OPIM 5604

Predictive Modeling Preprocessing Project 1

Team 3: Western Australia

(Team Members: Neehar Namjoshi, Nishita Rao, Pooja Shah, Shivam Kumar)

Table of Contents

1. Column Summary

Project 1 was started by undertaking an overview of all the columns which were available in the "Listings" JMP file for Western Australia. There was a total of 75 columns and 10,485 rows.

Upon going through the data set, the columns below summarize the project's initial findings. In accordance with the original data set, the table outlines:

(1) Column number

(2) Column name

(3) Activity-

- Retain the column.
- Hide and exclude the column.
- Modify the column, where a new column is created, and the original column is hidden and excluded

(4) Column description and rationale for the activity.

Table 1: Column Summary

Col No.	Column name	Activity	Column Description and Rationale
1	ID	Retained	This column indicates the ID of the individual property or Airbnb listing. It is a unique key and differentiates every single row from the others. Therefore the ID column is retained.
2	Listing_url	Hidden and excluded	This column indicates the url of the listing specified by the ID. There is no influence of listing url and hence is a non-useful data set to be considered as a predictor variable for the target. Hence it has been hidden and excluded from the analysis.
3	Scrape_id	Hidden and excluded	This column indicates the ID for when the scraping happened. Given there is no time reference to be put to the data, there is no impact of the scraping on the target variable; scrape ID will not influence the target and has been hidden and excluded.
4	Last_scraped	Hidden and excluded	This column indicates the date the data was last scrapped on. As the time references on the data set as mentioned above are not required, there is no need for the column and hence has been hidden and excluded from the analysis.

Col No.	Column name	Activity	Column Description and Rationale
5	Source	Hidden and excluded	This column indicates the source for scraping the data. There is no influence on the target variable, based on the source of scrapping, hence the column has been hidden and excluded.
6	Name	Hidden and excluded	This column describes the listing in slight detail and is a non - useful data set as a predictor variable to determine prices of properties as there is no influence this description can have on figuring property prices.
7	Description	Hidden and excluded	This column describes in greater detail about the property listing. It indicates the location, nearby locality details to figure out the place and can be used to trace the listing in question. However, it is of no use to find the prices of the listing and hence is being hidden and excluded from the analysis.
8	Neighborhood _overview	Hidden and excluded	This column describes the neighborhood where the property is. Better neighborhoods can help charge better prices for the property. The information from column "neighborhood cleansed" will be a better determinant with clear categories of the neighborhood. Hence, this column has been excluded and hidden from the analysis.
9	Picture_url	Hidden and excluded	This column describes the url of the picture of the property in question. As the prices cannot be determined by picture url and does not serve as a predictor, therefore it is excluded from the data set.
10	Host_id	Retained	This column indicates a unique ID for every single host. The column also has the same hosts for different properties. This predictor variable can help in the analysis of prices, if a particular host charges higher prices on average. Thus, it can help in modeling the target variable. Therefore, the column has been retained.
11	Host_url	Hidden and excluded	This column describes the url of the host. As url cannot help in building the model to determine prices, it needs to be dropped as it only gives the url of the host and is immaterial to the target variable.

Col No.	Column name	Activity	Column Description and Rationale
12	Host_name	Hidden and excluded	This column indicates the name of the host. As for distinction, the host ID is used already, therefore the host name is hidden and excluded to reduce the complexity.
13	host_since	Hidden and excluded	This column indicates the time since the host has been a “host” and hence has no influence on the price of the property. Hence, it is a non useful data column and is being hidden and excluded from the analysis.
14	host_location	Hidden and excluded	This column indicates the location of the host of the property in question. For determining the price, it does not matter which part of the world the host is located in and hence is being dropped from the data set.
15	host_about	Hidden and excluded	This column indicates a bit about the host. The host describes who they are and what they like. It gives no details about the property and does not help in determining the target variable. Hence, it is being dropped from the data set.
16	host_response_time	Modified	This column describes the time it takes for the host to respond to a customer request. It is categorized into different time references - under an hour, in a day, etc and can be a good data point to build user ratings and determine prices. Hence, it is being retained in the data set.
17	host_response_rate	Retained	This column describes the response rate of the host, if the host has been able to respond back to the customer 100% or any lesser. This can be a good predictor variable to understand how good the services of a particular listing are and succeedingly can help in determining prices. Therefore, it is being retained for analysis.
18	host_acceptance_rate	Hidden and excluded	This column indicates the acceptance rate of the customer requests and is at 100% or lesser. The acceptance rate if higher or lower cannot have an influence on the price a particular property may charge as it purely depends on the traffic inflow. Acceptance might be low during peak traffic and still the property might charge low or high based on their service. Hence, the column has been dropped from the analysis.

Col No.	Column name	Activity	Column Description and Rationale
19	host_is_superhost	Hidden and excluded	This column was hidden and excluded as it is a formula used by Airbnb to determine this. It seems to be a weighted average of the previous two columns, host_response_time and host_response_rate.
20	host_thumbnail_url	Hidden and excluded	This column provides an image of the host's profile picture on Airbnb. However, this column was hidden and excluded as it does not provide additional information or insight in determining the target variable, Price.
21	host_picture_url	Hidden and excluded	This column, similar to host_thumbnail_url, provides an image of the host's profile picture on Airbnb. However, this column was also hidden and excluded as it does not provide additional information or insight in determining the target variable, Price.
22	host_neigbourhood	Hidden and excluded	This column provides the host's neighborhood. It was hidden and excluded as more than 95% of the values are missing. Furthermore, this is not a useful column in predicting the target variable.
23	host_listings_count	Hidden and excluded	This column provides the number of host listings on the Airbnb website. However, this column was hidden and excluded as it does not provide additional information or insight in determining the target variable.
24	host_total_listings_count	Hidden and excluded	Similar to the column, host_listings_count, this column provides the total number of host listings on the Airbnb website. However, this column was also hidden and excluded as it does not provide additional information or insight in determining the target variable.
25	host_verifications	Hidden and excluded	This column provides the method used to verify the host, combinations of email and phone are seen in this column. However, this column was hidden and excluded as it contains redundant information that will be seen in column host_identity_verified. It is not imperative to know what method was used to verify the host as long as the host is verified and thereby legitimate.

Col No.	Column name	Activity	Column Description and Rationale
26	host_has_profile_pic	Hidden and excluded	This column provides if the host has a profile picture or not on the Airbnb website. This column was hidden and excluded as it does not provide additional information or insight in determining the target variable.
27	host_identity_verified	Modified	This column provides information regarding if the host is verified or not. This column was modified using dummy variables/indicator columns. The original column consists of true or false values, which is two dimensional. Thereby, a new column Host_Identity_Verified_True, is created that consists of 1s or 0s, that depicts if the host is verified or not. Accordingly, if a 0 is seen in the new modified column, then the host is not verified.
28	neighborhood	Hidden and excluded	This column provides the neighborhood that the property resides in. It was hidden and excluded as over 30% of the values are missing. Moreover, the next column, neighborhood_cleansed, depicts the same information in a more concise manner. This column also consists of redundant information as well.
29	neighbourhood_cleansed	Retained	This column provides the neighborhood the property is located in. It has no missing values and is retained. It gives us information regarding the neighborhood and location, which is the most important factor in determining real estate prices. This is a very important predictor variable in determining the target variable.
30	neighbourhood_group_cleansed	Hidden and excluded	This column is attempting to group neighborhoods together. However, it only consists of missing values. Therefore, it is a column that was hidden and excluded as there is no information that can be extrapolated.
31	Latitude	Hidden and excluded	This column provides information regarding the latitude of the property using WGS84. This column was hidden and excluded, as it is redundant information and not required to predict the target variable.

Col No.	Column name	Activity	Column Description and Rationale
32	Longitude	Hidden and excluded	This column provides information regarding the longitude of the property using WGS84. This column was hidden and excluded, as it is redundant information and not required to predict the target variable.
33	property_type	Hidden and excluded	This column provides information about the property itself. It provides information regarding what the host is trying to rent, e.g. a private room, the entire home/apartment, etc. While it is very important in determining price, the column is hidden and excluded as column room_type provides a more concise representation of the same information.
34	room_type	Retained	This column provides information regarding the host's intention on how they want to rent out their property. It consists of 4 values: shared room, private room, hotel room, and entire home/apartment. This is a very important column in determining the target variable, as it generally gives us insight regarding the size of the property being rented out. Therefore, this column is retained.
35	accommodates	Retained	This column provides information regarding the number of people that the host would accept for the property. This column is also retained, as it generally gives us insight into the size of the property, in combination with the column room_type. This will be a very important variable in predicting the target variable.
36	bathrooms	Hidden and excluded	This column provides information regarding the number of bathrooms in the property. Unfortunately, while important, most of the data was missing, so it had to be hidden and excluded. However, column *bathroom_texts* captures this information in a more concise manner, making the column bathrooms redundant as well.

Col No.	Column name	Activity	Column Description and Rationale
37	bathroom_texts	Modified	This column describes the total number of bathrooms in string. Certain rows indicate whether bathrooms are shared or private. This column has been modified and re-coded into a numeric column 'Total Bathrooms', which indicates the total number of bathrooms in the property. The information regarding the type of bathroom (private, shared) was available only for less than 10% of the data and hence was dropped.
38	bedrooms	Retained	This integer type column describes the total number of bedrooms in the listed property. It has been retained given its relation to the target variable. Higher bedrooms would generally indicate higher prices.
39	beds	Retained	This integer type column describes the total number of beds available in the listed property. It has been retained given its relation to the target variable price. Higher number of beds would generally indicate higher prices.
40	amenities	Hidden and excluded	This column had indecipherable data. Given any treatment, it was not possible to incorporate the data into the data set. Therefore, the column was excluded and hidden.
41	price	Retained	This column shows the daily price in Australian Dollars. This serves as the target variable of the data set and has been retained.
42	minimum_nights	Modified	This column describes the minimum number of nights required to accept a booking as listed by the host. Given its importance in determining the price for the listing, this column was retained.
43	maximum_nights	Retained	This column describes the maximum number of nights allowed for a customer to request in a booking as listed by the host. Given its importance in determining the price for the listing, this column was retained.

Col No.	Column name	Activity	Column Description and Rationale
44	minimum_minimum_nights	Hidden and excluded	This is a calculated column showing the smallest value of minimum nights. Due to the ambiguity of the calculation involved in deriving the column, it was hidden and excluded. Also, the column does not have an impact in determining the target variable given the columns: minimum and maximum nights.
45	maximum_minimum_nights	Hidden and excluded	This is a calculated column showing the largest value of minimum nights. Due to the ambiguity of the calculation involved in deriving the column, it was hidden and excluded. Also, the column does not have an impact in determining the target variable given the columns: minimum and maximum nights.
46	minimum_maximum_nights	Hidden and excluded	This is a calculated column showing the smallest value of maximum nights. Due to the ambiguity of the calculation involved in deriving the column, it was hidden and excluded. Also, the column does not have an impact in determining the target variable given the columns: minimum and maximum nights.
47	maximum_maximum_nights	Hidden and excluded	This is a calculated column showing the largest value of maximum nights. Due to the ambiguity of the calculation involved in deriving the column, it was hidden and excluded. Also, the column does not have an impact in determining the target variable given the columns: minimum and maximum nights.
48	minimum_nights_avg_ntm	Hidden and excluded	This calculated column shows the average number of minimum nights. As its inclusion in the data set will lead to data redundancy with the inclusion of the original column 'minimum_nights', this column was hidden and excluded.
49	maximum_nights_avg_ntm	Hidden and excluded	This calculated column shows the average number of maximum nights. As its inclusion in the data set will lead to data redundancy with the inclusion of the original column 'maximum_nights', this column was hidden and excluded.
50	calendar_updated	Hidden and excluded	This column was blank and fully comprised of missing values. Therefore, it was hidden and excluded.

Col No.	Column name	Activity	Column Description and Rationale
51	has_availability	Hidden and excluded	This column indicates TRUE/FALSE values to show whether the listing has availability. Given the similarity in data points, this column was hidden and excluded. More than 99% of data indicates the same response - TRUE and hence has no additional insight from the column.
52	availability_30	Hidden and excluded	This column indicated the availability of each listing 30 days in the future from the date the data was last scrapped. The column was hidden and excluded as a reference on the occupancy of the property and has information overlap with column availability_365 which is being retained.
53	availability_60	Hidden and excluded	This column indicated the availability of each listing 60 days in the future from the date that the data was last scrapped. The column was hidden and excluded as a reference on the occupancy of the property and has information overlap with column availability_365 which is being retained.
54	availability_90	Hidden and excluded	This column indicated the availability of each listing 90 days in the future from the date that the data was last scrapped. The column was hidden and excluded as a reference on the occupancy of the property and has information overlap with column availability_365 which is being retained.
55	availability_365	Retained	This column indicates the availability of each listing 365 days in the future from the date the data was last scrapped. A property's high occupancy rate indicates that the rates will be high. Hence, it will provide a more accurate estimate of the occupancy and will influence the price of the property. The column is retained given its relationship with the target variable.
56	calendar_last_scrapped	Retained	The calendar_last_scrapped column will serve as a baseline date to check the availability of each listing in the future. This column will be used by the availability_365 column to check the availability of each listing 365 days in the future.Therefore, this column is retained.

Col No.	Column name	Activity	Column Description and Rationale
57	number_of_reviews	Modified	This column indicates the number of reviews a listing has and this variable will impact the price of the property. The column is modified by using SHASH transformation to normalize the distribution. ;The new column is called SHASH Transformed to Normal number_of_reviews.
58	number_of_reviews_ltm	Hidden and Excluded	This column indicates the number of reviews the listing has had in the last 12 months. The column is a calculated and redundant column. The number_of_reviews column already contains all the reviews a property has. Therefore, it was hidden and excluded.
59	number_of_reviews_l30d	Hidden and Excluded	This column indicates the number of reviews the listing has had in the last 30 days.The column is a calculated and redundant column. The number_of_reviews column already contains all the reviews a property has. Therefore, it was hidden and excluded.
60	first_review	Hidden and Excluded	This column indicates the date of the first/oldest review. The first_review column does not have any impact on the price of a listing; so it was hidden and excluded.
61	last_review	Hidden and Excluded	This column indicates the date of the last/newest review. The first_review column does not have any impact on the price of a listing; so it was hidden and excluded.
62	review_score_rating	Modified	This column indicates the review score rating. There are 7 columns in the table which have different types of review data and to reduce the dimensionality, PCA was applied on those 7 columns. After reviewing the PCA, most of the data i.e 91.29% was present within the first 4 PCA's.The new columns formed in the table are Review_Scores_PCA1, Review_Scores_PCA2, Review_Scores_PCA3 and Review_Scores_PCA4. Therefore, the review_score_rating is hidden and excluded.

