SCHOOL OF ENGINEERING AND TECHNOLOGY

ASSIGNMENT / PROJECT SUBMISSION FORM

PROGRAMME: Bachelor of Information Systems (Honours) (Data Analytics)

SEMESTER: March 2022

SUBJECT: IST2024 Applied Statistics

DEADLINE: 19 July 2022, 1159 pm

INSTRUCTIONS TO CANDIDATES

- This is an individual report submission.

IMPORTANT
The University requires students to adhere to submission deadlines for any form of assessment. Penalties are applied in relation to unauthorized late submission of work.

Lecturer's Remark (Use additional sheet if required)

I Alicia Chong Tsui Ying (Student's Name) 20074290 (Student ID) received the assignment and read the comments.

...........Alicia........... (Signature/Date)
(13th July 2022)

Academic Honesty Acknowledgement

"I Alicia Chong Tsui Ying (Student's Name) verify that this paper contains entirely my own work. I have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, I have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. I realize the penalties *(refer to page 16, 5.5, Appendix 2, page 44 of the student handbook diploma and undergraduate programme)* for any kind of copying or collaboration on any assignment."

…......... Alicia............. (Student's signature / Date)
(13th July 2022)

Table of Contents

1.0 INTRODUCTION 2

2.0 DESCRIPTIVE ANALYSIS AND DATA PRE-PROCESSING 3

2.1 Observe Variables Metadata 3

2.2 Convert Categorical Variable bathroom_text to Numerical Variable 4

2.3 Generate Frequency Table for Categorical Variables 4

2.4 Convert Categorical Variable host_response_time to Numerical Variable 5

2.5 Summary Statistics of Numeric Variables 5

2.6 Distribution plot and box plot 6

3.0 STATISTICAL MODELLING AND ANALYSIS 9

3.1 Linear Regression: Explanatory Analysis on the Price of Property Rentals 9

3.1.1 Scatter Plot Matrix 10

3.1.2 Model selection 12

3.1.3 Linear Regression Analysis 13

3.3.4 Regression Diagnostic 19

3.2 Logistic Regression: Explanatory Analysis on host_is_superhost 21

3.2.1 Bivariate Analysis 21

3.2.2 Logistic Regression Analysis 24

3.3 ANOVA: Compare the means of review_scores_communication with different host_response_time_num 31

3.3.1 Descriptive Statistics Across Groups with Box and Whiskers Plot 31

3.3.2 Analysis of Variance (ANOVA) 32

4.0 CONCLUSION 34

5.0 APPENDIX 35

1.0 Introduction

For this project, a data set containing the records on short-terms property rentals for entire homes was given for critical analysis. As a basic overview, the given dataset has 30 columns and 2095 rows of data regarding information on host details, property details, property reviews information and reviews scores. Among the 30 columns, there are 4 nominal, 2 ordinal, 14 discrete, 8 continuous variables and 2 additional observation identifiers (*id*, *host_id*). The nominal variables are *host_is_superhost*, *host_has_profile_pic*, *host_identified_verified* and *property_type*; the ordinal variable are *host_response_time* and *bathrooms_text*; the discrete variables are *host_since, host_listings_count, accommodates, bedrooms, beds, minimum_nights, maximum_nights, availability_30, availability_60, availability_90, availability_365, number_of_reviews, number_of_reviews_ltm* and *number_of_reviews_130d*; the continuous variables are *price, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_communication, review_scores_location, review_scores_value and review_per_month*.

The analysis objectives of this project are to as follow:

1. To estimate the relationship between the daily price of property rentals (*price*) and other variables related to property details and review scores in this dataset.
2. To estimate the relationship between *host_is_superhost* and other variables related to the host details and review scores predictors.
3. To test whether the ratings score for ease of communication (*review_scores_communication*) is affected by the host's response time (*host_response_time_num*).

To achieve objective 1, linear regression analysis will be conducted as the response variable (*price)* is a numerical variable. To achieve objective 2, binary logistic regression analysis will be performed as the response variable (*host_is_superhost*) is a categorical variable. To reach objective 3, analysis of variance (ANOVA) will be conducted to test the relationship between the categorical variable (*host_response_time_num*) and numeric variable (*review_scores_communication*) by testing the difference between the population means of *review_scores_communication* grouped by *host_response_time_num*. SAS Studio is used as the SAS programming interface to perform analysis on our data set for this project.

2.0 Descriptive Analysis And Data Pre-processing

Before performing statistical modelling and analysis, descriptive analysis techniques are deployed to summarize and explore the behaviour of the data involved in the study. Statistical techniques such as frequency distribution, measures of central tendency and measures of dispersion were used. Furthermore, distribution plots and box plots are generated to visualize the distribution of values for numeric variables. Appropriate data pre-processing techniques were also deployed during the descriptive analysis procedure.

2.1 Observe Variables Metadata

To get an overview of the data set, we first observed the PROC CONTENTS table that reports metadata about the variables of our dataset that was interpreted by SAS studio (see Figure 1).

Variables in Creation Order

#	Variable	Type	Len	Format	Informat
1	id	Num	8	BEST12.	BEST32.
2	host_id	Num	8	BEST12.	BEST32.
3	host_since	Num	8	MMDDYY10.	MMDDYY10.
4	host_response_time	Char	14	$14.	$14.
5	host_is_superhost	Char	1	$1.	$1.
6	host_listings_count	Num	8	BEST12.	BEST32.
7	host_has_profile_pic	Char	1	$1.	$1.
8	host_identity_verified	Char	1	$1.	$1.
9	property_type	Char	26	$26.	$26.
10	accommodates	Num	8	BEST12.	BEST32.
11	bathrooms_text	Char	9	$9.	$9.
12	bedrooms	Num	8	BEST12.	BEST32.
13	beds	Num	8	BEST12.	BEST32.
14	price	Num	8	NLNUM12.	NLNUM32.
15	minimum_nights	Num	8	BEST12.	BEST32.
16	maximum_nights	Num	8	BEST12.	BEST32.
17	availability_30	Num	8	BEST12.	BEST32.
18	availability_60	Num	8	BEST12.	BEST32.
19	availability_90	Num	8	BEST12.	BEST32.
20	availability_365	Num	8	BEST12.	BEST32.
21	number_of_reviews	Num	8	BEST12.	BEST32.
22	number_of_reviews_ltm	Num	8	BEST12.	BEST32.
23	number_of_reviews_l30d	Num	8	BEST12.	BEST32.
24	review_scores_rating	Num	8	BEST12.	BEST32.
25	review_scores_accuracy	Num	8	BEST12.	BEST32.
26	review_scores_cleanliness	Num	8	BEST12.	BEST32.
27	review_scores_communication	Num	8	BEST12.	BEST32.
28	review_scores_location	Num	8	BEST12.	BEST32.
29	review_scores_value	Num	8	BEST12.	BEST32.
30	reviews_per_month	Num	8	BEST12.	BEST32.

Figure 1: Data set variables metadata (Code in Appendix Figure 1)

2.2 Convert Categorical Variable *bathroom_text* to Numerical Variable

Upon observation of Figure 1, it is identified that it would be appropriate to clean and convert the categorical variable *bathroom_text* into a numerical variable for further analysis. Figure 2 shows observations value of the *bathroom_text* variable and a new variable named bathrooms that holds the converted numerical values of the *bathroom_text* variable.

	bathrooms_text	bathrooms
1	1 bath	1
2	1 bath	1
3	1 bath	1
4	1 bath	1
5	1 bath	1
6	1 bath	1
7	3 baths	3
8	2.5 baths	2.5

Figure 2 Convert categorical variable bathroom_text to numerical variable (Code in Appendix Figure 2)

2.3 Generate Frequency Table for Categorical Variables

A frequency table is generated for each categorical variable, namely *host_is_superhost*, *host_has_profile_pic*, *host_identified_verified* and *property_type* (see Figure 3, 4, 5, 6 and 7).

host_is_superhost	Frequency	Percent
f	1103	52.65
t	992	47.35

Figure 3: Frequency Table for host_is_superhost (Code in Appendix Figure 3)

host_identity_verified	Frequency	Percent
f	362	17.28
t	1733	82.72

Figure 4: Frequency Table for host_identity_verified variable (Code in Appendix Figure 3)

host_has_profile_pic	Frequency	Percent
f	4	0.19
t	2091	99.81

Figure 5: Frequency Table for host_has_profile_pic variable (Code in Appendix Figure 3)

host_response_time	Frequency	Percent
a few days or	20	0.95
within a day	26	1.24
within a few h	65	3.10
within an hour	1984	94.70

Figure 6: Frequency Table for host_response_time variable (Code in Appendix Figure 3)

property_type	Frequency	Percent
Entire bungalow	7	0.33
Entire condominium (condo)	269	12.84
Entire cottage	33	1.58
Entire guest suite	83	3.96
Entire guesthouse	50	2.39
Entire loft	18	0.86
Entire place	2	0.10
Entire rental unit	758	36.18
Entire residential home	770	36.75
Entire serviced apartment	16	0.76
Entire townhouse	67	3.20
Entire villa	19	0.91
Tiny house	3	0.14

Figure 7: Frequency Table for property_type variable (Code in Appendix Figure 3)

2.4 Convert Categorical Variable *host_response_time* to Numerical Variable

It is observed that the variable levels of the *host_response_time* variable can be sorted to a particular order with "within an hour" being the least response time and "a few days or more" being the longest response time. Therefore, the *host_response_time* variable is encoded into to numeric variables. The values "within an hour", "within a few hour", "within a day" and "a few days or more" are encoded to the numbers 1 to 4 respectively. The encoded variable is then assigned to a new variable named *host_response_time_num* (see Appendix Figure 4 for code).

2.5 Summary Statistics of Numeric Variables

After pre-processing our data, the summary statistics for each numeric variables is generated. In Figure 8, the summary statistics table shows the basic statistical measures such as the mean, median, range, standard deviation, minimum, maximum, number of observations, and number of missing values of the variables. It is observed that there are quite a number of missing values for the variables *bedrooms*, *beds*, *review_scores_rating*, *review_scores_accuracy*, *review_scores_cleanliness*, *review_scores_communication*, *review_scores_location*, *review_scores_value and review_per_month*. By observing the mean, median, range, standard deviation, minimum and maximum statistics of the variables, we do not identify any data anomaly.

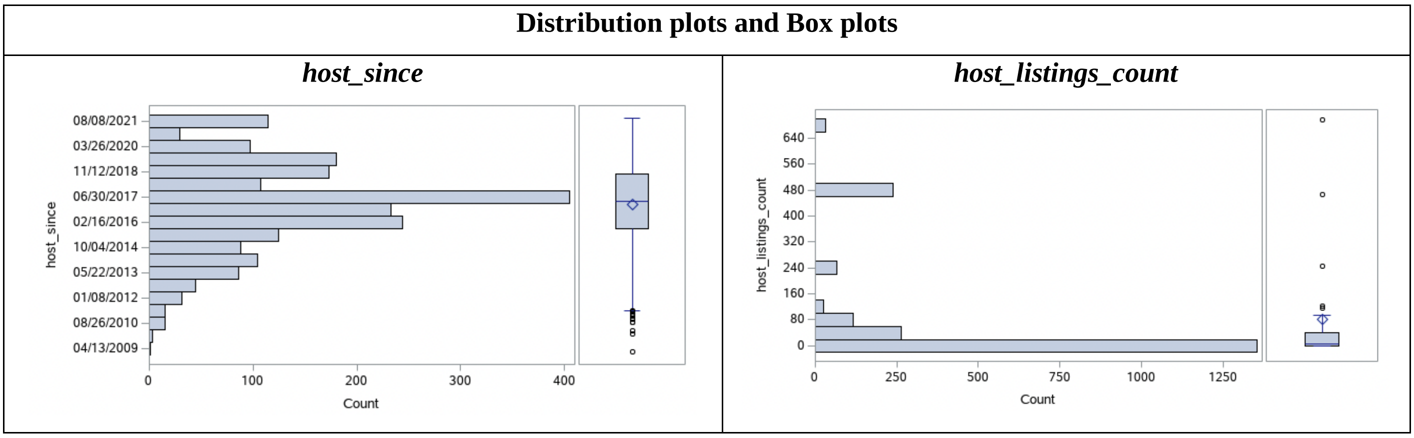
Summary Statistics for numeric variables

