Barcelona Airbnb Listings Data Preprocessing Report

Table of Contents

Contents

2. Column Summary

In project 1, our location was in Barcelona, Spain. We had to clean, remove, and modify the data through the SEMMA process. We had the initial sample with 75 columns and 17,231 rows. We explored and determined which columns were necessary in our analysis based on missing values, correlations, what type of variables (nominal, ordinal, continuous) they were and if they would be a good predictor for our target variable, price. We modified some of the columns to show numerical values to make them continuous that we marked down in the column summary table down below.

We went through each column documenting what columns were included or not and why due to the impact on the predictor variable, price.

Table 1: Column Summary

Col No.	Column name	Activity Chosen	Column Description and Reason
1	ID	Hide and Exclude	This has been disincluded from the dataset because it has no impact on price. The ID is the number given to each row and it is unique to each airbnb listing. While it is a numerical column it is just the unique identifier.
2	Listing_url	Hide and Exclude	The URL has been disincluded from the dataset. The url is a hyperlink to the listing for the airbnb and has no direct impact because it leads to a unique link of the listing.
3	Scrape_id	Hide and Exclude	The Scrape ID has been disincluded from the dataset. The scrape ID identifies when the data was extracted from the internet which would have zero impact on the price. It also has zero impact on the other columns therefore it is irrelevant.
3	Last_scraped	Hide and Exclude	The last Scraped has been disincluded from the dataset. The scrape ID identifies when the data was last extracted from the internet which would have zero impact on the price. It also has zero impact on the other columns therefore it is irrelevant.
5	Source	Hide and Exclude	The source has been disincluded from the dataset. There were two options for the source where the data was extracted from: city or previous scrape. This would have no impact on price because this correlates to how the data was extracted on the internet.
6	Name	Include	The name represents the type of airbnb in barcelona. It describes how many bedrooms, bathrooms, and baths there are. While it is a description and has no direct impact on price, we want to include this in dataset and

			modify it to extract the number of bedrooms into a new column.
7	Description	Hide and Exclude	The description describes where the listing is located as well as an introduction into the surrounding areas. There was no data to extract from this therefore we do not need to include it because there will be no correlations between the description and price.
8	Neighborhood _overview	Hide and Exclude	The neighborhood overview describes what the surrounding areas are and what some different activities are in the close by neighborhoods. While this may be a good description, there can be no data extracted for this column therefore we have decided to exclude it from the dataset.
9	Picture_url	Hide and Exclude	The picture URL is a link to see what each individual apartment looks like. While this (in theory) could be a good indicator for potential prospects to see what they are booking, there can be no sufficient data extracted from it so it is excluded from the dataset.
10	Host_id	Hide and Exclude	The host_id is a unique identifier for the host who is in charge of the rental unit. This was not included in the dataset because the ID is not correlated to anything in the dataset except for indicating the host therefore it will not have an impact on price.
11	Host_url	Hide and Exclude	The host_url is a link to the host for the different units. While it has a description of who they are, it does not have an impact on the price so it will be excluded from the dataset for insufficient data.
12	Host_name	Hide and Exclude	The host name describes the host name of the unit. There would be no correlation between a name and price of the listing therefore it has been excluded from the dataset. There was also no way to extract any data to have reasonable assumptions made.
13	host_since	Hide and Exclude	The host_since describes how long the host has been a host on airbnb renting that unit. This could have some importance, but we have decided to exclude it from the dataset because it wouldn’t matter if the host had been a host for a year or 10 years all that much to have a significant impact on price.
14	host_location	Hide and Exclude	The host_location describes where in Spain the host is located. Since the location of the host does not have an impact on the location of the unit it will not have an impact on price therefore we have decided to exclude it from our dataset.
15	host_about	Hide and Exclude	The host_about is a description of who the host is. While this may be good information to have so that the person who looks at the listing has an idea of who the hosts are, there won't be any sufficient data to pull from there that would be a good predictor of price, so we excluded it from the dataset.

16	host_response_time	Include	The host response time is how responsive the host is after receiving a message from the prospect about the rental unit. This could influence the target variable. It has been analyzed that having a lesser response time would have a higher mean price in most of the neighborhood.
17	host_response_rate	Hide and Exclude	The host response rate is how often the host responds to a message. We have chosen to exclude it because it could not have an impact on price because listings with same response rate might have price difference.Therefore, it won’t have any influence on target variable
18	host_acceptance _rate	Hide and Exclude	The host acceptance rate is how often the host accepts new prospects/customers to rent the airbnb. We have chosen to exclude it from our dataset due 3000 missing records and a high number or high acceptance rate therefore it would not have a significant impact on predicting the price.
19	host_is_superhost	Hide and Exclude	The host is super host is an indicator column of either “T” or “F” which indicates true or false. This means whether the host is the “owner” of the unit. This will not have a sufficient amount of data to support whether this has an impact on price therefore we did not include this in our dataset.
20	host_thumbnail_url	Hide and Exclude	The host thumbnail url is a link to show the picture or thumbnail” of the hosts. There is no data correlated with this therefore we cannot have this as a supportive column for predicting the price, so we excluded it from our dataset.
21	host_picture_url	Hide and Exclude	The host picture url is a link to show the picture of the hosts. There is no data correlated with this therefore we cannot have this as a supportive column for predicting the price, so we excluded it from our dataset.
22	host_neigbourhood	Hide and Exclude	The host neighborhood documents what neighborhood the listing is in. There is not sufficient information to see whether there will be an impact on price therefore we excluded it from our dataset.
23	host_listings_count	Hide and Exclude	The host listing count is the number of listings that each unique host has in spain. The count of these listings would not have a direct impact on the single price of a unit therefore we have excluded it from our dataset.
24	host_total_listings_ count	Hide and Exclude	The total listings count is the sum of all of the listings that each unique host in spain. The count of these listings would not have a direct impact on the single price of a unit therefore we have excluded it from our dataset.
25	host_verificatio ns	Hide and Exclude	The host verifications are the platform that they use to message their potential customers on. The platform would not indicate a difference in price therefore we can exclude it in our dataset.

