Data Analysis of the New York City Airbnb Dataset

New York City Airbnb Dataset / SAS Studio

Didem Bulut Aykurt

MIS540-1 – Introduction to Business Intelligence

Colorado State University-Global Campus

Dr. Alin Tomoiaga

October 6, 2022

Table of Contents

The New York City Airbnb’s Exploratory Data Analysis

A privately owned multinational corporation with its headquarters in San Francisco, Airbnb, Inc. runs an online trade and accommodation business that may be accessed through its apps and web apps. Subscribers of the website can use it to book or provide accommodation, generally guesthouses or travel opportunities. It is the world's most prominent domestic rental firm, with over 4 million listings spread across more than 81,000 cities and 191 nations.

One of the company’s customers, my family, and I. We always use the Airbnb web to choose the best price at a beautiful place. That company name is so unique that three words make the company name Air, Bed, and Breakfast. That is why I chose the company and want to see how the company increases loyalty. Feedback from customers and reviews are crucial in improving a customer's commitment to a company. They truly have the power to build or break a company. As a result, businesses must study the elements contributing to higher ratings and understand what the public thinks of the product.

To examine the relationship between written reviews and numerical ratings, I used SAS Studio to analyze customer ratings and forecast the key elements contributing to higher ratings. I then compared those findings to the numerical ratings. Additionally, I conducted a descriptive study to examine a few crucial factors that would be very beneficial for business, such as:

1. What New York neighborhoods are the most popular for Airbnb rentals?

2. Which Airbnb guests most love local neighborhoods?

For the analysis, a public database from the Airbnb platform was utilized. The dataset offers details on the characteristics of homes, review ratings, comments, and the availability of more than 10,000 listings in 2019. The Airbnb data was employed to execute visualizations, and SAS studio additionally carried out linear regression to identify the elements influencing higher ratings. SAS was also used to analyze consumer reviews.

This report is a data analysis of Airbnb in New York City. The information is divided into three milestones. Milestone 1 defines the business problem of Airbnb and some basic statistics such as data defined, descriptive statistics of the dataset, etc. Milestone project 2 describes four minimum business problems and creates alternate and null hypotheses for each business question. It also includes testing the ideas with an appropriate statistical test. Finally, milestone project three exploration of data visualization (such as customer ratings and reviews) and performing a predictive analysis technique.

The name of the dataset is Airbnb, an open data source available on Kaggle, and the variables required to address the business problem.

The business problem of Airbnb states that we can say which neighborhood has the highest prices range for the listings. From this, we can find out that the solution to the problem is to regulate the price of neighborhood hotels or rooms.

In this data set, there are 16 variables or columns, including 11 numerical variables and five-character variables.

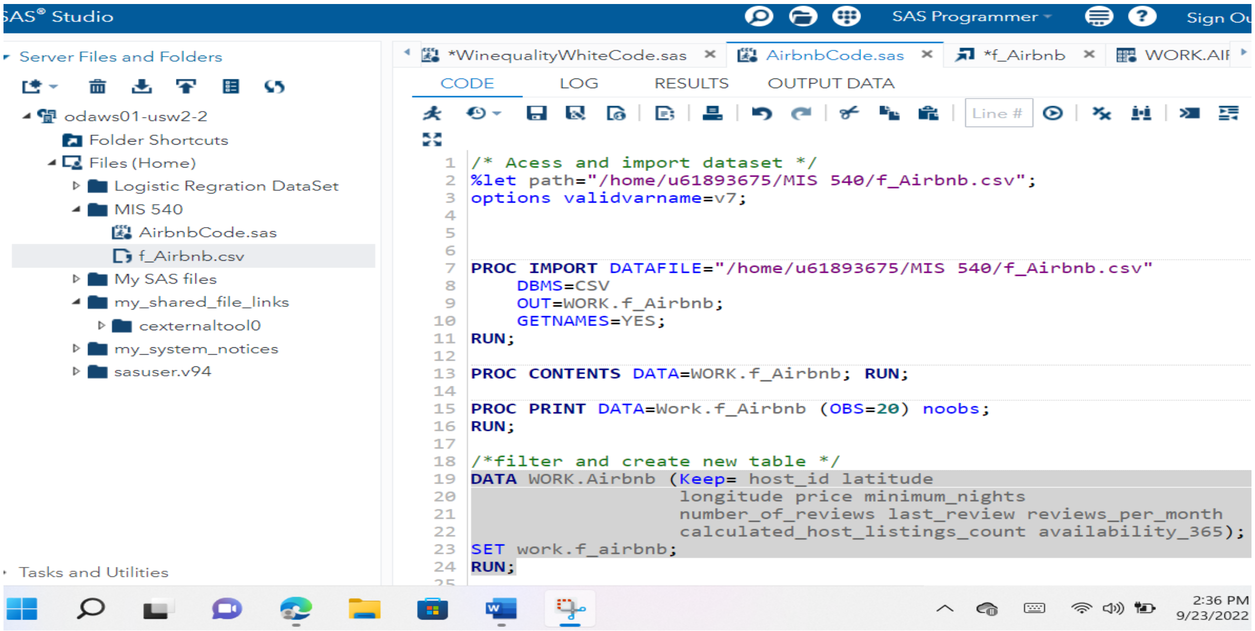
Data Description of Listings, Calendar, and Reviews

Variable	**Description**
ID	Listing id of the property
Name	Name of the property
Host_Id	Id of the property host
host_name	Name of the host property

Price	Price of the property
availability_365	Availability of property
Calculated_host_ _Listings_count	Total listings the host has
Neighbourhood group	The neighborhood of the property
neighborhood	The neighborhood of the property
Min_nights	Minimum number of nights required to book
Reviews_per_month	Average Number of reviews in a month
room_type	Type of the room
number_of_reviews	Total number of reviews
last_review	Date of the last review
Latitude	Location of the Latitude
Longitude	Location of the Longitude

SAS Description of the Airbnb data set is as follows:

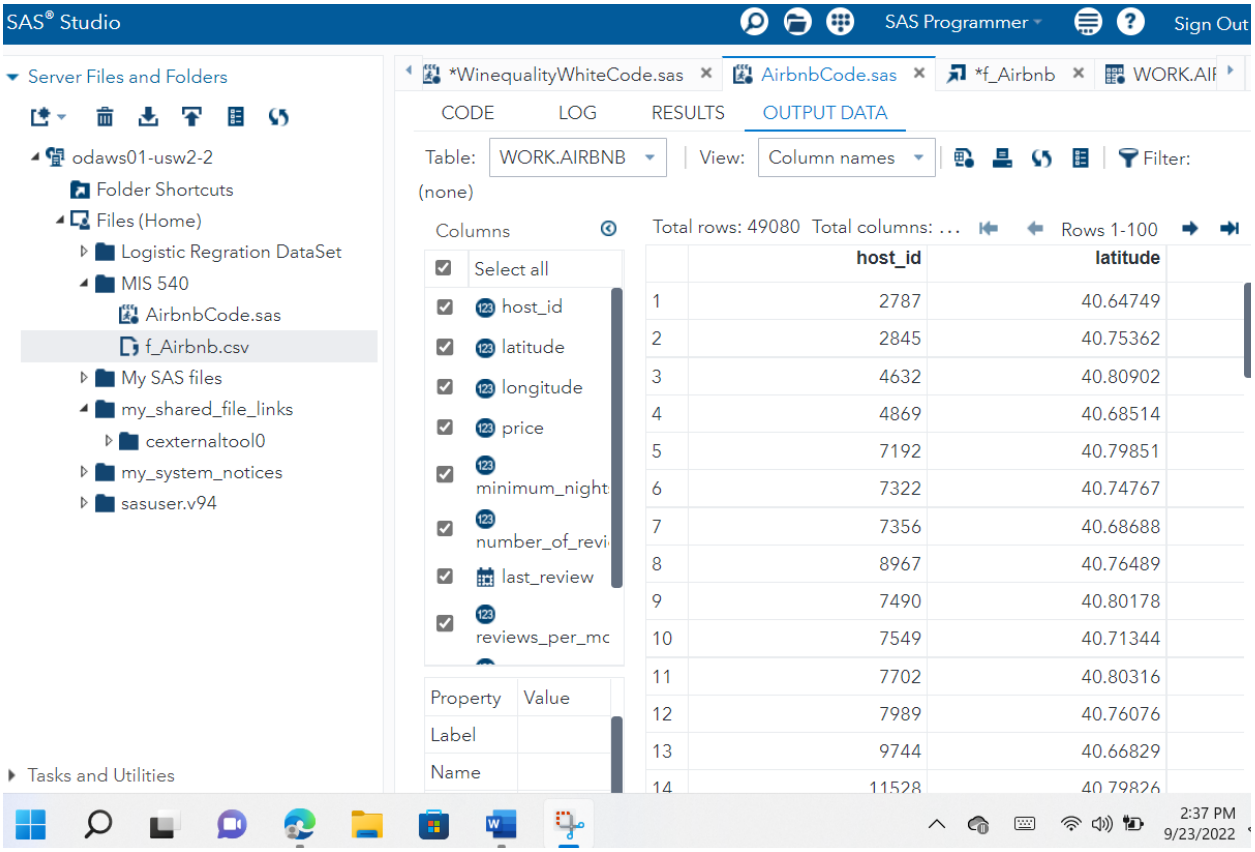
Figure 1: *In this part, we will use the SAS program to import and filter the dataset.*



```
/* Acess and import dataset */
%let path="/home/u61893675/MIS 540/f_Airbnb.csv";
options validvarname=v7;



PROC IMPORT DATAFILE="/home/u61893675/MIS 540/f_Airbnb.csv"
    DBMS=CSV
    OUT=WORK.f_Airbnb;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.f_Airbnb; RUN;

PROC PRINT DATA=Work.f_Airbnb (OBS=20) noobs;
RUN;

/*filter and create new table */
DATA WORK.Airbnb (Keep= host_id latitude
                  longitude price minimum_nights
                  number_of_reviews last_review reviews_per_month
                  calculated_host_listings_count availability_365);
SET work.f_airbnb;
RUN;
```