Col No.	Column name	Activity	Column Description and Rationale
63	review_score_accuracy	Modified	This column indicates the review score accuracy. There are 7 columns in the table which have different types of review data and to reduce the dimensionality, PCA was applied on those 7 columns. After reviewing the PCA, most of the data i.e 91.29% was present within the first 4 PCA's.The new columns formed in the table are Review_Scores_PCA1, Review_Scores_PCA2, Review_Scores_PCA3 and Review_Scores_PCA4. Therefore, the review_score_accuracy is hidden and excluded.
64	review_score_cleanliness	Modified	This column indicates the review score cleanliness. There are 7 columns in the table which have different types of review data and to reduce the dimensionality, PCA was applied on those 7 columns. After reviewing the PCA, most of the data i.e 91.29% was present within the first 4 PCA's.The new columns formed in the table are Review_Scores_PCA1, Review_Scores_PCA2, Review_Scores_PCA3 and Review_Scores_PCA4. Therefore, the review_score_cleanliness is hidden and excluded.
65	review_score_checkin	Modified	This column indicates the review score check in. There are 7 columns in the table which have different types of review data and to reduce the dimensionality, PCA was applied on those 7 columns. After reviewing the PCA, most of the data i.e 91.29% was present within the first 4 PCA's.The new columns formed in the table are Review_Scores_PCA1, Review_Scores_PCA2, Review_Scores_PCA3 and Review_Scores_PCA4.Therefore, the review_score_checkin is hidden and excluded.
66	review_score_communication	Modified	This column indicates the review score communication. There are 7 columns in the table which have different types of review data and to reduce the dimensionality, PCA was applied on those 7 columns. After reviewing the PCA, most of the data i.e 91.29% was present within the first 4 PCA's.The new columns formed in the table are Review_Scores_PCA1, Review_Scores_PCA2, Review_Scores_PCA3 and Review_Scores_PCA4. Therefore the review_score_communication is hidden and excluded.

Col No.	Column name	Activity	Column Description and Rationale
67	review_score_location	Modified	This column indicates the review score location. There are 7 columns in the table which have different types of review data and to reduce the dimensionality, PCA was applied on those 7 columns. After reviewing the PCA, most of the data i.e 91.29% was present within the first 4 PCA's.The new columns formed in the table are Review_Scores_PCA1, Review_Scores_PCA2, Review_Scores_PCA3 and Review_Scores_PCA4. Therefore, the review_score_location is hidden and excluded.
68	review_score_value	Modified	This column indicates the review score value. There are 7 columns in the table which have different types of review data and to reduce the dimensionality, PCA was applied on those 7 columns. After reviewing the PCA, most of the data i.e 91.29% was present within the first 4 PCA's.The new columns formed in the table are Review_Scores_PCA1, Review_Scores_PCA2, Review_Scores_PCA3 and Review_Scores_PCA4. Therefore, the review_score_value is hidden and excluded.
69	license	Hidden and Excluded	This column indicates the license/permit/registration number. The license column is hidden and excluded as most of the data was missing in the column.
70	Instant_bookable	Hidden and Excluded	This column indicates whether the guest can automatically book the listing without the host accepting their booking request. The target variable is not impacted by whether a reservation can be made automatically or not. So, the instant_bookable column has no impact on the target variable. Therefore, it is hidden and excluded.
71	Calculated_host_listings_count	Hidden and Excluded	This column indicates the number of listings that the host has in the current scrape based across city/region geography. The target variable is not impacted by the number of properties a host has listed. Therefore, this column is hidden and excluded.
72	Calculated_host_listings_count_entire_homes	Hidden and Excluded	This column indicates the number of entire home/apartment listings that the host has in the current scrape based across the city/region geography. The target variable of a listing is not impacted by the number of home/apartment that a host has listed. Therefore, this column is hidden and excluded.

Col No.	Column name	Activity	Column Description and Rationale
73	Calculated_host_listings_count_private_rooms	Hidden and Excluded	This column indicates the number of private room listings the host has in the current scrape, in the city/region geography. The target variable is not impacted by the number of private rooms a host has listed. Therefore, this column is hidden and excluded.
74	Calculated_host_listings_count_shared_rooms	Hidden and Excluded	This column indicates the number of shared room listings that the host has in the current scrape, across the city/region geography. The target variable will not be impacted by the number of shared rooms a host has listed. Therefore, this column is hidden and excluded.
75	Reviews_per_month	Hidden and Excluded	This column indicates the number of reviews that a listing has per month. The data present in this column is redundant and therefore, it is hidden and excluded.

2. Data Preprocessing:

In order to build a model, dataset 'Listings' was pre-processed using a range of data preprocessing techniques to reduce the complexity of the dataset and clean the data set. The following steps and techniques were used:

2.1 Hide and Exclude
2.2 Binning, Indicator Columns, and Re-code
2.3 Missing Values
2.4 Outliers
2.5 Reducing dimensionality

The detailed treatment of the data set has been outlined in the below sections along with appropriate screenshots to support the process.

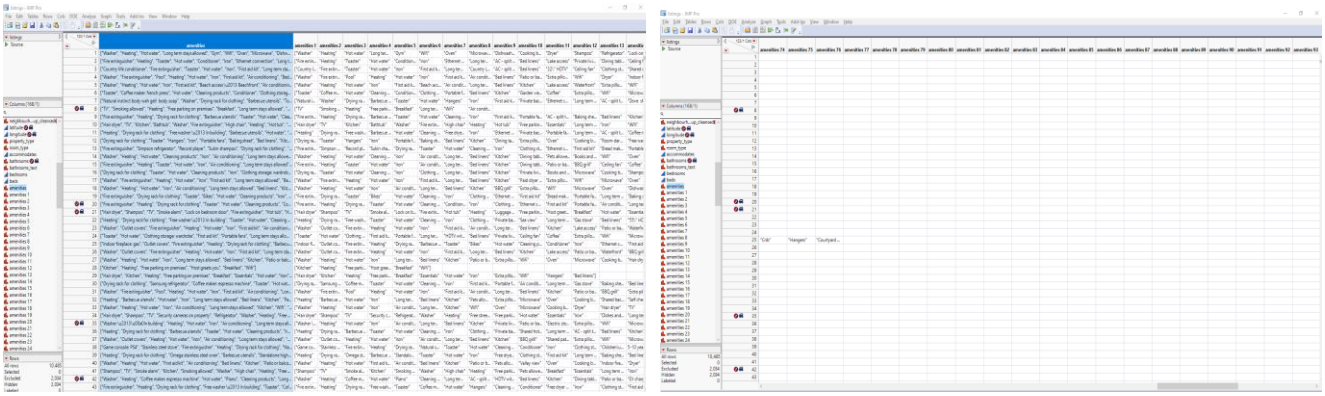
2.1 Hide and exclude

All columns were reviewed as mentioned in Section 3 Column Summary to ensure well-defined and appropriate columns were identified and included in the data set. All columns which were found to not be relevant or would not help with the prediction of the target variable were hidden and excluded.

A total of 42 columns were hidden and excluded from the data set. The remaining 33 columns were then further worked upon in the next section.

Figure 1 shows an example of column **'Amenities'** which included non - decipherable data sets when converted into columns for data processing. As a result, this column was hidden and excluded from the data set.

Figure 1: Screenshot of column Amenities



2.2 Binning, Indicator Columns, and Re-code

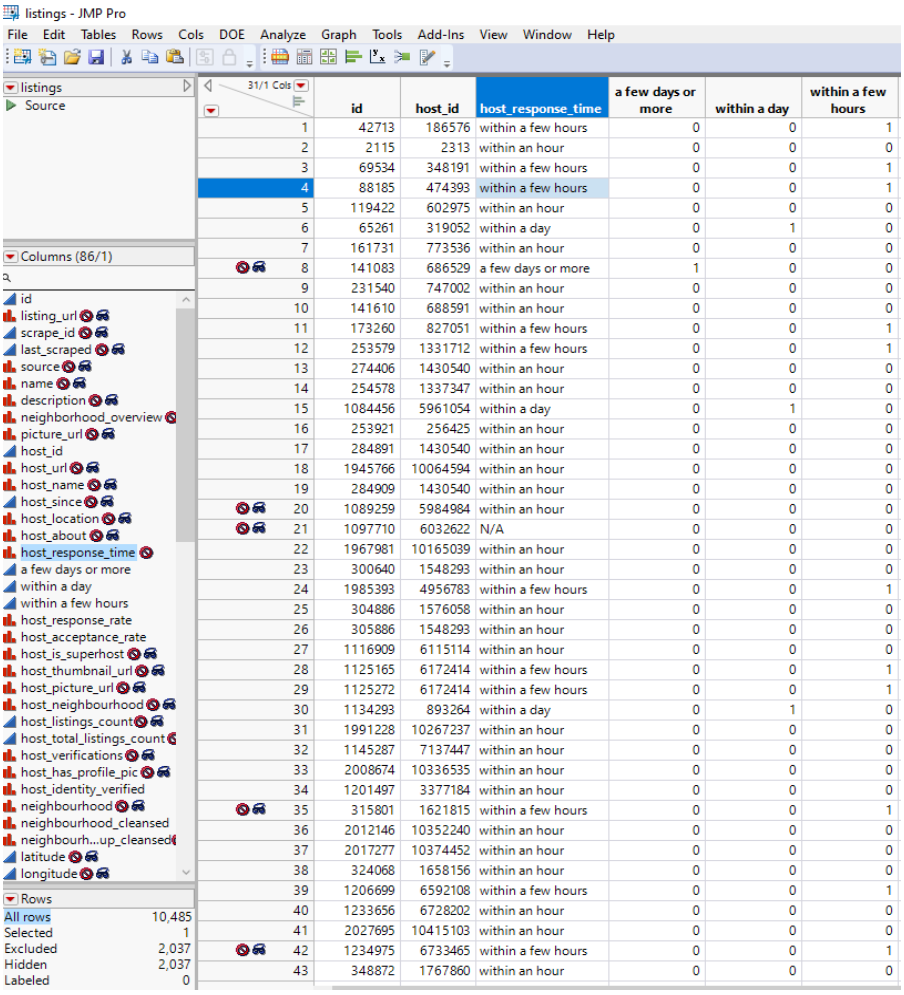
A total of 3 columns were assessed and treated in this section. These are:

(1) **host_response_time**

(2) **host_identity_verified**

(3) **bathroom_text**

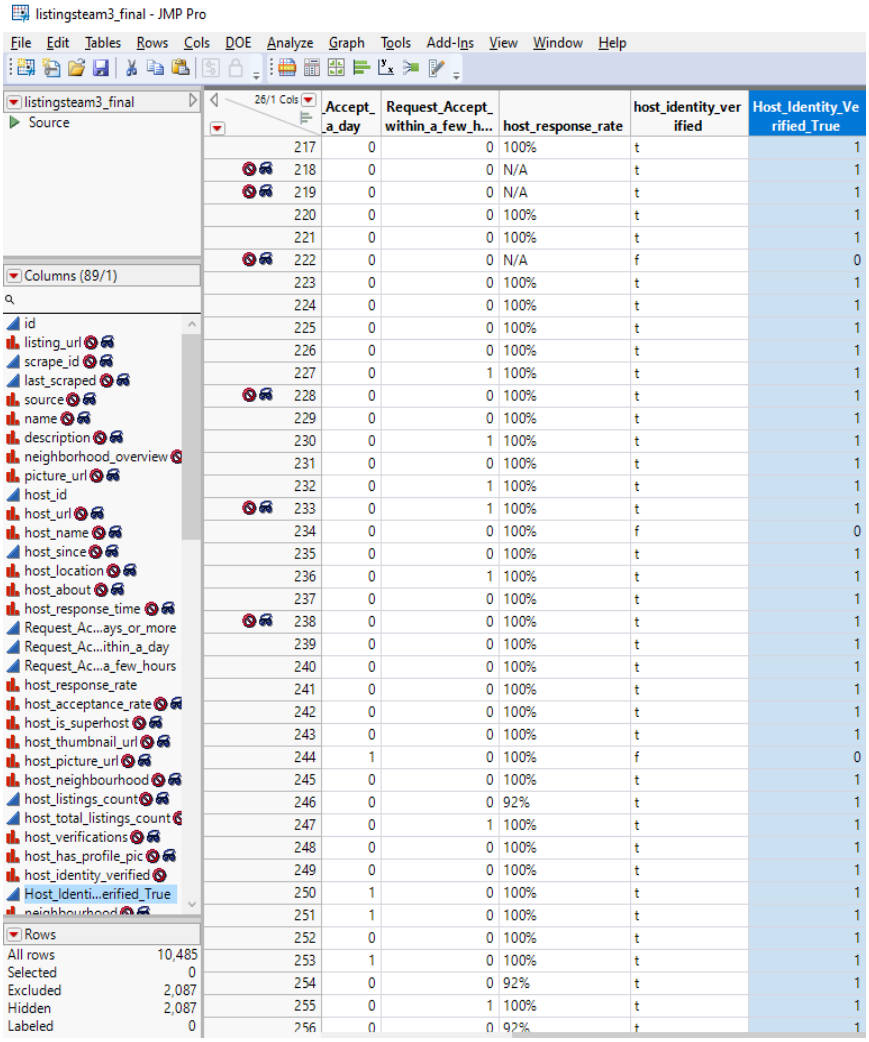
Figure 2: Screenshot of indicator columns breaking down host_response_time



	id	host_id	host_response_time	a few days or more	within a day	within a few hours
1	42713	186576	within a few hours	0	0	1
2	2115	2313	within an hour	0	0	0
3	69534	348191	within a few hours	0	0	1
4	88185	474393	within a few hours	0	0	1
5	119422	602975	within an hour	0	0	0
6	65261	319052	within a day	0	1	0
7	161731	773536	within an hour	0	0	0
8	141083	686529	a few days or more	1	0	0
9	231540	747002	within an hour	0	0	0
10	141610	688591	within an hour	0	0	0
11	173260	827051	within a few hours	0	0	1
12	253579	1331712	within a few hours	0	0	1
13	274406	1430540	within an hour	0	0	0
14	254578	1337347	within an hour	0	0	0
15	1084456	5961054	within a day	0	1	0
16	253921	256425	within an hour	0	0	0
17	284891	1430540	within an hour	0	0	0
18	1945766	10064594	within an hour	0	0	0
19	284909	1430540	within an hour	0	0	0
20	1089259	5984984	within an hour	0	0	0
21	1097710	6032622	N/A	0	0	0
22	1967981	10165039	within an hour	0	0	0
23	300640	1548293	within an hour	0	0	0
24	1985393	4956783	within a few hours	0	0	1
25	304886	1576058	within an hour	0	0	0
26	305886	1548293	within an hour	0	0	0
27	1116909	6115114	within an hour	0	0	0
28	1125165	6172414	within a few hours	0	0	1
29	1125272	6172414	within a few hours	0	0	1
30	1134293	893264	within a day	0	1	0
31	1991228	10267237	within an hour	0	0	0
32	1145287	7137447	within an hour	0	0	0
33	2008674	10336535	within an hour	0	0	0
34	1201497	3377184	within an hour	0	0	0
35	315801	1621815	within a few hours	0	0	1
36	2012146	10352240	within an hour	0	0	0
37	2017277	10374452	within an hour	0	0	0
38	324068	1658156	within an hour	0	0	0
39	1206699	6592108	within a few hours	0	0	1
40	1233656	6728202	within an hour	0	0	0
41	2027695	10415103	within an hour	0	0	0
42	1234975	6733465	within a few hours	0	0	1
43	348872	1767860	within an hour	0	0	0