26	host_has_profile_pic	Hide and Exclude	Host has a profile pic indicating whether the host has a picture attached to their airbnb profile. This is coded as either a "T" or "F" indicating true or false. The picture has no sufficient data to predict the price therefore we have excluded it from the dataset.
27	host_identity_verified	Include	The host identity verified is whether the host has been accurately verified. This can have an impact on price. For instance if a host who is verified in the same location, same amenities, there could be an impact if there is a host who is not verified. Some people may be inclined to pay more for someone who is verified versus not verified.
28	neighborhood	Hide and Exclude	The neighborhood is describing which neighborhood the unit is a part of. This is a generalized value (Barcelona) so it is not sufficient data to be extracted that would be a good predictor for price so we excluded it from our dataset.
29	neighbourhood_cleansed	Include	The neighborhood cleansed is describing which neighborhood the unit is a part of. For different locations there might be differentiations in pricing therefore it could be a good variable for predicting price.
30	neighbourhood_group_cleansed	Hide and Exclude	The neighborhood group cleansed is describing which neighborhood the unit is a part of. There is not sufficient data to be extracted as well as it is the same as neighborhood cleansed therefore we excluded it from our dataset due to redundancy.
31	Latitude	Hide and Exclude	Latitude is the coordinates where the listing is located. Based on the lack of sufficient data that can be extracted from the variable to be a good predictor of price we have decided to exclude it from our dataset.
32	Longitude	Hide and Exclude	Longitude is the coordinates where the listing is located. Based on the lack of sufficient data that can be extracted from the variable to be a good predictor of price we have decided to exclude it from our dataset.
33	property_type	Hide and Exclude	Property type is the type for the listing (private room in rental unit, entire home, etc). Property type is very similar to the room type so we would have redundant information therefore we excluded it from our final dataset.
34	room_type	Include	The room type is defined as what rooms are available for each unique listing. They can be a private room or the entire house/apartment. Because of the different sizes and what is offered at each room, this could be a good predictor in the amount for each unit rented therefore we have included it in our dataset.
35	accommodates	Include	Accommodates are the total amount of people who can be at each rental unit per booking. Because this counts for the total number of people (some with higher amounts or lower amounts) this could have an impact on the price, so we have decided to include this in our dataset.

36	bathrooms	Hide and Exclude	Bathrooms is the total count of all the bathrooms that are included in the unit. There is no data so we cannot use any information for the predictor variable, therefore we excluded it from our dataset.
37	bathroom_texts	Include	The bathroom texts are the type of bathrooms and the count associated in the unit (shared, bath, baths). Since we want to extract bathroom count from this column it is included in the data set
38	bedrooms	Hide and Exclude	The bedrooms are the total number of bedrooms for each individual unit. There are 5894 missing records therefore we excluded it from our dataset. Bedrooms could be a good predictor of price, so we extracted the count of bedrooms from the name column.
39	beds	Include	Beds is the total number of beds in each individual unit. Because the number of beds ranges for each unit that could have an impact on the price. For instance if the number of beds is higher that could indicate a higher price and vice versa so we have decided to include it in our dataset.
40	amenities	Include	The amenities are all the items such as kitchen, wifi, parking, etc that are included in each individual unit. This column would be a good predictor of price therefore we decided to modify and include it in our dataset.We have modified the data and grouped some items together. For modifying it we created a formula where if the unit has the item it labels it as a yes (1) or no (0). We have excluded some amenities items but have included tv, stove, refrigerator, kitchen, oven, wifi, parking, personal hygiene, essentials and clothing storage.
41	price	Include	The price determines the amount per unique listing in Spain. Since this is our predictor variable, in order to determine if the columns that we selected can be a good predictor we need to keep this in our dataset.
42	minimum_nights	Hide and Exclude	Minimum nights is defined as the minimum number of nights that a booking for each individual unit allows. For example if the unit has “4” listed that means that the customers who booked that listing have to stay there for a minimum of 4 days. We have decided to exclude this from our dataset as there was no substantial evidence that it would be a good predictor of price.
43	maximum_nights	Hide and Exclude	The maximum nights is defined as the maximum number of nights that a booking for each individual unit allows. For example if the unit has “30” listed that means that the customers who booked that listing have to stay there for a maximum of 30 days. We have decided to exclude this from our dataset as there was no substantial evidence that it would be a good predictor of price.
44	minimum_minimum_nights	Hide and Exclude	The minimum minimum nights refers to the lowest minimum night value from the calendar for 365 days in the future. The lowest value for the minimum nights would not have an impact on the price because it is

