ISSS602 Data Analytics Lab

IDEA on Airbnb in Singapore

Technical Report

Submitted By: Liu Cuiyi

Submitted data: 29/09/2019

Contents

1. Context

Airbnb, a service that connects travelers and homeowners with available rooms, offers a variety of accommodations. Different from other tourism websites, Airbnb, as a tourism platform directly facing rental landlords and tourists, provides a direct communication channel for landlords and tenants, and its relatively innovative form has attracted public attention. Singapore, one of the world's most competitive cities, has yet to legalize short-term rentals. The analysis of relevant data of Airbnb is helpful for us to analyze the possible impact of short-term housing rental on tourism and real estate industry in Singapore.

2. Data Preparation

The datasets of Airbnb are consisting of 6 files, namely “listings”, “listings_summary”, “reviews”, “reviews_summary”, “calendar” and “neighbourhoods”.

Upon examination of the datasets above, some data quality issues have been found and modified. Attach to Appendix 1 to have detailed information.

2.1. Data Quality Issues

2.1.1. Incorrect classification of character data as numeric data: As for “id”, “host_id” are used as unique identifier to differentiate between listings and hosts, there’s no need to get statistics from these variables, likewise variables “scrape_id”. Change data type from “Numeric” to “Character” and choose “Nominal” as modeling type.

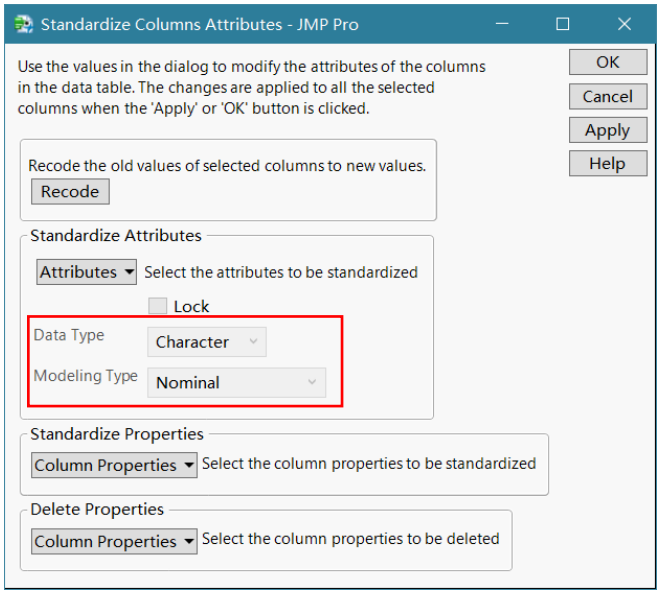


Figure 1

2.1.2. Incorrect classification of nemeric data as character data: The variable “host_response_rate” should be in percentage form to get statistics. Change its data type from “Character” to “Numeric” and choose “Continuous” as modeling type.

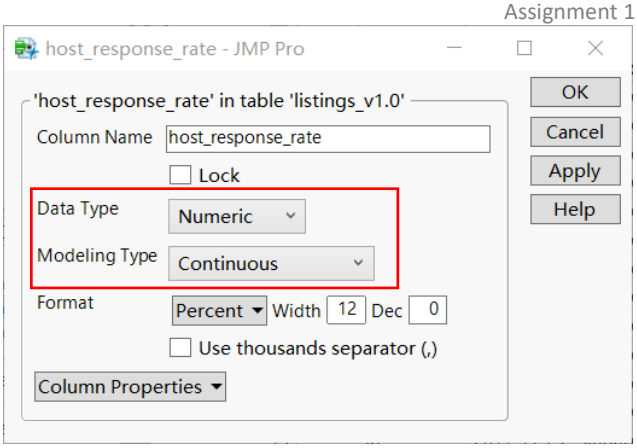


Figure 2

2.1.3. Incorrect modeling type of some variables: For doing further text analysis of listings' features, change modeling type of "summary", "space", "description", "neighborhood_overview", "notes", "transit", "access", "interaction", "house_rules" to "Unstructured Text" by standardize their attributes.

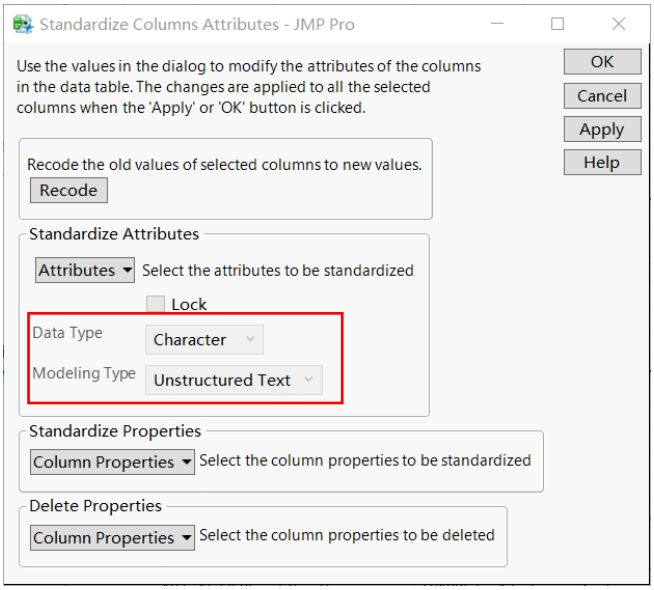


Figure 3

2.1.4. Incorrect data format of some variables: Correct data format of "latitude" and "longitude" as "latitude DMS" and "Longitude DMS". Standardize format of variables "price", "weekly_price", "monthly_price", "security_deposit", "cleaning_fee", "extra_people" as "Singapore Dollar ($)" and fix at 0 decimal places.

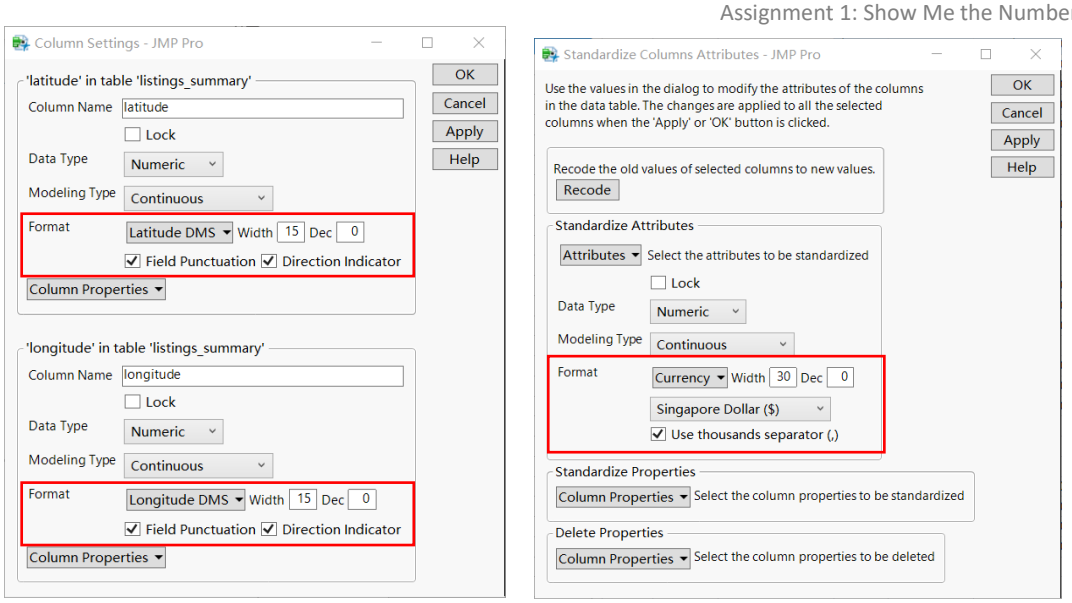


Figure 4

2.1.5. Overmuch missing values (cover 90% or more): Variables which contain over 90% missing values can't infer accurate statistics. These variables below except "weekly_price" and "monthly_price" should be excluded and hid for further analysis.

Columns	N	N Missing
id	8211	0
thumbnail_url	0	8293
medium_url	0	8293
xl_picture_url	0	8293
state	733	7543
square_feet	54	8157
weekly_price	585	7626
monthly_price	645	7566
license	4	8289
jurisdiction_names	1	8292

Figure 5

2.1.6. Missing value issue of variables related to hosts: Use tool "Columns Viewer" to show the quantity of missing values in each variable related to listings' host. The result shows that many variables have the same number of missing values. Select the 82 rows of missing values of the variable "host_name" to have detail look. The result shows that the 82 rows lose almost all the information related to host, which is unreasonable, so delete the 82 rows selected.

Columns	N	N Missing
host_id	8293	0
host_url	8293	0
host_name	8211	82
host_since	8211	82
host_location	8188	105
host_about	4975	3318
host_response_time	8211	82
host_response_rate	8211	82
host_acceptance_rate	8211	82
host_is_superhost	8211	82
host_thumbnail_url	8211	82
host_picture_url	8211	82
host_neighbourhood	7172	1121
host_listings_count	8211	82
host_total_listings_count	8211	82
host_verifications	8293	0
host_has_profile_pic	8211	82
host_identity_verified	8211	82

Columns	N	N Missing
host_id	82	0
host_url	82	0
host_name	0	82
host_since	0	82
host_location	0	82
host_about	0	82
host_response_time	0	82
host_response_rate	0	82
host_acceptance_rate	0	82
host_is_superhost	0	82
host_thumbnail_url	0	82
host_picture_url	0	82
host_neighbourhood	0	82
host_listings_count	0	82
host_total_listings_count	0	82
host_verifications	82	0
host_has_profile_pic	0	82
host_identity_verified	0	82

Figure 6

2.1.7. Incorrect "0" values of variables "host_listings_count" and "host_total_listings_count": there are "0" values in these two variables, to ensure accuracy, use "count" function to create a new formula column, named "host_listings_count_singapore", based on the variable "host_id". This new variable represents the quantity of listings each host owns in singapore. Replace "0" in original variables with corresponding value in the new variable. Figure 8 shows summary statistics of these three variables after modification.

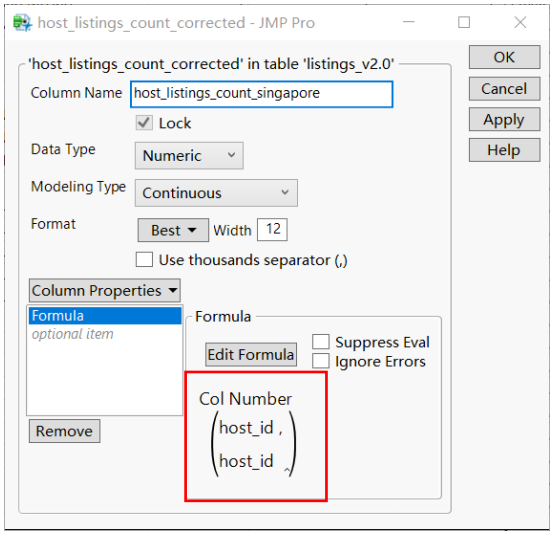
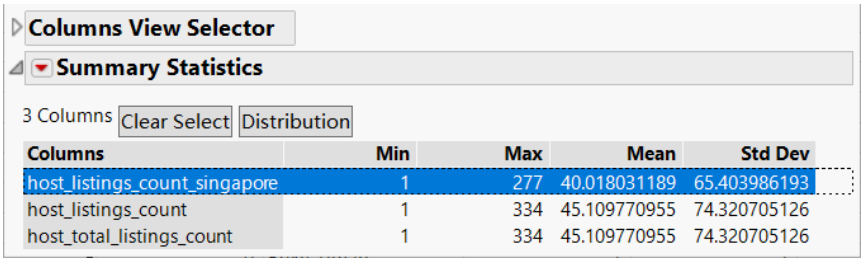


Figure 7



Columns	Min	Max	Mean	Std Dev
host_listings_count_singapore	1	277	40.018031189	65.403986193
host_listings_count	1	334	45.109770955	74.320705126
host_total_listings_count	1	334	45.109770955	74.320705126

Figure 8

2.1.8. Abnormal values of variable "host_since": To examine whether all the dates in "host_since" is earlier than those in "first_review", create a new formula column to examine this logic as shown in Figure 9. The result shows that there is one abnormal data (Figure 10). Replace this abnormal data with the date of the first review received by the host from all the listings owned.

