European Accommodation Analysis

Abstract

This report delved into the relationship between price and customer rating of different accommodation across different countries and websites, from a Kaggle dataset. I was able to find evidence towards a relationship between rating and price. What I hypothesised was that the higher the rating, the higher the price a customer would pay. During the linear regression analysis, there wasn't much proof of a positive linear relationship, however once I grouped the ratings into categories, we could see that original hypothesis come to light. The prices are in USD. The results were, from cheapest to expensive (Rating Groups):

1. Rating 2-6
    a. Median: $87; Mean: $108.33
2. Rating 6-8
    a. Median: $125; Mean: $146.04
3. Rating 8-10
    a. Median: $146; Mean: $181.07

Other information on Rating Categories:

Price Summary Table for Each Rating Category

rating_category	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis
2-6	174.0	108.327586	72.650574	25.0	69.25	87.0	119.0	657.0	3.557507	19.731532
6-8	1657.0	146.037417	81.386208	9.0	92.00	125.0	172.0	658.0	2.181552	6.992919
8-10	3904.0	181.070441	118.444677	19.0	96.00	146.0	236.0	664.0	1.444685	2.180365

Using Kruskal-Wallis test and Dunn's post-hoc test, I was able to successfully identify this pattern. Statistically significant results tell us that the groups are all different and I was able to order them. There will be more details in the report but essentially, I found that there was a relationship between rating and price.

Introduction

Holidays involve travel, visiting attractions, food, entertainment, and accommodation. Accommodation now comprises of hotels and more recently rental homes with the rise of Airbnb. This report will mainly explore the relationship between the cost of accommodation and the customer's rating of the accommodation. The site used and the city location will also be looked at. The dataset is obtained from Kaggle. We will use basic statistics, basic regression modelling and analysing different groups using ANOVA or Kruskal-Wallis test to help us form conclusions.

Data Description

The data from Kaggle was in JSON file format. There were 5 JSON files containing information from Paris, London, Madrid, Berlin, and Rome. The JSON files individually contain data from 3 different accommodation sites: Hotels.com, Airbnb and Booking.com. Each sites data contained slightly different tables, but they all had in common price in USD and a rating score. Appendix A will provide a link to the dataset found on Kaggle.

Cleaning Process

Essentially, during the cleaning process, the aim was to get as much numerical data out of the original set with the aim to explore price and customer rating data. I ended up with these columns:

	city	price_value	rating_score	site
0	Berlin	111	8.1	Booking.com
1	Berlin	68	6.8	Booking.com

If you are more interested in the total process, there is more information found in Appendix B as well as the Jupyter Notebook.

Analytical Limitations

The only inferences we can make are based off a certain period as the ads for the various accommodations were scraped around May. For example, if you're planning on going during September or October, this set of data may or may not be relevant and accurate for that time. This issue also removes another aspect of analysis which is time series trends. We won't be able to predict any trends in dates and times that will be the best to go.

Outliers

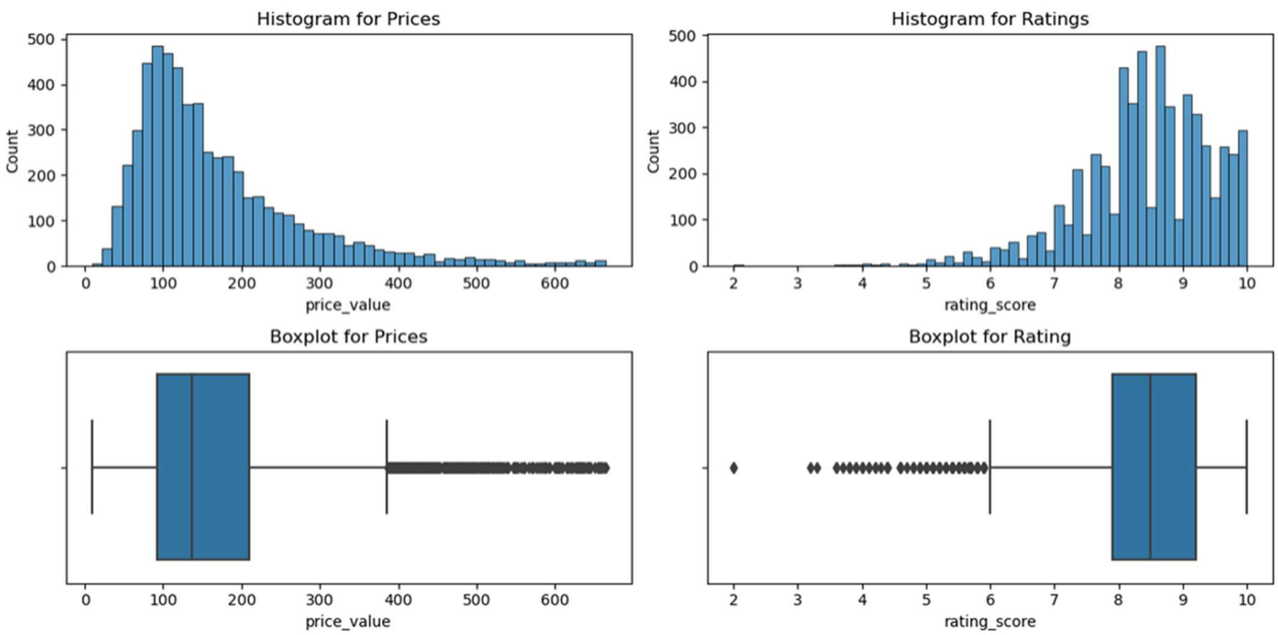
The main reason to remove outliers is to make the analysis more accurate as I wanted to focus on the average person. The process really focused on identifying and removing extreme outliers only in the price. There were some obvious ones that could be seen just by looking and then using the interquartile range (IQR) method, I used a high threshold number to specifically target 'extreme' outliers. Part of this process also involved having a look at what the outliers were and identifying where they mainly came from. In Appendix C and the Jupyter Notebook, there is more information on how the process works.

Analysis

The aim was to explore price and rating and see what their respective characteristics are. This will help us answer the question and maybe even find some other factors or interesting observations. Overall, there are a couple of key take aways from the EDA which were further explored using more advanced methods. These factors were the significance of what website you book with and even some locations were a lot pricier than others. Some of the main points and graphs will be below but the rest will be in Appendix D.

Exploratory Data Analysis (EDA)

During EDA, I was able to find out a lot about different groups and how they were distributed as well as creating a price vs rating scatterplot to see if there is a linear relationship present. The price data has a distinct right skewed distribution because of the very small number of luxury accommodation available. The rating data stopped at 10 and there were a few ratings below 6 potentially caused by one-off bad customer experiences which has caused a slight skew to the left for ratings.



	price_value	rating_score
count	5735.000000	5735.000000
mean	168.741412	8.415902
std	109.474724	1.039004
min	9.000000	2.000000
25%	93.000000	7.900000
50%	136.000000	8.500000
75%	210.000000	9.200000
max	664.000000	10.000000

The following tables will help us explore the characteristics of the booking sites as well as the cities.