			looking at the lowest minimum value therefore we have excluded it from our dataset.
45	maximum_minimum_nights	Hide and Exclude	The maximum minimum nights refers to the highest minimum night value from the calendar for 365 days in the future. The highest value for the minimum nights would not have an impact on the price because it is just looking at the highest minimum value therefore we have excluded it from our dataset.
46	minimum_maximum_nights	Hide and Exclude	The minimum maximum nights refers to the lowest maximum night value from the calendar for 365 days in the future. The lowest value for the maximum nights would not have an impact on the price because it is just looking at the lowest maximum value therefore we have excluded it from our dataset.
47	maximum_maximum_nights	Hide and Exclude	The maximum maximum nights refers to the highest maximum night value from the calendar for 365 days in the future. The highest value for the maximum nights would not have an impact on the price because it is just looking at the highest maximum value therefore we have excluded it from our dataset.
48	minimum_nights_avg_ntm	Hide and Exclude	The minimum nights average ntm refers to the average minimum_night value from the calendar for 356 in the future from when the customer selects it on the calendar. The average for the minimum nights would not have a substantial impact on price therefore we excluded it from the dataset.
49	maximum_nights_avg_ntm	Hide and Exclude	The maximum nights average ntm refers to the average maximum_night value from the calendar for 365 nights in the future from when the customer selects it on the calendar. The average for the maximum nights would not have a substantial impact on price therefore we excluded it from the dataset.
50	calendar_update d	Hide and Exclude	Calendar updated refers to if the host has their calendar updated on the listing. Although this might've been a good column to predict price there was no information in this column therefore we have to exclude it from our dataset.
51	has_availability	Hide and Exclude	Has availability refers to if the host of the listing has availability during "specific chosen times" that the customer selects. It will either show up as a "t" or "f". Just because the listing is available or not doesn't mean it will have an impact on price therefore we have excluded it from our dataset.
52	availability_30	Include	Availability 30 refers to the availability of the listing 30 days in the future as determined by the calendar. For example if the date is 12/5, it will show days 30 days after 12/5. We have decided to include this because the future availability would show us the demand for the units, therefore it would be a good predictor of price.

53	availability_60	Hide and Exclude	Availability 60 refers to the availability of the listing 60 days in the future as determined by the calendar. For example if the date is 12/5, it will show days 60 days after 12/5. We have decided to exclude this from our dataset due to strong correlation and redundancy with the availability_30 column.
54	availability_90	Hide and Exclude	Availability 60 refers to the availability of the listing 90 days in the future as determined by the calendar. For example if the date is 12/5, it will show days 90 days after 12/5. We have decided to exclude this from our dataset due to strong correlation and redundancy from the availability_30 column.
55	availability_365	Hide and Exclude	Availability 365 refers to the availability of the listing 365 days in the future as determined by the calendar. For example if the date is 12/5, it will show days 365 days after 12/5. We have decided to exclude this from our dataset due to strong correlation and redundancy from the availability_30 column.
56	calendar_last_scrapped	Hide and Exclude	Calendar last scrapped identifies when the calendar data was last extracted from the internet which would have zero impact on the price. It also has zero impact on the other columns therefore it is irrelevant.
57	number_of_reviews	Hide and Exclude	Number of reviews refers to the sum of the total number of reviews that each unique listing has. This variable would not have a substantial impact on predicting price because it is only looking at the number of reviews and not what kind of reviews it entails (good, bad, valid) therefore we have excluded it from our dataset.
58	number_of_reviews_ltm	Hide and Exclude	Number of reviews 1tm refers to the sum of the total number of reviews that each unique listing has up to 12 months. This variable would not have a substantial impact on predicting price because it is only looking at the number of reviews and not what kind of reviews it entails (good, bad, valid) therefore we have excluded it from our dataset.
59	number_of_reviews_l30d	Hide and Exclude	Number of reviews 130d (130 days) refers to the sum of the total number of reviews that each unique listing has up to 130 days. This variable would not have a substantial impact on predicting price because it is only looking at the number of reviews and not what kind of reviews it entails (good, bad, valid) therefore we have excluded it from our dataset.
60	first_review	Hide and Exclude	First review refers to the date when the first review for each unique listing for each unit. The date would have little to no impact on the prediction of price because it refers to the date and has no substantial data to support the price therefore we have excluded it from our dataset.
61	last_review	Hide and Exclude	Last review refers to the date when the first review for each unique listing for each unit. The date would have little to no impact on the prediction of price because it refers to the date and has no substantial data to support the price therefore we have excluded it from our dataset.

62	review_score_ra ting	Include	The review score rating refers to the rating (0-5) determined by the reviews for each unique unit listing. The average rating could have an impact on the price. For instance, a listing with a higher rating could mean that the price could be higher and vice versa for a lower rating. Because of this potential predictor or pierce we have included it in our dataset.
63	review_score_accuracy	Hide and Exclude	The review score accuracy refers to the accuracy of the review scores that each unit listing has. The accuracy of the review scores would not have a substantial impact on the predictor variable price because the hosts are not looking at the accuracy of the rating, just the actual ratings for their units therefore we have excluded it from the dataset.
64	review_score_cleanliness	Hide and Exclude	The review score cleanliness refers to the cleanliness(of the unit) of the review scores that each unit listing has. The cleanliness of the review scores would not have a substantial impact on the predictor variable price because the hosts are not looking at the cleanliness of the rating, just the actual ratings for their units therefore we have excluded it from the dataset.
65	review_score_checkin	Hide and Exclude	The review score checking refers to how the check-in is rated for each unit that the listing has. The check-in of the review scores would not have a substantial impact on the predictor variable price because the price is not going to change based on the check in therefore we have excluded it from the dataset.
66	review_score_communication	Hide and Exclude	The review score communication refers to how the communication between the host and the customer is rated for each unit that the listing has. The communication of the review scores would not have a substantial impact on the predictor variable price because the price is not going to change based on the type/how everything is communicated therefore we have excluded it from the dataset.
67	review_score_location	Hide and Exclude	The review score location refers to how the location was rated for each unit that the listing has. The location of the review scores would not have a substantial impact on the predictor variable price because the price is not going to change a significant amount based on the location, therefore we have excluded it from the dataset.
68	review_score_v alue	Hide and Exclude	The review score value refers to the overall “value” of each unique listing. The overall value would not have a significant impact on price therefore we have excluded it from our dataset.
69	license	Hide and Exlude	License indicates the license/permit/registration number. The license column is hidden and excluded because it has unique values and most of the data is missing.

