**CONTENTS**

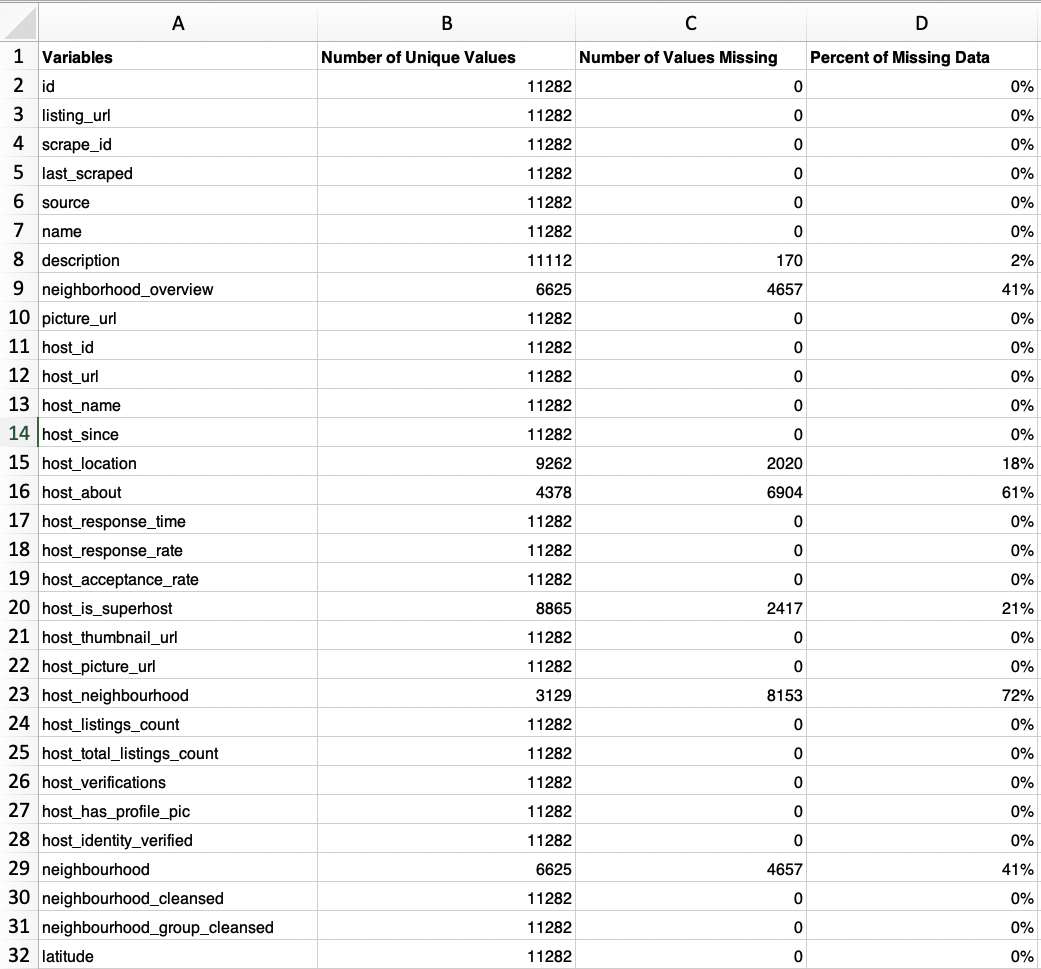
|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Title** | **Page. No** |
| 1 | Project overview | 1 |
| 2 | Variable summary | 1 |
| 3 | Variable visualization and exploration | 3 |
| 4 | Missing values | 18 |
| 5 | Outlier Analysis | 20 |
| 6 | Conclusion | 22 |

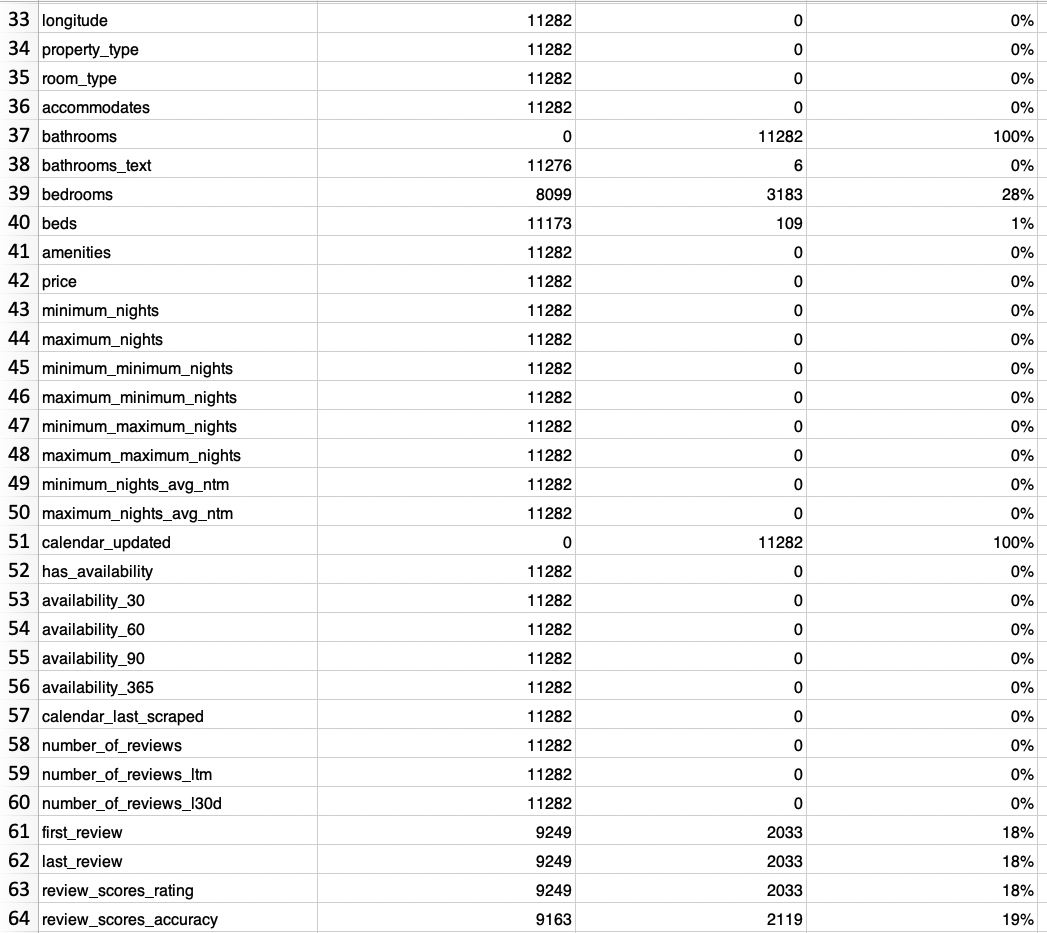
1. **Project Overview:**

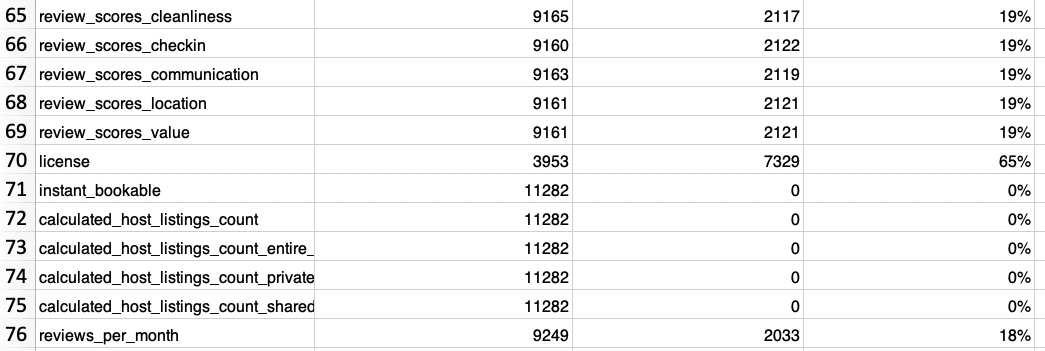
I was tasked with the project of preparing a dataset for subsequent modeling to predict the prices of Airbnb listings in Bordeaux, Nouvelle-Aquitaine, France. The dataset comprises essential information and metrics for Airbnb listings, encompassing a total of 75 columns and 11,282 rows. This data was sourced from the [Airbnb website](http://insideairbnb.com/get-the-data.html). Using JMP, each variable in the Airbnb dataset is analyzed in terms of correlation with the variable "price" as the target variable. Variables that are (1) numerical and highly correlated with "price" or (2) non-numerical but have identification value are maintained in the dataset.In the preprocessing phase, we thoroughly examined the variables that could influence pricing, addressed missing data, and handled outliers through a variety of techniques. A comprehensive summary of our process is detailed in the accompanying report.