This first one is the count:

Count Table

site / city	AirBnB	Booking.com	Hotels.com	Row Total
Berlin	244	499	493	1236
London	260	478	488	1226
Madrid	266	491	492	1249
Paris	232	486	497	1215
Rome	247	124	438	809
Column Total	1249	2078	2408	5735

This second will be the average price:

Average Price Table

site / city	AirBnB	Booking.com	Hotels.com	Row Average
Berlin	81.590164	120.537074	124.866126	108.997788
London	102.353846	218.242678	231.584016	184.060180
Madrid	97.556391	166.391039	153.831301	139.259577
Paris	91.612069	130.355967	222.362173	148.110070
Rome	79.850202	418.370968	301.636986	266.619385
Column Average	90.592534	210.779545	206.856120	169.409400

This third one will be average rating:

Average Rating Table

site / city	AirBnB	Booking.com	Hotels.com	Row Average
Berlin	9.552541	7.741683	8.316430	8.536885
London	9.471308	7.702301	8.156148	8.443252
Madrid	9.428797	8.219959	8.402033	8.683596
Paris	9.526207	7.539506	8.377867	8.481193
Rome	9.499595	8.213710	8.598174	8.770493
Column Average	9.495690	7.883432	8.370130	8.583084

We have not proved it yet in the EDA stage, but we can see to some extent that ratings and price are mainly impacted by which site is used. With price you can see that Booking.com and Hotels.com are very similar whilst Airbnb is over two times cheaper. Similar can be said about ratings, that each site has its own distributions. With the cities, we can see that there are some major differences in price and one reason we have seen already is the booking websites, but I believe there are some other factors outside of the scope of this dataset that have an influence. If you want to further explore the data, in Appendix D, there are some other tables and boxplots and violin plots that also show some of the above characteristics.

To go back to our original question of whether rating affects the price, I had a look at a scatterplot of price vs rating. Below you will see a graph that has some inklings of a linear pattern:

We can see that there is a lot more volume in the lower price range but there is a bit of a linear pattern and potentially with the right transformations or other techniques we might be able to make it a bit more linear. There won't be any non-linear regression in this analysis.

Linear Regression

To preface the linear regression analysis, all graphical and tabular data will be shown in Appendix E for more information. The dependent variable is Price, and the independent variable is Rating, and I performed one transformation.

The first model was price ~ rating. There were a couple of issues with this model, the first being that essentially all the assumptions were not met and secondly there was a very small R-squared value. This is telling us that rating does not explain price very well and to see if any major improvements could be made, I trialled and errored different transformations and nothing really worked too well. However, I decided to use a log transformation on the dependent variable which formed this model: log(price)~rating. Although this model created better assumptions and the AIC and BIC tests indicated that it was a better fitting model, the R-squared value moved more towards zero.

This can tell us that price and ratings do not have a linear relationship, in this specific context. I feel like this could be because prices in general are determined by more tangible factors such as location, competition, fees and so on, whereas rating is purely subjective and not everyone participates in the rankings. I believe that these factors affect the potential for a linear relationship and that's why I explored rating categories and the Kruskal-Wallis Test.

Kruskal-Wallis Test

The next analysis is to analyse the different groups found in the data but most importantly the rating categories. Normally we would use an ANOVA but doing Levene's Test (testing equal variance) and Shapiro-Wilk Test (testing normality) showed we needed to perform the non-parametric version, the Kruskal-Wallis Test (See Appendix F). Based off the results from the Kruskal-Wallis test as well as the post-hoc Dunn's test, all the group's prices are statistically significant in both tests meaning that every group is not equal to each other. However, the ratings ranking for each city is a bit different, we don't have enough evidence to support that Paris, Berlin and London's rating score are different but there was evidence of Rome and Madrid being not equal to all the other cities. Below are the rankings. All the Kruskal-Wallis test results are found in Appendix G.

Ranking the cities by cost (low to high):

1. Berlin
    a. Median: $105; Mean: $114.58
2. Madrid
    a. Median: $121; Mean: $146.78
3. Paris
    a. Median: $143; Mean: $160.59
4. London
    a. Median: $174.50; Mean: $198.98
5. Rome
    a. Median: $238; Mean: $251.81

Ranking the sites by cost (low to high):

1. Airbnb
    a. Median: $85; Mean: $90.83
2. Booking.com
    a. Median: $139; Mean: $173.92
3. Hotels.com
    a. Median: $180; Mean: $204.69

Ranking rating groups by cost (low to high):

4. Rating 2-6
    a. Median: $87; Mean: $108.33
5. Rating 6-8
    a. Median: $125; Mean: $146.04
6. Rating 8-10
    a. Median: $146; Mean: $181.07

Ranking cities by rating (High to Low; 3-5 can be interchangeable):

1. Rome
    a. Median: 9.0; Mean: 8.81
2. Madrid
    a. Median: 8.6; Mean: 8.55
3. Berlin
    a. Median: 8.4; Mean: 8.33
4. London

    a. Median: 8.4; Mean: 8.26
5. Paris
    a. Median: 8.3; Mean: 8.26

Ranking sites by rating (High to Low)

1. Airbnb
    a. Median: 9.6; Mean: 9.49
2. Hotels.com
    a. Median: 8.6; Mean: 8.37
3. Booking.com
    a. Median: 8.0; Mean: 7.83

Discussion

I believe the hypothesis I formed to be true because as we progressed up the rating groups, the mean and median prices increased. The group sizes were, 175 (2-6), 1678 (6-8) and 4034(8-10). A positive aspect to this is that 97% of all the data had ratings above 6 and 68.5% of the data had a rating above 8 meaning that the accommodation experience for customers overall was good. With the very different group sizes, I had to use the non-parametric Kruskal-Wallis Test. Throughout the analysis, there were also a lot of interesting factors to consider like which sites to use and which country as well as price and rating.

Airbnb was the site that the highest mean and median ratings, well above the other two booking sites. I am going to go off my personal experience with both hotels and Airbnb to try and figure out why we got these results. With Airbnb, I feel like it is mainly a place to stay and being satisfied is aligned with initial cleanliness and amenities, location, and owner communication. I believe what differentiates a hotel to an Airbnb is that hotels need to offer people that experience of feeling important and special. By providing restaurants and buffets, 24/7 customer service, facilities such as gym and swimming pools etc. In summary, there is a lot more to be critical of in a hotel compared to a Airbnb hence the rating differentials.

Ratings between countries were all relatively similar between Berlin, Paris, and London but Rome (1st) and Madrid (2nd) had statistical evidence to show that they were a little bit better. Compared to comparing the sites, the difference isn't as noticeable when comparing cities, and it shows that it doesn't matter where you go, your overall satisfaction of the accommodation isn't too much affected by which city you visit. Although we can conclude statistically that the accommodation that the customers rate the highest come from Airbnb, can we really conclude that it will provide the best experience, I think there are a couple more factors to try and find to be even more concrete in a conclusion.

