OPIM 5604 - Predictive Modeling

Project 1

Team 8 - Ayushi Bisht, Rohit Kamineni, Ryan Nelson, Satakshi Deshmukh, Sumit Pal, Supraja Bala

Summary:

Our approach involved several key steps. Initially, we examined the columns and their values to identify any meaningful patterns. We also transformed categorical variables into continuous ones and introduced new columns to enhance the dataset. Subsequently, we conducted a thorough assessment for missing values, removing columns with a high volume of missing data. To address these missing values, we employed an imputation method based on averages. We further identified and eliminated outliers from the dataset to enhance data quality. Finally, we performed Principal Component Analysis (PCA) on the refined dataset, allowing us to extract meaningful information while reducing data complexity and the number of columns significantly.

Columns and Explanations:

Id: Excluded as IDs are unique identifiers and are irrelevant to price.

listing_url: Excluded as each URL is unique and doesn’t contribute as a relevant factor in affecting price.

scrape_id: Excluded as IDs are unique identifiers and are irrelevant to price.

last_scraped: This column provides information that is once again irrelevant to the price. Hence, it’s excluded.

source: This column contains information that is similar to the last_scraped column. It’s information that can’t be used to predict price. Hence, this column is also excluded.

name: This column contains information that doesn’t follow a certain pattern or standard. It’s very vague information and can’t be used to evaluate price. Hence, it’s excluded.

description: A very subjective information that is almost unique as it doesn't follow any established pattern or standard. This information column cannot be quantified and had to be excluded.

neighborhood_overview: Very similar scenario to the description column. Not quantifiable and excluded.

picture_url: Excluded as each URL is unique and doesn't contribute as a relevant factor in affecting price.

host_id: Excluded as IDs are unique identifiers and are irrelevant to price.

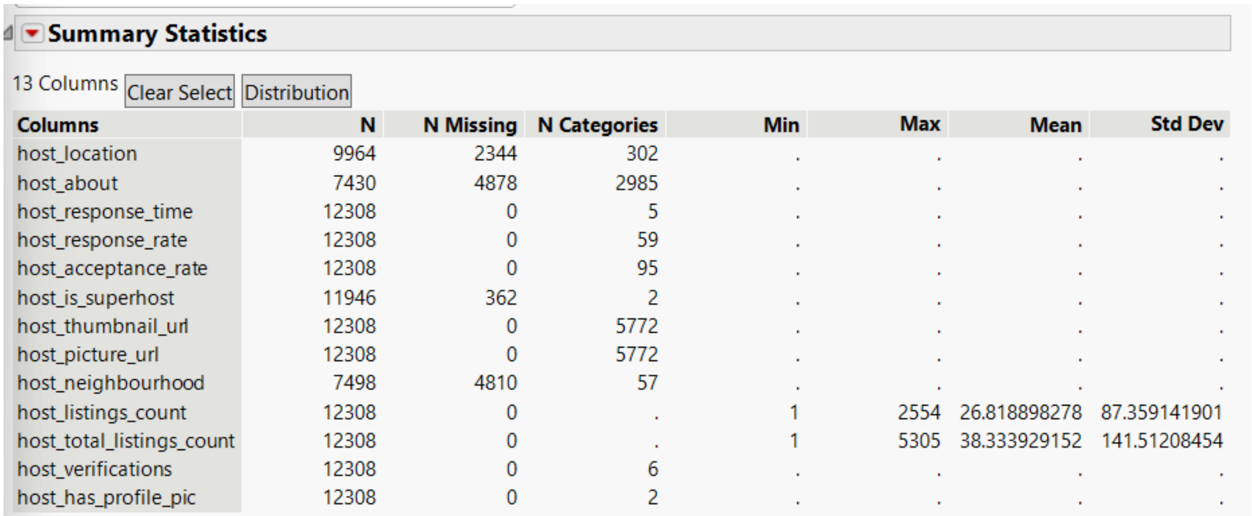
host_url: Excluded as each URL is unique and doesn't contribute as a relevant factor in affecting price. Excluded as each URL is unique and doesn't contribute as a relevant factor in affecting price. There are URLs repeating multiple times when a host has multiple listings, but it still remains irrelevant to price.

host_name: Excluded as a name is a unique identifier and isn't relevant to predicting price.

host_since: This column provides information that is irrelevant to the price. Hence, it's excluded.

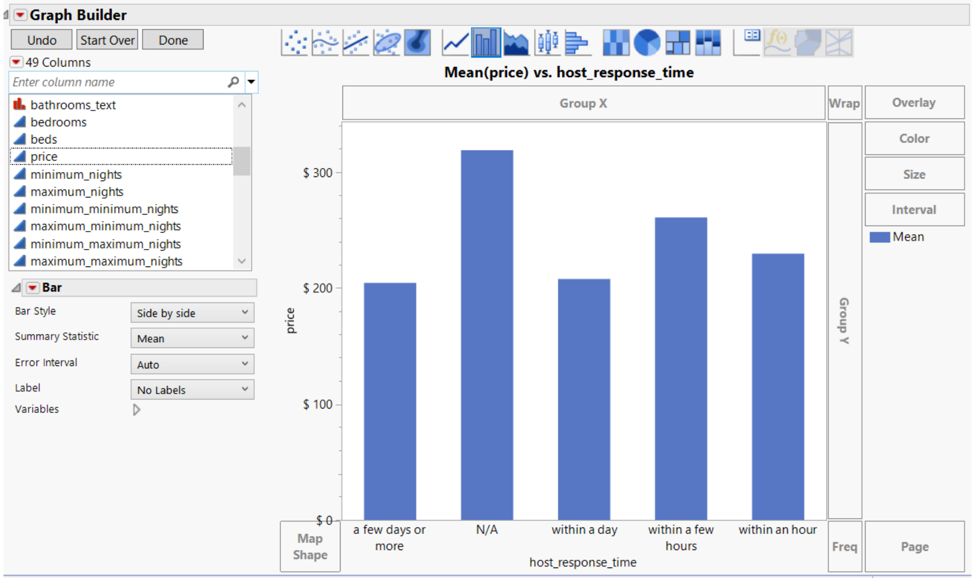
Host_location: Excluded the column as the location of the host does not relate with the price variable. Also it has 2344 missing values.

Host_about: Excluded as it's not related to price. Host_about has 4810 missing values which can not be imputed as the data is categorical.



Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev
host_location	9964	2344	302	.	.	.	.
host_about	7430	4878	2985	.	.	.	.
host_response_time	12308	0	5	.	.	.	.
host_response_rate	12308	0	59	.	.	.	.
host_acceptance_rate	12308	0	95	.	.	.	.
host_is_superhost	11946	362	2	.	.	.	.
host_thumbnail_url	12308	0	5772	.	.	.	.
host_picture_url	12308	0	5772	.	.	.	.
host_neighbourhood	7498	4810	57	.	.	.	.
host_listings_count	12308	0	.	1	2554	26.818898278	87.359141901
host_total_listings_count	12308	0	.	1	5305	38.333929152	141.51208454
host_verifications	12308	0	6	.	.	.	.
host_has_profile_pic	12308	0	2	.	.	.	.

Host_response_time: Kept as depending on how quick a host responds on the inquiries or bookings from a potential guest.



Bar

Bar Style: Side by side

Summary Statistic: Mean

Error Interval: Auto

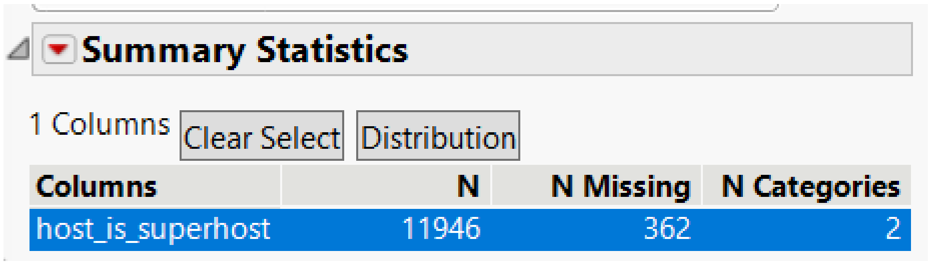
Label: No Labels

Variables

Host_response_rate: Keeping this column as higher response rates might command different price levels based on different hosts.

Host_acceptance_rate: Keeping this column as higher acceptance rates

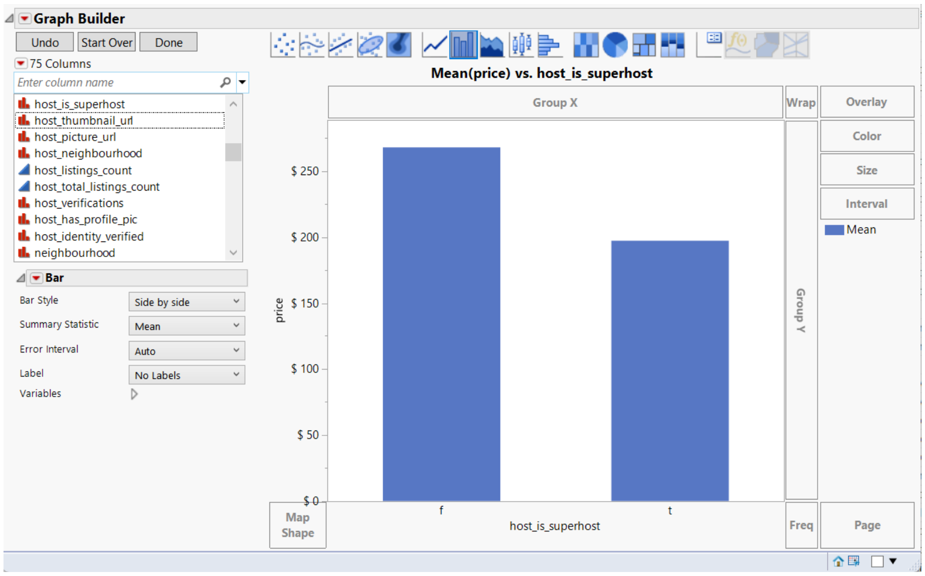
Host_is_superhost: This column is directly related to the price. We could add an indicator column to convert this categorical value to a continuous in order to analyze. This column has only 362 missing values which is about 3%. So we can exclude and hide those rows. Imputing the data on superhost is not ideal as it can directly affect the prices variable.



Summary Statistics

1 Columns

Columns	N	N Missing	N Categories
host_is_superhost	11946	362	2



Graph Builder

75 Columns

- 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
- neighbourhood

Bar

Bar Style: Side by side

Summary Statistic: Mean

Error Interval: Auto

Label: No Labels

Variables

Is_superhost: We have created this new indicator column which represents true for the host is superhost.

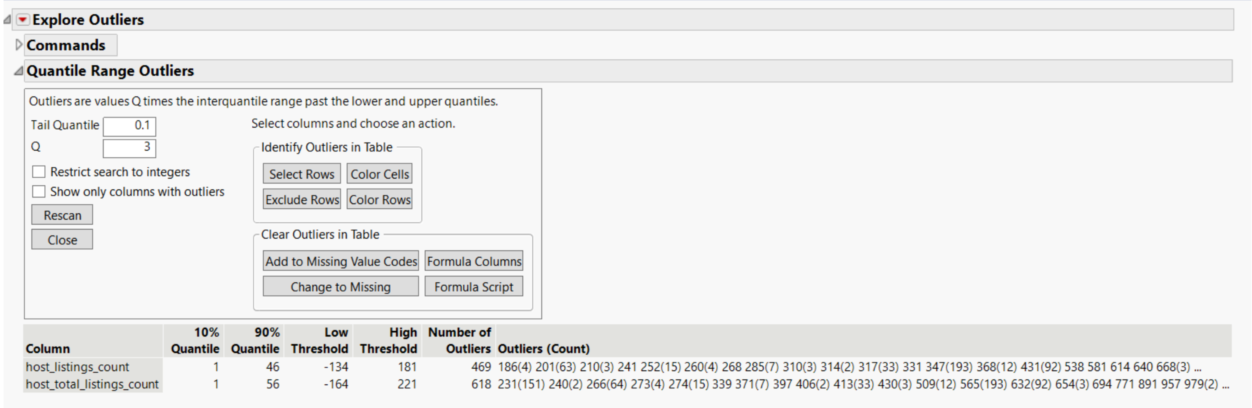
Host_thumbnail_url: Excluding as it is not related to the price.

Host_picture_url: Excluding as it is not related to the price.

Host_neighbourhood: Excluding as it is not related to the price.

Host_listing_count: Keep the column as the number of listings can affect the price variable.

Host_total_listing_count: Keep the column as the total number of listings have relation with the price variable.



Explore Outliers

Commands

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

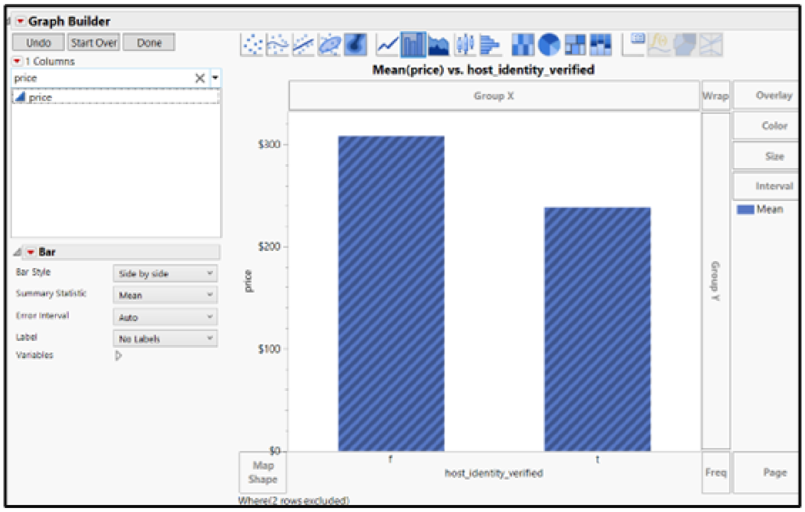
Select columns and choose an action.