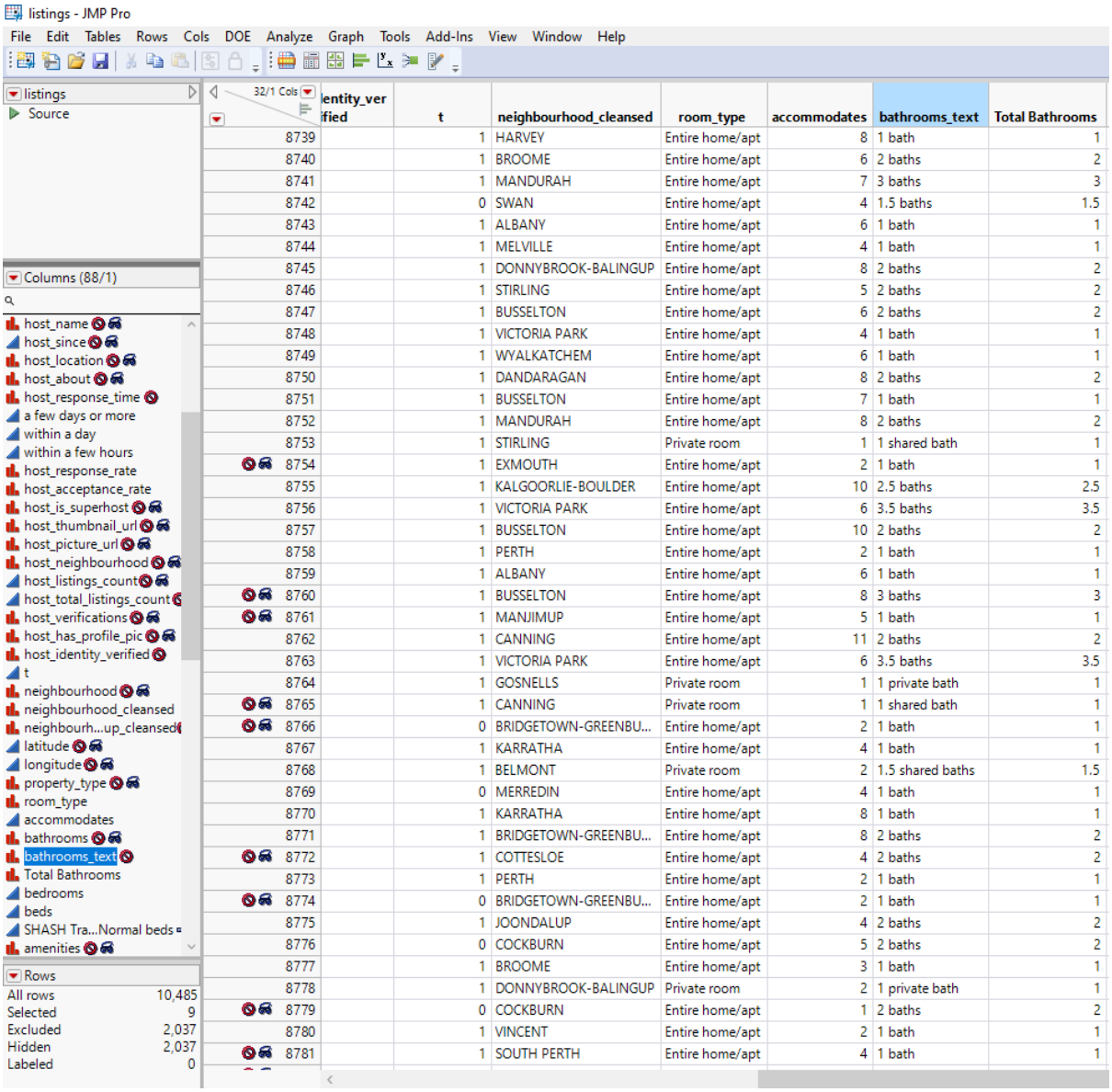
Here, the categorical variable, **Host_response_time,** is converted into a numeric variable by creating indicator columns. There was a total of 4 categories and by creating indicator columns, the columns were reduced to 3.

Figure 3: Screenshot of creating indicator columns for Host_identity_verified



For the categorical variable, **host_identity_verified**, a new indicator column was created called **Host_Identity_Verified_True**. The original column had a dimensionality of 2. Therefore, by creating a new indicator column, the total number of columns is reduced to 1, while retaining all the information.

Figure 4: Screenshot of re-coding bathroom_text and conversion into numerical column



	lentity_ver ified	t	neighbourhood_cleansed	room_type	accommodates	bathrooms_text	Total Bathrooms
8739		1	HARVEY	Entire home/apt	8	1 bath	1
8740		1	BROOME	Entire home/apt	6	2 baths	2
8741		1	MANDURAH	Entire home/apt	7	3 baths	3
8742		0	SWAN	Entire home/apt	4	1.5 baths	1.5
8743		1	ALBANY	Entire home/apt	6	1 bath	1
8744		1	MELVILLE	Entire home/apt	4	1 bath	1
8745		1	DONNYBROOK-BALINGUP	Entire home/apt	8	2 baths	2
8746		1	STIRLING	Entire home/apt	5	2 baths	2
8747		1	BUSSELTON	Entire home/apt	6	2 baths	2
8748		1	VICTORIA PARK	Entire home/apt	4	1 bath	1
8749		1	WYALKATCHEM	Entire home/apt	6	1 bath	1
8750		1	DANDARAGAN	Entire home/apt	8	2 baths	2
8751		1	BUSSELTON	Entire home/apt	7	1 bath	1
8752		1	MANDURAH	Entire home/apt	8	2 baths	2
8753		1	STIRLING	Private room	1	1 shared bath	1
8754		1	EXMOUTH	Entire home/apt	2	1 bath	1
8755		1	KALGOORLIE-BOULDER	Entire home/apt	10	2.5 baths	2.5
8756		1	VICTORIA PARK	Entire home/apt	6	3.5 baths	3.5
8757		1	BUSSELTON	Entire home/apt	10	2 baths	2
8758		1	PERTH	Entire home/apt	2	1 bath	1
8759		1	ALBANY	Entire home/apt	6	1 bath	1
8760		1	BUSSELTON	Entire home/apt	8	3 baths	3
8761		1	MANJIMUP	Entire home/apt	5	1 bath	1
8762		1	CANNING	Entire home/apt	11	2 baths	2
8763		1	VICTORIA PARK	Entire home/apt	6	3.5 baths	3.5
8764		1	GOSNELLS	Private room	1	1 private bath	1
8765		1	CANNING	Private room	1	1 shared bath	1
8766		0	BRIDGETOWN-GREENBU...	Entire home/apt	2	1 bath	1
8767		1	KARRATHA	Entire home/apt	4	1 bath	1
8768		1	BELMONT	Private room	2	1.5 shared baths	1.5
8769		0	MERREDIN	Entire home/apt	4	1 bath	1
8770		1	KARRATHA	Entire home/apt	8	1 bath	1
8771		1	BRIDGETOWN-GREENBU...	Entire home/apt	8	2 baths	2
8772		1	COTTESLOE	Entire home/apt	4	2 baths	2
8773		1	PERTH	Entire home/apt	2	1 bath	1
8774		0	BRIDGETOWN-GREENBU...	Entire home/apt	2	1 bath	1
8775		1	JOONDALUP	Entire home/apt	4	2 baths	2
8776		0	COCKBURN	Entire home/apt	5	2 baths	2
8777		1	BROOME	Entire home/apt	3	1 bath	1
8778		1	DONNYBROOK-BALINGUP	Private room	2	1 private bath	1
8779		0	COCKBURN	Entire home/apt	1	2 baths	2
8780		1	VINCENT	Entire home/apt	2	1 bath	1
8781		1	SOUTH PERTH	Entire home/apt	4	1 bath	1

Finally, in Figure 4, to break down the content of the **bathroom_text** column, a re-coding of the data was selected. A new column was created which was kept numeric in nature. Upon re-coding the content, this column was called **Total Bathrooms**. It reflects the total number of bathrooms available on the listed property.

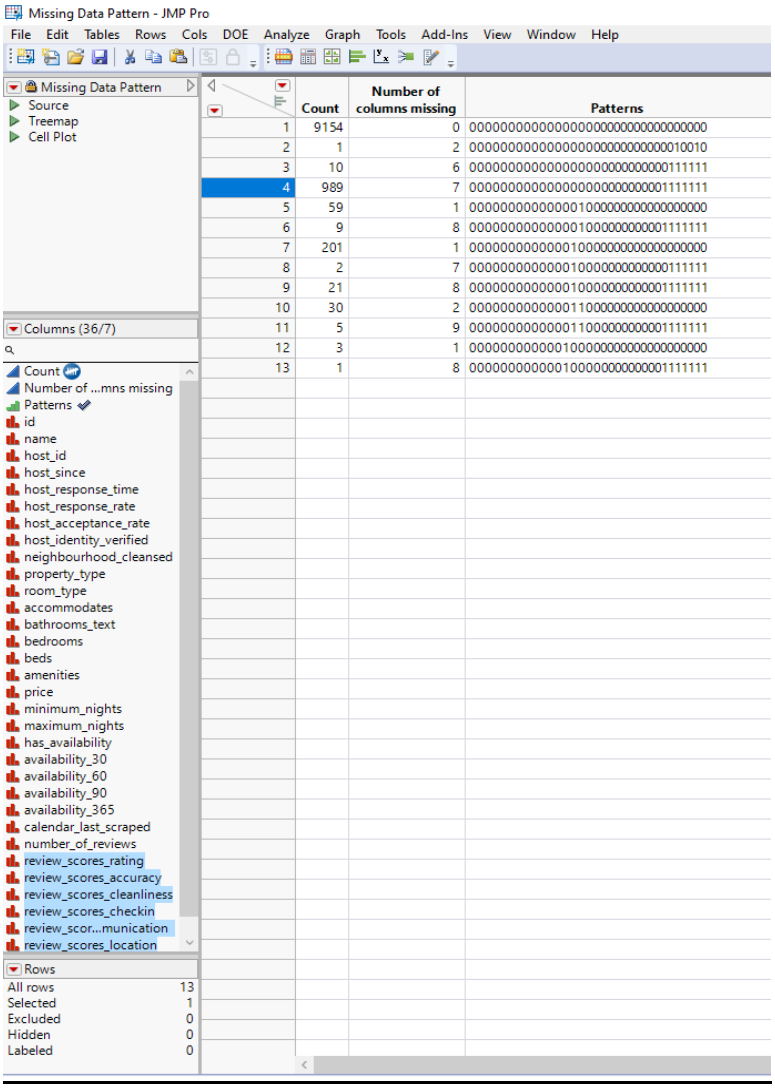
With this treatment, the most essential information about bathrooms was retained i.e. the number of bathrooms. Partial information such as whether the bathroom is shared or private was excluded given that less than 10% of the data provided such information.

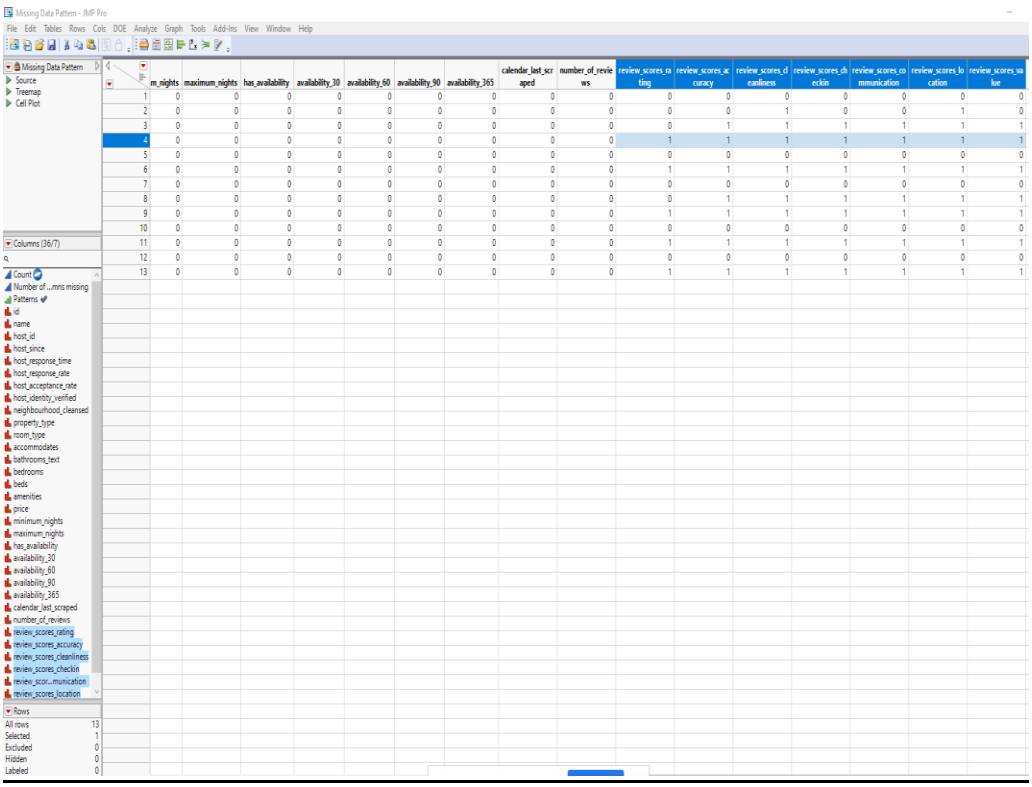
2.3 Missing Values

Hidden and exclusion

In order to further fix abnormalities in the data set, missing values in each column were identified and explored using the '*Missing Data Pattern*' function in JMP. Figure 5 shows the missing data pattern of the data.

Figure 5: Screenshot of identified missing value patterns





As shown in the above figure, 1331 rows were hidden and excluded from the data set. These are all the rows that are identified to have 1 or more missing values in each row. The remaining 9154 rows were used for the onward pre-processing of the dataset.

Some of the major columns with missing values have been outlined below:

- **Review scores for different parameters (e.g., accuracy, cleanliness, etc.)** could not be imputed as the missing values for each score appeared to be linked to a customer's rating. Given that the formula used for these columns was unclear, the decision to hide and exclude the identified missing values *(989 rows)* was made.
- For **bedrooms**, *201* rows of missing values were found. These values could not be imputed as the number of bedrooms can vary and is dependent on other predictor variables, such as "accommodates" or "beds". Hence, the column cannot be imputed based on the column. The missing rows were hidden and excluded from the data set.
- **Bathrooms_text** was identified to have *4* missing values. As the data set cannot derive which type of bath is included in each listing, the missing values were hidden and excluded.
- **Beds** were found to have *168* missing values. As it cannot be derived from any of the available variables, the missing values were hidden and excluded.