70	Instant_bookable	Include	Instant bookable indicates whether the guest can instantly book the listing. this data is an indicator of a commercial listing. Therefore it is included because the instant bookable has an influence on our target variable.
71	Calculated_host_listings_count	Hide and Exclude	The host listings count indicates the number of listings the host has in the current scrape, in the city/region geography. This column is hidden and excluded because the host's number of listings in the current scrape doesn't influence our target variable.
72	Calculated_host_listings_count_entire_homes	Hide and Exclude	This host listings count entire homes indicates the number of entire home/apartment listings that the host has in the current scrape based across the city/region geography. This column is hidden and excluded because the number of home/apartment that a host has listed doesn't influence our target variable.
73	Calculated_host_listings_count_private_rooms	Hide and Exclude	The host listings count private rooms indicates the number of Entire home/apt listings the host has in the current scrape, in the city/region geography. This column is hidden and excluded because the number of private rooms a host has listed doesn't influence our target variable.
74	Calculated_host_listings_count_shared_rooms	Hide and Exclude	The host listings count shared rooms indicates the number of Shared room listings the host has in the current scrape, in the city/region geography This column is hidden and excluded because the number of shared rooms a host has listed doesn't influence our target variable.
75	Reviews_per_month	Hide and Exclude	Reviews per month is the number of reviews per month divided by 100. This specific number would not have a substantial impact to predict price as we are more concerned with looking at the overall ratings therefore we have excluded it from our dataset.

3. Data Preprocessing

In order to build a model, the dataset ‘Barcelona Airbnb’ was preprocessed by using a range of data preprocessing techniques to clean the data and reduce the complexity of the dataset.

The following steps and techniques were used:

3.1. Hide and exclude
3.2. Columns Modification
3.3. Missing Values
3.4. Outliers

4.1. Hide and exclude

All columns were reviewed as mentioned in Section 1 Column Summary to ensure well-defined and appropriate columns were identified and included in the data set. All columns that were not relevant or would not help with the prediction of the model were hidden and excluded.

A total of 63 columns were hidden and excluded from the data set. 12 columns were then further worked upon (modified) or remained constant in the next section.

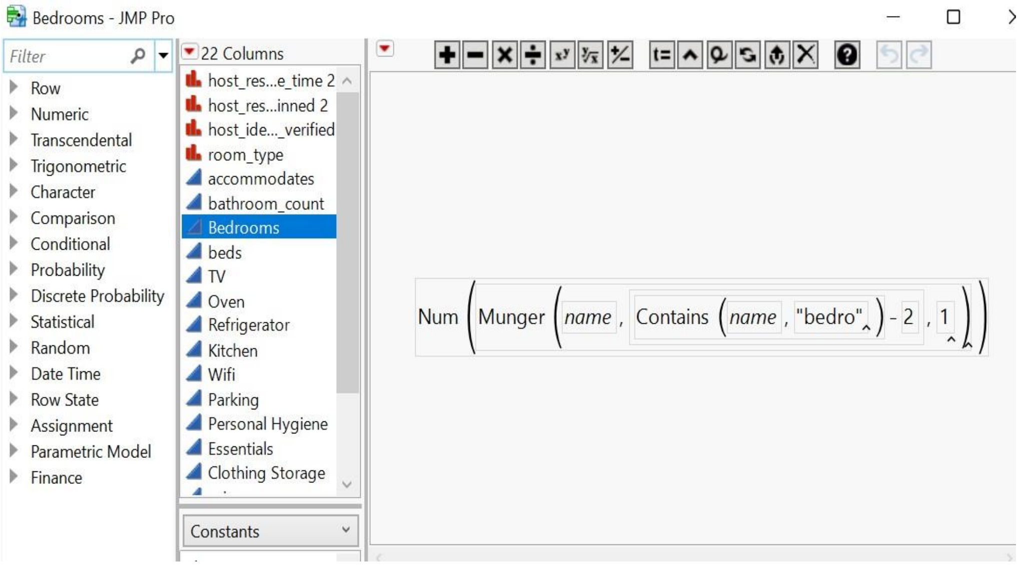
- Ø name
- Ø host_response_time
- Ø host_identity_verified
- Ø neighbourhood_cleansed
- Ø room_type
- Ø accommodates
- Ø bathrooms_text
- Ø beds
- Ø amenities
- Ø availability_30
- Ø review_scores_rating
- Ø instant_bookable

Note: name,bathrooms_text and amenities are included for now but later are modified into new columns

4.2. Columns Modification

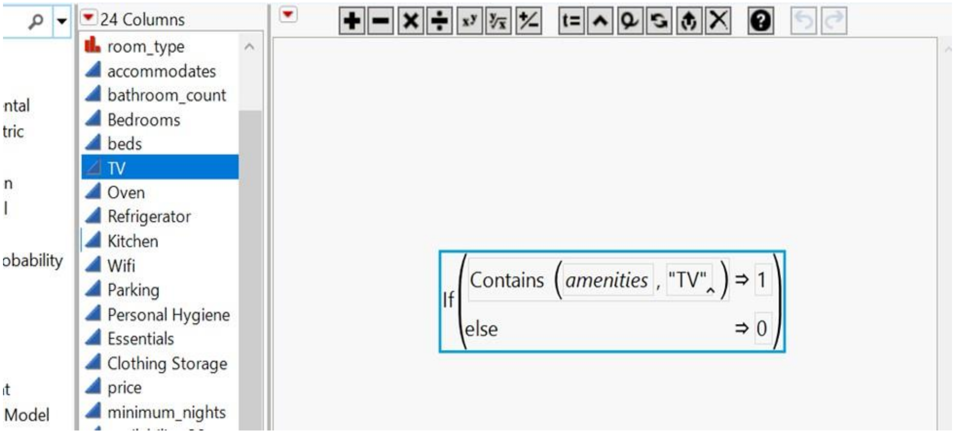
4.2.1 Formula Columns:

A. **Bedrooms:** Recode Formula - Since the original 'bedrooms' column has '5894' missing values we decided to extract the bedrooms count from the 'name ' column . We inputted a formula telling jmp to select the total number of bedrooms for each row and create a new ‘ Bedrooms’ column. The formula used to achieve this task is attached below.