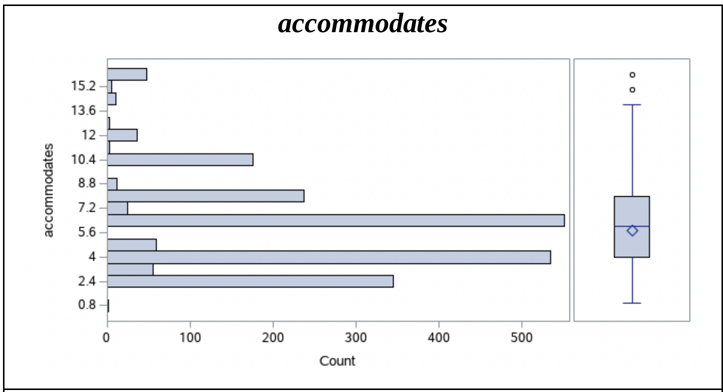
Variable	Mean	Median	Range	Std Dev	Minimum	Maximum	N	N Miss
host_listings_count	83.14	6.00	696.00	165.56	0.00	696.00	2095	0
host_response_time_num	1.08	1.00	3.00	0.40	1.00	4.00	2095	0
accommodates	5.71	6.00	15.00	3.02	1.00	16.00	2095	0
bathrooms	1.77	2.00	11.00	1.00	1.00	12.00	2095	0
bedrooms	2.42	2.00	22.00	1.51	1.00	23.00	1995	100
beds	3.34	3.00	29.00	2.46	1.00	30.00	2072	23
price	255.80	192.00	4077.00	230.38	45.00	4122.00	2095	0
minimum_nights	6.03	2.00	198.00	16.41	1.00	199.00	2095	0
maximum_nights	643.59	365.00	1123.00	471.59	2.00	1125.00	2095	0
availability_30	13.53	14.00	30.00	8.68	0.00	30.00	2095	0
availability_60	32.67	36.00	60.00	17.54	0.00	60.00	2095	0
availability_90	54.99	62.00	90.00	25.65	0.00	90.00	2095	0
availability_365	200.86	216.00	365.00	118.51	0.00	365.00	2095	0
number_of_reviews	56.74	30.00	563.00	74.23	0.00	563.00	2095	0
number_of_reviews_ltm	17.43	12.00	359.00	21.46	0.00	359.00	2095	0
number_of_reviews_l30d	1.41	1.00	12.00	1.79	0.00	12.00	2095	0
review_scores_rating	4.77	4.85	5.00	0.33	0.00	5.00	1880	215
review_scores_accuracy	4.82	4.90	4.00	0.31	1.00	5.00	1879	216
review_scores_cleanliness	4.80	4.87	4.00	0.29	1.00	5.00	1879	216
review_scores_communication	4.84	4.94	4.00	0.29	1.00	5.00	1879	216
review_scores_location	4.77	4.87	4.00	0.32	1.00	5.00	1879	216
review_scores_value	4.75	4.82	4.00	0.30	1.00	5.00	1879	216
reviews_per_month	3.62	2.09	98.65	6.18	0.03	98.68	1880	215

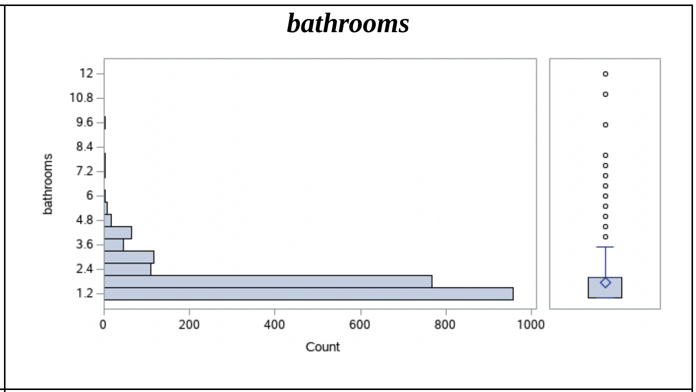
Figure 8: Summary Statistics for Numeric Variables (Code in Appendix Figure 5)

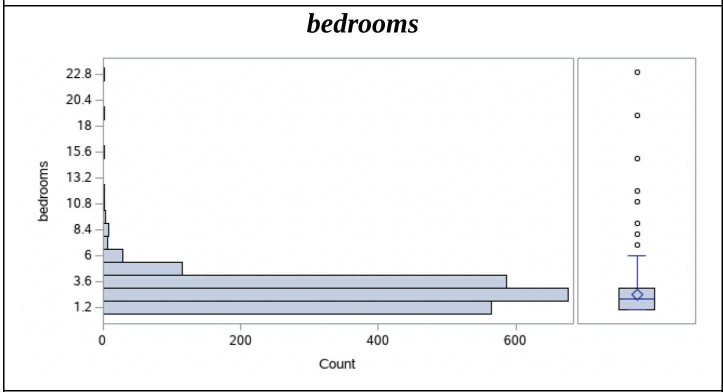
2.6 Distribution plot and box plot

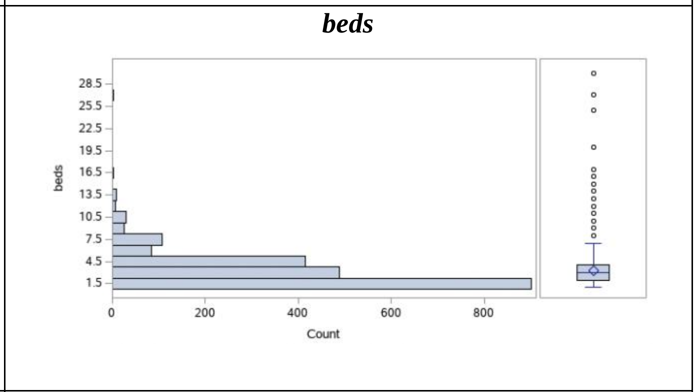
To visualize the distribution of values for each numeric value and detect outliers in our data, a distribution plot and box plot is generated for each numeric variable (see Figure 9). By observing the boxplots, it is apparent that all variables excluding the variables *availability_30, availability_60, availability_90, availability_365,* have some potential outliers. Therefore, the outliers have to be taken into considerations and further investigation on the outliers is needed to identify if the outliers are true outliers or outliers that is due to faulty data. Furthermore, it is observed that the variables *host_listings_count, bathrooms, bedrooms, beds, price, minimum_nights, number_of_reviews, number_of_reviews_ltm, number_of_reviews_130d, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_communication, review_scores_location, review_scores_value* and *review_per_month* have a highly skewed distribution.

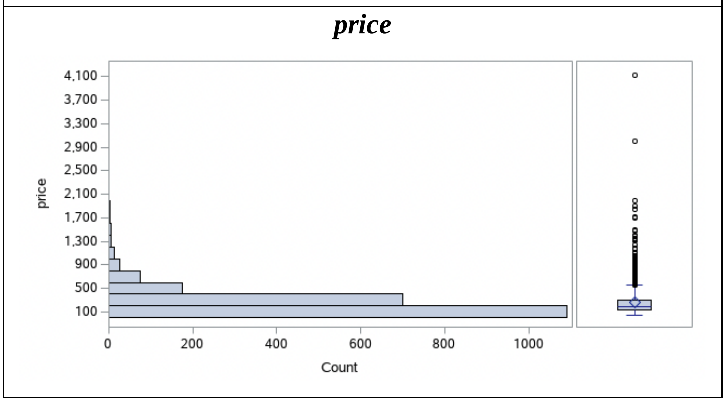


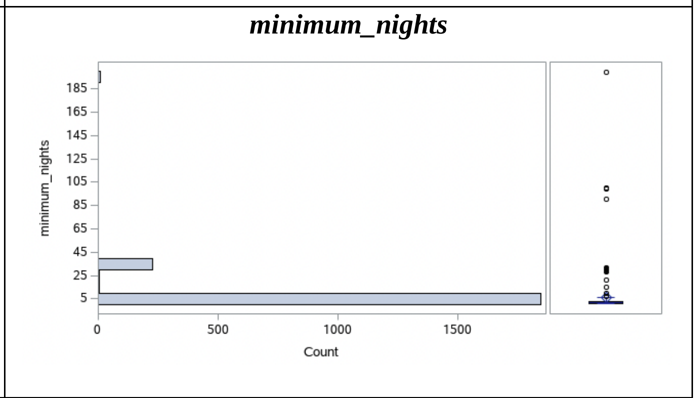

accommodates
accommodates
15.2
13.6
12
10.4
8.8
7.2
5.6
4
2.4
0.8
0
100
200
300
400
500
Count

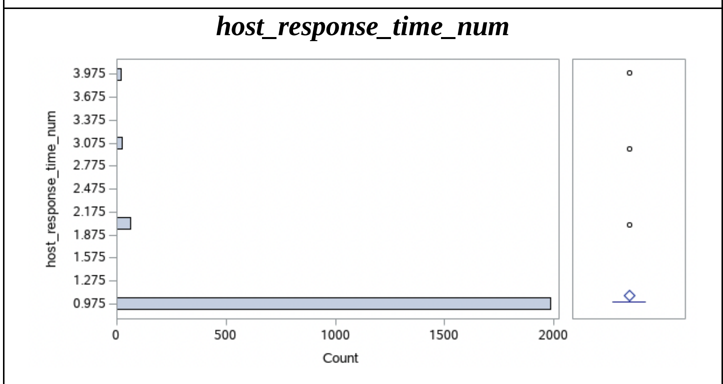

bathrooms
bathrooms
12
10.8
9.6
8.4
7.2
6
4.8
3.6
2.4
1.2
0
200
400
600
800
1000
Count

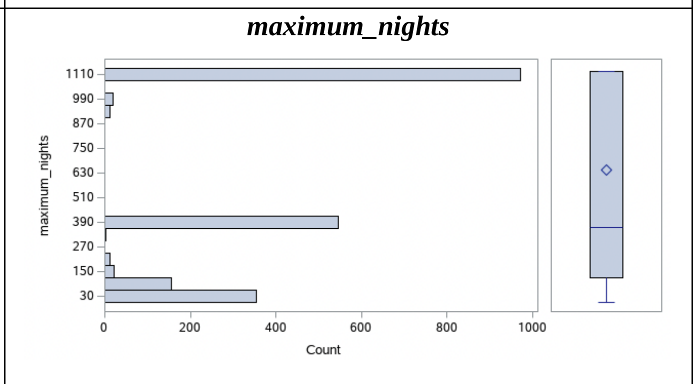

bedrooms
bedrooms
22.8
20.4
18
15.6
13.2
10.8
8.4
6
3.6
1.2
0
200
400
600
Count

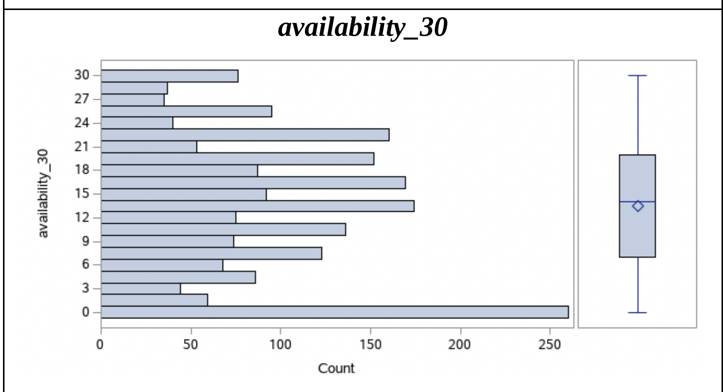

beds
beds
28.5
25.5
22.5
19.5
16.5
13.5
10.5
7.5
4.5
1.5
0
200
400
600
800
Count

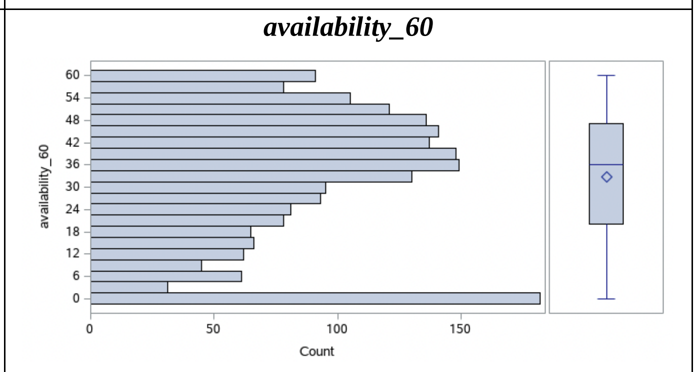

price
price
4,100
3,700
3,300
2,900
2,500
2,100
1,700
1,300
900
500
100
0
200
400
600
800
1000
Count


minimum_nights
minimum_nights
185
165
145
125
105
85
65
45
25
5
0
500
1000
1500
Count

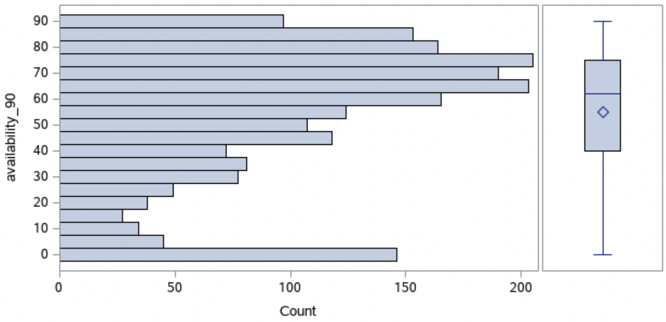

host_response_time_num
host_response_time_num
3.975
3.675
3.375
3.075
2.775
2.475
2.175
1.875
1.575
1.275
0.975
0
500
1000
1500
2000
Count


maximum_nights
maximum_nights
1110
990
870
750
630
510
390
270
150
30
0
200
400
600
800
1000
Count

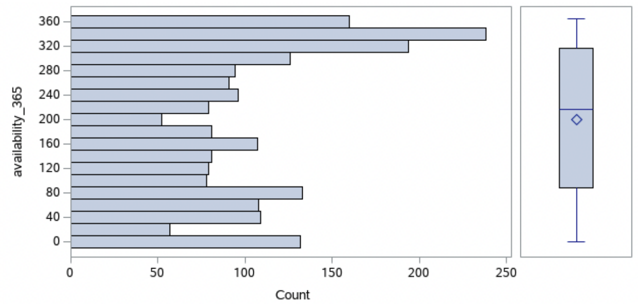

availability_30
availability_30
30
27
24
21
18
15
12
9
6
3
0
0
50
100
150
200
250
Count


availability_60
availability_60
60
54
48
42
36
30
24
18
12
6
0
0
50
100
150
Count

