KAUNAS UNIVERSITY OF
TECHNOLOGY, FACULTY OF
INFORMATICS

Artificial intelligence

Laboratory work No. 1

Name =wren francis alumkal thomas
IFU-0

KAUNAS, 2023

1. **Select (create) a dataset to perform this and other laboratory works. Your choice must be approved by the tutor. Selected Dataset: College**

Selected Dataset: airbnb

Link: http://data.insideairbnb.com/united-kingdom/england/london/2022-12-10/data/listings.csv.gz

Description: dataset for listings in london in 2022

Format: dataset trimed to 20 colums and 14000 rows

Columns:

host_id
host_acceptance_rate
accommodates
bedrooms
beds
price
minimum_nights
maximum_nights
availability_365
number_of_reviews
review_scores_rating
review_scores_cleanliness
reviews_per_month
host_response_time
host_is_superhost
room_type
bathrooms_text
instant_bookable
neighbourhood_cleansed
property_type

Task 1:

For each numeric type attribute calculate

1. total number of values,
2. percentage of missing values,
3. cardinality,
4. minimum (min) and maximum (max) values,
5. 1st and 3rd quartiles,
6. average,
7. median,
8. Standard deviation.
  -

For each *category* type attribute calculate:

9. total number of values,
10. percentage of missing values,
11. cardinality,
12. mode,
13. The frequency of the mode
14. Percentage value of the mode
15. Second mode value (mode 2),
16. Frequency value for Mode 2,
17. Percentage of Mode 2.

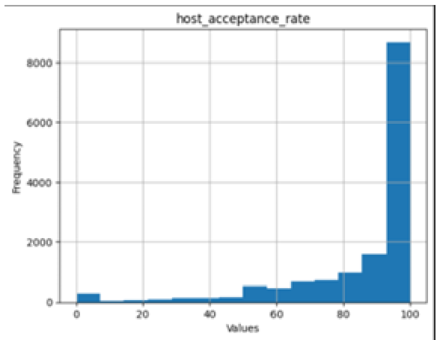
Continuous attribute name	total number of values	percentage of missing values	cardinality	min	max	1st quartile	3rd quartile	average	median	standard deviation
host_id	14829	0	8510	2594	489618073	10667138.5	128659370			
host_acceptance_rate	14622	1.4	94	0	100	80	100	86.083	96	21.2711
accommodates	14829	0	17	0	16	2	4	3.2316	2	2.04689
bedrooms	14134	4.69	10	0.9155	22	1	2	1.55184	1	0.90622
beds	14583	1.66	15	1	38	1	2	1.84452	1	1.28965
price	14829	0	881	12	1959	75	215	178.687	130	177.835
minimum_nights	14829	0	67	1	365	2	5	5.90479	3	13.4814
maximum_nights	14829	0	175	1	524855552	60	1125	36093.5	365	4319099
availability_365	14829	0	366	0	365	57	308	171.138	150	126.271
number_of_reviews	14829	0	375	0	956	2	34	31.4132	10	57.714
review_scores_rating	12597	15.05	160	0	5	4.67	5	4.74234	4.86	0.40562
review_scores_cleanliness	12582	15.15	172	0.401635	5	4.61	5	4.71749	4.84	0.40168
reviews_per_month	12596	15.06	671	0.01	26.05	0.39	1.7	1.27767	0.86	1.33457

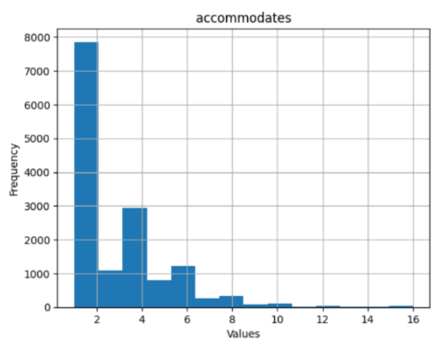
Categorical attribute name	total number of values	percentage of missing values	cardinality	Mode	Frequency value of the mode	Percentage value of the mode	2nd mode value	2nd mode frequency	Percentage of 2nd	
host_response_time	14767	0	3	within an hour	10415	70.52	within a few	2938	19.89	
host_is_superhost	14767	0	2	f	10471	70.9	t	4296	29.09	
room_type	14767	0	4	Entire home/apt	9363	63.4	Private room	5353	36.24	
bathrooms_text	14758	0.06	6	bath	10058	68.11	sharebath	3388	22.96	
instant_bookable	14767	0	2	f	11085	75.06	t	3682	24.93	
neighbourhood_cleansed	14829	0	33	Westminster	1701	11.51	Tower Haml	1148	7.77	
property_type	14829	0	69	Entire rental unit	4373	29.61	Entire cond	2720	18.419	

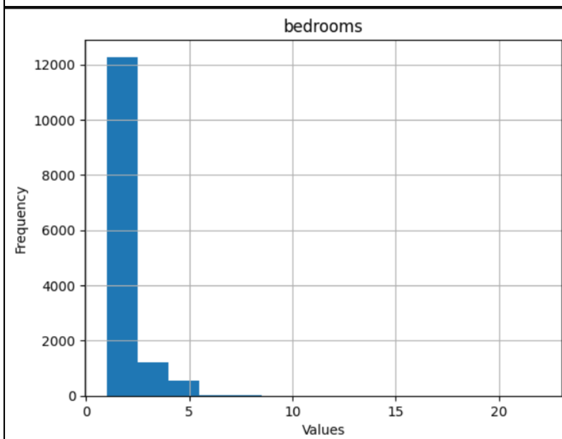
pav. 1 properties of attributes

Task 2:

Draw histograms of attributes. Provide descriptions of the distribution (eg, normal,exponential, etc.) and what conclusions can be drawn from it.

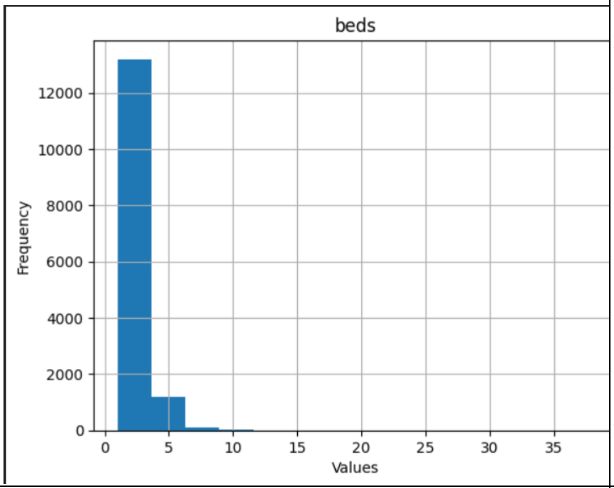
<table>
<tr><th>Histogram</th><th>Description</th></tr>
<tr><td>


