
Minimum	13.00	2.56
25%	64.00	4.16
50%	101.00	4.62
75%	160.60	5.08
Maximum	13,422	9.50

A Q-Q plot of the price, as shown in Figure 4-1(a), shows that the data is highly right skewed. To reduce the influence of those outliers which are crucial for data visualization and statistical analysis, we apply log transform to the price data, with the Q-Q plot shown in Figure 4-1(b). While the data is still right-skewed, it is considerably more normally distributed, which is also shown in Figure 4-2. It is also worth notice that the distribution has a short tail on the left, which means the minimum is higher than what a normal distribution would predict.

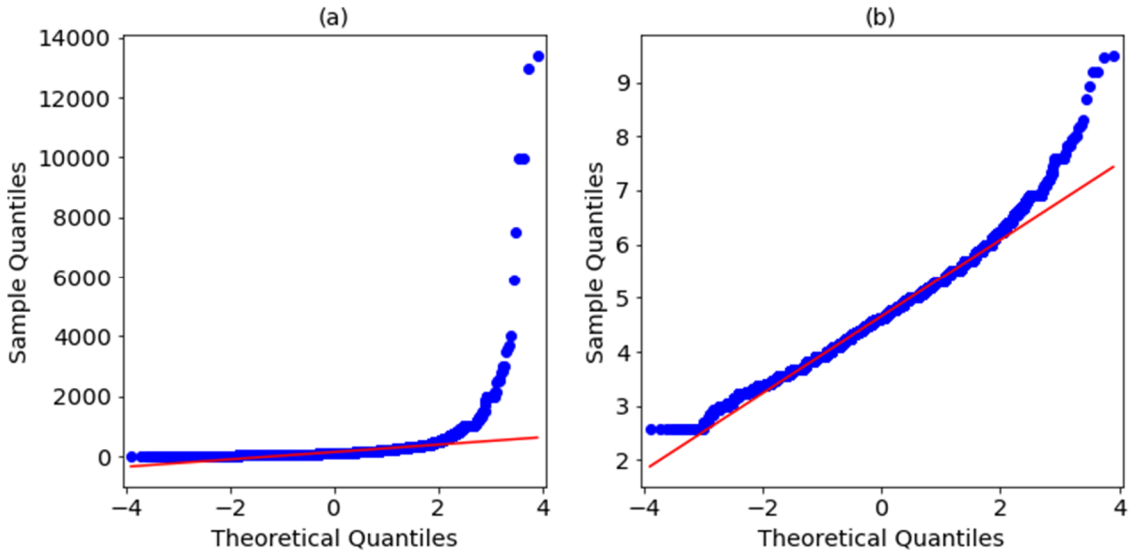

Figure 4-1 Quantile-quantile plot of (a) price; (b) log price

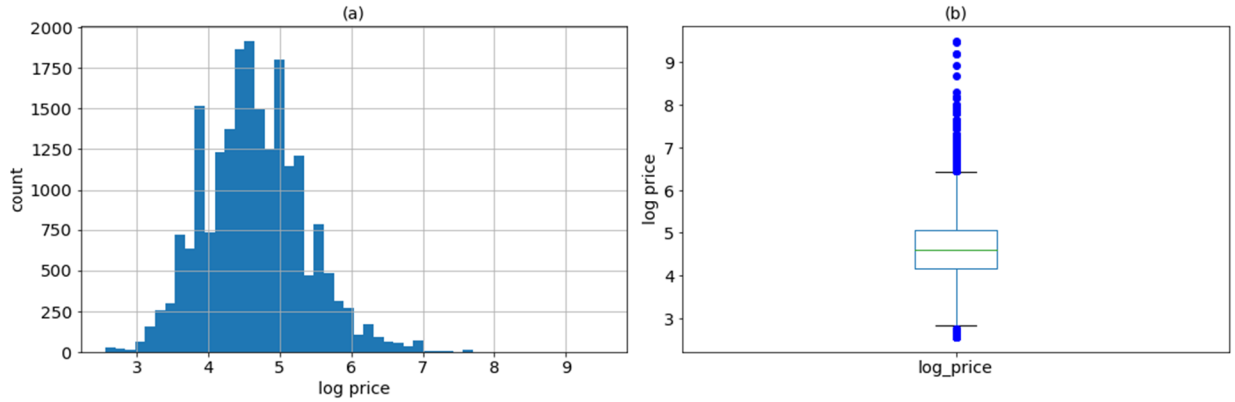

Figure 4-2 Distribution of log price: (a) histogram, (b) boxplot

4.2 Geographical Information

The count of listings in each region is shown in Figure 4-3(a). Overall, Toronto (Downtown, Central, East and West) consists of the most combined listings, while among them, Downtown Toronto has the most listing. The York region (York, North York and East York) has the 2nd most combined listings.

The listings price show considerable variable at each region. Downtown Toronto has the highest median price while Scarborough (a suburb northeast of Toronto) has the lowest median price.