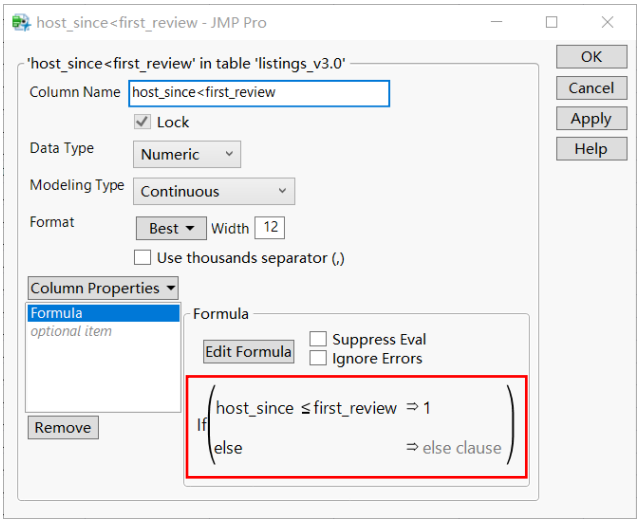


Figure 9

host_id	host_since	first_review	host_since<first_review
211434941	2018-08-24	2017-05-21	•

Figure 10

2.1.9. Outliers issue and abnormal value of variable "price": The distribution of "price" shows that there are overmuch outliers, we should exclude these unusual data from statistical analysis. Noticed that the 90% quantile of price value is $290, so select and delete the rows with a price higher than $290. Also, there are 3 observations of price valued 0, these rows need to be deleted.

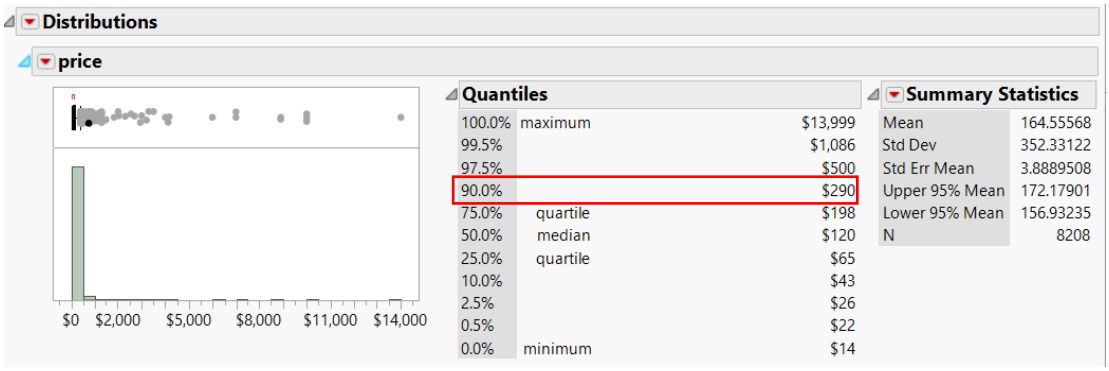


Figure 11

2.2. Data Quality Optimization

2.2.1. Change Boolean value to number "0" and "1": Variables "host_is_superhost", "host_has_profile_pic", "host_identity_verified" and others are expressed as Boolean

value, which can be replaced by numbers “0” and “1” to calculate mean or probability of an event. Change Boolean value “f” to “0” and “t” to “1“, then standardize their data type as “Numeric” and modeling type as “Continuous”.

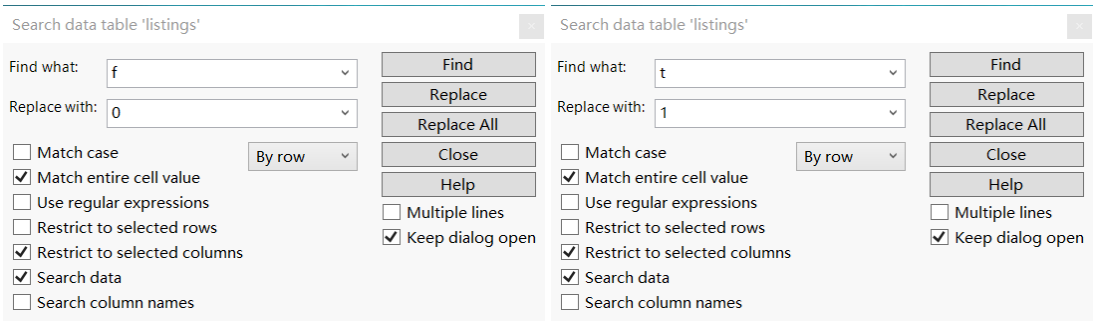


Figure 12

2.2.2. Recode and group similar values: For variable “host_location”, group locations which is in Singapore to find the regional distribution of hosts. Type keywords “SG”, “Singapore” and “sin” in turn to filter out values similar to “Singapore”. Figure 14 shows the result after grouping. For variable “cancellation_policy”, some categories under strict policy cover such small proportions that we could integrate these different categories of strict policy into one group (Figure 15).

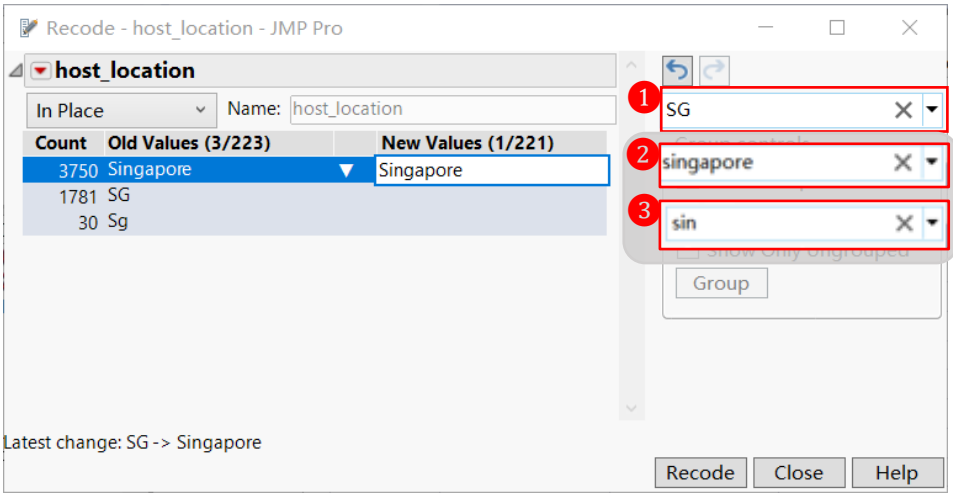


Figure 13

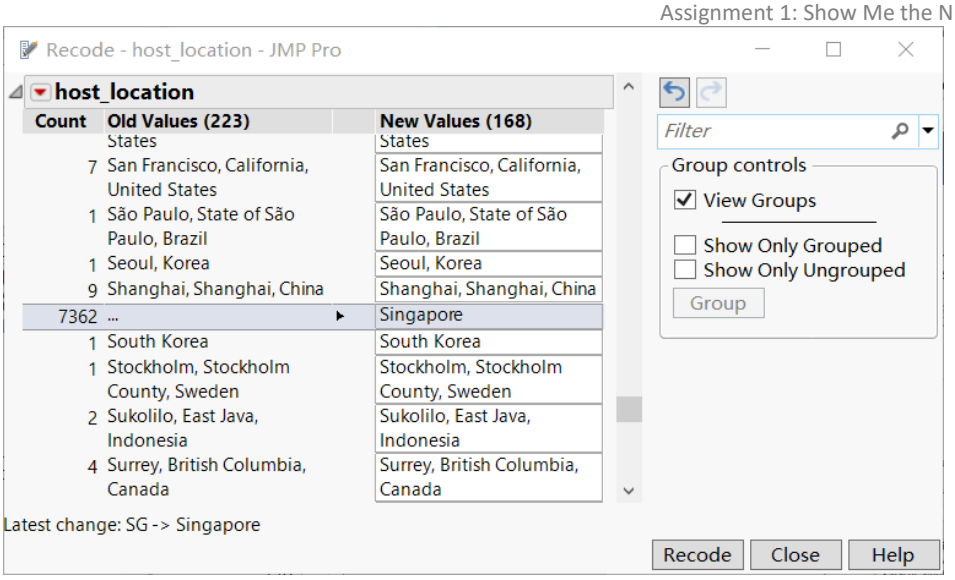


Figure 14

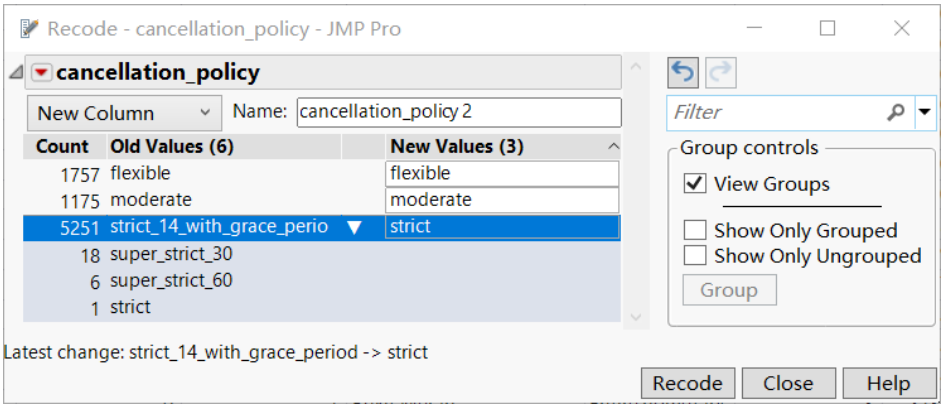


Figure 15

2.2.3. Unify data format: The format of variable "reviews_per_month" contain either 1 or 2 decimal places, to unify data format, fix it at 2 decimals.

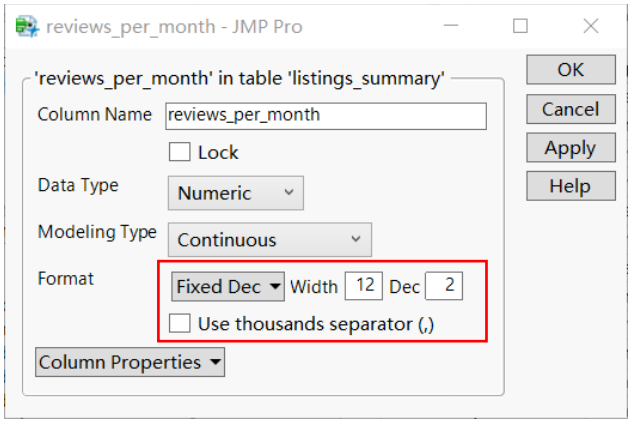


Figure 16

2.2.4. Organize data table: To make data table clearer and easy to observe, we can exclude and hide useless columns which contain redundant information or single value, including "street", "city", "market", "country_code", "country", etc.