1. **Variables Summary:**

Below is a table that displays (A) each variable in the data set, (B) the number of unique values for each variable, (C) number of values missing for each variable, (D) percentage of data missing for each variable. Twenty-two (22) variables in the data set are missing values.



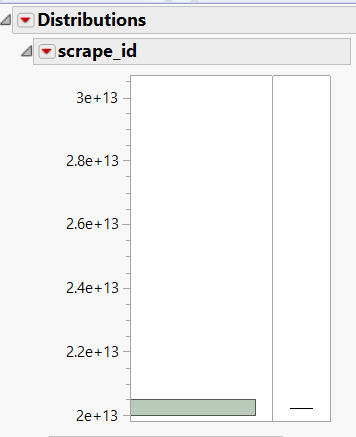




1. **Variable visualization and exploration:**

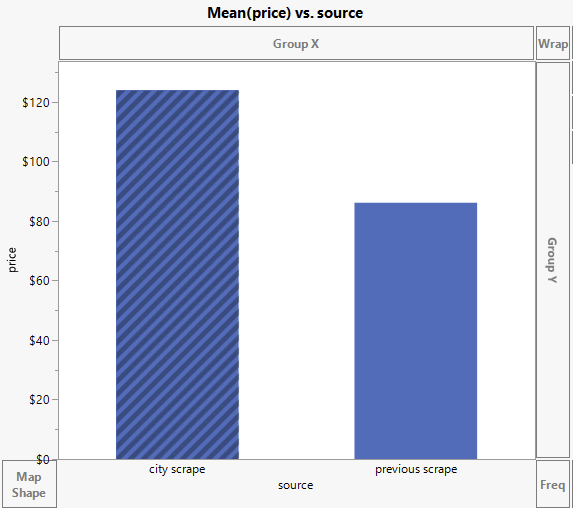
We conducted a comprehensive exploration of all the variables, considering both the nature of the data contained in each column and its potential influence on the target variable, 'price.' Additionally, we utilized visualization techniques to identify and address missing values within the dataset. As a result of this analysis, we made informed decisions about which variables to exclude or retain. Some columns were further categorized or recorded, and we also identified instances of 'hidden missing values.' A detailed summary for each variable is provided below:

1. **id:** which serves as Airbnb's distinct identifier for each listing, has been omitted from consideration as it has no bearing on the target variable.
2. **Listing\_url:** This column corresponds to a distinctive web URL for the listing, but it is not a factor that affects the target variable. Therefore, we have chosen to exclude this column from our analysis.
3. **Scrape\_id:** This column should be removed because it is the same across all fields. It will not provide any meaningful information therefore it should be excluded.



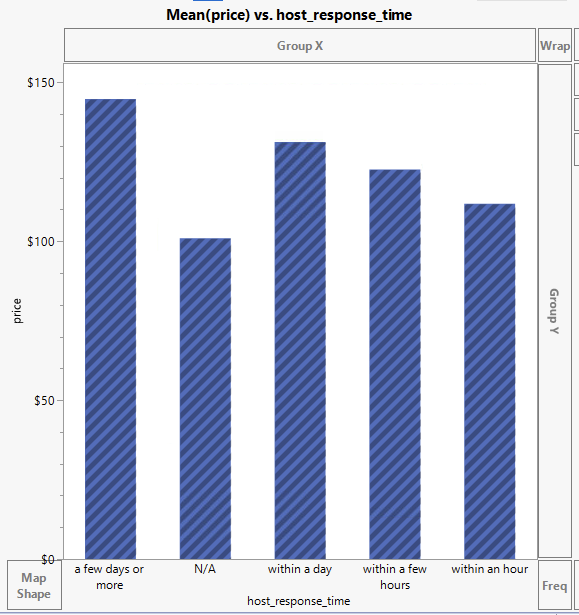
1. **Last\_scraped**: This particular column denotes the most recent date on which data for this listing was collected. Given that the numerical date value bears no relevance to the pricing, we have opted to omit this variable from our analysis.

1. **Source:** We willkeep this column because it will help differentiate a city location from a location elsewhere. We see a correlation between source and pricing.

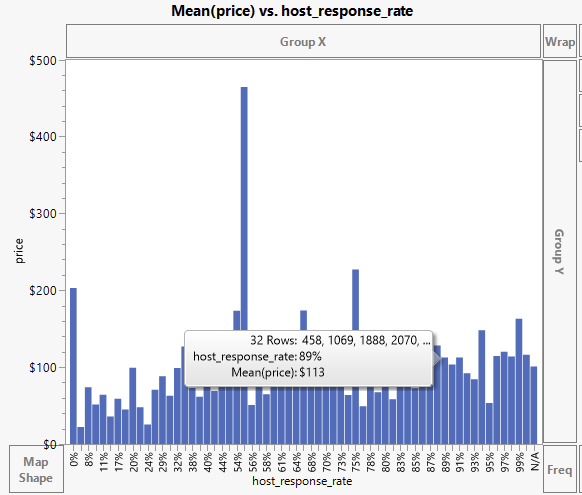


1. **Name:** This column contains unique names for each listing, and since it does not have any bearing on the price, we have chosen to exclude it from our analysis.
2. **Description:** This column provides a detailed description of each listing. It's considered a nominal variable with entirely unique values, and as such, it doesn't add meaningful information to our analysis. Therefore, we have excluded this column from our dataset.
3. **Neighborhood\_overview:** This should be removed as there is 41% of missing data. It contains the host description of the neighborhood. Therefore, we will remove this from our dataset.
4. **Picture\_url:** This column contains the URL for images of the Airbnb listings, and it has no influence on the target variable. As a result, we have omitted this column from our analysis.
5. **Host\_Id:** This column contains unique Airbnb IDs for hosts, and it does not have any effect on the target variable. Therefore, we have opted to exclude this column from our analysis.
6. **host-url:** This column contains unique Airbnb host URLS, and it does not have any effect on the target variable. Therefore, we have opted to exclude this column from our analysis.
7. **Host\_name:** This column contains unique Airbnb host names, and it does not have any effect on the target variable. Therefore, we have opted to exclude this column from our analysis.
8. **Host\_since:** This column records the registration date of the host with Airbnb. We recognize that this parameter, the host's experience, can have an impact on the price, as more experienced hosts may excel in terms of customer satisfaction. Therefore, we have retained this column for our analysis.
9. **Host location:** This column contains information about the host's city of residence. However, it does not have any discernible impact on the target variable, and as a result, it also contains many missing values. So we have chosen to exclude this column from our analysis.

