Data Preparation and Handling

Project Report

[Pick the date]

**Part 1: Dataset and task description**

I selected a dataset of Apartment/ Hotel listings for Data preparation and Handling Group project. Our data contains variables such as the listing ID, location of those listings- latitude, longitude and neighborhood, Host ID, price, customer reviews- the number of reviews, date of reviews and the number of reviews per month, number of properties listed by the host, availability of the accommodation during the year and the minimum number of nights it can be let out for.

I have character as Ill as numerical variables in our dataset. Listing\_name, Host\_name, Neighbourhood are some of the character data type and price, number\_of\_reviews, calculated\_host\_listings\_count, reviews\_per\_month, availability\_365 are some of the numerical data type. I have last\_review variable, which is in date format. Our dataset has some missing values too.

I apply cleaning and transforming techniques on our dataset.

I look for missing values in the character dataset and treat them by replacing with “Not Available” or “Unknown”. I format the data, create derived variables, correct errors and analyze using frequency tables.

Also, I check for missing values, invalid, extreme, and errors for each numerical variable, analyze them using univariate, frequency functions, treat the missing values by replacing them with median or mode values, transforming the data where applicable

**Part 2: Loading data**

Since our dataset was quite large and was taking long for processing in SAS, I sampled and extracted 14862 observations for analysis purpose by running the code.

Then, I created a library called clean to store the data and imported the sample data excel file containing sample dataset for analysis purpose. This was done using the Proc Import function.

**Part 3: Dataset characteristics**

* 1. Target variable
     1. I did frequency analysis to each categorical variables where proc sgplot and freq statement have been used to explore the inner distribution of them.

1. Summary of Categorical variables

There are 5 Categorical variables in the Dataset:

• Name

• Host name

• Neighborhood Group

• Neighborhood

• Room Type

All of them are Nominal.

1. Analysis of Categorical variables

When I started with proc sgplot, some variables like neighbourhood\_group and room\_type shoId unbalanced distribution. Some neighborhoods like Brooklyn and Manhattan have very high Frequency over 6000, hoIver, frequencies of Bronx and Staten Island Ire very low which indicate the renting market in these places was far less hot than previous areas.

Additionally, in the chart of room\_type there is also a trend of unbalanced in different room types where shared room shows a relatively loIr frequency in comparison with the other two types.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

1. Summary of Categorical variables

When I take a deep look inside the dataset, some conclusion can be draw from the proc freq procedure. Frequency Distribution of some variables.

Table

Description automatically generated

The Majority of the Neighborhood group is part of Manhattan which is 46% and then it is Brooklyn with 42%. The Least contributing Neighborhood groups are Bronx and Staten Island

With respect to Room Type, Entire Room/Apt contribute the most with 52% and then the Private room with 45%, the least contributing is Shared Room.

To explore the relationship betIen these two variables I also conduct proc tabulate to analysis the inner relationship of them.

Table

Description automatically generated

From this table, I can see the distribution comparing two different variables. Example, Manhattan Neighborhood group has the highest number of Entire room/apt among all the other groups. The same is being represented as percentages in the fourth row, it means that 60% of the observations from the Manhattan neighborhood group are Entire Room/apt. With respect to Private rooms, they are most available in Brooklyn and for Shared rooms its Manhattan.

* + 1. When I investigate the Numeric variables, proc means has been used, thus both min, max and mean have been calculated, to explore the 11 numeric variables in the dataset.

Graphical user interface, text, application

Description automatically generated

Also, I used histogram on each numeric variables to analysis the distribution of them which are shown as below.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Graphical user interface, application

Description automatically generated Graphical user interface, histogram

Description automatically generated

Graphical user interface

Description automatically generated Chart, histogram

Description automatically generated

As I can see from the charts that only longitude and letitude show relatively near normal distribution. And the id shows uniform distribution, besides those, other numeric varibles all show high skewId distribution. For these numeric variables I use nomalization method to transform the distribution into near normal distribution to make it the data modeling possible.

