Analysis of Airbnb data in Shanghai: Is price influenced by room type and reviews number?

STAT406 Final Project

Group Number: 6

Name: Gu Zheng	Student ID: 518370910190
Name: Xu Pengcheng	Student ID: 518370910177
Name: Ye Haolin	Student ID: 518370910178

Date: August 4, 2021

Contents

1 Introduction

Established in 2008, Airbnb has became one of the largest online accommodation providers so far. Nowadays, more than 800 million guests have already registered Airbnb, with 9 million listings across 100 thousand cities and more than 170 countries (Figure 1)[1]. In addition, Airbnb has diversified its services, such as the introduction of Airbnb Experiences and even agree to acquire Hotel Tonight, furthering its agenda to become an end-to-end travel platform. This platform has gained public and scholarly attention due to its disruptive effects on hospitality industry, impacts on housing markets, and legal conflicts over housing, taxation and consumer regulations.

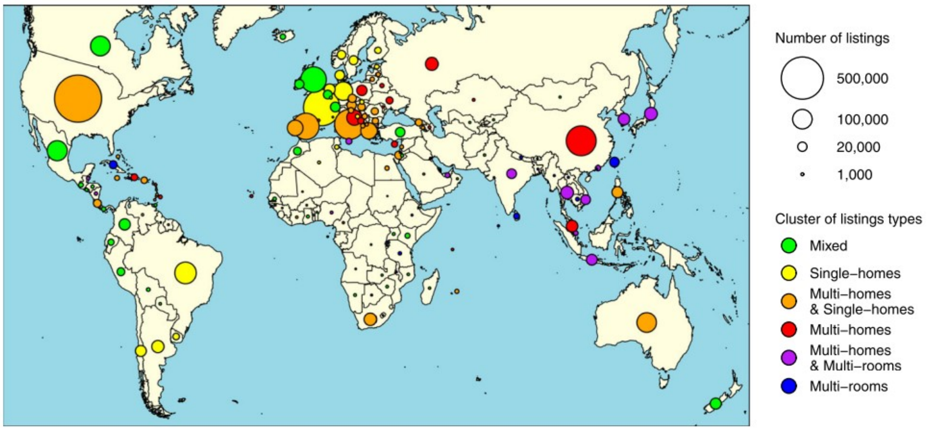

Figure 1: Distribution of active listings in 167 countries.

The emergence of Airbnb is unquestionably one of the most significant and transformative recent developments within the worldwide tourism sector. Therefore, several studies have already been conducted to explore the development and impact of this unique business pattern brought by Airbnb. However, these studies mostly focused on the influence and threaten posed by the phenomenon of Airbnb to the hotel sector, analyzing based on environment factors such as hospitality, leisure, sport, and tourism. For instance, Georgios Zervas et al., 2018 studied the economic influence of the sharing economy based on the case of Airbnb by analyzing Airbnb's introduction to the state of Texas and quantify its impact on the Texas hotel industry over the subsequent decade.[2] Other studies concentrated on how Airbnb make full use of local characteristics to achieve the success.

The purpose of this paper is to explore the statistic relationship between the room price and local factors, taking Shanghai as the study object. The factors taken into consideration will include environment factors such as latitude and longitude (since different location and environment will have a significant impact on the hotel pricing), humanity factors such as number of reviews and star rating (we initially assume that the response and comment from customers will positively influence the price), facility factors such as room type and hotel services (in general conditions better facility will result in higher price), and so on. Our result will give the customers reference on how to choose the best hotel while balancing the comfort and satisfactory with their economic abilities. To explore this topic, we will apply the statistics methods of permutation tests based on different kinds of test statistic.

The following report contains five main sections: *Data*, *Methods*, *Simulation*, *Analysis* and *Discussion*. We introduce the summary information and metrics for listings in Shanghai in *Data*, followed by an introduction to permutation tests in *Methods*. In *Simulation*, we compare the test statistic under different distribution, and choose our final test. We showcase the results of simulated test and comment on the null hypothesis in the *Analysis* section. Finally, in the *Discussion* part comments and future improvements will be put forward based on our research.

2 Data

2.1 Data Source

The data for this project is from the website (http://insideairbnb.com/get-the-data.html). The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site. Although the website contains data from more than a hundred countries, we are only interested in the data of Shanghai, China specially. Among all the available files, *listing.csv* includes the summary information and metrics for listings, which is most suitable for our analysis. In Figure 2, we show the data overview by plotting the geographical distribution of all the listings in terms of room price in the OpenStreetMap. It is not surprising to find that most available rooms are in the city center and most prices are within 1,000 yuan.