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
host_listings_count	1	46	-134	181	469	186(4) 201(63) 210(3) 241 252(15) 260(4) 268 285(7) 310(3) 314(2) 317(33) 331 347(193) 368(12) 431(92) 538 581 614 640 668(3) ...
host_total_listings_count	1	56	-164	221	618	231(151) 240(2) 266(64) 273(4) 274(15) 339 371(7) 397 406(2) 413(33) 430(3) 509(12) 565(193) 632(92) 654(3) 694 771 891 957 979(2) ...

Host_verifications: Not relevant to the price. It does not affect the price variable.

Host_has_profile pic: Excluding as it is not related to the price.

Host_identity_verified: No modification as it has direct relation with price.



Bar

Bar Style: Side by side

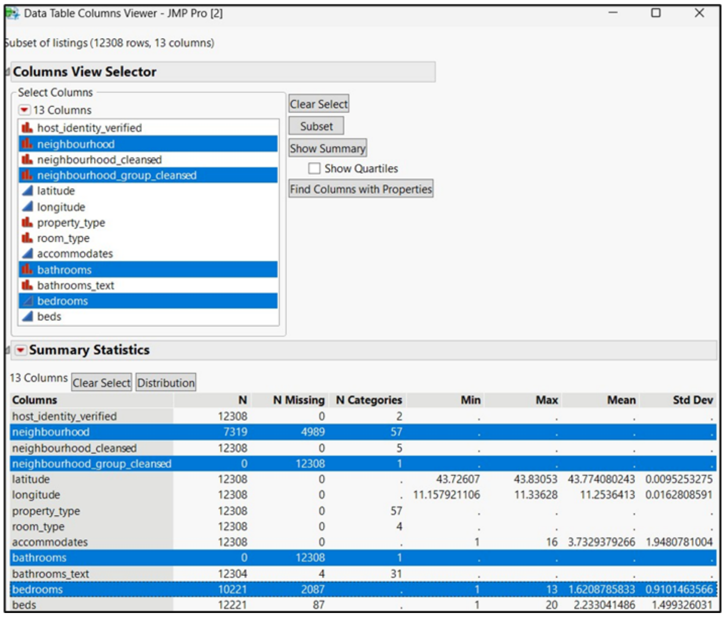
Summary Statistic: Mean

Error Interval: Auto

Label: No Labels

Variables

Neighborhood: Excluded this column as it has 4989 missing values which cannot be imputed as data is categorical.



Data Table Columns Viewer - JMP Pro [2]

Subset of listings (12308 rows, 13 columns)

Columns View Selector

Select Columns

13 Columns

- host_identity_verified
- neighbourhood
- neighbourhood_cleansed
- neighbourhood_group_cleansed
- latitude
- longitude
- property_type
- room_type
- accommodates
- bathrooms
- bathrooms_text
- bedrooms
- beds

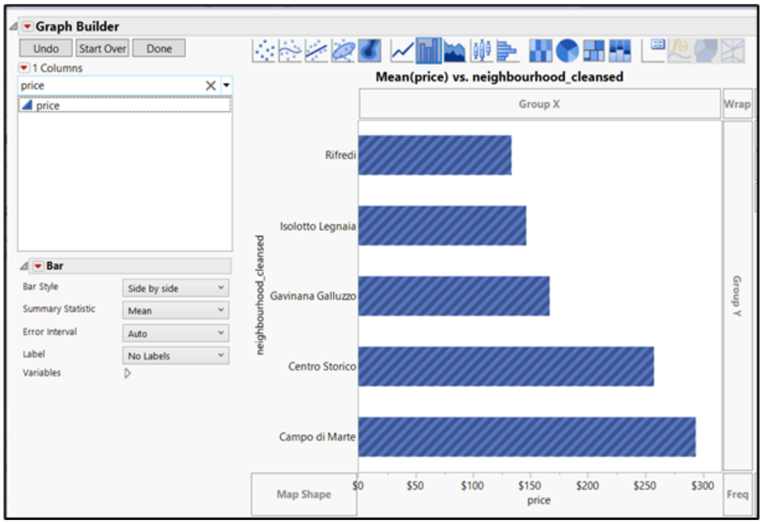
□ Show Quartiles

Summary Statistics

13 Columns

Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev
host_identity_verified	12308	0	2	.	.	.	.
neighbourhood	7319	4989	57	.	.	.	.
neighbourhood_cleansed	12308	0	5	.	.	.	.
neighbourhood_group_cleansed	0	12308	1	.	.	.	.
latitude	12308	0	.	43.72607	43.83053	43.774080243	0.0095253275
longitude	12308	0	.	11.157921106	11.33628	11.2536413	0.0162808591
property_type	12308	0	57	.	.	.	.
room_type	12308	0	4	.	.	.	.
accommodates	12308	0	.	1	16	3.7329379266	1.9480781004
bathrooms	0	12308	1	.	.	.	.
bathrooms_text	12304	4	31	.	.	.	.
bedrooms	10221	2087	.	1	13	1.6208785833	0.9101463566
beds	12221	87	.	1	20	2.233041486	1.499326031

neighbourhood_cleansed: Kept as is as direct relation with price and no missing values.



Bar Style: Side by side
Summary Statistic: Mean
Error Interval: Auto
Label: No Labels
Variables

Neighborhood_group_cleansed

Excluded this column as it has 0 available records.

Latitude: Kept as location of a place affects price.

Longitude: Kept as location of a place affects price.

property_type: Kept as type of property should have an impact on price.

room_type: Kept as room's type should have a direct impact on price.

Accommodates: Kept as capacity of property should have an impact on price

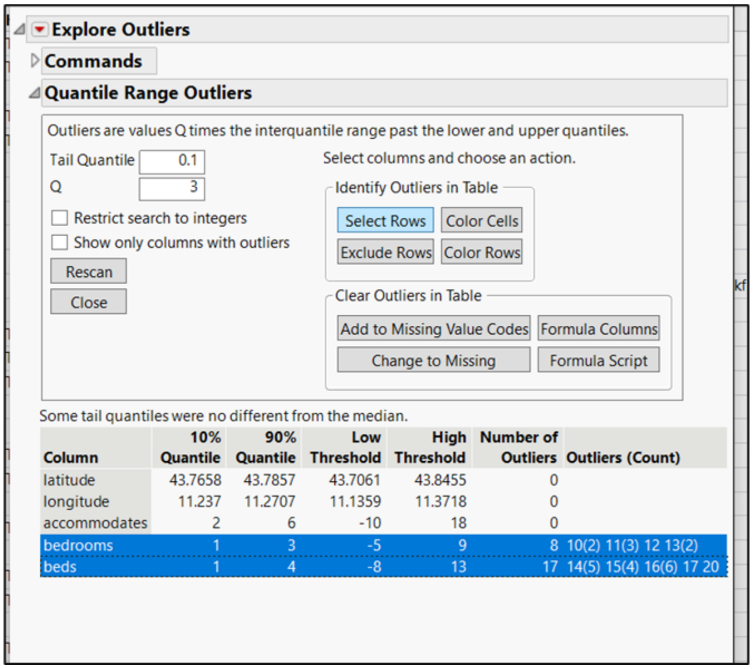
Bathrooms: Excluded this column as it has 0 available records.

Bathrooms_text: Kept as number and type of bathrooms should have an impact on price of property.

Bedrooms:

Imputed the data with average values, as 2087 values were missing.

There are 8 outliers as well, but not imputing its data as the properties with more number of bedrooms have higher prices which is understandable and enriches data quality for real life predictions.



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.

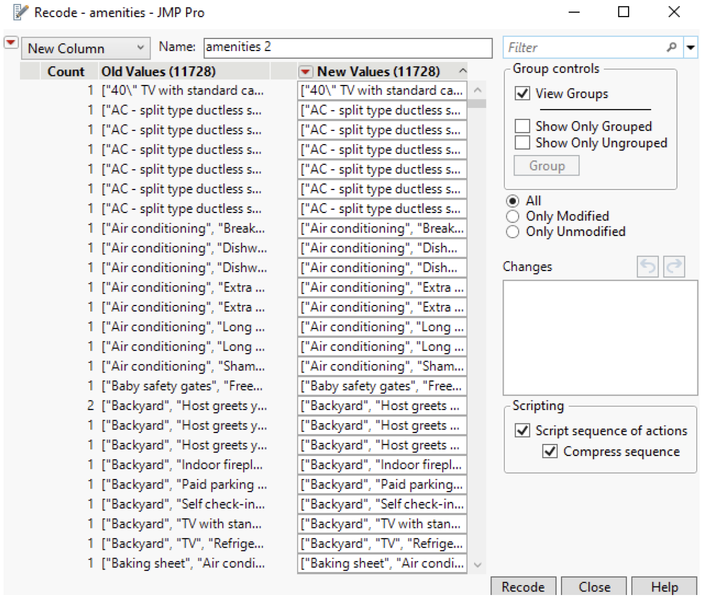
Some tail quantiles were no different from the median.

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
latitude	43.7658	43.7857	43.7061	43.8455	0	
longitude	11.237	11.2707	11.1359	11.3718	0	
accommodates	2	6	-10	18	0	
bedrooms	1	3	-5	9	8	10(2) 11(3) 12 13(2)
beds	1	4	-8	13	17	14(5) 15(4) 16(6) 17 20

Beds: Imputed the data with average values as 87 values were missing.

There are 17 outliers as well, but not imputing its data as it seems the properties with more no of beds have higher prices which is understandable and enriches data quality for real life prediction.

Amenities: Removing the variable a this is a categorical variable and it has no values which can be grouped together to make any meaningful pattern. Attached the screenshot.



Recode - amenities - JMP Pro

New Column   Name: amenities 2

Count	Old Values (11728)	New Values (11728)
1	["40\" TV with standard ca...	["40\" TV with standard ca...
1	["AC - split type ductless s...	["AC - split type ductless s...
1	["AC - split type ductless s...	["AC - split type ductless s...
1	["AC - split type ductless s...	["AC - split type ductless s...
1	["AC - split type ductless s...	["AC - split type ductless s...
1	["AC - split type ductless s...	["AC - split type ductless s...
1	["AC - split type ductless s...	["AC - split type ductless s...
1	["Air conditioning", "Break...	["Air conditioning", "Break...
1	["Air conditioning", "Dishw...	["Air conditioning", "Dish...
1	["Air conditioning", "Dishw...	["Air conditioning", "Dish...
1	["Air conditioning", "Extra ...	["Air conditioning", "Extra ...
1	["Air conditioning", "Extra ...	["Air conditioning", "Extra ...
1	["Air conditioning", "Long ...	["Air conditioning", "Long ...
1	["Air conditioning", "Long ...	["Air conditioning", "Long ...
1	["Air conditioning", "Sham...	["Air conditioning", "Sham...
1	["Baby safety gates", "Free...	["Baby safety gates", "Free...
2	["Backyard", "Host greets y...	["Backyard", "Host greets ...
1	["Backyard", "Host greets y...	["Backyard", "Host greets ...
1	["Backyard", "Host greets y...	["Backyard", "Host greets ...
1	["Backyard", "Indoor firepl...	["Backyard", "Indoor firepl...
1	["Backyard", "Paid parking ...	["Backyard", "Paid parking...
1	["Backyard", "Self check-in...	["Backyard", "Self check-in...
1	["Backyard", "TV with stan...	["Backyard", "TV with stan...
1	["Backyard", "TV", "Refrige...	["Backyard", "TV", "Refrige...
1	["Baking sheet", "Air condi...	["Baking sheet", "Air condi...