Check for N/A values in the data set

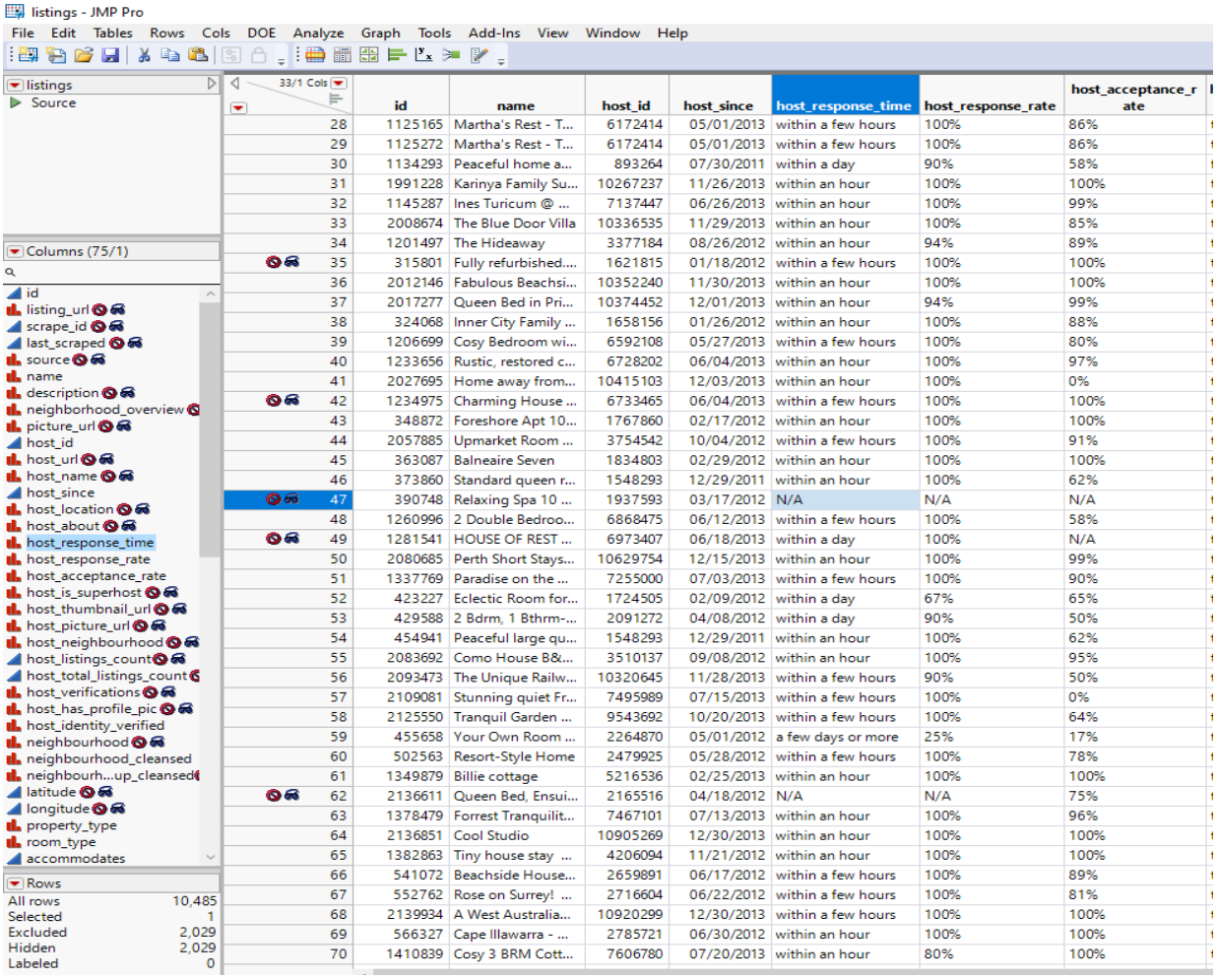
Given that there is a possibility of incorrect values being keyed into the data set, a quick check of the distribution of the columns was undertaken to check for such values. It was found that certain columns have values inputted as N/As.

For the **host_response_time** column, N/A rows were removed as that would not help to,, determine the target variable. Further, this column cannot be imputed using the data as the response time is specific to each property. Lastly, there is no reference to determine what is the response rate for a particular property. Hence, all rows with N/As were removed.

A similar treatment was given to columns **response_rate** and **acceptance_rate**. In total, these columns have 900 N/A values and cannot be predicted based on other values. Hence, these values were hidden and excluded.

Figure 6 shows the N/A inputs that were found in the data set.

Figure 6: Screenshot of hidden and excluded N/A rows



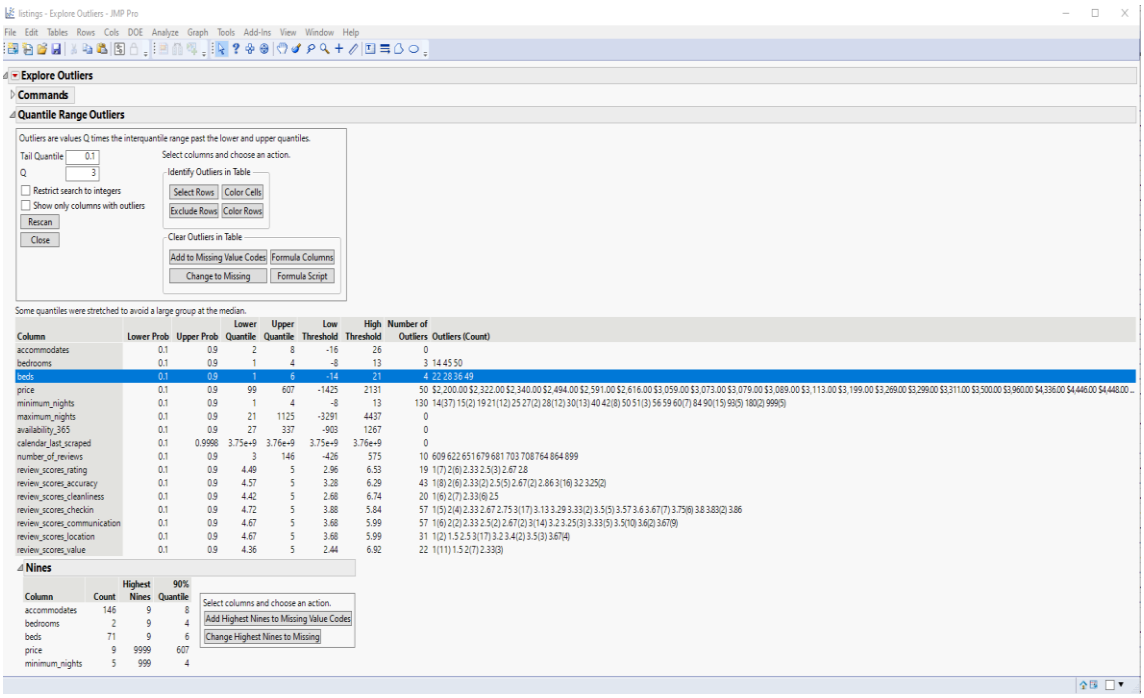
	id	name	host_id	host_since	host_response_time	host_response_rate	host_acceptance_rate
28	1125165	Martha's Rest - T...	6172414	05/01/2013	within a few hours	100%	86%
29	1125272	Martha's Rest - T...	6172414	05/01/2013	within a few hours	100%	86%
30	1134293	Peaceful home a...	893264	07/30/2011	within a day	90%	58%
31	1991228	Karinya Family Su...	10267237	11/26/2013	within an hour	100%	100%
32	1145287	Ines Turicum @ ...	7137447	06/26/2013	within an hour	100%	99%
33	2008674	The Blue Door Villa	10336535	11/29/2013	within an hour	100%	85%
34	1201497	The Hideaway	3377184	08/26/2012	within an hour	94%	89%
35	315801	Fully refurbished....	1621815	01/18/2012	within a few hours	100%	100%
36	2012146	Fabulous Beachsi...	10352240	11/30/2013	within an hour	100%	100%
37	2017277	Queen Bed in Pri...	10374452	12/01/2013	within an hour	94%	99%
38	324068	Inner City Family ...	1658156	01/26/2012	within an hour	100%	88%
39	1206699	Cosy Bedroom wi...	6592108	05/27/2013	within a few hours	100%	80%
40	1233656	Rustic, restored c...	6728202	06/04/2013	within an hour	100%	97%
41	2027695	Home away from...	10415103	12/03/2013	within an hour	100%	0%
42	1234975	Charming House ...	6733465	06/04/2013	within a few hours	100%	100%
43	348872	Foreshore Apt 10...	1767860	02/17/2012	within an hour	100%	100%
44	2057885	Upmarket Room ...	3754542	10/04/2012	within a few hours	100%	91%
45	363087	Balneaire Seven	1834803	02/29/2012	within an hour	100%	100%
46	373860	Standard queen r...	1548293	12/29/2011	within an hour	100%	62%
47	390748	Relaxing Spa 10 ...	1937593	03/17/2012	N/A	N/A	N/A
48	1260996	2 Double Bedroo...	6868475	06/12/2013	within a few hours	100%	58%
49	1281541	HOUSE OF REST ...	6973407	06/18/2013	within a day	100%	N/A
50	2080685	Perth Short Stays...	10629754	12/15/2013	within an hour	100%	99%
51	1337769	Paradise on the ...	7255000	07/03/2013	within a few hours	100%	90%
52	423227	Eclectic Room for...	1724505	02/09/2012	within a day	67%	65%
53	429588	2 Bdrm, 1 Bthrm-...	2091272	04/08/2012	within a day	90%	50%
54	454941	Peaceful large qu...	1548293	12/29/2011	within an hour	100%	62%
55	2083692	Como House B&...	3510137	09/08/2012	within an hour	100%	95%
56	2093473	The Unique Railw...	10320645	11/28/2013	within a few hours	90%	50%
57	2109081	Stunning quiet Fr...	7495989	07/15/2013	within a few hours	100%	0%
58	2125550	Tranquil Garden ...	9543692	10/20/2013	within a few hours	100%	64%
59	455658	Your Own Room ...	2264870	05/01/2012	a few days or more	25%	17%
60	502563	Resort-Style Home	2479925	05/28/2012	within a few hours	100%	78%
61	1349879	Billie cottage	5216536	02/25/2013	within an hour	100%	100%
62	2136611	Queen Bed, Ensui...	2165516	04/18/2012	N/A	N/A	75%
63	1378479	Forrest Tranquilit...	7467101	07/13/2013	within an hour	100%	96%
64	2136851	Cool Studio	10905269	12/30/2013	within an hour	100%	100%
65	1382863	Tiny house stay ...	4206094	11/21/2012	within an hour	100%	100%
66	541072	Beachside House...	2659891	06/17/2012	within a few hours	100%	89%
67	552762	Rose on Surrey! ...	2716604	06/22/2012	within a few hours	100%	81%
68	2139934	A West Australia...	10920299	12/30/2013	within a few hours	100%	100%
69	566327	Cape Illawarra - ...	2785721	06/30/2012	within an hour	100%	100%
70	1410839	Cosy 3 BRM Cott...	7606780	07/20/2013	within an hour	80%	100%

2.4 Outliers

2.4.1 Multivariate outlier detection:

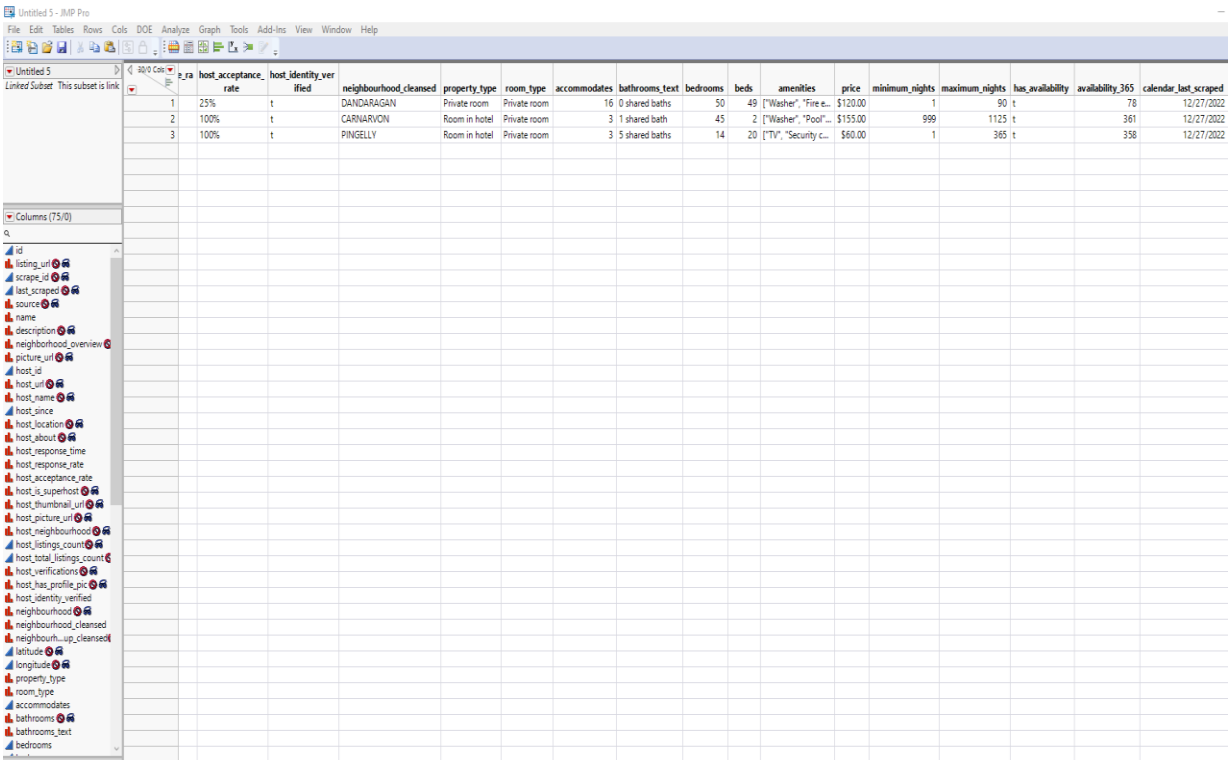
Upon identifying the rows and columns to be hidden and excluded in the preliminary analysis, the outlier function was used to identify and explore outliers across the variables. The multivariate outlier detection was used with a tail of 0.1 and a quantile interval of 3 to find the outliers as in the figure below.

Figure 7: Screenshot of the summary of outliers identified



For **bedrooms**, three rows were identified to feature outliers and two rows were hidden and excluded as a result. Upon further examination, Row 2 was retained as the host could have limited the number of guests to a smaller count to avoid dealing with large parties staying in their listing. Rows 1 and 3 have extreme data points which paint an unrealistic picture. This can be observed in the figure below.

Figure 8: Screenshot of outliers found in 'Bedrooms'

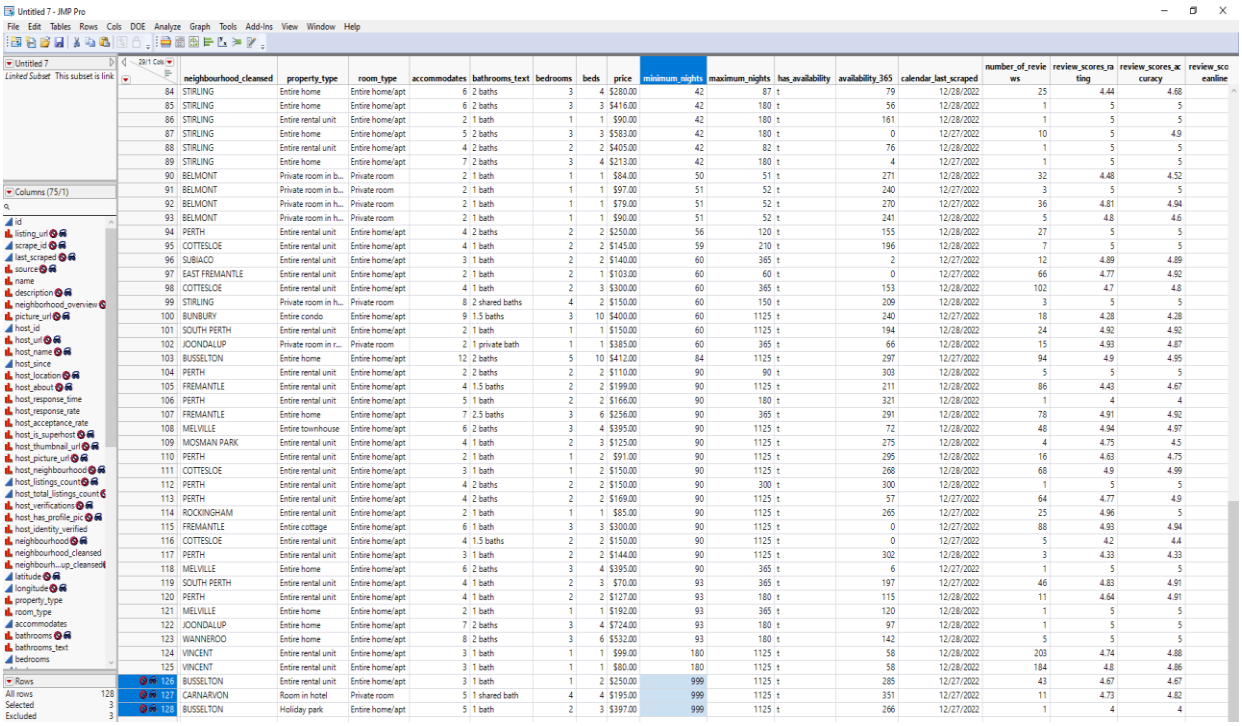


	e_ra	host_acceptance_ rate	host_identity_ver ified	neighbourhood_cleansed	property_type	room_type	accommodates	bathrooms_text	bedrooms	beds	amenities	price	minimum_nights	maximum_nights	has_availability	availability_365	calendar_last_scraped
1		25%	t	DANDARAGAN	Private room	Private room	16	0 shared baths	50	49	["Washer", "Fire e...	$120.00	1	90	t	78	12/27/2022
2		100%	t	CARNARVON	Room in hotel	Private room	3	1 shared bath	45	2	["Washer", "Pool"...	$155.00	999	1125	t	361	12/27/2022
3		100%	t	PINGELLY	Room in hotel	Private room	3	5 shared baths	14	20	["TV", "Security c...	$60.00	1	365	t	358	12/27/2022

Beds were found to have 4 outliers. These values were obscenely high and impractical. Therefore, they were hidden and excluded from the data set.