2.2 Key Variables

The dataset has 16 variables in total, which can be categorized into four types: identity information, environment factors, facility factors and humanity factors. In this project, we mainly focus on three variables which are worth to explore.

price: room price in RMB.

room_type: listing space type.

reviews_per_month: number of reviews per month.

Price is the most important variable we are interested in, since everyone pays attention to the price when they are booking a room. *room_type* is a typical categorical variable. *reviews_per_month* is the representative of numeric variables, as it is the humanity factor. To some extent, the number of monthly reviews represents the number of reservations and popularity of this host. The three selected variables contain as many different aspects as possible, and it is valuable to be analyzed in the housing market.

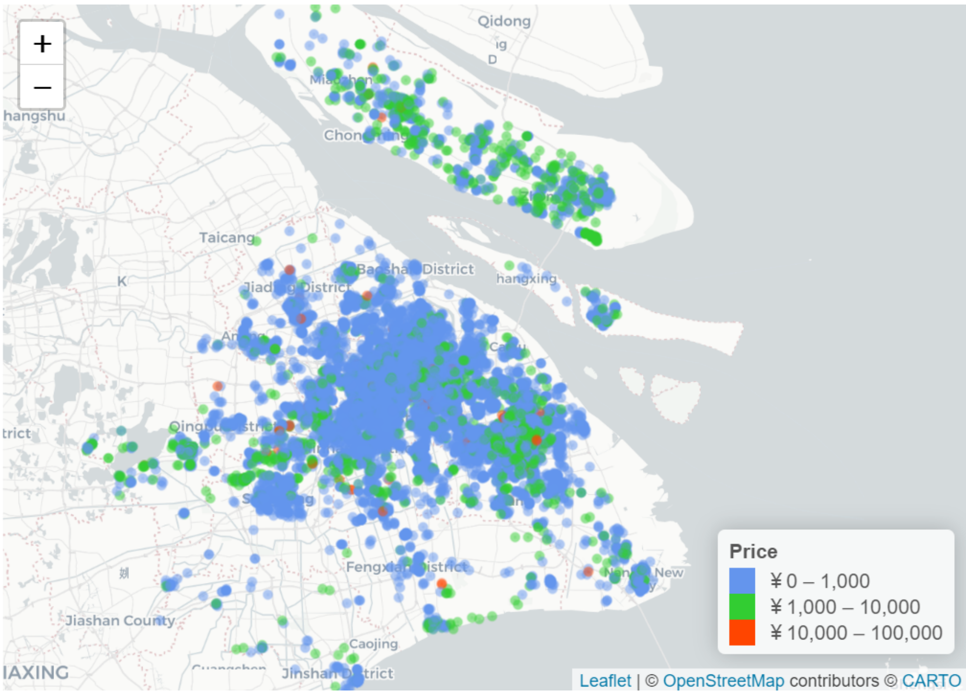

Figure 2: Location of Shanghai Airbnb listings in terms of room price.

2.3 Data Cleaning

There are 36294 rows with 16 variables in the raw data. However, it cannot directly meet our needs, so we have to make some changes to the original data:

(1) Delete *id*, *name*, *host_id*, *host_name* four variables, since they contain personal privacy and are insignificant for our research.

(2) Delete *neighbourhood_group* because it has no value for all the data and there doesn't exist the concept of blocks or neighbourhoods group in China.

(3) Delete the rows with $price = 0$ since it is not practical.

(4) Delete *hotel room* records for *room_type* variable. *hotel room* has only two records among all the data so it is difficult to analyze them.

(5) Change the data type of some features such as time from character to timestamp.

(6) Change the name of *neighbourhood* to totally English name for consistency.

(7) Append *reviews_per_month* column with zero for missing values because *number_of_reviews* equals to zero in these rows.

(8) Use log1p() function for *price* and *reviews_per_month*. The densities of these two variables don't follow the normal distribution, and more likely to have exponential distribution. For the serious skew data, log1p() returns log(1 + num) computed in a way that is accurate even when the value of number is close to zero.

After cleaning, we have 36288 rows of data with 11 variables in total.

3 Method

The three key variables have been shown in 2.2, and we are interested in whether the room price is influenced by the room type and reviews per month. In other words, we have to examine whether *price* is independent to *room_type* and *reviews_per_month*. Permutation tests will be used in both tests, while using different test statistic for categorical and numeric data.