Filter

Group controls
☑ View Groups
□ Show Only Grouped
□ Show Only Ungrouped
Group

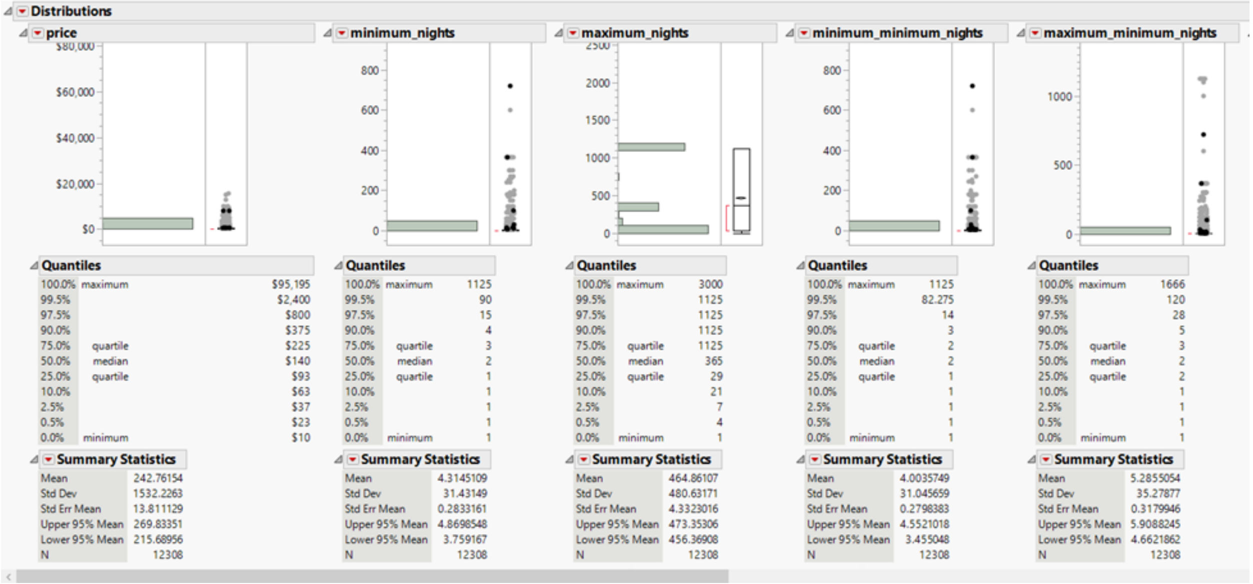
● All
○ Only Modified
○ Only Unmodified

Changes

Scripting
☑ Script sequence of actions
☑ Compress sequence

Recode   Close   Help

Finding Outlier in the below columns:

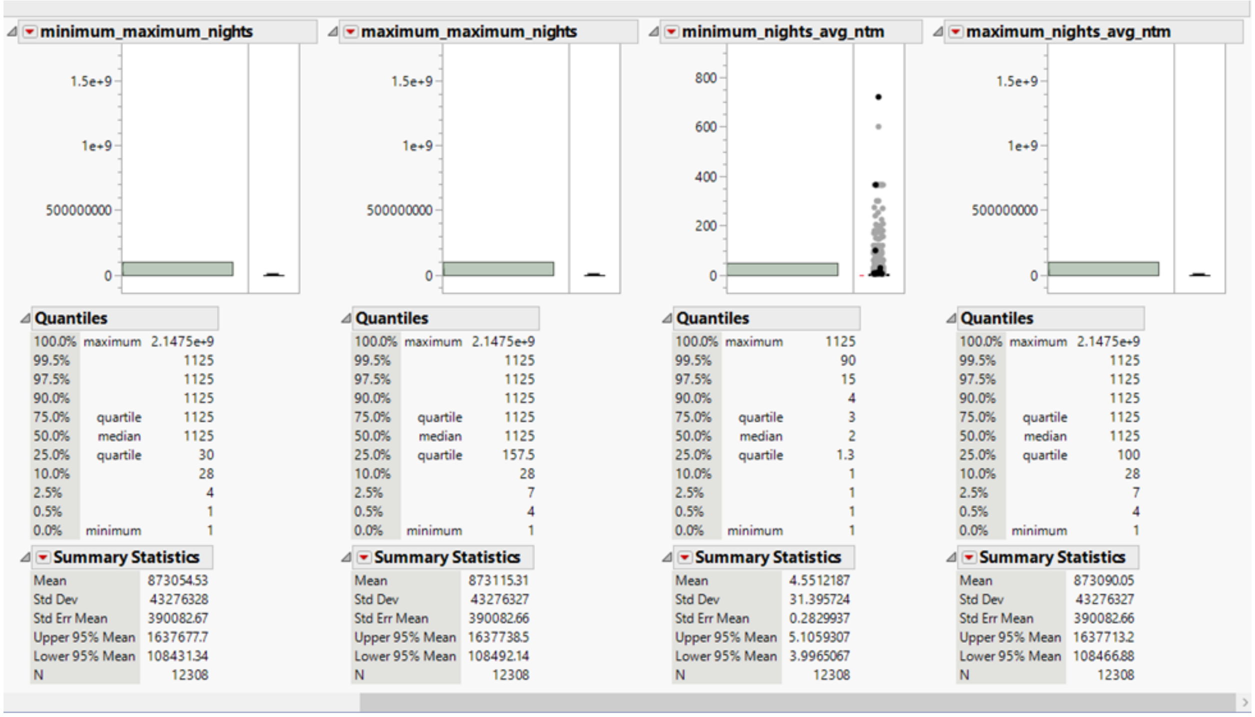


Quantiles

		price	minimum_nights	maximum_nights	minimum_minimum_nights	maximum_minimum_nights
100.0%	maximum	$95,195	1125	3000	1125	1666
99.5%		$2,400	90	1125	82.275	120
97.5%		$800	15	1125	14	28
90.0%		$375	4	1125	3	5
75.0%	quartile	$225	3	1125	2	3
50.0%	median	$140	2	365	2	2
25.0%	quartile	$93	1	29	1	2
10.0%		$63	1	21	1	1
2.5%		$37	1	7	1	1
0.5%		$23	1	4	1	1
0.0%	minimum	$10	1	1	1	1

Summary Statistics

	price	minimum_nights	maximum_nights	minimum_minimum_nights	maximum_minimum_nights
Mean	242.76154	4.3145109	464.86107	4.0035749	5.2855054
Std Dev	1532.2263	31.43149	480.63171	31.045659	35.27877
Std Err Mean	13.811129	0.2833161	4.3323016	0.2798383	0.3179946
Upper 95% Mean	269.83351	4.8698548	473.35306	4.5521018	5.9088245
Lower 95% Mean	215.68956	3.759167	456.36908	3.455048	4.6621862
N	12308	12308	12308	12308	12308



minimum_maximum_nights

Quantiles

100.0%	maximum	2.1475e+9
99.5%		1125
97.5%		1125
90.0%		1125
75.0%	quartile	1125
50.0%	median	1125
25.0%	quartile	30
10.0%		28
2.5%		4
0.5%		1
0.0%	minimum	1

Summary Statistics

Mean	873054.53
Std Dev	43276328
Std Err Mean	390082.67
Upper 95% Mean	1637677.7
Lower 95% Mean	108431.34
N	12308

maximum_maximum_nights

Quantiles

100.0%	maximum	2.1475e+9
99.5%		1125
97.5%		1125
90.0%		1125
75.0%	quartile	1125
50.0%	median	1125
25.0%	quartile	157.5
10.0%		28
2.5%		7
0.5%		4
0.0%	minimum	1

Summary Statistics

Mean	873115.31
Std Dev	43276327
Std Err Mean	390082.66
Upper 95% Mean	1637738.5
Lower 95% Mean	108492.14
N	12308

minimum_nights_avg_ntm

Quantiles

100.0%	maximum	1125
99.5%		90
97.5%		15
90.0%		4
75.0%	quartile	3
50.0%	median	2
25.0%	quartile	1.3
10.0%		1
2.5%		1
0.5%		1
0.0%	minimum	1

Summary Statistics

Mean	4.5512187
Std Dev	31.395724
Std Err Mean	0.2829937
Upper 95% Mean	5.1059307
Lower 95% Mean	3.9965067
N	12308

maximum_nights_avg_ntm

Quantiles

100.0%	maximum	2.1475e+9
99.5%		1125
97.5%		1125
90.0%		1125
75.0%	quartile	1125
50.0%	median	1125
25.0%	quartile	100
10.0%		28
2.5%		7
0.5%		4
0.0%	minimum	1

Summary Statistics

Mean	873090.05
Std Dev	43276327
Std Err Mean	390082.66
Upper 95% Mean	1637713.2
Lower 95% Mean	108466.88
N	12308

Price: This is our target variable, we will not make any change to this column.

Minimum_nights : The data in the given column is not making any meaningful pattern with the price hence excluding it.

Maximum_nights: The data in the given column is not making any meaningful pattern with the price hence excluding it.

Minimum_minimum_nights: The data in the given column is not making any meaningful pattern with the price hence excluding it.

Maximum_minimum_nights: The data in the given column is not making any meaningful pattern with the price hence excluding it.

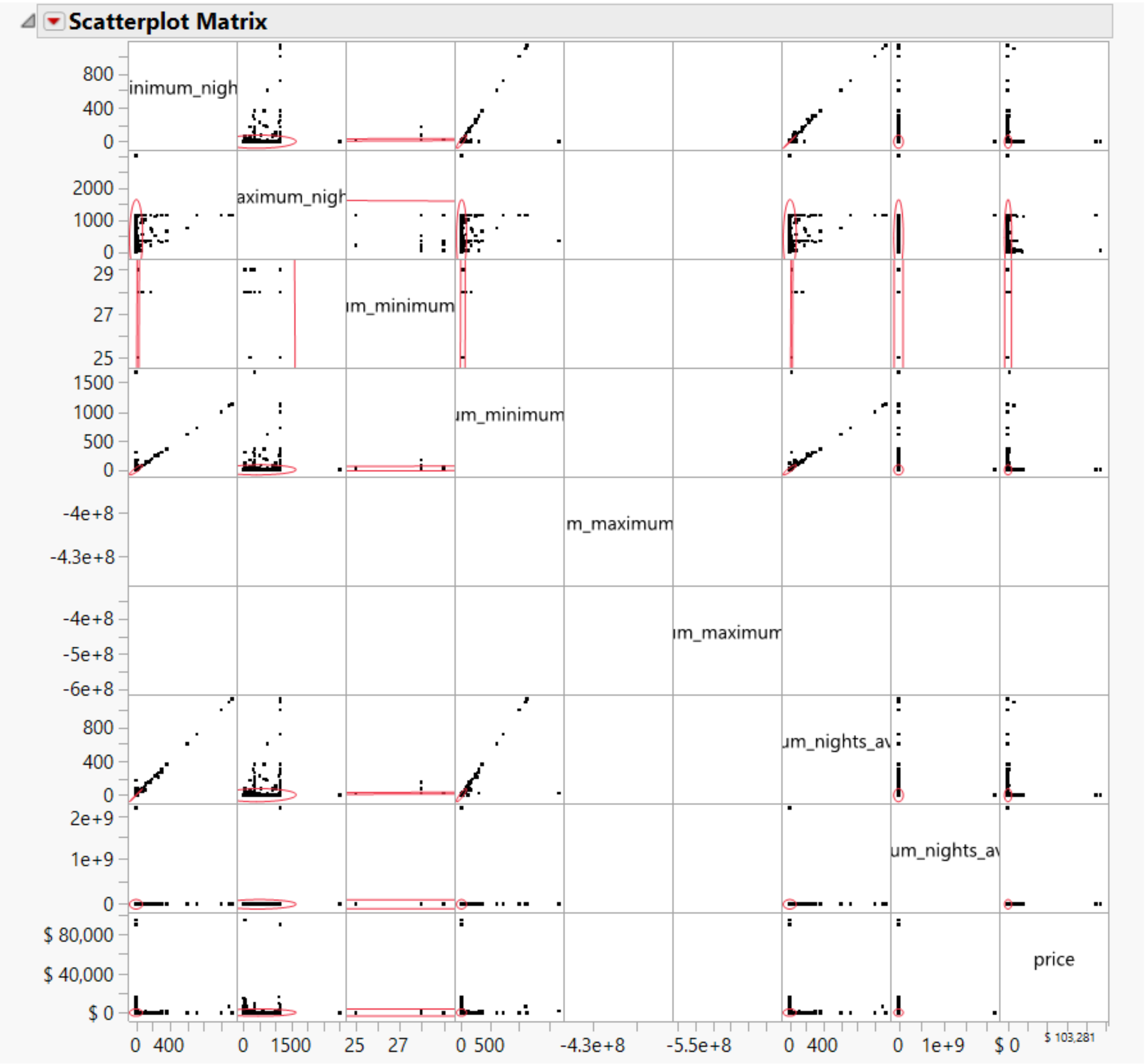
Minimum_maximum_nights: The data in the given column is not making any meaningful pattern with the price hence excluding it.

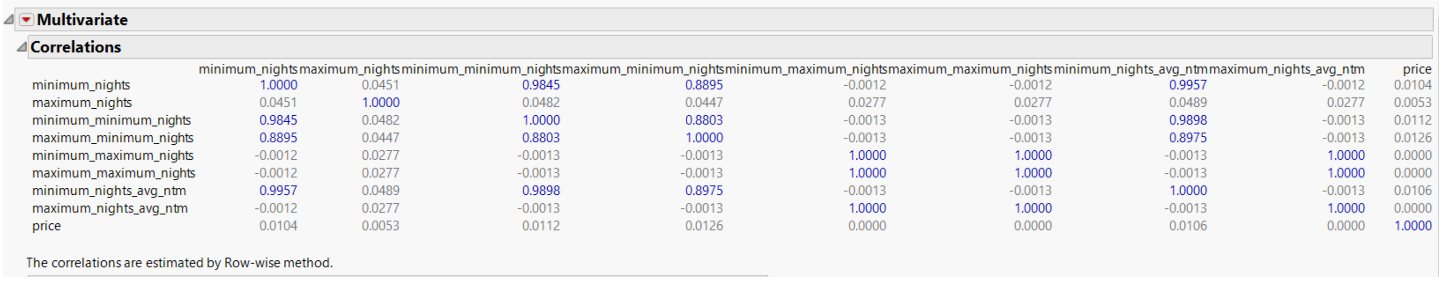
Maximum_maximum_nights: The data in the given column is not making any meaningful pattern with the price hence excluding it.

Minimum_nights_avg_ntm: The data in the given column is not making any meaningful pattern with the price hence excluding it.

Maximum_nights_avg_ntm: The data in the given column is not making any meaningful pattern with the price hence excluding it.

I have attached the below screenshot, which shows that there is no relation between the target variable and the columns listed above.