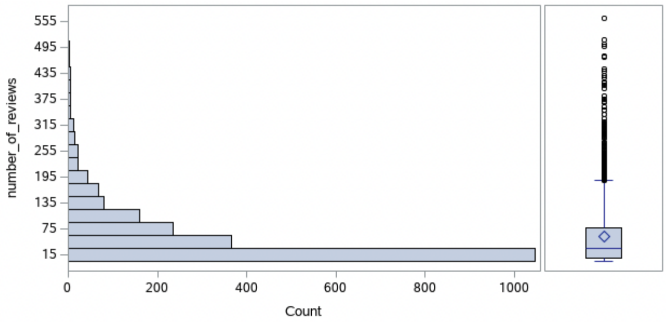
availability_90



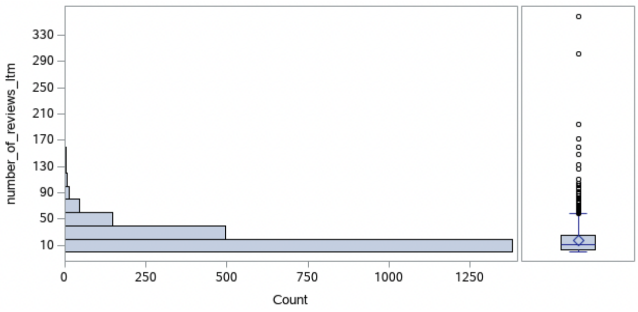
availability_365



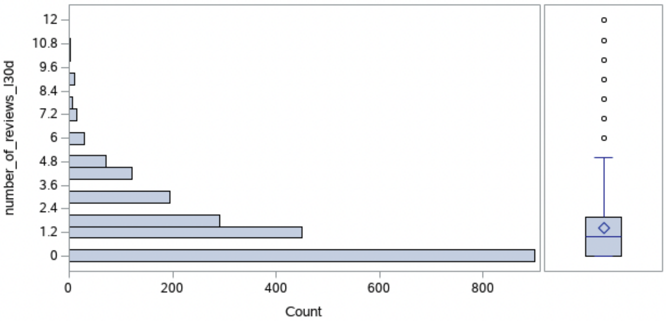
number_of_reviews



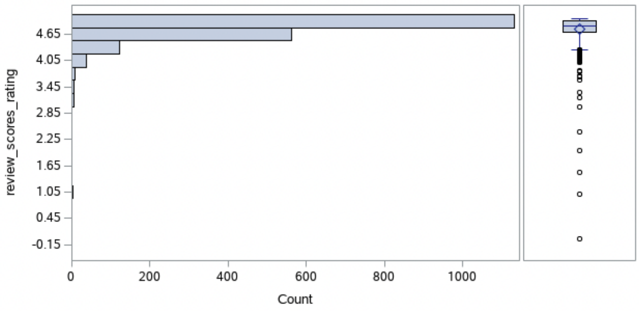
number_of_reviews_ltm



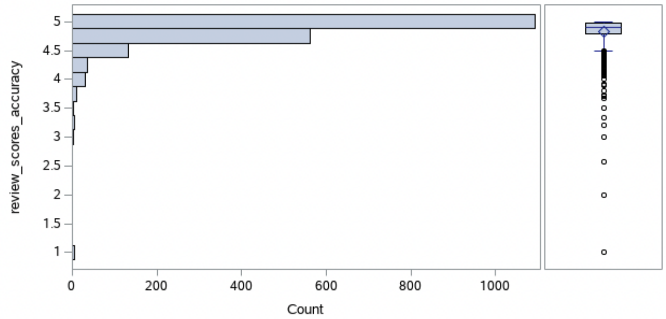
number_of_reviews_l30d



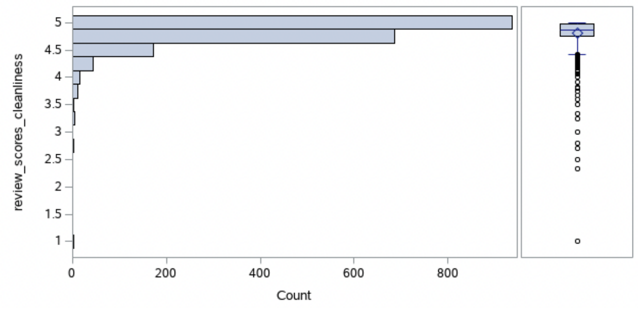
review_scores_rating



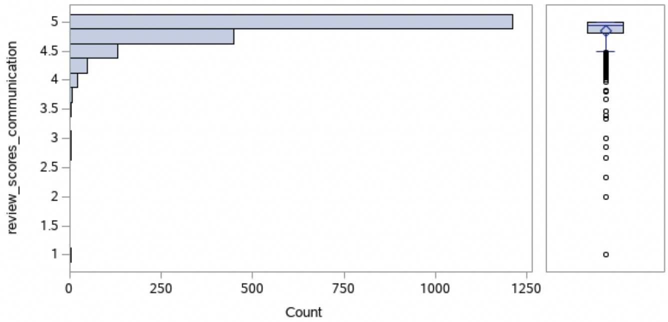
review_scores_accuracy



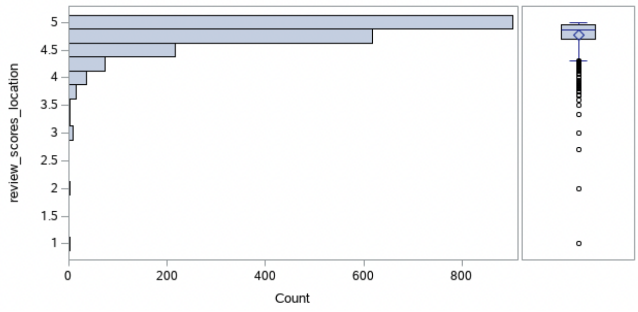
review_scores_cleanliness



review_scores_communication



review_scores_location



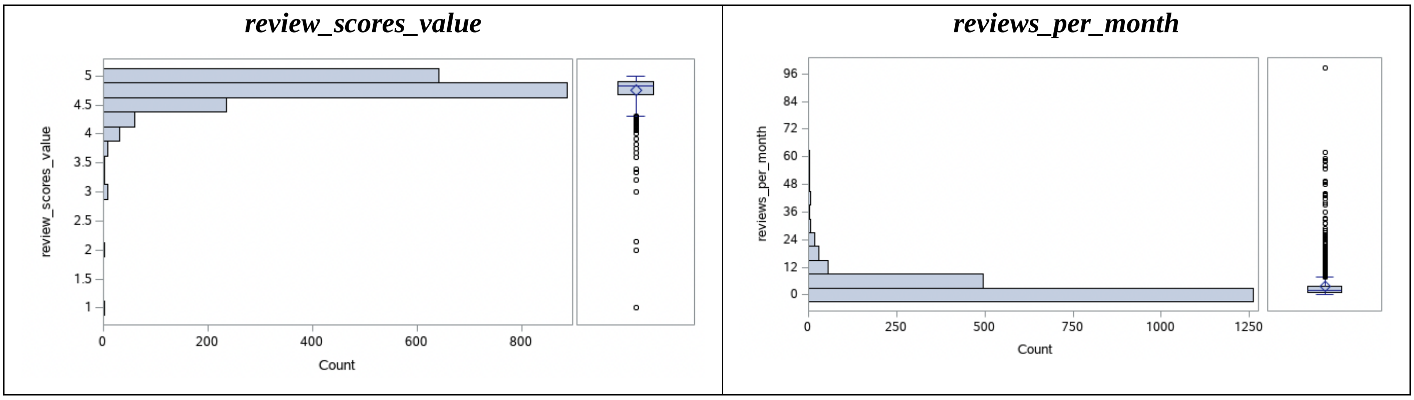


Figure 9: Distribution plots and Box plots for numeric variables (Code in Appendix Figure 6)

3.0 Statistical Modelling and Analysis

After performing descriptive analysis on our data set, statistical modelling and analysis is conducted to meet the objectives of this study. The following content in this section will be divided into 3 parts for 3 different statistical techniques:

1. **Linear Regression**: Explanatory Analysis on the Price of Property Rentals (*price*) and other variables related to property details and review scores
2. **Logistic Regression**: Explanatory Analysis on host_is_superhost and other variables related to the host details and review scores predictors.
3. **ANOVA**: Compare the means of review_scores_communication with different host_response_time_num

3.1 Linear Regression: Explanatory Analysis on the Price of Property Rentals

To achieve objective 1, linear regression analysis will be conducted as the response variable (*price)* is a numerical variable. This section will aim to estimate the relationship between price of property rentals and other potential variables that can predict the response variable such as *host_listings_count, accommodates, bathrooms, bedrooms, beds, availability_30, availability_60, availability_90, availability_365, number_of_reviews_130d, reviews_scores_rating, review_scores_accuracy, review_scores_communication, review_scores_location, review_scores_value, number_of_reviews_ltm, minimum_nights, maximum_nights, host_response_time_num, reviews_per_month, number_of_reviews* and *review_scores_cleanliness.*

3.1.1 Scatter Plot Matrix

Before performing statistical modelling to investigate the relationship between price of property rentals and other variables, a scatter plot matrix is constructed to investigate the linear relationships between variables and to check for outliers. As seen in Figure 10, the variable price and another 21 continuous variables are plotted against each other. It is observed that variables *accommodates*, *bedrooms*, *bathrooms* and *bath* are suggested to have a moderate linear correlation with *price*. Other variables such as *host_listings_count*, *availability_30*, *availability_60*, *availability_90*, *availability_365*, *minimum_nights*, *maximum_nights*, *number_of_reviews_ltm*, *reviews_per_month*, *number_of_reviews_130d*, *reviews_scores_rating*, *review_scores_accuracy*, *review_scores_communication*, *review_scores_location*, *review_scores_value*, *number_of_reviews* and *review_scores_cleanliness* do not seem to have a significant relation with price.

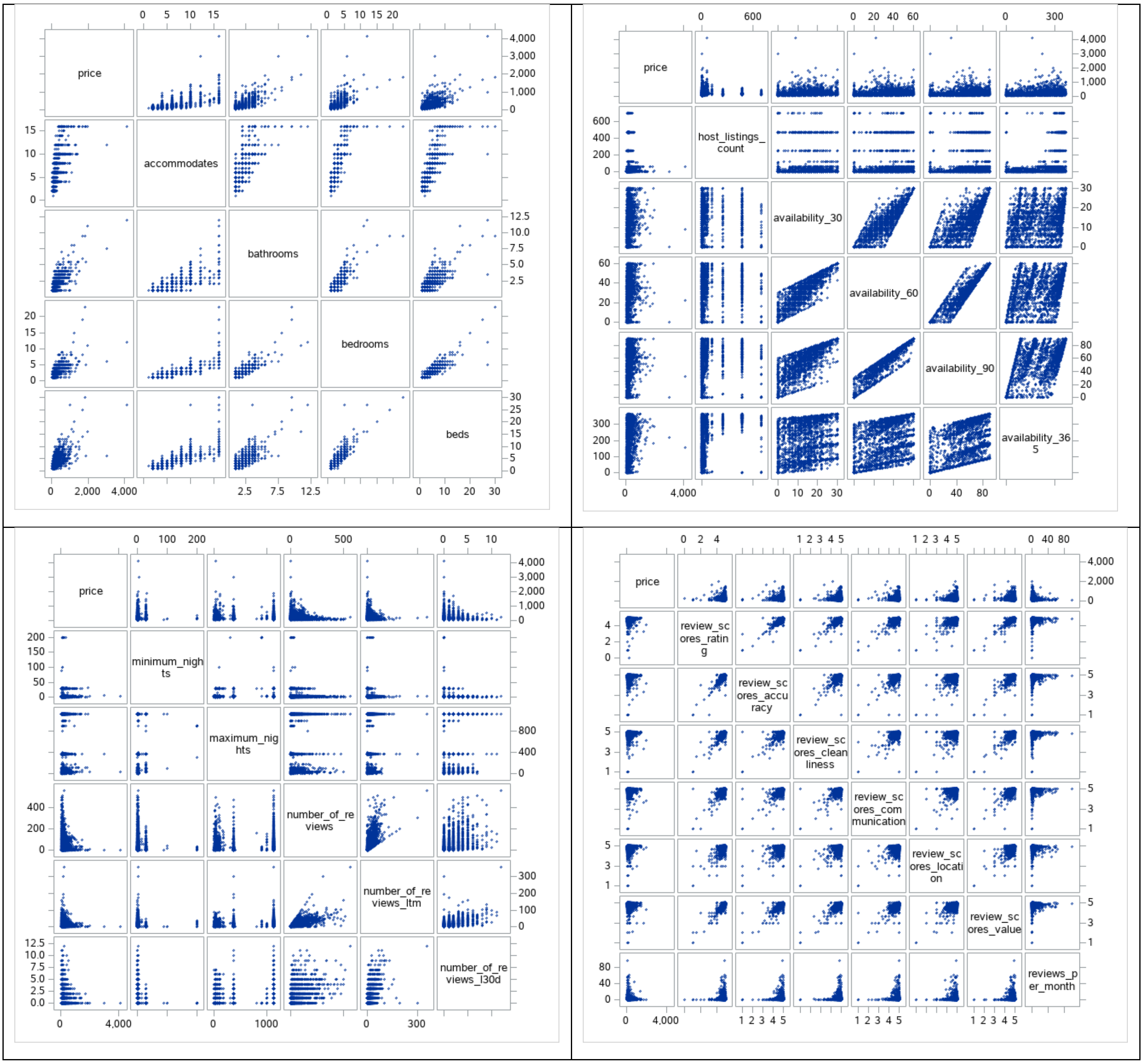


Figure 10: Scatter Plot matrix for numeric variables (Code in Appendix Figure 7)