Permutation test is a method of statistical inference by computing a test statistic on the dataset and then for many random permutations of sample data. Because of its freedom of overall distribution, it is widely used especially suitable for small sample data with unknown distribution, and certain hypothesis tests that are difficult to analyze with conventional methods. Permutation tests also condition on aspects of the sample such that the remaining data is invariant to permutation[3].

Considering the real-world data are usually not normally distributed, permutation tests have absolute advantage over other methods. In this project, we will deal with three sample problems for discrete data, and will test the independence between two numeric variables by permuting one.

3.1 Three Sample Problems

room_type has three types of values: entire home/apt, private room, and shared room. Suppose that they represent X, Y, Z separately, and we assume they are independent but not identical data, which means that

$$X_1, X_2, ..., X_n \overset{iid}{\sim} F, \quad Y_1, Y_2, ..., Y_m \overset{iid}{\sim} G, \quad Z_1, Z_2, ..., Z_p \overset{iid}{\sim} H$$

The null hypothesis is that three populations have the same distribution, and the alternative hypothesis is that they don't have the same distribution.

$$H_0 : F = G = H$$

For two sample problems, t-test is commonly used while it is unsuitable for three samples. Analysis of variance (ANOVA) is a statistical method that separates observed variance data into different components to use for additional tests. Both t-test and ANOVA examine whether group means differ from one another, but t-test compares two groups while ANOVA can do more than two groups. Therefore, ANOVA test will be used, and the detailed characteristics about this test will be mentioned in Simulation part.

Unlike the two samples, three sample problems doesn't have a unique representation of "difference of means" or "difference of medians", and we have to establish it by ourselves. Therefore, we assume the difference of means test is measured by

$$\Delta = \bar{X} - \frac{\bar{Y} - \bar{Z}}{2}$$

while the difference of medians test is

$$\delta = median(X) - \frac{median(Y) - median(Z)}{2}$$

3.2 Independent Tests for Numeric Data

In addition to three samples problems, we now have a single sample of pairs (X_i, Y_i), i = 1, ..., n, iid, where the variable X is *price* and the variable Y is *reviews_per_month.* The main goal is to judge whether these two variables are independent. We make the null hypothesis that X and Y are independent, while the alternative hypothesis is that they are not independent.

$$H_0 : F_{XY} = F_X F_Y$$

We could arbitrarily permute all the Y values and if the hypothesis is true, we should obtain the same distribution no matter which test statistic T(X, Y) we choose. We can test this hypothesis with a test statistic for pairs[4].

In practice, correlation may be useful for indicating a predictive relationship of interest and several methods exist that measure the degree of correlation. Sample correlation is a way to summarize the linear relationship between two variables. However, the relation between *price* and *reviews_per_month* doesn't seem to be linear. Therefore, we should consider generalizations of correlation to non-linear dependence. The two most common non-linear rank based correlation coefficients are Spearman's rank correlation coefficient and Kendall's rank correlation coefficient[5]. In this project, we will compare these two tests of correlation. Besides that, distribution-free methods based on ranks are also good tests for this kind of data.

4 Simulation

As we have stated in Method, permutation tests have the advantages that they don't invoke assumptions related to normality. Thanks to it, permutation tests allow for the use of almost all the test statistic. In this part, we will perform several simulations to show the operating characteristics of various tests like Type-I error and power with different underlying distributions, as well as their own confidence intervals.

We fix $n = m = 30$ for the number of random variables throughout the experiment and vary the distributions of X, Y and (Z) to follow normal distribution, log-normal distribution, skew normal distribution and Student's t-distribution. The permutation test is replicated for 1000 times. In addition, we perform a binomial test to provide a 99% confidence interval for the estimated Type-I error and power when $\alpha = 0.05$.

4.1 Test Statistic Comparison for Discrete Data

For discrete data, we compare the performance of three test statistic “Difference of Medians”, “Difference of Means”, and “ANOVA test”. All the information about Type-I error are shown in Figure 3, and those of power are shown in Figure 4.

