OPIM 5604 - PREDICTIVE MODELING

Group Assignment – Airbnb Data: Shanghai

Submitted by – Team 5

Agrawal, Shivani
Kalidindi, Peeyush Varma
Krishnan, Nandita
Sinha, Ayush
Xia, Liyun

This file documents all the processes related to data exploration and pre-processing for to Airbnb Shanghai data.

STEP 1: Data Exploration (Refer appendix for screenshots)

Airbnb – Shanghai, China - Columns 1 to 15

S. No	COLUMN NAME – *Modeling Type*	REASON for Inclusion/Exclusion	Exclusion
1	Id - *CONTINUOUS*	Unique identification number for each listing - Not relevant for modeling target variable	Yes
2	listing_url - *NOMINAL*	URL for listing not needed for modeling target variable.	Yes
3	scrape_id - *CONTINUOUS*	Not relevant for target variable - **Common for all listings - 20210731170350**	Yes
4	last_scraped - *CONTINUOUS*	Only 3 values with 1 month difference, doesn't give much information for correlation with the target variable	Yes
5	Name - *NOMINAL*	No correlation or connection with values of the target variable – Conversion from Chinese to English not proper.	Yes
6	Description - *NOMINAL*	No correlation or connection with values of the target variable – Conversion from Chinese to English not proper.	Yes
7	neighborhood_overview - *NOMINAL*	No correlation or connection with values of the target variable – Conversion from Chinese to English not proper.	Yes
8	picture_url - *NOMINAL*	No correlation or connection with values of the target variable - **All listings have pictures**	Yes
9	host_id - *CONTINUOUS*	No correlation or connection with values of the target variable	Yes
10	host_url - *NOMINAL*	No correlation or connection with values of the target variable	Yes
11	host_name - *NOMINAL*	No correlation or connection with values of the target variable	Yes
12	host_since - *CONTINUOUS*	**Host_since** is not relevant **to predict the target variable** since not linked to listing AGE	Yes
13	host_location - *NOMINAL*	**Host location** is not relevant **to predict the target variable** since not linked to listing location	Yes
14	host_about - *NOMINAL*	Not relevant for created model for Review_Score_Ratings for a listing	Yes
15	host_response_time - *NOMINAL*	**Can be used to model the correlation between RESPONSE_TIME vs RATINGs**	No

Airbnb – Shanghai, China - Columns 16 to 30

S. No	COLUMN NAME – *Modeling Type*	REASON for Inclusion/Exclusion	Exclusion
1	host_response_rate - *NOMINAL*	Give us the rate at which a host accepts booking requests - Can be used to model the correlation between host_response_rate vs RATINGs	No
2	host_acceptance_rate - *NOMINAL*	Can be used to model the correlation between host_acceptance_rate vs RATINGs	No
3	host_is_superhost - *NOMINAL*	Can be used to model the correlation between host_is_superhost vs RATINGs	No
4	host_thumbnail_url - *NOMINAL*	URL for listing not needed for modeling target variable.	Yes
5	host_picture_url - *NOMINAL*	URL for listing not needed for modeling target variable.	Yes
6	host_neighbourhood - *NOMINAL*	No correlation or connection with values of the target variable, there are too many kinds of neighborhood.	Yes
7	host_listings_count - *CONTINUOUS*	They are all same with host_total_listings_count.	Yes
8	host_total_listings_count - *CONTINUOUS*	Give us the number of listings the host has – Not relevant to the performance of individual listings.	Yes
9	host_verifications - *NOMINAL*	All the values are T. It is not helpful.	Yes
10	host_has_profile_pic - *NOMINAL*	~99.9% of the values are TRUE, hence we are excluding it since it will unnecessarily add complexity without adding any insights.	Yes
11	host_identity_verified - *NOMINAL*	All the values are T. It is not helpful.	Yes
12	Neighbourhood - *NOMINAL*	All the values are Shanghai. It is not helpful.	Yes
13	neighbourhood_cleansed - *NOMINAL*	Can be used to model the correlation between AREA/CITY vs RATINGs	No
14	neighbourhood_group_cleansed - *NOMINAL*	The neighborhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles - Can be used to model correlation for host_total_listings_count vs RATINGs	Yes
15	Latitude - *CONTINUOUS*	All the values are N/A. It is not helpful.	Yes