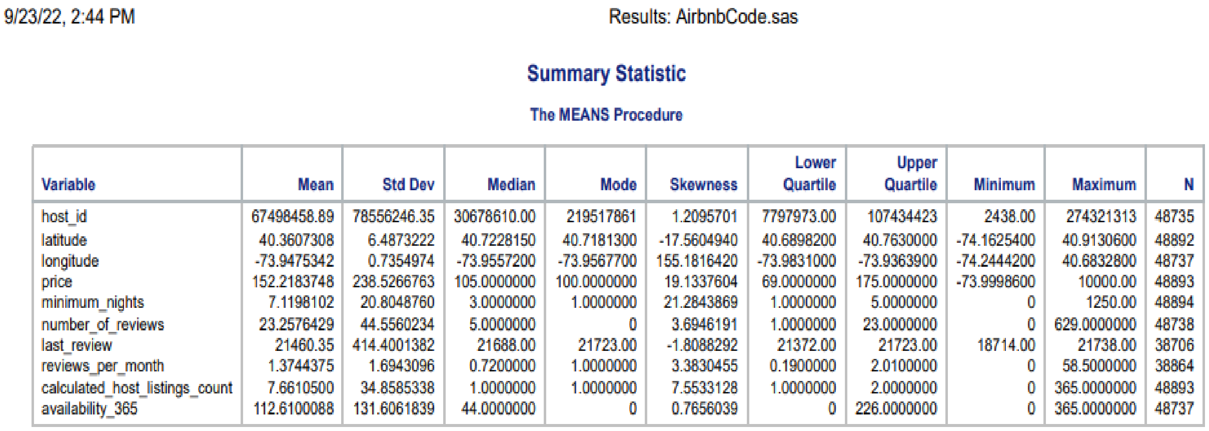
These variables 'id,’ 'host_name,’ and 'last_review' are not needed to address the business problem because these drop variables are irrelevant and insignificant to our investigation.

Figure 2: *The filtered dataset was created in SAS Studio.*



	host_id	latitude
1	2787	40.64749
2	2845	40.75362
3	4632	40.80902
4	4869	40.68514
5	7192	40.79851
6	7322	40.74767
7	7356	40.68688
8	8967	40.76489
9	7490	40.80178
10	7549	40.71344
11	7702	40.80316
12	7989	40.76076
13	9744	40.66829
14	11528	40.79826

Here is the SAS output, which describes the Airbnb summary statistics as follows:



9/23/22, 2:44 PM

Results: AirbnbCode.sas

Summary Statistic

The MEANS Procedure

Variable	Mean	Std Dev	Median	Mode	Skewness	Lower Quartile	Upper Quartile	Minimum	Maximum	N
host_id	67498458.89	78556246.35	30678610.00	219517861	1.2095701	7797973.00	107434423	2438.00	274321313	48735
latitude	40.3607308	6.4873222	40.7228150	40.7181300	-17.5604940	40.6898200	40.7630000	-74.1625400	40.9130600	48892
longitude	-73.9475342	0.7354974	-73.9557200	-73.9567700	155.1816420	-73.9831000	-73.9363900	-74.2444200	40.6832800	48737
price	152.2183748	238.5266763	105.0000000	100.0000000	19.1337604	69.0000000	175.0000000	-73.9998600	10000.00	48893
minimum_nights	7.1198102	20.8048760	3.0000000	1.0000000	21.2843869	1.0000000	5.0000000	0	1250.00	48894
number_of_reviews	23.2576429	44.5560234	5.0000000	0	3.6946191	1.0000000	23.0000000	0	629.0000000	48738
last_review	21460.35	414.4001382	21688.00	21723.00	-1.8088292	21372.00	21723.00	18714.00	21738.00	38706
reviews_per_month	1.3744375	1.6943096	0.7200000	1.0000000	3.3830455	0.1900000	2.0100000	0	58.5000000	38864
calculated_host_listings_count	7.6610500	34.8585338	1.0000000	1.0000000	7.5533128	1.0000000	2.0000000	0	365.0000000	48893
availability_365	112.6100088	131.6061839	44.0000000	0	0.7656039	0	226.0000000	0	365.0000000	48737

Business Question and Hypothesis

These are the *business problems* that will explore some key points which would be very helpful for business, such as:

1. Is there a difference in the room types based on the property's price?

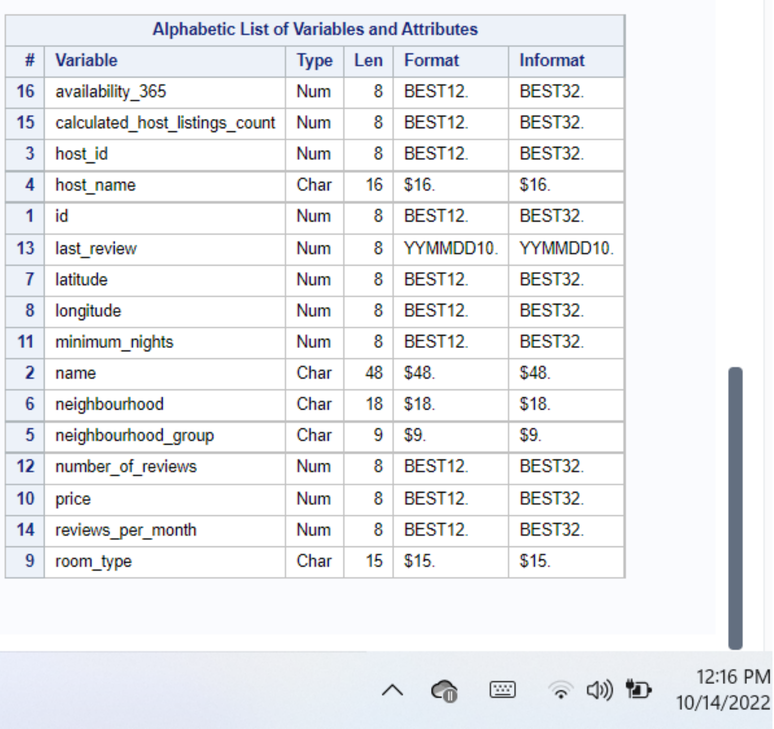
2. Is there a difference in the room types based on the total number of reviews?

3. Is there a difference in the neighborhood of the property based on the price of the property?

4. Is there a difference in the room types based on the property's availability?

The *organization strategic goal* is that the constant *goals* of Airbnb's strategy were to expand into new areas and deliver more inventory within the company's network.

SAS Description of the Airbnb data set is as follows:



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
16	availability_365	Num	8	BEST12.	BEST32.
15	calculated_host_listings_count	Num	8	BEST12.	BEST32.
3	host_id	Num	8	BEST12.	BEST32.
4	host_name	Char	16	$16.	$16.
1	id	Num	8	BEST12.	BEST32.
13	last_review	Num	8	YYMMDD10.	YYMMDD10.
7	latitude	Num	8	BEST12.	BEST32.
8	longitude	Num	8	BEST12.	BEST32.
11	minimum_nights	Num	8	BEST12.	BEST32.
2	name	Char	48	$48.	$48.
6	neighbourhood	Char	18	$18.	$18.
5	neighbourhood_group	Char	9	$9.	$9.
12	number_of_reviews	Num	8	BEST12.	BEST32.
10	price	Num	8	BEST12.	BEST32.
14	reviews_per_month	Num	8	BEST12.	BEST32.
9	room_type	Char	15	$15.	$15.

Now, we will create the null and alternative hypotheses for each business.