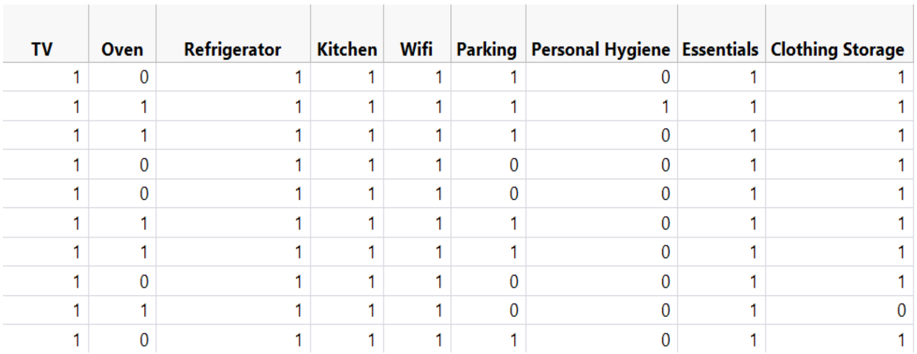


```
Num(Munger(name, Contains(name, "bedro") - 2, 1))
```

B. Amenities as Indicator Columns - In order to extract data useful information from the amenities column for price prediction, we had to create indicator columns by creating a unique formula for each column which “indicated” if the column contained that grouped or ungrouped item. For example in the column labeled “TV”, the rows that had a 1 contained the values “TV” in each row. The rows that had a 0 did not contain any values with “TV” in them. The screenshot of the formula used to achieve this task is attached below.



```
If(Contains(amenities, "TV") ⇒ 1
   else ⇒ 0)
```

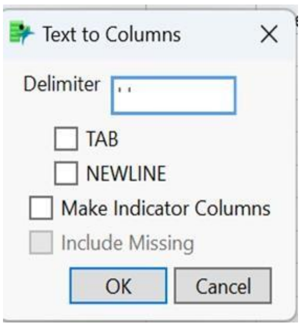


TV	Oven	Refrigerator	Kitchen	Wifi	Parking	Personal Hygiene	Essentials	Clothing Storage
1	0	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	0	1	1
1	0	1	1	1	0	0	1	1
1	0	1	1	1	0	0	1	1
1	1	1	1	1	1	0	1	1
1	1	1	1	1	1	0	1	1
1	0	1	1	1	0	0	1	1
1	1	1	1	1	0	0	1	0
1	0	1	1	1	1	0	1	1

4.2.2 Recoded Columns

A. Bathrooms: Since the original ‘bathrooms’ column has no data to be used, we decided to extract bathroom count information from the ‘bathrooms_text’ column and create a new ‘Bathrooms_count’ column.. We used the ‘Text to Column’ utility to achieve this task.

Delimiter used: ‘ ‘



Text to Columns

Delimiter ' '

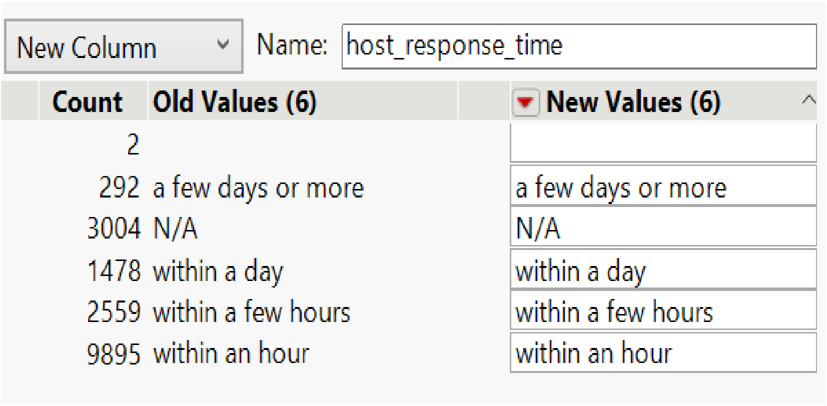
□ TAB

□ NEWLINE

□ Make Indicator Columns

□ Include Missing

B. host_response_time: There are 3004 ‘N/A’ values in the ‘host_response_time’ column, which are recoded into missing values. ‘Host_response_time’ is the newly recoded column.