Airbnb – Shanghai, China - Columns 31 to 45

S. No	COLUMN NAME – *Modeling Type*	REASON for Inclusion/Exclusion	Exclusion
1	longitude - *CONTINUOUS*	Column **neighbourhood_cleansed** can be utilized to judge the effect of area on the target variable -- Longitude precision may not be required.	Yes
2	property_type - *NOMINAL*	property_type is described by the host and hence can be subjective/vague. room_type, on the other hand, categorizes the property into three relevant types which can be used to predict the target variable.	Yes
3	room_type - *NOMINAL*	May have an impact target on target variable. Create 6 Indicator columns to display this information	No
4	accommodates - *CONTINUOUS*	Used in combination with price to generate a single column with the formula Price/Accommodates that generates more value in predicting target variable.	Yes
5	Bathrooms - *NOMINAL*	Blank Column	Yes
6	bathrooms_text - *NOMINAL*	Number of bathrooms doesn't add value by itself. Column Price/Accommodates to give required information	Yes
7	Bedrooms - *CONTINUOUS*	Bedrooms has a lot of missing values --- accommodates can be used instead, as it has a strong correlation	Yes
8	Beds - *CONTINUOUS*	Beds has a lot of missing values --- accommodates can be used instead, as it has a strong correlation	Yes
9	Amenities - *NOMINAL*	Created new 197 indicator columns to represent the data in amenities and checked correlation of each with target variable - correlation is low. PCA also provides value only by retaining at least 60 columns in turn adding complexity to the model	Yes
10	price - *CONTINUOUS*	Used in combination with Accommodates to generate a single column with the formula Price/Accommodates that generates more value in predicting target variable.	Yes
11	minimum_nights - *CONTINUOUS*	Minimum number of nights allowed in a stay might impact the target variable	No
12	maximum_nights - *CONTINUOUS*	Maximum number of nights allowed in a stay might impact the target variable	No
13	minimum_minimum_nights - *CONTINUOUS*	Provides same information as minimum_nights	Yes
14	maximum_minimum_nights - *CONTINUOUS*	Provides same information as minimum_nights	Yes
15	minimum_maximum_nights - *CONTINUOUS*	Irrelevant to targret_variable – low correlation	Yes

Airbnb – Shanghai, China - Columns 46 to 60

S. No	COLUMN NAME – *Modeling Type*	REASON for Inclusion/Exclusion	Exclusion
1	maximum_maximum_nights - *CONTINUOUS*	It describes the maximum nights a customer has stayed. If we look in a business perspective the number of nights stayed won't affect the rating of an Airbnb. So, this column won't be included.	Yes
2	minimum_nights_avg_ntm - *CONTINUOUS*	It describes the average minimum nights a customer has stayed. The ratings don't depend on the period a customer has stayed at the Airbnb. So, this column won't be included.	Yes
3	maximum_nights_avg_ntm - *CONTINUOUS*	It describes the average maximum nights a customer has stayed. The ratings don't depend on the period a customer has stayed at the Airbnb. So, we won't include.	Yes
4	calendar_updated - *CONTINUOUS*	As this column has 26977 missing values, it will be excluded.	Yes
5	has_availability - *NOMINAL*	Will be removing this column because the whole column contains only 1 value 't'.	Yes
6	availability_30 - *CONTINUOUS*	This column might represent the demand an Airbnb has. If there is no availability for the next 30 days, we can assume it has good ratings because high demand means a relatively good rating. So, we won't include.	No
7	availability_60 - *CONTINUOUS*	Same as availability_30 but for 60 days. So, this column will be included.	No
8	availability_90 - *CONTINUOUS*	Same as availability_30 but for 90 days. So, this column will be included.	No
9	availability_365 - *CONTINUOUS*	Same as availability_30 but for 365 days. So, this column will be included.	No
10	calendar_last_scraped - *NOMINAL*	This column won't be included as when the data was scraped doesn't affect the ratings.	Yes
11	number_of_reviews - *CONTINUOUS*	Will be not including this column because having a greater number of ratings does not point to better review_score_ratings. Additionally, high number of outliers.	Yes
12	number_of_reviews_ltm - *CONTINUOUS*	Will not be using this column because it represents the same information as the number_of_reviews.	Yes
13	number_of_reviews_l30d - *CONTINUOUS*	Will not be using this column because it represents the same information as the number_of_reviews.	Yes
14	first_review - *NOMINAL*	9807 missing values. Will not be using this column because it doesn't matter when the first review was for measuring the Review_Score_Ratings.	Yes
15	last review - *NOMINAL*	9807 missing values. Will not be using this column because it doesn't matter when the last review was for measuring the Review_Score_Ratings.	Yes

