**2. Data**

The dataset we use in this report contains price and relative features of 36,000 Airbnb listings in 6 major U.S. cities. This dataset contains all the information that a guest can find on the Airbnb platform, including the geographical information, properties and amenities, booking policy, reviews and host profile. Details are listed in the Table 1.

Table 1 Variables in the Airbnb dataset

|  |  |
| --- | --- |
| Geographical information | City  Neighborhood  Longitude and Latitude  Zip code |
| Properties | Accommodates  Number of beds  Number of bedrooms  Number of bathrooms  Property type  Bed type |
| Amenities | Amenities listed by host |
| Booking policy | Cancellation policy  Cleaning fee (logical)  Instant bookable |
| Reviews | First review date  Last review date  Number of reviews  Review score rating |
| Host profile | Profile picture (logical)  Identity verified (logical)  Host since date  Response rate |
| Data Source: AAE 722 |  |

**2.1 Data Exploration**

The six cites in the Airbnb data are Boston, Chicago, Washington DC, Los Angeles, New York City and San Francisco. Figure 1 shows that the price distribution are similar in those cities. There is no significant difference or pattern in prices between the West Coast, the East Coast and the Central Region. Figure 2 shows the geographical distribution on city level. Note that only one twentieth of observations in each city are shown on the map, avoiding overlapping of data points. Within each city, the listings are typically located along the main traffic, yet the price distribution has no obvious pattern. High prices appear in cities or suburbs, inland or coastal.

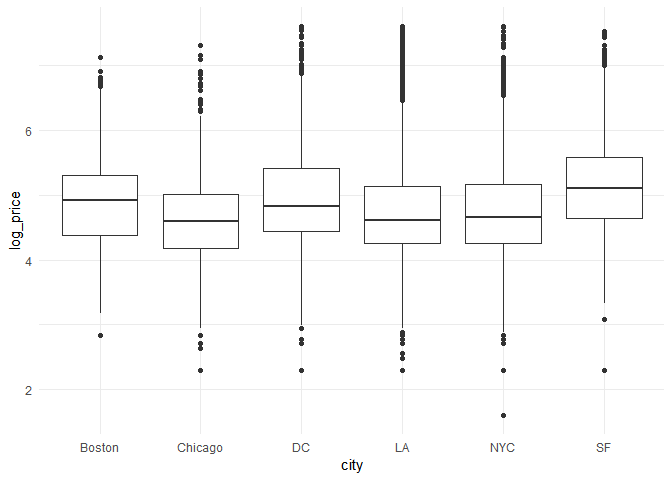


Figure 1 Price distributions across cities

Note: No significant difference or distribution pattern across the West Coast, the East Coast and the Central Region. Take the logarithm of the price to curve down the scale.