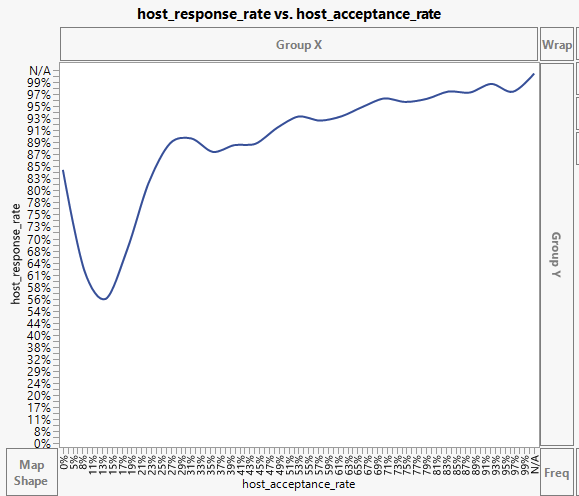
1. **Host\_about:** This column displays information about the host. It does not have any direct impact on the target variable (price) and therefore will be removed from our analysis.
2. **host \_ reponse\_time:** This column records the response time of hosts, and it has the potential to influence the price. The bar graph illustrates that all response time categories have different mean prices, suggesting a correlation between response time and pricing. Consequently, we have included this variable in our analysis, and it's important to note that the data type for this variable is nominal.



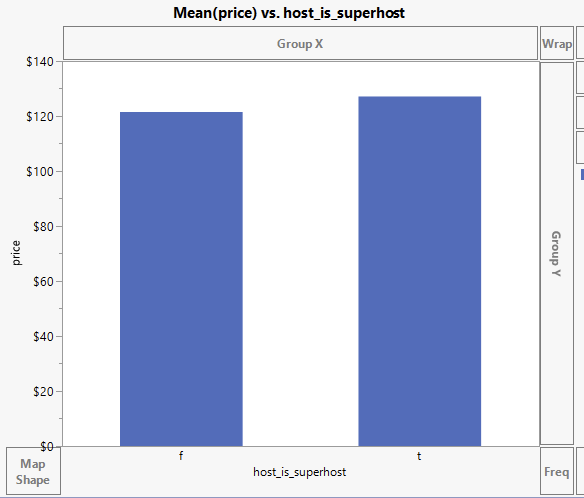
1. **Host\_response\_rate:** Host\_response\_rate is a crucial column that we will retain in our analysis because it provides valuable insights into the host's responsiveness, which can directly impact the target variable pricing.



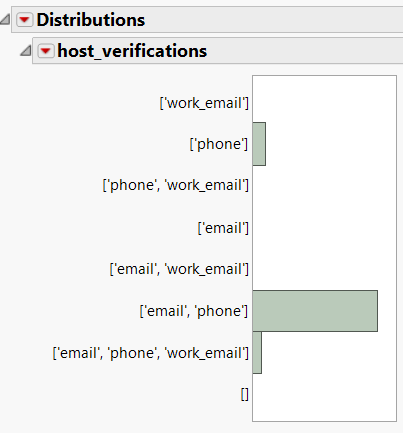
1. **Host\_ acceptance\_ rate:** Host\_acceptance\_rate is a crucial column that can directly impact the target variable pricing. As we see in the graph below, Host\_response\_rate and Host\_acceptance\_rate are positively correlated, we can exclude the Host\_acceptance\_rate from our analysis to reduce model complexity.



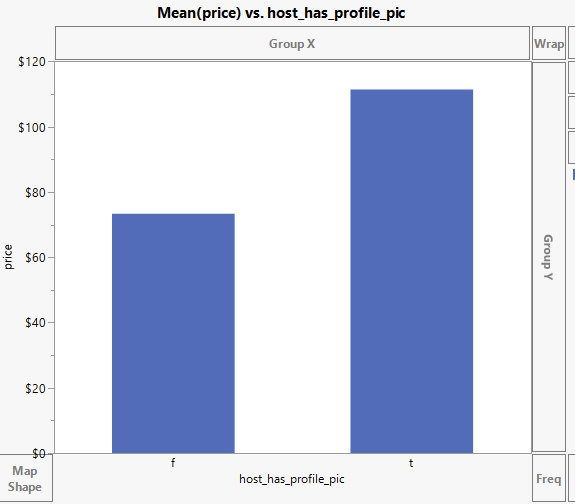
1. **host\_is \_superhost:** We have decided to remove this column because there is not much of a correlation between price. It also has 21% missing values and therefore it will be omitted from the model.



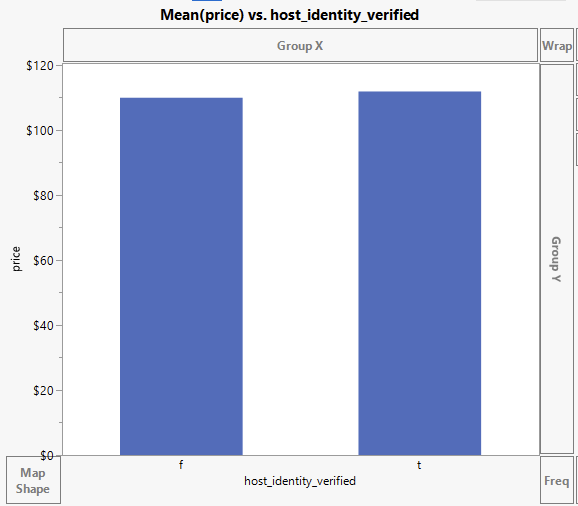
1. **Host thumbnail url**: This column contains the URL of the host's thumbnail image for the listing on Airbnb. However, it has no discernible impact on the target variable. As a result, we have decided to exclude this column from our analysis.
2. **Host\_picture\_url:** This column contains the URLs of the host's images for the Airbnb listing, but it does not have any influence on the target variable. In light of this, we have chosen to exclude this column from our analysis.
3. **Host\_neighborhood:** This column contains information about the host's neighborhood, but it is not expected to have an impact on the price. Additionally, it contains a high percentage of missing values (72%), which is why we have excluded this column from our analysis.
4. **Host\_listings\_count:** This column indicates the number of listings a host has on Airbnb. Since a host may have a combination of both low-priced and high-priced units in their portfolio, making it challenging to predict the price of a specific listing solely based on this information, we have decided to exclude this column from our analysis.
5. **Host\_total\_ listings\_count:** This column represents the total number of listings a host has on Airbnb. Given the similar approach to the host listing count, we have also excluded this column from our analysis, as it doesn't significantly contribute to our predictive modeling.
6. **Host\_verifications**: This column lists the methods by which a host verifies their customers, but it provides limited information as it only includes values of 'email' and 'phone.' Since this information does not have a meaningful impact on the price, we have excluded this column from our analysis.



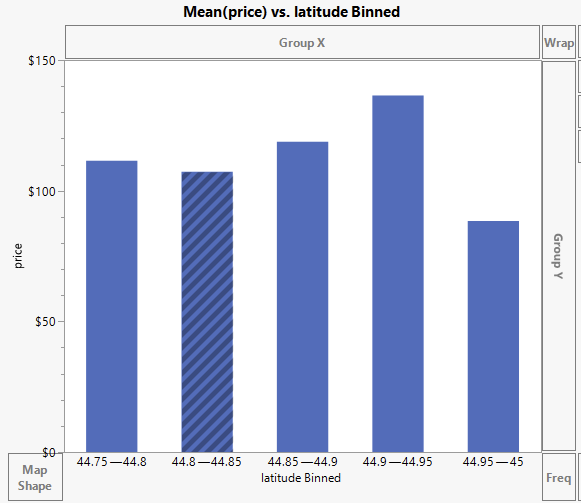
1. **host\_has\_ profile pic:** This column contains information about whether the host's profile has a profile picture or not. We have decided to keep this column because the correlation between price and profile picture as seen below.