Minimum_nights was found to have 128 outliers. However, only 3 values were considered outliers as they were very high in value (i.e. 999). A minimum stay of 999 days is not practical and is thereby hidden and excluded. No further action was taken on the remaining outliers as it was assumed that the values could have been set with guests staying for a longer minimum period of time.

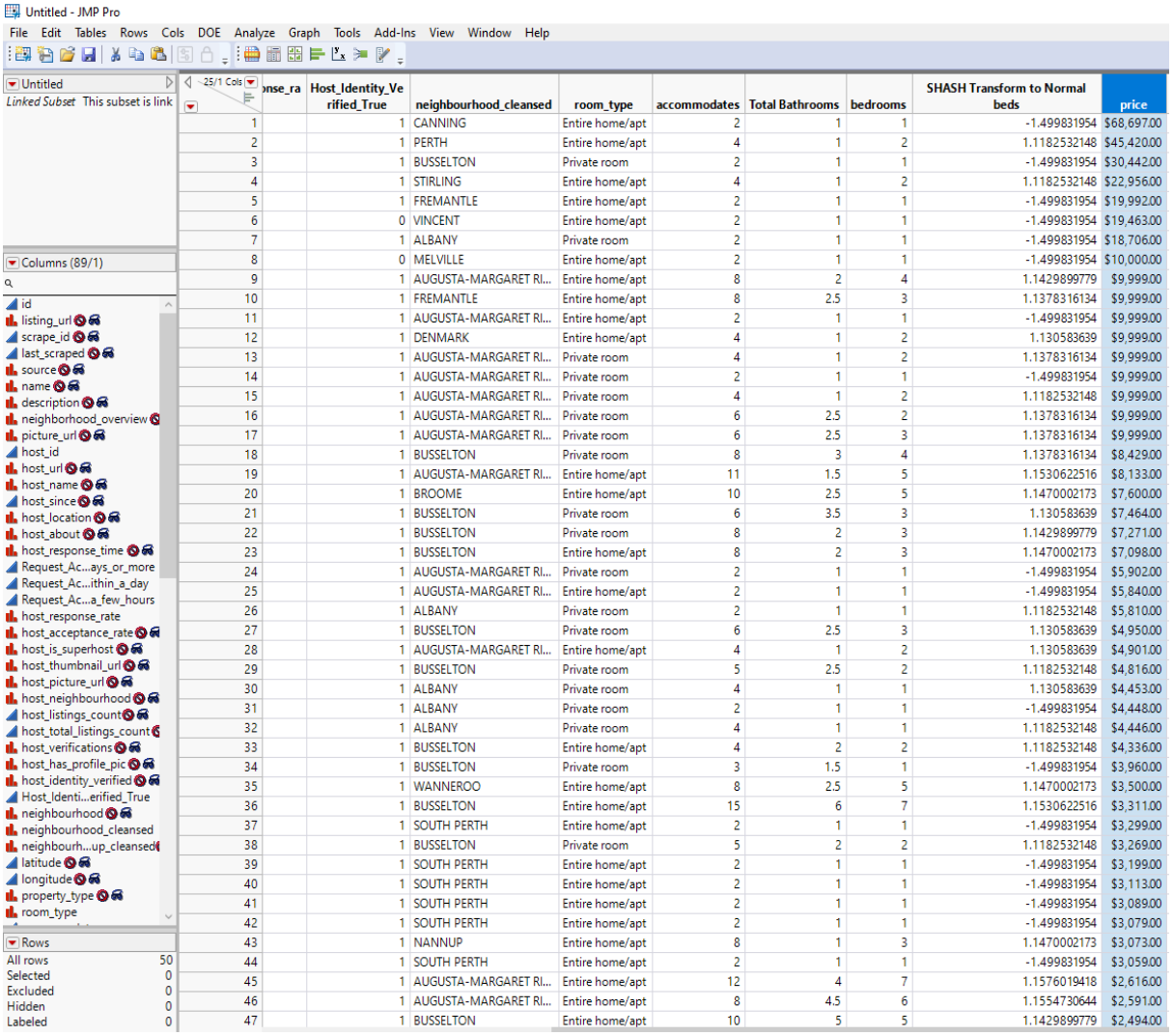
Figure 9: Screenshot of outliers identified in minimum_nights



	neighbourhood_cleansed	property_type	room_type	accommodates	bathrooms_text	bedrooms	beds	price	minimum_nights	maximum_nights	has_availability	availability_365	calendar_last_scraped	number_of_reviews	review_scores_rating	review_scores_accuracy	review_scoreanline
84	STIRLING	Entire home	Entire home/apt	6	2 baths	3	4	$280.00	42	87	t	79	12/28/2022	25	4.44	4.68	
85	STIRLING	Entire home	Entire home/apt	6	2 baths	3	3	$416.00	42	180	t	56	12/28/2022	1	5	5	
86	STIRLING	Entire rental unit	Entire home/apt	2	1 bath	1	1	$90.00	42	180	t	161	12/28/2022	1	5	5	
87	STIRLING	Entire home	Entire home/apt	5	2 baths	3	3	$583.00	42	180	t	0	12/27/2022	10	5	4.9	
88	STIRLING	Entire rental unit	Entire home/apt	4	2 baths	2	2	$405.00	42	82	t	76	12/28/2022	1	5	5	
89	STIRLING	Entire home	Entire home/apt	7	2 baths	3	4	$213.00	42	180	t	4	12/27/2022	1	5	5	
90	BELMONT	Private room in b...	Private room	2	1 bath	1	1	$84.00	50	51	t	271	12/28/2022	32	4.48	4.52	
91	BELMONT	Private room in b...	Private room	2	1 bath	1	1	$97.00	51	52	t	240	12/27/2022	3	5	5	
92	BELMONT	Private room in h...	Private room	2	1 bath	1	1	$79.00	51	52	t	270	12/27/2022	36	4.81	4.94	
93	BELMONT	Private room in h...	Private room	2	1 bath	1	1	$90.00	51	52	t	241	12/28/2022	5	4.8	4.6	
94	PERTH	Entire rental unit	Entire home/apt	4	2 baths	2	2	$250.00	56	120	t	155	12/28/2022	27	5	5	
95	COTTESLOE	Entire rental unit	Entire home/apt	4	1 bath	2	2	$145.00	59	210	t	196	12/28/2022	7	5	5	
96	SUBIACO	Entire rental unit	Entire home/apt	3	1 bath	2	2	$140.00	60	365	t	2	12/27/2022	12	4.89	4.89	
97	EAST FREMANTLE	Entire rental unit	Entire home/apt	2	1 bath	2	1	$103.00	60	60	t	0	12/27/2022	66	4.77	4.92	
98	COTTESLOE	Entire rental unit	Entire home/apt	4	1 bath	2	3	$300.00	60	365	t	153	12/28/2022	102	4.7	4.8	
99	STIRLING	Private room in h...	Private room	8	2 shared baths	4	2	$150.00	60	150	t	209	12/28/2022	3	5	5	
100	BUNBURY	Entire condo	Entire home/apt	9	1.5 baths	3	10	$400.00	60	1125	t	240	12/27/2022	18	4.28	4.28	
101	SOUTH PERTH	Entire rental unit	Entire home/apt	2	1 bath	1	1	$150.00	60	1125	t	194	12/28/2022	24	4.92	4.92	
102	JOONDALUP	Private room in r...	Private room	2	1 private bath	1	1	$385.00	60	365	t	66	12/28/2022	15	4.93	4.87	
103	BUSSELTON	Entire home	Entire home/apt	12	2 baths	5	10	$412.00	84	1125	t	297	12/27/2022	94	4.9	4.95	
104	PERTH	Entire rental unit	Entire home/apt	2	2 baths	2	2	$110.00	90	90	t	303	12/28/2022	5	5	5	
105	FREMANTLE	Entire rental unit	Entire home/apt	4	1.5 baths	2	2	$199.00	90	1125	t	211	12/28/2022	86	4.43	4.67	
106	PERTH	Entire rental unit	Entire home/apt	5	1 bath	2	2	$166.00	90	180	t	321	12/28/2022	1	4	4	
107	FREMANTLE	Entire home	Entire home/apt	7	2.5 baths	3	6	$256.00	90	365	t	291	12/28/2022	78	4.91	4.92	
108	MELVILLE	Entire townhouse	Entire home/apt	6	2 baths	3	4	$395.00	90	1125	t	72	12/28/2022	48	4.94	4.97	
109	MOSMAN PARK	Entire rental unit	Entire home/apt	4	1 bath	2	3	$125.00	90	1125	t	275	12/28/2022	4	4.75	4.5	
110	PERTH	Entire rental unit	Entire home/apt	2	1 bath	1	2	$91.00	90	1125	t	295	12/28/2022	16	4.63	4.75	
111	COTTESLOE	Entire rental unit	Entire home/apt	3	1 bath	1	2	$150.00	90	1125	t	268	12/28/2022	68	4.9	4.99	
112	PERTH	Entire rental unit	Entire home/apt	4	2 baths	2	2	$150.00	90	300	t	300	12/28/2022	1	5	5	
113	PERTH	Entire rental unit	Entire home/apt	4	2 baths	2	2	$169.00	90	1125	t	57	12/27/2022	64	4.77	4.9	
114	ROCKINGHAM	Entire rental unit	Entire home/apt	2	1 bath	1	1	$85.00	90	1125	t	265	12/27/2022	25	4.96	5	
115	FREMANTLE	Entire cottage	Entire home/apt	6	1 bath	3	3	$300.00	90	1125	t	0	12/27/2022	88	4.93	4.94	
116	COTTESLOE	Entire rental unit	Entire home/apt	4	1.5 baths	2	2	$150.00	90	1125	t	0	12/27/2022	5	4.2	4.4	
117	PERTH	Entire rental unit	Entire home/apt	3	1 bath	2	2	$144.00	90	1125	t	302	12/28/2022	3	4.33	4.33	
118	MELVILLE	Entire home	Entire home/apt	6	2 baths	3	4	$395.00	90	365	t	6	12/27/2022	1	5	5	
119	SOUTH PERTH	Entire rental unit	Entire home/apt	4	1 bath	2	3	$70.00	93	365	t	197	12/27/2022	46	4.83	4.91	
120	PERTH	Entire rental unit	Entire home/apt	4	1 bath	2	2	$127.00	93	180	t	115	12/28/2022	11	4.64	4.91	
121	MELVILLE	Entire home	Entire home/apt	2	1 bath	1	1	$192.00	93	365	t	120	12/28/2022	1	5	5	
122	JOONDALUP	Entire home	Entire home/apt	7	2 baths	3	4	$724.00	93	180	t	97	12/28/2022	1	5	5	
123	WANNEROO	Entire home	Entire home/apt	8	2 baths	3	6	$532.00	93	180	t	142	12/28/2022	5	5	5	
124	VINCENT	Entire rental unit	Entire home/apt	3	1 bath	1	1	$99.00	180	1125	t	58	12/28/2022	203	4.74	4.88	
125	VINCENT	Entire rental unit	Entire home/apt	3	1 bath	1	1	$80.00	180	1125	t	58	12/28/2022	184	4.8	4.86	
126	BUSSELTON	Entire rental unit	Entire home/apt	3	1 bath	1	2	$250.00	999	1125	t	285	12/27/2022	43	4.67	4.67	
127	CARNARVON	Room in hotel	Private room	5	1 shared bath	4	4	$195.00	999	1125	t	351	12/27/2022	11	4.73	4.82	
128	BUSSELTON	Holiday park	Entire home/apt	5	1 bath	2	3	$397.00	999	1125	t	266	12/27/2022	1	4	4	

For **Price**, 50 rows were found to be outliers in the data set. Given the target variable has an influence on the data set, the outliers were treated by hiding and excluding them. Extreme values of price indicate the listing can be considered as luxury properties which in turn can be used in another model.