3.1.2 Model selection

Model selection techniques is then deployed to select the most suitable variables for our model linear regression before constructing the model. The model selection procedure that are deployed are backward elimination and stepwise selection. As seen in Figure 11 and Figure 12, out of the 22 variables that are inputted into the linear regression model, only 14 variables are selected by the variables selection algorithm to be included into the model. The 14 variables that are suggested by both backward elimination and stepwise selection algorithm to be the most important variables to be included into the model to best fit the observed data are *host_listings_count, accommodates, bathrooms, bedrooms, beds, availability_30, availability_60, availability_365, number_of_reviews_130d, reviews_scores_rating, review_scores_accuracy, review_scores_communication, review_scores_location* and *review_scores_value.* The variables that are suggested to be removed from the linear regression model are *number_of_reviews_ltm, minimum_nights, host_response_time_num, reviews_per_month, availability_90, number_of_reviews, maximum_nights* and *review_scores_cleanliness.*

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-51.50426	55.26512	12418	0.87	0.3515
host_listings_count	-0.16789	0.02060	949676	66.42	<.0001
accommodates	9.47844	2.18012	270264	18.90	<.0001
bathrooms	91.88626	5.51820	3964439	277.27	<.0001
bedrooms	18.99329	5.50993	169896	11.88	0.0006
beds	8.89001	2.56253	172084	12.04	0.0005
availability_30	4.37667	0.79217	436436	30.52	<.0001
availability_60	-1.81742	0.41268	277303	19.39	<.0001
availability_365	0.18512	0.03079	517002	36.16	<.0001
number_of_reviews_l30d	-6.57503	1.65542	225557	15.78	<.0001
review_scores_rating	111.83599	23.25206	330762	23.13	<.0001
review_scores_accuracy	-50.00342	18.07816	109387	7.65	0.0057
review_scores_communication	-76.04817	15.67055	336732	23.55	<.0001
review_scores_location	114.81441	12.02095	1304338	91.23	<.0001
review_scores_value	-99.08733	20.36216	338582	23.68	<.0001

Bounds on condition number: 6.8024, 755.11

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	number_of_reviews_ltm	21	0.0000	0.6339	21.2333	0.23	0.6292
2	minimum_nights	20	0.0002	0.6337	19.9896	0.76	0.3845
3	host_response_time_num	19	0.0002	0.6335	18.8772	0.89	0.3461
4	reviews_per_month	18	0.0002	0.6333	17.8161	0.94	0.3325
5	availability_90	17	0.0003	0.6330	17.2087	1.39	0.2380
6	number_of_reviews	16	0.0003	0.6327	16.8529	1.64	0.1998
7	maximum_nights	15	0.0004	0.6323	16.5999	1.75	0.1864
8	review_scores_cleanliness	14	0.0004	0.6320	16.2904	1.69	0.1938

Figure 11: Output summary of Backward Elimination Model Selection Procedure (Code in Appendix Figure 8)

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	bathrooms		1	0.5309	0.5309	473.814	2005.42	<.0001
2	bedrooms		2	0.0338	0.5647	313.938	137.70	<.0001
3	review_scores_location		3	0.0085	0.5732	275.437	35.12	<.0001
4	review_scores_value		4	0.0143	0.5875	209.059	61.31	<.0001
5	availability_30		5	0.0076	0.5951	174.843	33.06	<.0001
6	host_listings_count		6	0.0070	0.6020	143.495	30.96	<.0001
7	accommodates		7	0.0069	0.6089	112.564	31.09	<.0001
8	availability_365		8	0.0043	0.6132	93.9246	19.69	<.0001
9	availability_60		9	0.0044	0.6177	74.8171	20.36	<.0001
10	number_of_reviews_l30d		10	0.0038	0.6214	58.6849	17.65	<.0001
11	review_scores_communication		11	0.0030	0.6244	46.3400	14.07	0.0002
12	review_scores_rating		12	0.0035	0.6279	31.6267	16.54	<.0001
13	beds		13	0.0024	0.6304	21.9465	11.63	0.0007
14	review_scores_accuracy		14	0.0016	0.6320	16.2904	7.65	0.0057

Figure 12: Output summary of Stepwise Selection Model Selection Procedure (Code in Appendix Figure 9)

3.1.3 Linear Regression Analysis

Linear Regression Analysis (Dependent Variable:Price)

Model: MODEL1
Dependent Variable: price

Number of Observations Read	2095
Number of Observations Used	1774
Number of Observations with Missing Values	321

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	43189789	3084985	215.76	<.0001
Error	1759	25150169	14298		
Corrected Total	1773	68339958			

Root MSE	119.57421	R-Square	0.6320
Dependent Mean	250.59808	Adj R-Sq	0.6291
Coeff Var	47.71553		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	-51.50426	55.26512	-0.93	0.3515	0	-159.89648	56.88797
host_listings_count	1	-0.16789	0.02060	-8.15	<.0001	1.34412	-0.20830	-0.12749
accommodates	1	9.47844	2.18012	4.35	<.0001	4.85400	5.20254	13.75433
bathrooms	1	91.88626	5.51820	16.65	<.0001	2.96618	81.06335	102.70918
bedrooms	1	18.99329	5.50993	3.45	0.0006	6.17322	8.18659	29.80000
beds	1	8.89001	2.56253	3.47	0.0005	3.99851	3.86409	13.91594
availability_30	1	4.37667	0.79217	5.52	<.0001	5.66633	2.82296	5.93037
availability_60	1	-1.81742	0.41268	-4.40	<.0001	6.27314	-2.62682	-1.00802
availability_365	1	0.18512	0.03079	6.01	<.0001	1.58295	0.12474	0.24550
number_of_reviews_l30d	1	-6.57503	1.65542	-3.97	<.0001	1.10941	-9.82183	-3.32824
review_scores_rating	1	111.83599	23.25206	4.81	<.0001	6.80239	66.23140	157.44058
review_scores_accuracy	1	-50.00342	18.07816	-2.77	0.0057	3.90253	-85.46037	-14.54648
review_scores_communication	1	-76.04817	15.67055	-4.85	<.0001	2.72027	-106.78303	-45.31331
review_scores_location	1	114.81441	12.02095	9.55	<.0001	1.86870	91.23755	138.39126
review_scores_value	1	-99.08733	20.36216	-4.87	<.0001	4.67455	-139.02392	-59.15075

Figure 13: Output results of Linear Regression Model (Code in Appendix Figure 10)

The output result of the regression model in Figure 13 is interpreted and analyzed. It is observed that our model has an R-Square value 0.6320. Therefore, 63.2% of the variation in property rental price is explained by the variation in *host_listings_count, accommodates, bathrooms, bedrooms, beds, availability_30, availability_60, availability_365, number_of_reviews_130d, reviews_scores_rating, review_scores_accuracy, review_scores_communication, review_scores_location* and *review_scores_value.* The Adjusted R-Square value is 0.6291. Therefore, 62.91% of the variation in property rental price is explained by the regression model adjusted for the number of independent variables and sample size. The coefficient of variation is 47.71, which is considered not bad, this suggests a moderately good model fit. Furthermore, the variance inflation factors (VIF) value suggest that there is no collinearity problem for the model since none of the VIF values for the variables are larger than 10.

The sample regression equation for the model is

$$\begin{aligned}\hat{y} = {} & -51.5043 - 0.1679x_1 + 9.4784x_2 + 91.8863x_3 + 18.9933x_4 + 8.89x_5 \\ & + 4.3767x_6 - 1.8174x_7 + 0.1851x_8 - 6.575x_9 + 111.836x_{10} \\ & - 50.0034x_{11} - 76.0482x_{12} + 114.8144x_{13} - 99.0873x_{14}\end{aligned}$$

Inference on Collective Influence

H_0: There is no linear relationship between the response variable and the explanatory variables.

H_1: There is a linear relationship between the response variable and at least one of the explanatory variables.

To determine the collective influence of the explanatory variables in this dataset, it is required to perform an overall F-test for the hypothesis testing procedure. Based on Figure 13, the F-value is 215.76 and the corresponding p-value is <0.0001, therefore the null hypothesis is rejected at the 0.05 level of significance (α = 0.05). There is sufficient evidence to conclude that at least one of the explanatory variables has a significant effect on the response variable. Next, the test for the significance of the individual regression coefficients is conducted to determine which explanatory variables have a significant effect on the response variable.

Inference for Individual Regression Coefficients & Confidence Interval Estimate for the Slope

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

where β_1 is the partial regression coefficient for X_1 (***host_listings_count***). The test statistic t-value for *host_listings_count* is -8.15 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *host_listings_count* has a significant relationship with *price*, controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_1 is (-0.2083, -0.1275). We are 95% confident that for every unit increase in *host_listings_count*, the predicted property rental daily price is estimated to decrease between $0.1275 to $0.2083.

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

where β_2 is the partial regression coefficient for X_2 (***accommodates***). The test statistic t-value for *accommodates* is 4.35 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *accommodates* has a significant relationship with *price,* controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_2 is (5.2025, 13.7543). We are 95% confident that for every unit increase in *accommodates,* the predicted property rental daily price is estimated to increase between $5.2025 to $13.7543.

$$H_0: \beta_3 = 0$$
$$H_1: \beta_3 \neq 0$$

where β_3 is the partial regression coefficient for X_3 (***bathrooms***). The test statistic t-value for *bathrooms* is 16.65 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *accommodates* has a significant relationship with *price,* controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_3 is (81.0634, 102.7092). We are 95% confident that for every unit increase in *bathrooms,* the predicted property rental daily price is estimated to increase between $81.0634to $102.7092.

$$H_0: \beta_4 = 0$$
$$H_1: \beta_4 \neq 0$$

where β_4 is the partial regression coefficient for X_4 (***bedrooms***). The test statistic t-value for *bedrooms* is 3.45 with corresponding p-value 0.0006, which is larger than 0.0001, H_0 is not rejected at significance level α = 0.05. There is insufficient evidence to conclude that *bedrooms* have a significant relationship with *price*, controlling for the other variables.

$$H_0: \beta_5 = 0$$
$$H_1: \beta_5 \neq 0$$

where β_5 is the partial regression coefficient for X_5 (***beds***). The test statistic t-value for *beds* is 3.47 with corresponding p-value 0.0006, which is larger than 0.0001, H_0 is not rejected at significance level α = 0.05. There is insufficient evidence to conclude that *beds* have a significant relationship with *price*, controlling for the other variables.

$$H_0: \beta_6 = 0$$
$$H_1: \beta_6 \neq 0$$

where β_6 is the partial regression coefficient for X_6 (***availability_30***). The test statistic t-value for *availability_30* is 5.52 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *availability_30* has a significant relationship with *price*, controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_6 is (2.823, 5.9304). We are 95% confident that for every unit increase in *availability_30*, the predicted property rental daily price is estimated to increase between $2.823 to $5.9304.

$$H_0: \beta_7 = 0$$
$$H_1: \beta_7 \neq 0$$

where β_7 is the partial regression coefficient for X_7 (***availability_60***). The test statistic t-value for *availability_60* is -4.4 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *availability_60* has a significant relationship with *price*, controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_7 is (-2.6268, -1.008). We are 95% confident that for every unit increase in *availability_60*, the predicted property rental daily price is estimated to decrease between $1.008 to $2.6268.

$$H_0: \beta_8 = 0$$
$$H_1: \beta_8 \neq 0$$

where β_8 is the partial regression coefficient for X_8 (***availability_365***). The test statistic t-value for *availability_365* is 6.01 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *availability_365* has a significant relationship with *price*, controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_8 is (0.1247, 0.2455). We are 95% confident that for every unit increase in *availability_365,* the predicted property rental daily price is estimated to increase between $0.1247 to $0.2455.

$$H_0: \beta_9 = 0$$
$$H_1: \beta_9 \neq 0$$

where β_9 is the partial regression coefficient for X_9 (***number_of_reviews_l30d***). The test statistic t-value for *number_of_reviews_l30d* is -3.97 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *number_of_reviews_l30d* has a significant relationship with *price*, controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_9 is (-9.8218, -3.3282). We are 95% confident that for every unit increase in *number_of_reviews_l30d*, the predicted property rental daily price is estimated to decrease between $3.3282 to $9.8218.

$$H_0: \beta_{10} = 0$$
$$H_1: \beta_{10} \neq 0$$

where β_{10} is the partial regression coefficient for X_{10} (***review_scores_rating***). The test statistic t-value for *review_scores_rating* is 4.81 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *review_scores_rating* has a significant relationship with *price*, controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_{10} is (66.2314, 157.4406). We are 95% confident that for every unit increase in *review_scores_rating,* the predicted property rental daily price is estimated to increase between $66.2314 to $157.4406.

$$H_0: \beta_{11} = 0$$
$$H_1: \beta_{11} \neq 0$$

where β_{11} is the partial regression coefficient for X_{11} (***review_scores_accuracy***). The test statistic t-value for *review_scores_accuracy* is -2.77 with corresponding p-value 0.0057, which is larger than 0.0001, H_0 is not rejected at significance level α = 0.05. There is insufficient evidence to conclude that *review_scores_accuracy* has a significant relationship with *price*, controlling for the other variables.

$$H_0: \beta_{12} = 0$$
$$H_1: \beta_{12} \neq 0$$

where β_{12} is the partial regression coefficient for X_{12} (***review_scores_communication***). The test statistic t-value for *review_scores_communication* is -4.85 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *review_scores_communication* has a significant relationship with *price*, controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_{12} is (-106.783, -45.3133). We are 95% confident that for every unit increase in *review_scores_communication*, the predicted property rental daily price is estimated to decrease between $45.3133 to $106.783.

$$H_0: \beta_{13} = 0$$
$$H_1: \beta_{13} \neq 0$$

where β_{13} is the partial regression coefficient for X_{10} (***review_scores_location***). The test statistic t-value for *review_scores_location* is 9.55 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *review_scores_location* has a significant relationship with *price*, controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_{13} is (91.2376, 138.3913). We are 95% confident that for every unit increase in *review_scores_location*, the predicted property rental daily price is estimated to increase between $91.2376 to $138.3913.

$$H_0: \beta_{14} = 0$$
$$H_1: \beta_{14} \neq 0$$

where β_{14} is the partial regression coefficient for X_{14} (***review_scores_value***). The test statistic t-value for *review_scores_value* is -4.87 with corresponding p-value < 0.0001, H_0 is rejected at significance level α = 0.05. There is sufficient evidence to conclude that *review_scores_value* has a significant relationship with *price*, controlling for the other variables. Controlling for other explanatory variables in the model, the 95% confidence interval for β_{14} is (-139.0239, -59.1508). We are 95% confident that for every unit increase in *review_scores_value*, the predicted property rental daily price is estimated to decrease between $59.1508 to $139.0239.