Airbnb – Shanghai, China - Columns 61 to 74

S. No	COLUMN NAME – *Modeling Type*	REASON for Inclusion/Exclusion	Exclusion
1	review_score_ratings - *CONTINUOUS*	It is a target variable; it has 9807 missing values so we excluded those rows right away as we cannot impute the target variable.	No
2	review_scores_accuracy - *CONTINUOUS*	Based on the Airbnb website, the target variable is calculated using some statistical combination (not average) of these 6 sub-ratings	No
3	review_scores_cleanliness - *CONTINUOUS*		
4	review_scores_checkin - *CONTINUOUS*		
5	review_scores_communication - *CONTINUOUS*		
6	review_scores_location - *CONTINUOUS*		
7	review_scores_value - *CONTINUOUS*		
8	License - *NOMINAL*	This variable has 26977(all) missing values so we are excluding this variable.	Yes
9	instant_bookable - *NOMINAL*	This variable indicates whether the guest can automatically book the listing without the host requiring to accept their booking request, important for the target variable.	No
10	calculated_host_listings_count - *CONTINUOUS*	The number of listings the host has in the current scrape, in the city/region geography, we are excluding since we're not keeping counts of entire private, shared rooms	Yes
11	calculated_host_listings_count_entire_homes - *CONTINUOUS*	The number of Entire home/apt listings the host has in the current scrape, in the city/region geography, so we are excluding it as the rating will not be same for every listing.	Yes
12	calculated_host_listings_count_private_rooms - *CONTINUOUS*	The number of Private room listings the host has in the current scrape, in the city/region geography, so we are excluding it as the rating will not be same for every listing.	Yes
13	calculated_host_listings_count_shared_rooms - *CONTINUOUS*	The number of Shared room listings the host has in the current scrape, in the city/region geography, so we are excluding it as the rating will not be same for every listing.	Yes
14	reviews_per_month - *CONTINUOUS*	It standardizes the total number of reviews of a listing by dividing it with the total duration of listing in months.	No