3. Data Analysis and Insights

3.1. Insights into booking volume of listings in the next year

3.1.1. Compare booking volume of different listings in a year

First, derive the booking volume of each listing in a year by subtracting values in the variable "availability_365" from 365 (Figure 17). Then, insert a formula column to sort these listings into 6 groups: "no booking", "1-30", "31-90", "91-180", "181-364", "fully booked" (Figure 18).

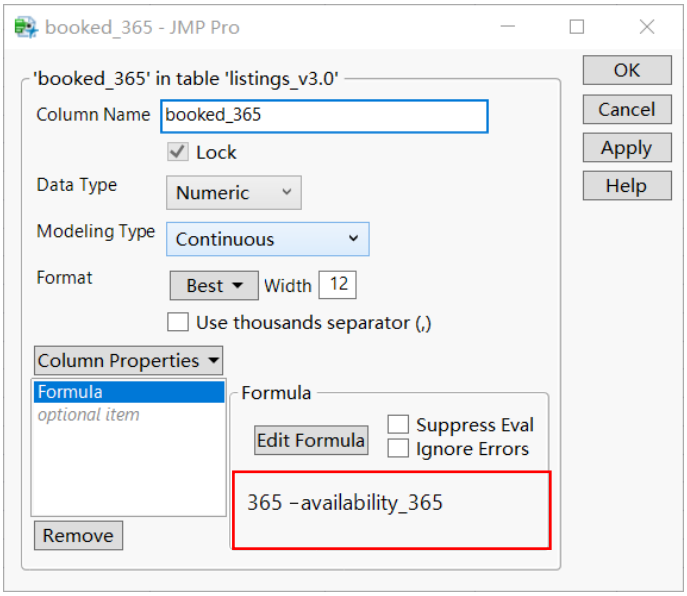


Figure 17

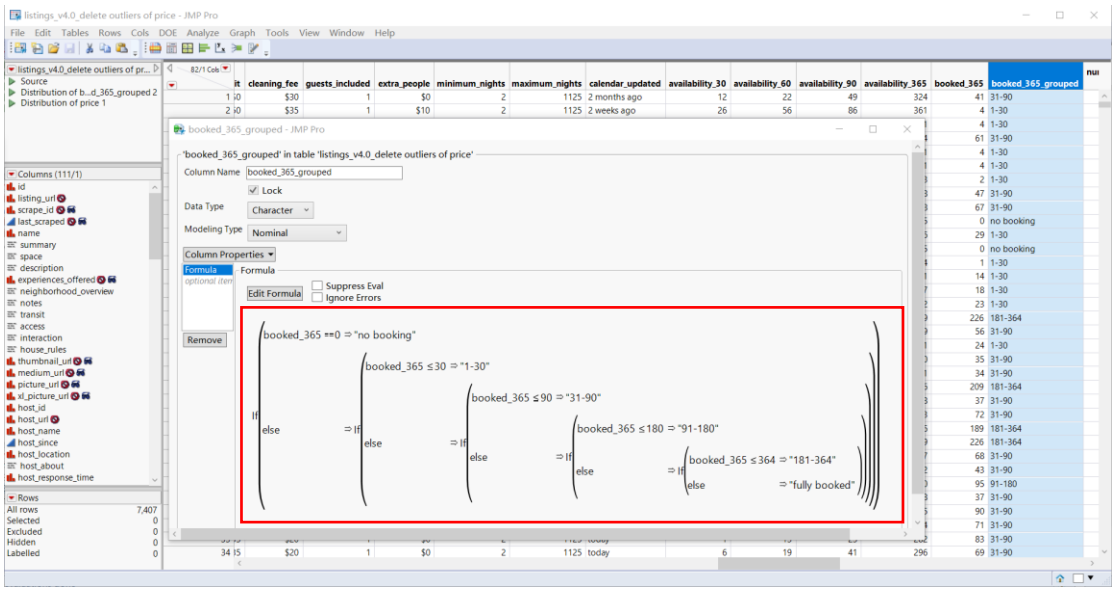


Figure 18

The table below shows that:

1) About 10% of the listings have no booking in a year, comparing with 17% are fully booked throughout a year.

2) About 43% of the listings are booked for over half a year, these listings should either be popular or have longer tenancy.

3) About half of the listings are booked for less than 90 days in sum. Taking no account of the "no booking" observations, about 41% of the listings in Singapore are certainly against the local policy of short-term tenancy.

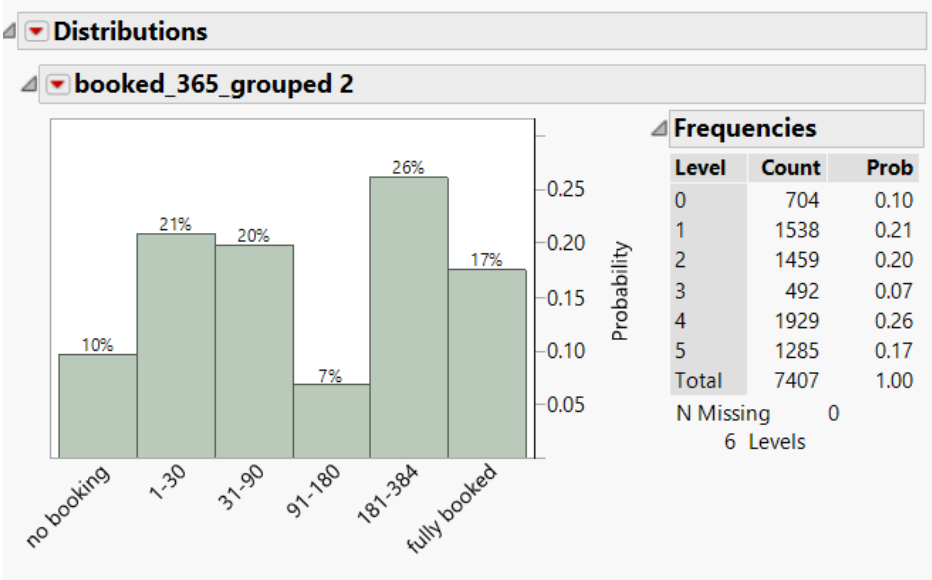


Figure 19

3.1.2. Influence factors of booking volume

In the long-term rental market, listings certainly have larger booking volume than those in the long-term rental market. To eliminate interference from tenancy length, we should shorten tenancy length of each leasing. For doing that, we assume one third of tenants would give review, and annual order volume for each listing is greater than 30, there should be more than 10 reviews received by each listing in a year, the conditions used to select applicable rows are showed below.

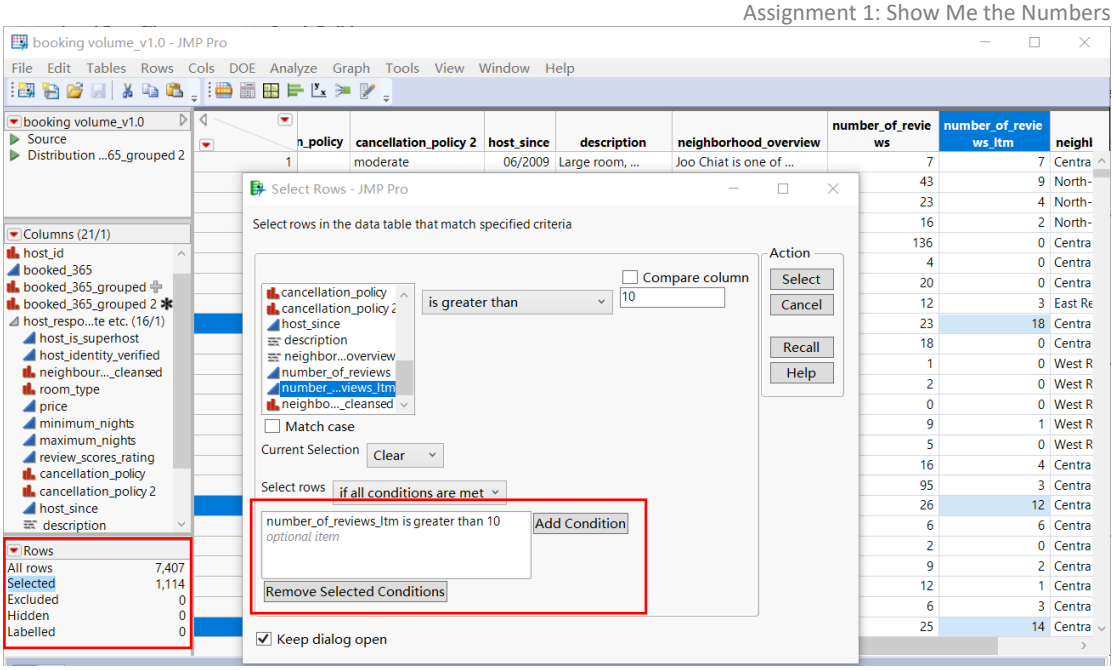


Figure 20

Considering the market based on such condition as a valid short-term rental market, the distribution of booking volume (Figure 21) shows that near 60% listings are booked for more than half a year.

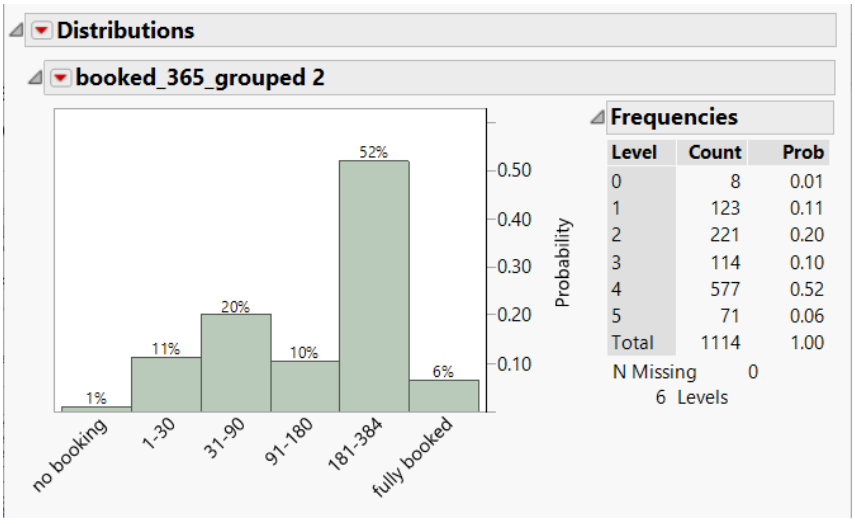


Figure 21

The chart below consists of 3 curve graphs:

1) the uppermost one shows that the average tenancy of the listings with booking volume over 180 days must be longer than those with booking volume less than 180 days, because the number of reviews should increase in accordance with the booking volume.

2) The last two graphs show that booking volume rise in accordance with review scores and superhost ratio that people like to book for a listing with a higher review score. Listings of superhost are more popular as well.