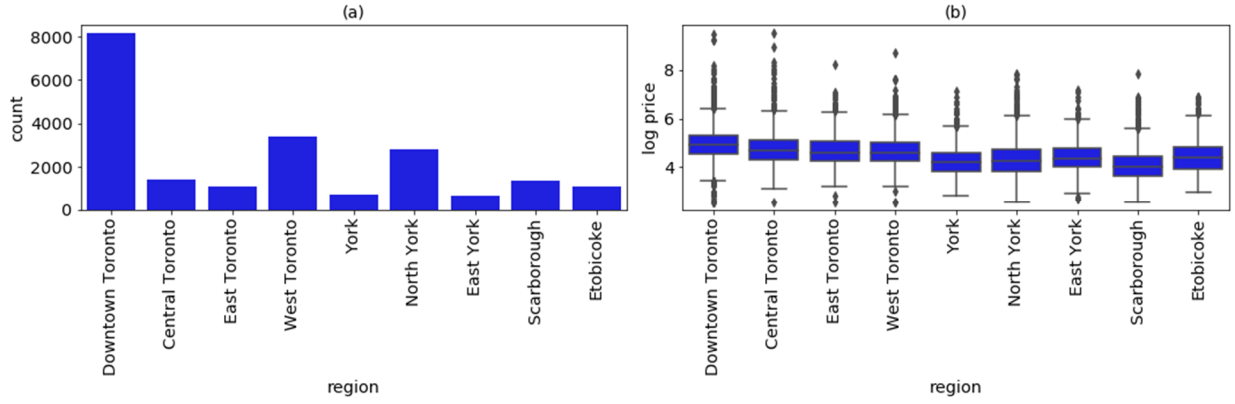

Figure 4-3 (a) Count of listings and (b) boxplot of log price in each region

4.3 Property Information

4.3.1 Property Type

The count of listings for each property type is shown in Figure 4-4.Overall, there are 30 different property types, with 16 of them with counts less than 10. The most common are Apartment, followed by Condominium and House.

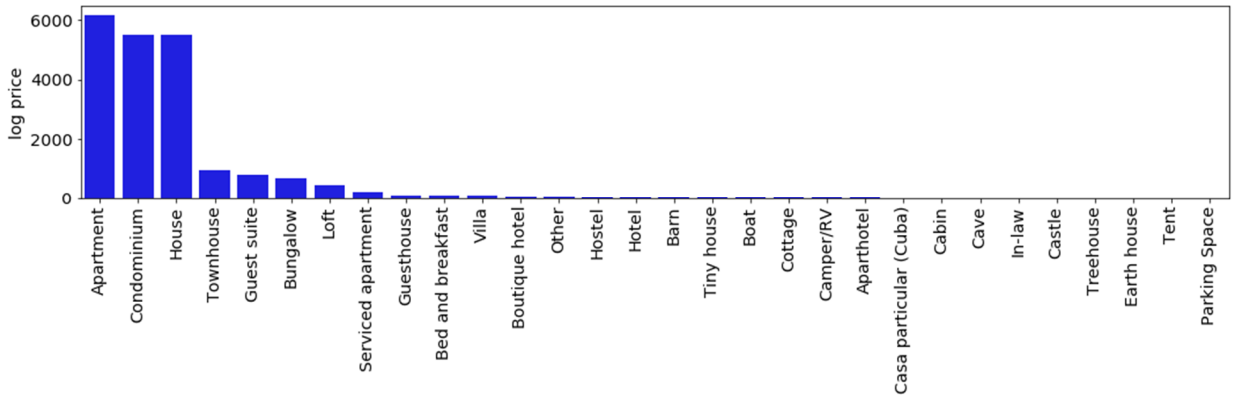

Figure 4-4 Count of listings for each property type

Bungalow is an interesting property type. According to Wikipedia: "Canada uses the definition of bungalow to mean a single-family dwelling that is one storey high". In other words, a bungalow is essentially a house. As such, this property type is assigned a value of "House".

The rest of the property types are assigned a value of "Other" to reduce granularity. Figure 4-5 shows that condominium has the highest median price while house has the lowest median price.

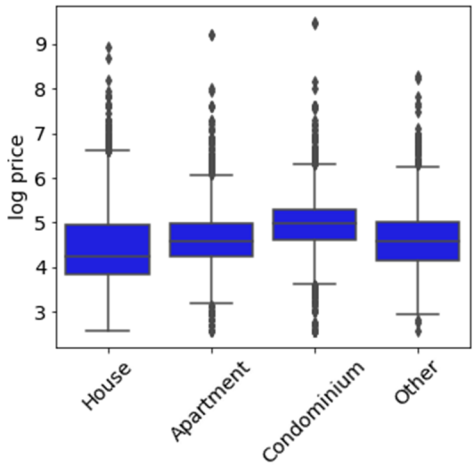

Figure 4-5 Distribution of log price for each property type

4.3.2 Room Type

The count of room type for each property type is shown in Figure 4-6(a). For houses, the most common room type is private room while for both apartment and condominium, the most common room type is entire home/apartment. Shared room is the least common for all property types.

As shown in Figure 4-6(b), the median price is the highest for the room type of entire home/apartment and lowest for shared room.

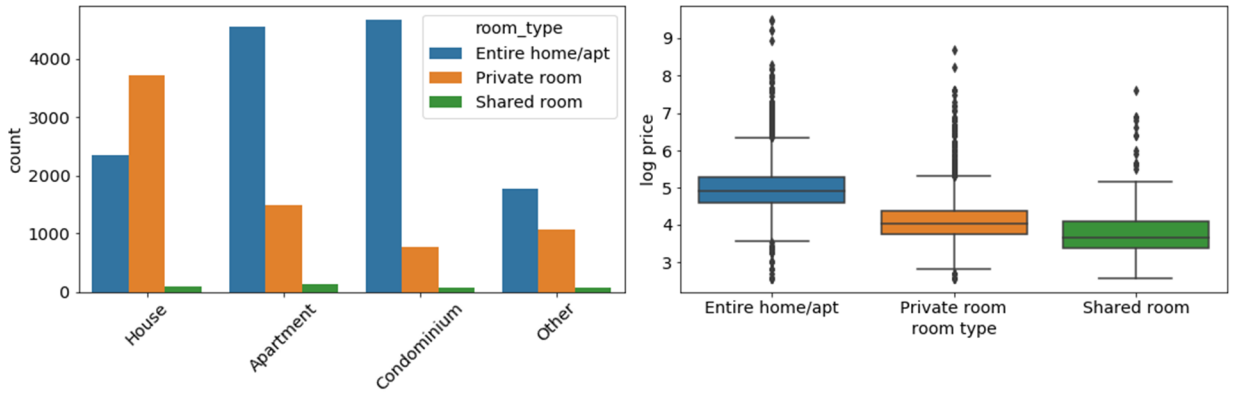

Figure 4-6 (a) Count of room type and (b) price distribution for each property type

4.3.3 Property Features

Features that fall into this category include accommodates, bathrooms, bedrooms, and beds, which are all numerical. A heat map of the correlations between the features and log price is shown in Figure 4-7. All features are somewhat correlated with log price. It is important to notice that the features are correlated with each other which is expected.