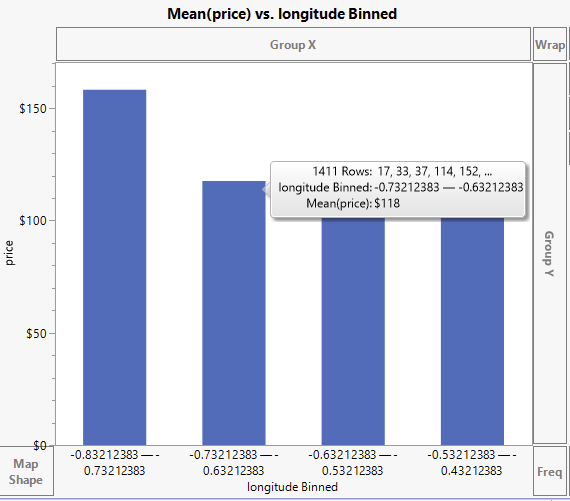
1. **host\_identfty \_verified:** This column indicates whether the host's profile has been verified or not. However, since the mean values for both categories in this variable in relation to price are nearly identical, suggesting that it does not have a significant impact on the price, we have decided to exclude this column from our analysis.



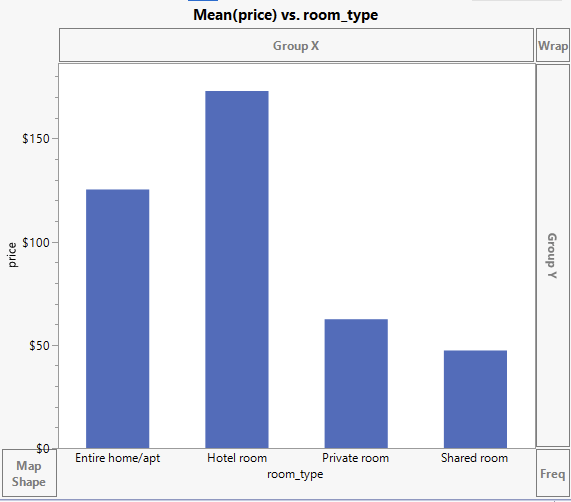
1. **Neighborhood:** This column contains data about the neighborhood of the listing, but it has a high percentage of missing data (41%). Since imputing missing values is not likely to be effective, and excluding rows with missing data would result in the loss of a significant number of rows, we have chosen to exclude this variable from our analysis.
2. **Neighbor\_cleansed:** This column contains geocoded location data for neighborhoods, and it is of nominal type with 62 categories. However, utilizing this variable for analysis would generate an additional 61 variables, leading to high computational complexity. To manage this complexity, we have decided to exclude this variable from our analysis.
3. **Neighbourhood\_group\_cleansed:** The latitude and longitude columns will result in the same as neighbourhood\_group\_ cleansed and it is of nominal type with 26 categories, utilizing this variable for analysis would generate an additional 25 variables. So, we have decided to exclude this variable from our analysis.
4. **Latitude:** This column contains the latitude of the geolocation for each listing. Although the original data type is continuous, we have converted it into a nominal variable through binning. We conducted binning because a range of latitude values can represent different locations, making it a potentially valuable parameter for price prediction. We have named the binned column 'Latitude Binned' and have decided to include it in our analysis.



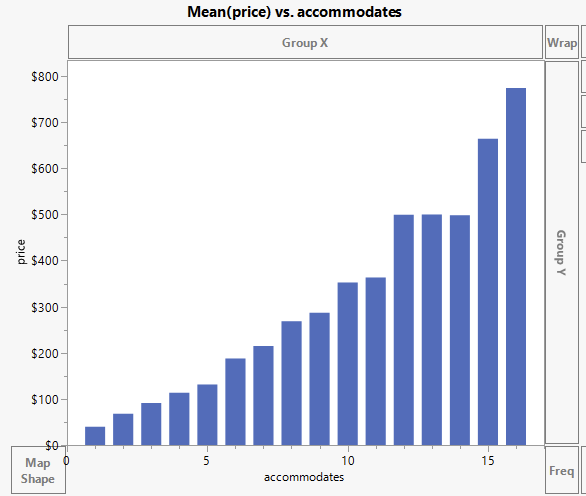
1. **Longitude:** This column pertains to the longitude of each listing's geolocation. Although originally a continuous data type, we've converted it into a nominal variable through binning. We decided to conduct this because it can show that distinct ranges of longitude can correspond to various locations, potentially offering valuable insights for predicting listing prices. As a result, we've labeled the binned column as 'Longitude Binned' and have chosen to incorporate it into our analysis.



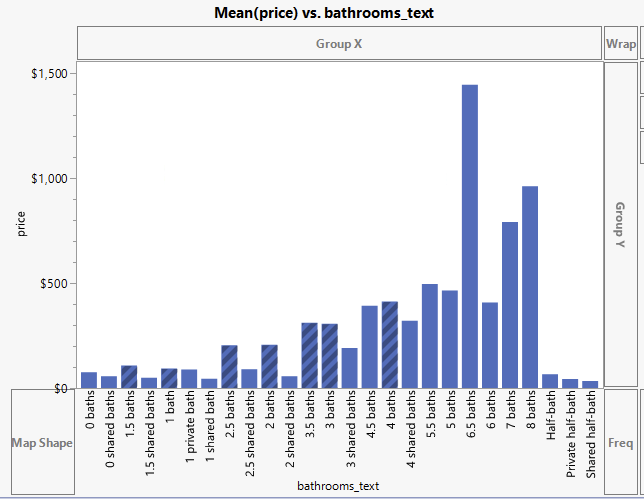
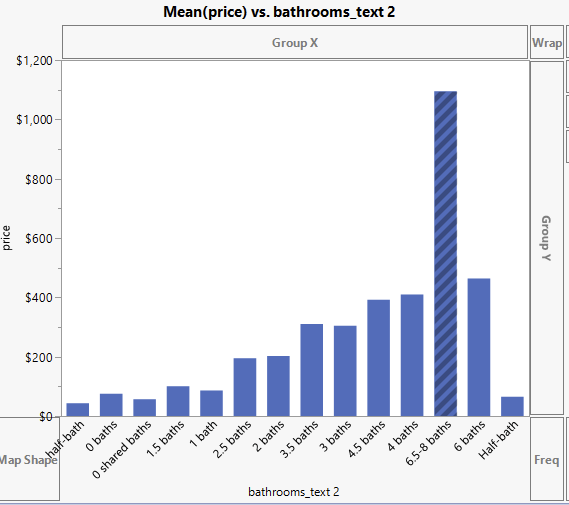
1. **Property\_type:** This column contains information about the property type of the listing, and it is expected to influence the price. However, it shares similar information with the 'Room\_type' variable, but it has a larger number of categories (64). Therefore, we have decided to exclude the 'Property\_type' variable from our analysis. Since we are still including 'Room\_type' in our analysis, this exclusion should not result in a significant loss of information.
2. **Room\_type:** This column provides information about the room type of the listing. An examination of the graph reveals significant variation between shared rooms and hotel homes, indicating that price is notably affected by the type of room. Therefore, we have included this variable in our analysis.



1. **Accommodations:** This column contains data about the maximum number of people a listing can accommodate. The graph demonstrates an observable increasing trend in price as the accommodation capacity grows. Given this trend, we have included this variable in our analysis, as it appears to have a significant impact on pricing.