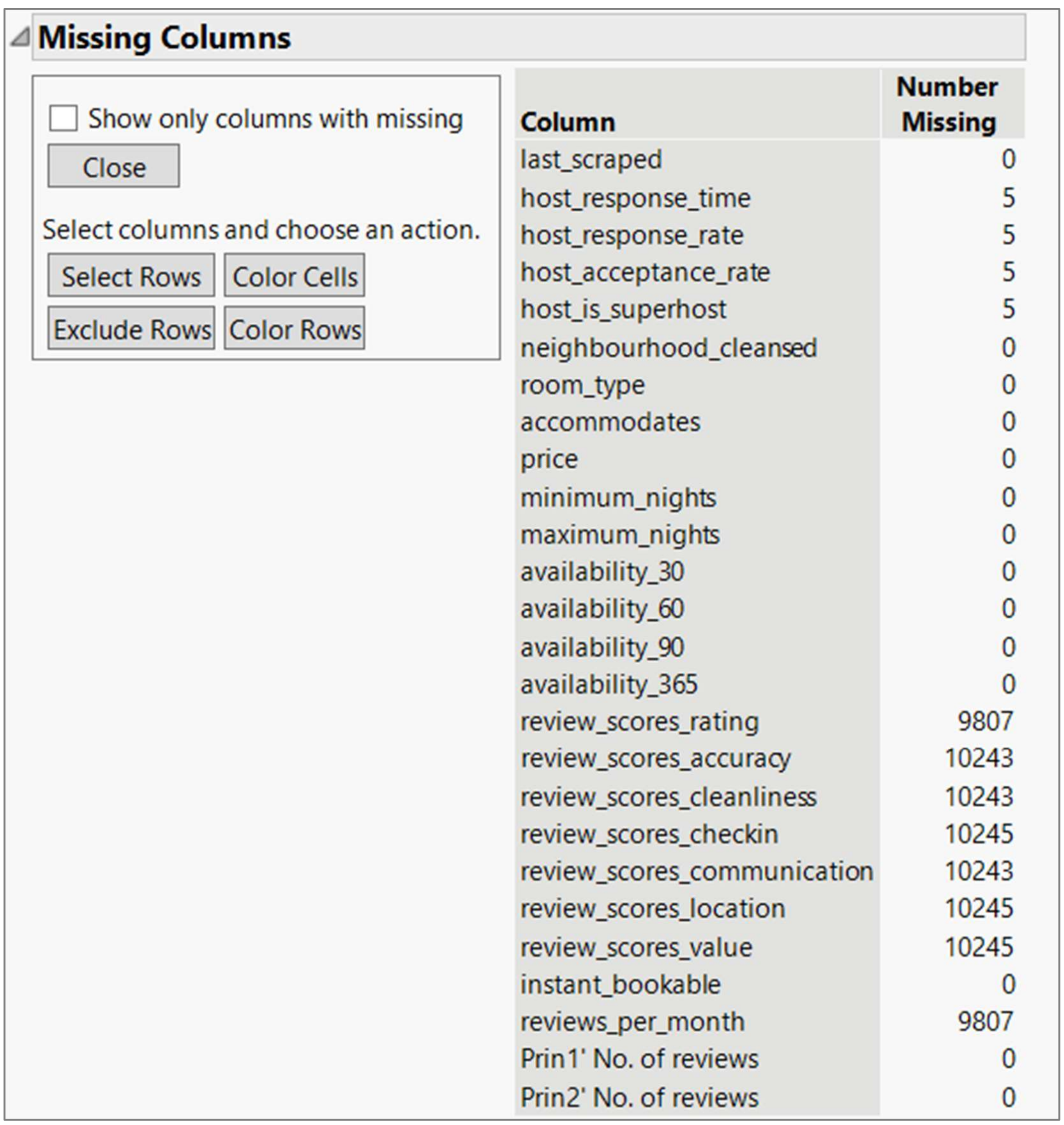
STEP 2: Variable Type Conversion

S.NO	Variable	Existing type	New type	Description
1	last_scraped	CONTINUOUS	NOMINAL	Only 3 unique values without order.
2	host_response_rate	NOMINAL	CONTINUOUS	Process – We used column info to update the data type to “numeric” and modeling type to “continuous”
3	host_acceptance_rate	NOMINAL	CONTINUOUS	Process – We used column info to update the data type to “numeric” and modeling type to “continuous”

STEP 3: Missing Values

PROCESS

1. Deleting **9,807** missing values from target variable – “**review_score_ratings”**



Column	Number Missing
last_scraped	0
host_response_time	5
host_response_rate	5
host_acceptance_rate	5
host_is_superhost	5
neighbourhood_cleansed	0
room_type	0
accommodates	0
price	0
minimum_nights	0
maximum_nights	0
availability_30	0
availability_60	0
availability_90	0
availability_365	0
review_scores_rating	9807
review_scores_accuracy	10243
review_scores_cleanliness	10243
review_scores_checkin	10245
review_scores_communication	10243
review_scores_location	10245
review_scores_value	10245
instant_bookable	0
reviews_per_month	9807
Prin1' No. of reviews	0
Prin2' No. of reviews	0

2. Additional 200 rows deleted
3. Imputing values
    - **host_response_time –**
        - 1,494 – N/A values
        - Since it’s categorical, we calculated the MODE (=” within an hour”) and imputed the value for the respective rows
    - **host_response_rate –**
        - 1,494 – N/A values
        - Since it’s a continuous variable, we calculated the MEAN (=0.96) and imputed the value for the respective rows
    - **host_acceptance_rate –**
        - 1,056 – N/A values
        - Since it’s a continuous variable, we calculated the MEAN (=0.95) and imputed the value for the respective rows.