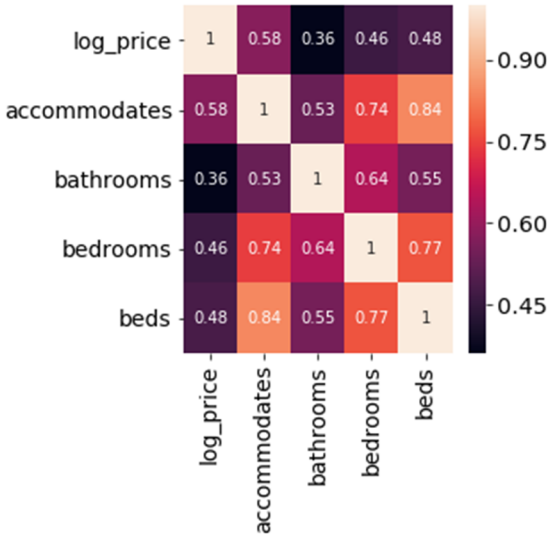

Figure 4-7 Correlations between property features and log price

4.3.4 Amenities

With 196 unique amenities, for this analysis, we pick the ones that (1) are intuitively non-trivial and may affect the price, and (2) have a balanced class, with we define as the majority class being less than 90% of the data. The chosen amenities are “bathtub”, “pets allowed”, “pool”, “gym”, “family/kid friendly”, “private entrance”, “free parking on premises”, and “air conditioning”.

As shown in Figure 4-8, for the eight chosen amenities, the median price is higher for listings with the amenity.

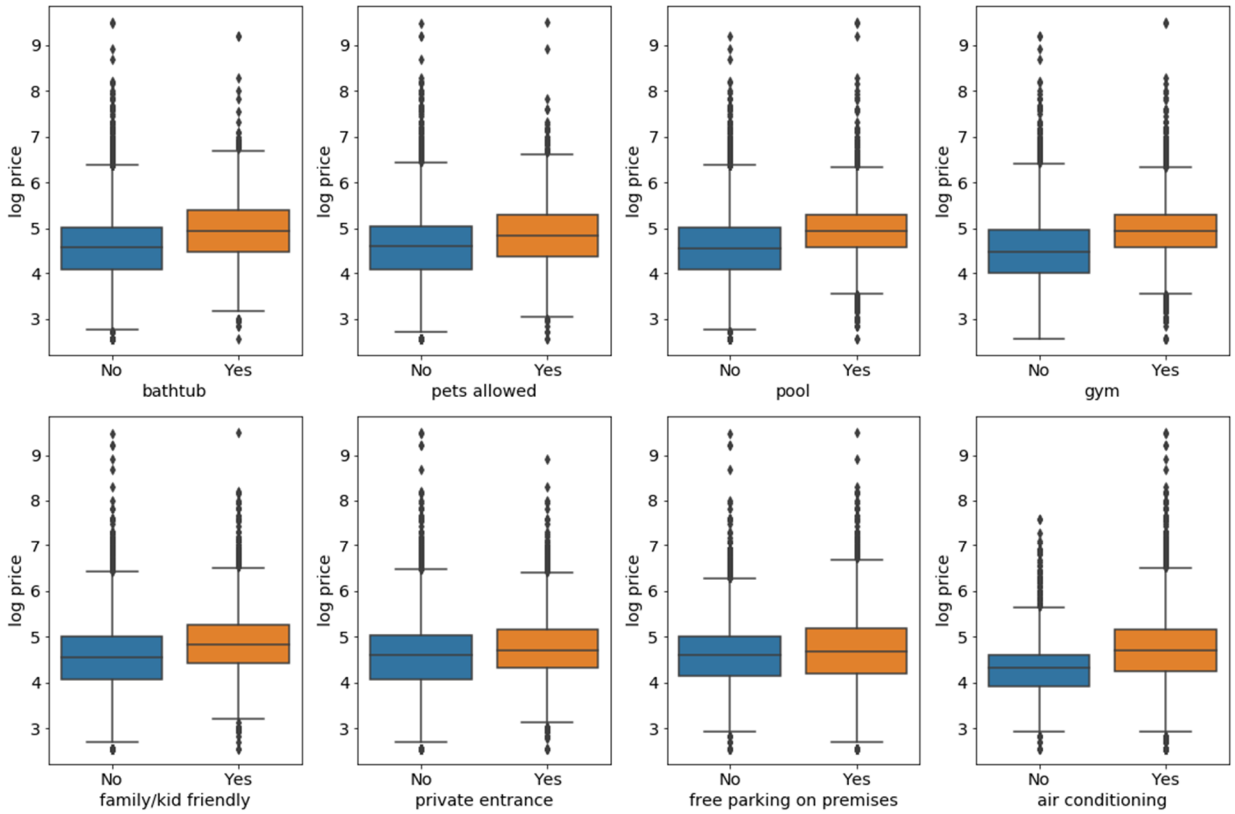

Figure 4-8 Distribution of log price for listings with and without amenities

4.4 Host Information

4.4.1 Superhost Status

According to Airbnb: "[s]uperhosts are experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests". As such, it is possible that a superhost status would command a higher price due to their good track records and reputations.