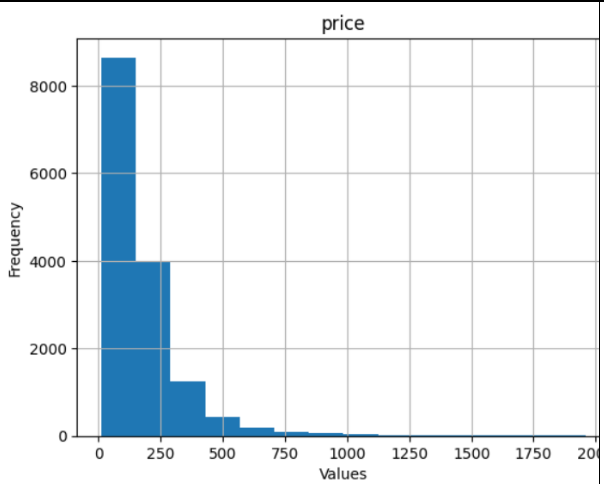
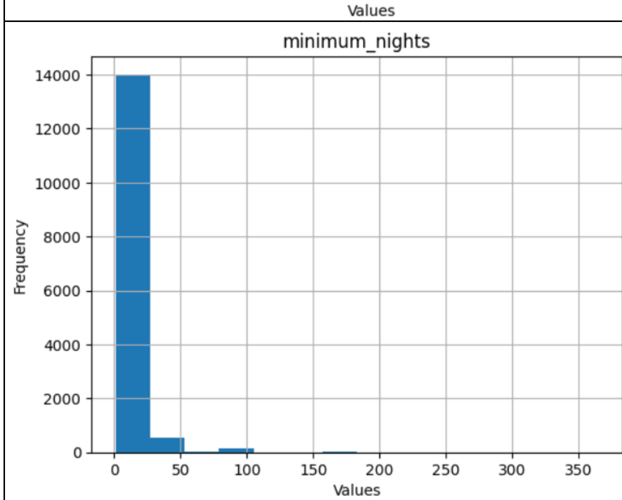
</td><td>Skewed left it seems most host acceptance rate is between 90 and 100%</td></tr>
<tr><td>


</td><td>Skewed right it seems most accommodates on interval 1 to 2 people</td></tr>
<tr><td>


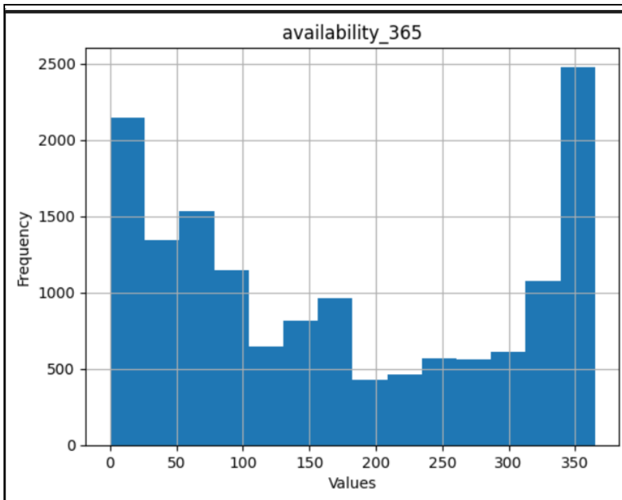
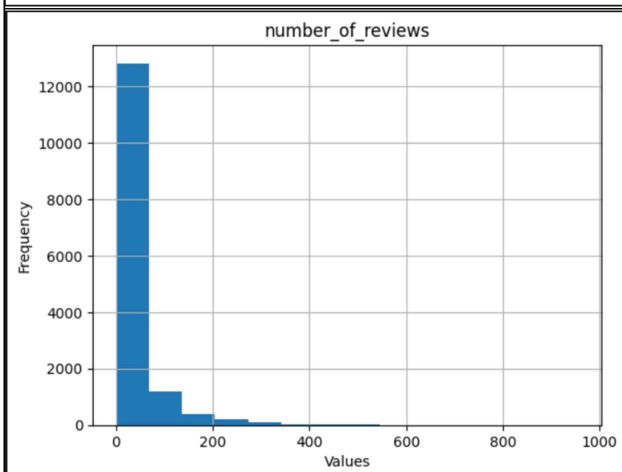
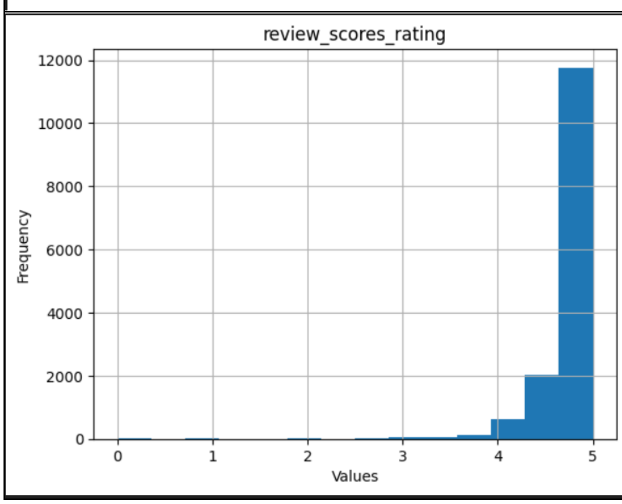
</td><td>Skewed right

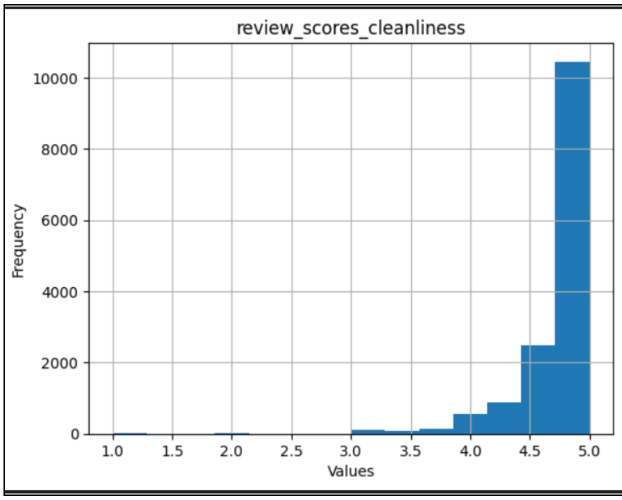
Most bedrooms between 1 and 2</td></tr>
</table>

<table>
<tr>
<td>
</td>
<td>Skewed right

Most beds between 1 and 2</td>
</tr>
<tr>
<td>
</td>
<td>Skewed right
Most prices between 0 and 125</td>
</tr>
<tr>
<td>
</td>
<td>Skewed right

Most minimum night between</td>
</tr>
</table>

<table>
<tr><td>
</td><td>Normal distribution</td></tr>
<tr><td>
</td><td>Skewed right, can see number of reviews mostly between 0 and 65</td></tr>
<tr><td>
</td><td>Skewed left can see most rating between4.5 and 5</td></tr>
</table>



Skewed left review score scores most between 4.5 and 5

Task 5:

Identify data quality problems: missing values, cardinality problems, outliers. Provide a plan for resolving these issues, which will be implemented programmatically (e.g., missing values for a categorical attribute based on the attribute estimate of the mode, extreme values beingremoved or corrected).