Questions are as follows:

- First business problem hypotheses are as follows:

  Null hypothesis: There is no difference between room types based on the prices of the property.

  Alternative Hypothesis: at least one group differs significantly from the overall mean price of the property.

  Based on the table 1 result (Appendix), the p-value of the effect is close to zero and less than the significance level, implying that 0.000 < 0.05. So, we can reject the null hypothesis in favor of an alternative idea. Therefore, we can conclude that there is a significant difference and that at least one room type differs significantly from the overall mean property price. In simple words, the solution to the business problem indicates that the room types (i.e., Entire home/apt, Private Room, and Shared room) differ for the property's price.

- Second business problem hypotheses are as follows:

  Null hypothesis: There is no difference between room types based on the total number of reviews.

  Alternative Hypothesis: at least one group differs significantly from the overall mean of the total number of reviews.

  Based on the table 2 result (Appendix), the p-value of the effect is close to zero and less than the significance level, implying that 0.000 < 0.05. So, we can reject the null hypothesis in favor of an alternative hypothesis. Therefore, we can conclude that there is a significant difference and that at least one room type differs significantly from the overall mean of the total number of reviews.

- Third business problem hypothesis is as follows:

Null hypothesis: There is no difference between the neighborhood of the property based on the price of the property

Alternative Hypothesis: at least one group differs significantly from the overall mean cost of the property.

Based on the table 3 result (Appendix), the p-value of the effect is close to zero and less than the significance level, implying that 0.000 < 0.05. So, we can reject the null hypothesis in favor of an alternative hypothesis. Therefore, we can conclude that there is a significant difference and that at least one neighborhood group differs significantly from the overall mean property price.

- Fourth business problem hypothesis is as follows:

  Null hypothesis: There is no difference between room types based on the property's availability.

  Alternative Hypothesis: at least one group differs significantly from the overall mean of availability of the property.

  Based on table 4 result (Appendix), the p-value of the effect is close to zero and less than the significance level, implying that 0.000 < 0.05. So, we can reject the null hypothesis in favor of an alternative hypothesis. Therefore, we can conclude that there is a significant difference and that at least one room type differs significantly from the overall mean of property availability.

All the hypotheses are framed according to the business question, and then we perform an ANOVA test for all the hypotheses: a one-way ANOVA analysis of variance. For the reason that one variable is categorical and the other is numerical or continuous, note that one thing, every definite class is more than 2, so that's why we can perform a one-way ANOVA analysis; if the flat style is less than or equal to 2, then we are not able to achieve this test. In this case, we have to perform an independent-sample t-test.

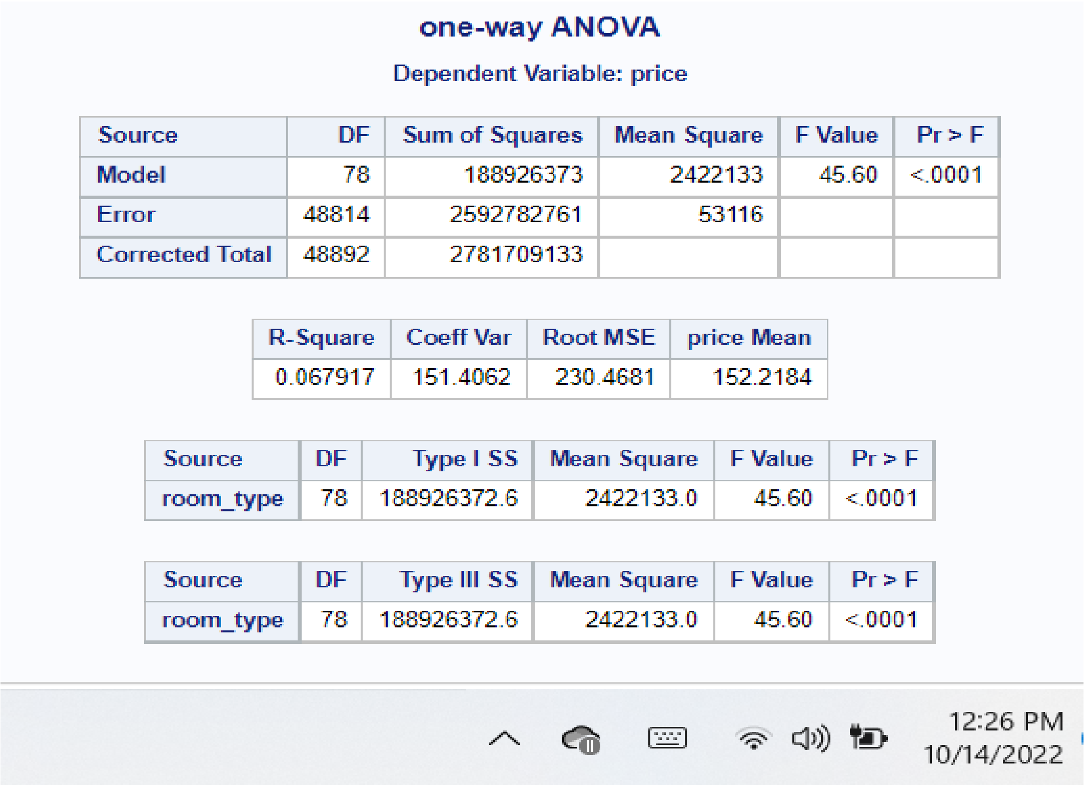
Visualization of Airbnb Dataset:

Concerns:

In milestone 2, I am facing an issue related to creating the hypotheses, like which hypothesis is more effective and implementing code is complex for ANOVA in this dataset. So, this is the concern in completing the portfolio project.

Appendix 2: (SAS Output for business problem hypotheses)

1.



one-way ANOVA

Dependent Variable: price

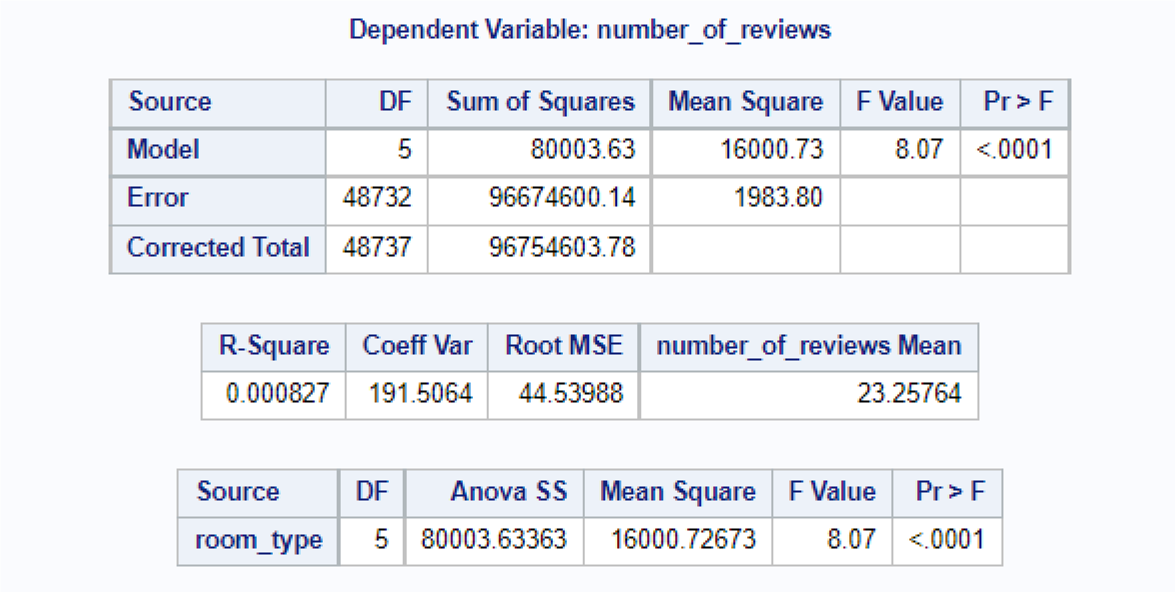
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	78	188926373	2422133	45.60	<.0001
Error	48814	2592782761	53116		
Corrected Total	48892	2781709133			

R-Square	Coeff Var	Root MSE	price Mean
0.067917	151.4062	230.4681	152.2184

Source	DF	Type I SS	Mean Square	F Value	Pr > F
room_type	78	188926372.6	2422133.0	45.60	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
room_type	78	188926372.6	2422133.0	45.60	<.0001

2.