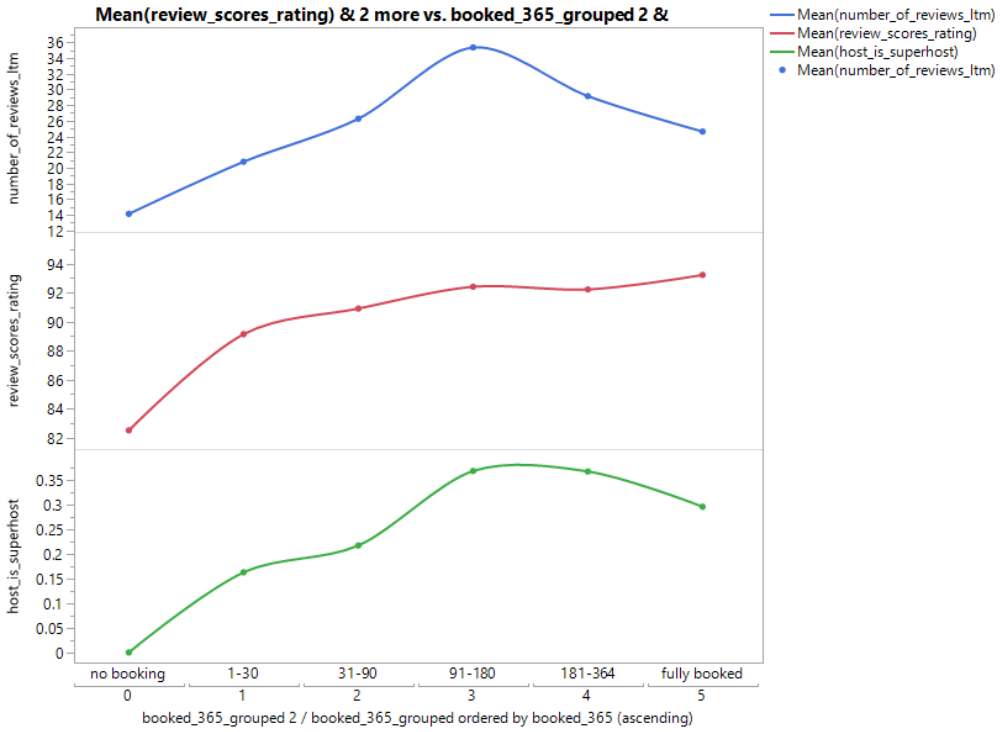


Figure 22

3.1.3. Explore the date distribution of booking volume in a year

In the data table “calendar”, we can get availability of each listing within a year. First, we need to prepare data and get variables “Month Year” and “booking status”. After deleting abnormal values of price, some listings offering unusual high price are removed from our analysis, this step is to minimal outliers’ influence on the mean of booking volume. To get the number of days booked every month, we can group date in accordance with month (Figure 23) and year and derive the booking status from variable “available” by the formula showed below (Figure 24).

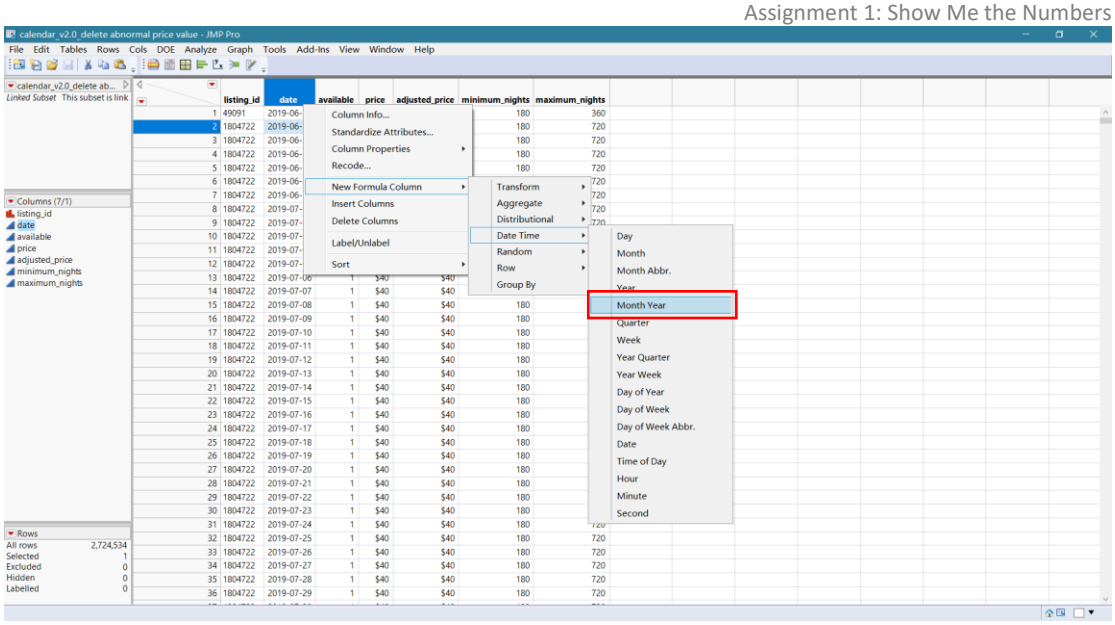


Figure 23

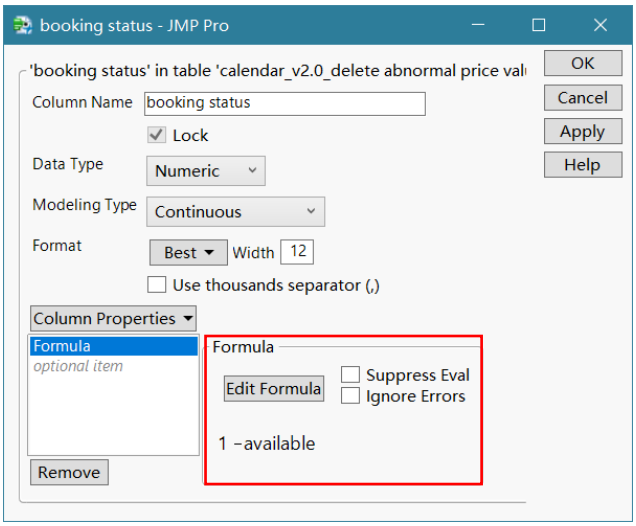


Figure 24

Then, get total booking volume of each listing in each month by grouping the same listing and the same date as shown below (Figure 25). The variable “Sum (booking status)” is the sum value we need (Figure 26).

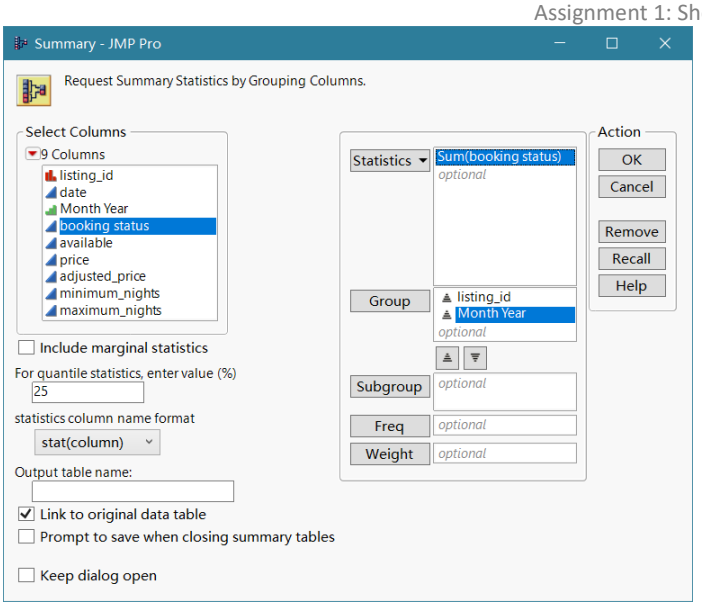


Figure 25

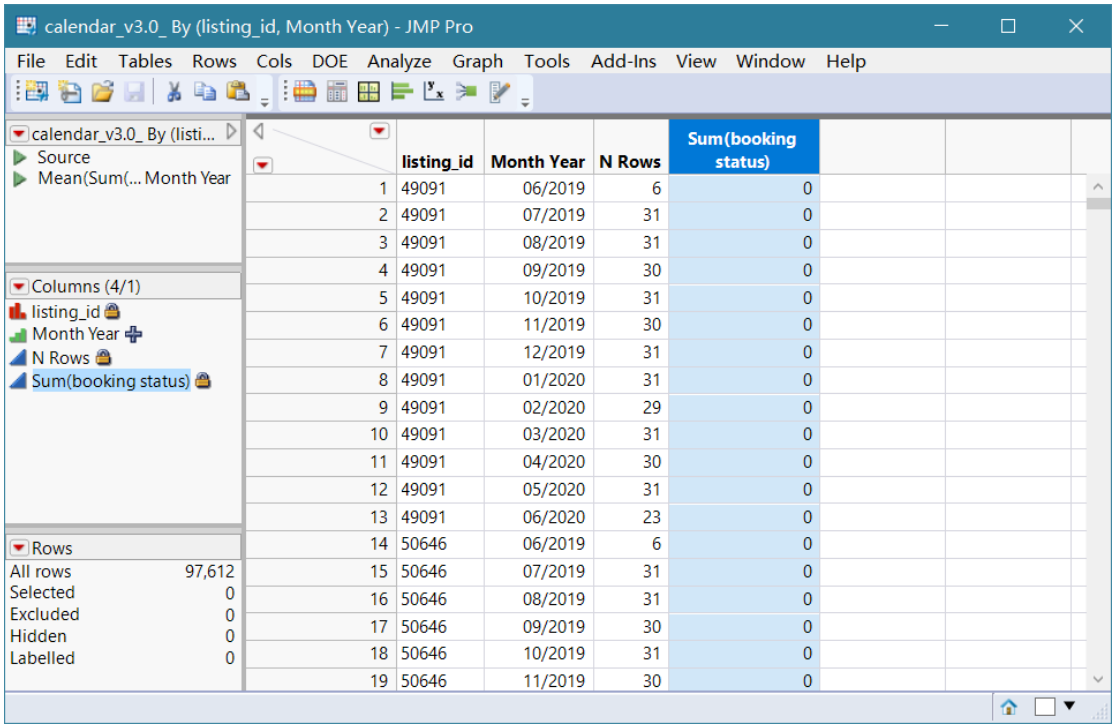


Figure 26

The bar charts below (Figure 27) shows the distribution of average booking volume in a year on monthly or daily basis. The vertical coordinates express the average number of booking days in a month for each listing.

1) Obviously, the coming month, July is the month with the largest booking volume, and September is the one with the lowest booking volume. November also shows a lower booking volume. Noticed we need to remove June from comparation, because these data are incomplete.

2) the booking volumes of months beside June, July, September and November fluctuate in a small range, within 1 day.

3) If we investigate the distribution on daily basis (Figure 28), we'll find out there's a dramatic dropping between July and September. The booking volume surge in October 2019 and January 2020 and drop in a few days followed by a stable fluctuation. So, January and October should be peak tourist seasons in Singapore. Knowing that Chinese cover the largest tourist proportion in Singapore, this trend is reasonable because Spring festival and National day are the longest holiday in China.