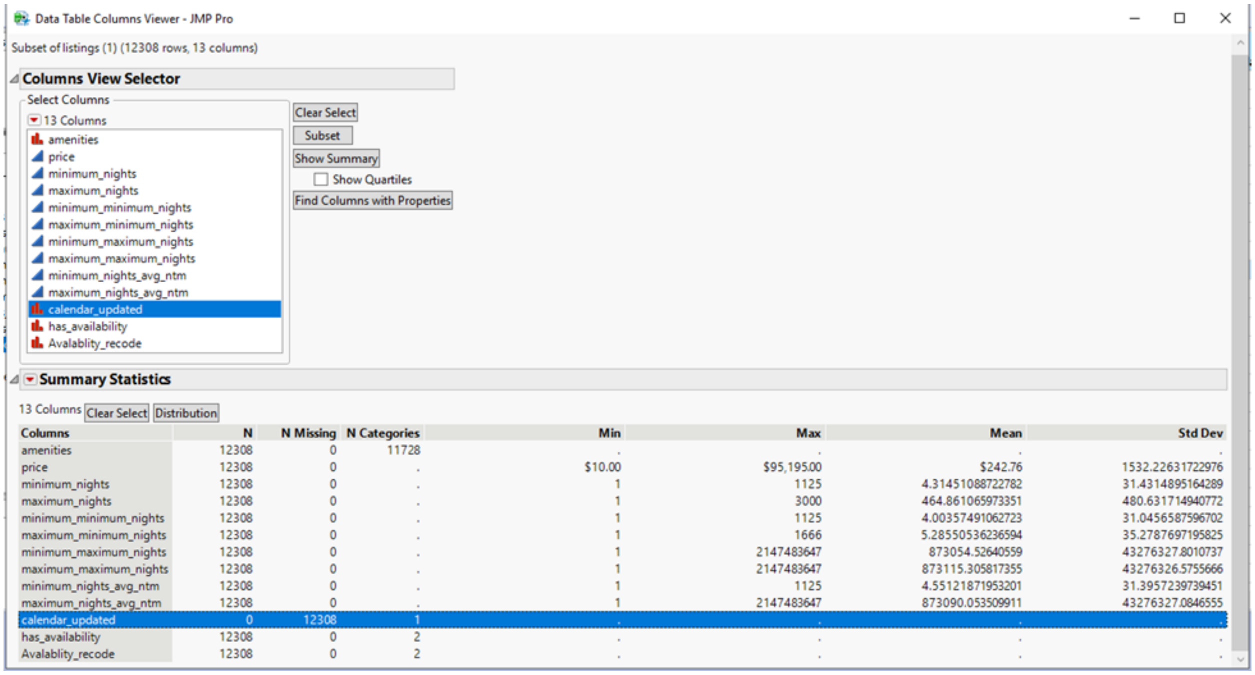
Multivariate

Correlations

	minimum_nights	maximum_nights	minimum_minimum_nights	maximum_minimum_nights	minimum_maximum_nights	maximum_maximum_nights	minimum_nights_avg_ntm	maximum_nights_avg_ntm	price
minimum_nights	1.0000	0.0451	0.9845	0.8895	-0.0012	-0.0012	0.9957	-0.0012	0.0104
maximum_nights	0.0451	1.0000	0.0482	0.0447	0.0277	0.0277	0.0489	0.0277	0.0053
minimum_minimum_nights	0.9845	0.0482	1.0000	0.8803	-0.0013	-0.0013	0.9898	-0.0013	0.0112
maximum_minimum_nights	0.8895	0.0447	0.8803	1.0000	-0.0013	-0.0013	0.8975	-0.0013	0.0126
minimum_maximum_nights	-0.0012	0.0277	-0.0013	-0.0013	1.0000	1.0000	-0.0013	1.0000	0.0000
maximum_maximum_nights	-0.0012	0.0277	-0.0013	-0.0013	1.0000	1.0000	-0.0013	1.0000	0.0000
minimum_nights_avg_ntm	0.9957	0.0489	0.9898	0.8975	-0.0013	-0.0013	1.0000	-0.0013	0.0106
maximum_nights_avg_ntm	-0.0012	0.0277	-0.0013	-0.0013	1.0000	1.0000	-0.0013	1.0000	0.0000
price	0.0104	0.0053	0.0112	0.0126	0.0000	0.0000	0.0106	0.0000	1.0000

The correlations are estimated by Row-wise method.

Calendar_updated : All Missing Values in this column, hence Excluding it from the data file. Attached screenshot.



Data Table Columns Viewer - JMP Pro

Subset of listings (1) (12308 rows, 13 columns)

Columns View Selector

Select Columns

13 Columns

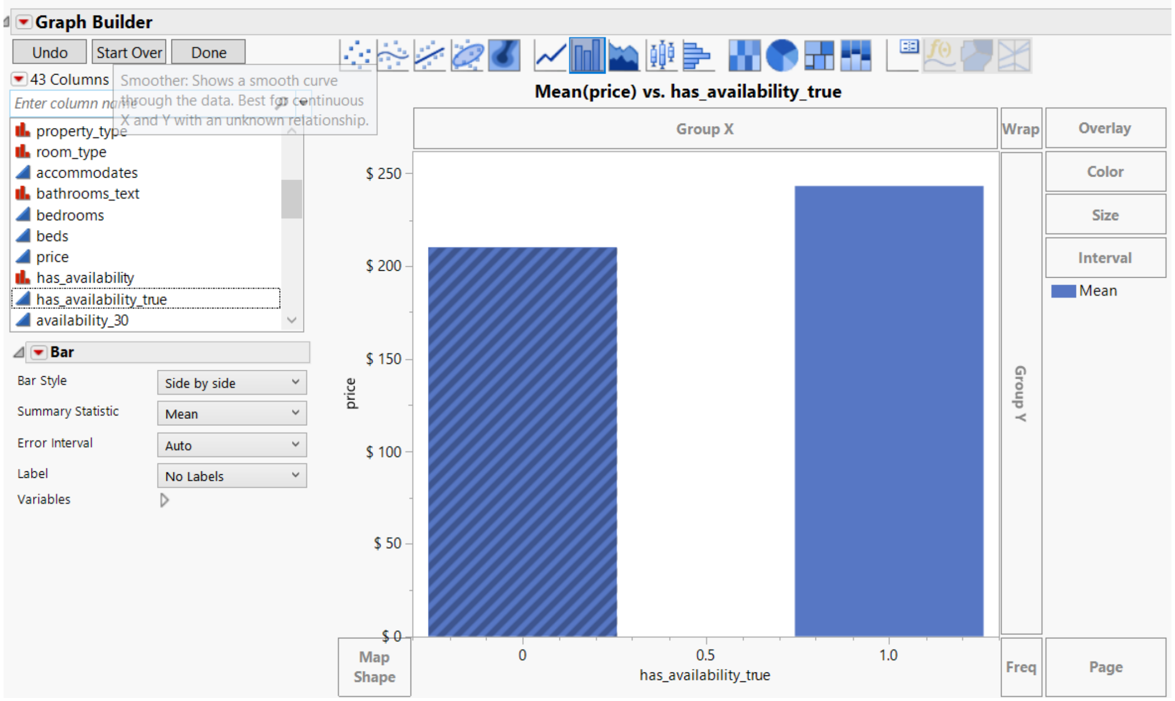
amenities
price
minimum_nights
maximum_nights
minimum_minimum_nights
maximum_minimum_nights
minimum_maximum_nights
maximum_maximum_nights
minimum_nights_avg_ntm
maximum_nights_avg_ntm
calendar_updated
has_availability
Avalablity_recode

Summary Statistics

13 Columns

Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev
amenities	12308	0	11728	.	.	.	.
price	12308	0	.	$10.00	$95,195.00	$242.76	1532.22631722976
minimum_nights	12308	0	.	1	1125	4.31451088722782	31.4314895164289
maximum_nights	12308	0	.	1	3000	464.861065973351	480.631714940772
minimum_minimum_nights	12308	0	.	1	1125	4.00357491062723	31.0456587596702
maximum_minimum_nights	12308	0	.	1	1666	5.28550536236594	35.2787697195825
minimum_maximum_nights	12308	0	.	1	2147483647	873054.52640559	43276327.8010737
maximum_maximum_nights	12308	0	.	1	2147483647	873115.305817355	43276326.5755666
minimum_nights_avg_ntm	12308	0	.	1	1125	4.55121871953201	31.3957239739451
maximum_nights_avg_ntm	12308	0	.	1	2147483647	873090.053509911	43276327.0846555
calendar_updated	0	12308	1	.	.	.	.
has_availability	12308	0	2	.	.	.	.
Avalablity_recode	12308	0	2	.	.	.	.

Has_availability: Created Indicator column for the values t and f, and created new Column **has_availablity_true**



Smoother: Shows a smooth curve through the data. Best for continuous X and Y with an unknown relationship.

Bar Style: Side by side

Summary Statistic: Mean

Error Interval: Auto

Label: No Labels

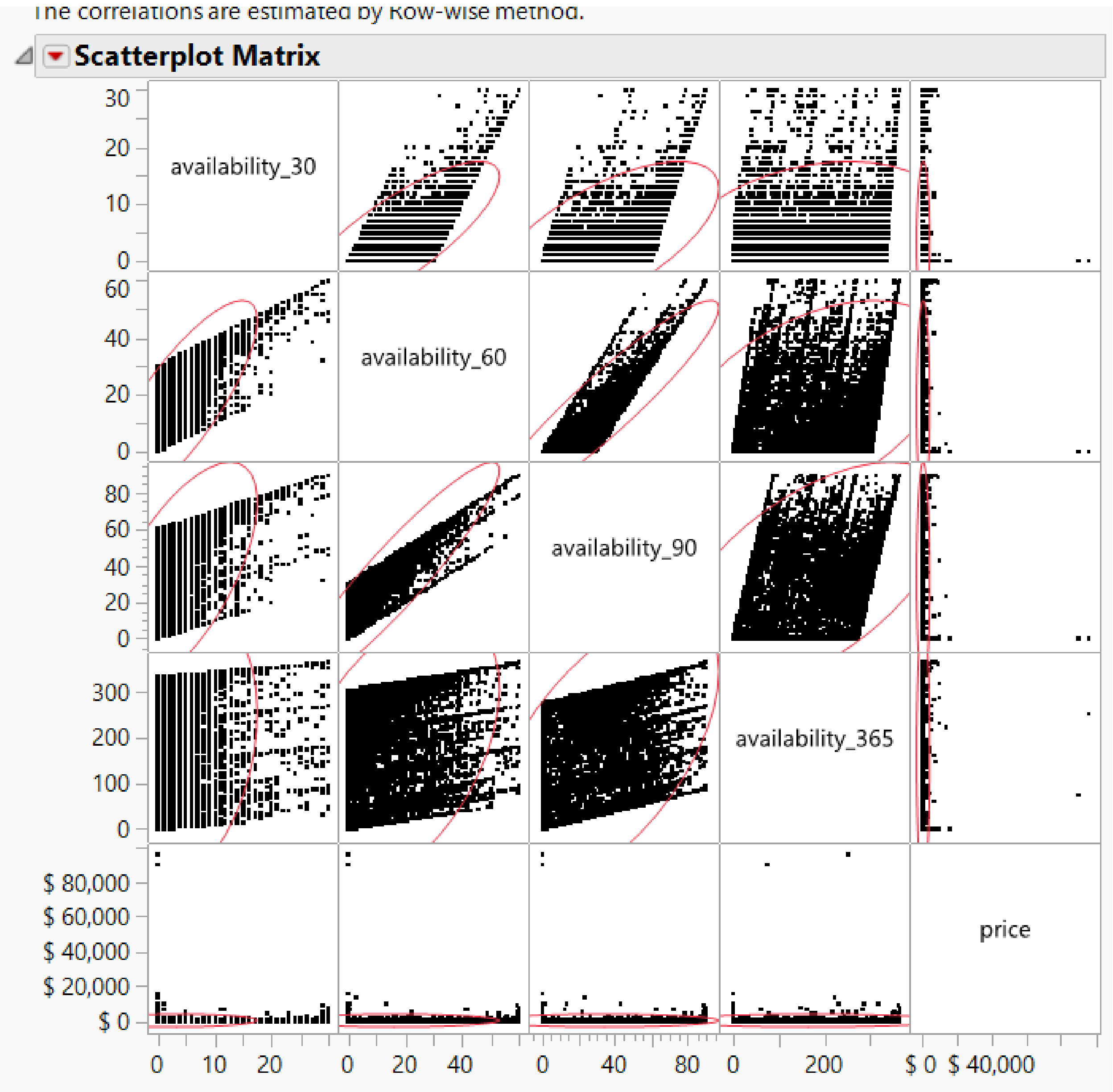
Availability_30: Has no relationship with price variables. That is why excluding it from the data set

Availability_60: Has no relationship with price variables. That is why excluding it from the data set.

Availablility_90: Has no relationship with price variables. That is why excluding it from the data set

Availablility_365: Has no relationship with price variables.

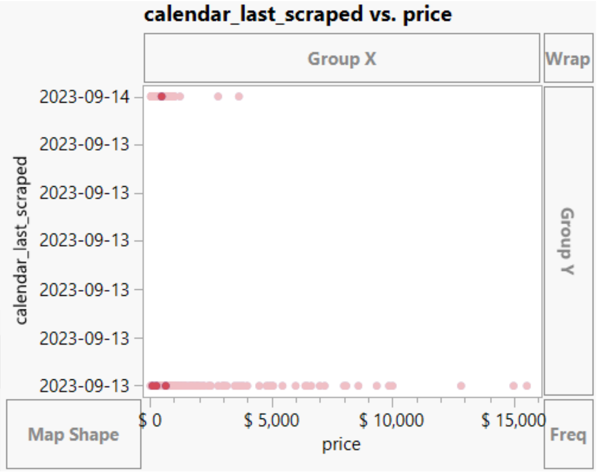
All the available columns have no relation with price variables. That's why excluding them.



The correlations are estimated by Row-wise method.