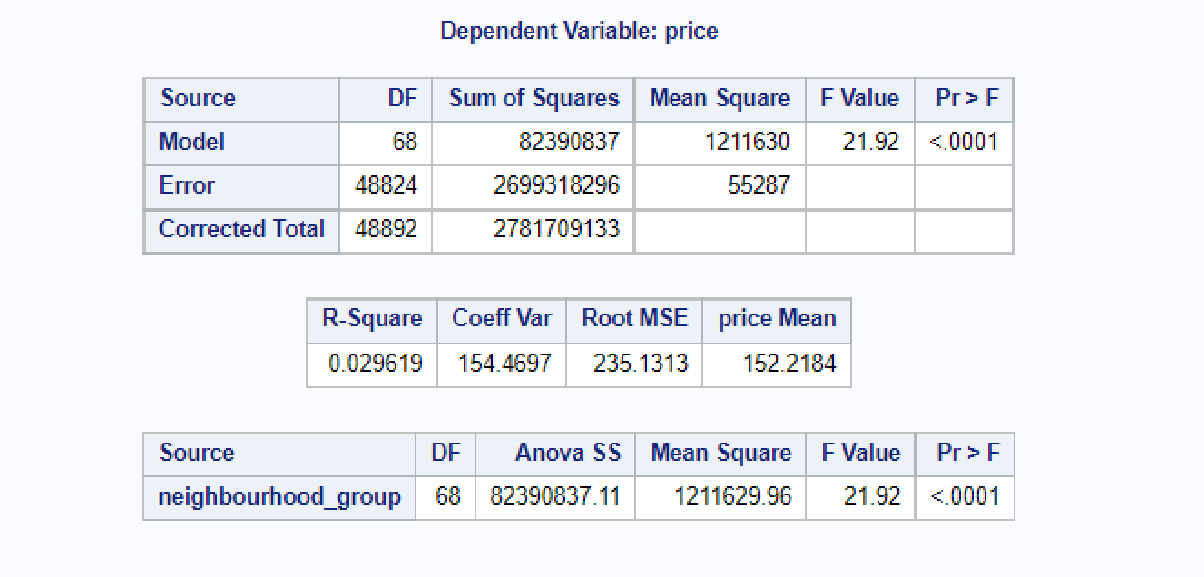
Dependent Variable: number_of_reviews

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	80003.63	16000.73	8.07	<.0001
Error	48732	96674600.14	1983.80		
Corrected Total	48737	96754603.78			

R-Square	Coeff Var	Root MSE	number_of_reviews Mean
0.000827	191.5064	44.53988	23.25764

Source	DF	Anova SS	Mean Square	F Value	Pr > F
room_type	5	80003.63363	16000.72673	8.07	<.0001

3.



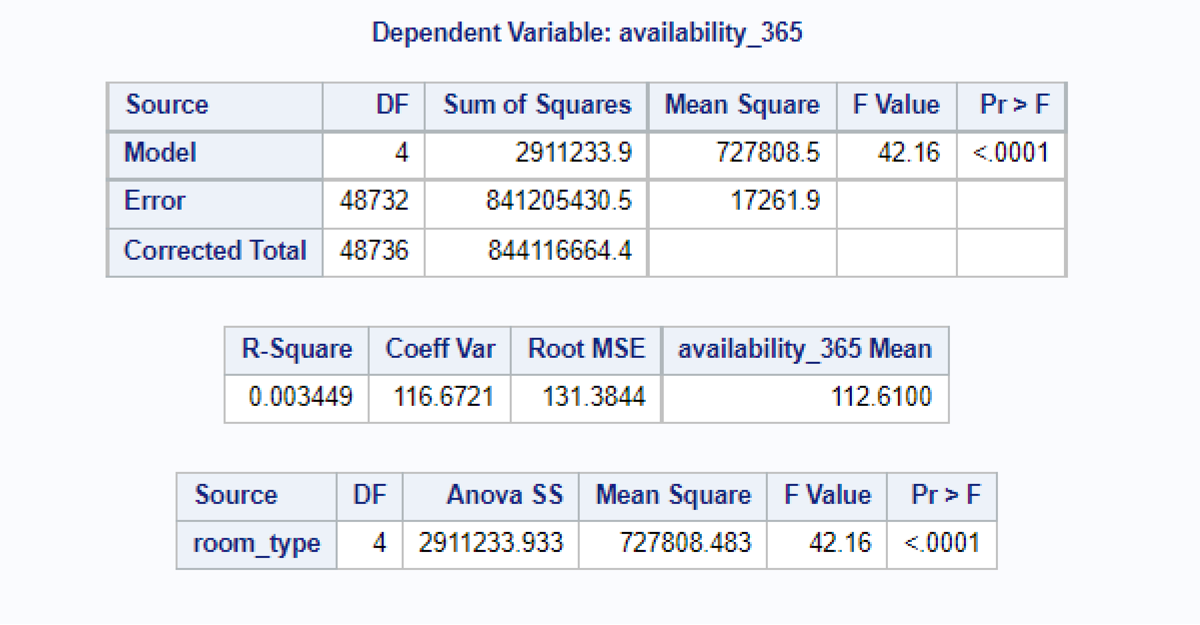
Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	68	82390837	1211630	21.92	<.0001
Error	48824	2699318296	55287		
Corrected Total	48892	2781709133			

R-Square	Coeff Var	Root MSE	price Mean
0.029619	154.4697	235.1313	152.2184

Source	DF	Anova SS	Mean Square	F Value	Pr > F
neighbourhood_group	68	82390837.11	1211629.96	21.92	<.0001

4.



Dependent Variable: availability_365

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2911233.9	727808.5	42.16	<.0001
Error	48732	841205430.5	17261.9		
Corrected Total	48736	844116664.4			

R-Square	Coeff Var	Root MSE	availability_365 Mean
0.003449	116.6721	131.3844	112.6100

Source	DF	Anova SS	Mean Square	F Value	Pr > F
room_type	4	2911233.933	727808.483	42.16	<.0001

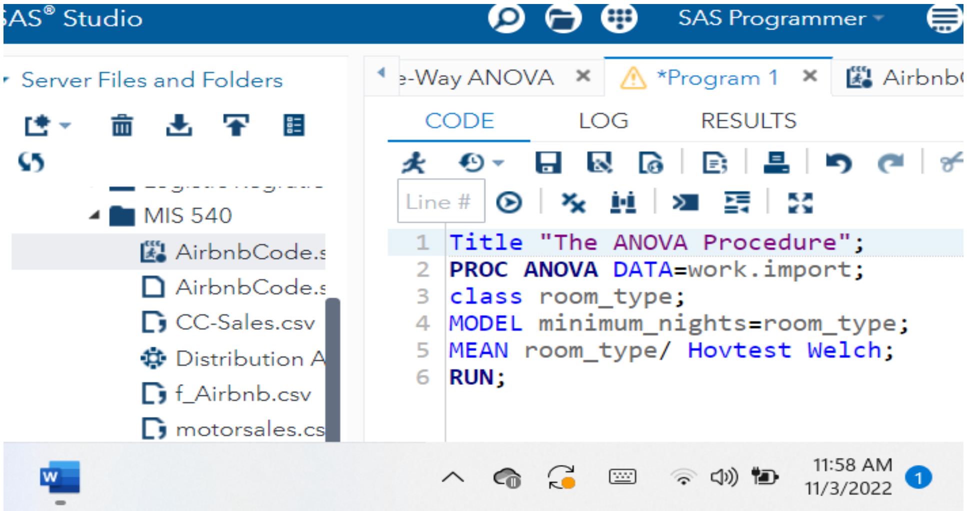
Analysis of findings concerning business question and hypothesis

For this milestone project, further define hypotheses. This section will highlight the hypotheses and business questions. It also includes the exploration and visualization of the Airbnb dataset. At the end of this milestone, will perform a predictive analysis of the dataset.

The business question is whether the room types differ based on the minimum number of nights required to book. The null hypothesis is that there is no difference between room types based on the minimum number of nights required to book, and the alternative is that there is a difference between the room types based on the minimum number of nights. Thus, the business question tells us that all the room types are different compared to the minimum number of nights required to book.