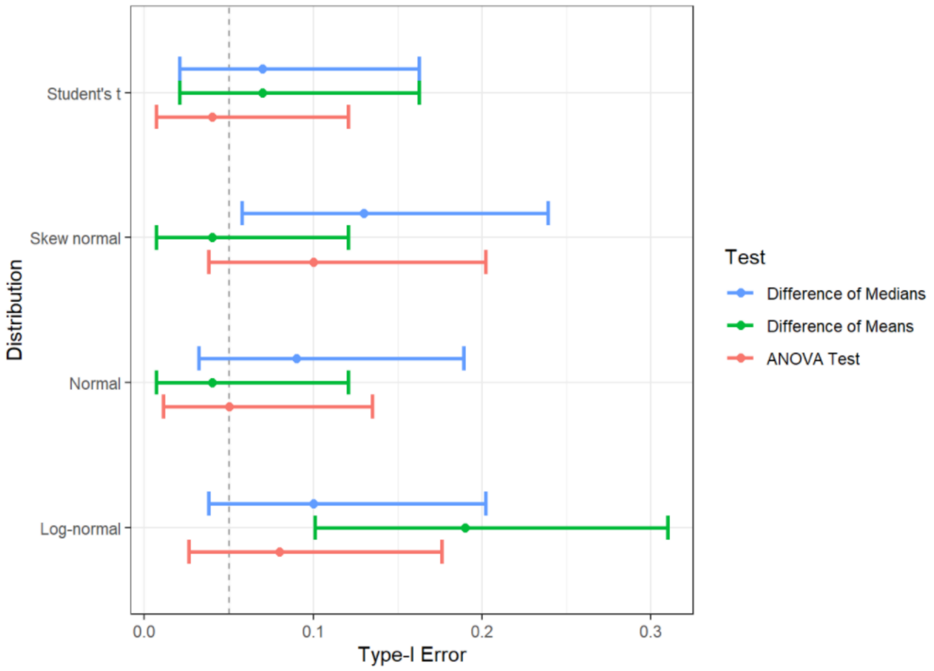

Figure 3: Type-I error of different tests on different distributions.

From Figure 3, all the estimated Type-I errors are close to the nominal 0.05 level and 0.05 is included in all the confidence intervals, which means that we would not reject the hypothesis that the test has size less than level. For skew normal and normal distribution, Difference of means has the smallest Type-I error, while ANOVA tests has the smallest for Student’s t-distribution and log-normal distribution.

Figure 4 indicates that the power differs a lot for different tests. ANOVA test has the best power value under any distribution, which is within our expectation. Difference of medians and difference of means have various performance on different distribution, and they don’t seem to have large difference.

Like t-test, ANOVA test has three assumptions: (1) independence assumption: the elements of one sample are not related to those of other samples; (2) normality assumption: samples are randomly drawn from the normally distributed populations with unknown population means; (3) equal variance assumption: the population variances of two groups are equal. However, our data may largely not follow the normal distribution, which violates the assumptions for ANOVA test. Even if ANOVA test performs the best among the three test statistic, we cannot choose it for its strict assumptions.

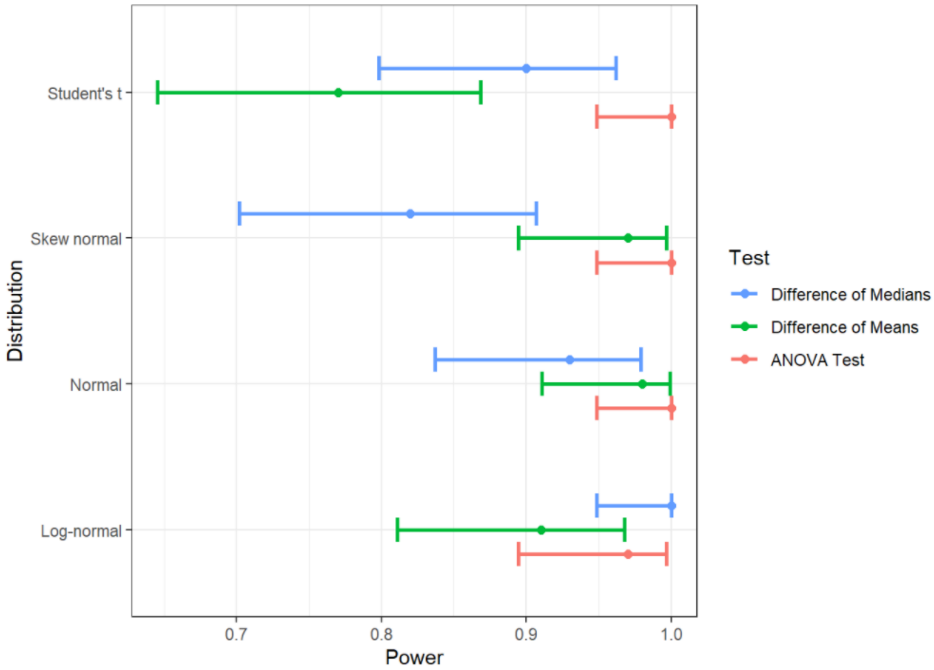

Figure 4: Power of different tests on different distributions.