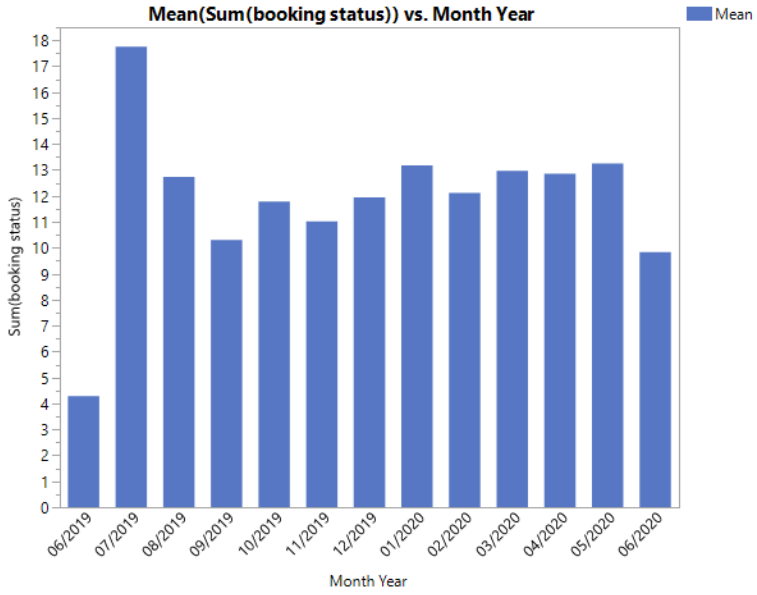


Figure 27

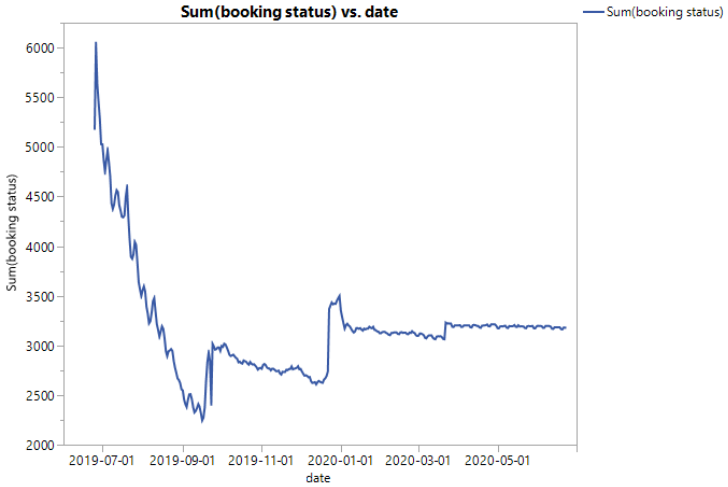


Figure 28

3.1.4. Further analysis into the hotel industry in Singapore

To further analyze whether the change of booking volume is in accordance with the change of tourist season, we can compare average booking volume with hotel's average occupancy rate.

We can find out off-season and peak tourist seasons in Singapore by exploring the occupancy rate of hotels in Singapore. Import hotel statistics from website of the Singapore Tourism Board (STB) to JMP and draw a relationship between month and average occupancy rate of hotels in Singapore in 2018 (Figure 29).

The graph shows that July, August and February are peak tourist seasons, while May and December are off-season. Such significant differences between months doesn't occur in the rental booking market, which implies that the rental market in Airbnb isn't a market only for tourists or other short-term tenants, this difference could be resulted from the longer average rental length in Airbnb. Besides, most people would not book a room months in advance, especially in the short-term rental market. July is the coming month that is approaching, so the booking volume should rise higher under this factor.

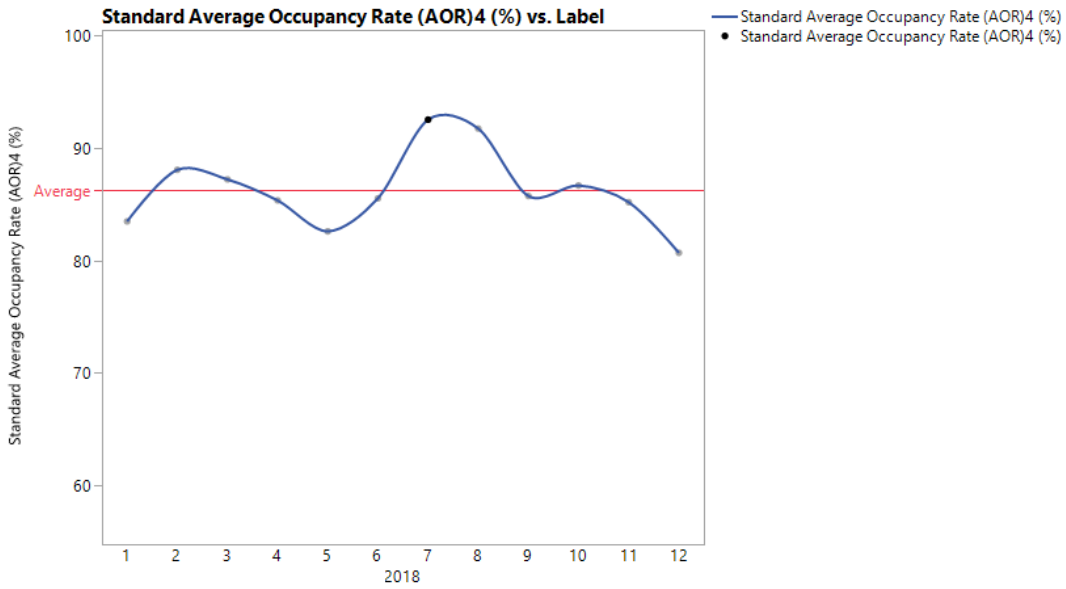


Figure 29

When we look at the tendency of average occupancy rate of hotels from 2012 to 2018 (Figure 30), it's obvious that average occupancy rate increased significantly in the last 2 years, whereas the number of available room-nights was increasing steadily, so gross lettings must increase dramatically in the last two years.

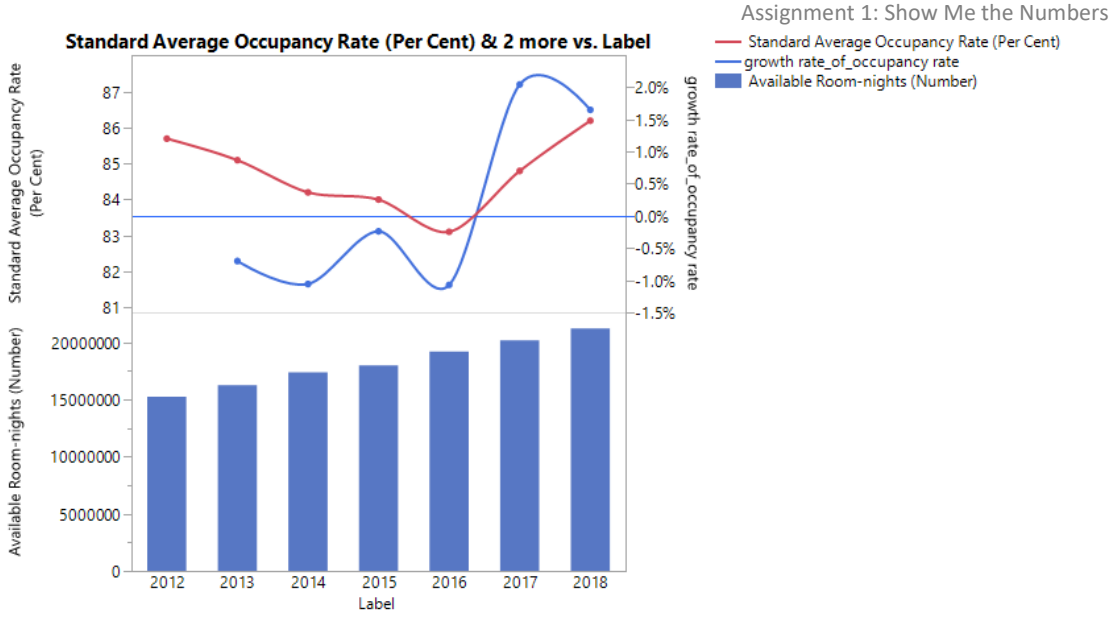


Figure 30

Comparing the trends shown above with the trend of reviews amount received by Airbnb in Singapore from 2012 to 2019 (Figure 31), we can find out that the number of reviews began to rise in 2015, so the opening of short-term rental market couldn't make a negative influence on the hotel industry before year 2015. As the development of Airbnb in Singapore after 2016, the trend of occupancy rate in hotels is accordant with the number of reviews received by Airbnb.

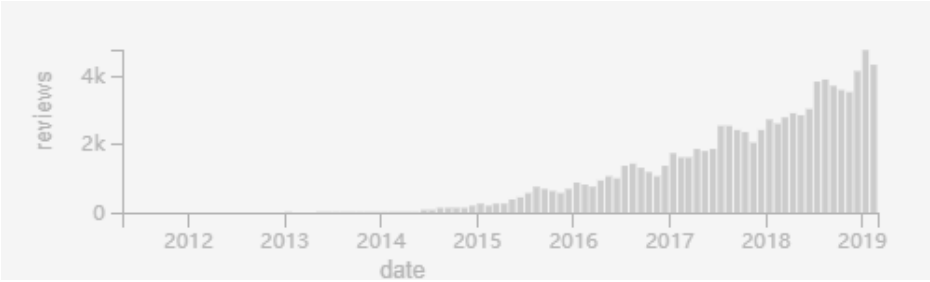


Figure 31

http://insideairbnb.com/singapore/?neighbourhood=&filterEntireHomes=false&filterHighlyAvailable=false&filterRecentReviews=false&filterMultiListings=false#

we also look at the variance of average room rate and room revenue of hotels in Singapore as shown below (Figure 32). Average room rate is drop whereas room revenue is rise from 2014 to 2017, it may because of an increasing tourist amount or an expansion in the number of hotels. The result is that opening short-term rental hasn't shown a negative influence on hotel industry and may exert positive effects on tourism.

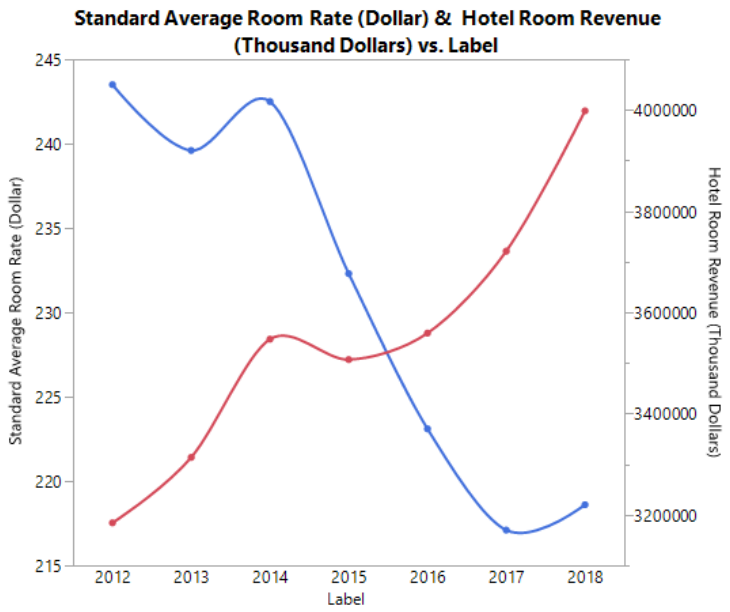


Figure 32

3.2. Insights into hosts in Airbnb

3.2.1. Hosts and tenants' identity verification

About 27% of the host in Airbnb had verified identity, and almost all the hosts don't require license and other verifications from tenants, which is against the policy.