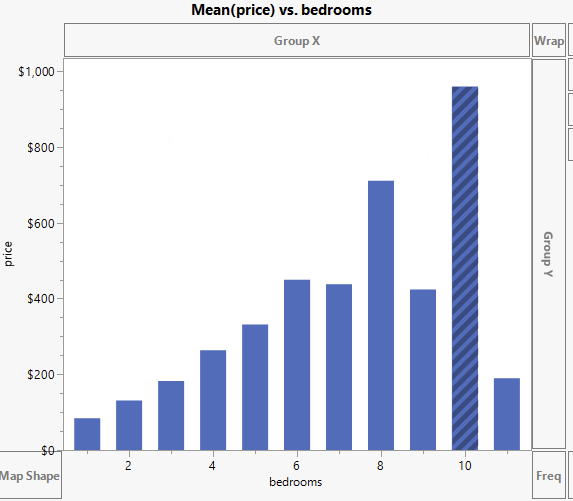
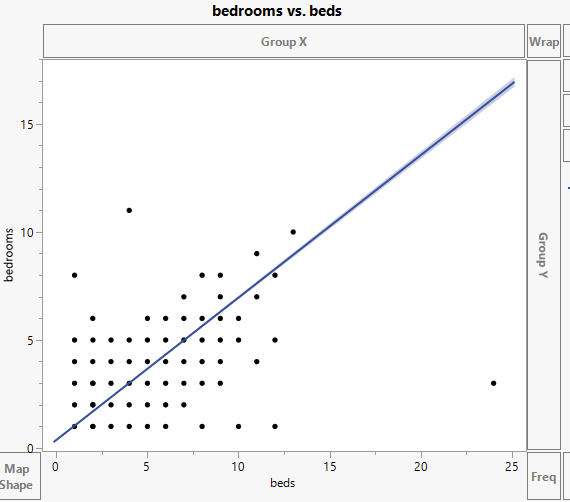


1. **Bathrooms:** This column is devoid of any records, and as a result, we have excluded it from our analysis.
2. **Bathrooms\_text:** This column provides information about the number of bathrooms in each listing. It is included in our analysis because it is expected to have an impact on the target variable. In its original format in JMP, it is categorized as a nominal variable with 27 categories, which can increase complexity. To address this, we have recoded the column into "Bathrooms\_text 2," which now has 14 categories and is considered an ordinal variable. The recording allows us to order the categories from highest to lowest, simplifying its use in our analysis.

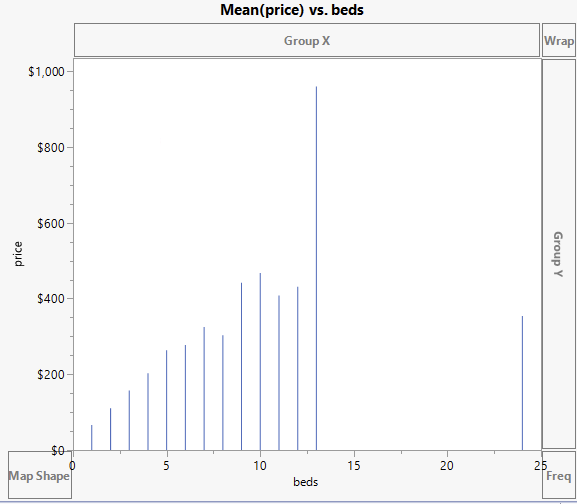


→ → →

1. **Bedrooms**: The decision to remove the "Bedrooms" column, which contains many missing values(28%), was made in favor of using the "Beds" column as a more reliable indicator of sleeping arrangements. Beds and bedrooms have a positive correlation.



1. **Beds:** We have decided to keep the beds column in our analysis. As seen below the more beds there are the higher the price showing a correlation between the two.

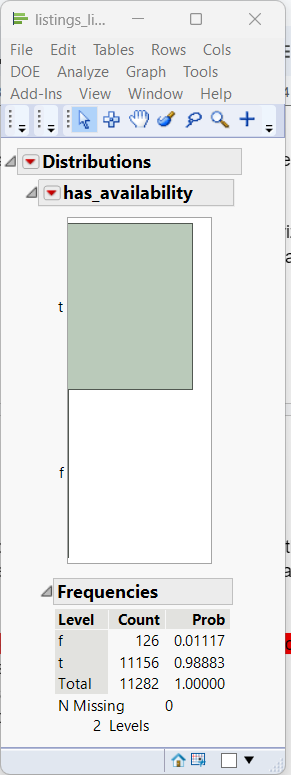


1. **Amenities:** This column lists the amenities offered by the listing, and it has the potential to impact the price. However, due to the high number of categories, which would increase complexity, we have chosen to exclude this variable from our analysis.
2. **price:** This is our target variable for the analysis.
3. **Minimum nights**: This column specifies the minimum number of nights required for a stay in the listing. Our analysis, along with the correlation factor, confirms that this variable does not have a substantial impact on the price. Therefore, we have chosen to exclude this column from our analysis.

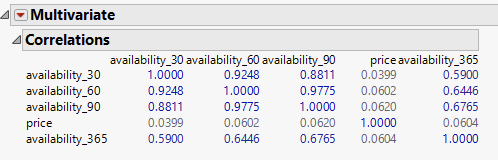
A screenshot of a computer

Description automatically generated

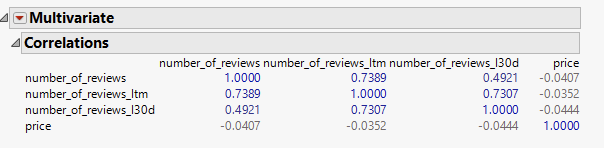
1. **Maximum nights**: This column records the maximum number of nights a listing allows for a stay. Our analysis, backed by the correlation factor, confirms that this variable does not exert a substantial influence on the price. Consequently, we have excluded this column from our analysis.
2. **Miniumum\_miumum\_ Nights:** The "Minimum Minimum Nights" column, which records the minimum allowable duration of stay for a listing, has similarly been excluded from our analysis. Our correlation analysis findings, like in the case of "Minimum Nights," confirm that this variable has minimal influence on pricing and therefore does not significantly contribute to our predictive model.
3. **Maximum minimum nights:** The "Maximum Minimum Nights" column, documenting the maximum allowable duration for the minimum stay for a listing, has been similarly excluded from our analysis. Our correlation analysis findings indicate that this variable has minimal to no influence on listing prices, substantiating its removal from the dataset.
4. **Minimium\_maximum night:** The "Minimum Maximum Night" column, documenting the minimum allowable duration for the maximum stay allowed for a listing, has also been excluded from our analysis. Our correlation analysis findings align with those of "MI Maximum Nights," indicating that this variable has minimal impact on listing prices, justifying its removal from the dataset.
5. **Maximum maximum nights:** The "Maximum Maximum Nights" column, specifying the maximum allowable duration for the maximum stay at a listing, has also been removed from our analysis. Our correlation analysis findings corroborate that this variable has minimal influence on listing prices, thus supporting its exclusion from the dataset.
6. **minimum\_nights\_avg\_ntm:** For "Minimum Nights Average NTM," we have chosen to remove this column from our analysis. Our examination of this variable revealed that it has minimal impact on pricing and doesn't significantly contribute to our predictive model.
7. **Maximum\_nights\_avg\_ntm:** For "Minimum Nights avg ntm," we have chosen to remove this column from our analysis. Our examination of this variable revealed that it has minimal impact on pricing and doesn't significantly contribute to our predictive model.
8. **Calendar\_updated:** This column lacks any records, and therefore, we have excluded it from our analysis.
9. **Has\_availability:** This column categorizes the availability of the listing as either 'yes' or 'no.' However, the column primarily consists of the same value, except for 126 records that have 'f' as their value, making the data nearly constant. Additionally, the mean price for both categories is almost identical. As a result, we have excluded this variable from our analysis.



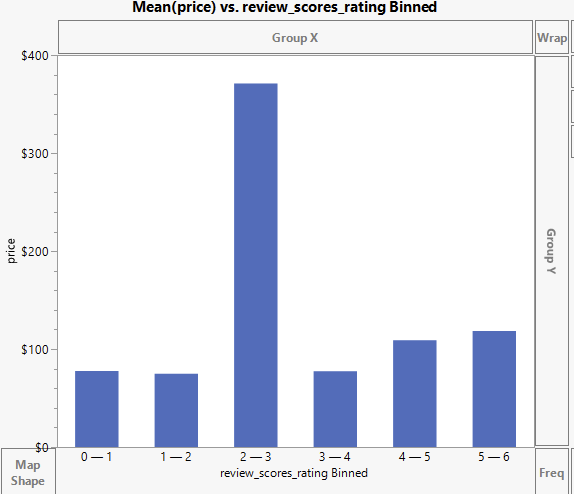
1. **Availability\_30:** This column specifies the number of days the listing is available in the next 30 days. However, when comparing it with the price, we observe that price does not vary significantly with this variable, and the correlation factor is low. As a result, we have opted to exclude this column from our analysis.