Comparing difference of means test and difference of medians test, we use difference of means test as our finally test statistic for discrete data.

Besides the tests mentioned before, there are also some other test statistic can be taken into consideration. For example, prices can be divided into several intervals so that we can use Chi-square test and Fisher's exact test by establishing a contingency table. However, chi-square test should only be applied when the expected frequency of any cell is at least 5, so it cannot be used here after experiment. Fisher's exact test[6] is valid under this situation. However, considering that establishing intervals for log-price lose much precision for the variable *price* in reality, we don't include these methods in the content.

4.2 Test Statistic Comparison for Numeric Data

For numeric data, we compare the performance of two test statistic "Kendall rank correlation test" and "Spearman rank correlation test". All the information about Type-I error are shown in Figure 5, and those of power are shown in Figure 6.

Figure 5 shows that the performance of Kendall rank correlation test and Spearman rank correlation test are very similar. The formal test have relatively smaller Type-I error under the normal distribution, and relatively larger value under the skew normal distribution.

Figure 6 indicates that the power of both tests are very small in general. Even if we make large difference to the alternative hypothesis, the power still has relatively

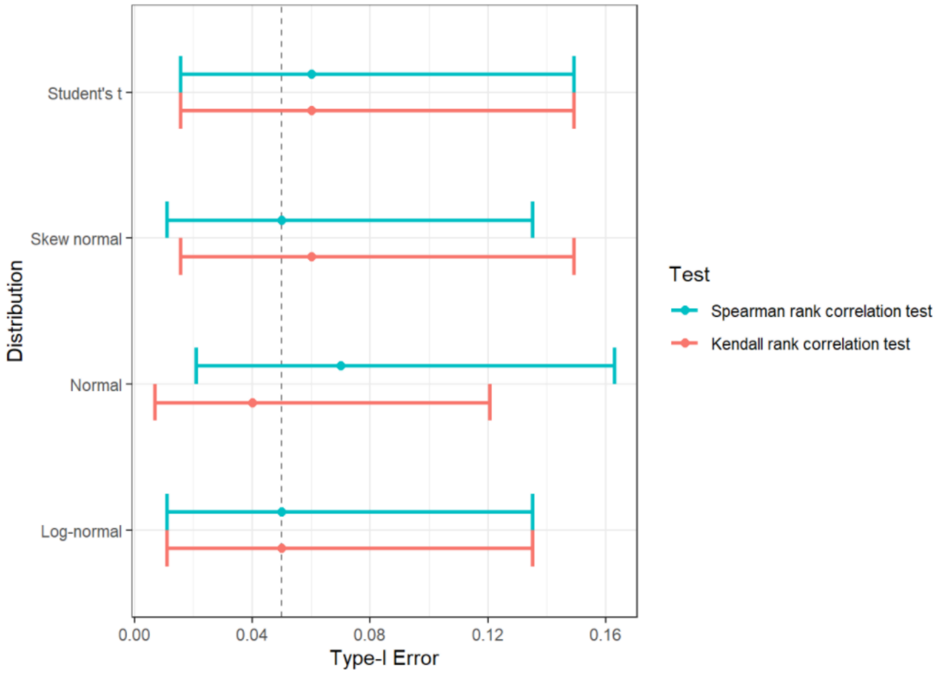

Figure 5: Type-I error of different tests on different distributions.