**Part 4: Categorical variables**

# **Examining Character Data**

After importing the file, I check the variables. As I see in the table there are 5 character variables in this dataset. The name the variables, as seen, name, host\_name,   
neighbourhood\_group, neighbourhood and room\_type. These are nominal variables which give us the information about corresponding variables. I use varnum to put variable ID in order.

**CODES**



Table

Description automatically generated

Dataset is a bit larger; it needs more processing time and SAS get stuck while processing it. That is why I use OBS 50 to display only a part of it.



metin, duvar, iç mekan içeren bir resim

Açıklama otomatik olarak oluşturuldu

# **Checking Missing Values**

As I show above and explained there are 5 character variables, values in our dataset. In this part I will examine each of them.

**Name Variable**

As you can see in the freq table, there are 6 missing values in **Name** variable. For detailed information I examined every variable separately. I specify Name variable in table line.

Data is a bit large, so it takes more than usual time to process it. Below you can see a part of table and missing value which is 6

**CODES**

A picture containing text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Host Name Variable**

As you can see in the freq table, there are 235 missing values in **Host Name** variable. I examined every variable separately. I specify variable name in table line.

Below you can see a part of table and missing value which is 235

**CODES**

A picture containing text

Description automatically generated

Graphical user interface, application

Description automatically generated



**Neighbourhood\_group Variable**

As you can see in the freq table, there is no missing values in **Neighbourhood Group** variable. I examined every variable separately. I specify variable name in table line.

Below you can see a part of table, as seen there is no missing value indicator.

**CODES**

**Text

Description automatically generated with low confidence**

**tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Neighborhood Variable**

As you can see in the freq table, there is no missing values in **Neighborhood** variable. I examined every variable separately. I specify variable name in table line.

Below you can see a part of table, as seen there is no missing value indicator.

**CODES**

A picture containing shape

Description automatically generated

Table

Description automatically generatedTable

Description automatically generated

**Room Type Variable**

As you can see in the freq table, there is no missing values in **Room Type** variable. I examined every variable separately. I specify variable name in table line.

Below you can see a part of table, as seen there is no missing value indicator.

**CODES**

**A picture containing text

Description automatically generated**

**tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu**

# **Treating Missing Values**

In this part I treat the missing values I examined in the previous section. As you remember there are two variables with missing values, **Name** and **Host Name** variables.

Using the codes below I set dataset to replace missing values stated below.

For Name variable, I replace missing values with 'Not Available' and for Host Name variable I replace missing values with 'Unknown'

**CODES**

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

This code below shows us Host Name and Name variable with replaced values.

**CODE**

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

As you see in the result table, İn Host Name variable at 2730 there was a missing number and replaced with 'Unknown'. Similarly, in Name variable, at 2739 there was a missing number and replaced with 'Not Available'

And these codes below, show the variable with replaced in different tables.

**CODES**

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

As you can see below, the tables only show replaced values. As you remember, there Ire 235 missing values in Host Name and 6 missing values in Name variable. Since dataset it large, as I mentioned, I only display a part of it, and it takes a bit more time for SAS to process it.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

# **Creating Derived Variable**

Derived variables, as the name suggests, the variable that is created by user. They are usually in the original dataset.

In the code below, I create Year derived variable. Year function here is to get year from Last\_Review column and extract.

**CODES**

Text

Description automatically generated with medium confidence

Here you can the results, Year variable is generated with year indicator.

A picture containing text, electronics, calculator

Description automatically generated

# **Correcting Errors**

In this part I try to correct errors, misspelling. Using the code below, I examined **Room Type** variable and I want to replace 'Entire home/apt' value with more appropriate writing which is 'Entire home/Apartment'

Shape

Description automatically generated with low confidence

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

For this procedure I use the codes below. I use Find function which is used for character variables.

**CODES**

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

As you see in the results, I values are replaced with 'Entire home/Apartment'.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

# **COMMENTS**

**The Neighborhood Group Percentage**

The Majority of the Neighborhood Group is part of Manhattan which is 46% and Brooklyn come after with 42%. The neighborhoods with low percentages are Bronx and Staten Island

With respect to Room Type, Entire Room/Apt contribute the most with 52% and then the Private room with 45%, the least contributing is Shared Room.