As shown in Figure 4-9(a), about 35% of listings are provided by superhosts. Overall, the median price is slightly higher for listings provided by superhosts (Figure 4-9(b)); however, the most expensive listings are provided by non-superhosts.

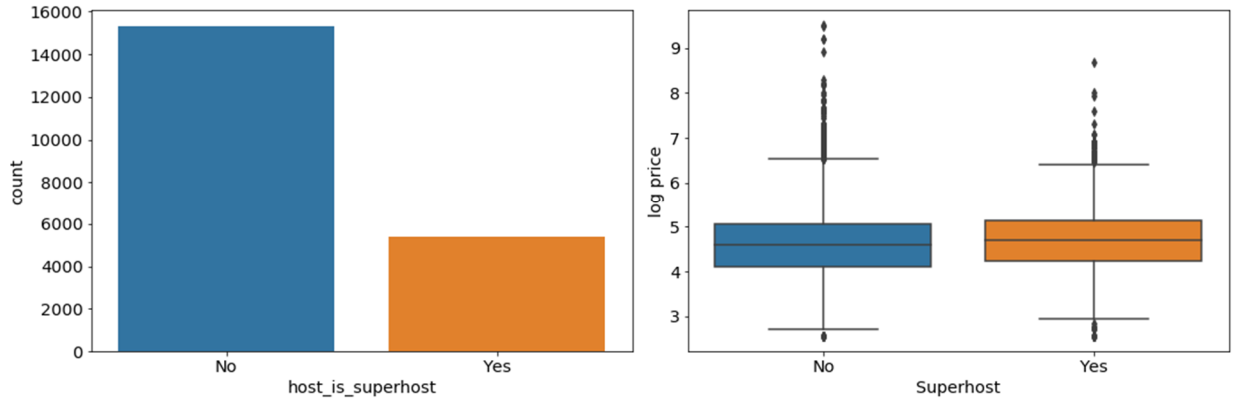

Figure 4-9 (a) Count of listings and (b) distribution of log price for superhost status

4.5 Booking Policy

4.5.1 Cleaning Fee

Cleaning fee is a one-time, non-refundable fee charged by the host, regardless of the duration of stay. It is an interesting feature because it can be part of the overall pricing strategy. From customers' point of view, a lower cleaning fee along with a higher nightly price may encourage shorter term stay, while a higher cleaning fee along with a lower nightly price may encourage longer term stay. Nonetheless, for this project, cleaning fee will be used solely as a predictor for price.

Due to the high skewness of cleaning fee, a log transformation is applied. As shown in Figure 4-10, there are two distinct populations of cleaning fee. On the left, there is no cleaning fee, either due to 0 or missing value. On the right, there is a positive correlation between log price and log cleaning fee.

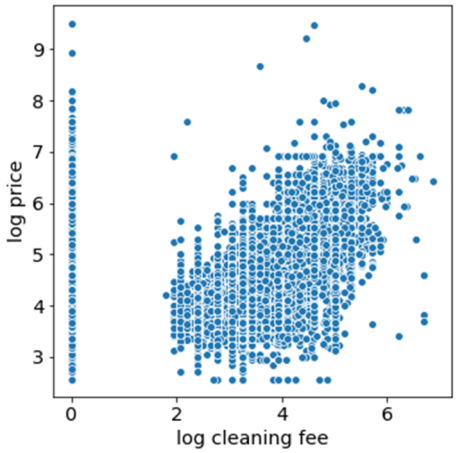

Figure 4-10 Log price vs log cleaning fee

4.5.2 Minimum Nights of stay

As shown in Figure 4-11(a), a majority of listings requires minimum nights of stay of 3 nights or less. There is no obvious difference among them in term of price (Figure 4-11(b)).

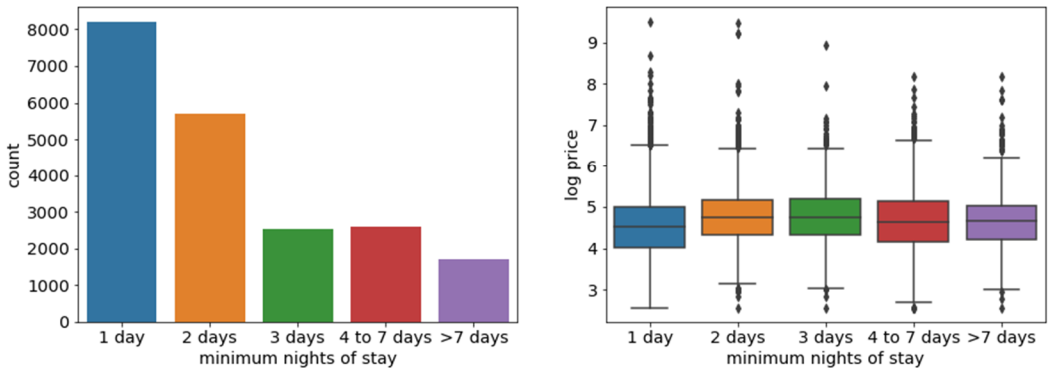

Figure 4-11 (a) Count of listings and (b) distribution of log price for minimum nights of stay

4.5.3 Numeric Features

Features that falling into this category include security deposit, cleaning fee, included number of guests, charge for extra people, minimum and maximum nights of stay.