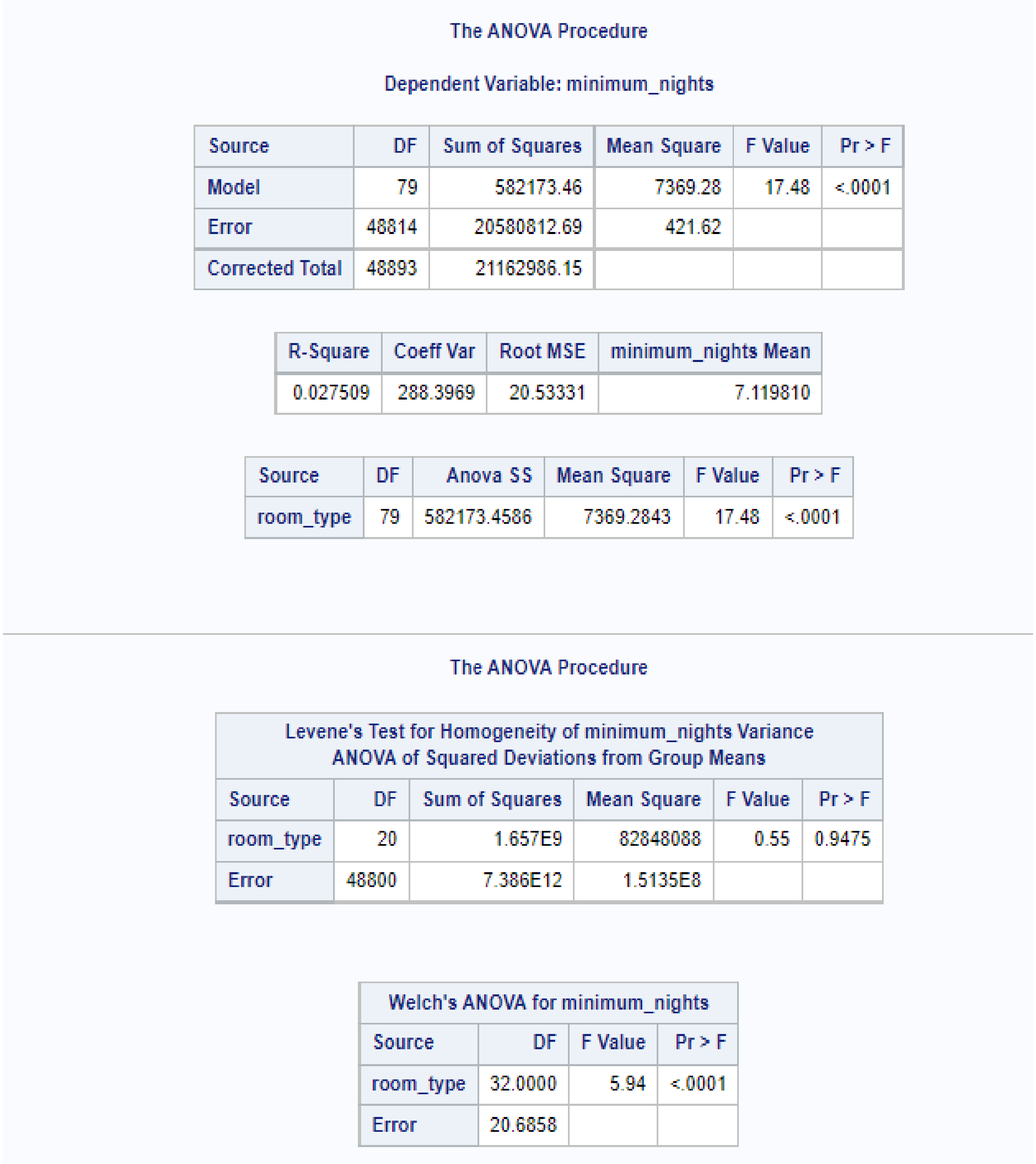
We have to use a one-way ANOVA analysis of variance to test the hypothesis.

SAS Output:



```
Title "The ANOVA Procedure";
PROC ANOVA DATA=work.import;
class room_type;
MODEL minimum_nights=room_type;
MEAN room_type/ Hovtest Welch;
RUN;
```

Based on the below output, the p-value of the test is close to zero, and the significance level is 0.05. We can say that the p-value of the test is less than the level of significance, which is 0.00 < 0.05. This means we can reject the null hypothesis and support an alternative one. Thus, we can easily conclude that there is a significant difference between the room types and the minimum number of nights required to book.



The ANOVA Procedure

Dependent Variable: minimum_nights

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	79	582173.46	7369.28	17.48	<.0001
Error	48814	20580812.69	421.62		
Corrected Total	48893	21162986.15			

R-Square	Coeff Var	Root MSE	minimum_nights Mean
0.027509	288.3969	20.53331	7.119810

Source	DF	Anova SS	Mean Square	F Value	Pr > F
room_type	79	582173.4586	7369.2843	17.48	<.0001

The ANOVA Procedure

Levene's Test for Homogeneity of minimum_nights Variance
ANOVA of Squared Deviations from Group Means

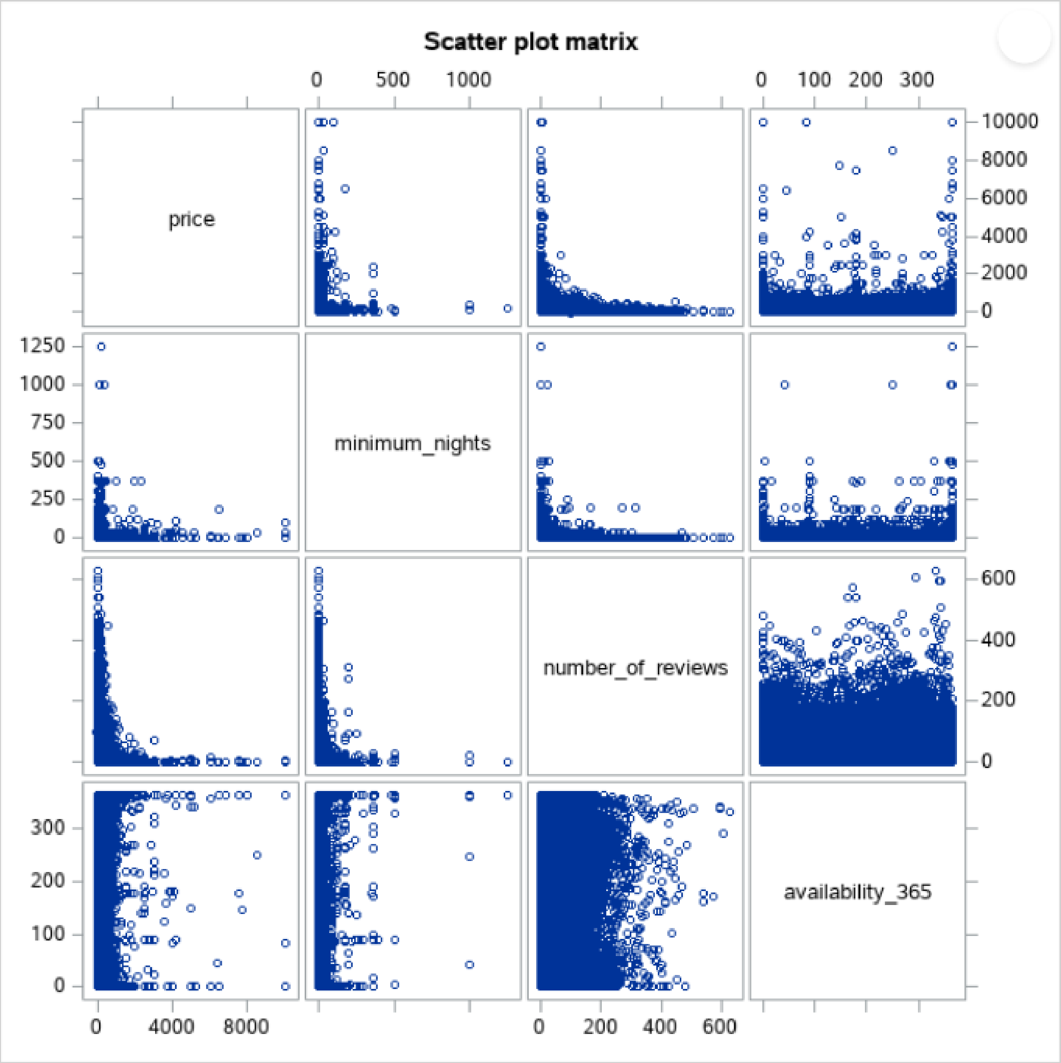
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
room_type	20	1.657E9	82848088	0.55	0.9475
Error	48800	7.386E12	1.5135E8		

Welch's ANOVA for minimum_nights

Source	DF	F Value	Pr > F
room_type	32.0000	5.94	<.0001
Error	20.6858		

Exploring and Visualizing Airbnb Dataset

11/3/22, 12:06 PM Results: Data Exploration

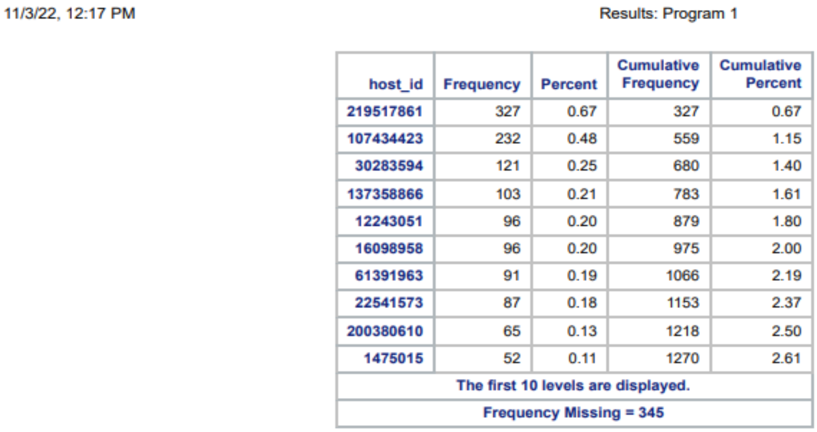


In the exploration, we want to know whether there is any significant correlation among the variables. So, based on the scatter plot matrix, we can conclude that there is no significant correlation among the variables, i.e., price, minimum_nights, number_of_reviews, and availability_365.