Figure 10: Screenshot of outliers identified in Price



		Host_Identity_Verified_True	neighbourhood_cleansed	room_type	accommodates	Total Bathrooms	bedrooms	SHASH Transform to Normal beds	price
1		1	CANNING	Entire home/apt	2	1	1	-1.499831954	$68,697.00
2		1	PERTH	Entire home/apt	4	1	2	1.1182532148	$45,420.00
3		1	BUSSELTON	Private room	2	1	1	-1.499831954	$30,442.00
4		1	STIRLING	Entire home/apt	4	1	2	1.1182532148	$22,956.00
5		1	FREMANTLE	Entire home/apt	2	1	1	-1.499831954	$19,992.00
6		0	VINCENT	Entire home/apt	2	1	1	-1.499831954	$19,463.00
7		1	ALBANY	Private room	2	1	1	-1.499831954	$18,706.00
8		0	MELVILLE	Entire home/apt	2	1	1	-1.499831954	$10,000.00
9		1	AUGUSTA-MARGARET RI...	Entire home/apt	8	2	4	1.1429899779	$9,999.00
10		1	FREMANTLE	Entire home/apt	8	2.5	3	1.1378316134	$9,999.00
11		1	AUGUSTA-MARGARET RI...	Entire home/apt	2	1	1	-1.499831954	$9,999.00
12		1	DENMARK	Entire home/apt	4	1	2	1.130583639	$9,999.00
13		1	AUGUSTA-MARGARET RI...	Private room	4	1	2	1.1378316134	$9,999.00
14		1	AUGUSTA-MARGARET RI...	Private room	2	1	1	-1.499831954	$9,999.00
15		1	AUGUSTA-MARGARET RI...	Private room	4	1	2	1.1182532148	$9,999.00
16		1	AUGUSTA-MARGARET RI...	Private room	6	2.5	2	1.1378316134	$9,999.00
17		1	AUGUSTA-MARGARET RI...	Private room	6	2.5	3	1.1378316134	$9,999.00
18		1	BUSSELTON	Private room	8	3	4	1.1378316134	$8,429.00
19		1	AUGUSTA-MARGARET RI...	Entire home/apt	11	1.5	5	1.1530622516	$8,133.00
20		1	BROOME	Entire home/apt	10	2.5	5	1.1470002173	$7,600.00
21		1	BUSSELTON	Private room	6	3.5	3	1.130583639	$7,464.00
22		1	BUSSELTON	Private room	8	2	3	1.1429899779	$7,271.00
23		1	BUSSELTON	Entire home/apt	8	2	3	1.1470002173	$7,098.00
24		1	AUGUSTA-MARGARET RI...	Private room	2	1	1	-1.499831954	$5,902.00
25		1	AUGUSTA-MARGARET RI...	Entire home/apt	2	1	1	-1.499831954	$5,840.00
26		1	ALBANY	Private room	2	1	1	1.1182532148	$5,810.00
27		1	BUSSELTON	Private room	6	2.5	3	1.130583639	$4,950.00
28		1	AUGUSTA-MARGARET RI...	Entire home/apt	4	1	2	1.130583639	$4,901.00
29		1	BUSSELTON	Private room	5	2.5	2	1.1182532148	$4,816.00
30		1	ALBANY	Private room	4	1	1	1.130583639	$4,453.00
31		1	ALBANY	Private room	2	1	1	-1.499831954	$4,448.00
32		1	ALBANY	Private room	4	1	1	1.1182532148	$4,446.00
33		1	BUSSELTON	Entire home/apt	4	2	2	1.1182532148	$4,336.00
34		1	BUSSELTON	Private room	3	1.5	1	-1.499831954	$3,960.00
35		1	WANNEROO	Entire home/apt	8	2.5	5	1.1470002173	$3,500.00
36		1	BUSSELTON	Entire home/apt	15	6	7	1.1530622516	$3,311.00
37		1	SOUTH PERTH	Entire home/apt	2	1	1	-1.499831954	$3,299.00
38		1	BUSSELTON	Private room	5	2	2	1.1182532148	$3,269.00
39		1	SOUTH PERTH	Entire home/apt	2	1	1	-1.499831954	$3,199.00
40		1	SOUTH PERTH	Entire home/apt	2	1	1	-1.499831954	$3,113.00
41		1	SOUTH PERTH	Entire home/apt	2	1	1	-1.499831954	$3,089.00
42		1	SOUTH PERTH	Entire home/apt	2	1	1	-1.499831954	$3,079.00
43		1	NANNUP	Entire home/apt	8	1	3	1.1470002173	$3,073.00
44		1	SOUTH PERTH	Entire home/apt	2	1	1	-1.499831954	$3,059.00
45		1	AUGUSTA-MARGARET RI...	Entire home/apt	12	4	7	1.1576019418	$2,616.00
46		1	AUGUSTA-MARGARET RI...	Entire home/apt	8	4.5	6	1.1554730644	$2,591.00
47		1	BUSSELTON	Entire home/apt	10	5	5	1.1429899779	$2,494.00

For **number_of_reviews** and **review_scores for different parameters (ex. accuracy, cleanliness, etc)**, the values have been retained with no change, as the range of these variables can be high or low numbers, as long as it is non-negative and lower than 5, which was verified.

2.4.2 Mahalanobis distance:

Additional outliers were identified using the Scatterplot matrix and derived from the Mahalanobis distances.

Figure 11: Mahalanobis Diagram

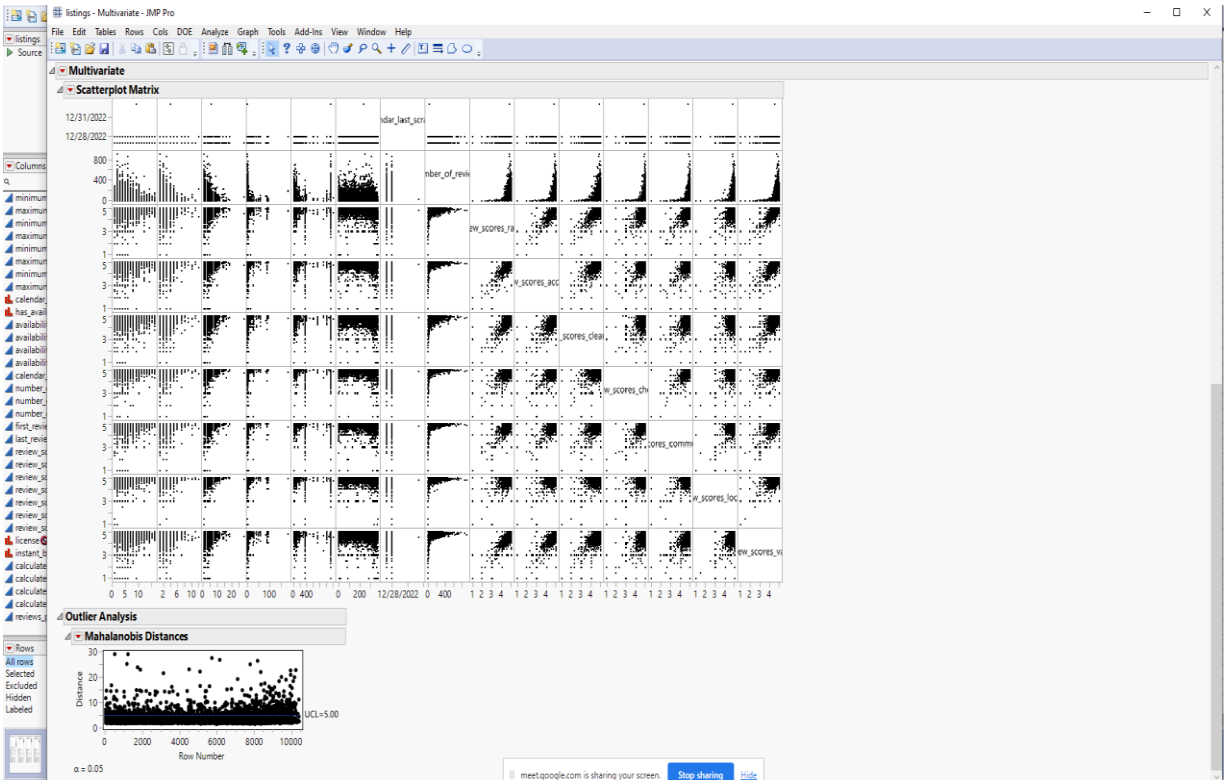
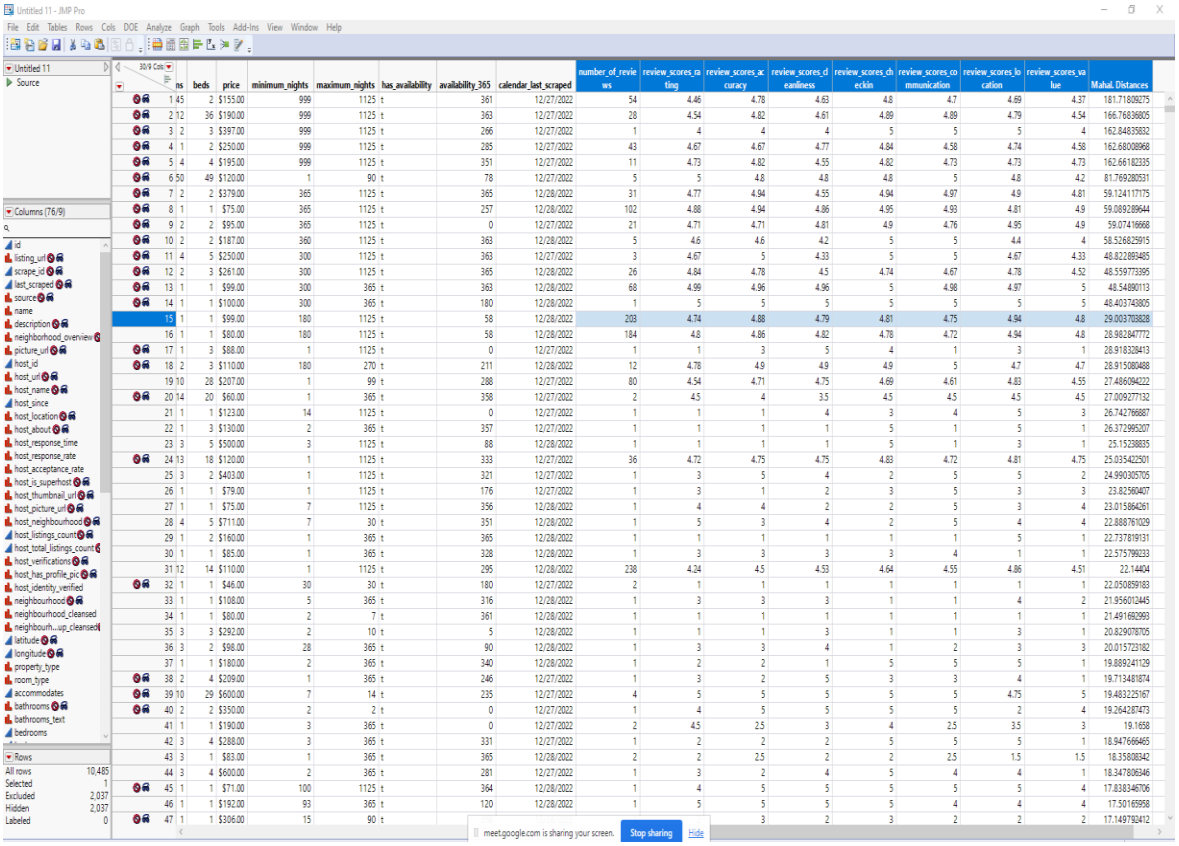


Figure 12: Outliers identified by Mahalanobis Diagram



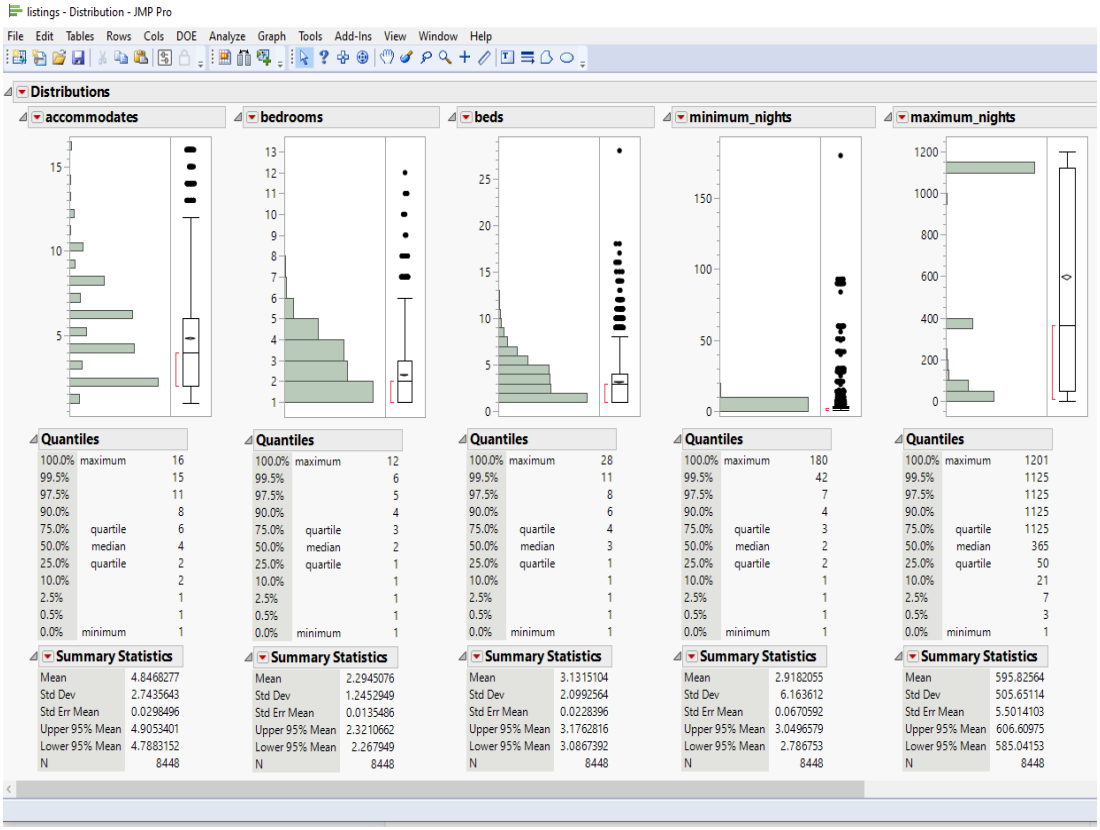
	beds	price	minimum_nights	maximum_nights	has_availability	availability_365	calendar_last_scraped	number_of_reviews	review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_location	review_scores_value	Mahal. Distances
1	45	2	$155.00	999	1125	t	361	12/27/2022	54	4.46	4.78	4.63	4.8	4.7	4.69	4.37	181.71808275
2	12	36	$190.00	999	1125	t	363	12/27/2022	28	4.54	4.82	4.61	4.89	4.89	4.79	4.54	166.76836805
3	2	3	$397.00	999	1125	t	266	12/27/2022	1	4	4	4	5	5	5	4	162.84835832
4	1	2	$250.00	999	1125	t	285	12/27/2022	43	4.67	4.67	4.77	4.84	4.58	4.74	4.58	162.68008968
5	4	4	$195.00	999	1125	t	351	12/27/2022	11	4.73	4.82	4.55	4.82	4.73	4.73	4.73	162.66182335
6	50	49	$120.00	1	90	t	78	12/27/2022	5	5	4.8	4.8	4.8	5	4.8	4.2	81.769280531
7	2	2	$379.00	365	1125	t	365	12/28/2022	31	4.77	4.94	4.55	4.94	4.97	4.9	4.81	59.124117175
8	1	1	$75.00	365	1125	t	257	12/28/2022	102	4.88	4.94	4.86	4.95	4.93	4.81	4.9	59.089289644
9	2	2	$95.00	365	1125	t	0	12/27/2022	21	4.71	4.71	4.81	4.9	4.76	4.95	4.9	59.07416668
10	2	2	$187.00	360	1125	t	363	12/28/2022	5	4.6	4.6	4.2	5	5	4.4	4	58.526825915
11	4	5	$250.00	300	1125	t	363	12/27/2022	3	4.67	5	4.33	5	5	4.67	4.33	48.822893485
12	2	3	$261.00	300	1125	t	365	12/28/2022	26	4.84	4.78	4.5	4.74	4.67	4.78	4.52	48.559773395
13	1	1	$99.00	300	365	t	363	12/28/2022	68	4.99	4.96	4.96	5	4.98	4.97	5	48.54890113
14	1	1	$100.00	300	365	t	180	12/28/2022	1	5	5	5	5	5	5	5	48.403749805
15	1	1	$99.00	180	1125	t	58	12/28/2022	203	4.74	4.88	4.79	4.81	4.75	4.94	4.8	29.003703828
16	1	1	$80.00	180	1125	t	58	12/28/2022	184	4.8	4.86	4.82	4.78	4.72	4.94	4.8	28.982847772
17	1	3	$88.00	1	1125	t	0	12/27/2022	1	1	3	5	4	1	3	1	28.918328413
18	2	3	$110.00	180	270	t	211	12/28/2022	12	4.78	4.9	4.9	4.9	5	4.7	4.7	28.915080488
19	10	28	$207.00	1	99	t	288	12/27/2022	80	4.54	4.71	4.75	4.69	4.61	4.83	4.55	27.486094222
20	14	20	$60.00	1	365	t	358	12/27/2022	2	4.5	4	3.5	4.5	4.5	4.5	4.5	27.009277132
21	1	1	$123.00	14	1125	t	0	12/27/2022	1	1	1	4	3	4	5	3	26.742766887
22	1	3	$130.00	2	365	t	357	12/27/2022	1	1	1	1	5	1	5	1	26.372995207
23	3	5	$500.00	3	1125	t	88	12/28/2022	1	1	1	1	5	1	3	1	25.15238835
24	13	18	$120.00	1	1125	t	333	12/27/2022	36	4.72	4.75	4.75	4.83	4.72	4.81	4.75	25.035422501
25	3	2	$403.00	1	1125	t	321	12/27/2022	1	3	5	4	2	5	5	2	24.990305705
26	1	1	$79.00	1	1125	t	176	12/27/2022	1	3	1	2	3	5	3	3	23.82560407
27	1	1	$75.00	7	1125	t	356	12/28/2022	1	4	4	2	2	5	3	4	23.015864261
28	4	5	$711.00	7	30	t	351	12/28/2022	1	5	3	4	2	5	4	4	22.888761029
29	1	2	$160.00	1	365	t	365	12/28/2022	1	1	1	1	1	1	5	1	22.737819131
30	1	1	$85.00	1	365	t	328	12/28/2022	1	3	3	3	3	4	1	1	22.575799233
31	12	14	$110.00	1	1125	t	295	12/28/2022	238	4.24	4.5	4.53	4.64	4.55	4.86	4.51	22.14404
32	1	1	$46.00	30	30	t	180	12/27/2022	2	1	1	1	1	1	1	1	22.050859183
33	1	1	$108.00	5	365	t	316	12/28/2022	1	3	3	3	1	1	4	2	21.956012445
34	1	1	$80.00	2	7	t	361	12/28/2022	1	1	1	1	1	1	1	1	21.491692993
35	3	3	$292.00	2	10	t	5	12/28/2022	1	1	1	3	1	1	3	1	20.829078705
36	3	2	$98.00	28	365	t	90	12/28/2022	1	3	3	4	1	2	3	3	20.015723182
37	1	1	$180.00	2	365	t	340	12/28/2022	1	2	2	1	5	5	5	1	19.889241129
38	2	4	$209.00	1	365	t	246	12/27/2022	1	3	2	5	3	3	4	1	19.713481874
39	10	29	$600.00	7	14	t	235	12/27/2022	4	5	5	5	5	5	4.75	5	19.483225167
40	2	2	$350.00	2	2	t	0	12/27/2022	1	4	5	5	5	5	2	4	19.264287473
41	1	1	$190.00	3	365	t	0	12/27/2022	2	4.5	2.5	3	4	2.5	3.5	3	19.1658
42	3	4	$288.00	3	365	t	331	12/27/2022	1	2	2	2	5	5	5	1	18.947666465
43	3	1	$83.00	1	365	t	365	12/28/2022	2	2	2.5	2	2	2.5	1.5	1.5	18.35808342
44	3	4	$600.00	2	365	t	281	12/27/2022	1	3	2	4	5	4	4	1	18.347806346
45	1	1	$71.00	100	1125	t	364	12/28/2022	1	4	5	5	5	5	5	4	17.838346706
46	1	1	$192.00	93	365	t	120	12/28/2022	1	5	5	5	5	4	4	4	17.50169958
47	1	1	$306.00	15	90	t					3	2	3	2	2	2	17.149792412