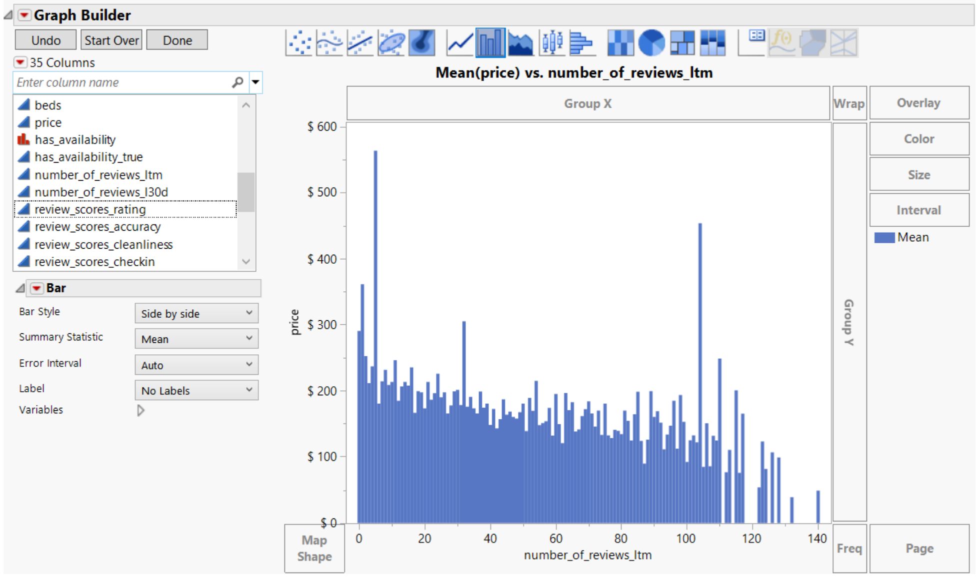
Scatterplot Matrix

Calendar last scraped: This column provides us little to no information. Because there's only two dates when the calendar was last scraped. Hence can be eliminated.



Number_of_reviews: There is no meaningful pattern between price and number of reviews.

Number_of_reviews_Itm: There exists a relationship with price.



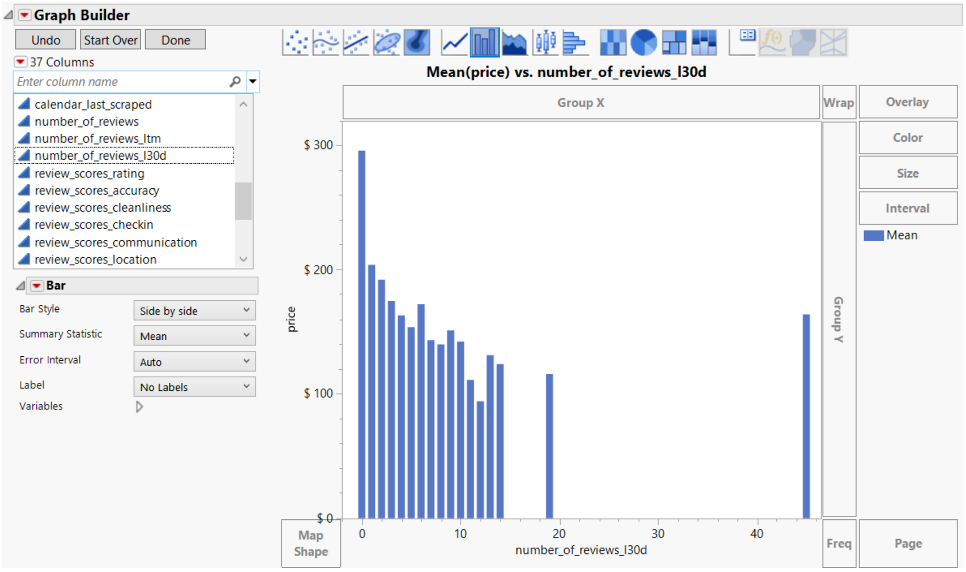
Bar Style: Side by side

Summary Statistic: Mean

Error Interval: Auto

Label: No Labels

number_of_reviews_I30d: It shows relation with price. We can see as the review increases the price decreases.



Bar Style: Side by side

Summary Statistic: Mean

Error Interval: Auto

Label: No Labels

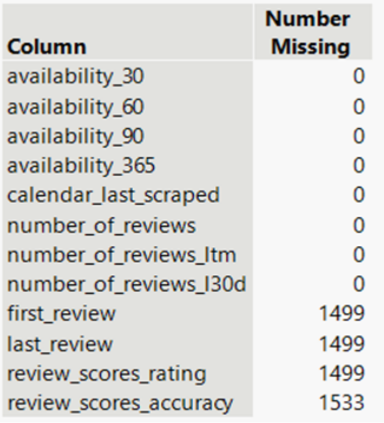
Variables

First_review: We are excluding this column because there's no correlation with price.

last_review : We are excluding this column because there's no correlation with price.

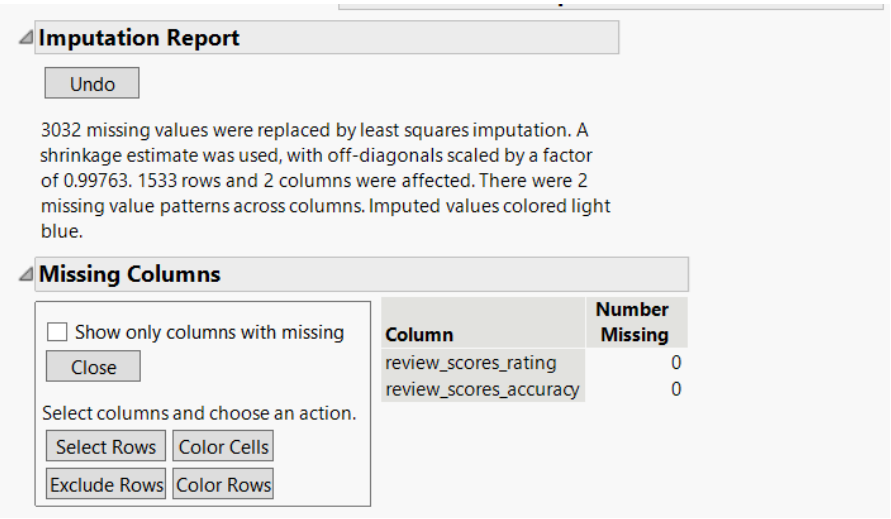
Review_scores_rating: Imputed the data on the missing values.

Review_scores_accuracy: Imputed the data on the missing values.



Column	Number Missing
availability_30	0
availability_60	0
availability_90	0
availability_365	0
calendar_last_scraped	0
number_of_reviews	0
number_of_reviews_ltm	0
number_of_reviews_l30d	0
first_review	1499
last_review	1499
review_scores_rating	1499
review_scores_accuracy	1533

Imputation would fit well in this case since only 1499 or 1533 rows of null values are present in each column out of 12,308 rows.



Imputation Report

3032 missing values were replaced by least squares imputation. A shrinkage estimate was used, with off-diagonals scaled by a factor of 0.99763. 1533 rows and 2 columns were affected. There were 2 missing value patterns across columns. Imputed values colored light blue.

Missing Columns

□ Show only columns with missing

Select columns and choose an action.

Column	Number Missing
review_scores_rating	0
review_scores_accuracy	0

Column Name: review_scores_cleanliness

Comments:

N Missing: 1533; Missing Data - Decided to impute due to <12% of rows

Outliers: 63; Outlier - Exclude outliers as this will affect price

Column Name: review_scores_checkin

N Missing: 1534

Outliers: 74

Comments:

Missing Data - Decided to impute due to <12% of rows

Outlier - Exclude outliers as this will affect price

Column Name: review_scores_communication

N Missing: 1534

Outliers: 85

Comments:

Missing Data - Decided to impute due to <12% of rows

Outlier - Exclude outliers as this will affect price

Column Name: review_scores_location

N Missing: 1534

Outliers: 66

Comments:

Missing Data - Decided to impute due to <12% of rows

Outlier - Exclude outliers as this will affect price

Column Name: review_scores_value

N Missing: 1534

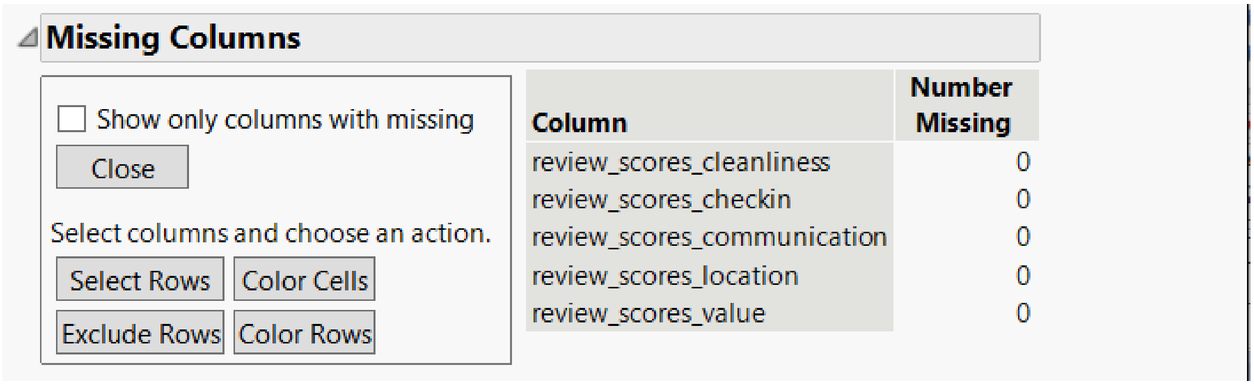
Outliers: 47

Comments:

Missing Data - Decided to impute due to <12% of rows

Outlier - Exclude outliers as this will affect price

After imputing missing data:



Select columns and choose an action.

Column	Number Missing
review_scores_cleanliness	0
review_scores_checkin	0
review_scores_communication	0
review_scores_location	0
review_scores_value	0

Column Name: license

N Missing: 10484

Outliers: #N/A

Comments:

Should exclude this column due to missing values and no effect on price

Column Name: instant_bookable

N Missing: 0

Outliers: #N/A

Comments:

Add an indicator column 'Instant_bookable_true to have 0 and 1. This was done to create a continuous variable associated with this character value type to be able to analyze.

Column Name: instant_bookable_true

Comments:

Add an indicator column 'Instant_bookable_true to have 0 and 1. This was done to create a continuous variable associated with this character value type to be able to analyze.

Column Name: calculated_host_listings_count

N Missing: 0

Outliers: 477

Comments:

Outliers - Could be possible to have hosts with a large amount of listings, this should not affect the price model

Decision - Exclude from price model, as this should not affect price"

Column Name: calculated_host_listings_count_entire_homes

N Missing: 0

Outliers: 477

Comments:

Outliers - Could be possible to have hosts with a large amount of listings, this should not affect the price model

Column Name: calculated_host_listings_count_private_rooms

N Missing: 0

Outliers: 45

Comments:

Outliers - Decided to leave alone due to the possibility of there being this amount of private rooms. The amount of outliers should not affect overall price model

Column Name: calculated_host_listings_count_shared_rooms

N Missing: 0

Outliers: 11

Comments:

Outliers - Decided to leave alone due to 11 shared rooms makes sense in a house

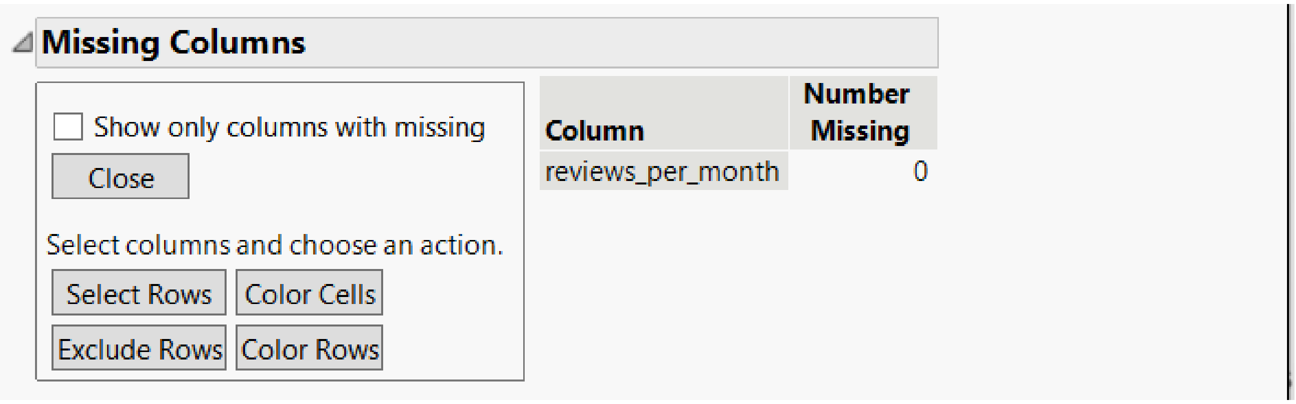
Column Name: reviews_per_month

N Missing: 1499

Outliers: 0

Comments:

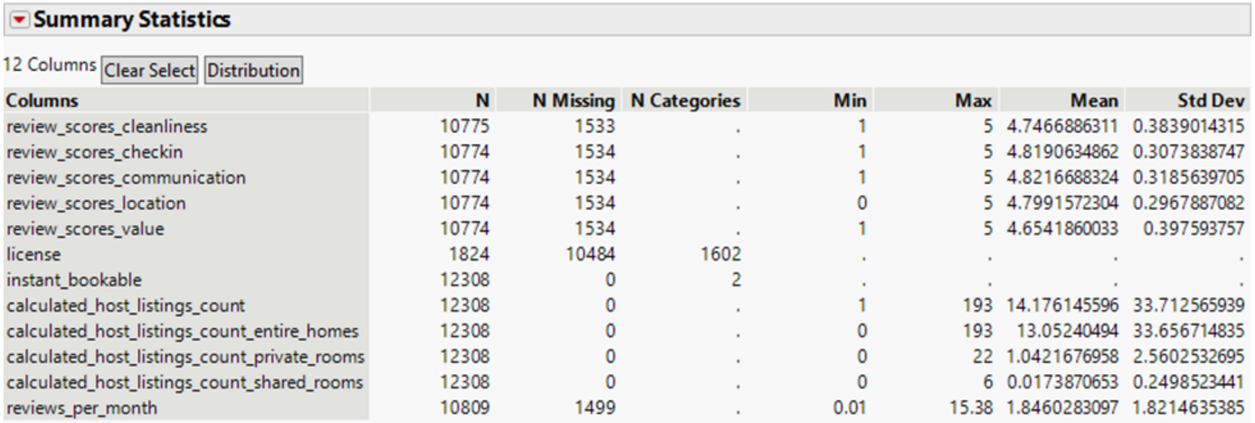
Missing Data - Decided to impute due to <12% of rows