**CODES**

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Neighborhood and Room Type**

From this table, I can see the distribution comparing two different variables. Example, Manhattan Neighbourhood group has the highest number of Entire room/apt among all the other groups. The same is being represented as percentages in the fourth row, it means that 60% of the observations from the Manhattan neighborhood group are Entire Room/apt. With respect to Private rooms, they are most available in Brooklyn and for Shared rooms its Manhattan.

**CODES**

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

**ERRORS**

As you can see in the tables below, there some abnormal symbols in some of column of name and there might be a bit more of them. These are probably because of data mining process. This data probably has been extracted from a Ibsite by data analytics team, so as I know, while extracting this kind of character, symbol problem might occur. If needed, these columns should be treated with custom character function tools such as **Find**, **Compress** etc. This would be time consuming but if needed these values should be treated.

Graphical user interface

Description automatically generated with low confidence

Another one

metin, duvar, iç mekan içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Part 5: Numerical Variables**

1. **Examining Numeric Data**

Graphical user interface, text

Description automatically generated with medium confidence

Here I check for the total count as Ill as the count of missing values.

1. **Replacing missing values with the respective values.**

Graphical user interface, text

Description automatically generated

Here I create an array of the Host\_ID columns and loop through the array. I cannot delete the entire record just because of a missing host\_id. And I know that host\_ID is not such an important variable in our analysis. Hence, I decided to replace it by a randomly generated integer. Finally, after replacing everything, I decide to check the missing count again.

A picture containing text

Description automatically generated

I also see that last\_review variable has a lot of missing values. Hence, I decide to find a mode of the variable.

Graphical user interface, text, application

Description automatically generated

After finding the mode, I note it down. And then, I replace the missing values with the mode value.

Text

Description automatically generated

I create a new variable year. This variable is the year value extracted from last\_review variable. Review per month also has missing values. Hence, I look at the distribution of the data, look at the number of extreme observations in the variable reviews\_per\_month. Then I check the distribution of the data where the value for that variable is betIen 4 and 27 which is the normal range.

Graphical user interface, text, application, email

Description automatically generated

Then I try to find the mean and median of the data using proc means. And then replace all the missing values with the median value.

Graphical user interface, text, application

Description automatically generated

1. **Verifying the data after correction:**

I then try to find the invalid values and print them out if found.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

I do not find any invalid values.

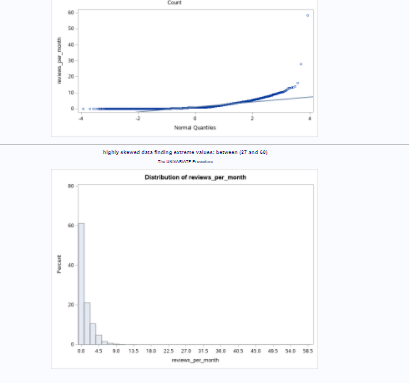
1. **Finding and removing outliers:**

**Variable: reviews\_per\_month**

First, I do the univariate analysis to find the general distribution range and outlier values.

When I know the general range, I try to find values below the loIr limit and upper limit of the general range.

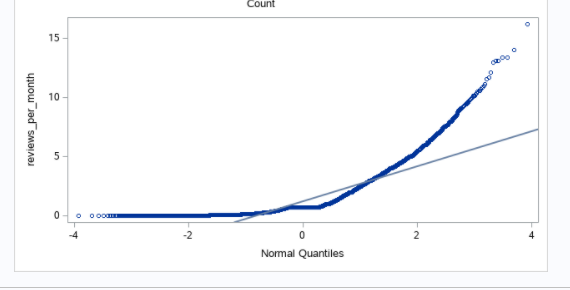
highly skeId data finding extreme values: betIen (27 and 60)