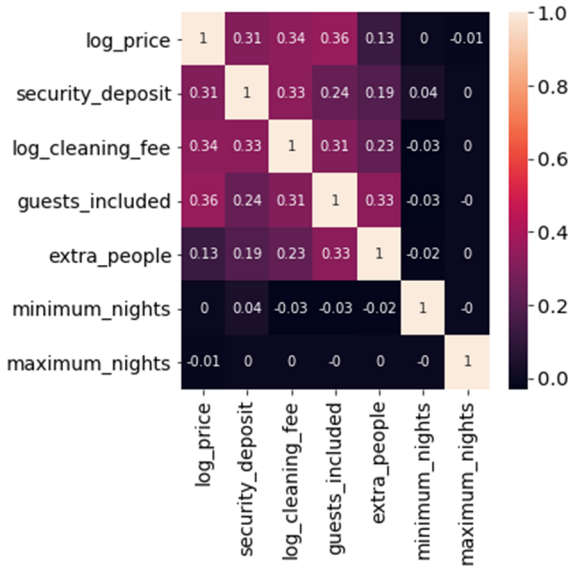

Figure 4-12 shows that security deposit, cleaning fees and included number of guests are correlated with log price.

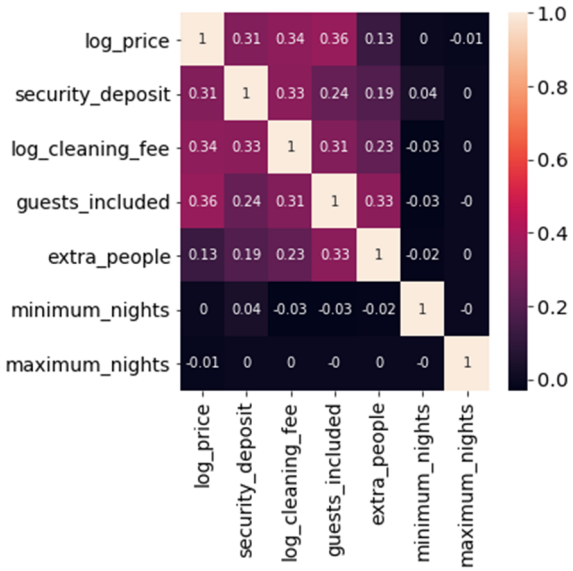

Figure 4-12 Correlations between booking policy features and log price

4.6 Availability

The dataset provides information on the listings' availabilities for the next 30, 60, 90 and 365 days. Figure 4-13 shows that while the availabilities are highly correlated with each other, they are not correlated with log price.

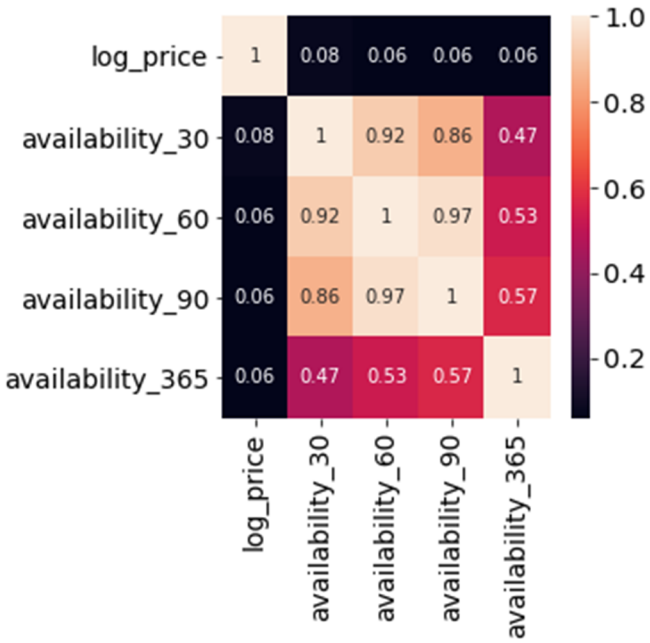

Figure 4-13 Correlation between availability and log price

4.7 Review Scores

Airbnb allows customers to to provide review scores of their experiences on various categories, including accuracy, cleanliness, check in, communication, and value. Also provided by the data are the number of reviews received by each listing, and the number of reviews of the last twelve months (ltm). Interestingly, Figure 4-14 shows that the neither the number of reviews nor review scores are correlated with log price.

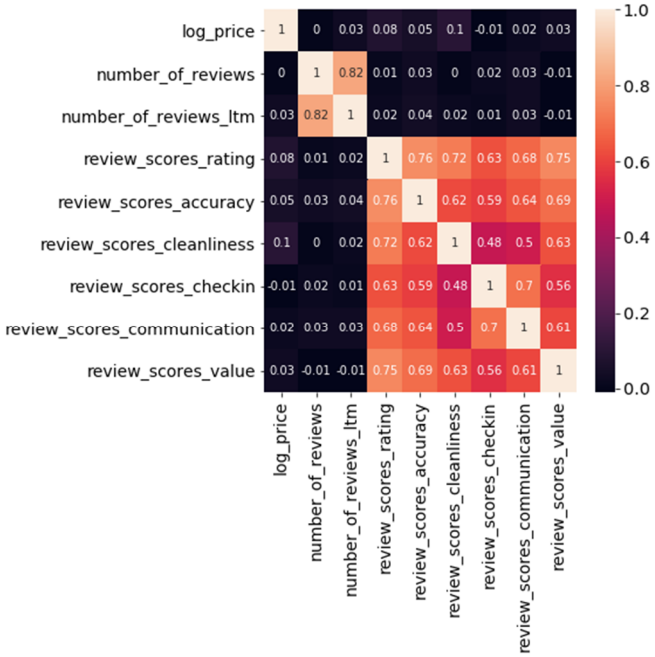

Figure 4-14 Correlation between review information and log price

4.8 Statistical Analysis of Categorical Features

In this section, we will conduct hypothesis tests on selected categorial features. The null hypothesis is that the distributions of log price are the same among all categories within a feature, where the alternative hypothesis is that the distributions are not the same. Since the log price is not normal, the Mann Whitney U (MWU) test will be used for features with binary response (1/0, or t/f, etc), where Kruskal-Wallis (KW) test will be used for multi-category response. The hypothesis tested by both MWU test and KW test is whether the samples from the different categories are taken from the same population. If they are not, then the feature may be of use to predict the price.