Number of missing values in continues attributes

host_acceptance_rate	1.4% missing value
bedroom	4.69% missing value
beds	1.66% missing value
review_scores_rating	15.05% missing value
review_scores_cleanliness	15.15% missing value
reviews_per_month	15.06% missing value

Number of missing values in categorical attribute

bathrooms_text	0.06%

As the missing characteristics for bedrooms, beds, and accommodations are highly correlated, there is no missing value for the accommodations element. To forecast the missing value of the other columns, we apply linear regression.
using library sklearn.linear_model in the Python scikit-learn machine learning package that provides tools for linear regression and related models.

host_acceptance_rate   cannot be filled with regression as it doesn’t have any strong correlation to other attributes . fill missing values with the median for it not to affect the distribution of values
we can see that the charecteristics have not changed after imputation

reviews per month are empty because the value of the matching row in the number of reviews column is 0, which can result in reviews per month being 0.

Use 0 in place of any missing values in "reviews per month"
Review scores cleanliness and Review scores rating's missing values should be filled up with the mean score for those columns.

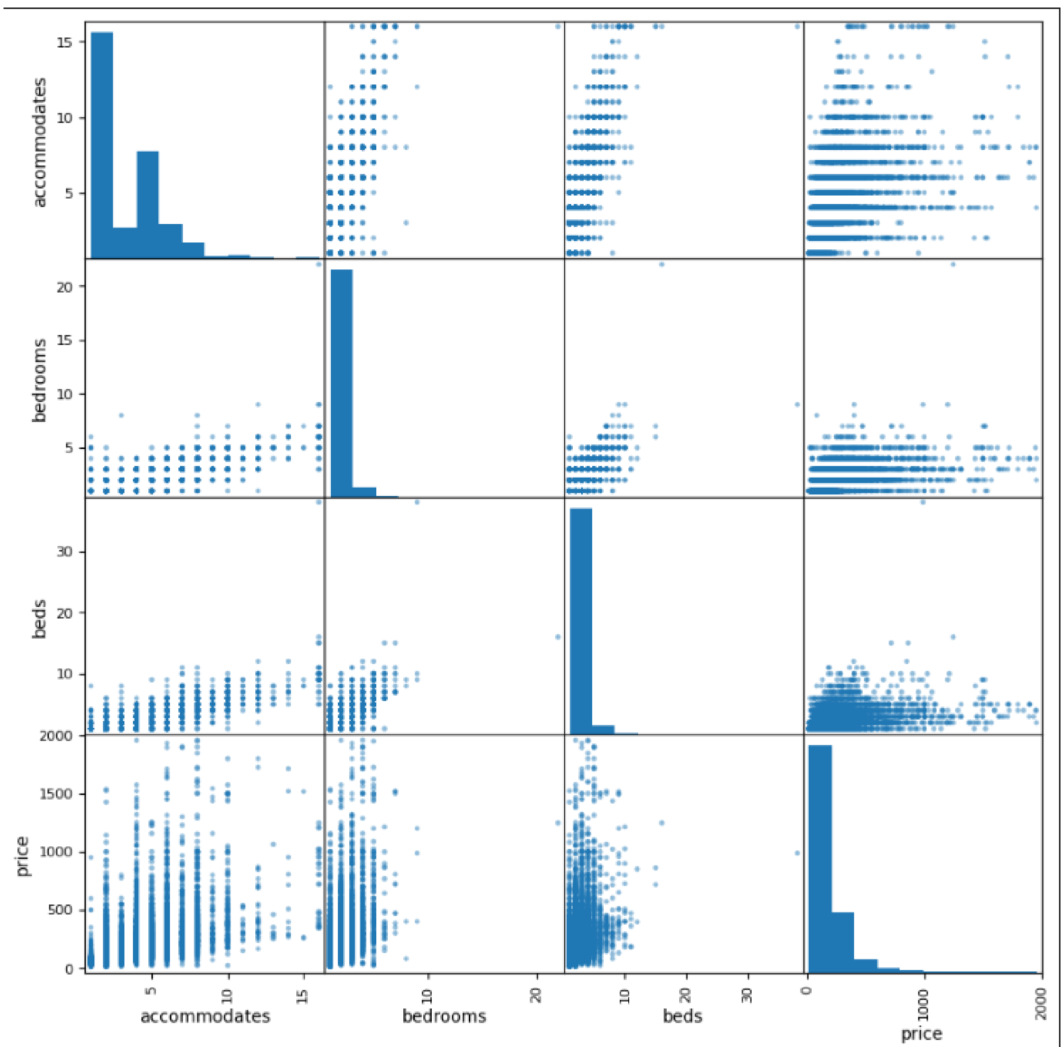
since bathroom text is a categorical variable, we can use mode to fill the missing values

Task 6:

Investigate relationships between attributes using visualization techniques: For continuous type attributes:

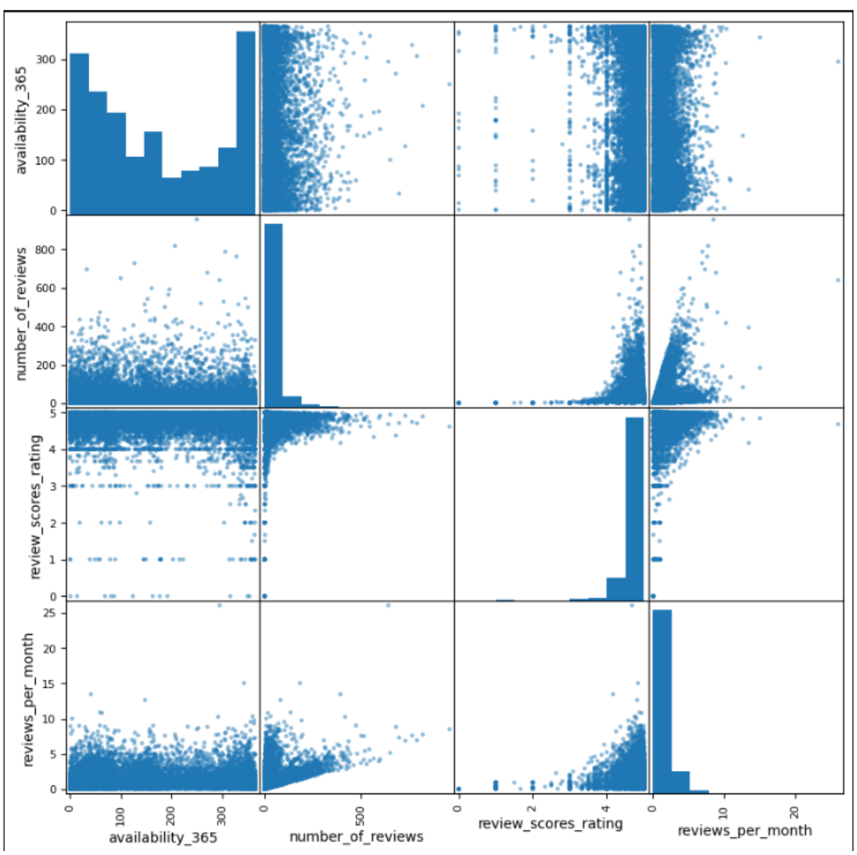
Example of few strongly correlated attributes

	accommodates	bedrooms	beds	price
accommodates	1	0.83266	0.846915	0.576423
bedrooms	0.83266	1	0.808493	0.558551
beds	0.846915	0.808493	1	0.475015
price	0.576423	0.558551	0.475015	1