Missing Columns

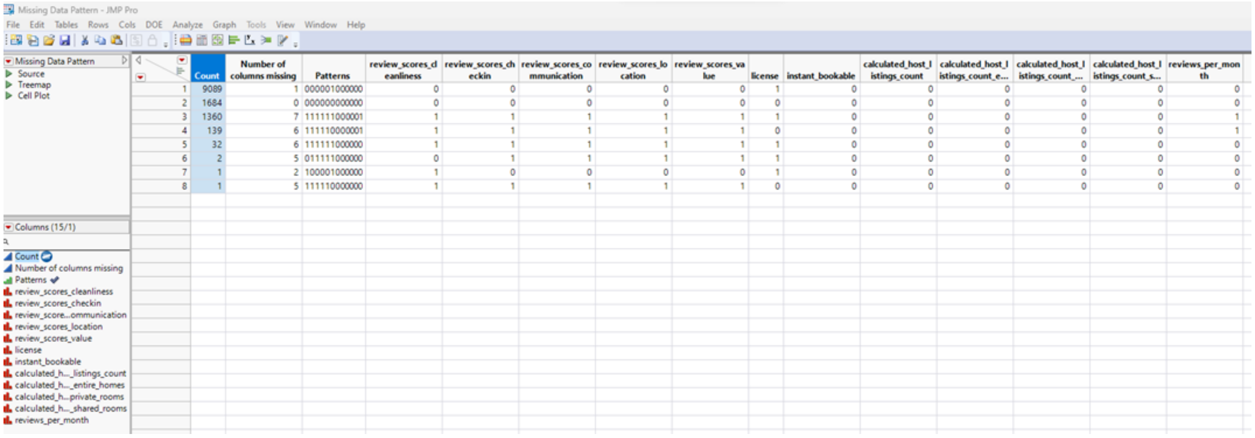
Select columns and choose an action.

Column	Number Missing
reviews_per_month	0

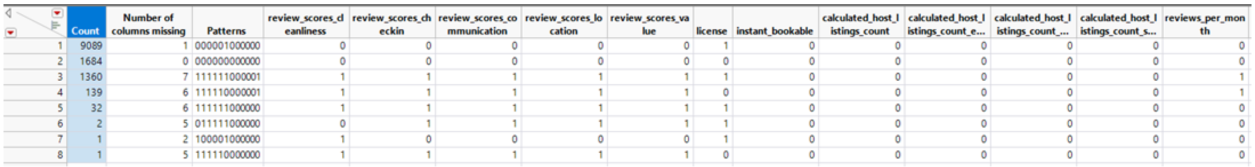


Summary Statistics

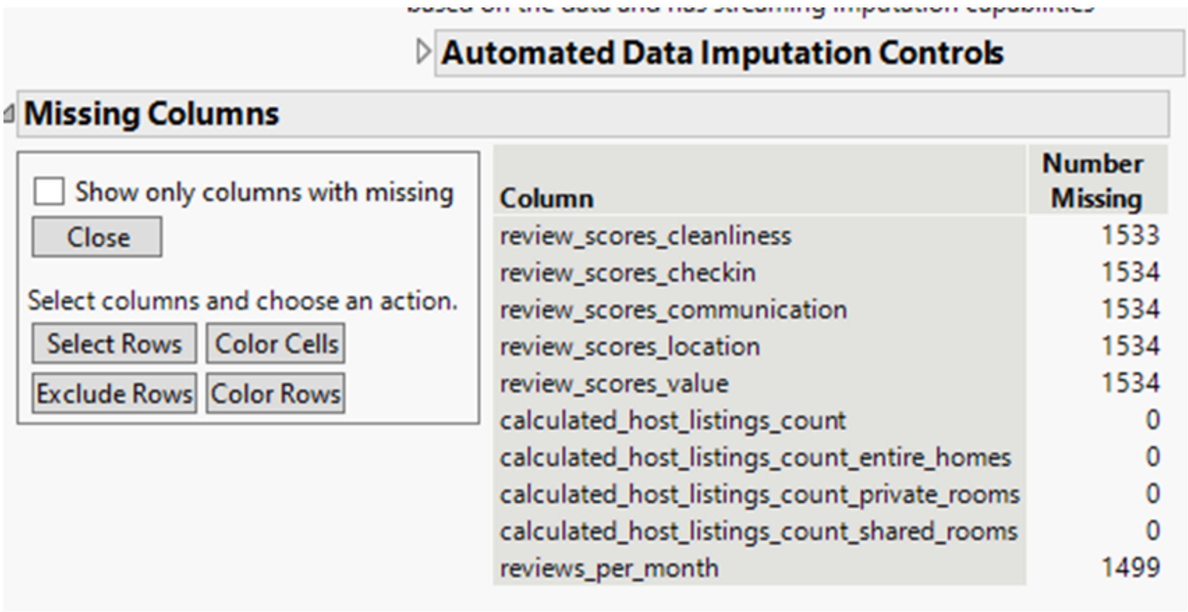
Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev
review_scores_cleanliness	10775	1533	.	1	5	4.7466886311	0.3839014315
review_scores_checkin	10774	1534	.	1	5	4.8190634862	0.3073838747
review_scores_communication	10774	1534	.	1	5	4.8216688324	0.3185639705
review_scores_location	10774	1534	.	0	5	4.7991572304	0.2967887082
review_scores_value	10774	1534	.	1	5	4.6541860033	0.397593757
license	1824	10484	1602	.	.	.	.
instant_bookable	12308	0	2	.	.	.	.
calculated_host_listings_count	12308	0	.	1	193	14.176145596	33.712565939
calculated_host_listings_count_entire_homes	12308	0	.	0	193	13.05240494	33.656714835
calculated_host_listings_count_private_rooms	12308	0	.	0	22	1.0421676958	2.5602532695
calculated_host_listings_count_shared_rooms	12308	0	.	0	6	0.0173870653	0.2498523441
reviews_per_month	10809	1499	.	0.01	15.38	1.8460283097	1.8214635385



	Count	Number of columns missing	Patterns	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_location	review_scores_value	license	instant_bookable	calculated_host_listings_count	calculated_host_listings_count_e...	calculated_host_listings_count_...	calculated_host_listings_count_s...	reviews_per_month
1	9089	1	000001000000	0	0	0	0	0	1	0	0	0	0	0	0
2	1684	0	000000000000	0	0	0	0	0	0	0	0	0	0	0	0
3	1360	7	111111000001	1	1	1	1	1	1	0	0	0	0	0	1
4	139	6	111110000001	1	1	1	1	1	0	0	0	0	0	0	1
5	32	6	111111000000	1	1	1	1	1	1	0	0	0	0	0	0
6	2	5	011111000000	0	1	1	1	1	1	0	0	0	0	0	0
7	1	2	100001000000	1	0	0	0	0	1	0	0	0	0	0	0
8	1	5	111110000000	1	1	1	1	1	0	0	0	0	0	0	0



	Count	Number of columns missing	Patterns	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_location	review_scores_value	license	instant_bookable	calculated_host_listings_count	calculated_host_listings_count_e...	calculated_host_listings_count_...	calculated_host_listings_count_s...	reviews_per_month
1	9089	1	000001000000	0	0	0	0	0	1	0	0	0	0	0	0
2	1684	0	000000000000	0	0	0	0	0	0	0	0	0	0	0	0
3	1360	7	111111000001	1	1	1	1	1	1	0	0	0	0	0	1
4	139	6	111110000001	1	1	1	1	1	0	0	0	0	0	0	1
5	32	6	111111000000	1	1	1	1	1	1	0	0	0	0	0	0
6	2	5	011111000000	0	1	1	1	1	1	0	0	0	0	0	0
7	1	2	100001000000	1	0	0	0	0	1	0	0	0	0	0	0
8	1	5	111110000000	1	1	1	1	1	0	0	0	0	0	0	0



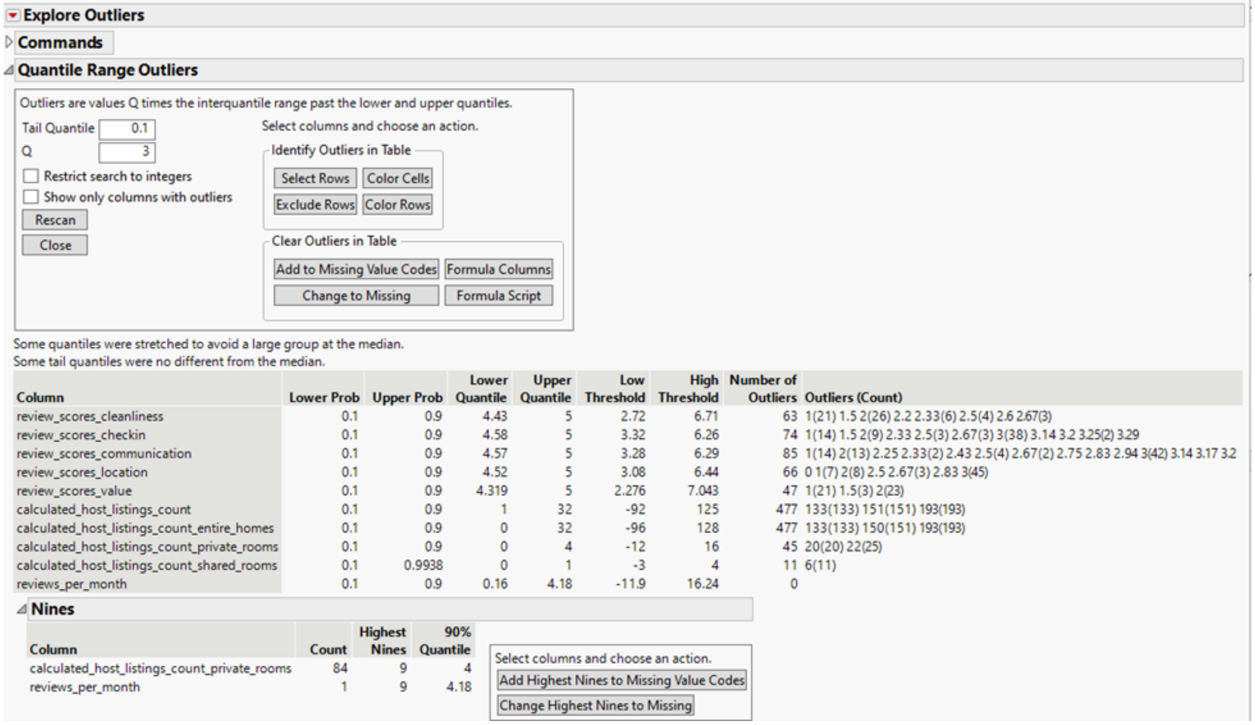
based on the data and has streaming imputation capabilities

▷ Automated Data Imputation Controls
◿ Missing Columns
□ Show only columns with missing
Close
Select Rows
Color Cells
Exclude Rows
Color Rows

Select columns and choose an action.

Column	Number Missing
review_scores_cleanliness	1533
review_scores_checkin	1534
review_scores_communication	1534
review_scores_location	1534
review_scores_value	1534
calculated_host_listings_count	0
calculated_host_listings_count_entire_homes	0
calculated_host_listings_count_private_rooms	0
calculated_host_listings_count_shared_rooms	0
reviews_per_month	1499

Outliers:



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

Tail Quantile 0.1

Q 3

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.