The results of the tests are shown in

Appendix IV. For all features, the null hypothesis is rejected, i.e. log price does not have time same distribution among all categories in each feature. It is important to note that, however, with such a large sample size (> 10,000), the null hypothesis will be rejected virtually all the time. As such, this hypothesis testing may not be the most useful. A better approach maybe to use machine learning models to identify features that are useful for predicting nightly price which will be the focus of the next step.

5 Conclusion and Next Steps

In this report, we have detailed the steps of data cleaning and wrangling. We have also explored the data through visualizations and statistical methods.

Features that shows promise in predicting prices include geographical location, property types and room types, number of accommodates, beds, bedrooms, cleaning fee, security deposit, number of guests included, and provision of amenities. It is important to note that features not correlated with price can still be useful of prediction because of interaction among features.

The next step of this project is to develop, evaluate and select machine learning models to predict price. Ultimately, the model can be used by hosts to set a reasonable price for their listings' prices based on their features.

Appendix I
Features List

Host information	Property Information	Booking information and policy
host_id	name	price
host_url	summary	weekly_price
host_name	space	monthly_price
host_since	description	security_deposit
host_location	experiences_offered	cleaning_fee
host_about	neighborhood_overview	guests_included
host_response_time	notes	extra_people
host_response_rate	transit	minimum_nights
host_acceptance_rate	access	maximum_nights
host_is_superhost	interaction	minimum_minimum_nights
host_thumbnail_url	house_rules	maximum_minimum_nights
host_picture_url	street	minimum_maximum_nights
host_neighbourhood	neighbourhood	maximum_maximum_nights
host_listings_count	neighbourhood_cleansed	minimum_nights_avg_ntm
host_total_listings_count	neighbourhood_group_cleansed	maximum_nights_avg_ntm
host_verifications	city	requires_license
host_has_profile_pic	state	license
host_identity_verified	zipcode	jurisdiction_names
calculated_host_listings_count	market	instant_bookable
calculated_host_listings_count_entire_homes	smart_location	is_business_travel_ready
calculated_host_listings_count_private_rooms	country_code	cancellation_policy
calculated_host_listings_count_shared_rooms	country	require_guest_profile_picture
	latitude	require_guest_phone_verification
	longitude	
	is_location_exact	
	property_type	
	room_type	
	accommodates	
	bathrooms	
	bedrooms	
	beds	
	bed_type	
	amenities	
	square_feet	

Availability	**Airbnb listing information**	**Reviews**	**Web scraping information**
calendar_updated	id	number_of_reviews	scrape_id
has_availability	listing_url	number_of_reviews_ltm	last_scraped
availability_30	thumbnail_url	first_review	
availability_60	medium_url	last_review	
availability_90	picture_url	review_scores_rating	
availability_365	xl_picture_url	review_scores_accuracy	
calendar_last_scraped		review_scores_cleanliness	
		review_scores_checkin	
		review_scores_communication	
		review_scores_location	
		review_scores_value	
		reviews_per_month	

Appendix II

List of Features with Missing Values

Feature	Missing Count	Missing Percentage
thumbnail_url	20765	100
medium_url	20765	100
host_acceptance_rate	20765	100
neighbourhood_group_cleansed	20765	100
xl_picture_url	20765	100
jurisdiction_names	20763	99.99036841
license	20761	99.98073682
square_feet	20609	99.24873585
monthly_price	18987	91.43751505
weekly_price	18678	89.94943414
notes	10967	52.81483265
host_about	8757	42.17192391
access	7986	38.45894534
interaction	7716	37.15868047
neighborhood_overview	7291	35.11196725
transit	7086	34.12472911
house_rules	6576	31.66867325
space	5774	27.80640501
host_response_time	4984	24.00192632
host_response_rate	4984	24.00192632
security_deposit	4902	23.60703106
review_scores_location	4287	20.64531664
review_scores_value	4284	20.63086925
review_scores_checkin	4282	20.62123766
review_scores_accuracy	4281	20.61642186
review_scores_communication	4279	20.60679027
review_scores_cleanliness	4279	20.60679027
review_scores_rating	4271	20.56826391
last_review	3982	19.17649892
reviews_per_month	3982	19.17649892
first_review	3982	19.17649892
cleaning_fee	3384	16.29665302
host_neighbourhood	2520	12.13580544
summary	664	3.197688418
zipcode	365	1.757765471
description	346	1.66626535
market	37	0.178184445

state	28	0.134842283
beds	23	0.110763304
host_location	19	0.09150012
bathrooms	15	0.072236937
bedrooms	8	0.038526366
host_identity_verified	5	0.024078979
host_has_profile_pic	5	0.024078979
host_total_listings_count	5	0.024078979
host_listings_count	5	0.024078979
host_picture_url	5	0.024078979
host_thumbnail_url	5	0.024078979
host_is_superhost	5	0.024078979
host_since	5	0.024078979
host_name	5	0.024078979
city	1	0.004815796
neighbourhood	1	0.004815796
name	1	0.004815796