With price, I believe a similar reasoning behind the differences can be down to the sites. However, there is a little bit more influence based off the city that the accommodation is located. Airbnb, seems steady no matter which country you are in but once we look at the hotel booking sites, that's where the variation comes in. Rome has the lowest Airbnb price mean but the hotel prices from the dataset in Rome are way higher than all the other cities. To further analyse, it would require a bit more research which I don't have the time or resources to do. Things like hotel competition, country's regulations, tourism policy etc.

Some other things to mention are the outliers. In the initial removal, there were 5 obvious points that I removed as the prices were too high and 4/5 were from Rome and Booking.com. Excluding the 5 outliers manually removed, I identified 152 outliers and 122 of them all came from Rome and all 152 were from hotel booking sites. None of the data points from Airbnb or Berlin got removed. With all these factors considered, I think there is enough to conclude the report.

Conclusion

To conclude the analysis, I can say that I was able to show that there is a relationship between price and rating. Although there was a large difference in the group's sizes, I think that has more to do with the distribution of ratings as most people were generally satisfied. It can be summarised by these statistics:

```
Medians:
Group 8-10: 146.0
Group 6-8: 125.0
Group 2-6: 87.0

Means:
Group 8-10: 181.0704405737705
Group 6-8: 146.0374170187085
Group 2-6: 108.32758620689656
```

I also delved into why there were differences in prices and ratings. The main difference was the actual difference between Airbnbs and hotels and the different uses, and expectations of each type of accommodation. At a hotel you're probably expecting a bit in terms of restaurants, daily customer service, etc… whereas in an Airbnb, if it is clean, functioning and the owner has good communication you're general very satisfied. I also saw different cities may have different rules and regulations or the countries have very different tourism policies which influence hotel prices.

Going back to the limitations, this data only covers 10–30-day span of accommodation ads. I think this analysis is useful and can bring someone some insight into accommodation prices in May-June time periods. Like all experiments, this will require a lot of repetition and maybe I will revisit this later and get some more data to analyse again. With the current analysis, we can take away that for the cheapest accommodation fares, use Airbnb because that has by far the cheapest rates. If you would prefer to go to a hotel, use Booking.com for cheaper prices compared to Hotels.com. Hotels seem to be more affected by the hypothesis which will lead to the final point which is lower rated hotels will be cheaper.

Appendix

A – Data Link

https://www.kaggle.com/datasets/mykhailozub/500-hotels-from-airbnb-booking-and-hotelscom?select=Berlin.json

B – Data Cleaning

All the columns in each accommodation site's page:

```
Column Comparison:

Columns in Hotels.com Data:
['title', 'isAd', 'location', 'snippet', 'paymentOptions', 'highlightedAmenities', 'price', 'rating', 'link']

Columns in Booking.com Data:
['thumbnail', 'title', 'stars', 'preferredBadge', 'promotedBadge', 'location', 'subwayAccess', 'sustainability', 'distance
FromCenter', 'highlights', 'price', 'rating', 'link']

Columns in Airbnb Data:
['thumbnail', 'title', 'subtitles', 'price', 'rating', 'link']
```

Here are those different dataframe's first rows:

Hotels.com:

	title	isAd	location	snippet	paymentOptions	highlightedAmenities	price	rating	link
0	Moxy Berlin Ostbahnhof	True	Friedrichshain	{'title': 'Lifestyle Hotel close to Ostbahnhof...	[]	[]	{'currency': '$', 'value': 107, 'withTaxesAndC...	{'score': 8.8, 'reviews': 596}	https://www.hotels.com/ho497828896/moxy-berlin...

Airbnb:

	thumbnail	title	subtitles	price	rating	link
0	https://a0.muscache.com/im/pictures/miso/Hosti...	Private room in Tempelhof	[Privatzimmer in Tempelhofer Feld, 1 bed, Jul ...	{'currency': '$', 'value': 31, 'period': 'night'}	5	https://www.airbnb.com/rooms/647664199858827562

Booking.com:

	thumbnail	title	stars	preferredBadge	promotedBadge	location	subwayAccess	sustainability	distanceFromCenter	highlights
0	https://cf.bstatic.com/xdata/images/hotel/squa...	Scandic Berlin Kurfürstendamm	NaN	True	True	Charlottenburg-Wilmersdorf, Berlin	True	Travel Sustainable property	3.2	[Standard Double Room, Beds: 1 double or 2 twi...

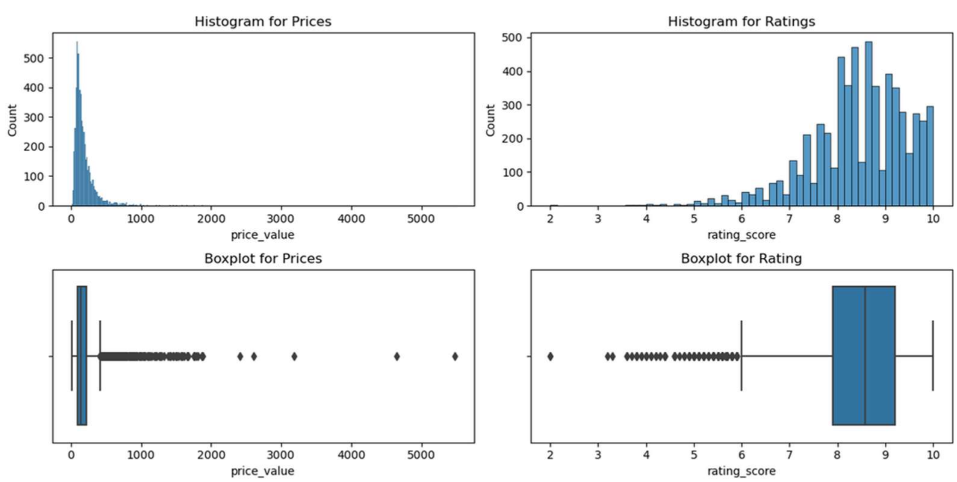
price	rating	link
{'currency': 'US$', 'value': 111, 'taxesAndCha...	{'score': 8.1, 'scoreDescription': 'Very Good'...	https://www.booking.com/hotel/de/scandic-kurfu...

The common data points in all 3 were price and rating. Some other potential data points to explore could have been ‘distanceFromCenter’ from Booking.com and Airbnb has a date that the booking is available for in the ‘subtitles’ column. The decision was made to just focus on price and rating. With a myriad of steps, the final columns and output will be shown below.

	city	price_value	rating_score	site
0	Berlin	111	8.1	Booking.com
1	Berlin	68	6.8	Booking.com
2	Berlin	104	8.4	Booking.com
3	Berlin	100	8.3	Booking.com
4	Berlin	135	8.3	Booking.com
...	...	...	...	...
5887	Madrid	218	8.0	Hotels.com
5888	Madrid	229	8.4	Hotels.com
5889	Madrid	287	8.8	Hotels.com
5890	Madrid	125	8.0	Hotels.com
5891	Madrid	178	10.0	Hotels.com