****

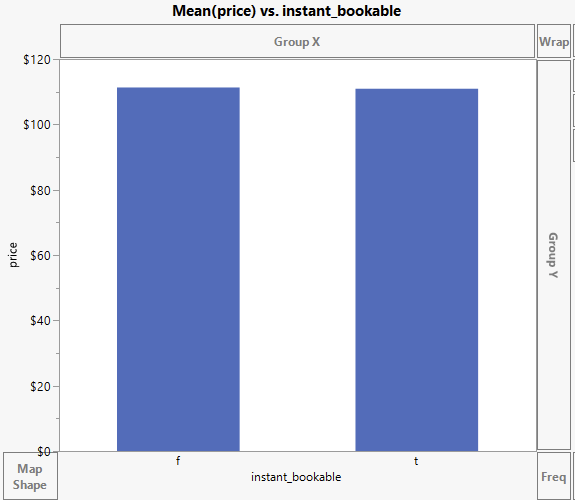
1. **Availability\_60:** The "Availability 60" column indicates the number of days the listing is available over the next 60 days. We found that the price exhibits minimal variation with respect to this variable, and the correlation factor is low. Consequently, we have made the decision to exclude the "Availability 60" column from our analysis.
2. **Availability\_90:** The "Availability 90" column, which signifies the number of days the listing is available over the next 90 days, also displays limited price variation and a low correlation factor when analyzed. Therefore, we have chosen to remove the "Availability 90" column from our analysis, as it does not significantly influence listing prices.
3. **Availability\_365:** Similarly, the "Availability 365" column, representing the number of days the listing is available over the course of a year, demonstrates minimal price variation and a low correlation factor in our analysis. Hence, we have decided to exclude the "Availability 365" column from our analysis, as it does not significantly impact listing prices.
4. **Calendar last scraped:** This column denotes the date when this listing was scraped. Since this information does not have a discernible impact on the price, we have chosen to exclude this variable from our analysis.
5. **Number of reviews:** This column records the total number of reviews for the listing. Upon comparison with the price, it becomes evident that they have a very low correlation. Consequently, we have excluded this column from our analysis.



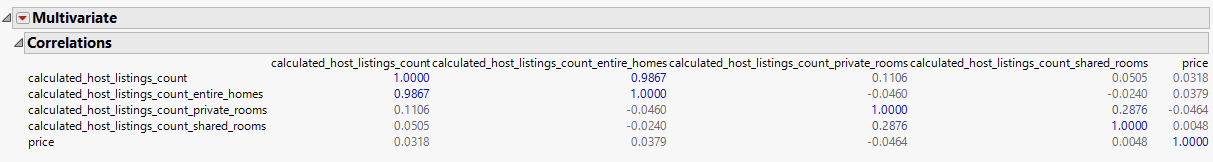
1. **Number of reviews ltm:** The "Number of Reviews LTM" column, documenting the total number of reviews for the listing in the last twelve months, similarly displays a very low correlation with price in our analysis. As a result, we have chosen to remove this column from our analysis, as it has little influence on listing prices.
2. **Number\_of\_reviews l130d**: The "Number of Reviews L130D" column denotes the total count of reviews a listing has accumulated in the last 130 days, similarly displays a very low correlation with price in our analysis. As a result, we have chosen to remove this column from our analysis, as it has little influence on listing prices.
3. **First review:** This column contains the date when the first review was posted for the listing. However, as it does not appear to have a significant impact on the price, we have decided to exclude this variable from our analysis.
4. **Last review:** This column indicates the date when the latest review was posted for the listing. However, as it does not have a notable impact on the price, we have chosen to exclude this variable from our analysis.
5. **Review\_scores\_rating:** This column represents the average customer rating for the listing, derived from various aspects such as accuracy, cleanliness, check-in, communication, location, and value. To make it more manageable, we have binned the data into six categories, allowing for a more effective comparison of mean prices. Upon analyzing the graph depicting the relationship between price and the binned review score rating, it's evident that there is a substantial variation in prices across different ratings. Therefore, we have included this variable in our analysis, presenting it as binned data with a changed data type from continuous to ordinal for enhanced analytical clarity.



1. **Review\_scores\_accuracy:** This column contains the average customer rating for the accuracy of the parameters committed. However, we have excluded this variable from our analysis as it is already encompassed within the broader 'Review\_scores\_rating' variable, avoiding redundancy in our analysis.
2. **Review\_scores\_cleanliness:** The "Review Scores for Cleanliness" column, which represents the average customer rating for the cleanliness of listings, has been omitted from our analysis. This decision was made to avoid redundancy, as this information is already included within the broader variable "Review\_scores\_rating." Hence, we have chosen to focus on the more comprehensive variable such as review\_score\_rating to streamline our analysis.
3. **Review\_scores\_checkin:** The "Review Scores Check-in" column typically reflects the average customer rating for the check-in process of listings. In the context of our analysis, we have chosen not to include this specific column. This decision was made to prevent redundancy because the information regarding the check-in experience is already encompassed within the more comprehensive variable "Review Scores Rating."
4. **Review\_scores\_communication:** The "Review Scores for Communication" column, reflecting the average customer rating for the communication aspects of listings, has also been excluded from our analysis. This was done to prevent redundancy, as this information is encompassed within the broader "Review\_scores\_rating" variable. By focusing on the overarching variable, we maintain a more streamlined and comprehensive approach in our analysis.
5. **Review\_scores\_location:** The "Review Scores for Location" column, representing customer ratings specific to the location of listings, has been excluded from our analysis. This decision was made to prevent redundancy, as this information is already encapsulated within the broader "Review\_scores\_rating" variable. Our aim is to maintain a more streamlined and comprehensive approach in our analysis, focusing on the overarching rating variable.
6. **Review\_scores\_value:** The "Review Scores for Value" column, which records customer ratings pertaining to the value of listings, has also been omitted from our analysis. To streamline our approach and avoid redundancy, we have chosen to concentrate on the broader "Review\_scores\_rating" variable, which encompasses overall ratings, including value considerations.
7. **License:** This column contains the license, permit, or registration numbers for the listing. However, with more than 65% of the records missing in this column, we have chosen to exclude it from our analysis due to the significant amount of missing data.
8. **Instant\_bookable:** We have decided to remove the "Instant Bookable" column from our dataset. This decision is based on our assessment that "Instant Bookable" does not significantly impact the prediction of listing prices, and therefore, its exclusion will help streamline our analysis and modeling efforts.



1. **Calculated\_host\_listing\_count:** This column indicates the number of properties the host has listed in a city at the time of the last data scrape. However, it does not appear to have a significant impact on the price, and as such, we have excluded this variable from our analysis.