Appendix III

List of Amenities

Amenity	Count	Percentage of Listings
brick oven	1	0.004819
pool toys	1	0.004819
tennis court	1	0.004819
hammock	1	0.004819
washer / dryer	2	0.009638
alfresco bathtub	2	0.009638
mobile hoist	2	0.009638
heat lamps	2	0.009638
private gym	2	0.009638
ground floor access	2	0.009638
pool cover	2	0.009638
ceiling hoist	3	0.014457
private pool	3	0.014457
touchless faucets	3	0.014457
standing valet	3	0.014457
air purifier	4	0.019276
fax machine	4	0.019276
outdoor kitchen	4	0.019276
mountain view	5	0.024095
projector and screen	5	0.024095
private hot tub	5	0.024095
bidet	6	0.028914
steam oven	6	0.028914
sauna	6	0.028914
stand alone steam shower	7	0.033733
private bathroom	7	0.033733
heated towel rack	7	0.033733
jetted tub	7	0.033733
fire pit	8	0.038552
double oven	8	0.038552
beach view	8	0.038552
wine cooler	10	0.04819
mudroom	11	0.053009
amazon echo	11	0.053009
shared hot tub	11	0.053009
high-resolution computer monitor	11	0.053009
ski-in/ski-out	11	0.053009
electric profiling bed	12	0.057829
murphy bed	14	0.067467

warming drawer	15	0.072286
sun loungers	15	0.072286
mini fridge	17	0.081924
hbo go	17	0.081924
shared pool	18	0.086743
firm mattress	19	0.091562
outdoor parking	20	0.096381
printer	21	0.1012
dvd player	21	0.1012
pool with pool hoist	22	0.106019
shared gym	26	0.125295
day bed	29	0.139752
exercise equipment	32	0.154209
heated floors	33	0.159028
gas oven	34	0.163848
shower chair	37	0.178305
kitchenette	39	0.187943
ceiling fan	39	0.187943
bathtub with bath chair	42	0.2024
table corner guards	42	0.2024
fixed grab bars for toilet	46	0.221676
pillow-top mattress	48	0.231314
terrace	50	0.240952
sound system	53	0.255409
formal dining area	54	0.260228
rain shower	54	0.260228
espresso machine	56	0.269867
other pet(s)	57	0.274686
memory foam mattress	57	0.274686
soaking tub	57	0.274686
convection oven	69	0.332514
balcony	69	0.332514
walk-in shower	72	0.346971
central air conditioning	74	0.356609
baby monitor	74	0.356609
outdoor seating	76	0.366247
en suite bathroom	78	0.375885
breakfast table	91	0.438533
beach essentials	92	0.443352
fireplace guards	100	0.481904
fixed grab bars for shower	104	0.501181
smart tv	129	0.621657
beachfront	137	0.660209

netflix	141	0.679485
roll-in shower	145	0.698762
changing table	147	0.7084
window guards	151	0.727676
hot water kettle	161	0.775866
ev charger	181	0.872247
baby bath	182	0.877066
stair gates	195	0.939714
outlet covers	232	1.118018
pocket wifi	299	1.440894
game console	299	1.440894
babysitter recommendations	313	1.508361
toilet paper	328	1.580647
bath towel	328	1.580647
body soap	328	1.580647
bedroom comforts	333	1.604742
bathroom essentials	333	1.604742
wide clearance to shower	387	1.86497
toilet	387	1.86497
crib	396	1.908342
full kitchen	400	1.927618
children’s dinnerware	404	1.946894
cat(s)	420	2.023999
disabled parking spot	421	2.028818
dog(s)	473	2.279408
suitable for events	480	2.313142
extra space around shower and toilet	509	2.452894
wide doorway to guest bathroom	564	2.717941
waterfront	631	3.040817
accessible-height toilet	658	3.170932
smoking allowed	673	3.243217
high chair	755	3.638379
accessible-height bed	778	3.749217
building staff	831	4.004626
children’s books and toys	867	4.178112
pack ’n play/travel crib	877	4.226302
wide entryway	908	4.375693
smart lock	915	4.409426
cleaning before checkout	933	4.496169
wide entrance	1032	4.973254
extra space around bed	1050	5.059997
pets live on this property	1076	5.185292
room-darkening shades	1094	5.272035

doorman	1142	5.503349
flat path to guest entrance	1160	5.590092
wheelchair accessible	1165	5.614187
wide entrance for guests	1308	6.303311
wide hallways	1308	6.303311
ethernet connection	1312	6.322587
single level home	1375	6.626187
lake access	1378	6.640644
buzzer/wireless intercom	1542	7.430967
other	1627	7.840586
bbq grill	1646	7.932148
breakfast	1850	8.915233
well-lit path to entrance	1938	9.339309
24-hour check-in	1958	9.43569
free street parking	2015	9.710375
indoor fireplace	2051	9.883861
keypad	2136	10.29348
translation missing: en.hosting_amenity_49	2251	10.84767
no stairs or steps to enter	2337	11.26211
bathtub	2475	11.92714
private living room	2592	12.49096
paid parking on premises	2599	12.5247
pets allowed	2604	12.54879
garden or backyard	2660	12.81866
safety card	2702	13.02106
translation missing: en.hosting_amenity_50	2799	13.48851
hot tub	3115	15.01132
host greets you	3239	15.60889
lockbox	3359	16.18717
pool	3819	18.40393
luggage dropoff allowed	3913	18.85692
patio or balcony	4587	22.10496
cable tv	4735	22.81818
paid parking off premises	4759	22.93383
internet	4998	24.08559
extra pillows and blankets	5835	28.11913
dishwasher	5875	28.31189
lock on bedroom door	6244	30.09012
gym	6295	30.33589
long term stays allowed	6471	31.18404
family/kid friendly	6560	31.61293
private entrance	7006	33.76223
coffee maker	7184	34.62002

self check-in	7229	34.83688
bed linens	7875	37.94998
first aid kit	7924	38.18611
cooking basics	8121	39.13546
oven	8174	39.39087
free parking on premises	8327	40.12819
stove	8521	41.06308
microwave	8671	41.78594
elevator	8671	41.78594
dishes and silverware	8700	41.92569
refrigerator	9431	45.44841
fire extinguisher	10742	51.76618
hot water	12253	59.04776
iron	14644	70.57009
tv	14662	70.65684
hair dryer	15306	73.7603
laptop friendly workspace	15611	75.23011
dryer	16542	79.71664
shampoo	16602	80.00578
carbon monoxide detector	16655	80.26119
washer	16797	80.9455
hangers	17590	84.767
air conditioning	17728	85.43203
kitchen	19127	92.17387
smoke detector	19520	94.06776
essentials	19736	95.10867
heating	20091	96.81943
wifi	20366	98.14467