Looking at the Mahalanobis Distances in descending order, there were many rows found to be away from the Upper Control Limit as can be seen in Figure 12. A decision was made to exclude the first 15 rows, as the distance values were unusually higher than the remaining data.

2.4.3 Transformation of variables:

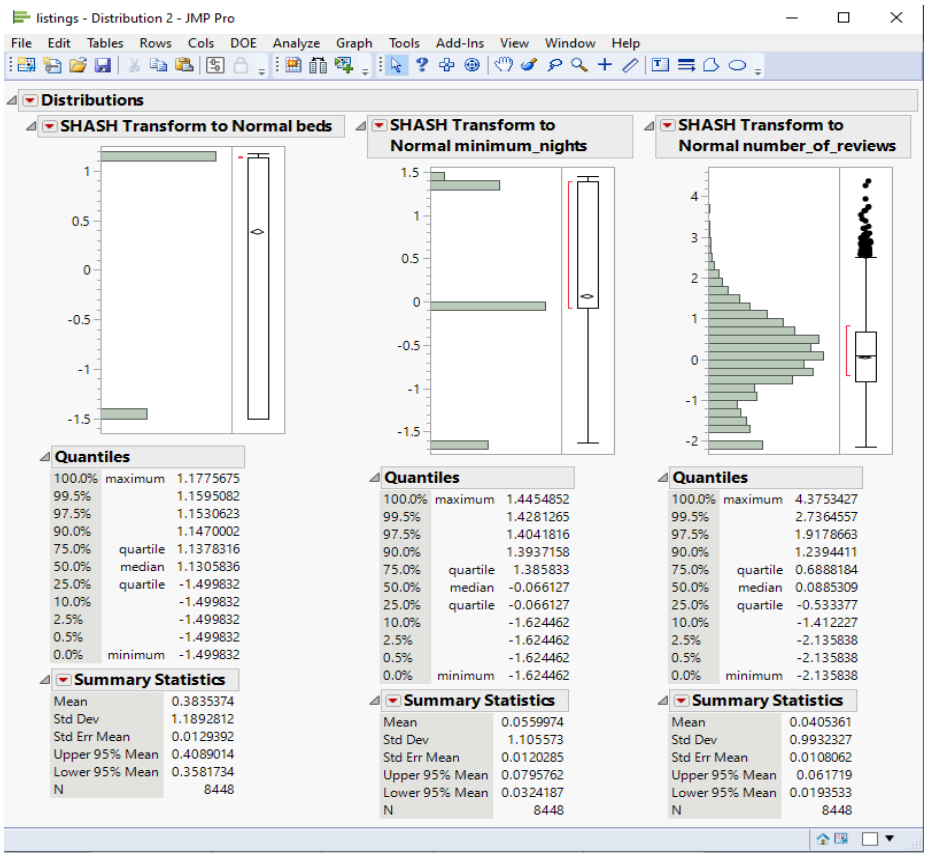
To further clean the columns with outliers, the histogram distribution was viewed for all continuous variables. **Bedrooms, minimum_nights, and number_of_reviews** were identified to have a lot of outliers. The majority of the data lay beyond the whiskers of the box plot. Therefore, these three variables were transformed to smooth out the data and capture the points, which were currently identified as outliers.

Figure 13: Screenshot of identified variables with outliers prior to transformation



After applying the 'fit all' distribution command on the 3 identified variables, the SHASH distribution was identified to be the best distribution which transforms the data closest to the normal distribution. They were found to have the lowest AICC and the resultant distribution has smoothened the data as can be seen in the figure below.

Figure 14: Screenshot of identified variables with outliers post-transformation



2.5 Reducing Dimensionality:

The data set has 7 columns that contain reviews on different parameters: rating, accuracy, cleanliness, checkin, communication, location, and value. These 7 columns have similar information and a decision was made to reduce the dimensionality. Therefore, the Principal Components Analysis technique was applied. Reviewing the eigenvalues, most of the data i.e. 91.29% is present within the first 4 PCAs. So, these 4 PCAs were saved to the dataset. The new columns created in the dataset are **Review_Scores_PCA1, Review_Scores_PCA2, Review_Scores_PCA3, and Review_Scores_PCA4.**

Figure 15: Screenshot of identified variables with outliers post-transformation

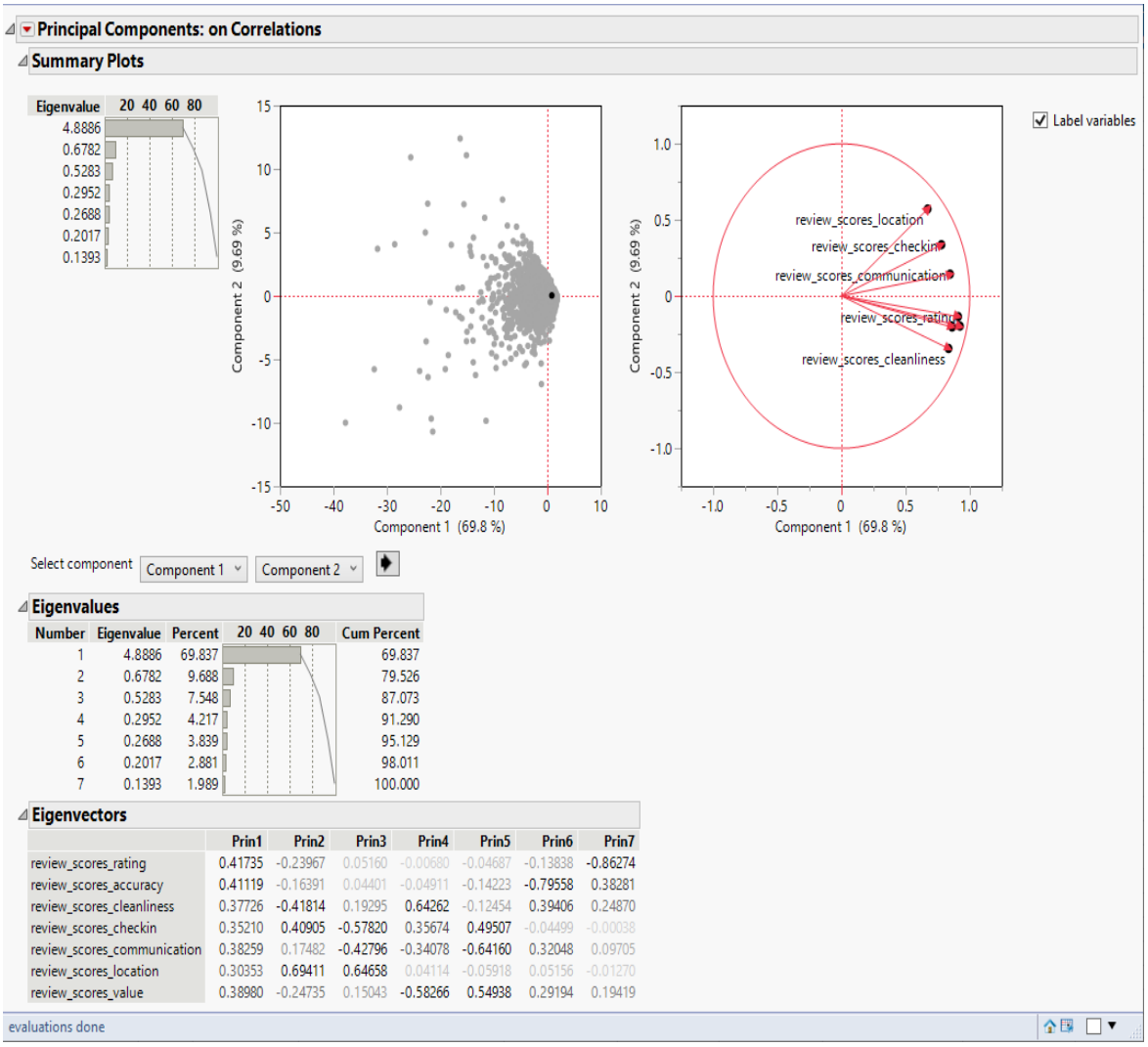
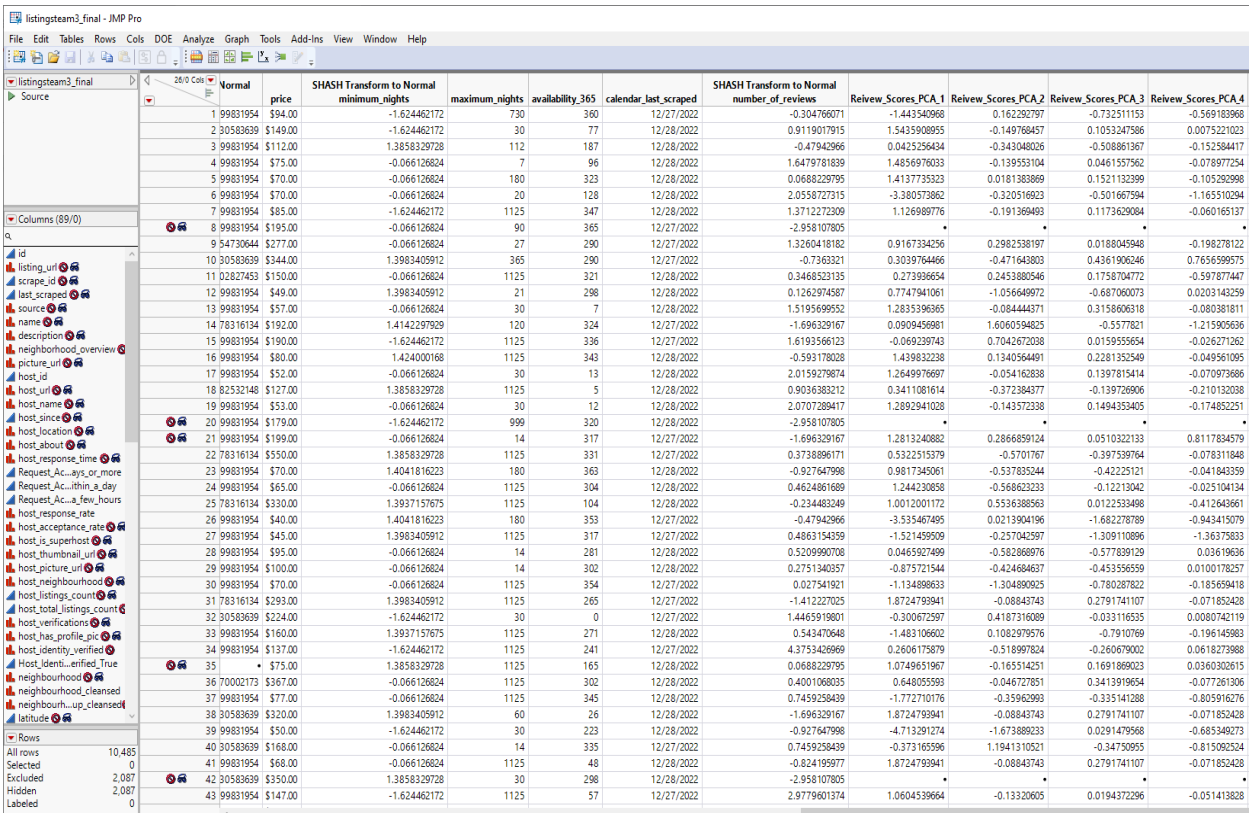


Figure 16: Principal Component Analysis



listingsteam3_final - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help

listingsteam3_final
Source

Columns (89/0)
id
listing_url
scrape_id
last_scraped
source
name
description
neighborhood_overview
picture_url
host_id
host_url
host_name
host_since
host_location
host_about
host_response_time
Request_Ac...ays_or_more
Request_Ac...ithin_a_day
Request_Ac...a_few_hours
host_response_rate
host_acceptance_rate
host_is_superhost
host_thumbnail_url
host_picture_url
host_neighbourhood
host_listings_count
host_total_listings_count
host_verifications
host_has_profile_pic
host_identity_verified
Host_Identi...erified_True
neighbourhood
neighbourhood_cleansed
neighbourh...up_cleansed
latitude