3.3.4 Regression Diagnostic

To verify that our F-test and t-test in hypothesis testing for our linear regression model are reliable, it is necessary to deploy regression diagnostics to ensure that the standard regression assumptions are satisfied. Regression diagnostics plots such as the Normal Quantile-Quantile (Q-Q) Plot, Studentized Deleted Residuals (RStudent) plot, Cook's Distance (Cook's D) plot, Difference in Fit (DFFit) plot and Difference in Beta (DFBeta) plot is generated to check for the normality of the residuals as well as to identify high leverage points and outliers that are potential influential data.

Based on the residuals against the normal quantiles (Q-Q) plot in Figure 14, it is observed that there is no serious violation of the normality assumption although there is a slight deviation at the tails of the data. Based on the kernel density plot in Figure 14, it is observed that the density curve is slightly skewed to the right, but it is not significant to the extent of violating the normality assumption. This conclusion is not contradicted by the quantile-quantile plot.

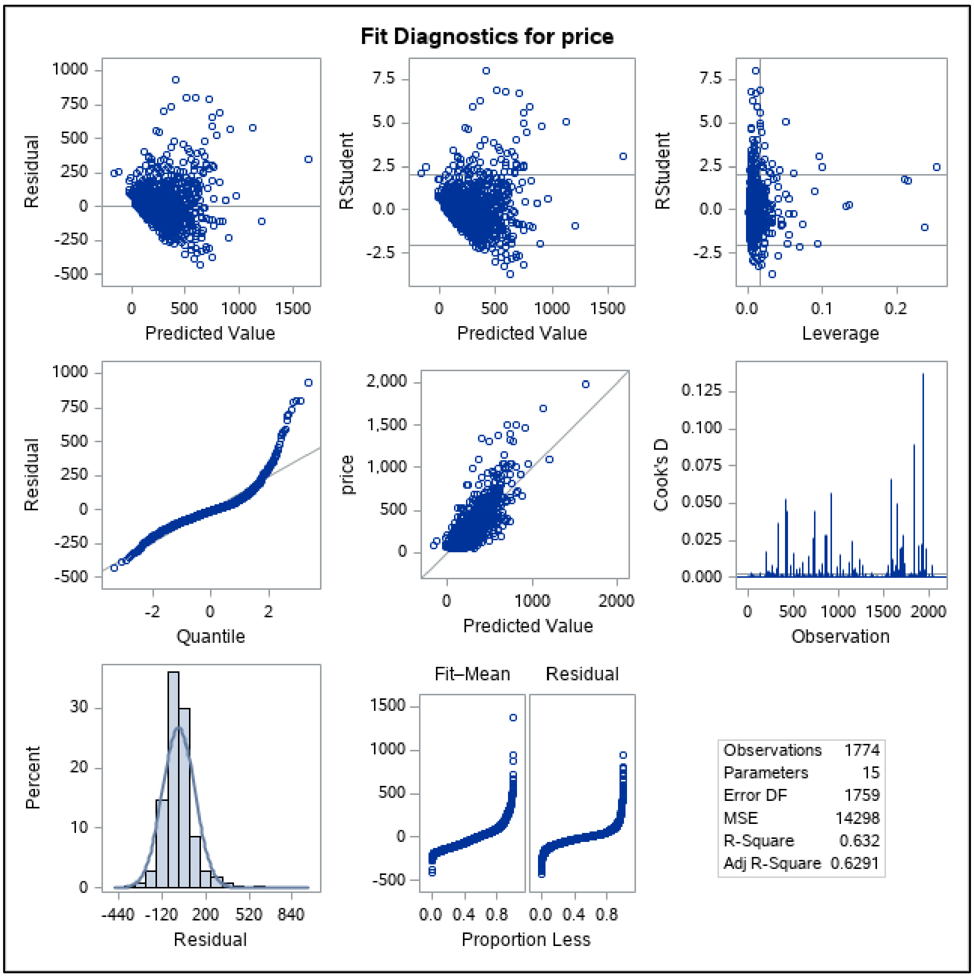
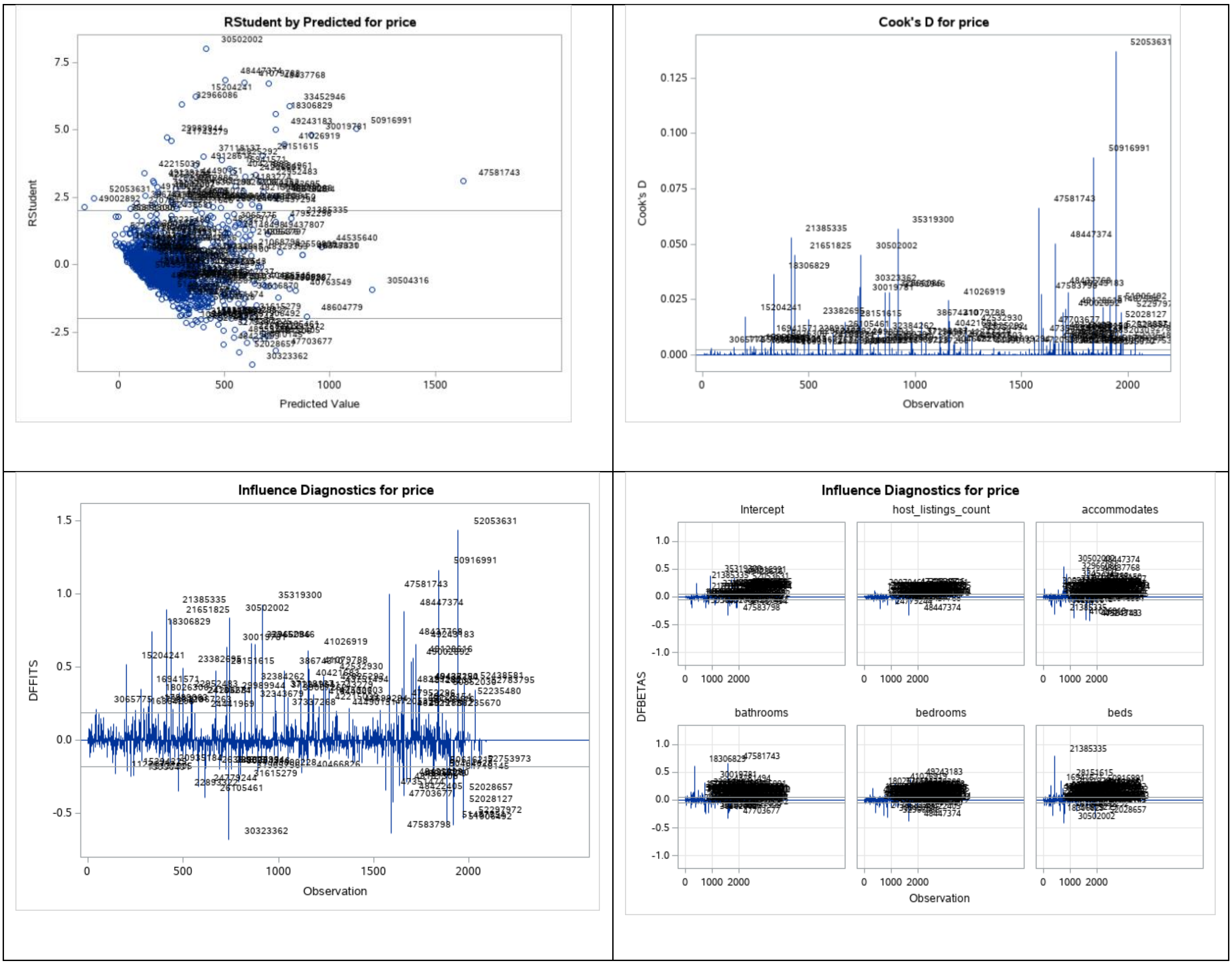


Figure 14: Fit Diagnostic for price (Code in Appendix Figure 10)

To get a closer look of the RStudent Plot and Cook's D plot in Figure 14, a larger version of the plot is generated in Figure 15. In addition to the RStudent Plot and Cook's D plot, the DFFit Plot and DFBeta Plot are also generated to identify high leverage points and outliers that are potential influential data. In Figure 15, the RStudent plot shows a significant number of observations beyond two standard errors from the mean of 0. The Cook's D plot and DFFit plot shows that there are several potential influential observations in the dataset, particularly observations #52053631, #50916991 and #47581743. To see which parameters these influential points might influence the most, the DFBeta plot is examined. Based on the DFBeta plot, observation #52053631is influential because of its effects on *review_scores_communication, review_scores_accuracy and review_scores_rating;* #50916991 is influential because of its effects on *review_scores_location;* observation #47581743 is influential because of its effects on *bathrooms.* These observations were analysed to ensure that they are not faulty data. After inspection of the suspicious influential points, no faulty data was found; therefore, no observations were removed.



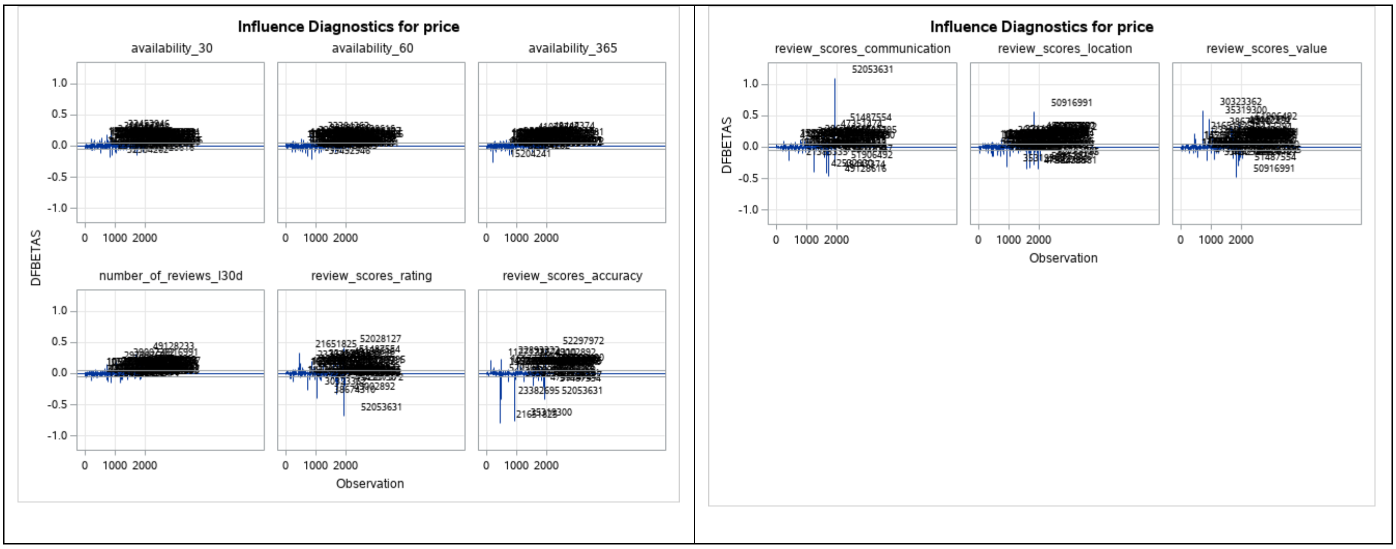


Figure 15: RStudent Plot, Cook's D plot, DFFit Plot, DFBeta Plot for price (Code in Appendix Figure 11)

3.2 Logistic Regression: Explanatory Analysis on *host_is_superhost*

The second objective of our study is to estimate the relationship between *host_is_superhost* and other variables related to the host details and review scores predictors. As such, binary logistic regression analysis is performed with the variable *host_is_superhost* as the response variable and the variables *host_since*, *host_response_time_num*, *host_listings_count*, *host_has_profile_pic*, *host_identity_verifi and review_scores_value* as the predictor variables.