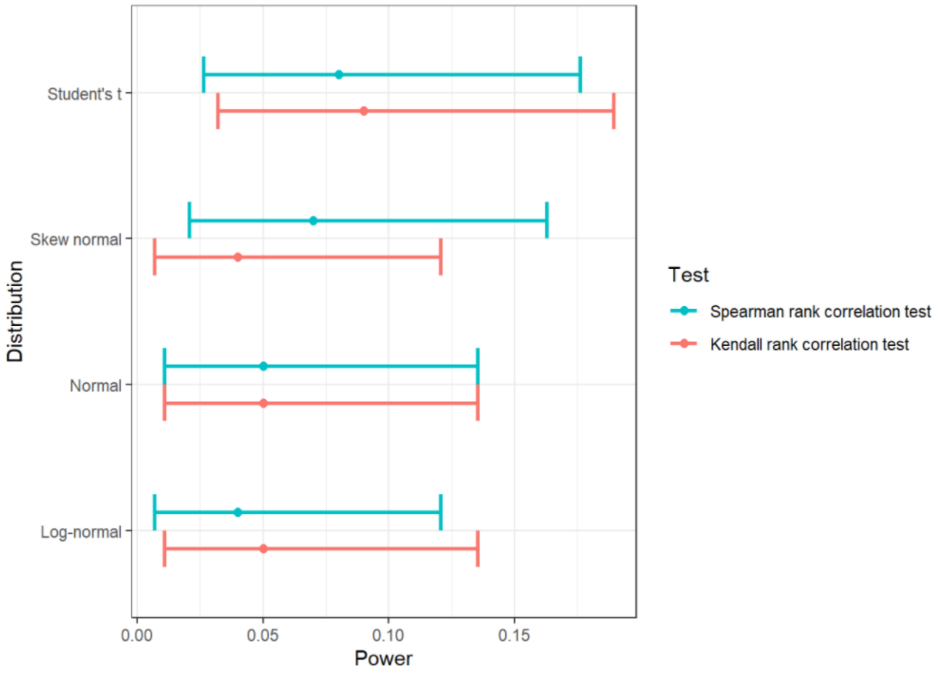

Figure 6: Power of different tests on different distributions.

small value. In comparison, Kendall rank correlation has larger power under the log-normal distribution and Student's t-distribution, while smaller power under other two distributions.

Comparing these two test statistic, Kendall rank correlation is calculated based on concordant and discordant pairs, thus it is insensitive to error and P-values are more accurate with smaller sample sizes. Spearman's rank correlation is calculated based on deviations, so it is more sensitive to errors and discrepancies in data[7]. However, in most of the situations, the interpretations of Kendall's tau and Spearman's rank correlation coefficient are very similar and thus invariably lead to the same inferences. Considering this, we will use Kendall rank correlation as the test statistic for numeric data.

5 Analysis

5.1 Permutation Test for ***room_type***

room_type has three types of values after cleaning, 20520 records for *Entire home/apt*, 14784 records for *Private room*, and 984 records for *Shared room.* Figure 7 shows the density plot of log-price in terms of different room type. The three samples don't seem to follow the normal distribution, which is within our expectation. In particular, *Shared room* has a serious right skew and relatively lower means than another two types, which may be caused by its small sample size compared with other two types. The Central Limit Theorem enables the samples of *Entire home/apt* and *Private room* to be more likely normal even if it isn't.

Applying the permutation test with difference of mean test statistic, we get the distribution of permutation results in Figure 8.

The observed difference of mean is the red line in Figure 8, which is 0.855. Since all the permutation results are smaller than the observed data, so the resulting p-value is zero. Therefore, we reject the null hypothesis that the three samples have the same distribution. Different distributions mean that the log-price of each room type are various distributed, so the variable *price* is dependent of the variable *room_type.*

5.2 Permutation Test for ***reviews_per_month***

reviews_per_month is a seriously right-skewed data, and it is reasonable in reality since the number of reviews for most hosts may not be very large. Most values of *reviews_per_month* are only at most one or two. Even if we apply the logarithmic conversion, it still has very long tails. The scattered plot of logprice and logreviews is shown in Figure 9.

Applying the permutation test with Kendall rank correlation test statistic, we obtain the distribution of permutation results in Figure 10.

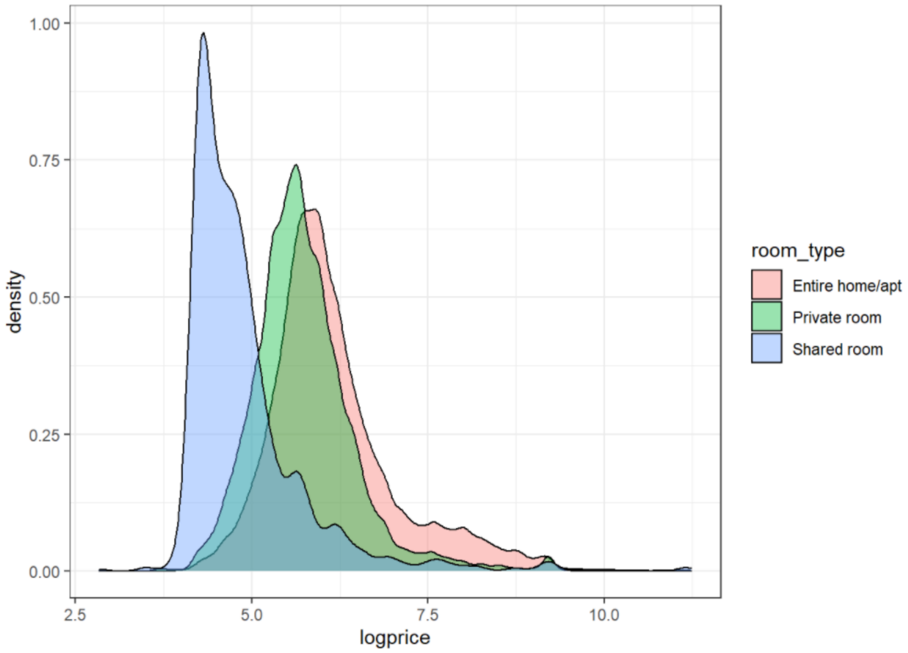

Figure 7: Density distribution of log-price in terms of room type.