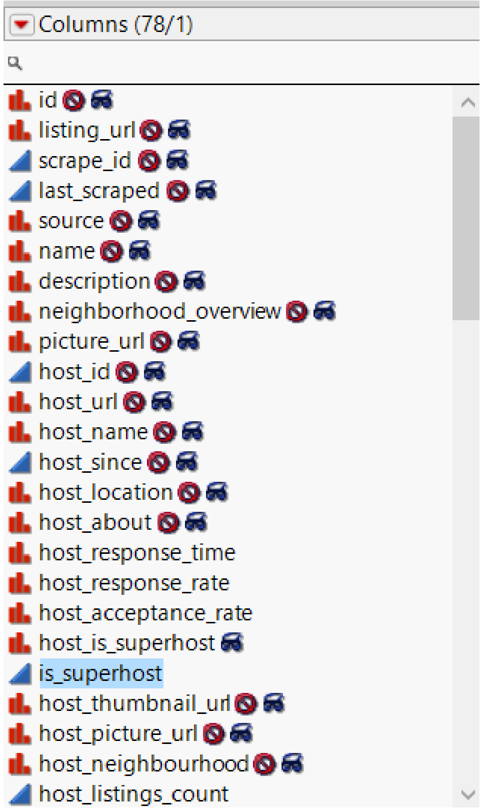
Column	Lower Prob	Upper Prob	Lower Quantile	Upper Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
review_scores_cleanliness	0.1	0.9	4.43	5	2.72	6.71	63	1(21) 1.5 2(26) 2.2 2.33(6) 2.5(4) 2.6 2.67(3)
review_scores_checkin	0.1	0.9	4.58	5	3.32	6.26	74	1(14) 1.5 2(9) 2.33 2.5(3) 2.67(3) 3(38) 3.14 3.2 3.25(2) 3.29
review_scores_communication	0.1	0.9	4.57	5	3.28	6.29	85	1(14) 2(13) 2.25 2.33(2) 2.43 2.5(4) 2.67(2) 2.75 2.83 2.94 3(42) 3.14 3.17 3.2
review_scores_location	0.1	0.9	4.52	5	3.08	6.44	66	0 1(7) 2(8) 2.5 2.67(3) 2.83 3(45)
review_scores_value	0.1	0.9	4.319	5	2.276	7.043	47	1(21) 1.5(3) 2(23)
calculated_host_listings_count	0.1	0.9	1	32	-92	125	477	133(133) 151(151) 193(193)
calculated_host_listings_count_entire_homes	0.1	0.9	0	32	-96	128	477	133(133) 150(151) 193(193)
calculated_host_listings_count_private_rooms	0.1	0.9	0	4	-12	16	45	20(20) 22(25)
calculated_host_listings_count_shared_rooms	0.1	0.9938	0	1	-3	4	11	6(11)
reviews_per_month	0.1	0.9	0.16	4.18	-11.9	16.24	0	

◿ Nines
Add Highest Nines to Missing Value Codes
Change Highest Nines to Missing

Column	Count	Highest Nines	90% Quantile
calculated_host_listings_count_private_rooms	84	9	4
reviews_per_month	1	9	4.18

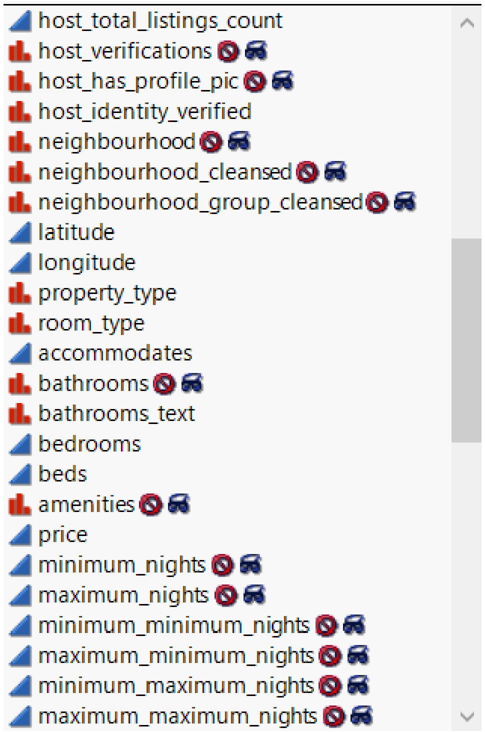
Select columns and choose an action.

Total 30 columns we have finalized at the end in which 24 are continuous and 6 are categorical.

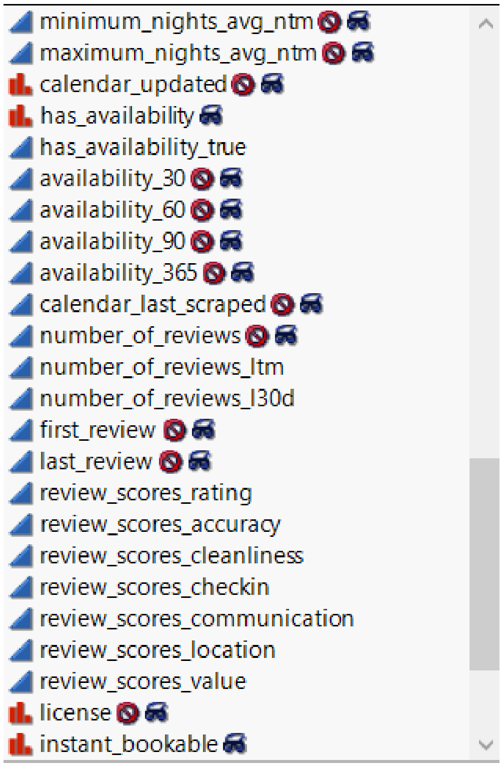


Columns (78/1)

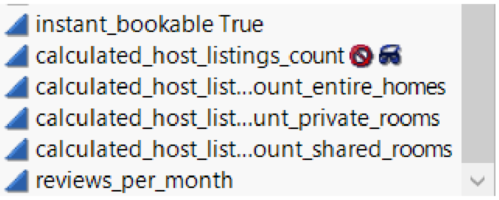
id
listing_url
scrape_id
last_scraped
source
name
description
neighborhood_overview
picture_url
host_id
host_url
host_name
host_since
host_location
host_about
host_response_time
host_response_rate
host_acceptance_rate
host_is_superhost
is_superhost
host_thumbnail_url
host_picture_url
host_neighbourhood
host_listings_count



```
host_total_listings_count
host_verifications ⊘ 👓
host_has_profile_pic ⊘ 👓
host_identity_verified
neighbourhood ⊘ 👓
neighbourhood_cleansed ⊘ 👓
neighbourhood_group_cleansed ⊘ 👓
latitude
longitude
property_type
room_type
accommodates
bathrooms ⊘ 👓
bathrooms_text
bedrooms
beds
amenities ⊘ 👓
price
minimum_nights ⊘ 👓
maximum_nights ⊘ 👓
minimum_minimum_nights ⊘ 👓
maximum_minimum_nights ⊘ 👓
minimum_maximum_nights ⊘ 👓
maximum_maximum_nights ⊘ 👓
```

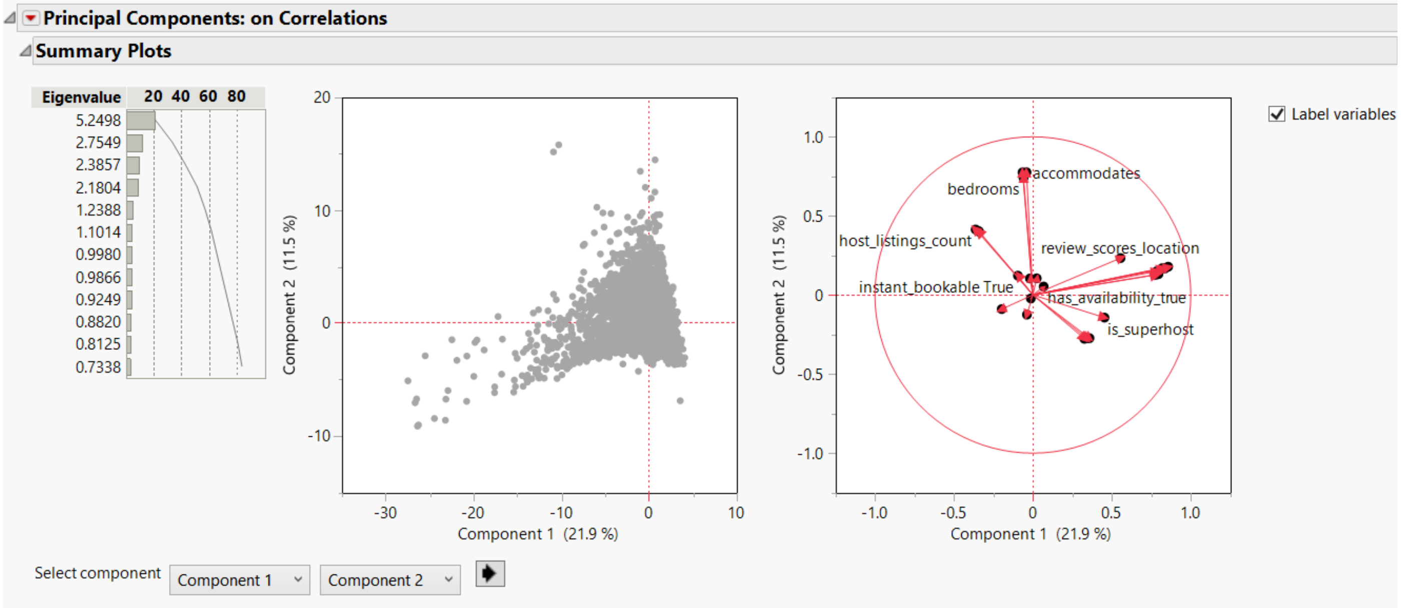


```
minimum_nights_avg_ntm ⊘ 👓
maximum_nights_avg_ntm ⊘ 👓
calendar_updated ⊘ 👓
has_availability 👓
has_availability_true
availability_30 ⊘ 👓
availability_60 ⊘ 👓
availability_90 ⊘ 👓
availability_365 ⊘ 👓
calendar_last_scraped ⊘ 👓
number_of_reviews ⊘ 👓
number_of_reviews_ltm
number_of_reviews_l30d
first_review ⊘ 👓
last_review ⊘ 👓
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value
license ⊘ 👓
instant_bookable 👓
```



instant_bookable True
calculated_host_listings_count
calculated_host_list...ount_entire_homes
calculated_host_list...unt_private_rooms
calculated_host_list...ount_shared_rooms
reviews_per_month

PCA: Principal component Analysis has been done on the 24 continuous columns of the data set.



Principal Components: on Correlations

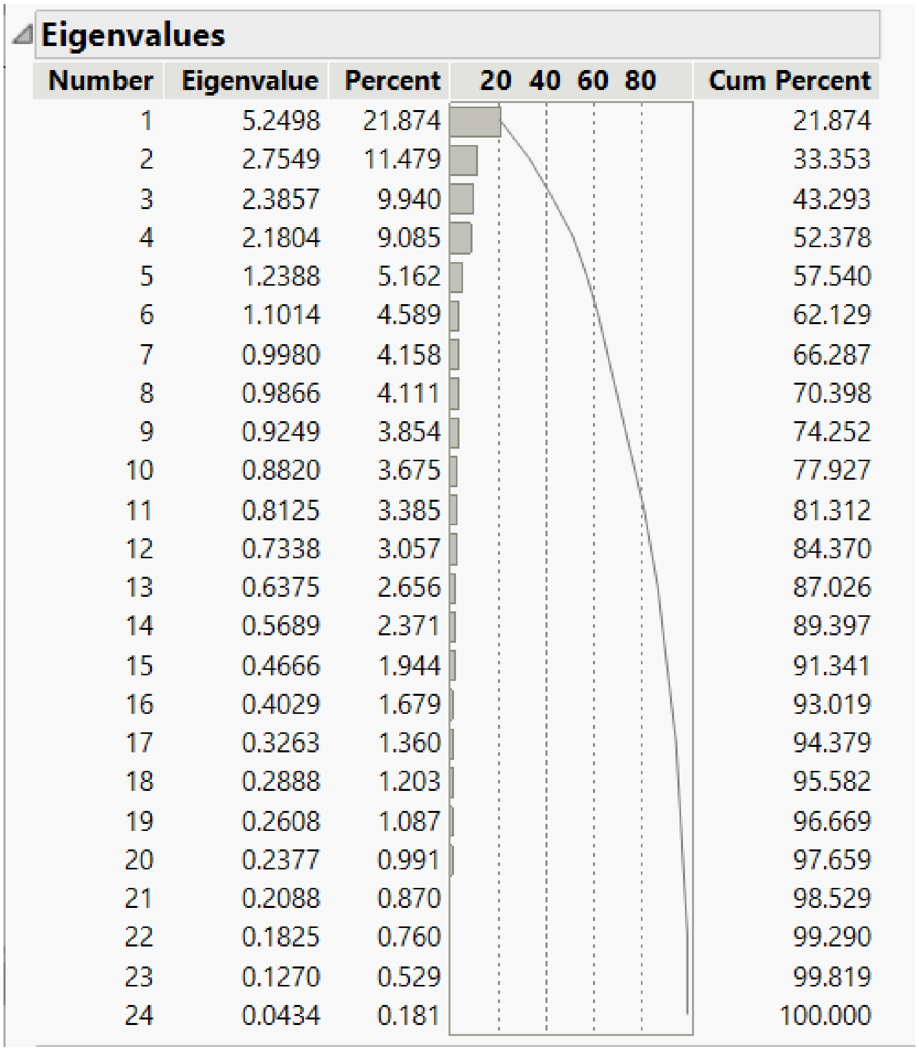
Summary Plots

Eigenvalue
5.2498
2.7549
2.3857
2.1804
1.2388
1.1014
0.9980
0.9866
0.9249
0.8820
0.8125
0.7338

☑ Label variables

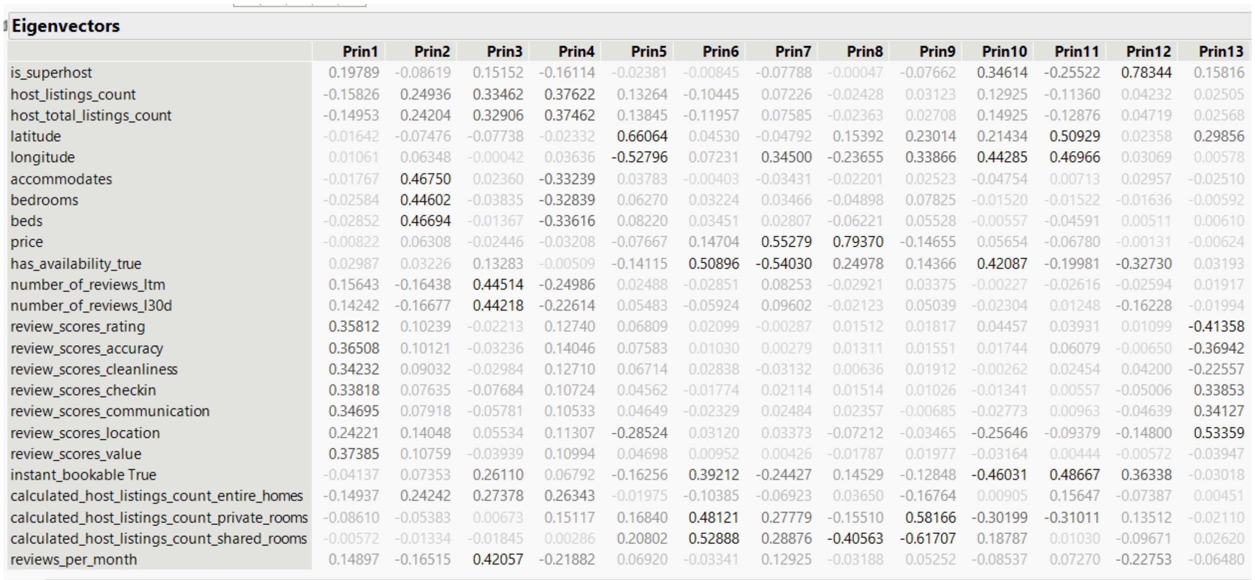
Select component Component 1 Component 2