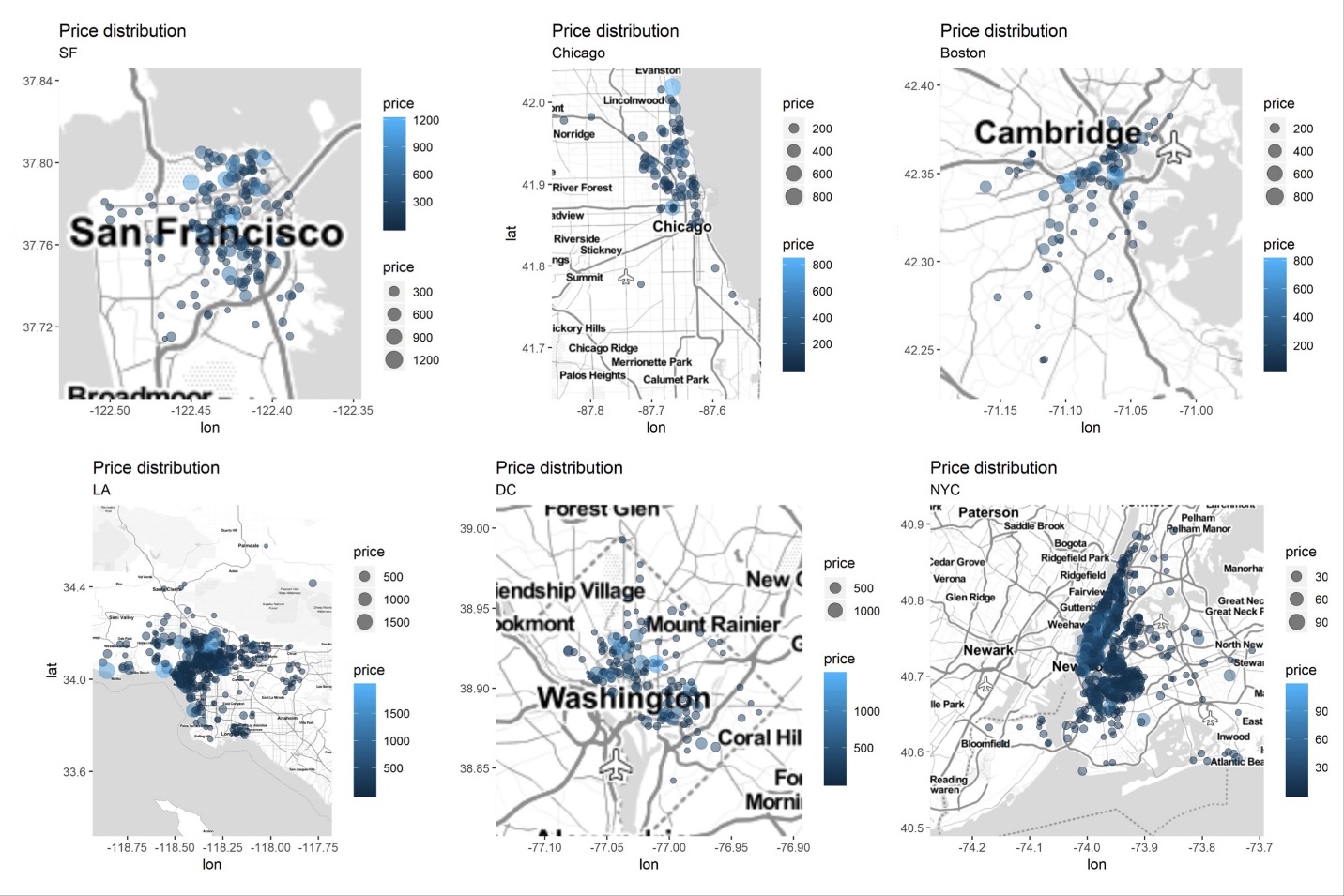


Figure 2 Price distribution within each city

Note: In order to avoid overlapping of data points, we randomly sampled one-twentieth of observations within each city to draw bubble map. The distribution of price is geographically random. Larger, lighter-colored bubbles present listings with higher price. Note that the scales of legend are different in each city.

Indeed, location is not the only factor of price. Downtown dorms may be cheaper than suburban villas. Other properties matter. Figure 3 shows the positive correlation of the price and the number of properties, while the variety is large especially where the property numbers are small, which suggests that after controlling for the size of listings, other affecting factors exits.