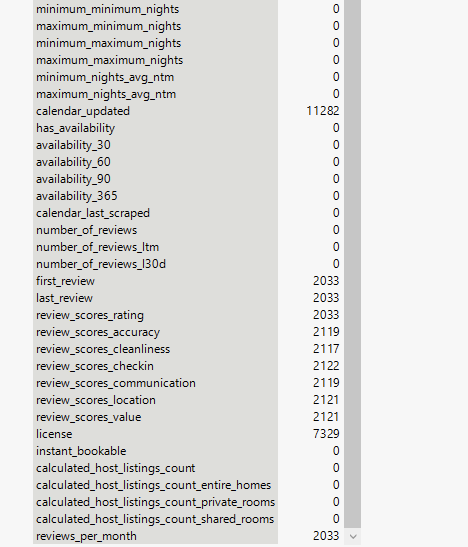
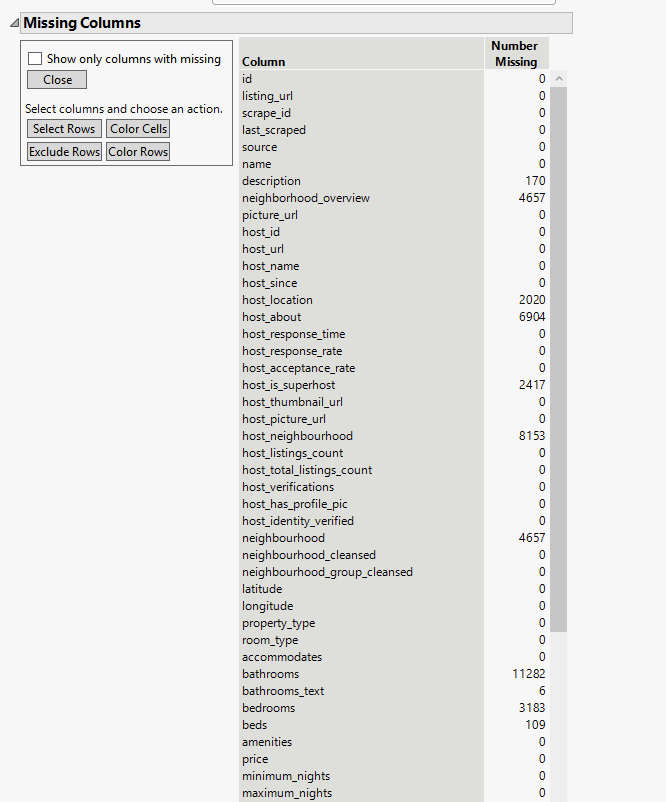


1. **Calculated\_host\_listing\_count\_entire\_homes:** The "Calculated Host Listing Count Entire Homes" column, which represents the number of entire homes listed by the host in a city at the time of the last data scrape, has been excluded from our analysis. Our rationale for this exclusion is that it does not appear to have a substantial impact on listing prices.
2. **Calculated\_host\_listing\_count\_private\_rooms:** Similarly, the "Calculated Host Listing Count Private Rooms" column, which denotes the number of private rooms listed by the host in a city at the time of the last data scrape, has been excluded from our analysis. Our decision to remove this variable is based on our observation that it does not exert a significant impact on listing prices. By excluding this feature, we aim to enhance the efficiency of our analysis and modeling, focusing on the most influential predictors of Airbnb listing prices.
3. **Calculated\_host\_listings\_count\_shared\_rooms:** The "Calculated Host Listing Count Shared Rooms" column, representing the number of shared rooms listed by the host in a city at the time of the last data scrape, has also been excluded from our analysis. This decision is rooted in our observation that this variable does not significantly influence listing prices.
4. **Reviews per month**: This column represents the number of reviews a listing has received over the years, calculated as a monthly average. Since the number of reviews is not expected to have a significant impact on the price, we have chosen to exclude this variable from our analysis.



1. **Missing Values**

The initial dataset has 75 variables including our target variable “price” and the dataset exhibits a substantial number of missing values, which we have explored through a missing data pattern analysis. In particular, the columns "calender\_updated" and "bathroom" are entirely devoid of values, while the "bathroom\_text" column has only 6 missing values. In total, there are 23 columns in the dataset with missing values.



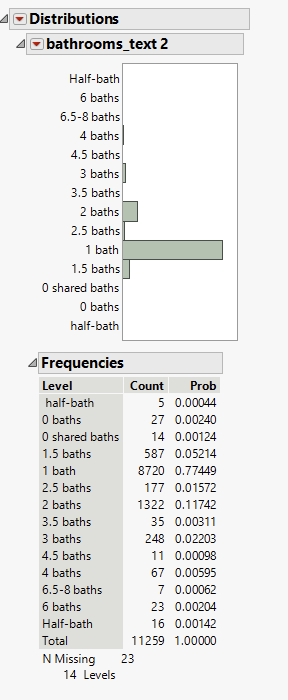
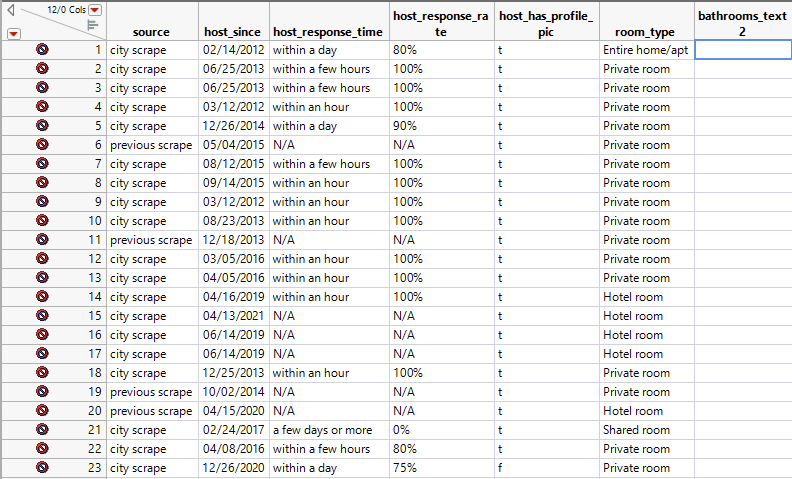
After excluding the columns based on our analysis, we have 11 variable columns that we can use for the model to predict the “price”. We have missing values only in the recoded bathroom\_text column. As seen below: ↓

A screenshot of a computer

Description automatically generated

**Handling the missing values:**

Since there are only 23 missing out of 11,282 values, we have decided to omit these rows from our analysis. The impact of this decision will allow for more precise and reliable results in our model.

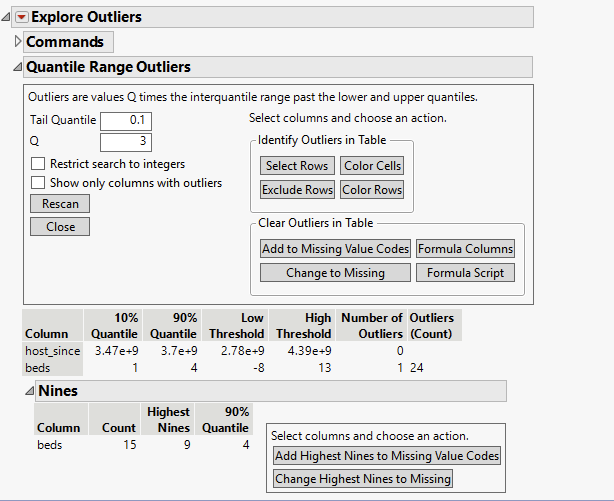


**5. Outlier Analysis:**

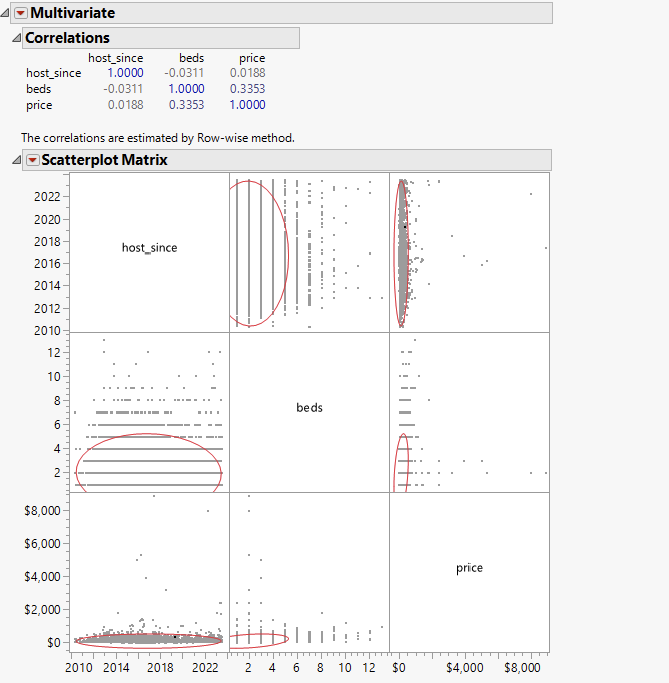
We have diligently addressed the task of removing outliers from both continuous and categorical variables within the dataset. Our approach has involved implementing various methods and techniques, as detailed below.