3.2.1 Bivariate Analysis

Prior to moving on to the fully specified model, bivariate summaries of the *host_is_superhost* variable and the individual predictors are examined to understand the associations between them. Figure 16 shows a bar chart which compares *host_is_superhost* and *host_response_time_num*. It is observed that the value count true (t) is slightly higher then value count false (f) for variable *host_is_superhost* grouped by host_response_time_num. In Figure 17, the bar chart of *host_is_superhost* versus *host_has_profile_pic* shows that majority of the hosts has a profile picture and all host who is a superhost has a profile picture. Based on the bar chart of *host_is_superhost* versus *host_identity_verified* in Figure 18, it is observed that the value count false (f) is slightly higher then value count true (t) for variable *host_is_superhost* grouped by *host_identity_verifi*. *Figure 19 illustrates a bar chart of host_is_superhost* versus *host_listing_count*. It is observed that the majority of the hosts who are a superhost have relatively less property listing count whereas the majority of the hosts who are not a superhost host have relatively more property listing count. Figure 20 shows a histogram of *host_is_superhost* versus *host_since*. It is observed that the distribution of superhost-host count seems to peak higher than non-superhost-host when *host_since* is before 2017 whereas the count distribution of non-superhost-host seems to peak higher than superhost-host when *host_since* is after 2017. This suggest that a host is more likely to be a superhost when *host_since* is before 2017 and a host is more likely to not be a superhost when *host_since* is after 2017. This may also suggest that the earlier a host starts hosting, the larger the possibility that a host is a superhost.

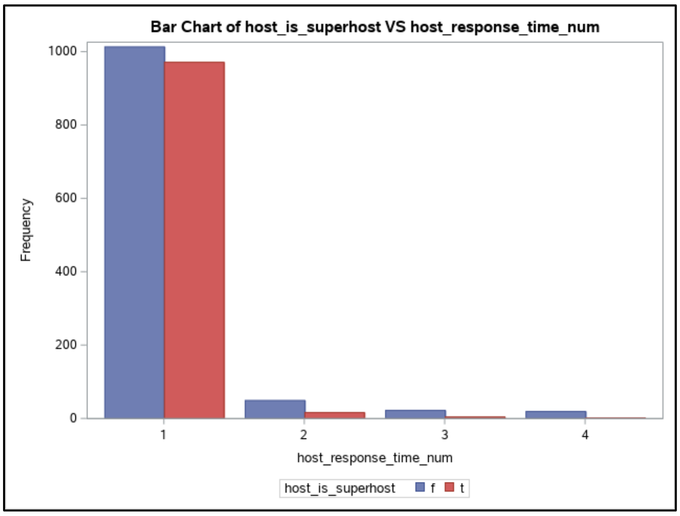


Figure 16: Bar Chart of host_is_superhost VS host_response_time_num (Code in Appendix Figure 12)

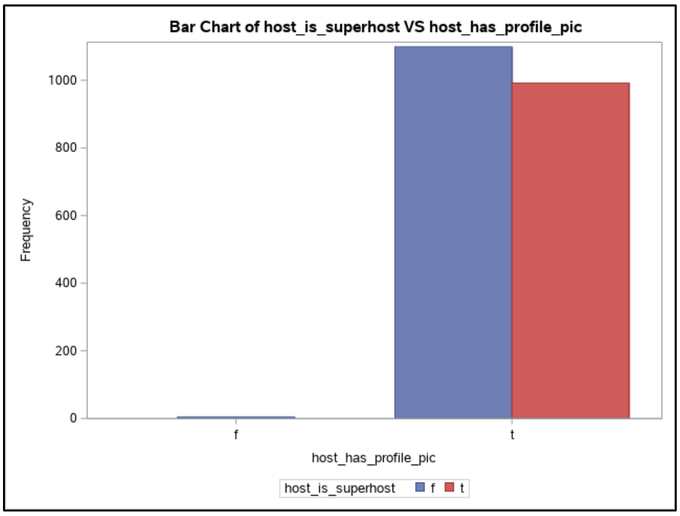


Figure 17: Bar Chart of host_is_superhost VS host_has_profile_pic (Code in Appendix Figure 12)

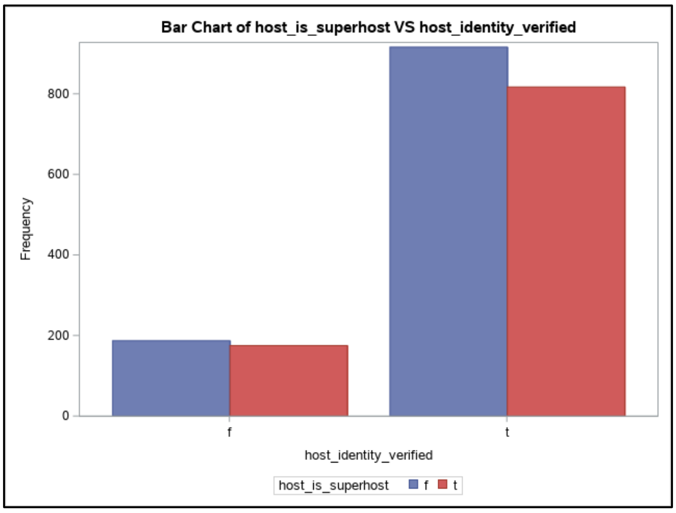


Figure 18: Bar Chart of host_is_superhost VS host_identity_verified (Code in Appendix Figure 12)

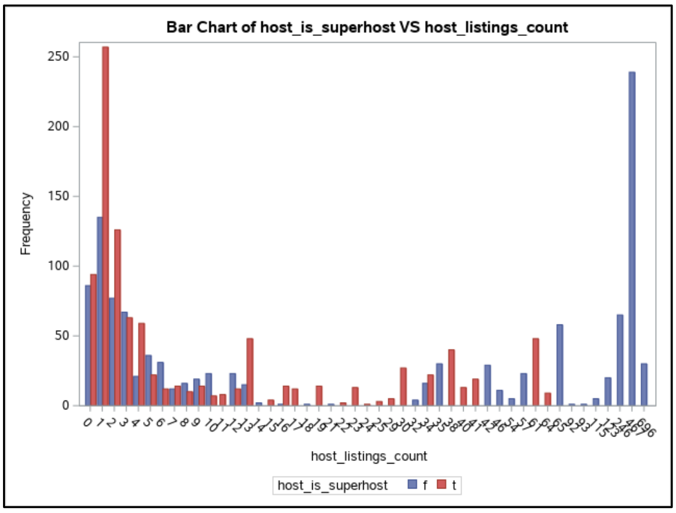


Figure 19: Bar Chart of host_is_superhost VS host_listing_count (Code in Appendix Figure 12)

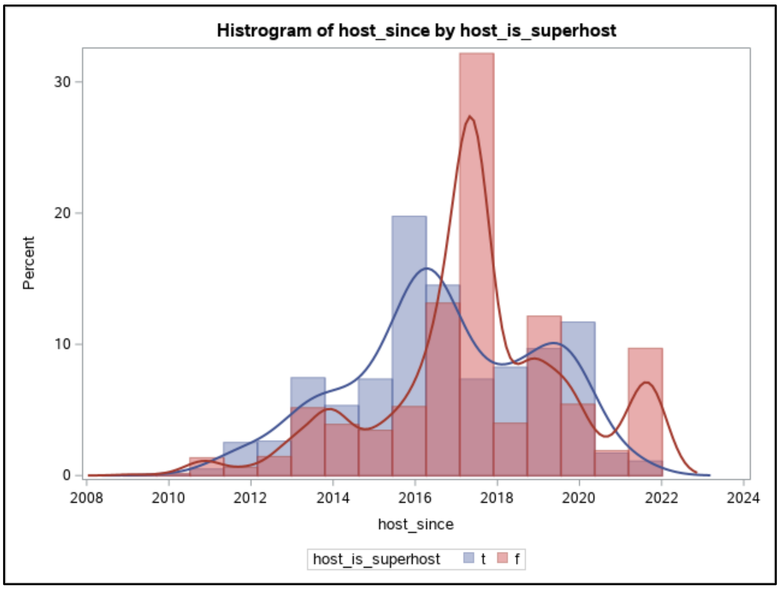


Figure 20: Histogram of host_is_superhost VS host_since (Code in Appendix Figure 13)

3.2.2 Logistic Regression Analysis

Figure 21 provides information of the model, data set, the response variable, the number of response levels, the type of model, the algorithm used to obtain the parameter estimates, and the number of observations read and used in this model. Variable *host_is_superhost* has two response level, which are either true (t) or false (f), therefore the model is assumed to be "binary logit".

Logistic Regression Analysis (Dependent Variable: host_is_superhost)

The LOGISTIC Procedure

Model Information	
Data Set	STATDATA.PROPERTYCLEANED
Response Variable	host_is_superhost
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	2095
Number of Observations Used	1879

Response Profile		
Ordered Value	host_is_superhost	Total Frequency
1	f	930
2	t	949

Probability modeled is host_is_superhost='t'.

Note: 216 observations were deleted due to missing values for the response or explanatory variables.

Figure 21: Model information & Response Profile of logistic regression model (Code in Appendix Figure 14)

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	2606.655	1954.171
SC	2612.193	1992.940
-2 Log L	2604.655	1940.171

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	664.4845	6	<.0001
Score	459.6076	6	<.0001
Wald	228.3064	6	<.0001

Figure 22: Model Fit Statistics & Testing Global Null Hypothesis table of logistic regression model (Code in Appendix Figure 14)

The Model Fit Statistics table in Figure 22 provides three goodness-of-fit measures, namely Akaike's Information Criterion (AIC) test, Schwarz criterion (SC) test and the -2LogL test. By comparing these test values for the "Intercept Only" column and the "Intercept and Covariates" column, we can observe that the "Intercept and Covariates" column has a smaller value, this imply that this logistic regression model is a good model to fit the data set.

<u>Inference on Collective Influence</u>

$H_0: All\ the\ regression\ coefficients\ are\ 0$

$H_1: At\ least\ one\ of\ the\ regression\ coefficient\ is\ not\ 0$

Based on the output results of the Testing Global Null Hypothesis Table in Figure 22, H_0 is rejected since the p-values for all three tests, namely the Likelihood ratio test, Score test and Wald test are <0.0001. At the 0.05 significance level, collectively the predictor variables are significant, indicating at least one of the predictors in the model is useful in predicting whether a host is a superhost.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
host_since	1	10.8199	0.0010
host_response_time_n	1	39.5837	<.0001
host_listings_count	1	59.6846	<.0001
host_has_profile_pic	1	0.0004	0.9832
host_identity_verifi	1	5.6249	0.0177
review_scores_value	1	124.6312	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-10.5013	1.9257	29.7366	<.0001
host_since		1	-0.00021	0.000065	10.8199	0.0010
host_response_time_n		1	-1.2599	0.2003	39.5837	<.0001
host_listings_count		1	-0.0142	0.00184	59.6846	<.0001
host_has_profile_pic	f	1	-10.7784	511.8	0.0004	0.9832
host_identity_verifi	f	1	-0.3387	0.1428	5.6249	0.0177
review_scores_value		1	3.5449	0.3175	124.6312	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
host_since	1.000	1.000	1.000
host_response_time_n	0.284	0.192	0.420
host_listings_count	0.986	0.982	0.989
host_has_profile_pic f vs t	<0.001	<0.001	>999.999
host_identity_verifi f vs t	0.713	0.539	0.943
review_scores_value	34.636	18.589	64.538

Figure 23: Type 3 Analysis of Effects, Analysis of Maximum Likelihood Estimates and Odds Ratio Estimates table of logistic regression model (Code in Appendix Figure 14)

From the Analysis of Maximum Likelihood Estimates table in Figure 23, we obtain the parameter estimates of $\beta_0 = -10.5013$, $\beta_1 = -0.00021$, $\beta_2 = -1.2599$, $\beta_3 = -0.0142$, $\beta_4 = -10.7784$, $\beta_5 = -0.3387$ and $\beta_6 = 3.5449$. Given that reference cell coding was used in this analysis, each effect is measured against the reference level.

Logistic Regression Model

$$logit(p) = \beta_0 + \beta_1 X_{host\_since} + \beta_2 X_{host\_response\_time\_num} + \beta_3 X_{host\_listing\_count} + \beta_4 X_{host\_has\_profile\_pic} + \beta_5 X_{host\_identity\_verifi} + \beta_6 X_{review\_scores\_values}$$

Sample Logistic Regression Equation

$$\text{logit}(\hat{p}) = -10.5013 - 0.00021 * X_{host\_since} - 1.2599 * X_{host\_response\_time\_num} - 0.0142 * X_{host\_listing\_count} - 10.7784 * X_{host\_has\_profile\_pic} - 0.3387 * X_{host\_identity\_verifi} + 3.5449 * X_{\text{review\_scores\_values}}$$

Inference for Individual Regression Coefficients

Based on the Type 3 Analysis of Effect Table in Figure 23, let

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

where β_1 is the partial regression coefficient for X_{host_since}. The test statistic Wald Chi-Square for *host_since* is 10.8199 with corresponding p-value is 0.0010, which is > 0.0001, null hypothesis is not rejected at significance level α = 0.05. *host_since* is not significant in predicting whether a host is a superhost, controlling for the other variables.