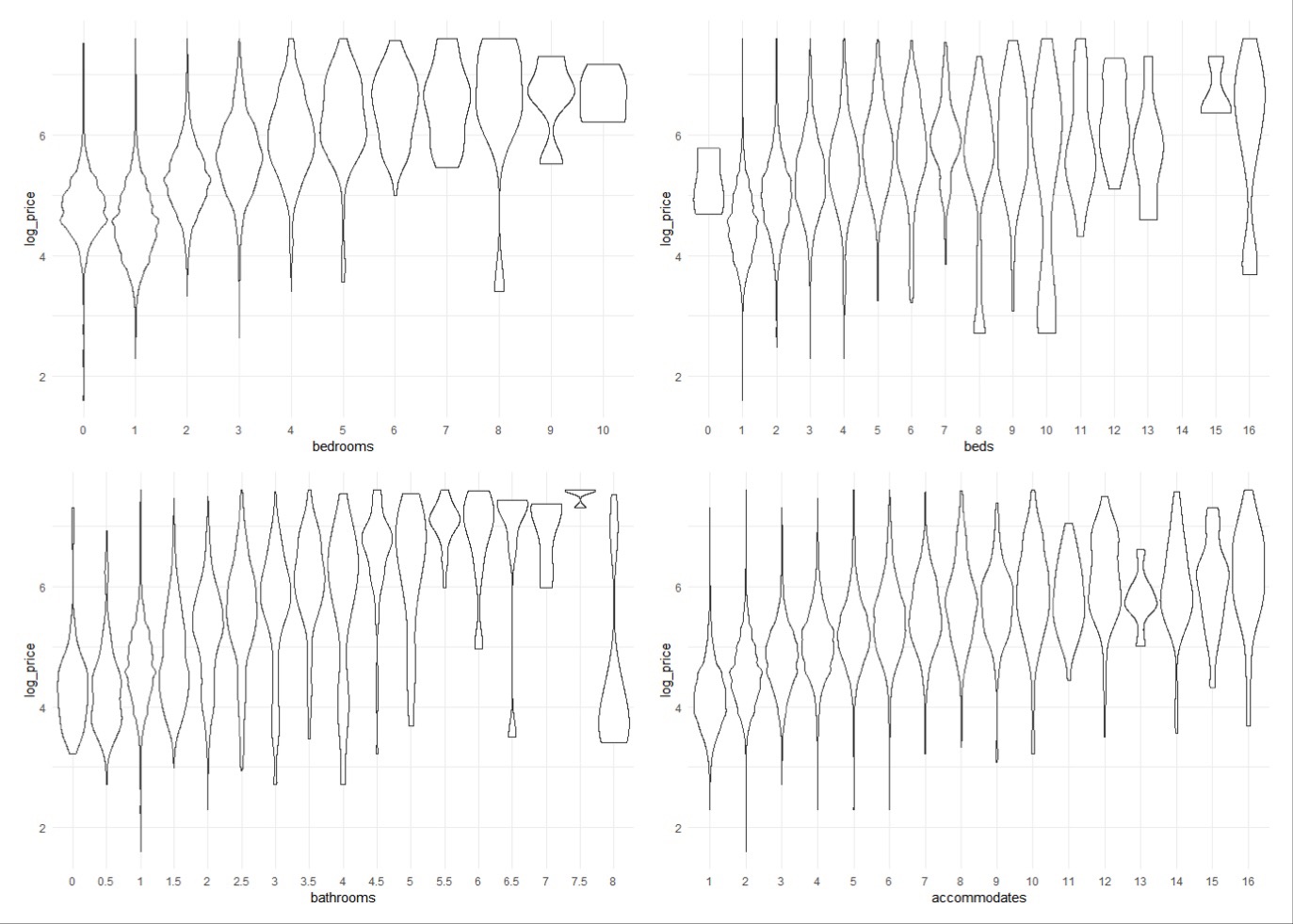


Figure 3 Price distributions against properties

Note: Price distributions against the number of bedrooms, beds, bathrooms and accommodates. Take the logarithm of the price to curve down the scale. Observations contains missing value are removed (detailed NA process discussed in section 2.2).

Considering the effect of amenities, we use word cloud to show the frequency at which the amenities appear in the host description, as shown in Figure 4. From highest to lowest, the top 10 listed are wireless Internet (34428), kitchen (32553), heating (32191), smoke detector (29841), essentials (29632), air conditioning (26623), hangers (23282), carbon monoxide detector (22874), shampoo (22170), and laptop friendly workspace (21184). Some of them are rigid needs of users, such as access to the Internet, air conditioning and shampoo. Some are basic facilities, which may not affect user choice but are necessary to list, such as smoke detector, carbon monoxide detector. Others like essentials and laptop friendly workspace, do not have clear definition, yet do affect impression of the listing.



Figure 4 Word Cloud of top listed amenities

For further understanding the impact of the amenities, we used box plots to show the price distribution between groups with and without certain amenity. The “certain” amenities are selected based on the frequency of listed plus some intuition. Except for the first aid kit, the price of the listings with certain amenity in all groups is higher than those without that amenity.