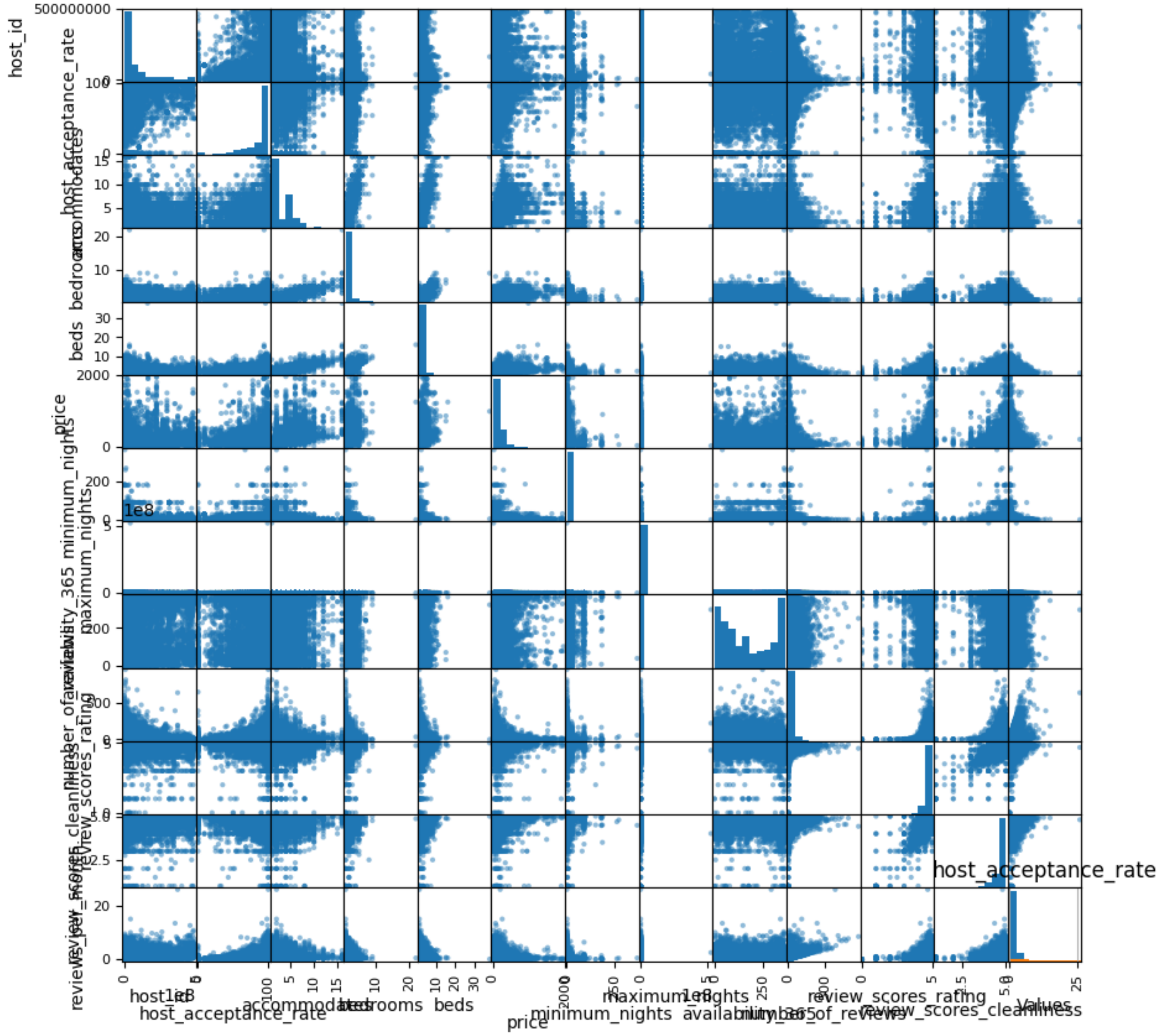


Indicating either a negative or positive correlation, correlation values vary from -1 to 1. The matrix's observations reveal significant positive connections between bedrooms and accommodates as well as between accommodates and beds. In contrast to host acceptance rate, which has a weak positive connection with price and a weak negative correlation with bedrooms, price has a weak positive association with accommodations, bedrooms, and beds. The matrix reveals information about the connections between the variables.

Example of few weakly correlated attributes:



SPLOM diagram (Scatter Plot Matrix).:

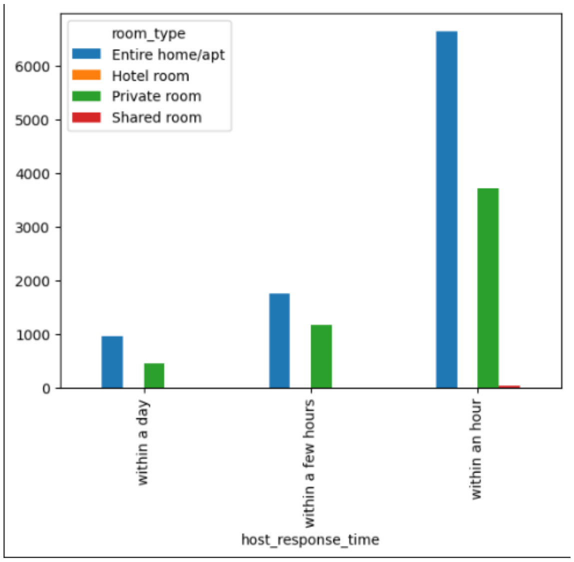


	host_acceptance_rate	accommodates	bedrooms	beds	price	minimum_nights	maximum_nights	availability_365	number_of_reviews	review_scores_rating	review_scores_cleanliness	reviews_per_month
host_acceptance_	1	0.04050228	-0.0201	0.013323	0.042622	-0.02968445	0.005036188	-0.000651637	0.118761465	-0.080467704	-0.050410551	0.264361693
accommodates	0.04050228	1	0.83266	0.846915	0.576423	0.011791417	-0.008965748	0.004436945	-0.09289603	-0.044708679	-0.068419784	-0.053502475
bedrooms	-0.020103567	0.832659905	1	0.808493	0.558551	0.033850482	-0.005130119	-0.015305433	-0.123723032	-0.007354915	-0.037256783	-0.10653978
beds	0.013323111	0.846914578	0.808493	1	0.475015	0.001812593	-0.005430218	0.002749902	-0.075124866	-0.024913457	-0.053647442	-0.06123009
price	0.042622043	0.576422524	0.558551	0.475015	1	0.038755078	-0.007059756	0.080680117	-0.147737171	0.010271383	0.022614855	-0.108468073
minimum_nights	-0.02968445	0.011791417	0.03385	0.001813	0.038755	1	-0.00298282	0.073486681	-0.082452437	-0.006436477	-0.003148522	-0.148695503
maximum_nights	0.005036188	-0.008965748	-0.00513	-0.00543	-0.00706	-0.00298282	1	0.008148645	0.027326181	-0.005997178	-0.015732253	0.008434284
availability_365	-0.000651637	0.004436945	-0.01531	0.00275	0.08068	0.073486681	0.008148645	1	-0.042409004	-0.109722744	-0.066362705	0.019516761
number_of_revie	0.118761465	-0.09289603	-0.12372	-0.07512	-0.14774	-0.082452437	0.027326181	-0.042409004	1	0.034281955	0.055841361	0.383428705
review_scores_ra	-0.080467704	-0.044708679	-0.00735	-0.02491	0.010271	-0.006436477	-0.005997178	-0.109722744	0.034281955	1	0.773604049	0.045314845
review_scores_cl	-0.050410551	-0.068419784	-0.03726	-0.05365	0.022615	-0.003148522	-0.015732253	-0.066362705	0.055841361	0.773604049	1	0.048741848
reviews_per_mor	0.264361693	-0.053502475	-0.10654	-0.06123	-0.10847	-0.148695503	0.008434284	0.019516761	0.383428705	0.045314845	0.048741848	1

For categorical attributes:

Using the bar plot type diagram, give some (2-3) examples of attribute dependency

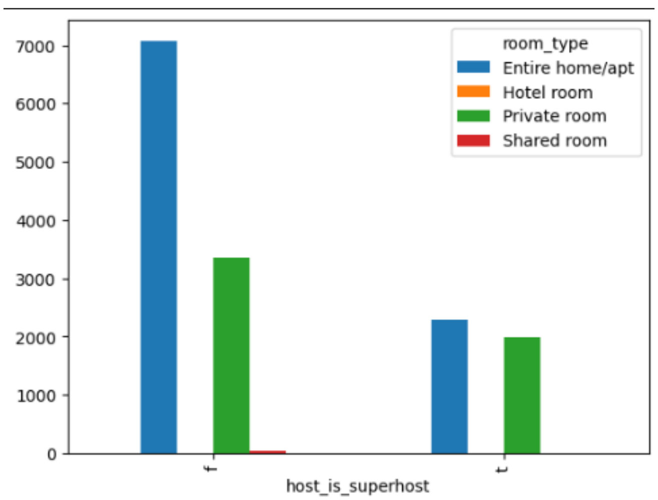
1. host_response_and time room_type



We can see that when host response time drops, the number of postings for each type of accommodation generally rises, suggesting that hosts with quicker responses typically have more listings accessible.

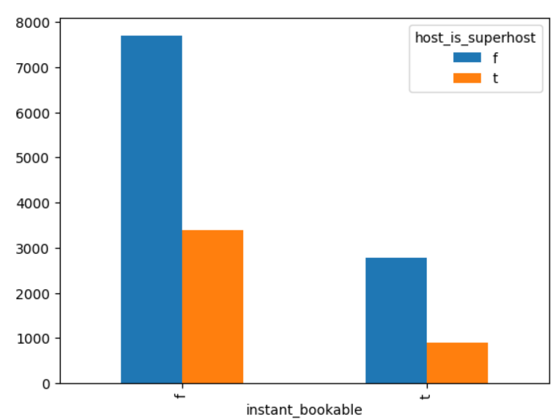
We can see that for each level of host response time, the most common room type is "Entire home/apt", followed by "Private room" and "Shared room

2. host_is_superhost room_type



We can conclude that non-superhosts list more complete homes/apartments and individual rooms than do superhosts. Yet, compared to non-superhosts, superhosts list more private rooms and only shared room is listed by non

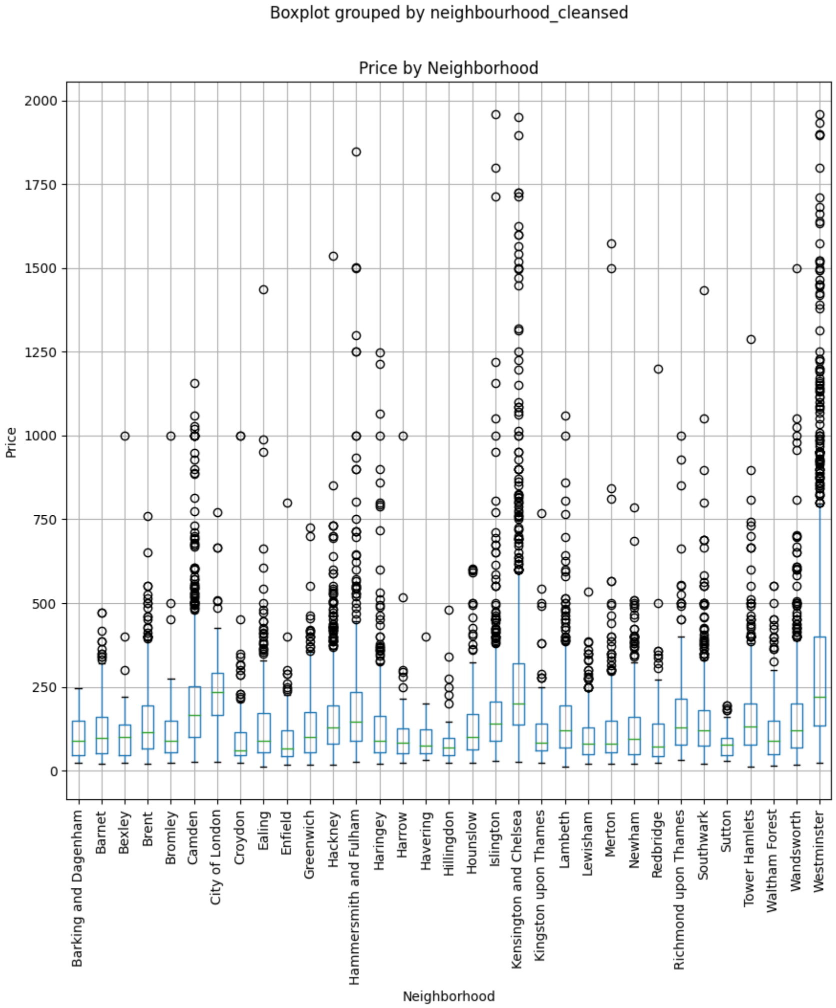
3. instant_bookable and host is superhost



This indicates that a higher proportion of non-superhosts tend to not offer instant booking compared to superhosts.