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

where β_2 is the partial regression coefficient for $X_{host\_response\_time\_num}$.The test statistic Wald Chi-Square for *host_response_time_num* is 39.5837 with corresponding p-value < 0.0001, null hypothesis is rejected at significance level α = 0.05. *host_response_time_num* is significant in predicting whether a host is a superhost, controlling for the other variables.

$$H_0: \beta_3 = 0$$
$$H_1: \beta_3 \neq 0$$

where β_3 is the partial regression coefficient for $X_{host\_listing\_count}$.The test statistic Wald Chi-Square for *host_listing_count* is 59.6846 with corresponding p-value < 0.0001, null hypothesis is rejected at significance level α = 0.05. *host_listing_count* is significant in predicting whether a host is a superhost, controlling for the other variables.

$$H_0: \beta_4 = 0$$
$$H_1: \beta_4 \neq 0$$

where β_4 is the partial regression coefficient for $X_{host\_has\_profile\_pic}$. The test statistic Wald Chi-Square for *host_has_profile_pic* is 0.0004 with corresponding p-value is 0.9832, which is > 0.0001, null hypothesis is not rejected at significance level α = 0.05. *host_has_profile_pic* is not significant in predicting whether a host is a superhost, controlling for the other variables.

$$H_0: \beta_5 = 0$$
$$H_1: \beta_5 \neq 0$$

where β_5 is the partial regression coefficient for $X_{host\_identity\_verifi}$. The test statistic Wald Chi-Square for *host_identity_verifi* is 5.6249 with corresponding p-value is 0.0177, which is >

0.0001, null hypothesis is not rejected at significance level α = 0.05. *host_identity_verifi* is not significant in predicting whether a host is a superhost, controlling for the other variables.

$$H_0: \beta_6 = 0$$
$$H_1: \beta_6 \neq 0$$

where β_6 is the partial regression coefficient for $X_{review\_scores\_values}$. The test statistic Wald Chi-Square for *review_scores_values* is 124.6312 with corresponding p-value < 0.0001, null hypothesis is rejected at significance level α = 0.05. *review_scores_values* is significant in predicting whether a host is a superhost, controlling for the other variables.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	80.9	Somers' D	0.617
Percent Discordant	19.1	Gamma	0.617
Percent Tied	0.0	Tau-a	0.309
Pairs	882570	c	0.809

Odds Ratios		
Effect	Unit	Estimate
host_listings_count	10.0000	0.868

Figure 24: Association of Predicted Probabilities and Observed Responses and Odds Ratios table of logistic regression model (Code in Appendix Figure 14)

Based on the Association of Predicted Probabilities and Observed Responses Table in Figure 24, the c (concordance) statistics has a value of 0.809, indicating that 80.9% of the positive and negative response pairs (*host_is_superhost*) are correctly sorted using *host_since*, *host_response_time_num*, *host_listing_count*, *host_has_profile_pic*, *host_identity_verifi* and *review_scores_values*. This shows a strong ability for *host_since*, *host_response_time_num*, *host_listing_count*, *host_has_profile_pic*, *host_identity_verifi* or *review_scores_values* to discriminate between whether a host is a superhost.

The Odds Ratios table in Figure 24 shows that a number of 10 increase in *host_listing_count* is associated with a (1-0.868)% = 13.2% decrease in the odds of a host being a superhost. This suggest that the larger the *host_listing_count*, the less likely a host is to be a superhost.

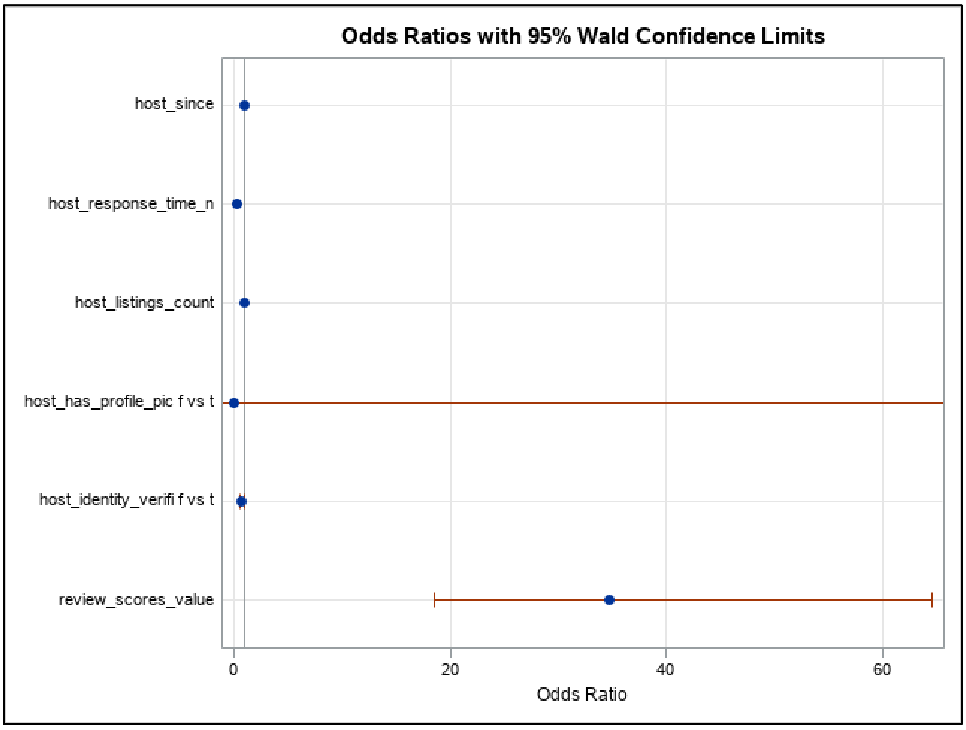


Figure 25: Odds Ratio plot with 95% Wald Confidence Limits of logistic regression model (Code in Appendix Figure 14)

Figure 25 shows the odds ratio plot for the Walk confidence limit of our mode. Based on the Odds Ratio Estimates table in Figure 24, for 95% confidence interval, we are confident that the true odds ratio of *host_since* falls between 1.000 and 1.000; the true odds ratio of *host_response_time_num* falls between 0.192 and 0.420; the true odds ratio of *host_listings_count* falls between 0.982 and 0.989; the true odds ratio of *host_has_profile_pic* falls between <0.001 and >999.999; the true odds ratio of *host_identity_verifi* falls between 0.539 and 0.943; the true odds ratio of *review_scores_value* falls between 18.589 and 64.538. In Figure 25, it is observed that the estimates of *host_response_time_num*, *host_listings_count and host_identity_verifi* are less then 1 whereas the estimates of *review_scores_value* is greater than 1. Both estimates of *host_since* and *host_has_profile_pic* intersect the reference line at odds ratio = 1, which indicates ratios that are not significantly different from 1, the effect of these two variables are not significant at the 0.05 significance level.

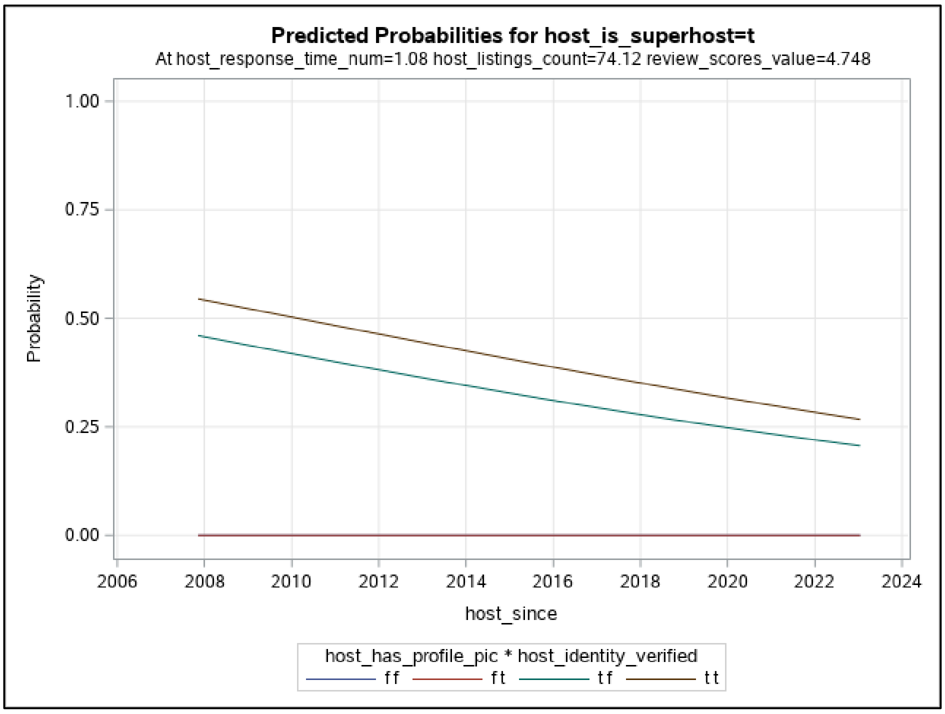


Figure 26: Effects Plot of logistic regression model (Code in Appendix Figure 14)

The effects plot in Figure 26 shows the probability of whether a host is a superhost across all combinations of categories and levels of all three predictor variables. It is observed that the probability of *host_is_superhost* is true decreases with the increase in the year for *host_since*, therefore, this suggest that the earlier a host starts hosting, the larger the probability that a host is a superhost. Furthermore, this plot suggest that a host who has a profile pic and has identity verified have the highest probability to be a superhost. Following that, the condition for a host to have the second largest probability to be a superhost is to have a profile pic and host identified not verified. The condition of a host not having a profile pic but have identified verified and the condition of a host who neither has a profile pic nor have their identity verified has little to no probability of being a superhost.

3.3 ANOVA: Compare the means of *review_scores_communication* with different *host_response_time_num*

Our third objective of this study is to test whether the ratings score for ease of communication (*review_scores_communication*) is affected by the host's response time (*host_response_time_num*). To reach this objectve, analysis of variance (ANOVA) will be conducted to test the relationship between the categorical variable (*host_response_time_num*) and numeric variable (*review_scores_communication*) by testing the difference between the population means of *review_scores_communication* grouped by *host_response_time_num*.

3.3.1 Descriptive Statistics Across Groups with Box and Whiskers Plot

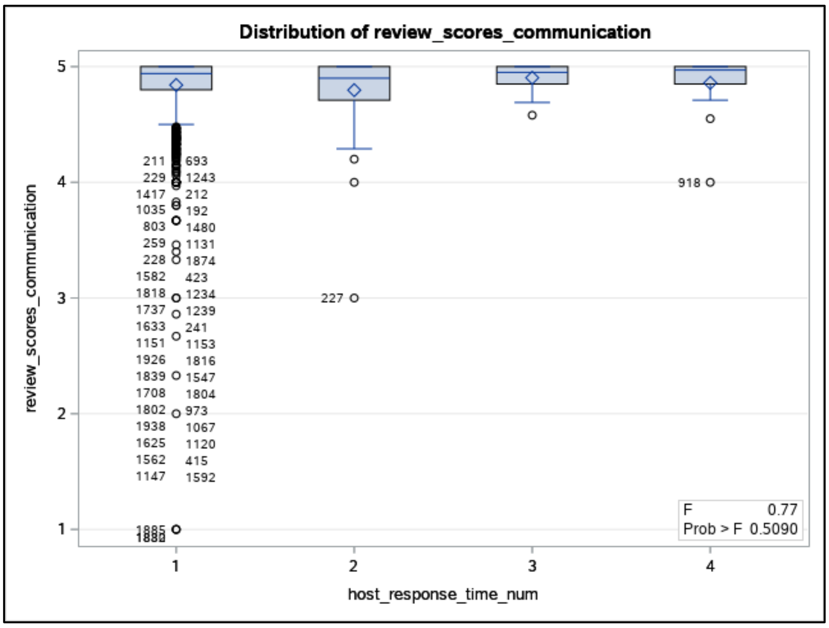


Figure 27: Box and Whiskers Plot of review_scores_communication grouped by host_response_time_num (Code in Appendix Figure 15)

Figure 27 shows the box and whiskers plot of the *review_scores_communication* grouped by *host_response_time_num*. By observing the plot, there is no significant difference between the boxes, all boxes are situated near the value 5 of *review_scores_communication.* It is suggested that the four host_response_time_num value may result in the same mean of the *review_scores_communication*. However, it is also observed that the values of *review_scores_communication* with the host_response_time_num = 1 are more scattered, ranging from the value 1 to 5 of *review_scores_communication*.