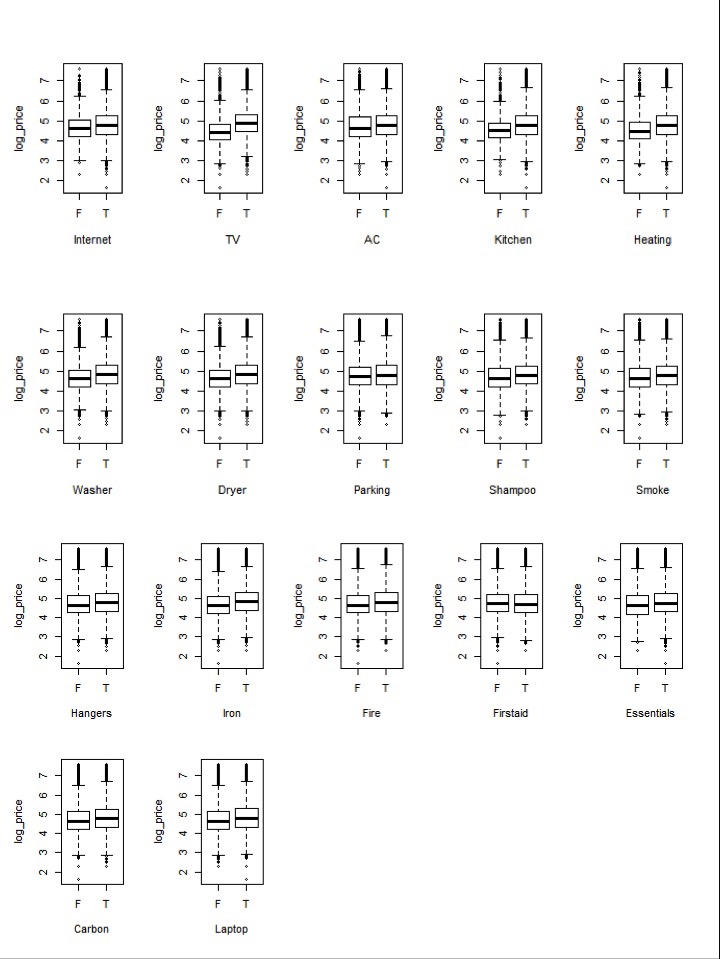


Figure Price distributions against amenities

Note: In general, the price of listings with certain amenity is higher than those without that amenity.

Airbnb users usually refer to reviews and the host profile when selecting a listing. Whereas in Figure 6, we do not see the clear correlations of price with reviews and host profile. Whereas Figure 7 indicates that price differences exist between those without review and those with at least one review, which suggests that hosts tend to lower the price after the first booking out.

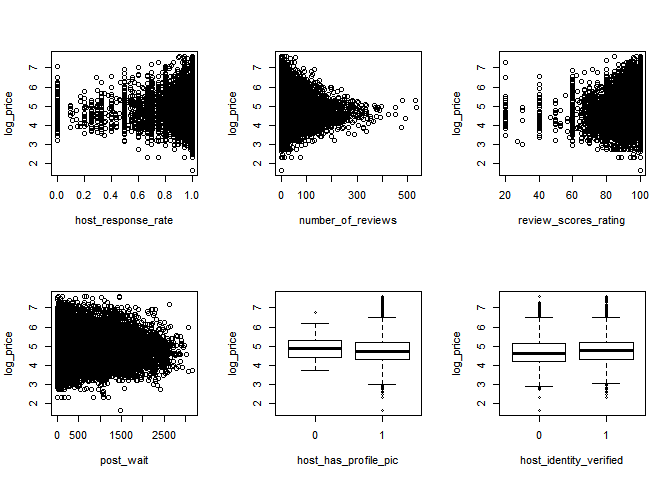


Figure 6 Price distributions against review and host profile

Note: As discussed in section 2.2, post\_wait is the day length from host register date to the first review date, measuring the probability of a listing being booked out. No clear correlations of price with reviews and host profile can be observed from graphs.