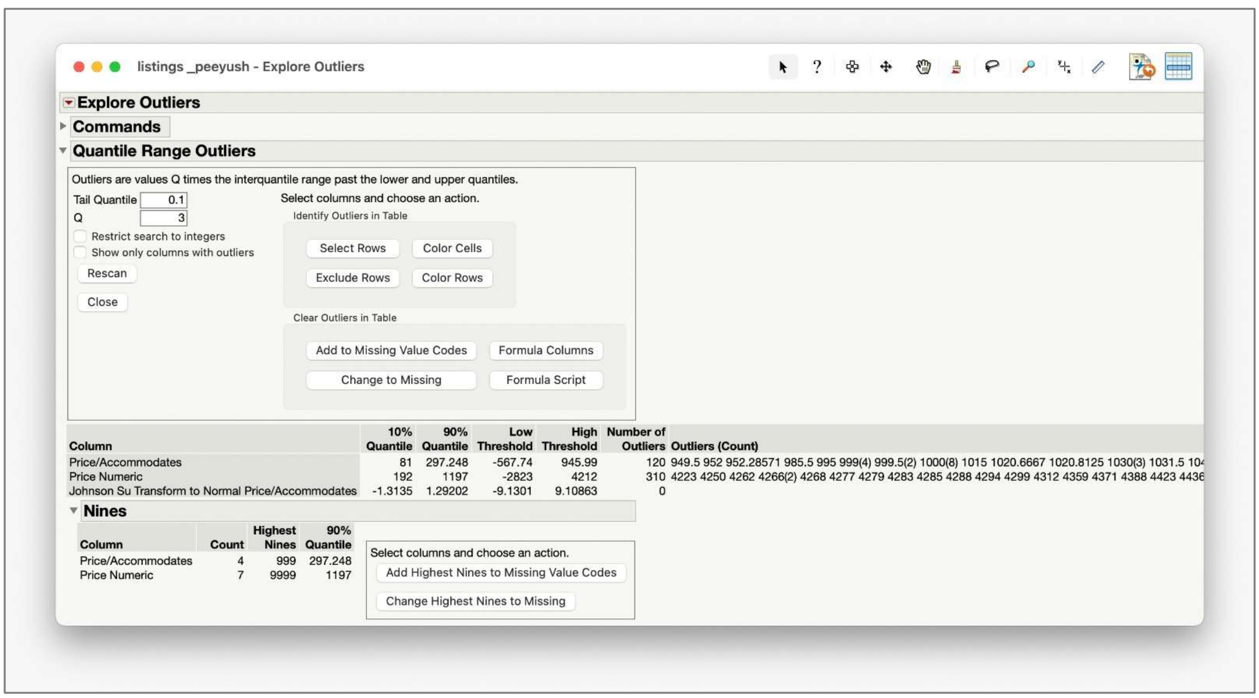
STEP 4: Outlier Analysis

1. **Step 1** – Check for the distribution
2. **Step 2** – Apply transformation
3. **Step 3** – Select the best fit transformation
4. **Step 4** – Save the transformed column to the dataset

VARIABLE	METHOD	BEST FIT
host_response_rate	VARIABLE TRANSFORM	SHASH
host_acceptance_rate	VARIABLE TRANSFORM	SHASH
min_nights	VARIABLE TRANSFORM	SHASH
price/accommodates***	VARIABLE TRANSFORM	JOHNSON SU
max_night	Exclude Records*	NA
Reviews_per_month	Exclude Records**	NA

** There were **only 2 outliers**, hence we **excluded it from the dataset**.*
*** Excluding outliers for **max_night** removed the **only 3 outliers** from **Reviews_per_month** as well*
**** Created using Price and Accommodates – Variable still had 120 outliers that were tackled using transformation*

<u>Outlier Analysis for "Price/Accommodates"</u>



STEP 5: Dummy Variables

(The following variable columns were created as CONTINUOUS, we changed them to NOMINAL)

1. **host_response_time –**

    - 4 variables columns created as dummy variables
    - We keep only 3 and hide "**Within an hour**" since we only need (n-1) columns
    - We hide "host_response_time"

2. **host_is_superhost –**

    - Text values for variables needs to be converted to 2 dummy variables
    - We keep only 1 and hide "**False**" since we only need (n-1) columns
    - We hide "host_is_superhost"

3. **neighbourhood_cleansed –**

    - 16 variable columns created as dummy variables
    - We keep only 15 and hide "**Jiading District**" since we only need (n-1) columns
    - We hide "neighbourhood_cleansed"

4. **room_type** –

    - 3 variable columns created as dummy variables
    - We keep only 2 and hide "**shared room**" since we only need (n-1) columns
    - We hide "**room_type**"