Descriptive Statistic

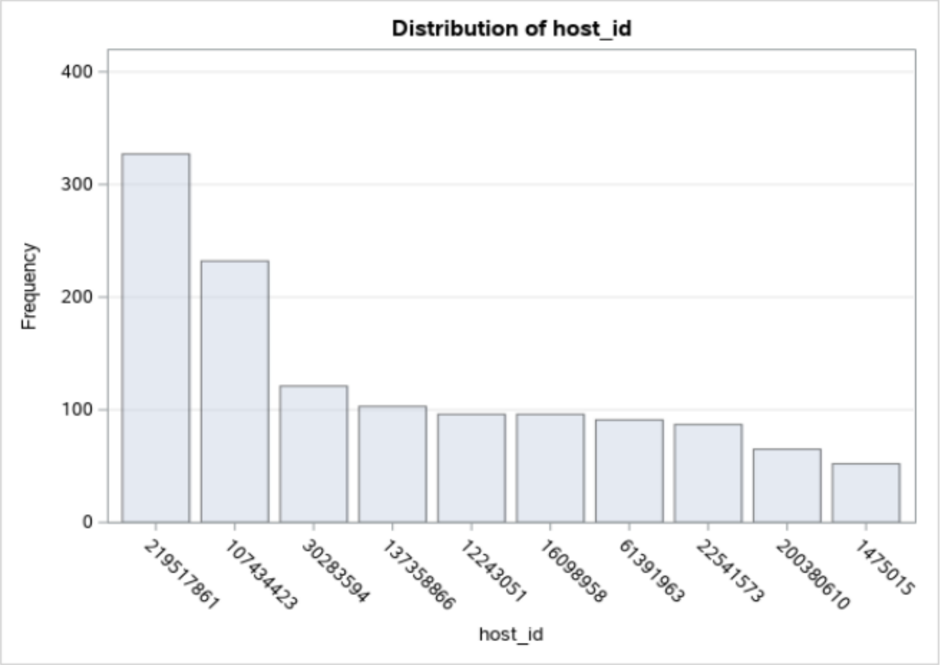
One-Way Frequencies

The **Frequency** column shows how many data points fell into the product category. The **Percent** column specifies the percentage of data points in that category. The **Cumulative Frequency** column indicates the adding all the numbers in the Frequency column above and includes the current row. The last column on the table is the **Cumulative Percent** shows the adding all the Percent columns up to the current row. The host_id dataset has a missing value of 345.

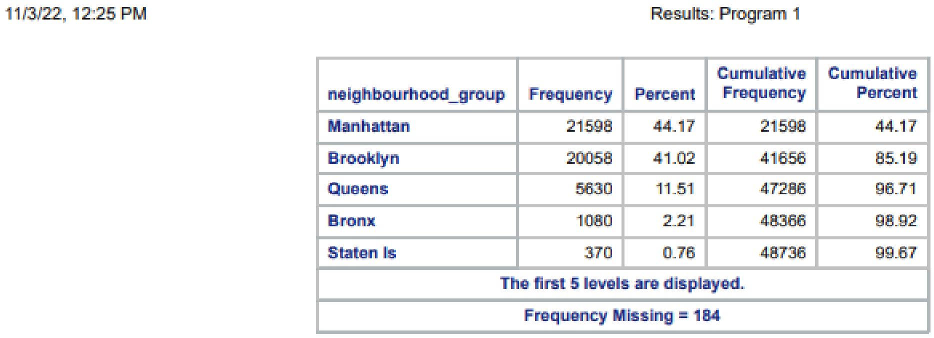


11/3/22, 12:17 PM Results: Program 1

host_id	Frequency	Percent	Cumulative Frequency	Cumulative Percent
219517861	327	0.67	327	0.67
107434423	232	0.48	559	1.15
30283594	121	0.25	680	1.40
137358866	103	0.21	783	1.61
12243051	96	0.20	879	1.80
16098958	96	0.20	975	2.00
61391963	91	0.19	1066	2.19
22541573	87	0.18	1153	2.37
200380610	65	0.13	1218	2.50
1475015	52	0.11	1270	2.61
The first 10 levels are displayed.				
Frequency Missing = 345				

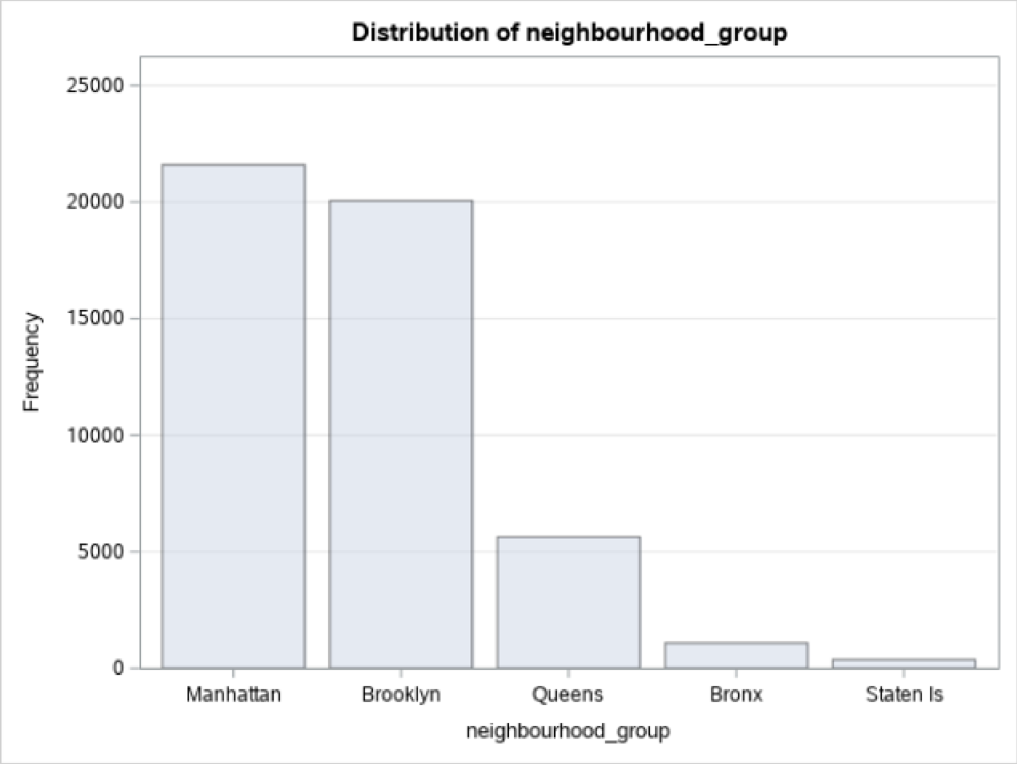


From the above chart, it’s interesting to note that the top 10 hosts with the most listings have a good distribution. More than 300+ listings are on the first host. On the other side, out of 10 hosts, we can observe that 6 of them have fewer than 100 listings.