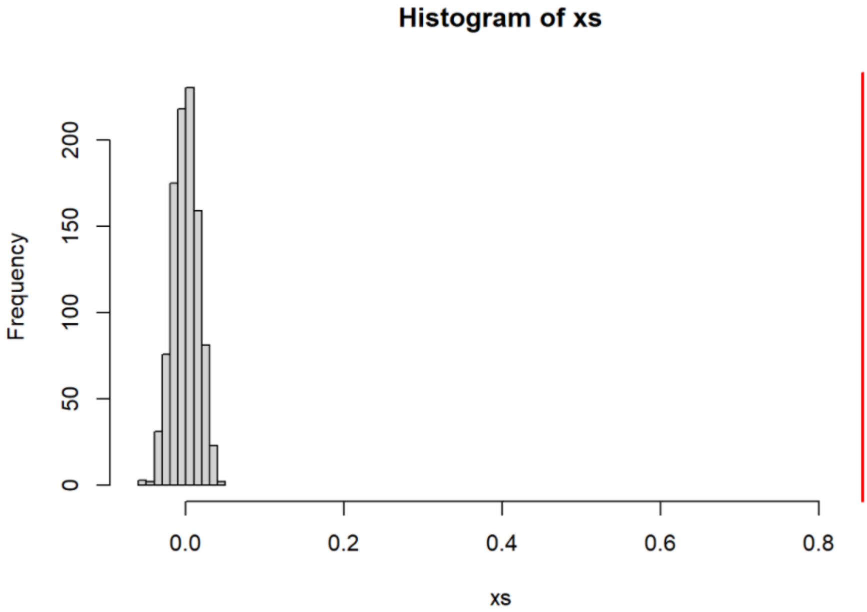

Figure 8: Difference of mean test on room type.

The observed Kendall rank correlation is the red line in Figure 10, which is 1.964×10^{-7}. Similar to the previous part, all the permutation results are larger than the observed data, so the resulting p-value is zero. Therefore, we reject the null hypothesis that these two variables are independent, which leads to the conclusion that the variable *price* is dependent of the variable *reviews_per_month*.

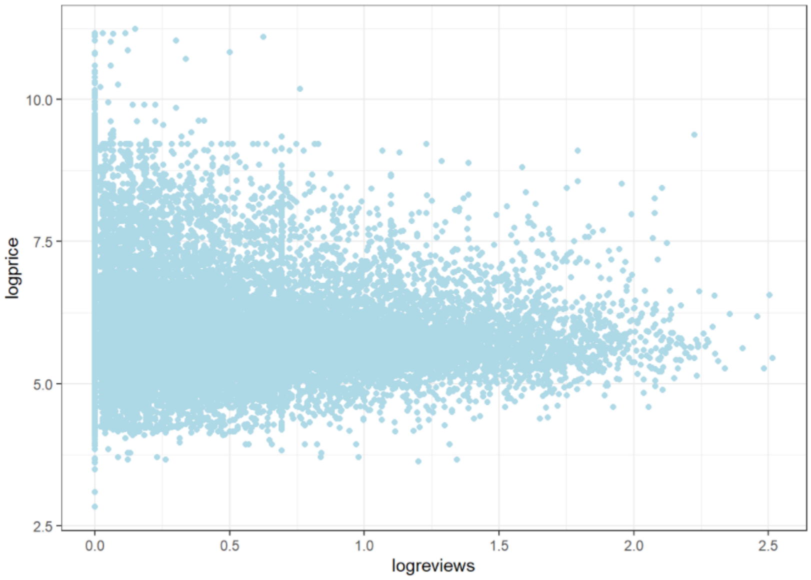

Figure 9: Scattered distribution of logprice and logreviews.