New Column Name: host_response_time

Count	Old Values (6)	New Values (6)
2		
292	a few days or more	a few days or more
3004	N/A	N/A
1478	within a day	within a day
2559	within a few hours	within a few hours
9895	within an hour	within an hour

After performing these modifications now, we have 17230 rows and 21 columns in total and worked upon in later sections.

list of the columns:

Ø Host_response_time

Ø host_identity_verified

Ø neighbourhood_cleansed

Ø room_type

Ø accommodates

Ø Bathrooms_count

Ø Bedrooms

Ø beds

Ø TV

Ø Oven

Ø Refrigerator

Ø Kitchen

Ø Wifi

Ø Parking

Ø Personal Hygiene

Ø Essentials

Ø Clothing Storage

Ø price

Ø availability_30

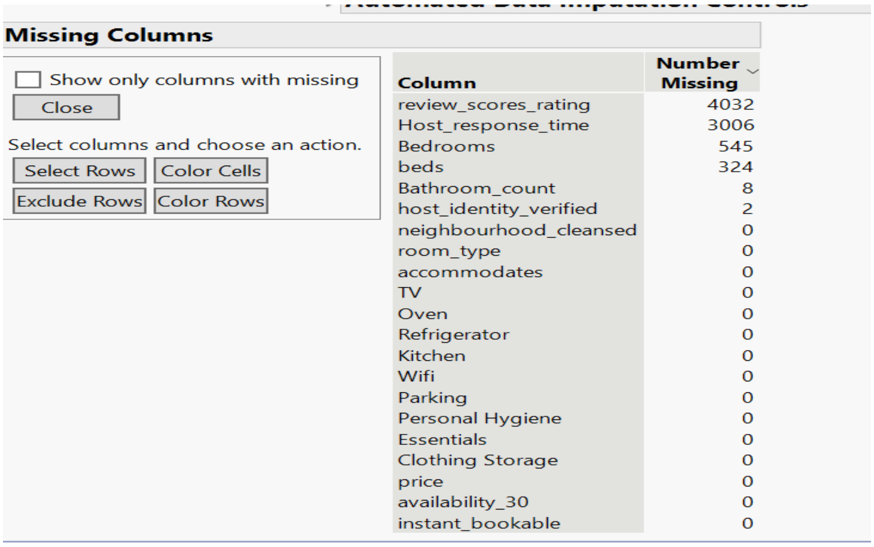
Ø review_scores_rating

Ø instant_bookable

4.3. Missing Values

4.3.1 Explore Missing Values

We attached the screenshot showing missing records in each column after performing the ‘Explore Missing Values’ analysis.



Missing Columns

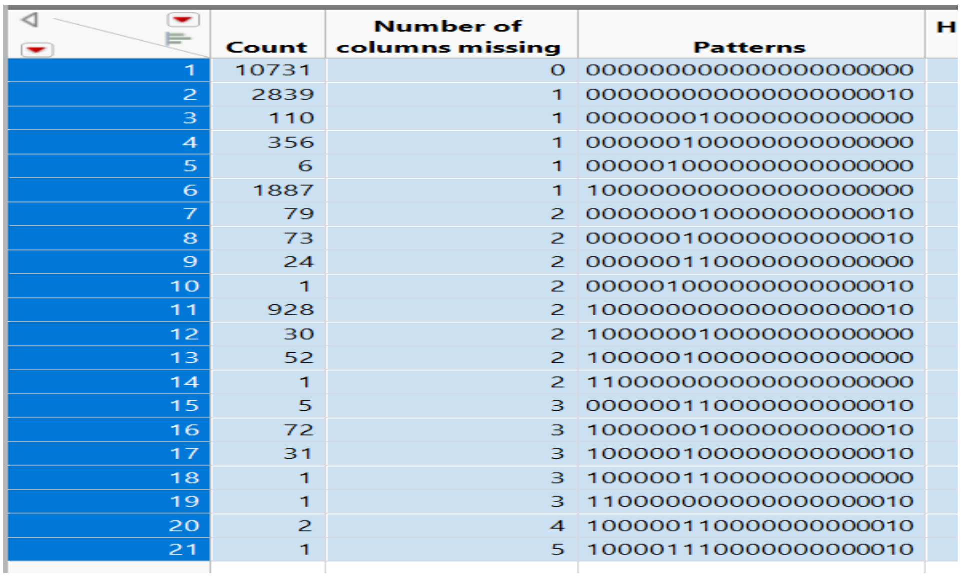
□ Show only columns with missing

Column	Number Missing
review_scores_rating	4032
Host_response_time	3006
Bedrooms	545
beds	324
Bathroom_count	8
host_identity_verified	2
neighbourhood_cleansed	0
room_type	0
accommodates	0
TV	0
Oven	0
Refrigerator	0
Kitchen	0
Wifi	0
Parking	0
Personal Hygiene	0
Essentials	0
Clothing Storage	0
price	0
availability_30	0
instant_bookable	0

From the above table we can see that there are a considerable number of missing values in review_scores_rating and Host_response_time columns. Since, Imputation of huge data in a single column doesn’t help to build an accurate model, we tried to get the missing data pattern and exclude the records with missing variables.

4.3.2 Missing Data Pattern

The below screenshot shows the count and pattern of missing data in all the records. The highlighted records are excluded and hidden from the data set.

