**Trip Advisor Data Analysis**

**Data Analysis**

Math Check

Being a mathematician, I could not help myself and had to check the math on my scraped data. I used Excel for this. I summed the values from the 5 different rating categories and compared this value to the Total Number of Reviews. Out of the 82 observations, my sum was greater for just three of the sites (by 1, 2, and 2).

Next, I checked the average review score with my own average calculation (rounded to 0.5 as TripAdvisor does) and these checked out exactly.

**Summary Statistic Observations**

*Assumption*: All reviews are from valid guests.

For the TripAdvisor site data scraped:

**The Mean Average Review is 4.05**

This is rather high on a 5 point scale. When we sum the review columns, we see that the 27869 guests gave the highest rating of excellent which is more than double the next rating of Very Good (12223) and more than all of the lower categories combined (19406). This surprised me because I believed that people who were upset with their experience were more likely to take the time to leave a review.

The distribution of the number of each type of review can be seen here in this bar chart.

This Box-And-Whisker plot gives us an image of how the customer reviews are distributed into quartiles.

\* For the sites that included a star rating.



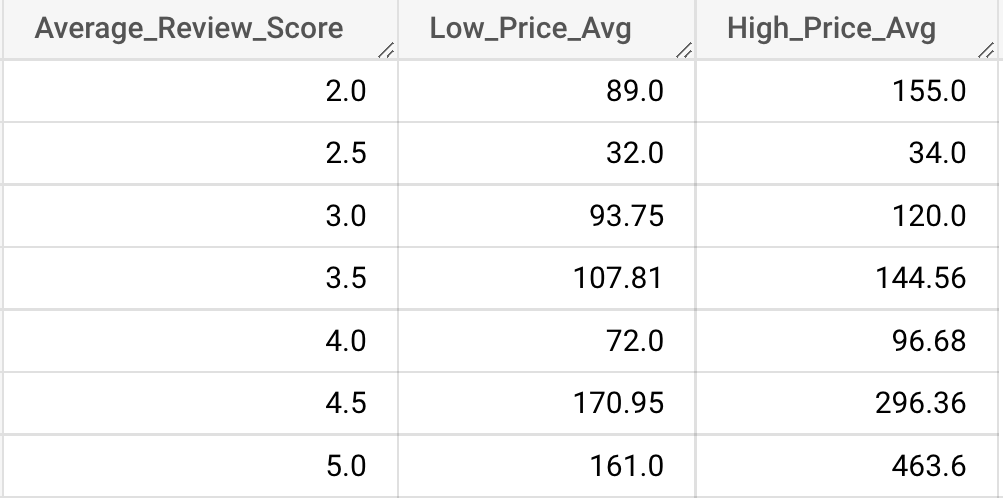
**Star Ratings and Average Customer Reviews Compared to Averages of Low and High Prices**

The highest and second highest star ratings have the highest and second highest ‘high price averages’ and ‘low price averages’, respectively. This pattern does not follow for the two lower ratings assigned. But for these higher ratings it is typical for something at a higher price to be considered to have a higher value.

Table

Description automatically generated

Similar findings can be seen with the highest and second highest ‘high price averages’ and ‘low price averages’ for the 2 best customer average review scores.



**Star Rating to Avg Review Score**

\* For the sites that included a star rating.

I was curious if there was a relationship between the assigned star rating and the calculated average review score. You can see from this scatter plot that there is a positive correlation. Perhaps the third parting supplying the star rating has a similar way of assigning a score, uses customer reviews from one or more websites, or reports from experience of scouts.

**Avg of Low and High Price to Number of Rooms**

\* For the sites that included a price range.

I was also curious to see if there was a relationship between the average of the price range and the number of rooms. While there does seem to be a correlation here, it is best observed at the extremes where we see the highest prices (around $600 and $700) in lodgings with under 50 rooms and the lowest prices (about $100 or less) in lodgings with 250 or more rooms. Perhaps the venues with less rooms offer guests more personal attention. They could be in quaint remote areas or bed and breakfasts with some small-town or nature experiences. And perhaps the hotels with more rooms are large chains that can afford huge properties and offer discounts.

**Considerations for further exploration**

Interested in trying out some urllib functionality for scraping because it seems more powerful than Requests and beautifulsoup

**Resources**

[Choose the Best Python Web Scraping Library for Your Application](https://towardsdatascience.com/choose-the-best-python-web-scraping-library-for-your-application-91a68bc81c4f) (data scraping method selection)

[How To Work with Web Data Using Requests and Beautiful Soup with Python 3](https://www.digitalocean.com/community/tutorials/how-to-work-with-web-data-using-requests-and-beautiful-soup-with-python-3)   
[Cannot Web Scraping Tripadvisor](https://stackoverflow.com/questions/71181932/cannot-web-scraping-tripadvisor) (addressing issues with scraping TripAdvisor)  
[Beautiful Soup Nested Tag Search](https://stackoverflow.com/questions/46510966/beautiful-soup-nested-tag-search) (handling nested tags you want to scrape)