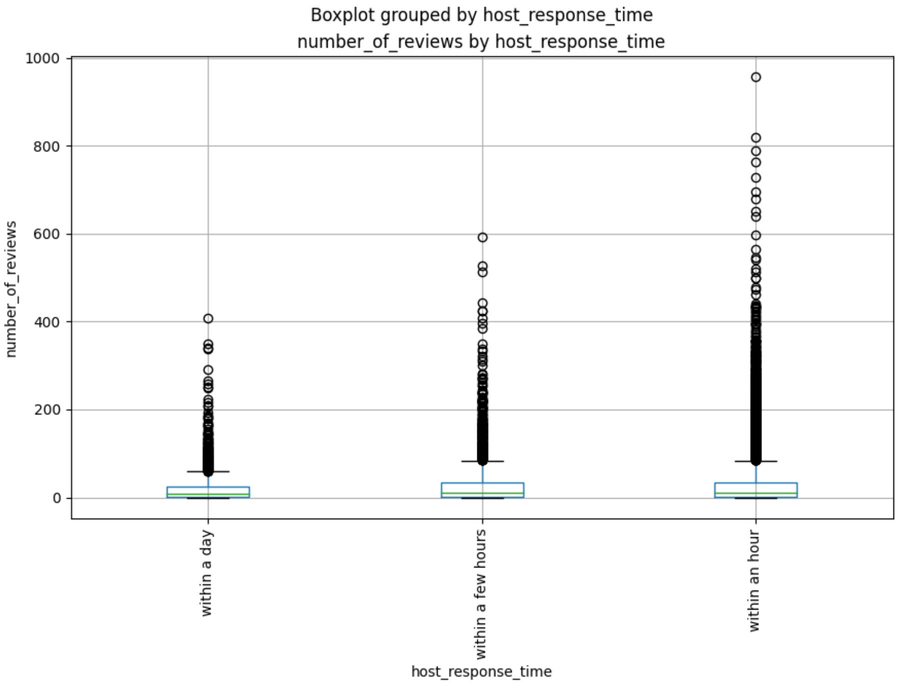
between categorical and continuous type variables.

1.price and neighbourhood category



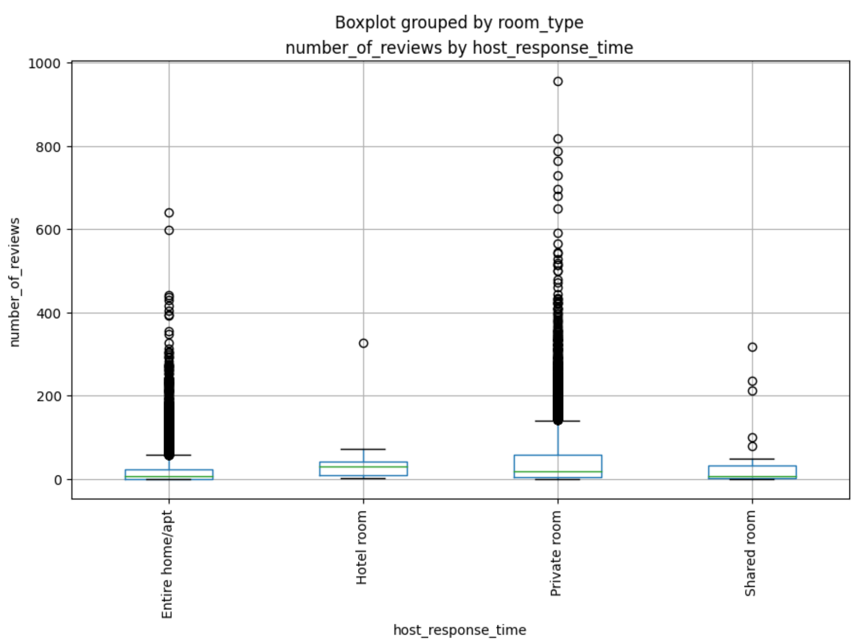
Here from the boxplot we can see that westminister and kingstyon have the highest number of listings for the hishest price

2.host response time and number of reviews



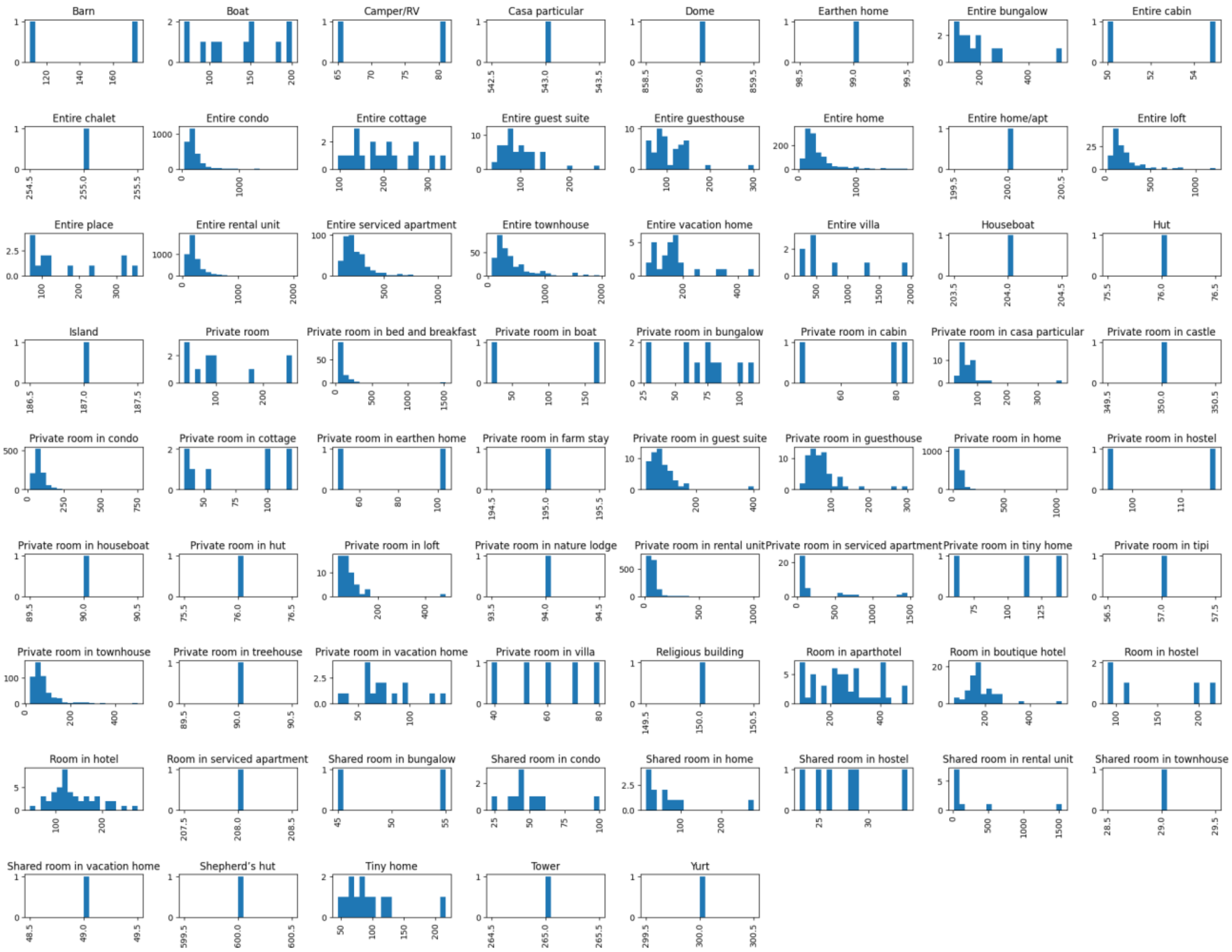
there are more outliersfor number of reviews for the listing which the host respond within an hour and it decreases to the others which mean thehost that reponds fast will get more reviews