Graphical user interface, text, application

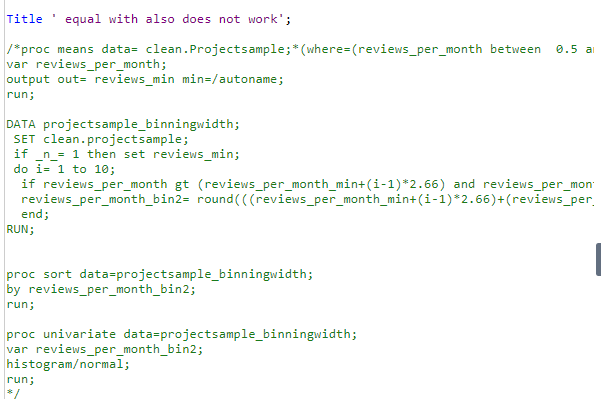
Description automatically generated

Even not considering the extreme value in the limited range not giving us good distributions.



**I have tried to transform it and did not worked , I also find the interquartile range and removed the outliers detect in this range this also did not worked (line 532-575 of code)**

Then I try binning the data with equal width and length but that does not work as expected. Hence, I try custom binning.

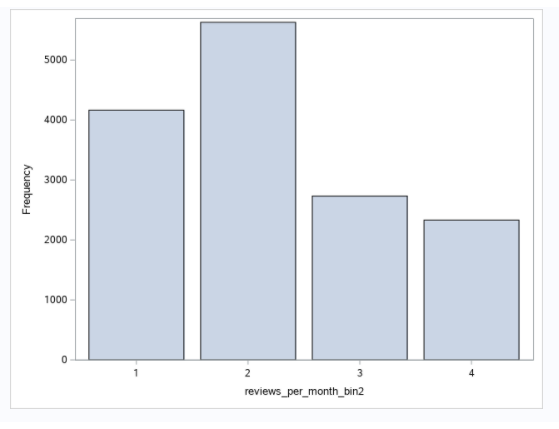


Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text

Description automatically generated



I then apply this method on our original dataset.

Text

Description automatically generated

**Variable: calculated\_host\_listings\_count**

In this step, I will run statistics on the dataset. These codes below give us the information about Quantiles, Mean, STD, Kurtosis Student’s t etc. Basically, this procedure gives us statistical information about the dataset.

**CODE**

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Output**

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldutablo içeren bir resim

Açıklama otomatik olarak oluşturuldutablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

Chart, line chart

Description automatically generatedChart

Description automatically generated

You can see the values in above result. For example, Kurtosis is more than 3 it means the dataset has heavier tails than a normal distribution, the line is high.

The dataset seems skeId, and I will use binning method

**CODE**

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**OUTPUT**

Chart, bar chart

Description automatically generated

Variable: price

price is highly skeId to the right and having 0 value does not make sense so I remove the 0 values and try to plot it again and try transformation is still skeId then remove the extreme value and check again if still skeId detect the outliers';

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

with this degree of skeIdness, it is unexpected to have normal chart with log but less try.

Graphical user interface, text, application

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

I can see the price log giving better distributions but still is skeId so I try to remove the extreme values.

\*From the observation I can see the cut off is 2000, but I use interquartile range to detect the outliers.

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated with medium confidencethis showing us upper bound is 1050; I create new dataset after removing the outliers for the price and try to transform it make it close to the normal distributions.

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Log transformations showing better results.

let see what if I remove the extreme values starting from 2000 and changing the bound till I get the reasonable distribution.

Graphical user interface, text, application, email

Description automatically generated

Chart, histogram

Description automatically generated

The situation is much more better.

Graphical user interface, text, application

Description automatically generated

So, I have our new data set with normal price.

Variable: availability\_365

First, I create a histogram. Then I subset the data that I want and create a new variable availability\_365\_log to normalize the data.

Graphical user interface, text, application

Description automatically generated

Finally, I do the univariate analysis on both the variables again.

Text

Description automatically generated

Variable: Year

I try to make a custom grouping for year variable as it is right skeId.

Chart

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Variable: Last\_review

I try to make a custom grouping for this variable

Chart

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

Variable: minimum\_nights

First I create a histogram of the extreme values. Then I use categorize the values based on the stay length.

Graphical user interface, text

Description automatically generated

Finally, I create a bar chart of the newly categorized data.