3.3.2 Analysis of Variance (ANOVA)

Analysis of Variance (Review_score_communication VS Host_response_time type)

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
host_response_time_num	4	1 2 3 4

Number of Observations Read	2095
Number of Observations Used	1879

Analysis of Variance (Review_score_communication VS Host_response_time type)

The GLM Procedure

Dependent Variable: review_scores_communication

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.1995898	0.0665299	0.77	0.5090
Error	1875	161.3649653	0.0860613		
Corrected Total	1878	161.5645551			

R-Square	Coeff Var	Root MSE	review_scores_communication Mean
0.001235	6.060361	0.293362	4.840671

Source	DF	Type I SS	Mean Square	F Value	Pr > F
host_response_time_n	3	0.19958981	0.06652994	0.77	0.5090

Source	DF	Type III SS	Mean Square	F Value	Pr > F
host_response_time_n	3	0.19958981	0.06652994	0.77	0.5090

Figure 28: Output results of Analysis of Variance (Code in Appendix Figure 15)

Let μ_i be the population mean *review_scores_communication*

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_1: At\ least\ one\ of\ the\ \mu_i\ is\ different, i = 1, 2, 3, 4$$

Based on the analysis of variance table in Figure 28, the reported f-value is 0.77, and the corresponding p-value is 0.5090, which is greater than 0.05, therefore, we do not reject H_0 at the 0.05 level of significance (α = 0.05). There is insufficient evidence to conclude that there is statistically significant difference between the means of *review_scores_communication*. The four different *host_response_time_num* value result in the same mean *review_scores_communication*. Furthermore, it is observed that the R-Square value of our model is 0.0012, therefore, *host_response_time_num* explains about 0.12% of the variability of *review_scores_communication*. The total mean of the *review_scores_communication* is 4.8407 and the Root mean square error (RMSE) is 0.0665.

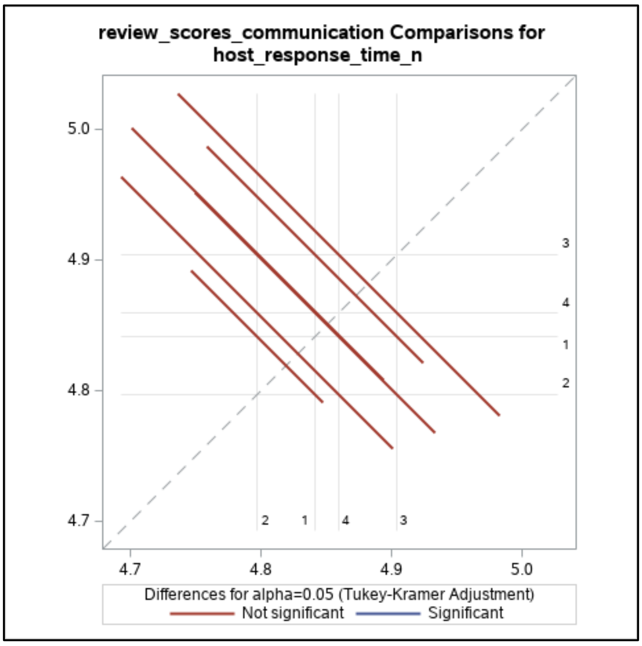


Figure 29: Diffogram plot of review_scores_communication (Code in Appendix Figure 15)

Figure 29 shows the diffogram plot of *review_scores_communication* comparison for *host_response_time_num*. It is observed that all the confidence limit for the difference cross the diagonal equivalence line, therefore, there is no significant difference between *host_response_time_num* 1 to 4.

4.0 Conclusion

In summary, the objectives of this study are to estimate the relationship between the daily price of property rentals and other variables related to property details and review scores; to estimate the relationship between *host_is_superhost* and other variables related to the host details and review scores predictors; and to test whether the ratings score for ease of communication is affected by the host's response time. For the first objective, linear regression analysis was conducted and it was found that 63.2% of the variation in property rental price is explained by the variation in *host_listings_count, accommodates, bathrooms, bedrooms, beds, availability_30, availability_60, availability_365, number_of_reviews_130d, reviews_scores_rating, review_scores_accuracy, review_scores_communication, review_scores_location* and *review_scores_value.* Controlling for the other variables, the variables that has a significant relationship with *price* are *host_listings_count, accommodates, bathrooms, availability_30, availability_60, availability_365, number_of_reviews_130d, reviews_scores_rating, review_scores_communication, review_scores_location* and *review_scores_value.* For the second objective, logistic regression analysis was conducted and it was found that 80.9% of the positive and negative response pairs (*host_is_superhost*) are correctly sorted using *host_since, host_response_time_num, host_listing_count, host_has_profile_pic, host_identity_verifi* and *review_scores_values.* Controlling for the other variables, the variables that has a significant relationship with *host_is_superhost* are *host_response_time_num*, *host_listings_count*, and *review_scores_value.* For the third objective, analysis of variance (ANOVA) is performed and it is found that there is insufficient evidence to conclude that there is statistically significant difference between the means of *review_scores_communication* of different *host_response_time_num.* Therefore, the ratings score for ease of communication is not affected by the host's response time.

5.0 Appendix

```
/*Exploring data*/
proc contents data=STATDATA.PropertyImport varnum;
Run;
```

Appendix Figure 1 : Code for output in Figure 1

```
data STATDATA.PropertyCleaned;
    set STATDATA.PropertyImport;

    *Extract numerical value from bathrooms_text char type variable;
    bathrooms = input(compress(bathrooms_text, '.', 'kd'), 8.);

run;
```

Appendix Figure 2: Code for output in Figure 2

```
/* Frequency Table for categorical variables */
title 'Frequency Table for caategorical variable';
proc freq data=STATDATA.PropertyImport nlevels;
    tables host_is_superhost  host_identity_verified
           host_response_time host_has_profile_pic
           property_type / missing nocum;
run;
title;
```

Appendix Figure 3: Code for output in Figure 3

```
data STATDATA.PropertyCleaned;
    set STATDATA.PropertyImport;

    * encode categorical variable host_response_time to numeric;
    if host_response_time = 'within an hour' then host_response_time = 1;
        else if host_response_time = 'within a few h' then host_response_time = 2;
        else if host_response_time = 'within a day' then host_response_time = 3;
        else if host_response_time = 'a few days or' then host_response_time = 4;

    host_response_time_num = input(host_response_time, 14.);
    drop host_response_time;

run;
```

Appendix Figure 4: Code for encoding host_response_time categorical variable to numeric variable

```
/* Summary Statistics for Numeric variables*/
ods noproctitle;
ods graphics / imagemap=on;
proc means data=STATDATA.PropertyCleaned maxdec=2
                chartype mean median range std min max n nmiss vardef=df;
    var host_listings_count host_response_time_num accommodates
        bathrooms bedrooms beds price minimum_nights
        maximum_nights availability_30 availability_60 availability_90
        availability_365 number_of_reviews number_of_reviews_ltm
        number_of_reviews_l30d review_scores_rating review_scores_accuracy
        review_scores_cleanliness review_scores_communication
        review_scores_location review_scores_value reviews_per_month;
title 'Summary Statistics for numeric variables';
run;
```

Appendix Figure 5: Code for output in Figure 8

```
/* Univariate Analysis on numeric variables*/
proc univariate data=statdata.PropertyCleaned plot;
    title 'Univariate Analysis of each variable';
    var host_since host_listings_count accommodates bathrooms bedrooms
        beds price minimum_nights host_response_time_num
        maximum_nights availability_30 availability_60 availability_90
        availability_365 number_of_reviews number_of_reviews_ltm
        number_of_reviews_l30d review_scores_rating review_scores_accuracy
        review_scores_cleanliness review_scores_communication
        review_scores_location review_scores_value reviews_per_month;
    id id;
run;
```

Appendix Figure 6: Code for output in Figure 9

```
/* Scatter Plot Matrix 1*/
proc sgscatter data=STATDATA.PropertyCleaned;
               matrix price accommodates bathrooms bedrooms beds;
run;

/* Scatter Plot Matrix 2*/
proc sgscatter data=STATDATA.PropertyCleaned;
               matrix price host_listings_count availability_30
                      availability_60 availability_90 availability_365;
run;

/* Scatter Plot Matrix 3*/
proc sgscatter data=STATDATA.PropertyCleaned;
               matrix price minimum_nights maximum_nights
                      number_of_reviews number_of_reviews_ltm
                      number_of_reviews_l30d;
run;

/* Scatter Plot Matrix 4*/
proc sgscatter data=STATDATA.PropertyCleaned;
               matrix price review_scores_rating review_scores_accuracy
                      review_scores_cleanliness review_scores_communication
                      review_scores_location review_scores_value reviews_per_month;
run;
```

Appendix Figure 7: Code for output in Figure 10

```
/* Model Selection with Backward Elimination*/
ods graphics on;
proc reg data=STATDATA.PropertyCleaned PLOTS(MAXPOINTS=none);
    model price = host_listings_count accommodates bathrooms bedrooms beds minimum_nights
        host_response_time_num maximum_nights availability_30 availability_60
        availability_90 availability_365 number_of_reviews number_of_reviews_ltm
        number_of_reviews_l30d review_scores_rating review_scores_accuracy
        review_scores_cleanliness review_scores_communication
        review_scores_location review_scores_value reviews_per_month/ selection=backward;
    title 'Property data: Backward elimination results';
run;
ods graphics off;
```

Appendix Figure 8: Code for output in Figure 11

```
/* Model Selection with Stepwise Selection */
ods graphics on;
proc reg data=STATDATA.PropertyCleaned PLOTS(MAXPOINTS=none);
    model price = host_listings_count accommodates bathrooms bedrooms beds minimum_nights
        host_response_time_num maximum_nights availability_30 availability_60 availability
        availability_365 number_of_reviews number_of_reviews_ltm
        number_of_reviews_l30d review_scores_rating review_scores_accuracy
        review_scores_cleanliness review_scores_communication
        review_scores_location review_scores_value reviews_per_month / selection=stepwise;
title 'Property data: Stepwise selection results';
run;
title;
ods graphics off;
```

Appendix Figure 9: Code for output in Figure 12

```
/* Multiple Linear Regression Analysis (After model selection, left 14 predictors)*/
* removed minimum_nights, maximum_nights, availability_90, number_of_reviews
  review_scores_cleanliness, number_of_reviews_ltm, reviews_per_month, host_response_time
proc reg data=STATDATA.PropertyCleaned;
    title 'Linear Regression Analysis (Dependent Variable:Price)';
        model price = host_listings_count accommodates bathrooms bedrooms beds
                      availability_30 availability_60 availability_365
                      number_of_reviews_l30d review_scores_rating
                      review_scores_accuracy review_scores_communication
                      review_scores_location review_scores_value / clb vif;
    id id;
run;
```

Appendix Figure 10: Code for output in Figure 13 and Figure 14

```
/************************* DIAGNOSTIC PLOTS  **************************************/
ods graphics on;
proc reg data = STATDATA.PropertyCleaned PLOTS(MAXPOINTS=none)
    plots(only label) = rstudentbypredicted
    plots(only label) = cooksd
    plots(only label) = dffits
    plots(only label) = dfbetas
;
model price = host_listings_count accommodates bathrooms bedrooms beds
              availability_30 availability_60 availability_365
              number_of_reviews_l30d review_scores_rating review_scores_accuracy
              review_scores_communication review_scores_location
              review_scores_value;
id id;
run;
```

Appendix Figure 11: Code for output in Figure 15

```
/* Bar Chart of host_is_superhost VS host_response_time_num */
proc sgplot data = STATDATA.PropertyCleaned;
    title 'Bar Chart of host_is_superhost VS host_response_time_num';
  vbar host_response_time_num / group = host_is_superhost groupdisplay = cluster;
run;

/* Bar Chart host_is_superhost VS host_has_profile_pic*/
proc sgplot data = STATDATA.PropertyCleaned;
    title 'Bar Chart of host_is_superhost VS host_has_profile_pic';
  vbar host_has_profile_pic / group = host_is_superhost groupdisplay = cluster;
run;

/* Bar Chart of host_is_superhost VS host_identity_verified*/
proc sgplot data = STATDATA.PropertyCleaned;
    title 'Bar Chart of host_is_superhost VS host_identity_verified';
  vbar host_identity_verified / group = host_is_superhost groupdisplay = cluster;
run;

/* Bar Chart of host_is_superhost VS host_listings_count*/
proc sgplot data = STATDATA.PropertyCleaned;
    title 'Bar Chart of host_is_superhost VS host_listings_count';
  vbar host_listings_count / group = host_is_superhost groupdisplay = cluster;
run;
```

Appendix Figure 12: Code for output in Figure 16, Figure 17, Figure 18, Figure 19

```
/* Histrogram of host_since by host_is_superhost */
proc sgplot data=STATDATA.PropertyCleaned;
title "Histrogram of host_since by host_is_superhost";
  histogram host_since / group=host_is_superhost transparency=0.5;
  density host_since / type=kernel group=host_is_superhost;
run;
```

Appendix Figure 13: Code for output in Figure 20

```
/*Multiple Logistic Regression*/
ods graphics on;
proc logistic data=STATDATA.PropertyCleaned
    plots(only)=(oddsratio effect);
    class host_has_profile_pic
          host_identity_verified / param=ref;
    model host_is_superhost (event='t') = host_since
                                          host_response_time_num
                                          host_listings_count
                                          host_has_profile_pic
                                          host_identity_verified
                                          review_scores_value
                                          ;
    units host_listings_count = 10;
    title 'Logistic Regression Analysis (Dependent Variable: host_is_superhost)';
run;
```

Appendix Figure 14: Code for output in Figure 21, Figure 22, Figure 23, Figure 24, Figure 25 and Figure 26

```
/* ANOVA (review_scores_communication VS Host_response_time type)*/
ods graphics on;
proc glm data=STATDATA.PROPERTYCLEANED  plots=(residuals diagnostics);
title 'Analysis of Variance (Review_score_communication VS Host_response_time type)';
    class host_response_time_num;
    model review_scores_communication = host_response_time_num;
    lsmeans host_response_time_num / adjust = tukey;
    means host_response_time_num / hovtest =levene;
run;
```

Appendix Figure 15: Code for output in Figure 27, Figure 28 and Figure 29