3.Host_response _time and number of reviews



We can see number of reviews for private rooms have more outliers has more reviews than other types

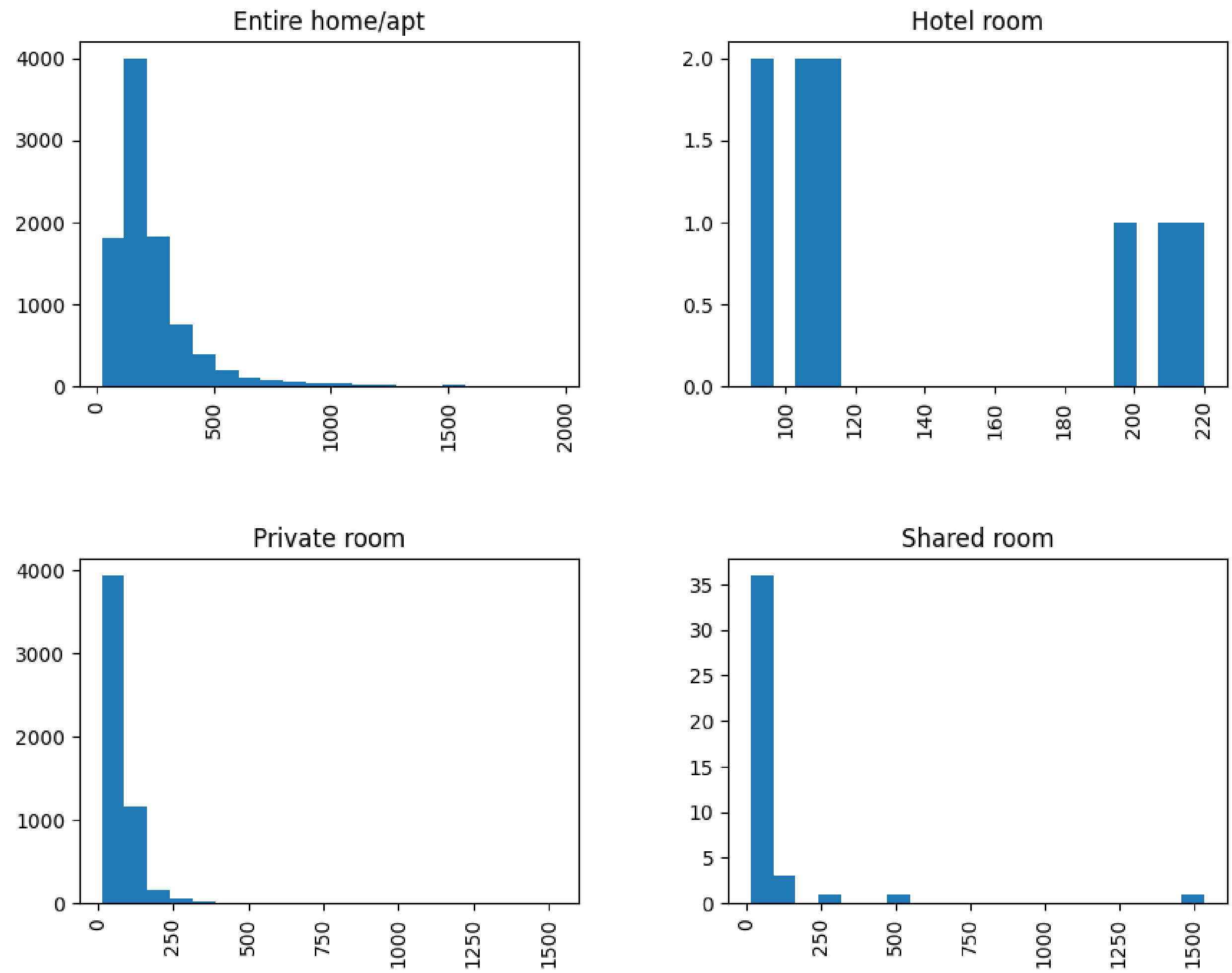
4.price and property type:



There are more offerings for complete homes, condos, and villas as the price range rises. Hostels and rental properties with shared rooms are prevalent in the lower price range of (0, 100). Less shared rooms are advertised as the price range rises, while individual rooms in houses and apartments are more prevalent.

Listings for entire islands, castles, and towers are found in the most costly price range of (900, 1000), indicating that these are high-end, luxurious properties.

5.room type and prices



According to graphmost of the postings are "Entire home/apt" types, and the majority of them are priced between ($10.05 and $109.35]. There aren't many postings between (1666.95 and 1764.3) and above, and those that are generally of the "Entire home/apt" type.

On the other side, "Hotel room" type listings are quite uncommon and typically have lower costs. Although there is a wide range of "private room" type listings, the majority of them are in the lower price level. Listings for "shared rooms" are like wise quite uncommon, and their costs are typically lower.

Task 7:

Calculate the covariance and correlation values between continuous attributes and graphically represent the correlation matrix

	Unnamed:	host_acce	accommo	bedrooms	beds	price	minimum_	maximum_	availability	number_o	review_sc	review_sc	reviews_per_month
Unnamed:	4.77E+08	71082.04	3136.119	1309.063	778.5422	469635.4	-3035.94	-1.2E+09	220896.1	-636811	-241.135	-337.065	2633.243
host_acce	71082.04	452.4612	1.764527	-0.39054	0.364355	161.035	-8.42128	465930.8	-1.75007	146.5294	-0.55443	-0.35111	6.558164
accommo	3136.119	1.764527	4.189756	1.565925	2.243583	209.823	0.325384	-79263.7	1.146788	-10.9742	-0.03147	-0.04764	-0.18562
bedrooms	1309.063	-0.39054	1.565925	0.821236	0.959531	91.63964	0.413432	-20568.7	-1.75445	-6.50927	-0.0024	-0.01123	-0.14363
beds	778.5422	0.364355	2.243583	0.959531	1.663196	109.4169	0.031578	-30500	0.447063	-5.60509	-0.01118	-0.02375	-0.10726
price	469635.4	161.035	209.823	91.63964	109.4169	31625.38	92.91436	-5422515	1811.711	-1516.31	0.582231	1.268718	-31.5976
minimum_	-3035.94	-8.42128	0.325384	0.413432	0.031578	92.91436	181.7493	-173683	125.0978	-64.1536	-0.0266	-0.01287	-2.90768
maximum_	-1.2E+09	465930.8	-79263.7	-20568.7	-30500	-5422515	-173683	1.87E+13	4444094	6811666	-9687.13	-25153.3	51642.78
availability	220896.1	-1.75007	1.146788	-1.75445	0.447063	1811.711	125.0978	4444094	15944.44	-309.061	-4.70073	-2.81294	-1.71084
number_o	-636811	146.5294	-10.9742	-6.50927	-5.60509	-1516.31	-64.1536	6811666	-309.061	3330.906	0.720372	1.161267	32.52593
review_sc	-241.135	-0.55443	-0.03147	-0.0024	-0.01118	0.582231	-0.0266	-9687.13	-4.70073	0.720372	0.139867	0.099851	0.02083
review_sc	-337.065	-0.35111	-0.04764	-0.01123	-0.02375	1.268718	-0.01287	-25153.3	-2.81294	1.161267	0.099851	0.137034	0.022167
reviews_p	2633.243	6.558164	-0.18562	-0.14363	-0.10726	-31.5976	-2.90768	51642.78	-1.71084	32.52593	0.02083	0.022167	1.722089