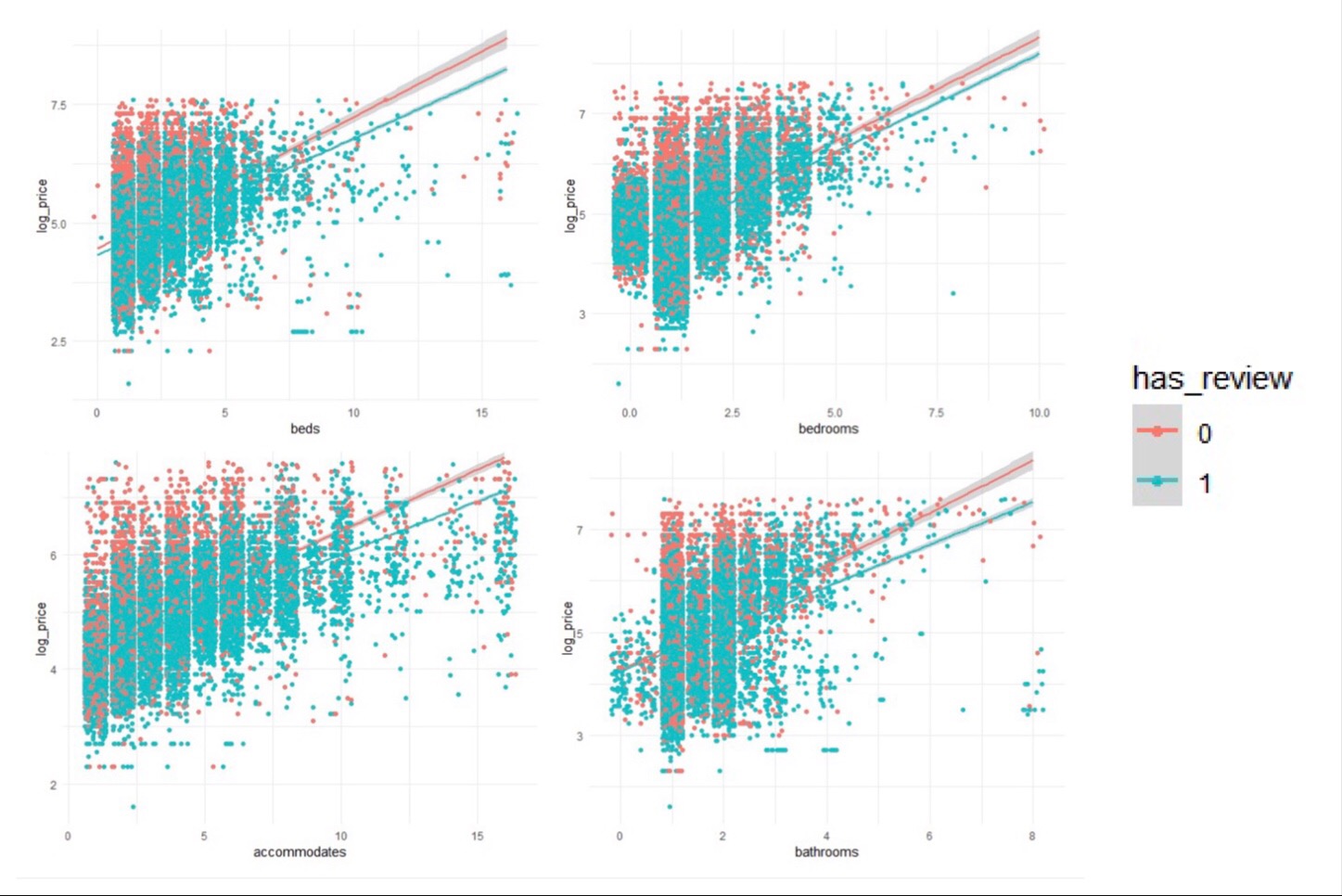


Figure 7 Price comparison of listings with/without reviews

Note: Red for listings with no review; Green for listings with at least one review. The solid lines are simple linear regressions of log price and numbers of beds, bedrooms, accommodates and bathrooms, respectively. Price differences exist between those without review and those with at least one review.