**Continuous variables – Outlier analysis:**

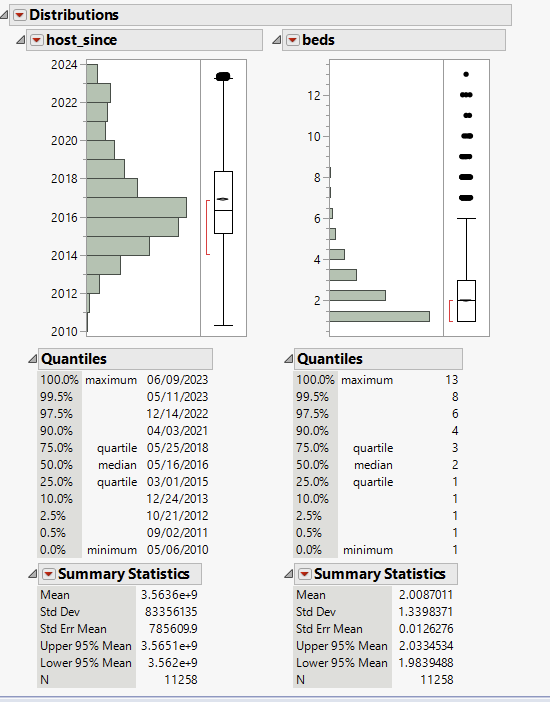
We've employed the quantile range outliers function to identify outliers in continuous variables. Specifically, we applied this function to the "host\_since" and "beds" variables, setting the tail quantile to 0.1 and the Q parameter to 3. Our analysis revealed that there is one outlier in the "beds" variable. To ensure the robustness of our model, we have chosen to exclude this single outlier, as it has the potential to influence the modeling process.



We attempted to confirm the presence of any additional outliers by creating a multivariate scatter plot distribution. Since no clear pattern or outliers could be detected from this plot, we did not identify any further outliers using this approach. The correlation factor between the target variable, "price," and the other two predictor variables is very low. This suggests that there may not be a strong relationship between these variables.



While exploring outlier analysis for continuous variables, we examined the distribution of both variables and observed their box plots. It became apparent from the box plots that there are still numerous potential outliers present. One potential approach to address this issue would be to use a continuous fit function to transform the data, which may help include most of these outliers within the tails of the box plots.



We have decided to transform the beds column using a continuous fit function option and save the new column. SHASH distribution is found to be more suitable for this column as shown below.

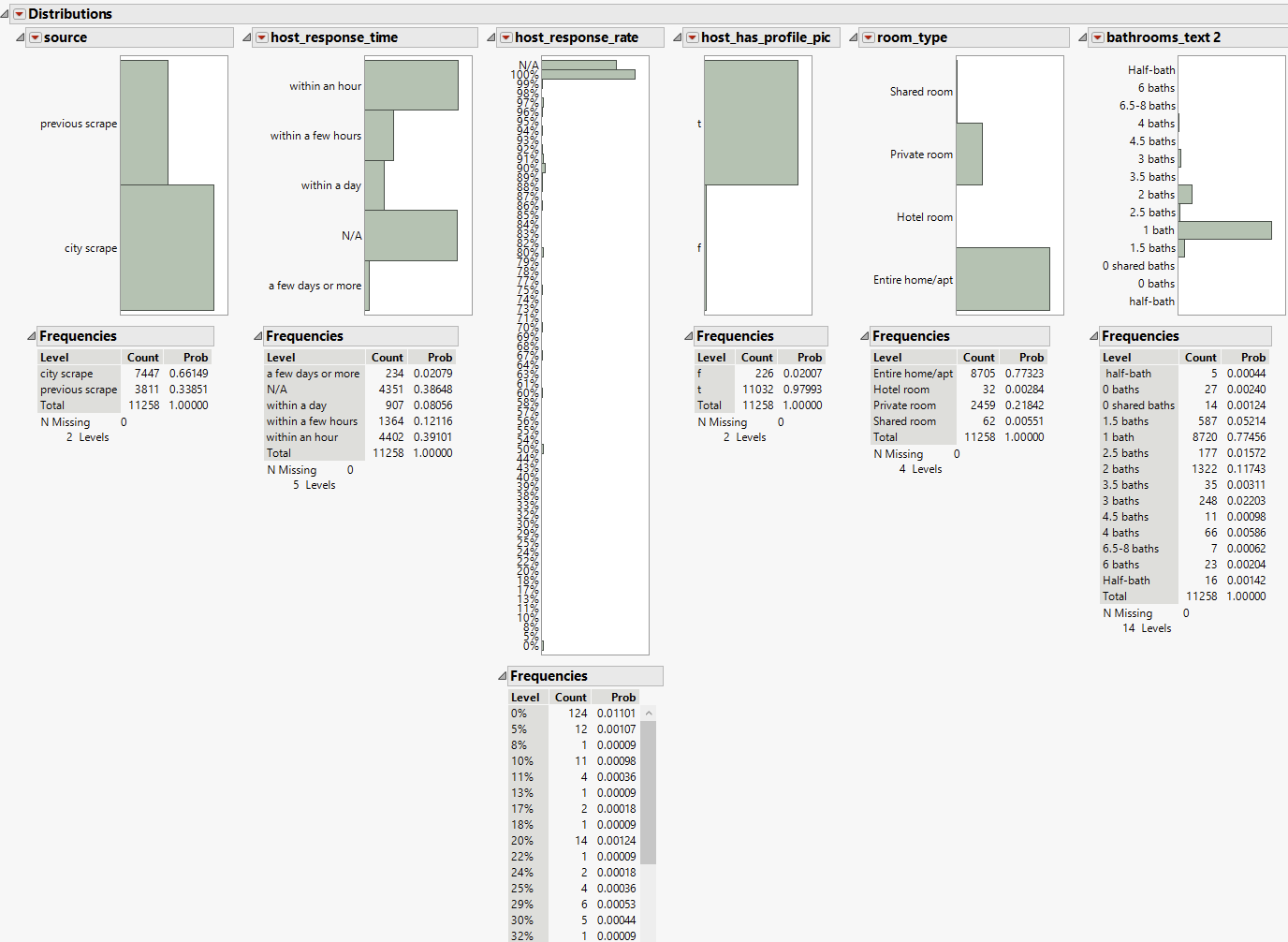
A screenshot of a graph

Description automatically generated A screenshot of a computer

Description automatically generated

**Categorical Variables – Outlier Analysis**

In the categorical variables, we noticed that some categories in certain variables have very low counts. This distribution of low counts can be observed in the charts provided below for those specific variables. But based on our analysis, we have decided not to exclude any rows from these categorical variables as they are correlated with the price.



1. **Conclusion:**

The processed data now consists of 11,259 rows and 12 columns. During the data preparation phase, we made decisions about which variables to exclude based on their importance, and we included variables that we deemed significant using various visualization and exploration techniques. We also addressed missing data through exclusion and imputation and handled outliers as necessary. It's important to note that we have not imputed or transformed the target variable, "price," in the dataset.