Inferential Statistics

Through exploratory data analysis and the usage of scatterplots we have found some relationships and correlations that are worth looking at. Bedrooms and Baths are correlated not surprisingly which could cause some bias in our model because of multi-collinearity. Another interesting pattern I saw was number of reviews is correlated to value of the scores. The more reviews an AirBNB property has the higher the reviews of the property which makes sense because the more feedback a host is receiving the more likely they are working on whatever issues they have gotten complaints for. Another correlation is between the guests included and accommodates columns, as they are positively correlated. This is also a trend that is not surprising but should be taken into account when we parse our data through the machine learning algorithms.

There are also some other interesting correlations on the heat map, there seems to be 1.0 strength correlation between number of reviews, review accuracy, review score, cleanliness score, location score, and check-in scores. The other strong correlations come from the relationships discussed previously with bedrooms, bathrooms, beds, accommodation number and guests included. The other data seems to have very weak correlations with each other at .25 and below. It is peculiar because one would think that review score or cleanliness score would be a driving factor of price. If the property is well-maintained and cleaned that should be reflective on the price. The largest driving factor of price seems to be accommodates, bedrooms and bathrooms but are there other factors at play here? I think this data set would benefit geo-spatial dictionaries for the latitude and longitude columns (For example, where in relation is this property to things like public transportation, airport, landmarks, restaurants etc.)