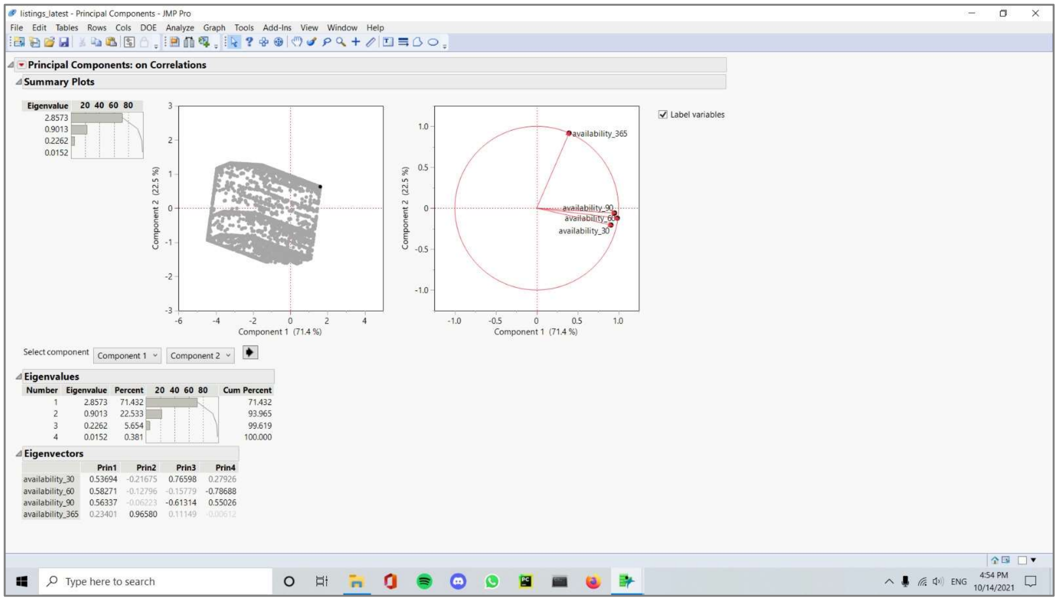
5. **instant_bookable –**

    - Text values for variables needs to be converted to 2 dummy variables
    - We keep only 1 and hide "**False**" since we only need (n-1) columns
    - We hide "**instant_bookable**"

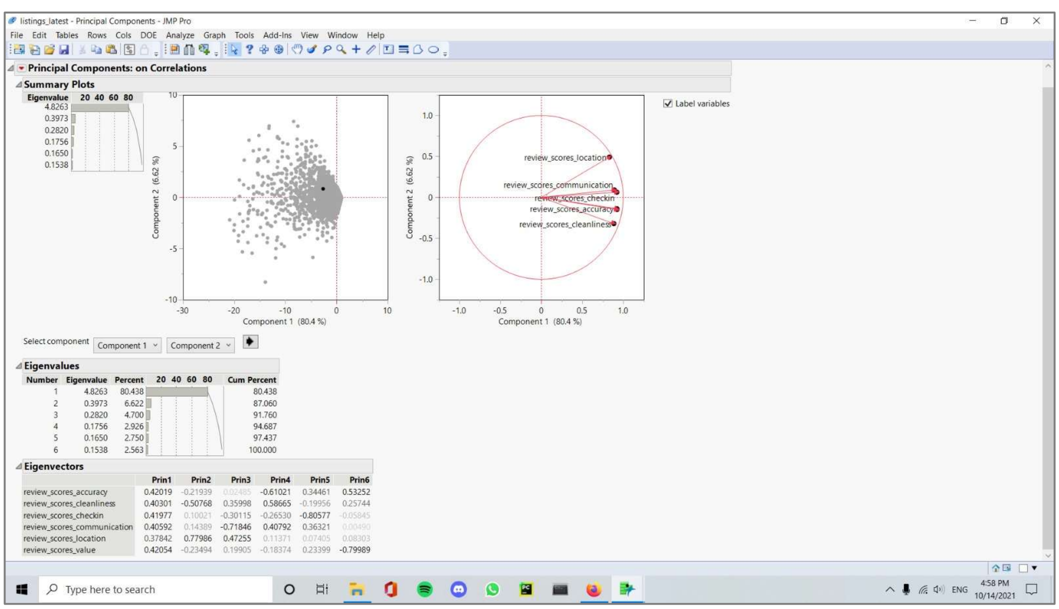
STEP 6: Reducing Dimensionality

We ran the **Principal Components Analysis** on <u>two</u> sets of variables with the intention of reducing attributes needed for predictive modeling.