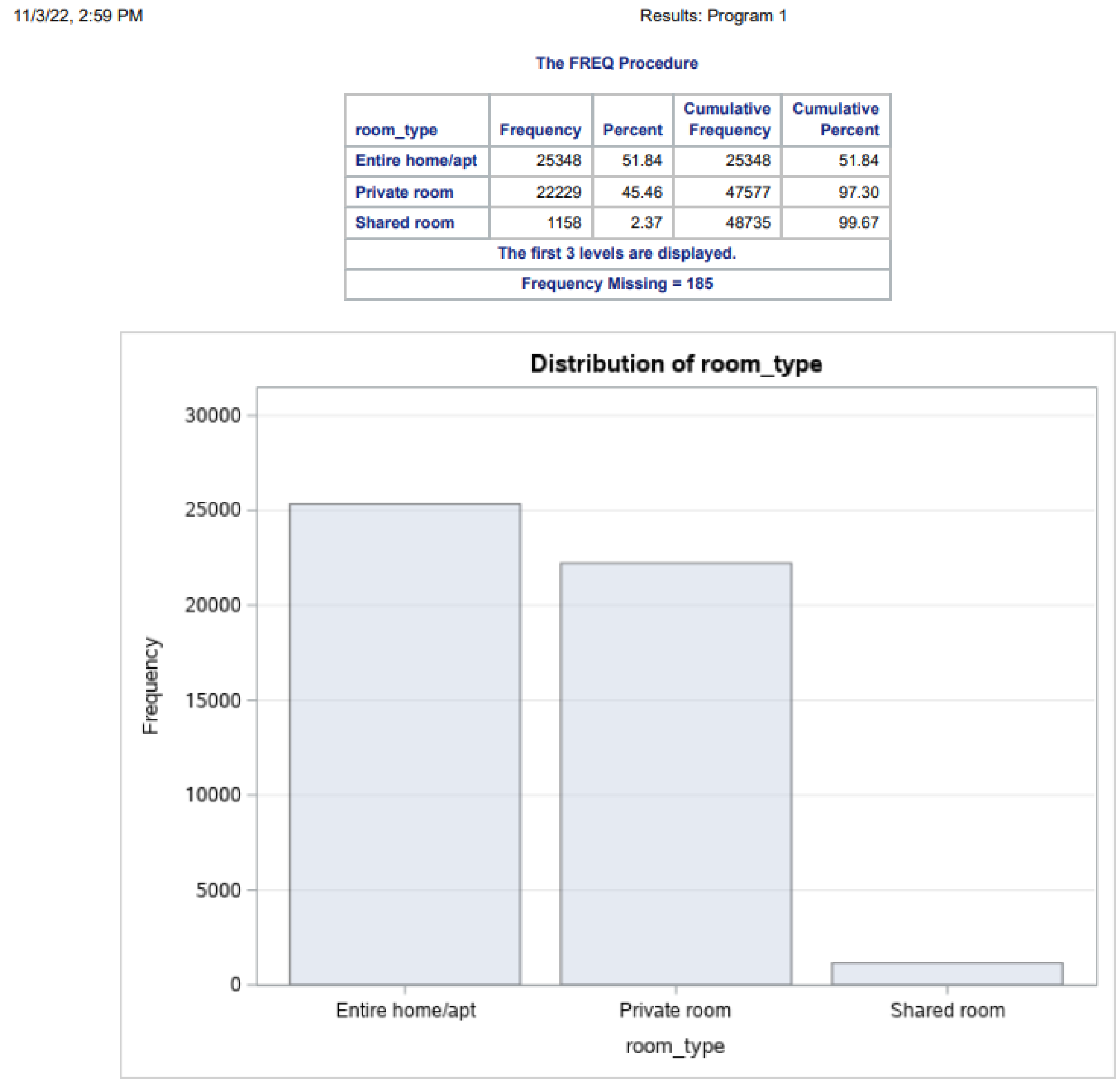


11/3/22, 12:25 PM Results: Program 1

neighbourhood_group	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Manhattan	21598	44.17	21598	44.17
Brooklyn	20058	41.02	41656	85.19
Queens	5630	11.51	47286	96.71
Bronx	1080	2.21	48366	98.92
Staten Is	370	0.76	48736	99.67
The first 5 levels are displayed.				
Frequency Missing = 184				



In the above chart, we can see that there is a good distribution between the top 5 neighborhood groups with the most listings. Manhattan group has more than 2000+ listings.



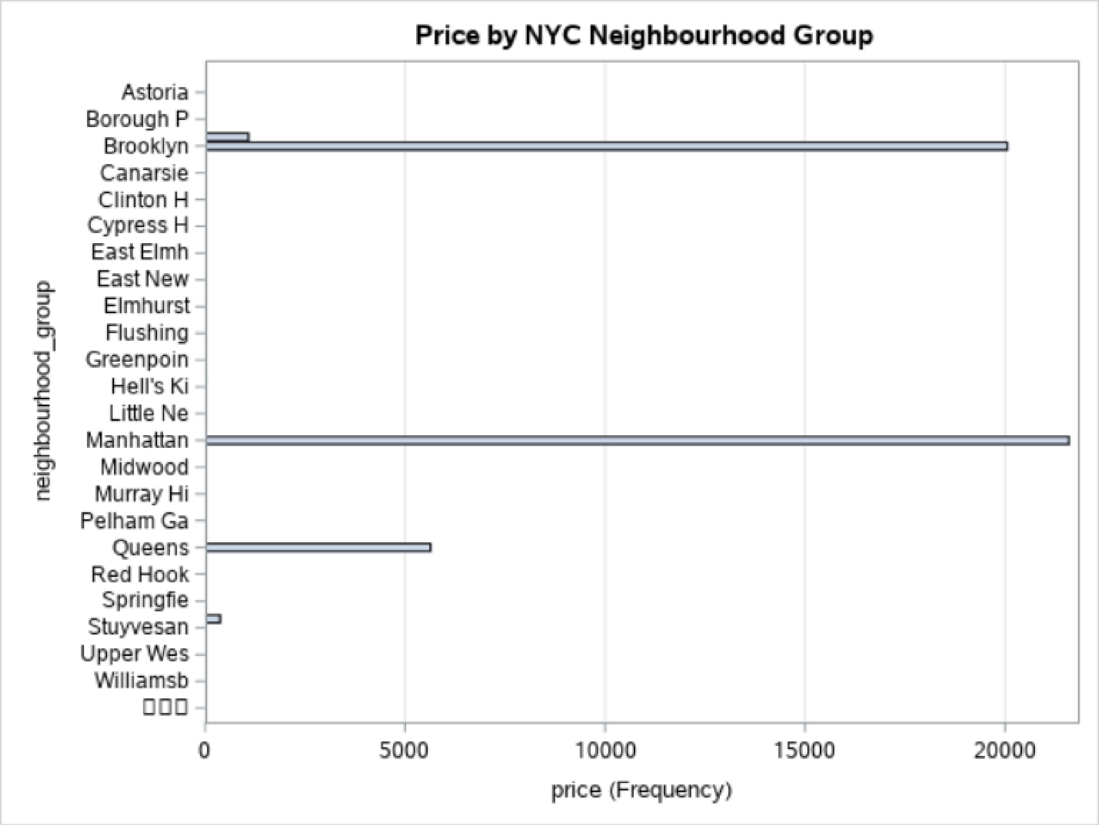
11/3/22, 2:59 PM Results: Program 1

The FREQ Procedure

room_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Entire home/apt	25348	51.84	25348	51.84
Private room	22229	45.46	47577	97.30
Shared room	1158	2.37	48735	99.67
The first 3 levels are displayed.				
Frequency Missing = 185				

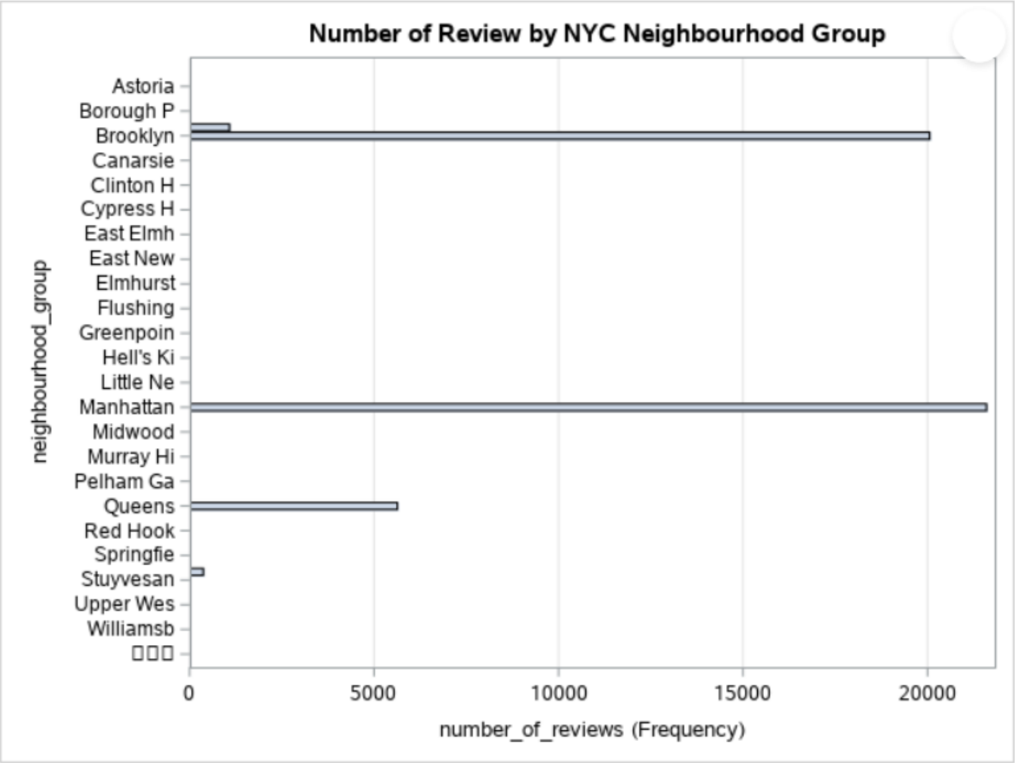
We can see from the aforementioned data that the room types with the most listings are distributed fairly. The entire home/apartment includes more than 2500+ listings.