Throughout the statistical description, it is hard to say that prices of listings has a definite relationship with any certain feature, nor say which group of features is the determinant of price. In the meanwhile, it is nearly impossible to describe some subjective factors in the data, such as room pictures, host profile pictures, etc. The price of Airbnb listings is complex. It cannot be determined by a simple, sparse model. That is partly the reason why we want apply machine learning method to predict Airbnb listings price.

**2.2 Data Processing**

**2.2.1 Variable transferring**

In the dataset, we have 3 date-type variables: host since (the register date of the host), first review date and last review date. Those dates contain information about how popular the property is, how much experience the host has, how trustworthy the reviews are. Nevertheless, directly apply regression method on date-type data can be meaningless. Thus, we construct a new variable “post wait” as the date length between host since and first review, meaning to measure the probability of a listing booking out. Until the first guest post their review, one list is chosen by user without any review information. During this period, the better the quality of the listing property itself, the sooner the house can be booked out.

For the amenities of listing, we transfer it into a series of dummy variable, as shown in Figure 5. We also create a variable “number of amenities”, counting the total number of listed amenities. ­­The distribution of new variables “post wait” and “number of amenities” are shown in Figure 8.

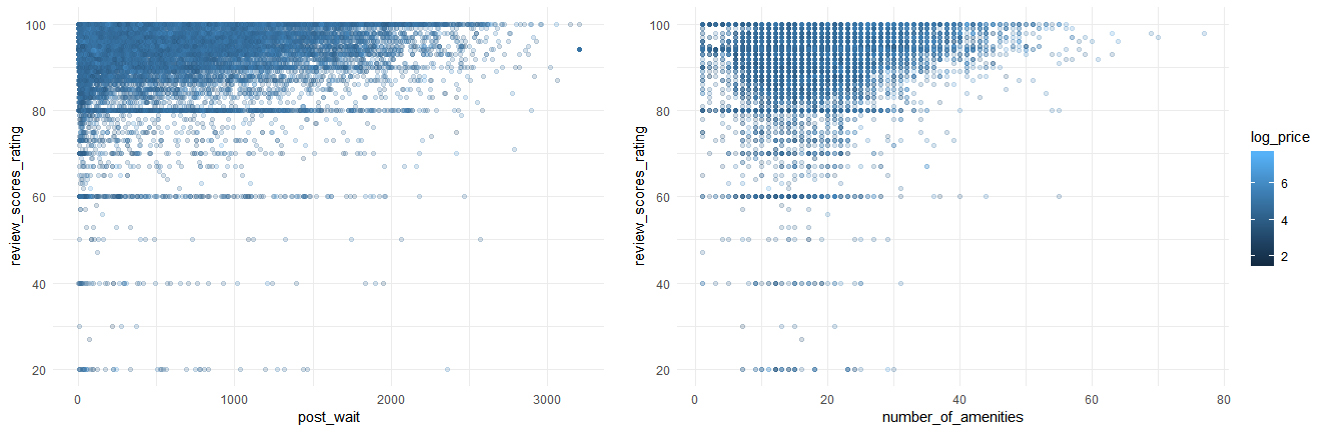


Figure 8 Distribution of new variables

**2.2.2 Dealing with missing values**

We have 8,333 observations containing missing values out of 36,000. We define missing values into 2 types: random and systematical missing values. Random missing values are rare, randomly happened, including bathrooms (95), bedrooms (41), beds (59), host has profile pic (91), host identity verified (91), and host since (91). Systematical missing values are more common, typically because of no review, which includes: first review (7717), last review (7696), and review scores rating (8143).

We simply dropped random NA's. For systematical one's, we assign sample mean to review scores rating, and sample max to post wait, which is the day-difference between host since and first review. Then remove first review and last review, since we won’t use those variables in further study.

**2.3 Data splitting**

We randomly divide three-quarters of the data as training data, and the rest one quarter as testing data. The splitting is the same throughout 5 prediction methods for comparison.