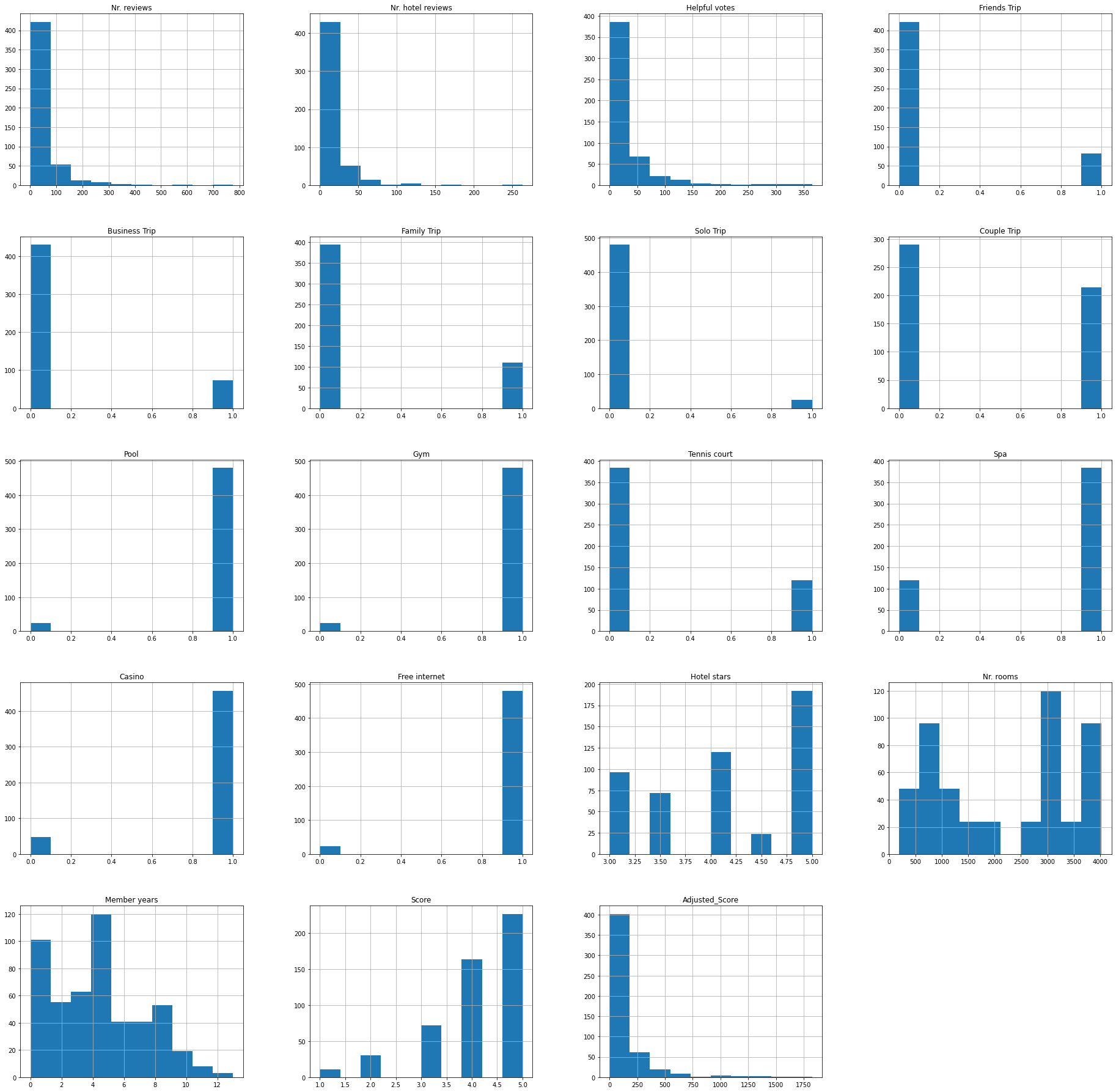
For my final project, I have decided to research on what makes a hotel successful in terms of its ratings. For this intent, I found a dataset on UCI which contains 504 TripAdvisor reviews for 21 hotels in Vegas.

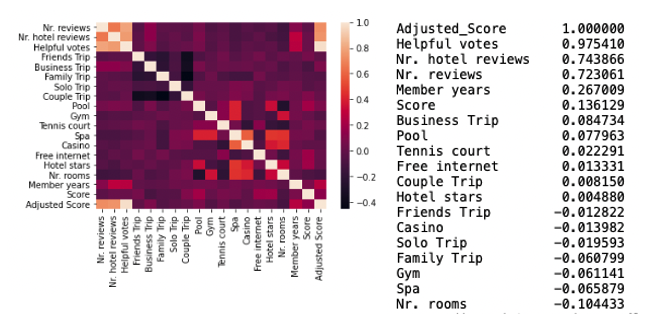
The dataset contains various attributes such as:

* Nr. Reviews = How many reviews the users have submitted on TripAdvisor
* Nr. Hotel reviews = How many hotel reviews the users have submitted on TripAdvisor
* Helpful votes = How many likes the review got
* Score = The score the user gave for that review
* Traveler type = Nature of trip (business, personal)
* Pool = Has pool
* Gym = Has Gym
* Tennis Court = Has Tennis Court
* Spa = Has Spa
* Casino = Has Casino
* Free Internet = Has Free Internet
* Hotel name = name of hotel
* Hotel stars = Amount of stars hotel has
* Nr. Rooms = Number of room hotel has
* Member years = How long the user has been a member on TripAdvisor

To adjust for the number of helpful votes, a new attribute was created named “Adjusted Score” which multiplied score by helpful votes. The reason behind this is that a helpful vote is assumed to be a duplicate review, meaning that whoever upvoted that review would leave the same review. Some member years are also negative, they will all be replaced with a zero.



Initial Exploration of the data, reveals that most reviewers are fairly new to TripAdvisor as the distribution of member years is right skewed. It is also worth noting that most hotels have a pool, gym, spa, casino, and free internet and only a few have tennis courts. The shape of the distribution of score is left skewed, but right skewed once adjusted for the number of helpful votes.



After analyzing the correlation of the data, we can observe that the attributes of guests overall have a greater impact on the review score compared to the amenities the hotel provides. This can be seen by how many years of membership and overall reviews greatly positively correlate to the adjusted score. It is also worth noting that business trips generally leave better reviews, while family trips leave worst reviews. The number of stars also correlate positively to the adjusted score, but not as much if the hotel has a pool, tennis court, and free internet. It is also worth noting that having a huge number of rooms negatively correlate to the adjusted score, along with the hotel having a spa, gym, and a casino.

Linear Regression was also performed for this analysis, once with all of the data, and another where the dataset was split in two.

All of the reviews

Adjusted Score = -205.786 + 1.188 (Nr. Reviews) + 3.908 (Nr. Hotel reviews) + 12.178 (Friends Trip) – 12.351 (Business Trip) – 7.478 (Family Trip) – 4.411 (Solo Trip) + 12.061 (Couple Trip) + 31.119 (Pool) + 22.717 (Gym) + 31.695 (Tennis Court) – 44.355 (Spa) + 46.343 (Casino) + 32.170 (Free Internet) + 20.516 (Hotel Stars) + 0.00281 (Nr. Rooms) + 5.850 (Member Years)

Accuracy = 0.671

Split into test and train 50/50

Adjusted Score = -179.952 + 1.159 (Nr. reviews) + 3.988 (Nr. hotel reviews) + 20.324 (Friends Trip) + 3.402 (Business Trip) – 6.904 (Family Trip) – 16.417 (Solo Trip) – 0.404 (Couple Trip) + 57.471 (Pool) – 29.892 (Gym) + 16.879 (Tennis court) – 44.199 (Spa) + 62.997 (Casino) + 70.272 (Free Internet) + 5.714 (Hotel stars) + 0.00688 (Nr. rooms) + 7.192 (Member years)

Accuracy = 0.590

As we can see, the first model seems to generate a better accuracy, however, the second model should be considered too to avoid overfitting.

In conclusion, in order for a hotel to have a better overall social media presence, they should focus more on attracting hotel critics rather than improving on amenities. If a hotel has not had one, they should know that guests prefer pool, tennis court, and free internet, over casino, gym, and spa. The hotel should also be aware that business-oriented guests tend to leave better reviews than other types, and that guests tend to leave a better rating is the hotel has lesser number of rooms.