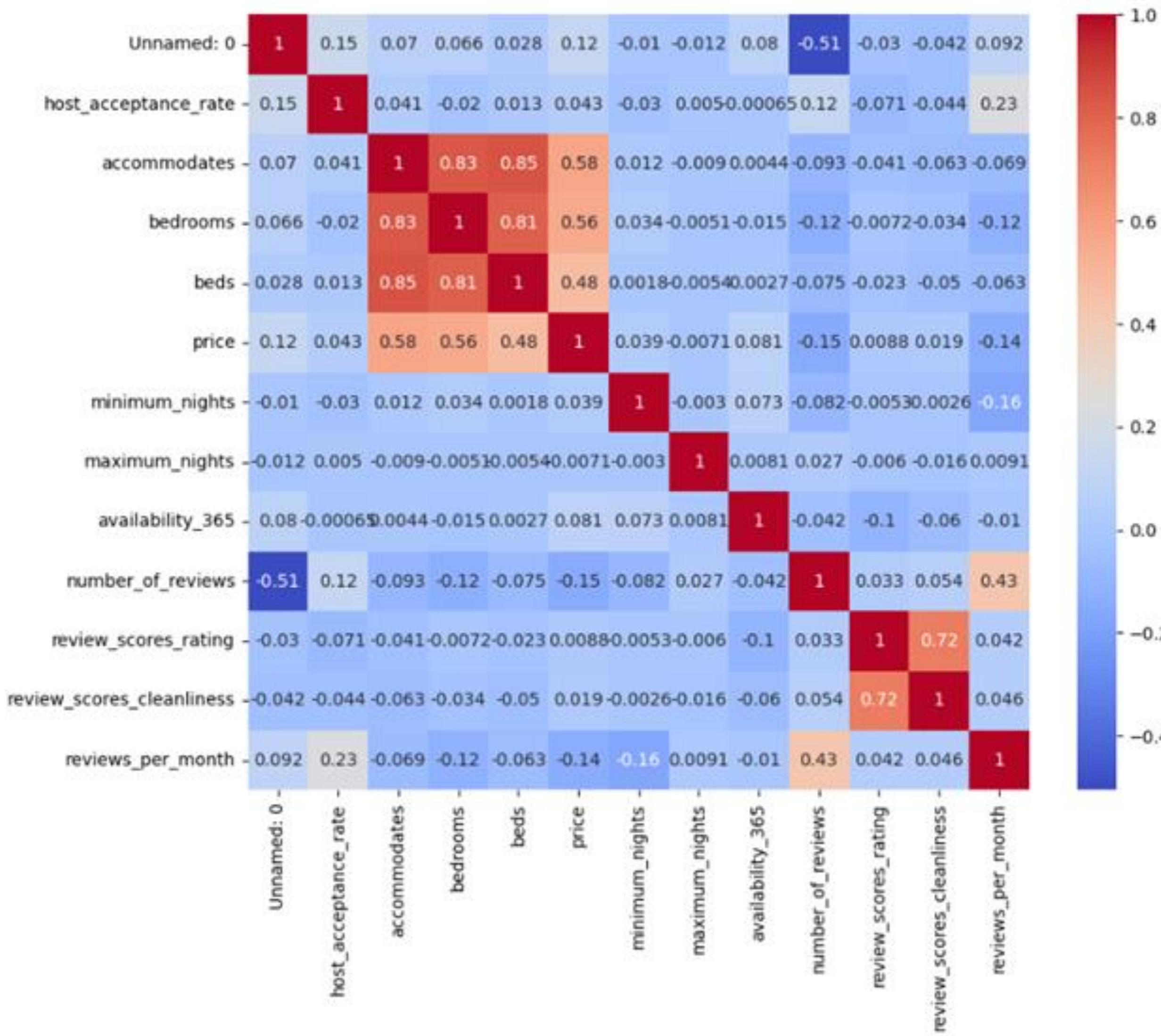
pav. 2covarience matrix

The covariance matrix shows the covariance values between each pair of attributes. The diagonal of the matrix shows the variance of each attribute. Covariance is a measurement of the correlation between two variables. Positive covariance between two variables indicates that they are more likely to move together than negative covariance does. The following details can be obtained from the covariance matrix: From top left to bottom right, the diagonal of the matrix displays the variance of each attribute. For instance, the pricing feature's variance is 31625,37973. The correlation between price and the number of reviews is -636811.0218, which shows that as an ad's price rises, less people will leave reviews.

While covariance calculates the combined variation of two variables, correlation describes the degree and direction of the linear relationship between two variables.
The analysis revealed that while some variables—such as lodging, bedrooms, and beds—had strong relationships with one another, others had little to no correlation at all. Price and some variables were shown to have positive correlations, while price and some other variables were found to have negative correlations. Also, while the covariance of some variables was positive (i.e., they tended to rise or fall together), the covariance of other variables was negative (indicating that when one variable increases, the other decreases).

When viewed as a whole, the research highlights the need of using both covariance and correlation to more fully comprehend the relationships between variables in a dataset.

Heatmap:


Unnamed: 0
host_acceptance_rate
accommodates
bedrooms
beds
price
minimum_nights
maximum_nights
availability_365
number_of_reviews
review_scores_rating
review_scores_cleanliness
reviews_per_month
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

8.Perform data normalization.

Normalisation done using the formula

(df2_norm[item]-df2_norm[item].min())/(df2_norm[item].max()-df2_norm[item].min())

9. Convert categorical variables to continuous type variables.

Done using label encoder , fit_transform from sklearn.preprocessing library

Conclusion:

In summary, we analyzed a large Airbnb dataset for London in 2022, identifying data quality issues and relationships between attributes through various calculations and visualizations. Our analysis provides valuable insights for data-driven decision making