We can see that 95% of accuracy can be achieved by reducing the column to 18



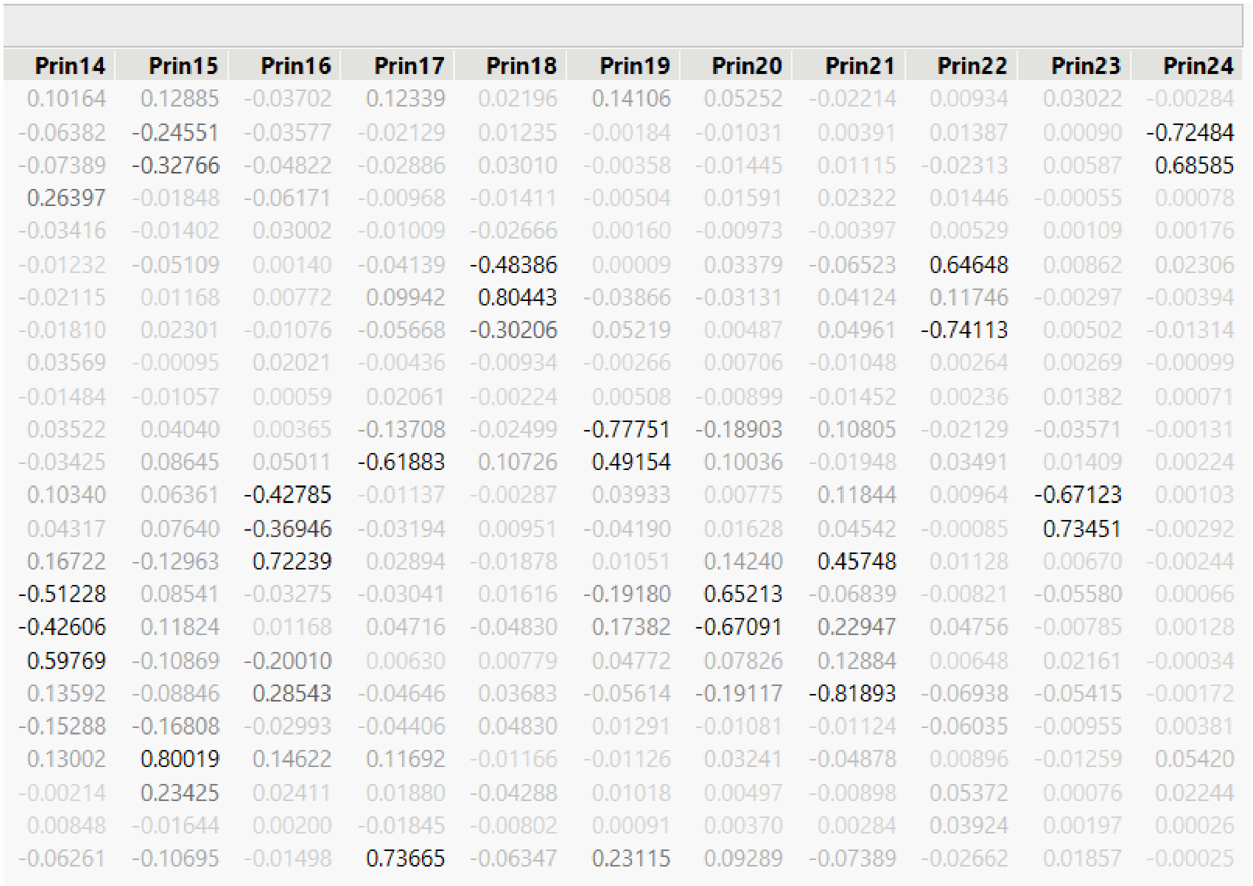
Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	5.2498	21.874		21.874
2	2.7549	11.479		33.353
3	2.3857	9.940		43.293
4	2.1804	9.085		52.378
5	1.2388	5.162		57.540
6	1.1014	4.589		62.129
7	0.9980	4.158		66.287
8	0.9866	4.111		70.398
9	0.9249	3.854		74.252
10	0.8820	3.675		77.927
11	0.8125	3.385		81.312
12	0.7338	3.057		84.370
13	0.6375	2.656		87.026
14	0.5689	2.371		89.397
15	0.4666	1.944		91.341
16	0.4029	1.679		93.019
17	0.3263	1.360		94.379
18	0.2888	1.203		95.582
19	0.2608	1.087		96.669
20	0.2377	0.991		97.659
21	0.2088	0.870		98.529
22	0.1825	0.760		99.290
23	0.1270	0.529		99.819
24	0.0434	0.181		100.000



Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13
is_superhost	0.19789	-0.08619	0.15152	-0.16114	-0.02381	-0.00845	-0.07788	-0.00047	-0.07662	0.34614	-0.25522	0.78344	0.15816
host_listings_count	-0.15826	0.24936	0.33462	0.37622	0.13264	-0.10445	0.07226	-0.02428	0.03123	0.12925	-0.11360	0.04232	0.02505
host_total_listings_count	-0.14953	0.24204	0.32906	0.37462	0.13845	-0.11957	0.07585	-0.02363	0.02708	0.14925	-0.12876	0.04719	0.02568
latitude	-0.01642	-0.07476	-0.07738	-0.02332	0.66064	0.04530	-0.04792	0.15392	0.23014	0.21434	0.50929	0.02358	0.29856
longitude	0.01061	0.06348	-0.00042	0.03636	-0.52796	0.07231	0.34500	-0.23655	0.33866	0.44285	0.46966	0.03069	0.00578
accommodates	-0.01767	0.46750	0.02360	-0.33239	0.03783	-0.00403	-0.03431	-0.02201	0.02523	-0.04754	0.00713	0.02957	-0.02510
bedrooms	-0.02584	0.44602	-0.03835	-0.32839	0.06270	0.03224	0.03466	-0.04898	0.07825	-0.01520	-0.01522	-0.01636	-0.00592
beds	-0.02852	0.46694	-0.01367	-0.33616	0.08220	0.03451	0.02807	-0.06221	0.05528	-0.00557	-0.04591	0.00511	0.00610
price	-0.00822	0.06308	-0.02446	-0.03208	-0.07667	0.14704	0.55279	0.79370	-0.14655	0.05654	-0.06780	-0.00131	-0.00624
has_availability_true	0.02987	0.03226	0.13283	-0.00509	-0.14115	0.50896	-0.54030	0.24978	0.14366	0.42087	-0.19981	-0.32730	0.03193
number_of_reviews_ltm	0.15643	-0.16438	0.44514	-0.24986	0.02488	-0.02851	0.08253	-0.02921	0.03375	-0.00227	-0.02616	-0.02594	0.01917
number_of_reviews_l30d	0.14242	-0.16677	0.44218	-0.22614	0.05483	-0.05924	0.09602	-0.02123	0.05039	-0.02304	0.01248	-0.16228	-0.01994
review_scores_rating	0.35812	0.10239	-0.02213	0.12740	0.06809	0.02099	-0.00287	0.01512	0.01817	0.04457	0.03931	0.01099	-0.41358
review_scores_accuracy	0.36508	0.10121	-0.03236	0.14046	0.07583	0.01030	0.00279	0.01311	0.01551	0.01744	0.06079	-0.00650	-0.36942
review_scores_cleanliness	0.34232	0.09032	-0.02984	0.12710	0.06714	0.02838	-0.03132	0.00636	0.01912	-0.00262	0.02454	0.04200	-0.22557
review_scores_checkin	0.33818	0.07635	-0.07684	0.10724	0.04562	-0.01774	0.02114	0.01514	0.01026	-0.01341	0.00557	-0.05006	0.33853
review_scores_communication	0.34695	0.07918	-0.05781	0.10533	0.04649	-0.02329	0.02484	0.02357	-0.00685	-0.02773	0.00963	-0.04639	0.34127
review_scores_location	0.24221	0.14048	0.05534	0.11307	-0.28524	0.03120	0.03373	-0.07212	-0.03465	-0.25646	-0.09379	-0.14800	0.53359
review_scores_value	0.37385	0.10759	-0.03939	0.10994	0.04698	0.00952	0.00426	-0.01787	0.01977	-0.03164	0.00444	-0.00572	-0.03947
instant_bookable True	-0.04137	0.07353	0.26110	0.06792	-0.16256	0.39212	-0.24427	0.14529	-0.12848	-0.46031	0.48667	0.36338	-0.03018
calculated_host_listings_count_entire_homes	-0.14937	0.24242	0.27378	0.26343	-0.01975	-0.10385	-0.06923	0.03650	-0.16764	0.00905	0.15647	-0.07387	0.00451
calculated_host_listings_count_private_rooms	-0.08610	-0.05383	0.00673	0.15117	0.16840	0.48121	0.27779	-0.15510	0.58166	-0.30199	-0.31011	0.13512	-0.02110
calculated_host_listings_count_shared_rooms	-0.00572	-0.01334	-0.01845	0.00286	0.20802	0.52888	0.28876	-0.40563	-0.61707	0.18787	0.01030	-0.09671	0.02620
reviews_per_month	0.14897	-0.16515	0.42057	-0.21882	0.06920	-0.03341	0.12925	-0.03188	0.05252	-0.08537	0.07270	-0.22753	-0.06480



Prin14	Prin15	Prin16	Prin17	Prin18	Prin19	Prin20	Prin21	Prin22	Prin23	Prin24
0.10164	0.12885	-0.03702	0.12339	0.02196	0.14106	0.05252	-0.02214	0.00934	0.03022	-0.00284
-0.06382	-0.24551	-0.03577	-0.02129	0.01235	-0.00184	-0.01031	0.00391	0.01387	0.00090	-0.72484
-0.07389	-0.32766	-0.04822	-0.02886	0.03010	-0.00358	-0.01445	0.01115	-0.02313	0.00587	0.68585
0.26397	-0.01848	-0.06171	-0.00968	-0.01411	-0.00504	0.01591	0.02322	0.01446	-0.00055	0.00078
-0.03416	-0.01402	0.03002	-0.01009	-0.02666	0.00160	-0.00973	-0.00397	0.00529	0.00109	0.00176
-0.01232	-0.05109	0.00140	-0.04139	-0.48386	0.00009	0.03379	-0.06523	0.64648	0.00862	0.02306
-0.02115	0.01168	0.00772	0.09942	0.80443	-0.03866	-0.03131	0.04124	0.11746	-0.00297	-0.00394
-0.01810	0.02301	-0.01076	-0.05668	-0.30206	0.05219	0.00487	0.04961	-0.74113	0.00502	-0.01314
0.03569	-0.00095	0.02021	-0.00436	-0.00934	-0.00266	0.00706	-0.01048	0.00264	0.00269	-0.00099
-0.01484	-0.01057	0.00059	0.02061	-0.00224	0.00508	-0.00899	-0.01452	0.00236	0.01382	0.00071
0.03522	0.04040	0.00365	-0.13708	-0.02499	-0.77751	-0.18903	0.10805	-0.02129	-0.03571	-0.00131
-0.03425	0.08645	0.05011	-0.61883	0.10726	0.49154	0.10036	-0.01948	0.03491	0.01409	0.00224
0.10340	0.06361	-0.42785	-0.01137	-0.00287	0.03933	0.00775	0.11844	0.00964	-0.67123	0.00103
0.04317	0.07640	-0.36946	-0.03194	0.00951	-0.04190	0.01628	0.04542	-0.00085	0.73451	-0.00292
0.16722	-0.12963	0.72239	0.02894	-0.01878	0.01051	0.14240	0.45748	0.01128	0.00670	-0.00244
-0.51228	0.08541	-0.03275	-0.03041	0.01616	-0.19180	0.65213	-0.06839	-0.00821	-0.05580	0.00066
-0.42606	0.11824	0.01168	0.04716	-0.04830	0.17382	-0.67091	0.22947	0.04756	-0.00785	0.00128
0.59769	-0.10869	-0.20010	0.00630	0.00779	0.04772	0.07826	0.12884	0.00648	0.02161	-0.00034
0.13592	-0.08846	0.28543	-0.04646	0.03683	-0.05614	-0.19117	-0.81893	-0.06938	-0.05415	-0.00172
-0.15288	-0.16808	-0.02993	-0.04406	0.04830	0.01291	-0.01081	-0.01124	-0.06035	-0.00955	0.00381
0.13002	0.80019	0.14622	0.11692	-0.01166	-0.01126	0.03241	-0.04878	0.00896	-0.01259	0.05420
-0.00214	0.23425	0.02411	0.01880	-0.04288	0.01018	0.00497	-0.00898	0.05372	0.00076	0.02244
0.00848	-0.01644	0.00200	-0.01845	-0.00802	0.00091	0.00370	0.00284	0.03924	0.00197	0.00026
-0.06261	-0.10695	-0.01498	0.73665	-0.06347	0.23115	0.09289	-0.07389	-0.02662	0.01857	-0.00025