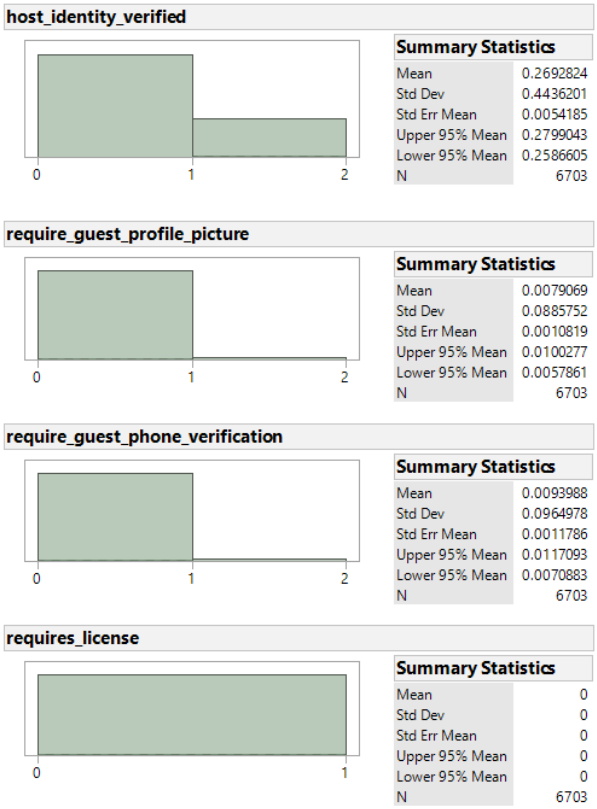


Figure 33

3.2.2. The difference between superhosts and non-superhosts

The graphs below show that about 90% of the hosts located in Singapore, and the remain 10% should have more than one house and would not share a living experience with tenants. About 15% are superhosts, whose listings got a higher average review scores rating (). However, superhosts own more listings on average than those who are not superhosts. Furthermore, to investigate average number of listings per host based on room type, the bar chart shows that superhosts are more likely to rent out entire house than rent out private or shared room.

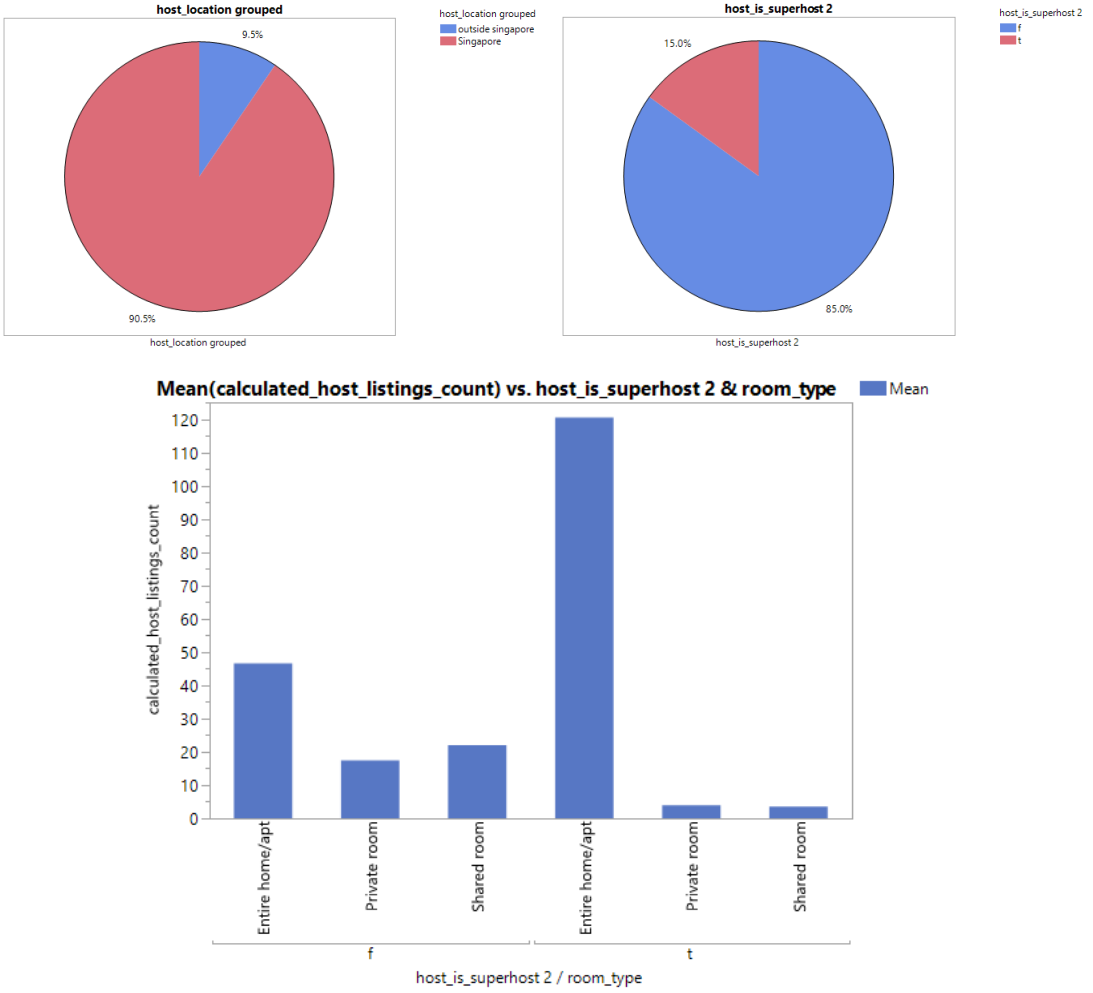


Figure 34

On average, the listings owned by superhosts received a higher review scores in six different aspects “accuracy”, “cleanliness”, “checkin”, “communication”, location” and “value”. Combined with the results above, superhosts are likely to rent out more than one listing but get higher comments from tenants.

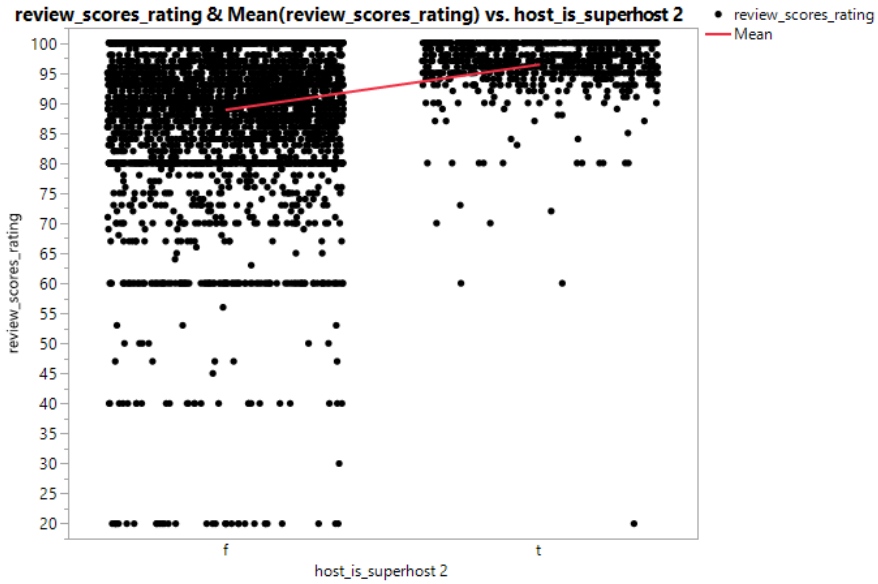


Figure 35

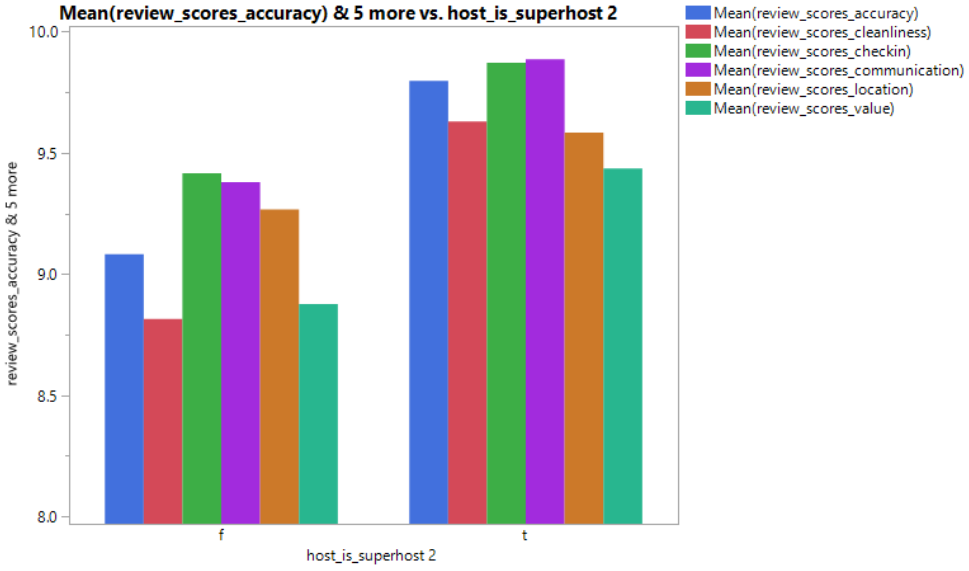


Figure 36

3.3. Insights into booked listings

3.3.1. Exploring text in variable “description” and “house rules”

To find out the most concerned features of listings by tenants and major house rules of the listings in Airbnb. To summarize these key words, use tool “text explorer” to find out the most phrase used and set minimum characters per word at 10 to get an informative Word Cloud, we can get results that:

1) the most concerned features by tenants are related to bed size, transportation, facilities offered, housekeeping, surrounding restaurants, surrounding attractions and so on.

2) the most concerned rule set by hosts is related to noisy making, that most hosts are concerned about keep a quiet and peaceful environment to neighborhoods and communities.

Term and Phrase Lists

Phrase	Count	N
queen size	2185	2
cooking utensils	1863	2
mrt station	1826	2
washing machine	1585	2
walking distance	1383	2
mins walk	1345	2
air conditioning	1336	2
hair dryer	1271	2
ironing board	1191	2
fully furnished	1164	2
dining table	1149	2

Phrase	Count	N
please refrain	399	2
11 pm	386	2
10 pm	383	2
house rules	324	2
loud noises	309	2
making loud	280	2
noises at night	275	3
inside room	275	2
live in the apartment	274	4
loud noises at night	274	4
please refrain from making	274	4
refrain from making loud	274	4
general please refrain	274	3
making loud noises	274	3

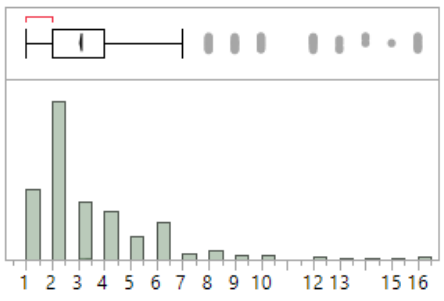
Figure 37

3.3.2. Accommodates of booked listings

Group accommodates volume into 2 groups: not over 6 tenants' group and over 6 tenants' group. Get the distribution of accommodates and draw a relationship between accommodates and price per night as shown below.