C – Outlier Identification Process

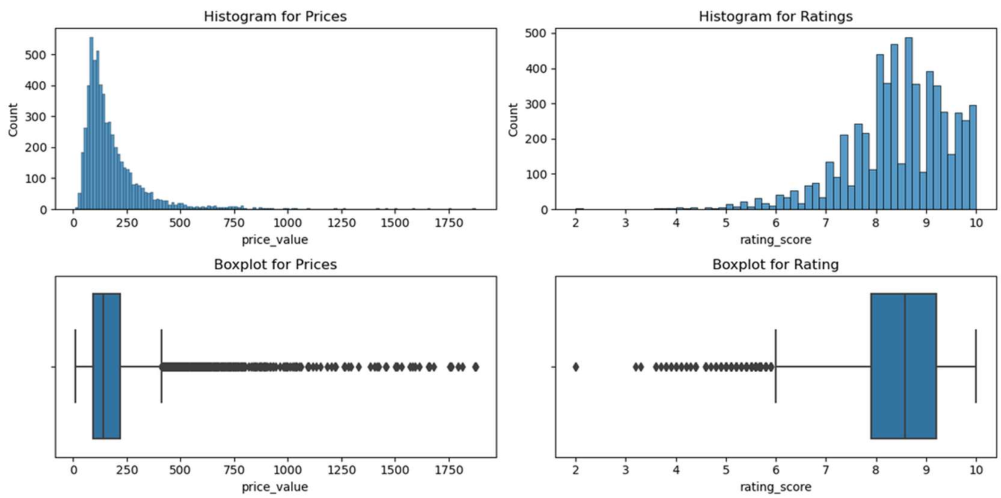
Below is the initial data:



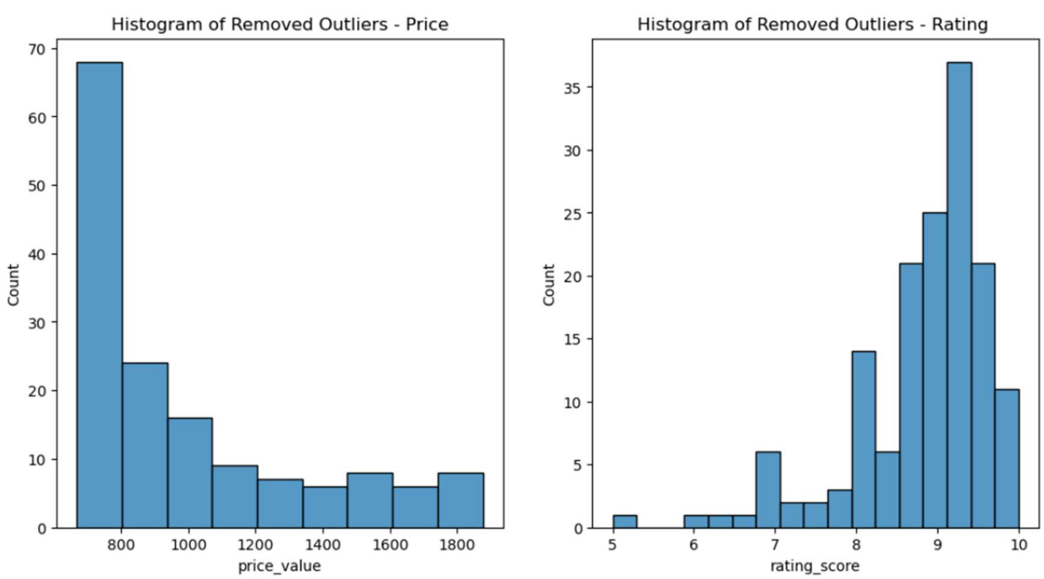
In the boxplot for price, you can see 5 distinct points at the right end of the plot, and I decided to remove these before further reducing using interquartile range (IQR) method. Those points were these:

	city	price_value	rating_score	site
1577	Rome	5476	8.0	Booking.com
1638	Rome	4653	9.4	Booking.com
1572	Rome	3182	8.4	Booking.com
1470	Rome	2606	9.0	Booking.com
4829	London	2406	8.0	Hotels.com

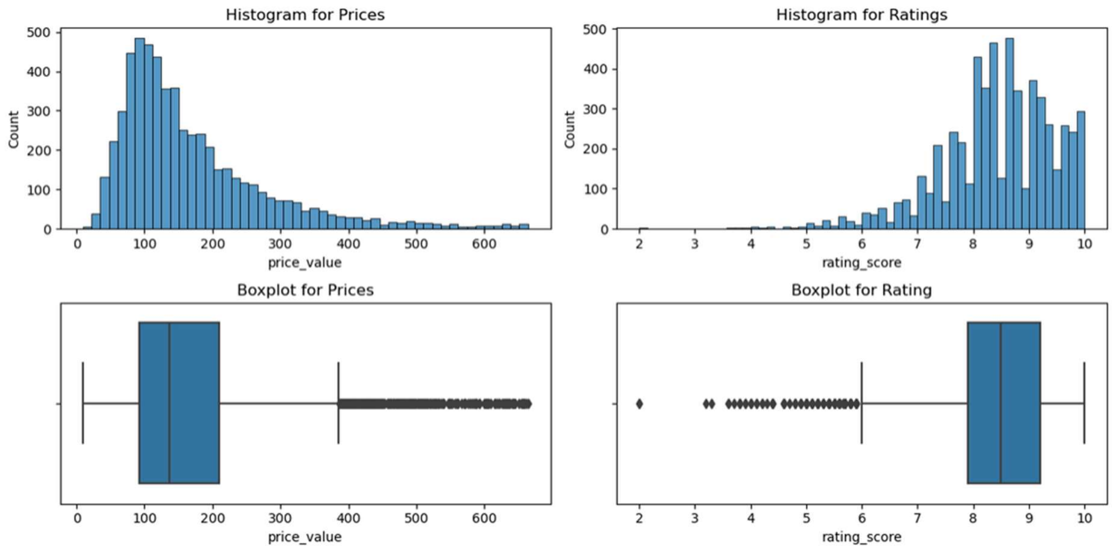
It's interesting to see that the top 4 points are all from the same site and country. Now here are the updated graphs:



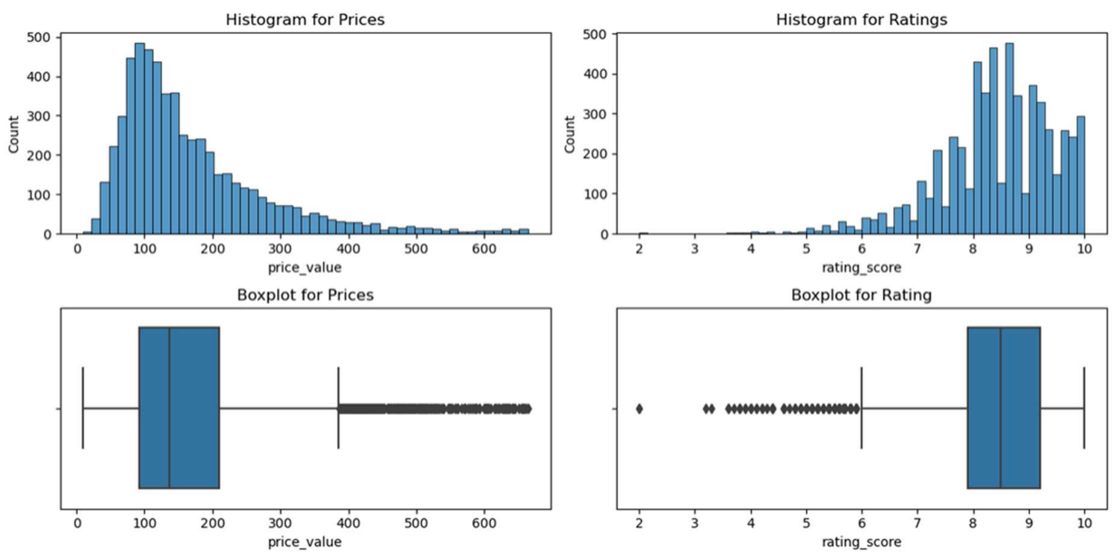
Now we can see that there are less obvious outliers, and it seems the nature of this data is that it is skewed to the right. Using the IQR method, I really wanted to focus a bit more on the common person's holiday. My threshold was quite high (3.5) as I only wanted the one considered extreme to be removed. Below are graphs showing what got removed:



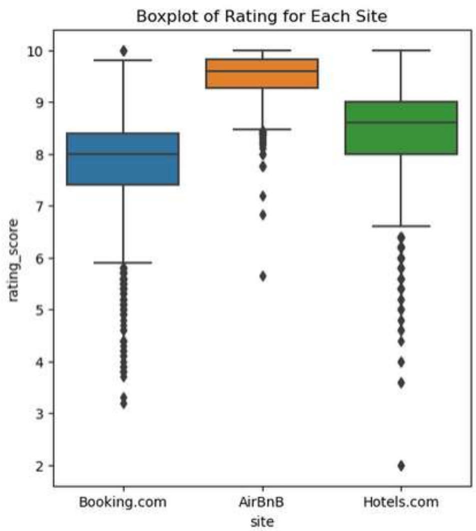
Now we can see that we'll only be working with accommodation priced roughly $700 and below. By still including some of the more expensive accommodations, it will make the analysis a bit more accurate since the nature of the marketplace of holiday accommodation will include the more upper end places. Here is the final refined data:

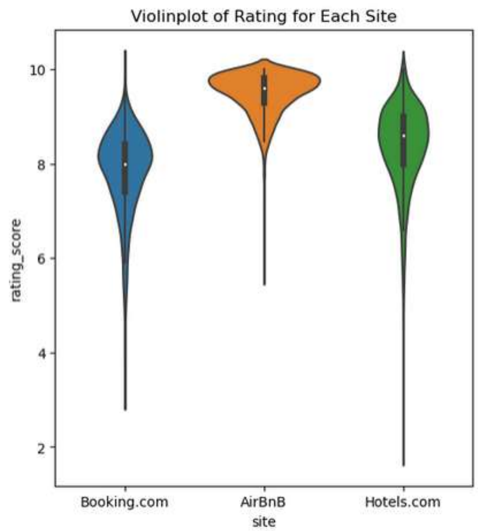


D – Exploratory Data Analysis Tables and Graphs



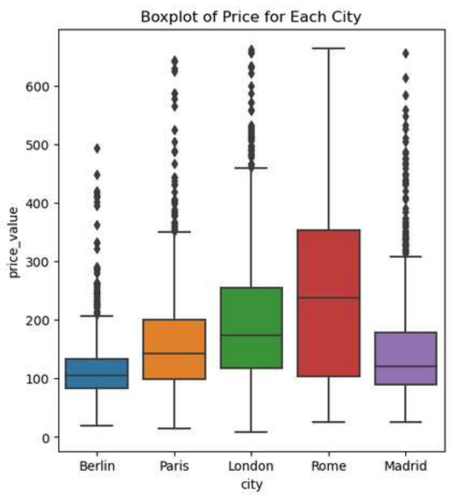
```
Rating Summary for Each Site

              count      mean       std   min   25%  50%   75%   max  \
site
AirBnB       1249.0  9.493915  0.446838  5.66  9.28  9.6  9.82  10.0
Booking.com  2078.0  7.826516  0.928293  3.20  7.40  8.0  8.40  10.0
Hotels.com   2408.0  8.365365  0.905036  2.00  8.00  8.6  9.00  10.0

             skewness  kurtosis
site
AirBnB      -1.710162  6.147000
Booking.com -1.291555  2.520890
Hotels.com  -1.411687  4.072082
```

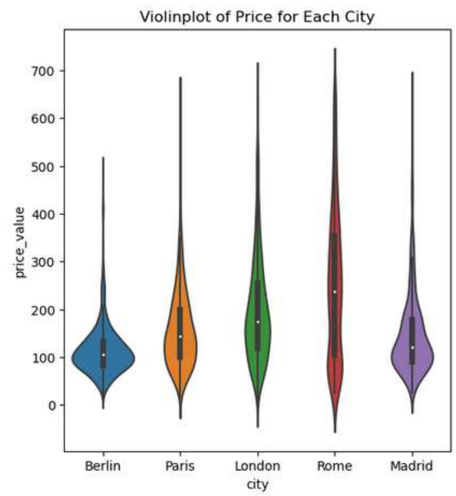




Price Summary Table for Each City

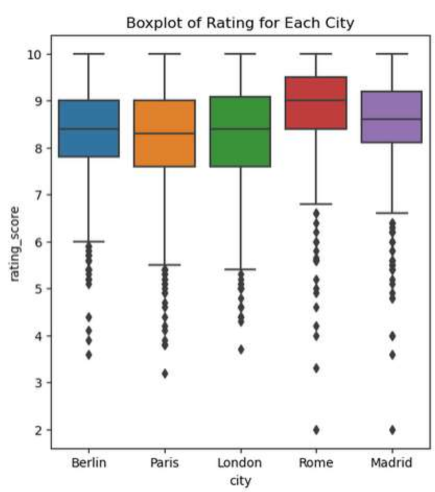
city	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis
Berlin	1236.0	114.575243	52.669361	19.0	83.00	105.0	133.00	493.0	2.315873	9.528976
London	1226.0	198.976346	112.566469	9.0	118.25	174.5	254.75	662.0	1.194951	1.625855
Madrid	1249.0	146.783827	86.215385	25.0	90.00	121.0	178.00	655.0	1.915274	5.007804
Paris	1215.0	160.593416	88.557891	15.0	99.50	143.0	200.00	643.0	1.681906	4.932163
Rome	809.0	251.814586	157.179031	26.0	103.00	238.0	354.00	664.0	0.603512	-0.328649

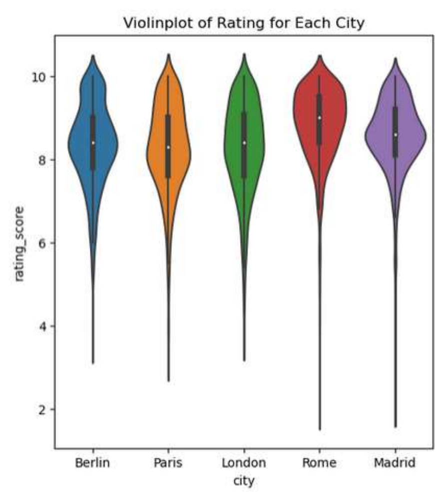




Rating Summary for Each City

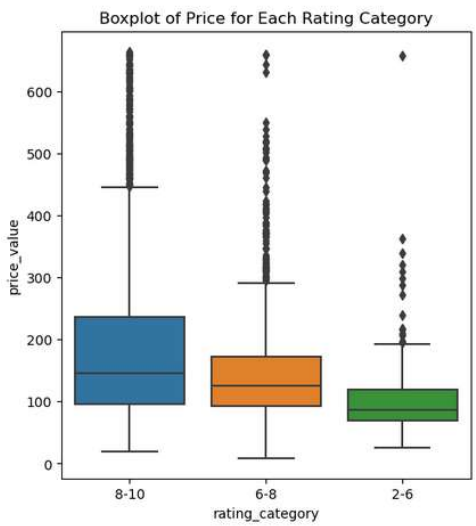
city	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis
Berlin	1236.0	8.328414	1.031212	3.6	7.8	8.4	9.000	10.0	-0.672041	0.821118
London	1226.0	8.258108	1.108666	3.7	7.6	8.4	9.075	10.0	-0.759898	0.597681
Madrid	1249.0	8.549127	0.895764	2.0	8.1	8.6	9.200	10.0	-1.395012	4.996952
Paris	1215.0	8.261794	1.085095	3.2	7.6	8.3	9.010	10.0	-0.811930	1.324243
Rome	809.0	8.814462	0.944444	2.0	8.4	9.0	9.500	10.0	-1.878589	7.041350

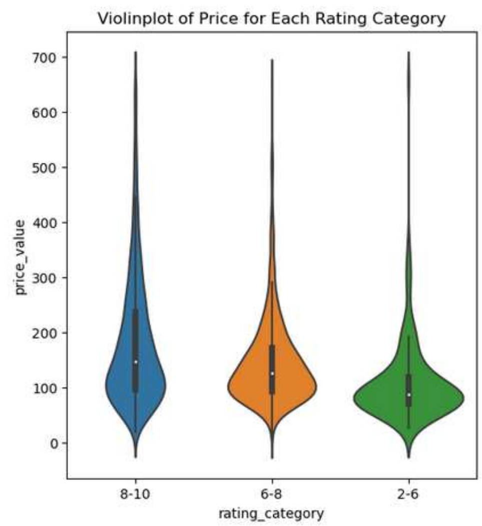




Price Summary Table for Each Rating Category

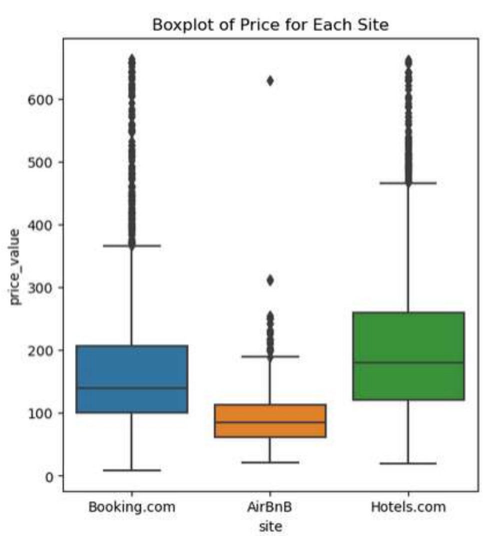
rating_category	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis
2-6	174.0	108.327586	72.650574	25.0	69.25	87.0	119.0	657.0	3.557507	19.731532
6-8	1657.0	146.037417	81.386208	9.0	92.00	125.0	172.0	658.0	2.181552	6.992919
8-10	3904.0	181.070441	118.444677	19.0	96.00	146.0	236.0	664.0	1.444685	2.180365

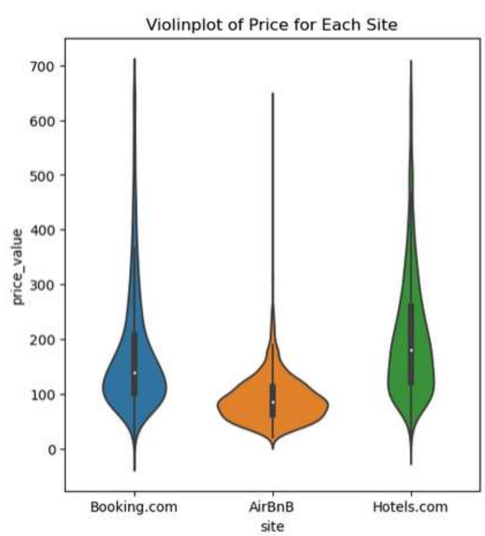




Price Summary Table for Each Site

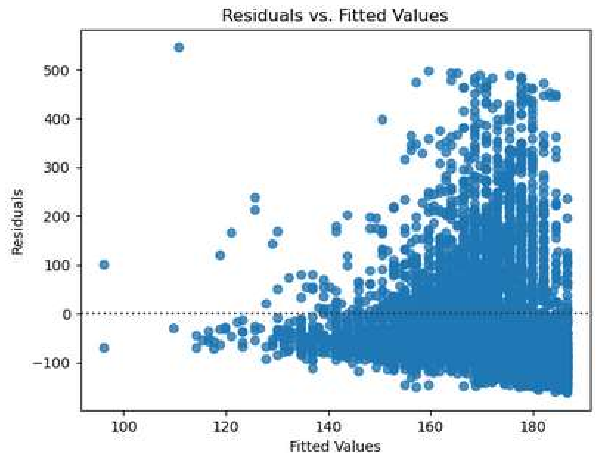
site	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis
AirBnB	1249.0	90.830264	42.829492	21.0	62.0	85.0	113.0	629.0	2.567087	21.314002
Booking.com	2078.0	173.915784	111.642226	9.0	101.0	139.0	207.0	664.0	1.891430	3.965211
Hotels.com	2408.0	204.687708	111.257971	19.0	121.0	180.0	259.0	662.0	1.295710	1.824151

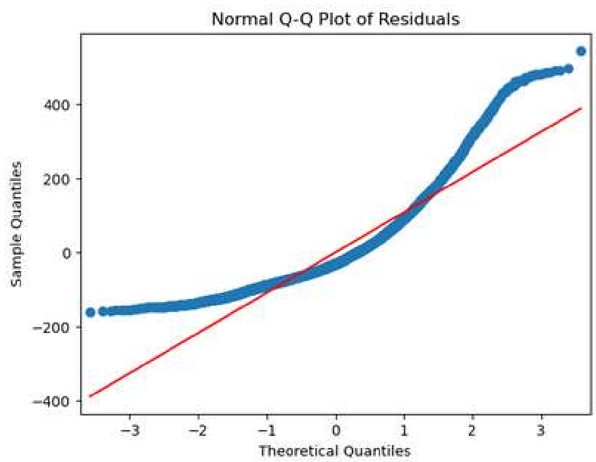


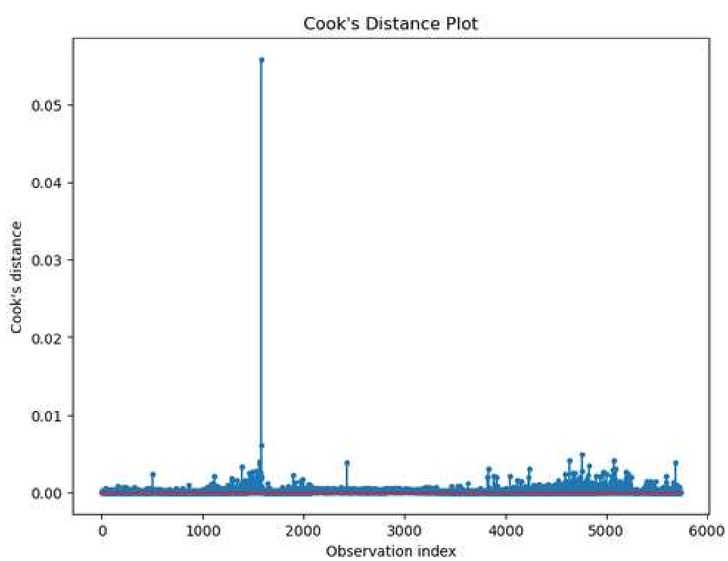


E – Linear Regression Assumptions and Models

Non-transformed Original Linear Regression



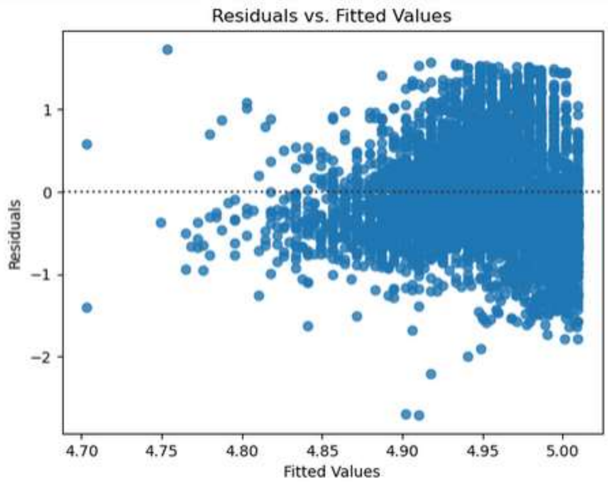


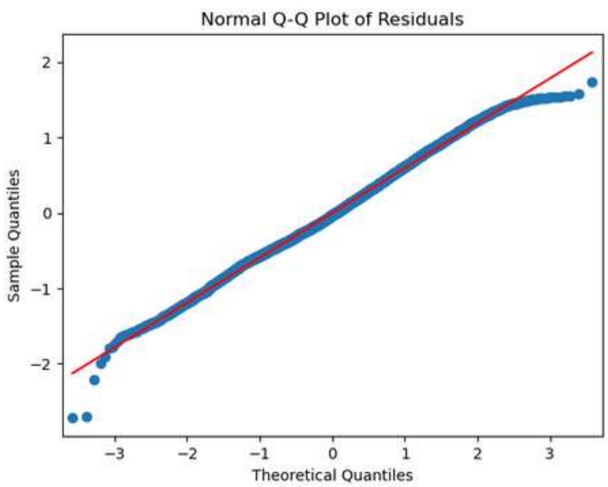


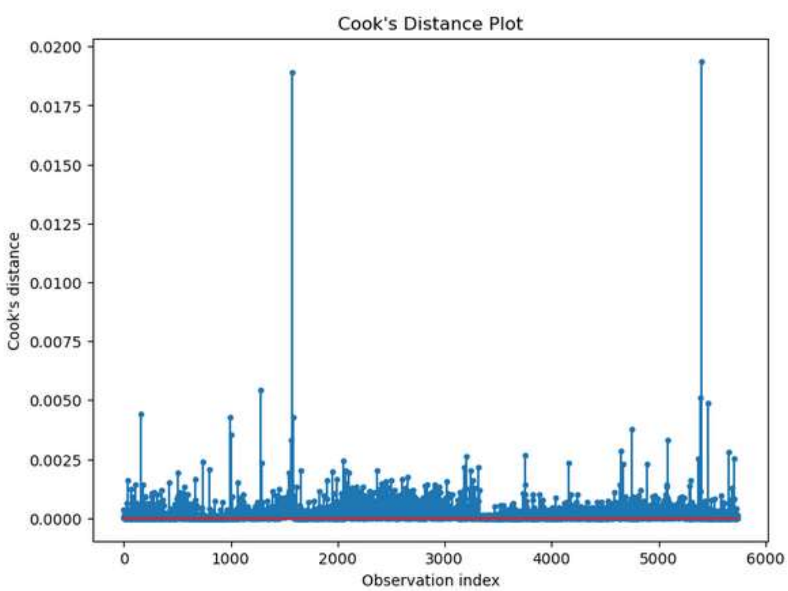
Rating predicting Price

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:            price_value   R-squared:                       0.012
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     66.96
Date:                Thu, 24 Aug 2023   Prob (F-statistic):           3.38e-16
Time:                        18:18:40   Log-Likelihood:                -35034.
No. Observations:                5735   AIC:                         7.007e+04
Df Residuals:                    5733   BIC:                         7.008e+04
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           73.4615     11.732      6.262      0.000      50.463      96.460
rating_score    11.3214      1.384      8.183      0.000       8.609      14.034
==============================================================================
Omnibus:                     1750.127   Durbin-Watson:                   0.970
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4917.359
Skew:                           1.618   Prob(JB):                         0.00
Kurtosis:                       6.179   Cond. No.                         70.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Log Transformed Linear Regression







```
                            OLS Regression Results
==============================================================================
Dep. Variable:            price_value   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     25.61
Date:                Thu, 24 Aug 2023   Prob (F-statistic):           4.30e-07
Time:                        18:21:31   Log-Likelihood:                -5156.2
No. Observations:                5735   AIC:                         1.032e+04
Df Residuals:                    5733   BIC:                         1.033e+04
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            4.6268      0.064     72.184      0.000       4.501       4.752
rating_score     0.0383      0.008      5.061      0.000       0.023       0.053
==============================================================================
Omnibus:                        2.129   Durbin-Watson:                   0.924
Prob(Omnibus):                  0.345   Jarque-Bera (JB):                2.168
Skew:                           0.042   Prob(JB):                        0.338
Kurtosis:                       2.956   Cond. No.                         70.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

F – Shapiro-Wilk and Levene's Test

```
City and Price
Group: Berlin
Shapiro-Wilk Test
Test Statistic: 0.8324776887893677
P-value: 4.066123580748085e-34

Group: Paris
Shapiro-Wilk Test
Test Statistic: 0.8854522109031677
P-value: 4.714863078323316e-29

Group: London
Shapiro-Wilk Test
Test Statistic: 0.9185727834701538
P-value: 3.6025894569669144e-25

Group: Rome
Shapiro-Wilk Test
Test Statistic: 0.9447432160377502
P-value: 8.362596807819607e-17

Group: Madrid
Shapiro-Wilk Test
Test Statistic: 0.8376408219337463
P-value: 7.41911055469579e-34

Levene's Test
Test Statistic: 245.12462657492824
P-value: 1.2116820910014845e-194
```

```
City and Rating
Group: Berlin
Shapiro-Wilk Test
Test Statistic: 0.9655061364173889
P-value: 1.4865639982549785e-16

Group: Paris
Shapiro-Wilk Test
Test Statistic: 0.9579793214797974
P-value: 3.2747586210558365e-18

Group: London
Shapiro-Wilk Test
Test Statistic: 0.9595345258712769
P-value: 6.051286506836283e-18

Group: Rome
Shapiro-Wilk Test
Test Statistic: 0.8699321746826172
P-value: 2.2063100920683153e-25

Group: Madrid
Shapiro-Wilk Test
Test Statistic: 0.9193646907806396
P-value: 2.763476632969127e-25

Levene's Test
Test Statistic: 24.81894505586204
P-value: 2.1101882550924412e-20
```

```
Site and Price
Group: Booking.com
Shapiro-Wilk Test
Test Statistic: 0.8109065294265747
P-value: 4.344025239406933e-44

Group: AirBnB
Shapiro-Wilk Test
Test Statistic: 0.8635073304176331
P-value: 1.3138436882445763e-31

Group: Hotels.com
Shapiro-Wilk Test
Test Statistic: 0.9003838300704956
P-value: 5.4517124390987e-37

Levene's Test
Test Statistic: 208.23888828904774
P-value: 4.9837166426056825e-88
```

```
Rating Category and Price
Group: 8-10
Shapiro-Wilk Test
Test Statistic: 0.874171257019043
P-value: 0.0

Group: 6-8
Shapiro-Wilk Test
Test Statistic: 0.8172882795333862
P-value: 8.108991914400922e-40

Group: 2-6
Shapiro-Wilk Test
Test Statistic: 0.6817498207092285
P-value: 6.533768964019582e-18

Levene's Test
Test Statistic: 106.16684588749516
P-value: 5.318632267719711e-46
```

```
Site and Rating
Group: Booking.com
Shapiro-Wilk Test
Test Statistic: 0.9181024432182312
P-value: 3.7332627943520726e-32

Group: AirBnB
Shapiro-Wilk Test
Test Statistic: 0.8715369701385498
P-value: 7.671531921036987e-31

Group: Hotels.com
Shapiro-Wilk Test
Test Statistic: 0.914094090461731
P-value: 6.337226650583487e-35

Levene's Test
Test Statistic: 156.8783403894428
P-value: 4.6549229233479106e-67
```

G – Kruskal-Wallis and Dunn's Test Results

Price per City

```
Kruskal-Wallis Test - Price per City
Test Statistic: 745.7292867246719
P-value: 4.361715016711756e-160

Medians:
Berlin: 105.0
Paris: 143.0
London: 174.5
Rome: 238.0
Madrid: 121.0

Means:
Berlin: 114.5752427184466
Paris: 160.5934156378601
London: 198.9763458401305
Rome: 251.81458590852904
Madrid: 146.78382706164933

Dunn's Test Results:
              Berlin         London         Madrid          Paris  \
Berlin  1.000000e+00  2.562871e-104  5.210910e-19  5.861648e-42
London  2.562871e-104  1.000000e+00  7.885959e-38  8.126765e-16
Madrid  5.210910e-19   7.885959e-38  1.000000e+00  2.158301e-06
Paris   5.861648e-42   8.126765e-16  2.158301e-06  1.000000e+00
Rome    2.035386e-118  1.463606e-04  1.323159e-52  5.114234e-28

                Rome
Berlin  2.035386e-118
London   1.463606e-04
Madrid   1.323159e-52
Paris    5.114234e-28
Rome     1.000000e+00
```

Price per Site

```
Kruskal-Wallis Test - Price per Site
Test Statistic: 1454.6825755309694
P-value: 0.0

Medians:
Booking.com: 139.0
AirBnB: 85.0
Hotels.com: 180.0

Means:
Booking.com: 173.9157844080847
AirBnB: 90.8302642113691
Hotels.com: 204.68770764119603

Dunn's Test Results:
                   AirBnB    Booking.com    Hotels.com
AirBnB       1.000000e+00  6.726196e-164  0.000000e+00
Booking.com  6.726196e-164  1.000000e+00  5.991161e-31
Hotels.com   0.000000e+00   5.991161e-31  1.000000e+00
```

Price per Rating Category

```
Kruskal-Wallis Test - Price per Rating Category
Test Statistic: 168.60664161258697
P-value: 2.4408043447886735e-37

Medians:
Group 8-10: 146.0
Group 6-8: 125.0
Group 2-6: 87.0

Means:
Group 8-10: 181.0704405737705
Group 6-8: 146.0374170187085
Group 2-6: 108.32758620689656

Dunn's Test Results:
               2-6           6-8          8-10
2-6   1.000000e+00  3.370468e-13  9.550351e-27
6-8   3.370468e-13  1.000000e+00  1.767993e-17
8-10  9.550351e-27  1.767993e-17  1.000000e+00
```

Rating Score per City

```
Kruskal-Wallis Test - Rating per City
Test Statistic: 238.42236855963924
P-value: 2.0285483422541933e-50

Medians:
Berlin: 8.4
Paris: 8.3
London: 8.4
Rome: 9.0
Madrid: 8.6

Means:
Berlin: 8.3284142394822
Paris: 8.261794238683127
London: 8.25810766721044
Rome: 8.814462299134734
Madrid: 8.549127301841473

Dunn's Test Results:
              Berlin        London        Madrid         Paris          Rome
Berlin  1.000000e+00  3.432081e-01  3.673374e-08  1.725838e-01  1.208249e-32
London  3.432081e-01  1.000000e+00  1.156493e-10  6.763812e-01  4.429843e-37
Madrid  3.673374e-08  1.156493e-10  1.000000e+00  7.405888e-12  2.094005e-12
Paris   1.725838e-01  6.763812e-01  7.405888e-12  1.000000e+00  4.752851e-39
Rome    1.208249e-32  4.429843e-37  2.094005e-12  4.752851e-39  1.000000e+00
```

Rating Score per Site

```
Kruskal-Wallis Test - Rating per Site
Test Statistic: 2547.3306451585404
P-value: 0.0

Medians:
Booking.com: 8.0
AirBnB: 9.6
Hotels.com: 8.6

Means:
Booking.com: 7.826515880654475
AirBnB: 9.493915132105686
Hotels.com: 8.365365448504983

Dunn's Test Results:
                    AirBnB   Booking.com     Hotels.com
AirBnB        1.000000e+00  0.000000e+00  5.421264e-278
Booking.com   0.000000e+00  1.000000e+00   1.836290e-77
Hotels.com   5.421264e-278  1.836290e-77   1.000000e+00
```