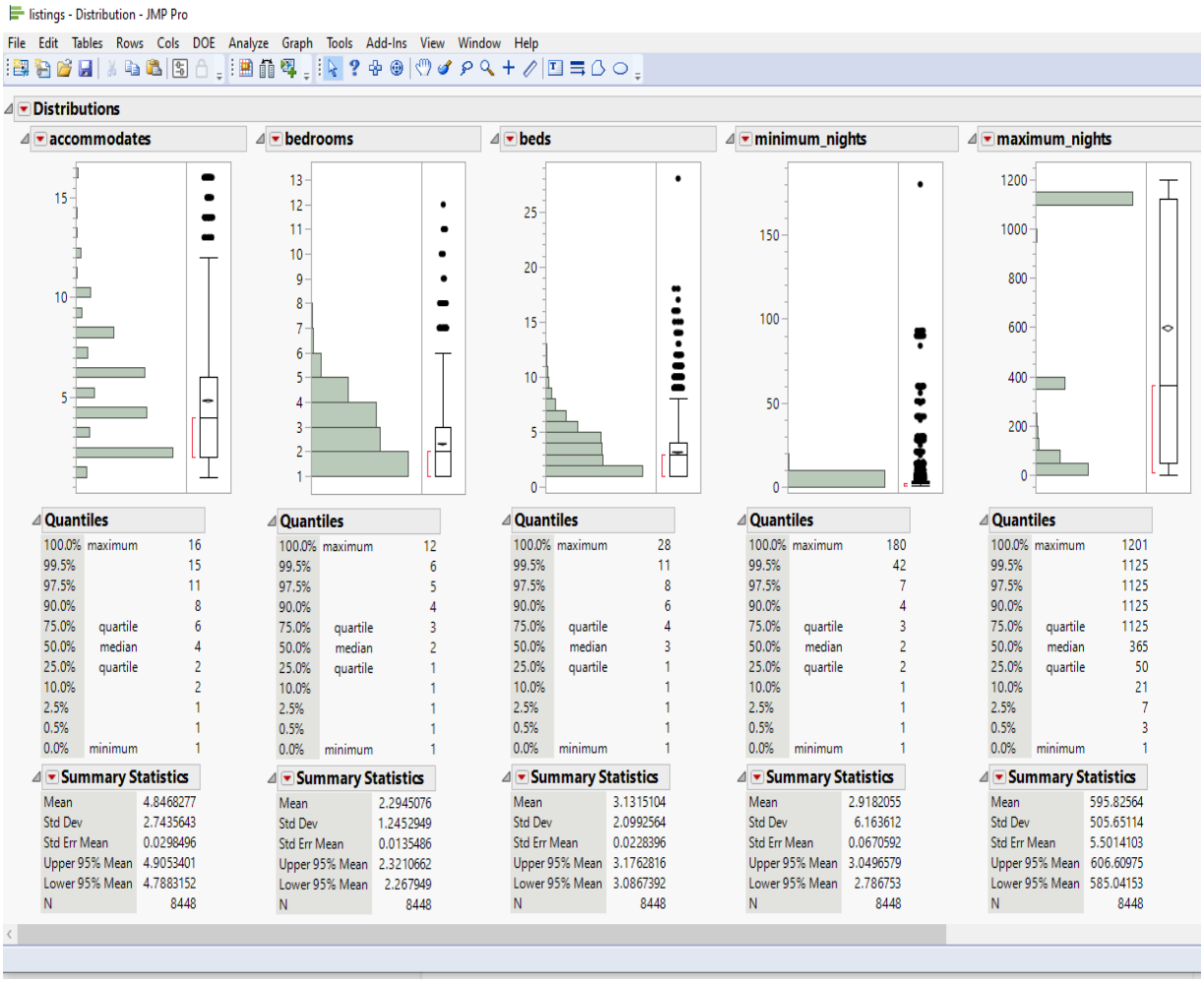
Rows
All rows 10,485
Selected 0
Excluded 2,087
Hidden 2,087
Labeled 0

	Normal	price	SHASH Transform to Normal minimum_nights	maximum_nights	availability_365	calendar_last_scraped	SHASH Transform to Normal number_of_reviews	Reivew_Scores_PCA_1	Reivew_Scores_PCA_2	Reivew_Scores_PCA_3	Reivew_Scores_PCA_4
1	99831954	$94.00	-1.624462172	730	360	12/27/2022	-0.304766071	-1.443540968	0.162292797	-0.732511153	-0.569183968
2	30583639	$149.00	-1.624462172	30	77	12/28/2022	0.9119017915	1.5435908955	-0.149768457	0.1053247586	0.0075221023
3	99831954	$112.00	1.3858329728	112	187	12/28/2022	-0.47942966	0.0425256434	-0.343048026	-0.508861367	-0.152584417
4	99831954	$75.00	-0.066126824	7	96	12/28/2022	1.6479781839	1.4856976033	-0.139553104	0.0461557562	-0.078977254
5	99831954	$70.00	-0.066126824	180	323	12/28/2022	0.0688229795	1.4137735323	0.0181383869	0.1521132399	-0.105292998
6	99831954	$70.00	-0.066126824	20	128	12/28/2022	2.0558727315	-3.380573862	-0.320516923	-0.501667594	-1.165510294
7	99831954	$85.00	-1.624462172	1125	347	12/28/2022	1.3712272309	1.126989776	-0.191369493	0.1173629084	-0.060165137
8	99831954	$195.00	-0.066126824	90	365	12/27/2022	-2.958107805	.	.	.	.
9	54730644	$277.00	-0.066126824	27	290	12/27/2022	1.3260418182	0.9167334256	0.2982538197	0.0188045948	-0.198278122
10	30583639	$344.00	1.3983405912	365	290	12/27/2022	-0.7363321	0.3039764466	-0.471643803	0.4361906246	0.7656599575
11	02827453	$150.00	-0.066126824	1125	321	12/28/2022	0.3468523135	0.273936654	0.2453880546	0.1758704772	-0.597877447
12	99831954	$49.00	1.3983405912	21	298	12/28/2022	0.1262974587	0.7747941061	-1.056649972	-0.687060073	0.0203143259
13	99831954	$57.00	-0.066126824	30	7	12/28/2022	1.5195699552	1.2835396365	-0.084444371	0.3158606318	-0.080381811
14	78316134	$192.00	1.4142297929	120	324	12/27/2022	-1.696329167	0.0909456981	1.6060594825	-0.5577821	-1.215905636
15	99831954	$190.00	-1.624462172	1125	336	12/27/2022	1.6193566123	-0.069239743	0.7042672038	0.0159555654	-0.026271262
16	99831954	$80.00	1.424000168	1125	343	12/28/2022	-0.593178028	1.439832238	0.1340564491	0.2281352549	-0.049561095
17	99831954	$52.00	-0.066126824	30	13	12/28/2022	2.0159279874	1.2649976697	-0.054162838	0.1397815414	-0.070973686
18	82532148	$127.00	1.3858329728	1125	5	12/28/2022	0.9036383212	0.3411081614	-0.372384377	-0.139726906	-0.210132038
19	99831954	$53.00	-0.066126824	30	12	12/28/2022	2.0707289417	1.2892941028	-0.143572338	0.1494353405	-0.174852251
20	99831954	$179.00	-1.624462172	999	320	12/28/2022	-2.958107805	.	.	.	.
21	99831954	$199.00	-0.066126824	14	317	12/27/2022	-1.696329167	1.2813240882	0.2866859124	0.0510322133	0.8117834579
22	78316134	$550.00	1.3858329728	1125	331	12/27/2022	0.3738896171	0.5322515379	-0.5701767	-0.397539764	-0.078311848
23	99831954	$70.00	1.4041816223	180	363	12/28/2022	-0.927647998	0.9817345061	-0.537835244	-0.42225121	-0.041843359
24	99831954	$65.00	-0.066126824	1125	304	12/28/2022	0.4624861689	1.244230858	-0.568623233	-0.12213042	-0.025104134
25	78316134	$330.00	1.3937157675	1125	104	12/28/2022	-0.234483249	1.0012001172	0.5536388563	0.0122533498	-0.412643661
26	99831954	$40.00	1.4041816223	180	353	12/27/2022	-0.47942966	-3.535467495	0.0213904196	-1.682278789	-0.943415079
27	99831954	$45.00	1.3983405912	1125	317	12/27/2022	0.4863154359	-1.521459509	-0.257042597	-1.309110896	-1.36375833
28	99831954	$95.00	-0.066126824	14	281	12/28/2022	0.5209990708	0.0465927499	-0.582866976	-0.577839129	0.03619636
29	99831954	$100.00	-0.066126824	14	302	12/28/2022	0.2751340357	-0.875721544	-0.424684637	-0.453556559	0.0100178257
30	99831954	$70.00	-0.066126824	1125	354	12/27/2022	0.027541921	-1.134898633	-1.304890925	-0.780287822	-0.185659418
31	78316134	$293.00	1.3983405912	1125	265	12/27/2022	-1.412227025	1.8724793941	-0.08843743	0.2791741107	-0.071852428
32	30583639	$224.00	-1.624462172	30	0	12/27/2022	1.4465919801	-0.300672597	0.4187316089	-0.033116535	0.0080742119
33	99831954	$160.00	1.3937157675	1125	271	12/28/2022	0.543470648	-1.483106602	0.1082979576	-0.7910769	-0.196145983
34	99831954	$137.00	-1.624462172	1125	241	12/27/2022	4.3753426969	0.2606175879	-0.518997824	-0.260679002	0.0618273988
35	.	$75.00	1.3858329728	1125	165	12/28/2022	0.0688229795	1.0749651967	-0.165514251	0.1691869023	0.0360302615
36	70002173	$367.00	-0.066126824	1125	302	12/28/2022	0.4001068035	0.648055593	-0.046727851	0.3413919654	-0.077261306
37	99831954	$77.00	-0.066126824	1125	345	12/28/2022	0.7459258439	-1.772710176	-0.35962993	-0.335141288	-0.805916276
38	30583639	$320.00	1.3983405912	60	26	12/28/2022	-1.696329167	1.8724793941	-0.08843743	0.2791741107	-0.071852428
39	99831954	$50.00	-1.624462172	30	223	12/28/2022	-0.927647998	-4.713291274	-1.673889233	0.0291479568	-0.685349273
40	30583639	$168.00	-0.066126824	14	335	12/27/2022	0.7459258439	-0.373165596	1.1941310521	-0.34750955	-0.815092524
41	99831954	$68.00	-0.066126824	1125	48	12/28/2022	-0.824195977	1.8724793941	-0.08843743	0.2791741107	-0.071852428
42	30583639	$350.00	1.3858329728	30	298	12/28/2022	-2.958107805	.	.	.	.
43	99831954	$147.00	-1.624462172	1125	57	12/27/2022	2.9779601374	1.0604539664	-0.13320605	0.0194372296	-0.051413828

3. Partitioning the data

Post pre-processing the data, cleansing the outliers and missing data values, Figure 17 shows the distribution across different columns.

Figure 17: Screenshot of identified variables with outliers prior to transformation



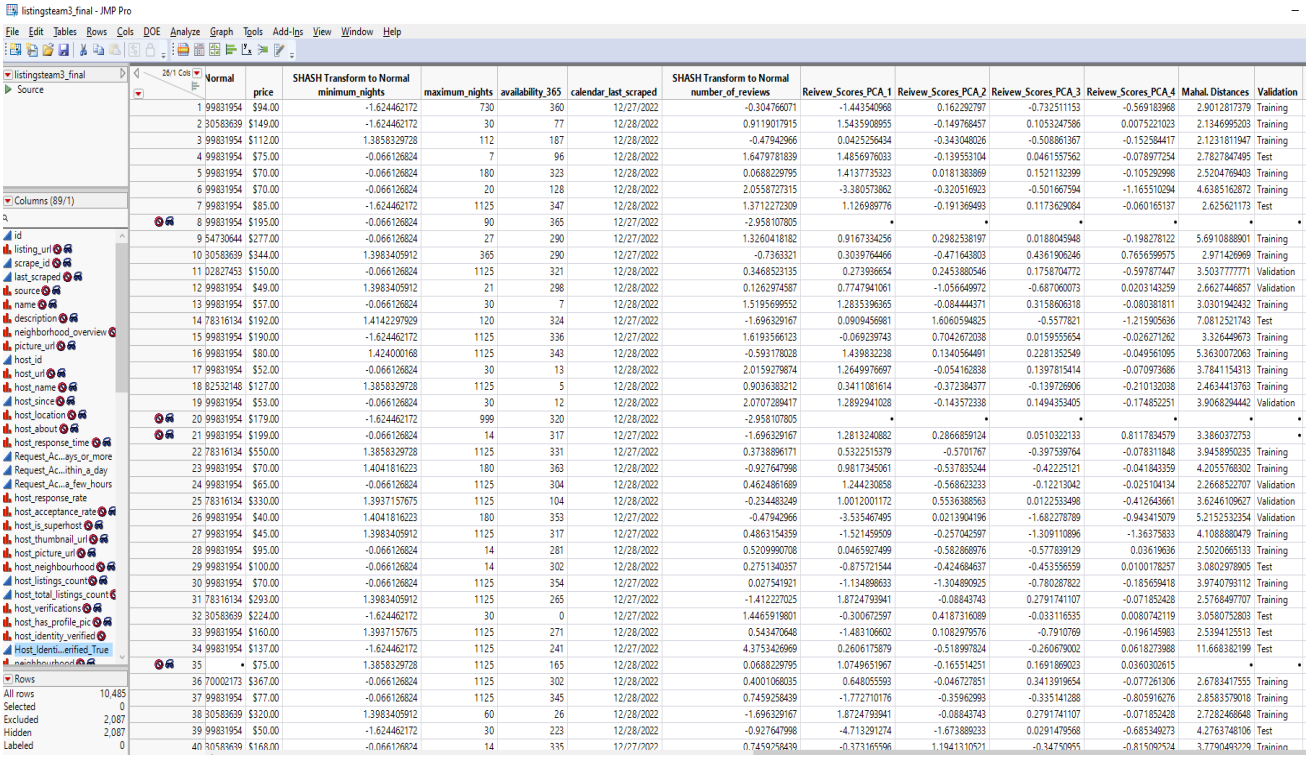
With the refined dataset, data can now be partitioned into 3 sets: Training, Validation, and test data. This will help to build models, cross-check, and validate the resultant model with the validation data set and finally, predict based on the test data.

The following split between the 3 partitions has been executed:

- Training: 60%
- Validation: 20%
- Test: 20%

Figure 18 shows the final data set, which was created upon cleaning the data

Figure 18: Screenshot of the partition done on the pre-processed data set



4. Conclusion

The initial Western Australia data set contained **10485 rows x 75 columns**. Through pre-processing, the data set has been cleansed and reduced to **8398 rows x 26 columns**. The same has been achieved by reducing data dimensionality, excluding the missing values, addressing all N/As in the data set, smoothening the outliers (*transformations*), which will have an impact on determining the target variable, and lastly addressing outliers, if any (*through Mahalanobis / Multivariate, etc.*). Now, the data set is to be modeled and run through the partitions to test on the test data.

Disclaimer: The work contained and presented here is our work and our work alone.

* * * * *