1. **SET 1** - Availability_30, Availability_60, Availability_90, Availability_365
   a. **Results** –
      i. Variance – **93.96%** using 2 Principal Components
      ii. Hence, we were able to reduce the variables needed for modeling from 4 down to 2

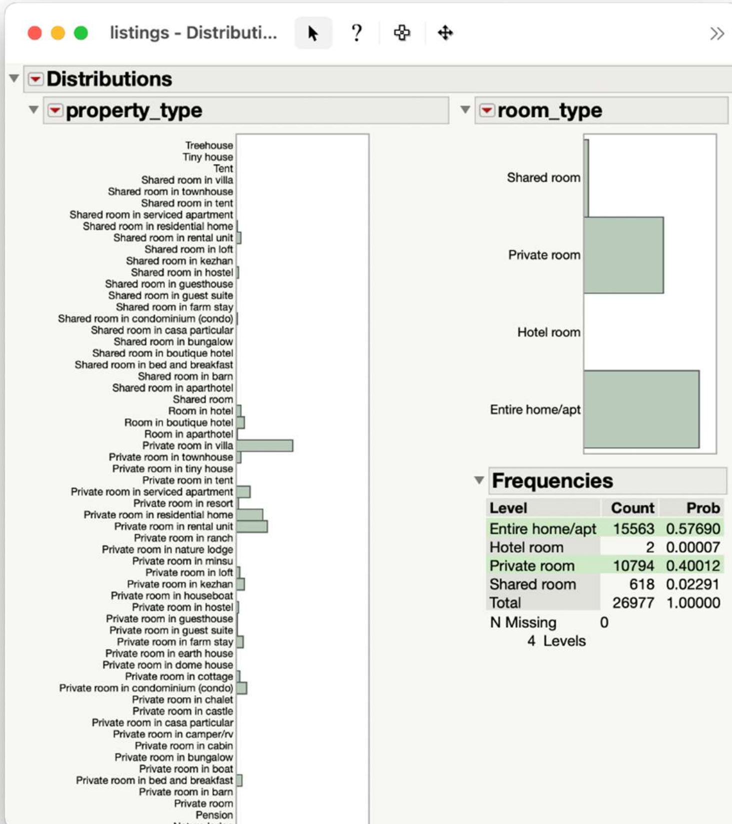


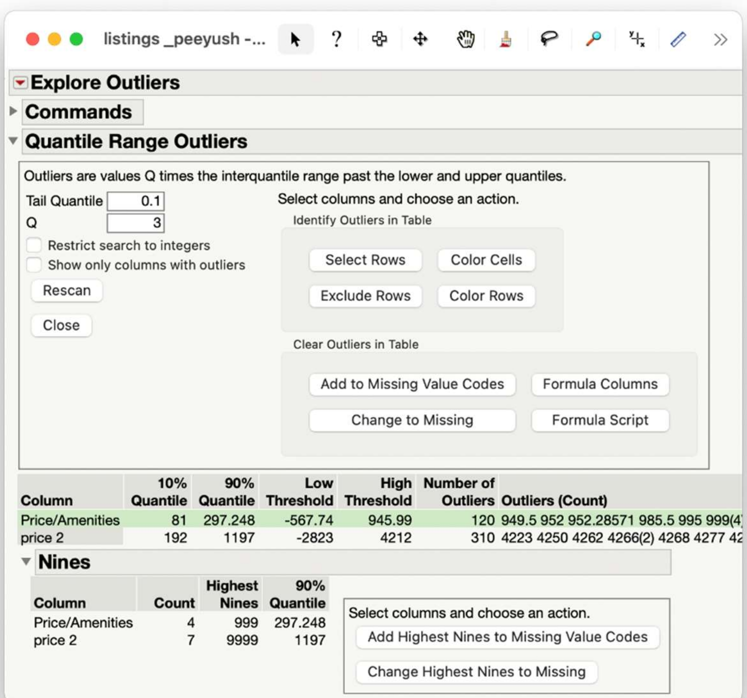
2. **SET 2** - Review_scores_accuracy, Review_scores_cleanliness, Review_scores_checkin, Review_scores_communication, Review_scores_location, Review_scores_value
   a. **Results** –
      i. Variance – **91.76%** using 3 Principal Components
      ii. Hence, we were able to reduce the variables needed for modeling from 6 down to 3