On average, every listing accommodates 3 tenants, and there still have 5.7% of listings accommodate more than 6 tenants, which is against the short-term rental policy. By looking through the change of average price according to accommodates number (Figure 40), in general, the average listing price per night rise with an increase in accommodates under nine.

accommodates



Quantiles

100.0%	maximum	16
99.5%		13
97.5%		8.4
90.0%		6
75.0%	quartile	4
50.0%	median	2
25.0%	quartile	2
10.0%		1
2.5%		1
0.5%		1
0.0%	minimum	1

Summary Statistics

Mean	3.1208414
Std Dev	2.1615629
Std Err Mean	0.0264018
Upper 95% Mean	3.1725973
Lower 95% Mean	3.0690855
N	6703

Figure 38

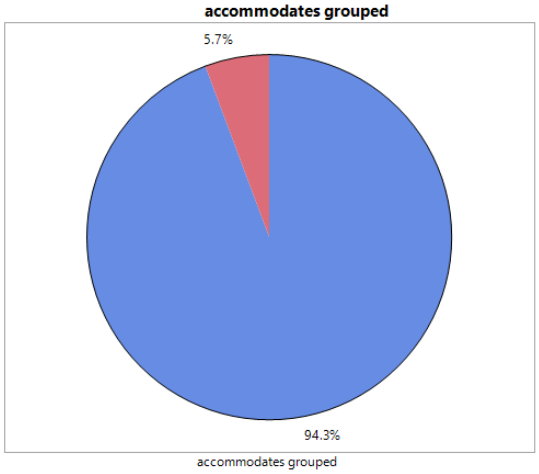


Figure 39

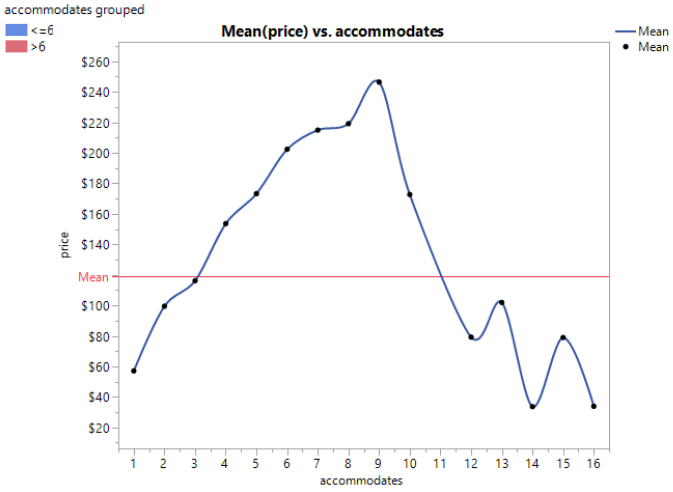


Figure 40

3.3.3. Room type of listings in Airbnb

The pie chart shows that only 5.9% of the listings offered by hosts who share a house with tenants. About half of the listings are renting entire apartments. On average, these listings are offered for 119 dollars per night.

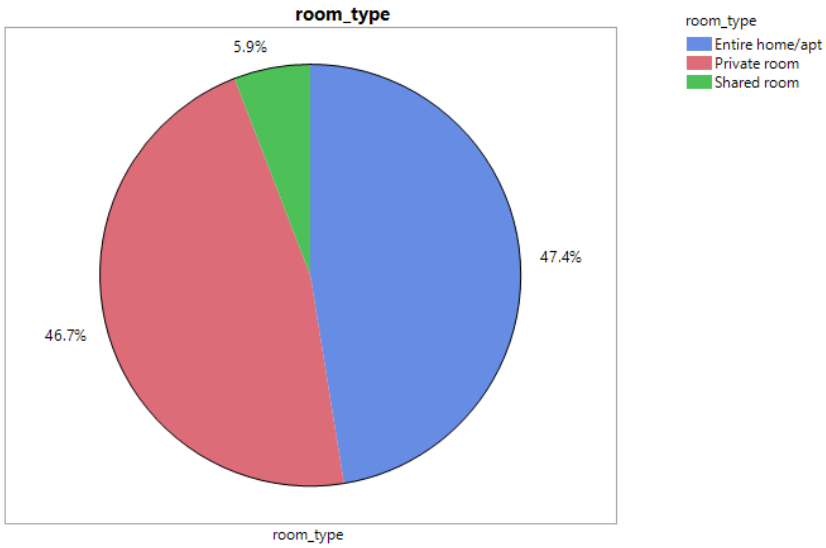
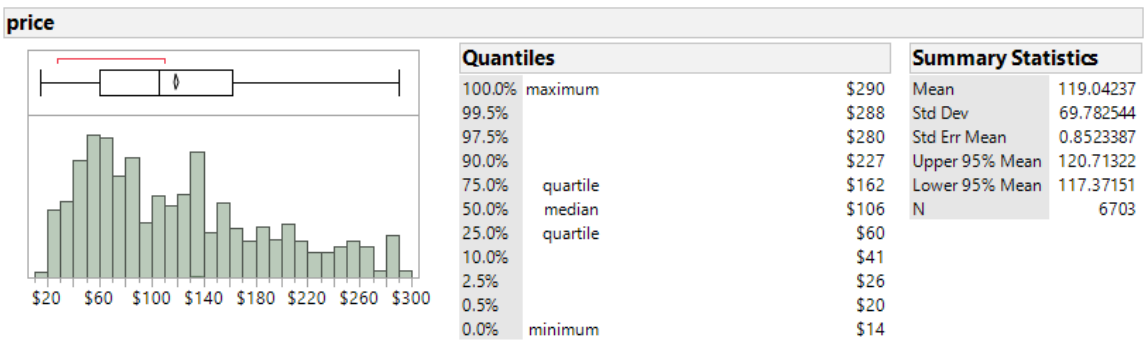


Figure 41



Quantiles		
100.0%	maximum	$290
99.5%		$288
97.5%		$280
90.0%		$227
75.0%	quartile	$162
50.0%	median	$106
25.0%	quartile	$60
10.0%		$41
2.5%		$26
0.5%		$20
0.0%	minimum	$14

Summary Statistics	
Mean	119.04237
Std Dev	69.782544
Std Err Mean	0.8523387
Upper 95% Mean	120.71322
Lower 95% Mean	117.37151
N	6703

Figure 42

3.4. Insights into reviews

3.4.1. Exploring tenants' comment to Airbnb

By exploring text meaning in the variable "comments" in table "reviews", we can find that most reviews received by Airbnb are positive, like "comfortable", "convenience", highly recommended", "great host", "nice place" and so on. This result shows that many tenants had good experience in Airbnb. Airbnb is advanced in short-term rental market in Singapore and get good comments from the tenants. It also indicates that short-term rental market is popular to tourists that it could give a positive influence on the tourism industry in Singapore.

Term and Phrase Lists

Phrase	Count	N
place to stay	5656	3
mrt station	5603	2
great location	5375	2
great place	4322	2
walking distance	3541	2
good location	3451	2
highly recommend	3355	2
great host	3159	2
nice place	2927	2
bus stop	2817	2
come back	2585	2
highly recommended	2572	2
great stay	2183	2
definitely recommend	2038	2
good place	2035	2
recommend this place	2023	3

Figure 43

3.4.2. Total number of reviews received so far

The table list the top 20 tenants who give the largest number of reviews from May 2011 onwards. The largest number of reviews received from the same tenants is 22 in total. About 93% of tenants just had given one piece of comments until June 2019.

reviewer_id	reviewer_name	N Rows
56833080	YatShan	22
3289968	Stephen	19
6406727	David	17
55688569	Vignes	16
10818291	Jerome	15
73609029	Nick	14
20732518	Tomi	13
17273265	Eun-Byul	12
21453171	Miranda	12
49863961	Ermin	12
125653187	Keerthiraj	12
841659	Hans	11
156145799	カズミ	11
195298862	Thit	11
11194254	Matthieu	10
16491329	Yuko	10
44074843	Andreas	10
66815639	Nai	10
97681492	Alex	10
104849977	Ilango	10

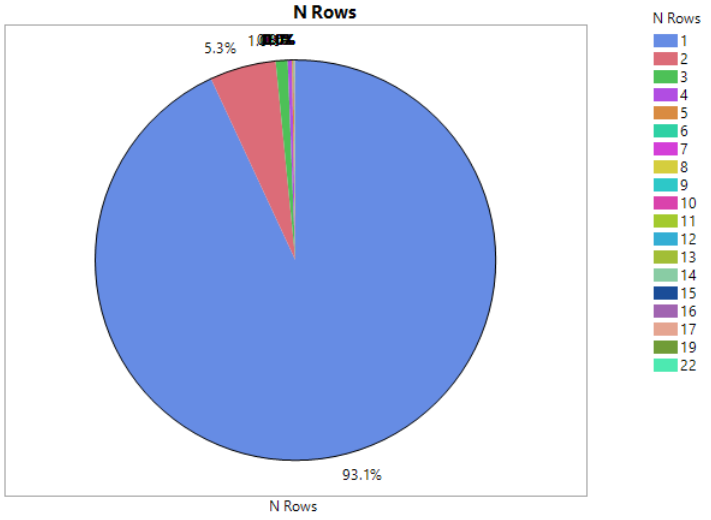


Figure 44

4. Confirmative analysis

4.1. Hypothesis testing 1

To confirm the conclusion that the average tenancy of the listings with booking volume over 180 days must be longer than those with booking volume less than 180 days, we need to hypothesize that listings with larger booking volume should receive more booking orders, as a result, they should receive more reviews in a year.

By doing this, first separate listings into two groups: listings with booking volume over 180 days and below 180 days. The distribution of reviews received by listings with a year booking volume below 180 days shows that the average number of reviews is 4.79.

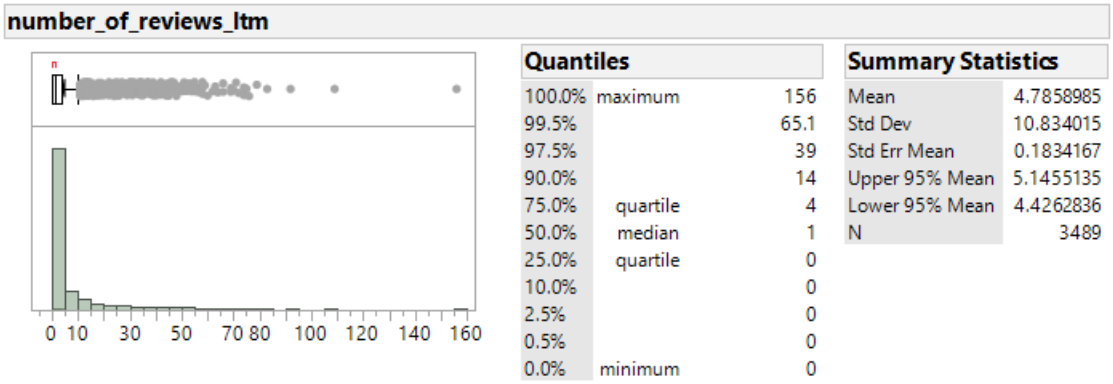


Figure 45

Do the hypothesis testing with the listings having a booking volume over 180 days as shown below.

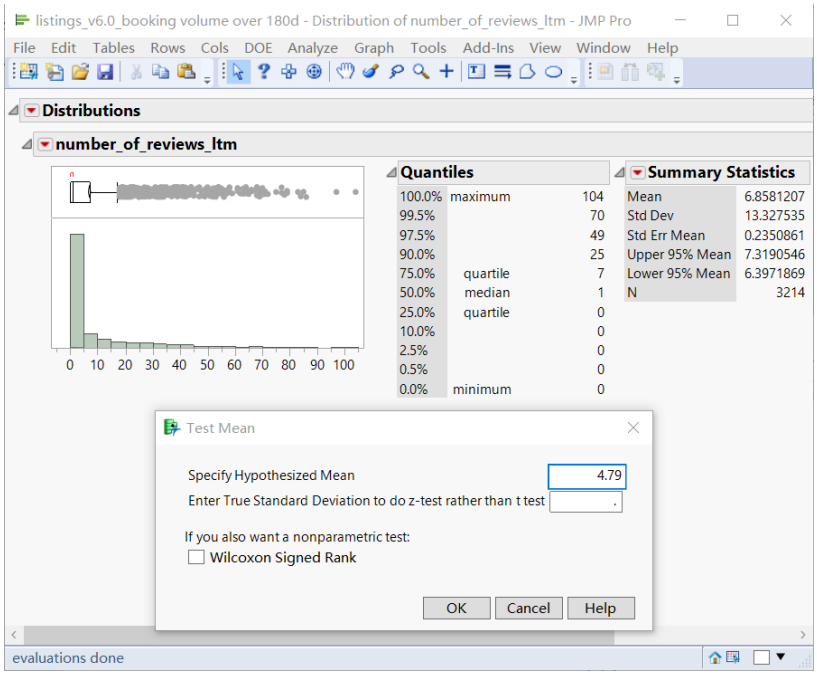
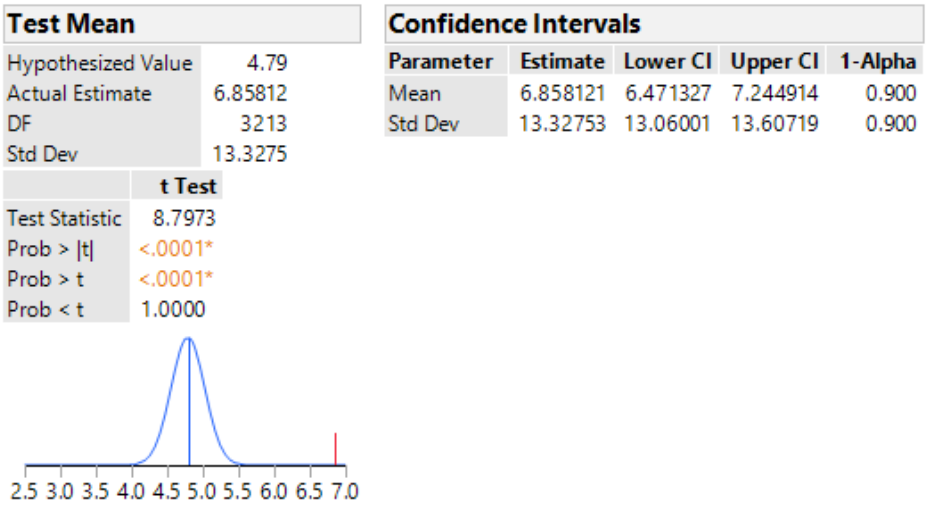


Figure 46

The result shows that the p value is much lower than 0.05, so we could reject the hypothesis that the listings with larger booking volume would receive the same number of booking orders or reviews as the listings with lower booking volume.



Test Mean

Hypothesized Value	4.79
Actual Estimate	6.85812
DF	3213
Std Dev	13.3275

	t Test
Test Statistic	8.7973
Prob > |t|	<.0001*
Prob > t	<.0001*
Prob < t	1.0000

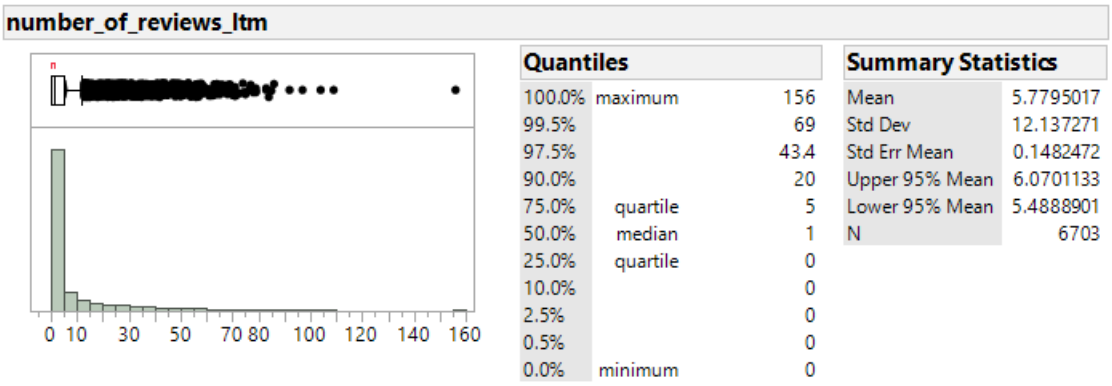
Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	6.858121	6.471327	7.244914	0.900
Std Dev	13.32753	13.06001	13.60719	0.900

Figure 47

4.2. Hypothesis testing 2

To verify the hypothesis that listings of more than one booking received more reviews, we use hypothesis testing shown below. The average number of reviews received in a year for the listings with at least one booking is 5.78.



number_of_reviews_ltm

Quantiles

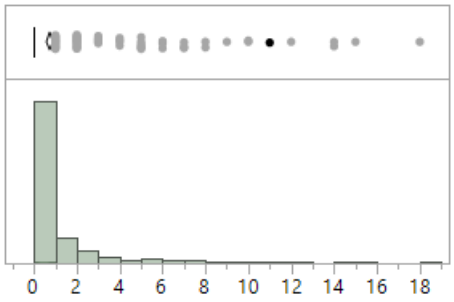
100.0%	maximum	156
99.5%		69
97.5%		43.4
90.0%		20
75.0%	quartile	5
50.0%	median	1
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	5.7795017
Std Dev	12.137271
Std Err Mean	0.1482472
Upper 95% Mean	6.0701133
Lower 95% Mean	5.4888901
N	6703

Figure 48

The result below shows that we got a lower enough p value to reject the hypothesis that listings of more than one booking received the same number of reviews.



Quantiles		
100.0%	maximum	18
99.5%		14.475
97.5%		7
90.0%		2
75.0%	quartile	0
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	0.7045455
Std Dev	2.0101257
Std Err Mean	0.0757595
Upper 95% Mean	0.8532874
Lower 95% Mean	0.5558036
N	704

Test Mean	
Hypothesized Value	5.78
Actual Estimate	0.70455
DF	703
Std Dev	2.01013

	t Test
Test Statistic	-66.994
Prob > |t|	<.0001*
Prob > t	1.0000
Prob < t	<.0001*

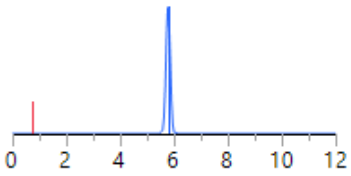
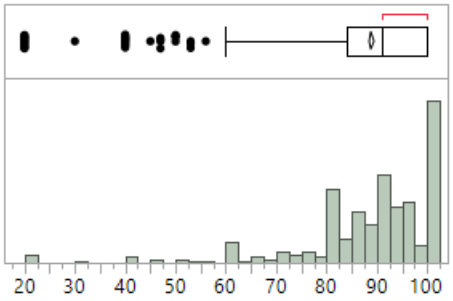


Figure 49

4.3. Hypothesis testing 3

To verify that superhosts should have an average higher review score than non-superhosts, we need to get the average review scores rating of superhosts, which is 88.84.

review_scores_rating



Quantiles		
100.0%	maximum	100
99.5%		100
97.5%		100
90.0%		100
75.0%	quartile	100
50.0%	median	91
25.0%	quartile	84
10.0%		77
2.5%		60
0.5%		20
0.0%	minimum	20

Summary Statistics	
Mean	88.842423
Std Dev	12.692667
Std Err Mean	0.2150373
Upper 95% Mean	89.264034
Lower 95% Mean	88.420811
N	3484

Figure 50

The result shows that we could reject the hypothesis that superhosts have the same average review score as non-superhosts.

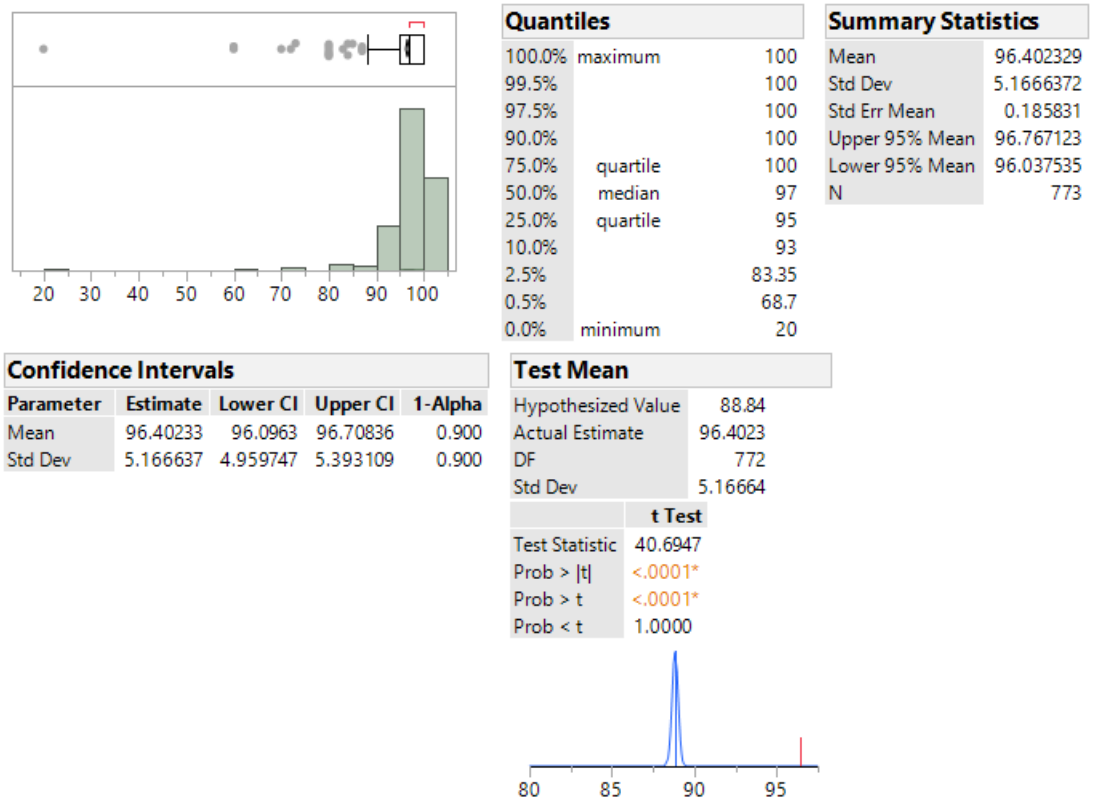


Figure 51

Data source

1. InsideAirbnb:
   http://insideairbnb.com/get-the-data.html
   http://insideairbnb.com/singapore/?neighbourhood=&filterEntireHomes=false&filterHighlyAvailable=false&filterRecentReviews=false&filterMultiListings=false
2. the Singapore Tourism Board (STB):
   https://www.stb.gov.sg/content/stb/en/statistics-and-market-insights/tourism-statistics/hotel-statistics.html
3. SingStat:
   https://www.tablebuilder.singstat.gov.sg/publicfacing/createDataTable.action?refId=