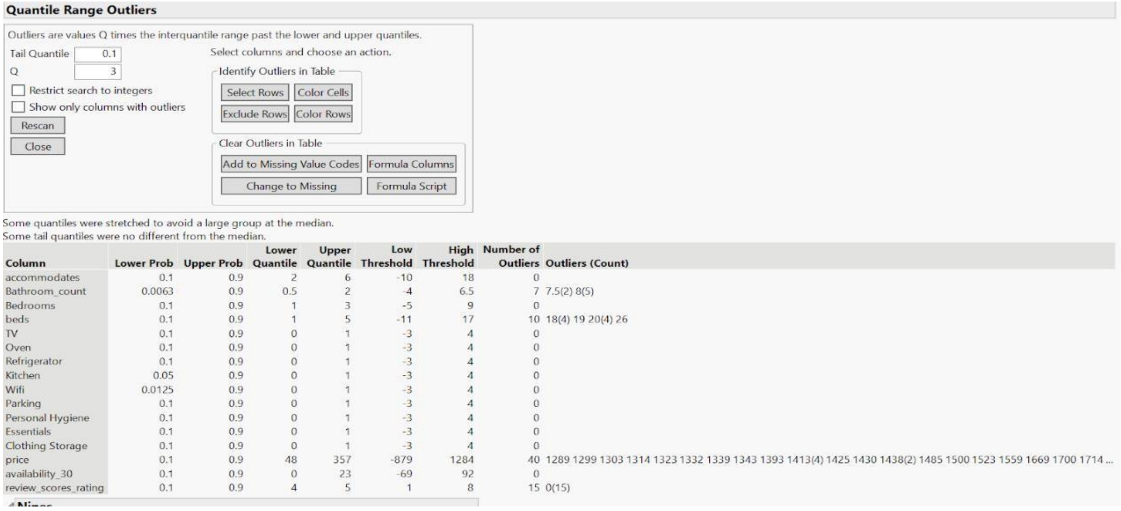


	Count	Number of columns missing	Patterns	H
1	10731	0	000000000000000000000	
2	2839	1	000000000000000000010	
3	110	1	000000010000000000000	
4	356	1	000000100000000000000	
5	6	1	000001000000000000000	
6	1887	1	100000000000000000000	
7	79	2	000000010000000000010	
8	73	2	000000100000000000010	
9	24	2	000000110000000000000	
10	1	2	000001000000000000010	
11	928	2	100000000000000000010	
12	30	2	100000010000000000000	
13	52	2	100000100000000000000	
14	1	2	110000000000000000000	
15	5	3	000000110000000000010	
16	72	3	100000010000000000010	
17	31	3	100000100000000000010	
18	1	3	100000110000000000000	
19	1	3	110000000000000000010	
20	2	4	100000110000000000010	
21	1	5	100001110000000000010	

In order to enhance the data integrity and analytical robustness of the report, a comprehensive approach was undertaken to systematically identify and eliminate missing values from the dataset.

4.4. Outliers

4.4.1 Univariate Outliers: Below screenshot shows the summary of univariate outliers in each column after performing ‘Explore Outliers’ analysis.



Quantile Range Outliers

Outliers are values Q times the interquantile range past the lower and upper quantiles.

Tail Quantile: 0.1

Q: 3

☐ Restrict search to integers

☐ Show only columns with outliers

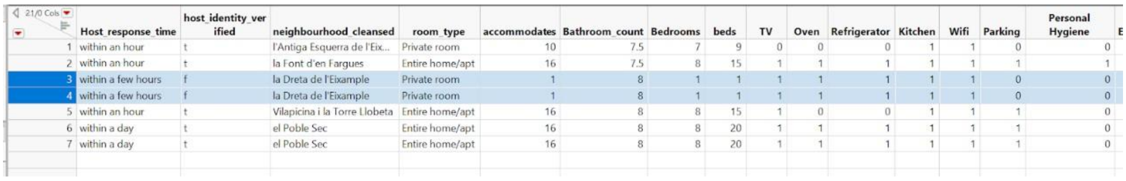
Select columns and choose an action.

Some quantiles were stretched to avoid a large group at the median.
Some tail quantiles were no different from the median.

Column	Lower Prob	Upper Prob	Lower Quantile	Upper Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
accommodates	0.1	0.9	2	6	-10	18	0	
Bathroom_count	0.0063	0.9	0.5	2	-4	6.5	7	7.5(2) 8(5)
Bedrooms	0.1	0.9	1	3	-5	9	0	
beds	0.1	0.9	1	5	-11	17	10	18(4) 19 20(4) 26
TV	0.1	0.9	0	1	-3	4	0	
Oven	0.1	0.9	0	1	-3	4	0	
Refrigerator	0.1	0.9	0	1	-3	4	0	
Kitchen	0.05	0.9	0	1	-3	4	0	
Wifi	0.0125	0.9	0	1	-3	4	0	
Parking	0.1	0.9	0	1	-3	4	0	
Personal Hygiene	0.1	0.9	0	1	-3	4	0	
Essentials	0.1	0.9	0	1	-3	4	0	
Clothing Storage	0.1	0.9	0	1	-3	4	0	
price	0.1	0.9	48	357	-879	1284	40	1289 1299 1303 1314 1323 1332 1339 1343 1393 1413(4) 1425 1430 1438(2) 1485 1500 1523 1559 1669 1700 1714 ...
availability_30	0.1	0.9	0	23	-69	92	0	
review_scores_rating	0.1	0.9	4	5	1	8	15	0(15)

The outliers in Bathroom_count,Bedrooms,price and review_scores_rating columns are identified and taken care of in the below sections.

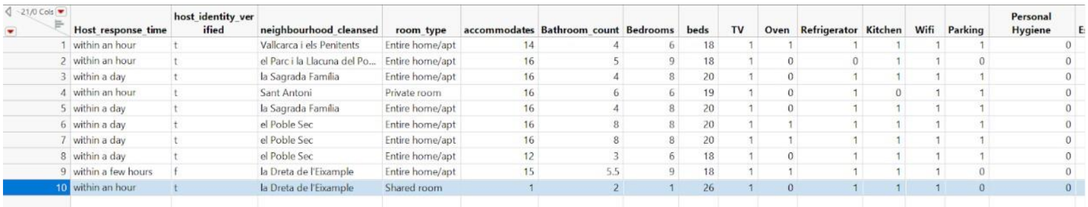
A. Bathroom_count:



	Host_response_time	host_identity_verified	neighbourhood_cleansed	room_type	accommodates	Bathroom_count	Bedrooms	beds	TV	Oven	Refrigerator	Kitchen	Wifi	Parking	Personal Hygiene	E
1	within an hour	t	l'Antiga Esquerra de l'Eix...	Private room	10	7.5	7	9	0	0	0	1	1	0	0	
2	within an hour	t	la Font d'en Fargues	Entire home/apt	16	7.5	8	15	1	1	1	1	1	1	1	
3	within a few hours	f	la Dreta de l'Eixample	Private room	1	8	1	1	1	1	1	1	1	0	0	
4	within a few hours	f	la Dreta de l'Eixample	Private room	1	8	1	1	1	1	1	1	1	0	0	
5	within an hour	t	Vilapicina i la Torre Llobeta	Entire home/apt	16	8	8	15	1	0	0	1	1	1	0	
6	within a day	t	el Poble Sec	Entire home/apt	16	8	8	20	1	1	1	1	1	1	0	
7	within a day	t	el Poble Sec	Entire home/apt	16	8	8	20	1	1	1	1	1	1	0	

From the above screenshot we can see that there are 7 outliers in the column. After analyzing, we identified Bathroom_count=8 as a potential outlier and decided to exclude the highlighted 2 records in the screenshot because it’s not practically possible for a room with 1 bedroom and can accommodate only 1 person to have 8 bathrooms. Hence excluded the highlighted 2 rows in the above screen shot.

B. Beds:



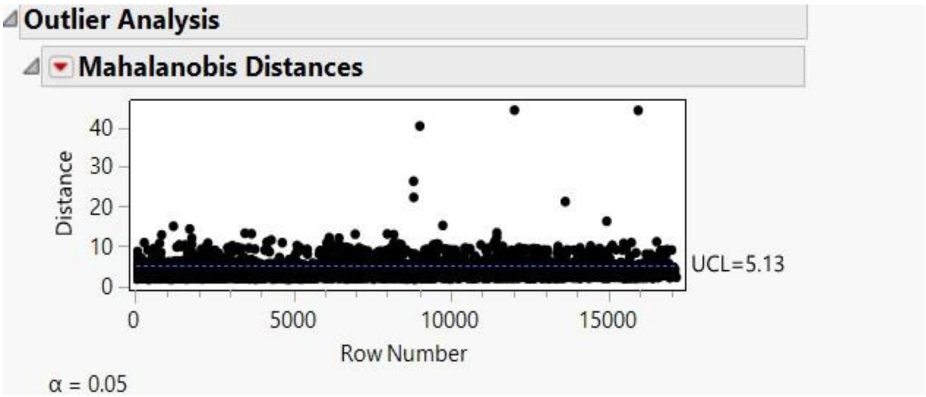
	Host_response_time	host_identity_verified	neighbourhood_cleansed	room_type	accommodates	Bathroom_count	Bedrooms	beds	TV	Oven	Refrigerator	Kitchen	Wifi	Parking	Personal Hygiene	E
1	within an hour	t	Vallcarca i els Penitents	Entire home/apt	14	4	6	18	1	1	1	1	1	1	0	
2	within an hour	t	el Parc i la Llacuna del Po...	Entire home/apt	16	5	9	18	1	0	0	1	1	0	0	
3	within a day	t	la Sagrada Família	Entire home/apt	16	4	8	20	1	0	1	1	1	1	0	
4	within an hour	t	Sant Antoni	Private room	16	6	6	19	1	0	1	0	1	1	0	
5	within a day	t	la Sagrada Família	Entire home/apt	16	4	8	20	1	0	1	1	1	1	0	
6	within a day	t	el Poble Sec	Entire home/apt	16	8	8	20	1	1	1	1	1	1	0	
7	within a day	t	el Poble Sec	Entire home/apt	16	8	8	20	1	1	1	1	1	1	0	
8	within a day	t	el Poble Sec	Entire home/apt	12	3	6	18	1	0	1	1	1	1	0	
9	within a few hours	f	la Dreta de l'Eixample	Entire home/apt	15	5.5	9	18	1	1	1	1	1	0	0	
10	within an hour	t	la Dreta de l'Eixample	Shared room	1	2	1	26	1	0	1	1	1	0	0	

There are 10 outliers in beds column. After analyzing, we identified 26 beds as a potential outlier and decided to exclude the highlighted record in the screenshot because it’s not practically possible for a room with 1 bedroom and can accommodate only 1 person to have 26 beds. Hence excluded the highlighted 2 rows in the above screens shot.

C. **Price:** Price variable values are not considered as outliers because it is our target variable where we consider the given value of the data according to the other requirements of the variables.

D. **review_score_rating:** The outliers shown in the screenshot 2.4.1 are considered because the review score rating of those 15 rows is 0 which states that there are no reviews for those 15 Airbnb’s but they can be priced considering the other features.

4.4.2 Multivariate outliers



There are a lot of multivariate outliers that are identified which are above UCL but we decided to exclude only the ones that seem to be too extreme. Hence, we excluded 3 records.

4. Conclusion

The initial Barcelona, Catalonia, Spain Airbnb data set contains 17230 rows x 75 columns. Through data preprocessing, we excluded 6505 rows and 63 columns, the data set has been cleansed and reduced to 10725 rows x 21 columns.

The same has been achieved by reducing data dimensionality, excluding the missing values, detaching and smoothening the outliers, which will have an impact on the target variable, and lastly addressing outliers.

In conclusion, the handling of missing values, outliers, and data normalization were all successfully addressed throughout the data preprocessing step. These crucial stages have established a strong framework for the data analysis that will follow, guaranteeing the quality and dependability of the conclusions drawn from the tidy and organized information.