11/3/22, 12:44 PM Results: Program 1



We can see a few things about the pricing pattern for Airbnb in NYC with a bar chart. As a starting point, Manhattan has the broadest range of pricing for the listings, with a mean observation of $150, followed by Brooklyn with $90 per night. The Bronx is the least expensive of the three, with distributions that seem relatively comparable in Queens and Staten Island. As an illustration, the Bronx looks to have lower standards of living than Manhattan, known for being among the world's most expensive cities.

11/3/22, 1:02 PM Results: Program 1

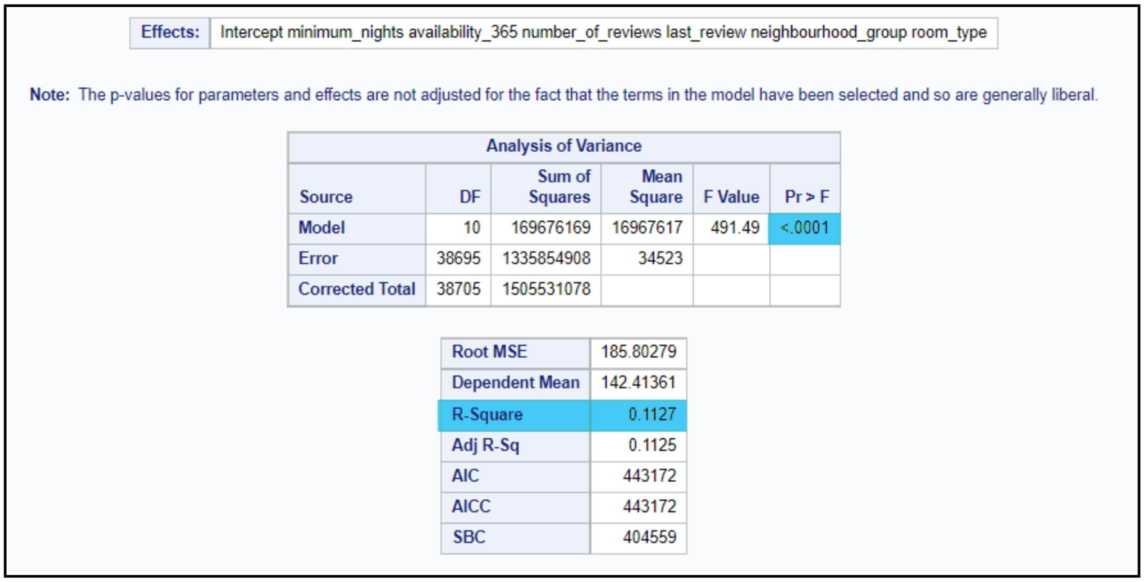


The above chart shows a few things about the pattern of the number of reviews for Airbnb in NYC with a bar chart. Manhattan has the broadest range of reviews for the listings as a starting point, and Brooklyn has the second-highest, most comprehensive range of reviews. We can treat without bar line neighborhoods with the least number of reviews.

Predictive Analysis

In this analysis section, we will create predictive modeling with the help of a mathematical process to predict future events/outcomes for Airbnb through a regression technique and analyze patterns in a given set of input data.

Here is the SAS program and its output:



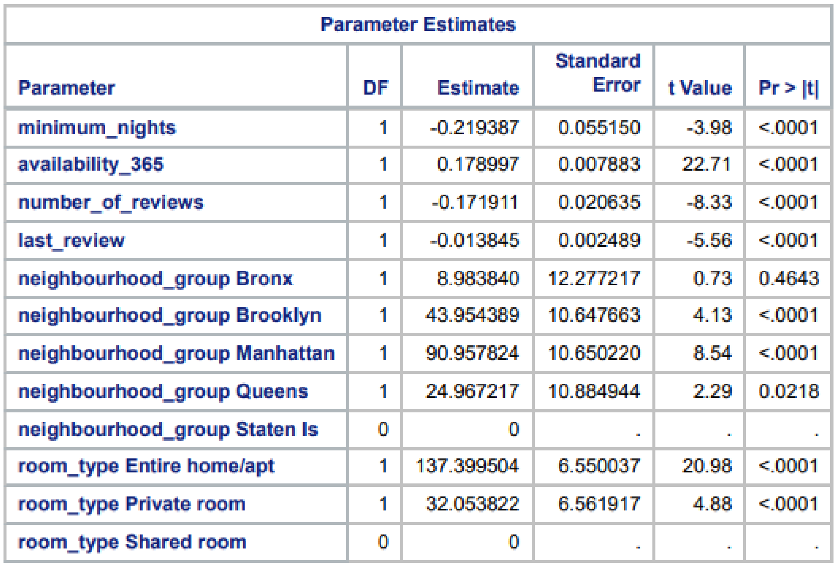
Effects:	Intercept minimum_nights availability_365 number_of_reviews last_review neighbourhood_group room_type

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	169676169	16967617	491.49	<.0001
Error	38695	1335854908	34523		
Corrected Total	38705	1505531078			

Root MSE	185.80279
Dependent Mean	142.41361
R-Square	0.1127
Adj R-Sq	0.1125
AIC	443172
AICC	443172
SBC	404559

11/3/22, 1:09 PM Results: Program 1



Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
minimum_nights	1	-0.219387	0.055150	-3.98	<.0001
availability_365	1	0.178997	0.007883	22.71	<.0001
number_of_reviews	1	-0.171911	0.020635	-8.33	<.0001
last_review	1	-0.013845	0.002489	-5.56	<.0001
neighbourhood_group Bronx	1	8.983840	12.277217	0.73	0.4643
neighbourhood_group Brooklyn	1	43.954389	10.647663	4.13	<.0001
neighbourhood_group Manhattan	1	90.957824	10.650220	8.54	<.0001
neighbourhood_group Queens	1	24.967217	10.884944	2.29	0.0218
neighbourhood_group Staten Is	0	0	.	.	.
room_type Entire home/apt	1	137.399504	6.550037	20.98	<.0001
room_type Private room	1	32.053822	6.561917	4.88	<.0001
room_type Shared room	0	0	.	.	.

The regression analysis is used to predict the price of the property of Airbnb. In research, our dependent variable is price; on the other hand, the independent variables are minimum_nights, availability_365, number_of_reviews, last_review, neighbourhood_group (categorical variable), and room_type (categorical variable).

In the model above, stepwise regression was utilized, which is the iterative process of building a regression model step by step while selecting explanatory variables to be included in the regression analysis. Additionally, the possible informative factors are successively added or removed during each step, and a statistically significant difference is tested.

Look at the Pr<|t| column; as we observe the p-values of all the variables, there is only one dummy variable, i.e., neighbourhood_group Bronx is not statistically significant because the significance level is more than the p-value; the rest of the variables are statistically significant.

As we see in the output, two variables (i.e., from room type, it is “shared room” and from a neighborhood group, it is “Staten Is”) coefficients are zero for the reason that these variables are a benchmark category of dummy variables.

Model Diagnostic

The r-square of the model is very low, i.e., 11%, which implies that 11% of the variation in the dependent variable, i.e., the price of the property, can be explained by the variation in the independent variables. We can conclude that the model is not appropriate for the data to the fitted regression line. On the other hand, the overall model’s p-value is less than the significance level, which means that the model is statistically significant.

Recommendations for further analysis are to add relevant variables such as Review_scores_rating, the Cancellation policy of property, Security deposit, Host is Superhost, etc. If applicable or appropriate variables are added to the model, there is a higher chance of a more elevated r square. This means that predictive analysis would give us better and more insightful results.

References

Bode, O., Toader, V., & Rus, R. (2022). Pricing Strategies of Porto’s Airbnb New Listings. International Conference On Tourism Research, 15(1), 425-432. https://doi.org/10.34190/ictr.15.1.249

Guggilla, Chakraborty, P. Price Recommendation Engine for Airbnb. Support.sas.com. Retrieved from https://support.sas.com/resources/papers/proceedings17/1326-2017.pdf

Kas, J., Delnoij, J., Corten, R., & Parigi, P. (2022). Trust spillovers in the sharing economy: Does international Airbnb experience foster cross-national trust? Journal Of Consumer Behaviour, 21(3), 509-522. https://doi.org/10.1002/cb.2014

SAS Documentation. Support.sas.com. (2022). Retrieved from https://support.sas.com/en/documentation.html

Vlogger, M., Pforr, C., Stawinoga, A., Taplin, R., & Matthews, S. (2018). Who adopts the Airbnb innovation? An analysis of international visitors to Western Australia. Tourism Recreation Research, 43(3), 305-320. https://doi.org/10.1080/02508281.2018.1443052