Appendix IV

Hypothesis Test Results for Categorical Features

```
Feature: property_type_simple
Categories: ['House', 'Apartment', 'Condominium', 'Other']
Test: Kruskal-Wallis Test
The test statistics is 2314.7076216390383 and the p-value is 0.0.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.


Feature: room_type
Categories: ['Entire home/apt', 'Private room', 'Shared room']
Test: Kruskal-Wallis Test
The test statistics is 8741.637250440977 and the p-value is 0.0.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.


Feature: bed_type
Categories: ['Real Bed', 'Futon', 'Pull-out Sofa', 'Airbed', 'Couch']
Test: Kruskal-Wallis Test
The test statistics is 112.18115634515216 and the p-value is 2.49307793005
0478e-23.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.


Feature: city_fsa
Categories: ['Downtown Toronto', 'West Toronto', 'North York', 'Central To
ronto', 'Scarborough', 'Etobicoke', 'East Toronto', 'York', 'East York']
Test: Kruskal-Wallis Test
The test statistics is 3470.9880841479116 and the p-value is 0.0.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.


Feature: cancellation_policy
Categories: ['strict_14_with_grace_period', 'moderate', 'flexible', 'super
_strict_30']
Test: Kruskal-Wallis Test
The test statistics is 680.2052665448703 and the p-value is 4.113393137407
8247e-147.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.


Feature: instant_bookable
Categories: ['f', 't']
Test: Mann Whitney U Test
The test statistics is 46584347.0 and the p-value is 7.988604469599921e-30
.
```

```
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.


Feature: is_business_travel_ready
Categories: ['f']
Only one category has count above the threshold of 20. No test was perform
ed.


Feature: host_response_time
Categories: ['within an hour', 'within a few hours', 'within a day', 'a fe
w days or more']
Test: Kruskal-Wallis Test
The test statistics is 32.84918503752064 and the p-value is 3.465425031283
587e-07.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.


Feature: host_is_superhost
Categories: ['f', 't']
Test: Mann Whitney U Test
The test statistics is 38860336.5 and the p-value is 7.892123382249833e-13
.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.


Feature: host_has_profile_pic
Categories: ['t', 'f']
Test: Mann Whitney U Test
The test statistics is 449090.0 and the p-value is 0.011299338772144845.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.


Feature: host_identity_verified
Categories: ['f', 't']
Test: Mann Whitney U Test
The test statistics is 48516470.5 and the p-value is 0.0002208933204255774
6.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution


Feature: elevator
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 32136941.5 and the p-value is 0.0.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.
```

Feature: fire extinguisher
Categories: [1, 0]
Test: Mann Whitney U Test
The test statistics is 51607257.0 and the p-value is 3.0290338100345455e-07.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.

Feature: private entrance
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 41594725.0 and the p-value is 2.319847642147087e-58.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.

Feature: first aid kit
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 50688442.0 and the p-value is 0.37630187550722954.
Null hypothesis NOT rejected. There is not significant evidence the samples are not from the same distribution.

Feature: hot water
Categories: [1, 0]
Test: Mann Whitney U Test
The test statistics is 50525114.0 and the p-value is 0.0001446619928103638 7.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.

Feature: lock on bedroom door
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 27213554.5 and the p-value is 0.0.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.

Feature: cooking basics
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 43645602.0 and the p-value is 7.756227046188195e-74.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.

Feature: stove
Categories: [0, 1]

Test: Mann Whitney U Test
The test statistics is 44870356.5 and the p-value is 1.850638949509985e-65
.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.

Feature: long term stays allowed
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 39354844.5 and the p-value is 4.175335562542532e-66
.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.

Feature: microwave
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 47951470.0 and the p-value is 1.3771747211817397e-2
5.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.

Feature: refrigerator
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 48872390.0 and the p-value is 4.749782420897668e-26
.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.

Feature: oven
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 42634141.0 and the p-value is 2.252623077029503e-96
.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.

Feature: family/kid friendly
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 35114415.0 and the p-value is 6.468205540731169e-17
9.
Null hypothesis rejected. There is significant evidence the not all sample
s are from the same distribution.

Feature: dishes and silverware
Categories: [0, 1]

Test: Mann Whitney U Test
The test statistics is 46203736.0 and the p-value is 1.2975105640857729e-48.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.

Feature: self check-in
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 41919091.0 and the p-value is 1.5641052248945282e-64.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.

Feature: gym
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 28224153.5 and the p-value is 0.0.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.

Feature: coffee maker
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 39340759.5 and the p-value is 3.683219060795258e-116.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.

Feature: free parking on premises
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 48194141.5 and the p-value is 3.287582020327982e-17.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.

Feature: bed linens
Categories: [0, 1]
Test: Mann Whitney U Test
The test statistics is 47215516.5 and the p-value is 4.3759109922733146e-17.
Null hypothesis rejected. There is significant evidence the not all samples are from the same distribution.