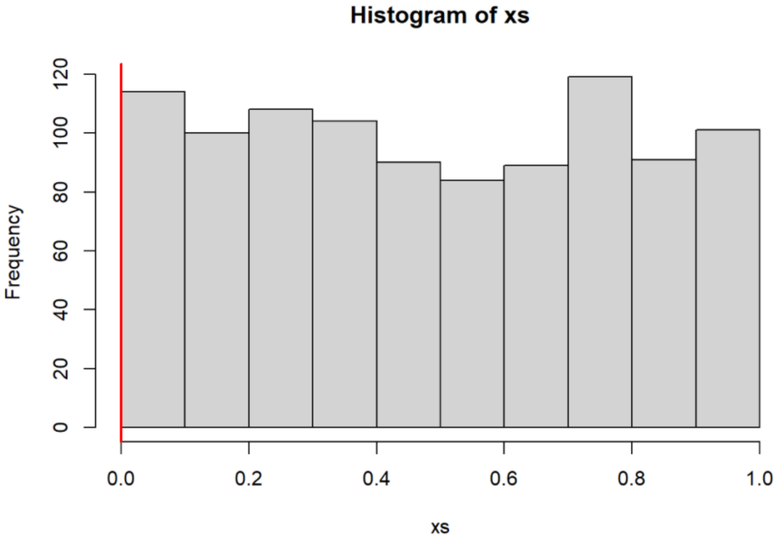

Figure 10: Kendall rank correlation test on reviews per month.

6 Discussion

Nowadays, the tourism industry has gradually developed and people's need to live outside is increasing. Airbnb provides an online marketplace for lodging and vacation rentals. Analyzing the Airbnb open data can not only help the customers to find the most cost-effective house, but also provide suggestions to the hosts about how to increase revenue under the limited conditions.

In this project, we deal with the Airbnb open data in Shanghai, where is the city we are most familiar with. Despite many variables in the dataset, we take three typical variables we are most interested in: price, room type and reviews number, which separately represents identity information, facility factor and humanity factor. We come up with the question that: is price influenced by room type and reviews number? In order to solve it, we have to figure out whether room price is independent of room type, and whether room price is independent of reviews number.

Randomization techniques are usually more powerful than non-parametric rank transformation tests like Kruskal Wallis test. Their main strength is that they don't invoke assumptions related to normality. Non-parametric methods in general (bootstrap, jacknife, randomization) can handle data from any distribution. Therefore, we use permutation tests in this project mainly because it doesn't need the assumptions of normality.

Since room price and reviews number are numeric data while room type is discrete data, so we apply different test statistic based on different situations. Difference of Medians, Difference of Means, and ANOVA test are used for discrete data, while Kendall rank correlation test and Spearman rank correlation test are applied to numeric data. After compare the performance of different tests under different distributions, we choose Difference of means test and Kendall rank correlation test as our final test statistic.

In the analysis part, we apply the permutation test with corresponding test statistic to different kinds of data, and finally obtain the results that the price is dependent of room type, as well as the reviews per month.

There are still some insufficient in this project. Firstly, the test statistic for numeric data can have more choices, such as the distribution-free methods based on ranks. Secondly, in the simulation part, we could test the test statistic on more different distributions such as exponential distribution. Thirdly, we only take three variables as examples to apply the permutation tests, we can further work it on all the valid variables. After that, we can establish a generalized linear model based on the dependent variables of price, use cross validation to optimize the model, and finally make prediction on the room price based on the given conditions.

A Reference

[1] C. Adamiak, “Current state and development of Airbnb accommodation offer in 167 countries”, *Current Issues in Tourism*, 2019, doi:10.1080/13683500.2019.1696758

[2] Zervas, Georgios, Davide Proserpio, and John W. Byers. ”The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry.” Journal of marketing research 54.5 (2017): 687-705.

[3] M. Fredrickson, “lecture-week08_permutation_randomization_basics-flat.pdf”, *UMJI-SJTU*, Shanghai. Accessed July, 18, 2021.

[4] M. Fredrickson, “lecture-week08_permutation_randomization_advanced-flat.pdf”, *UMJI-SJTU*, Shanghai. Accessed July, 18, 2021.

[5] Wang, Y., Li, Y., Cao, H. et al, “Efficient test for nonlinear dependence of two continuous variables”, *BMC Bioinformatics*, 16, 260, 2015, doi:10.1186/s12859-015-0697-7.

[6] Fisher, R. A. (1922). ”On the interpretation of 2 from contingency tables, and the calculation of P”. Journal of the Royal Statistical Society. 85 (1): 87–94. doi:10.2307/2340521. JSTOR 2340521.

[7] “Kendall’s Tau and SPEARMAN’S Rank Correlation Coefficient.” *Statistics Solutions*, 5 May 2021, www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/kendalls-tau-and-spearmans-rank-correlation-coefficient.