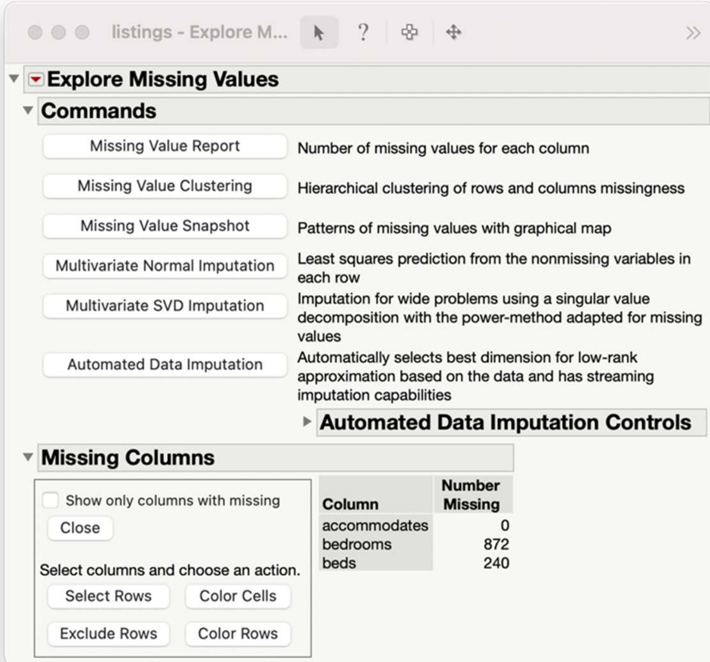


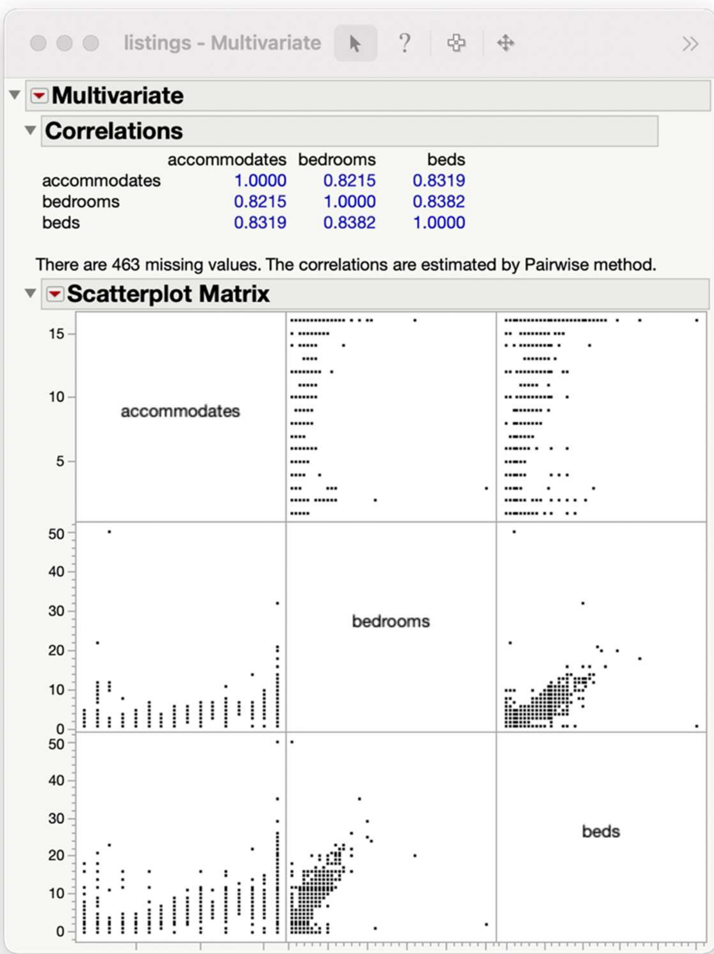
Appendix

Columns 31-45:






listings - Explore M...
Explore Missing Values
Commands
Missing Value Report
Number of missing values for each column
Missing Value Clustering
Hierarchical clustering of rows and columns missingness
Missing Value Snapshot
Patterns of missing values with graphical map
Multivariate Normal Imputation
Least squares prediction from the nonmissing variables in each row
Multivariate SVD Imputation
Imputation for wide problems using a singular value decomposition with the power-method adapted for missing values
Automated Data Imputation
Automatically selects best dimension for low-rank approximation based on the data and has streaming imputation capabilities
Automated Data Imputation Controls
Missing Columns
Show only columns with missing
Close
Select columns and choose an action.
Select Rows
Color Cells
Exclude Rows
Color Rows
Column
Number Missing
accommodates 0
bedrooms 872
beds 240


listings - Multivariate
Multivariate
Correlations
accommodates bedrooms beds
accommodates 1.0000 0.8215 0.8319
bedrooms 0.8215 1.0000 0.8382
beds 0.8319 0.8382 1.0000
There are 463 missing values. The correlations are estimated by Pairwise method.
Scatterplot Matrix
accommodates
bedrooms
beds
15
10
5
50
40
30
20
10
0

Columns 46-60: