Milestone Report

Overview:

As of 2020, Airbnb has 7M+ listings worldwide and 2M+ people staying on Airbnb per night. “We continue to see incredible growth in the Bay Area - even while tourism is essentially flat in the region - with 44 percent growth year-over-year in guest arrivals in the five counties surrounding San Francisco.” (new.airbnb.com) Airbnb has become an important source of supplemental income and the company is only continuing to grow, leading to a healthy competition among Airbnb hosts.

Issue:

San Francisco has over 7,800 listings and over 3,700 long term and hotel listings (as of Feb. 2019). How can a host stand out from the crowd and increase revenue? This analysis aims to determine the most important characteristics for a top listing and provide an ideal price range for rentals based on those factors. (edit factors later) The increase in revenue for hosts will lead to increased revenue for Airbnb, happier hosts and an improved business model for pricing listings.

Questions to explore:

- How are rental properties distributed?
- Which neighborhoods consist of the highest/lowest prices?
- What listing features are most important for guests?
- How do bookings fluctuate based on season?
- What keywords differentiate a favorable vs. unfavorable listing?
- What is the ideal listing range based on features and amenities?

Approach:

Data Wrangling: load data into Juptyer notebook and “clean” the dataset
Data Story: create tables and graphs to better depict data
Statistical Data Analysis:
Milestone Report:
**update over time

Data:

The data is provided on http://insideairbnb.com/

Deliverables:

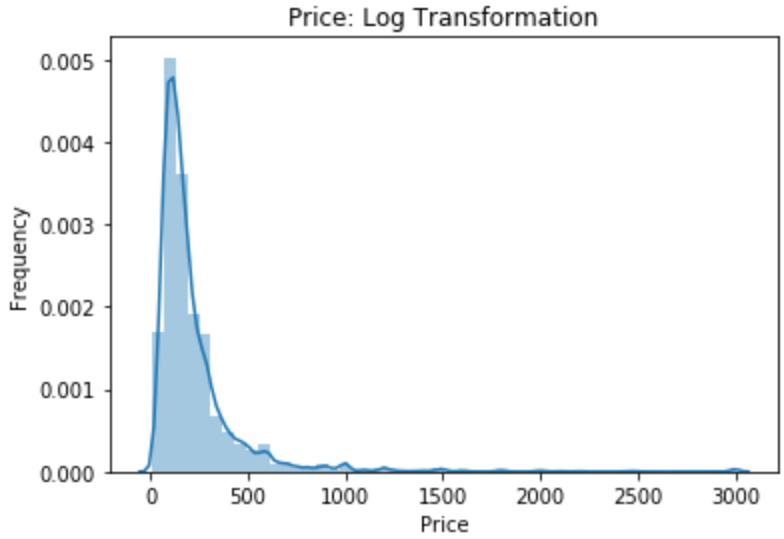
- Code in Jupyter notebook
- Report that outlines steps taken to determine conclusion
- Slide deck for presentation

Data Wrangling

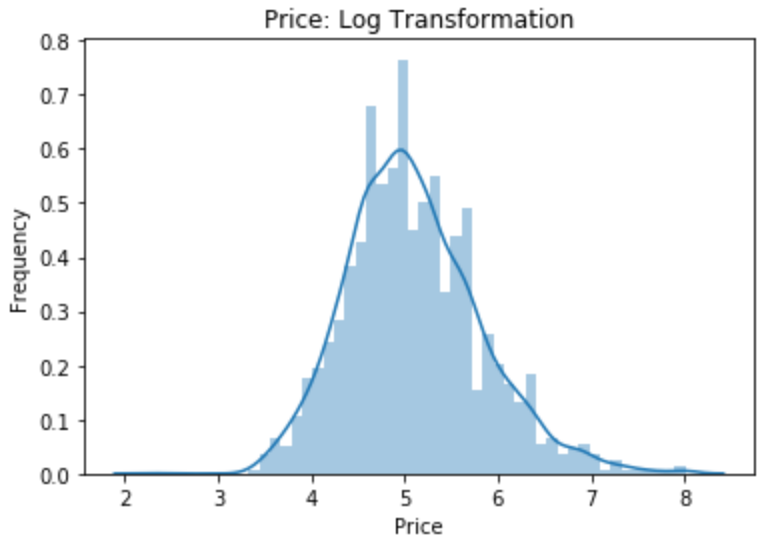
Collecting Data:
The data was sourced from http://insideairbnb.com/get-the-data.html which is an independent, non-commercial website that utilizes publicly available information about San Francisco's Airbnbs. The listings.csv.gz file was obtained and the raw dataset contained 17 columns and 8533 records. The data was compiled on December 4th, 2019.

Exploratory Data Analysis

Price

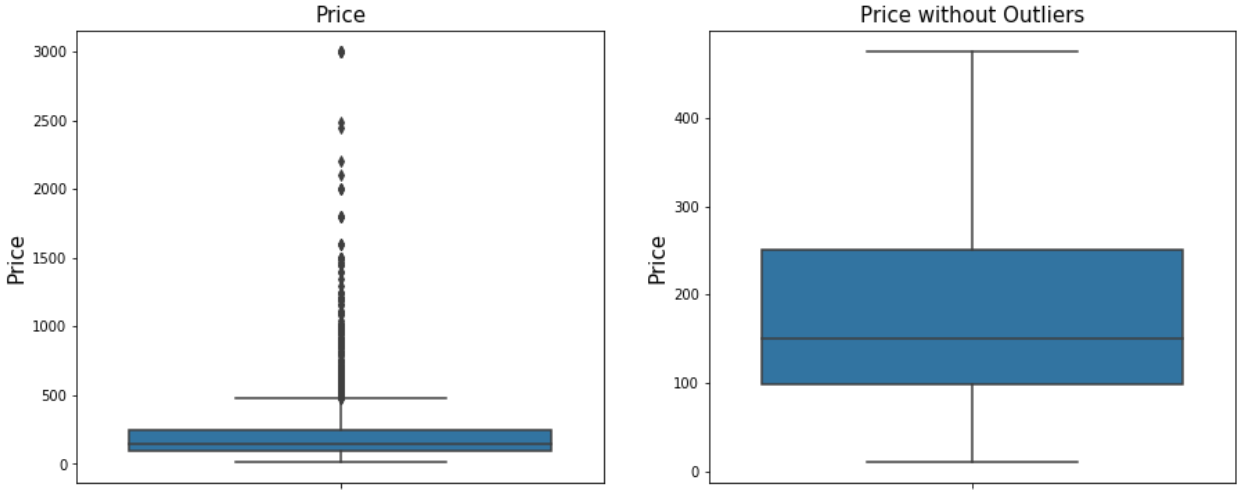


The price histogram shows strongly right-skewed data with a long tail. The majority of prices are very low in comparison to the few high-priced Airbnbs which skew the data.



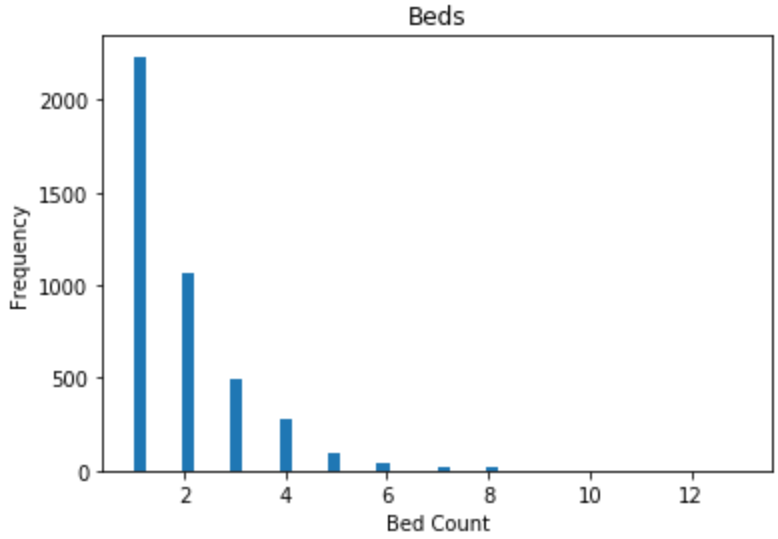
When the price is log-transformed, it becomes much more normally distributed.

Normality test on logged price:
NormaltestResult(statistic=233.08008658455662, pvalue=2.43950817269063e-51)

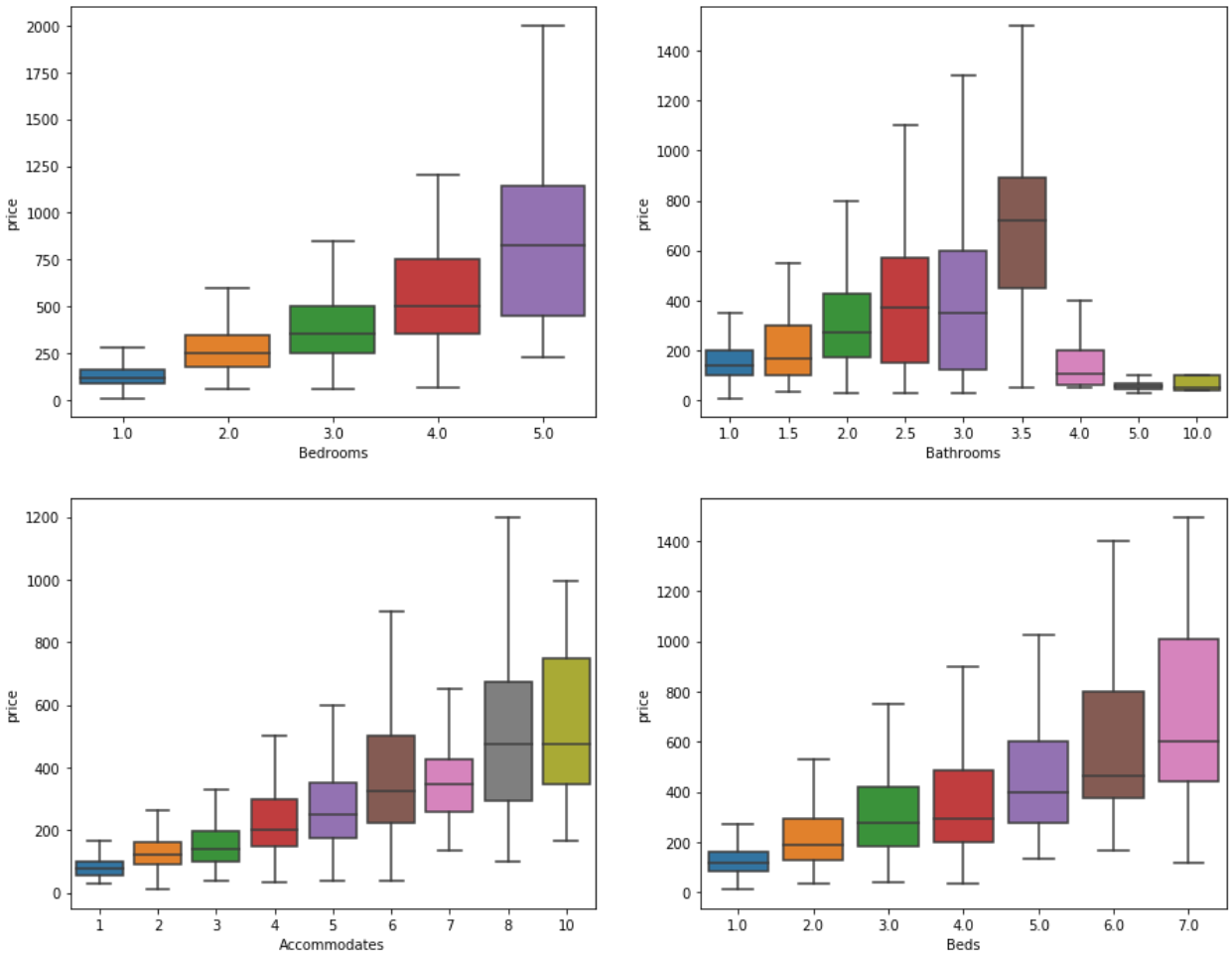


The boxplots above show how highly right-skewed the data is. The boxplot without outliers better depicts the price range between $100 and $250

Accommodations

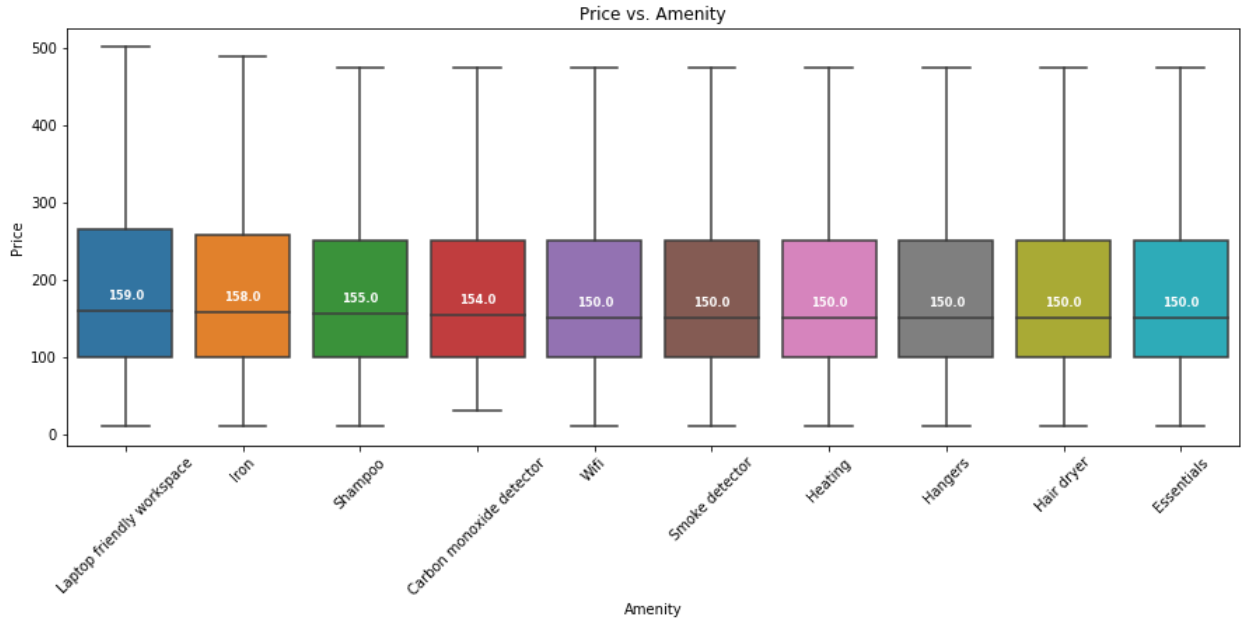


The vast majority listings offer one bed, again showing Airbnb strongly caters to solo and coupled guests. The amount of listings decreases as each bed count increases.



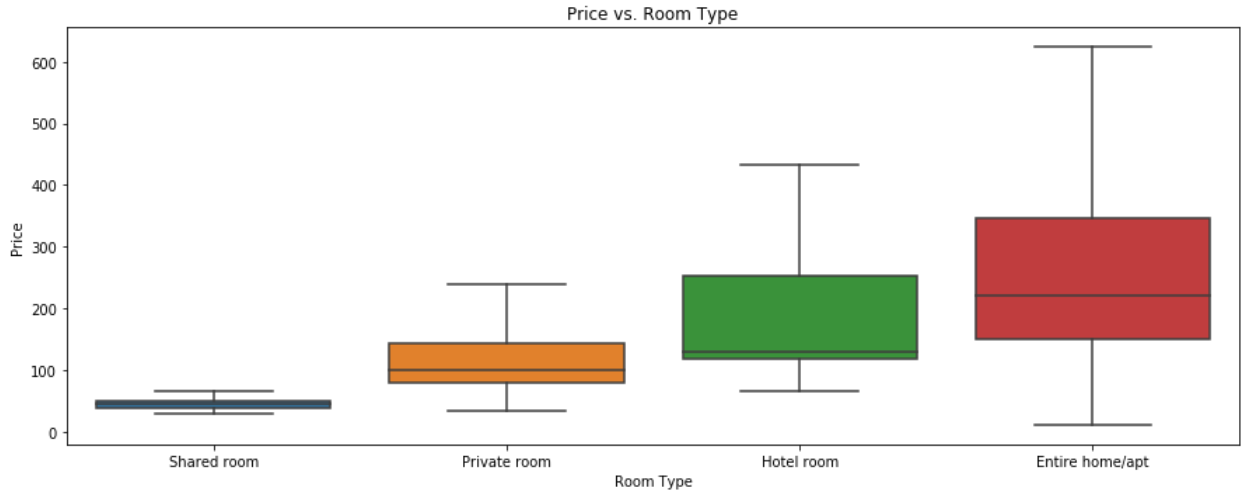
Bathrooms appear to be less correlated to price compared to bedrooms, accommodates and beds. All values with a sample size less than 20 were omitted from the boxplots. In addition, all features are significantly correlated with price based on the p-values obtained from the Point Biserial Correlation.

Amenities



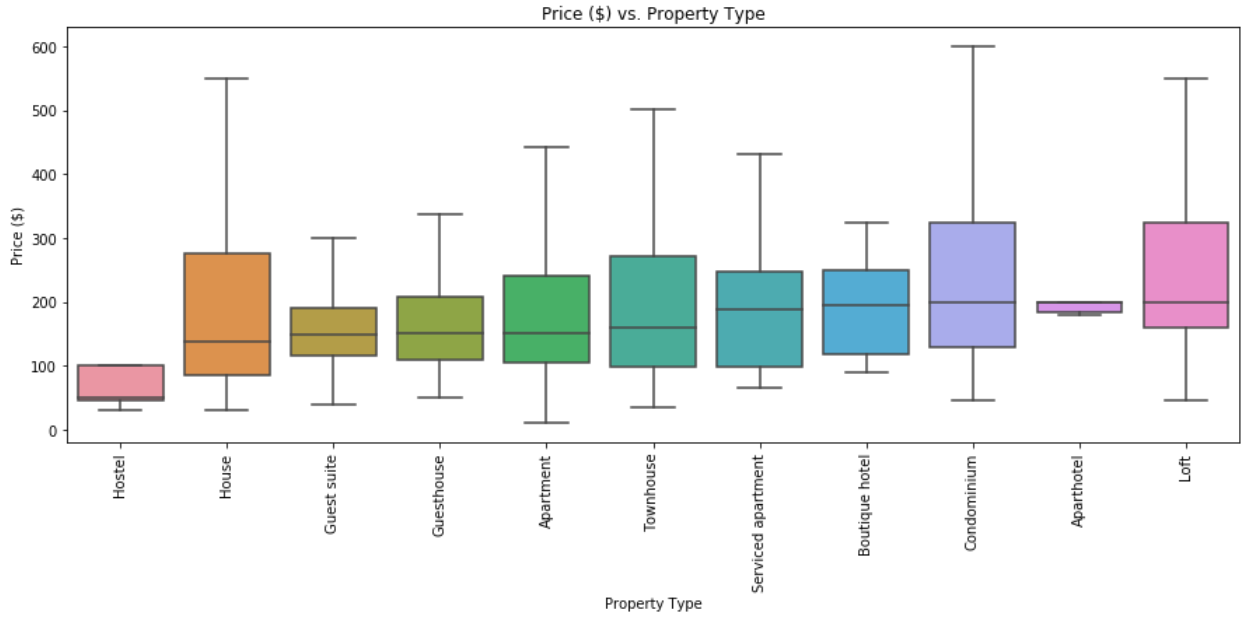
There is no visible difference between the top 10 amenities. All amenities except wifi and essentials are all significantly correlated with price based on the p-values obtained from the Point Biserial Correlation.

Room Type

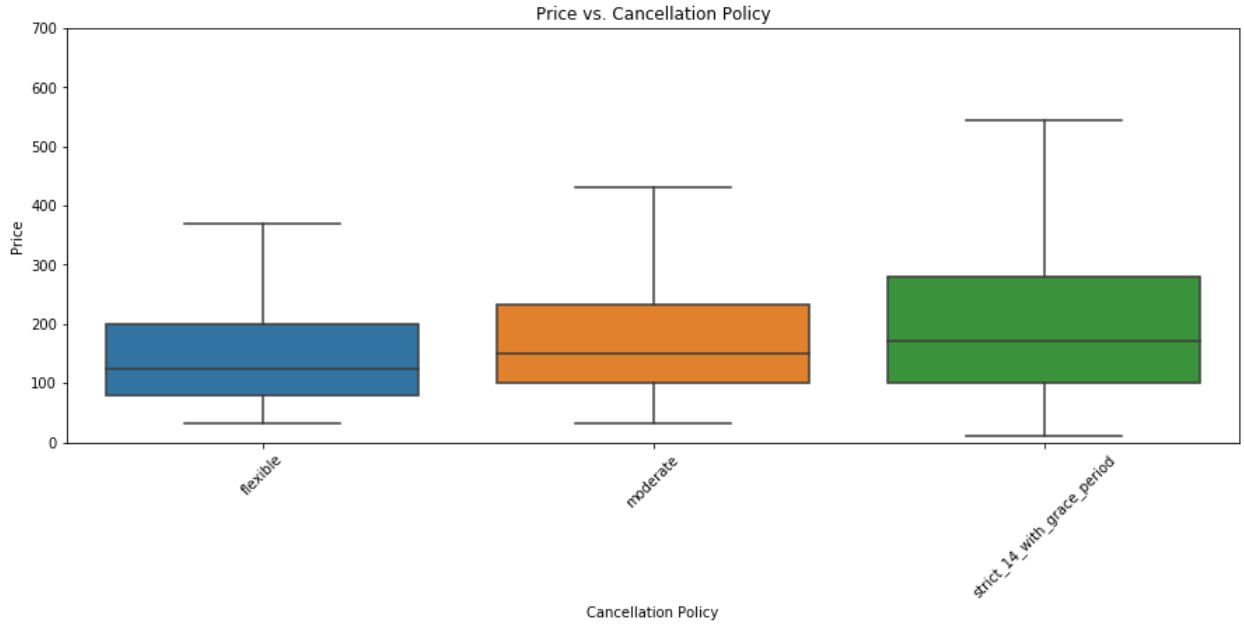


The size of room type appears to be highly correlated with price. Price increases as more privacy is given based on room type space.

Property Type

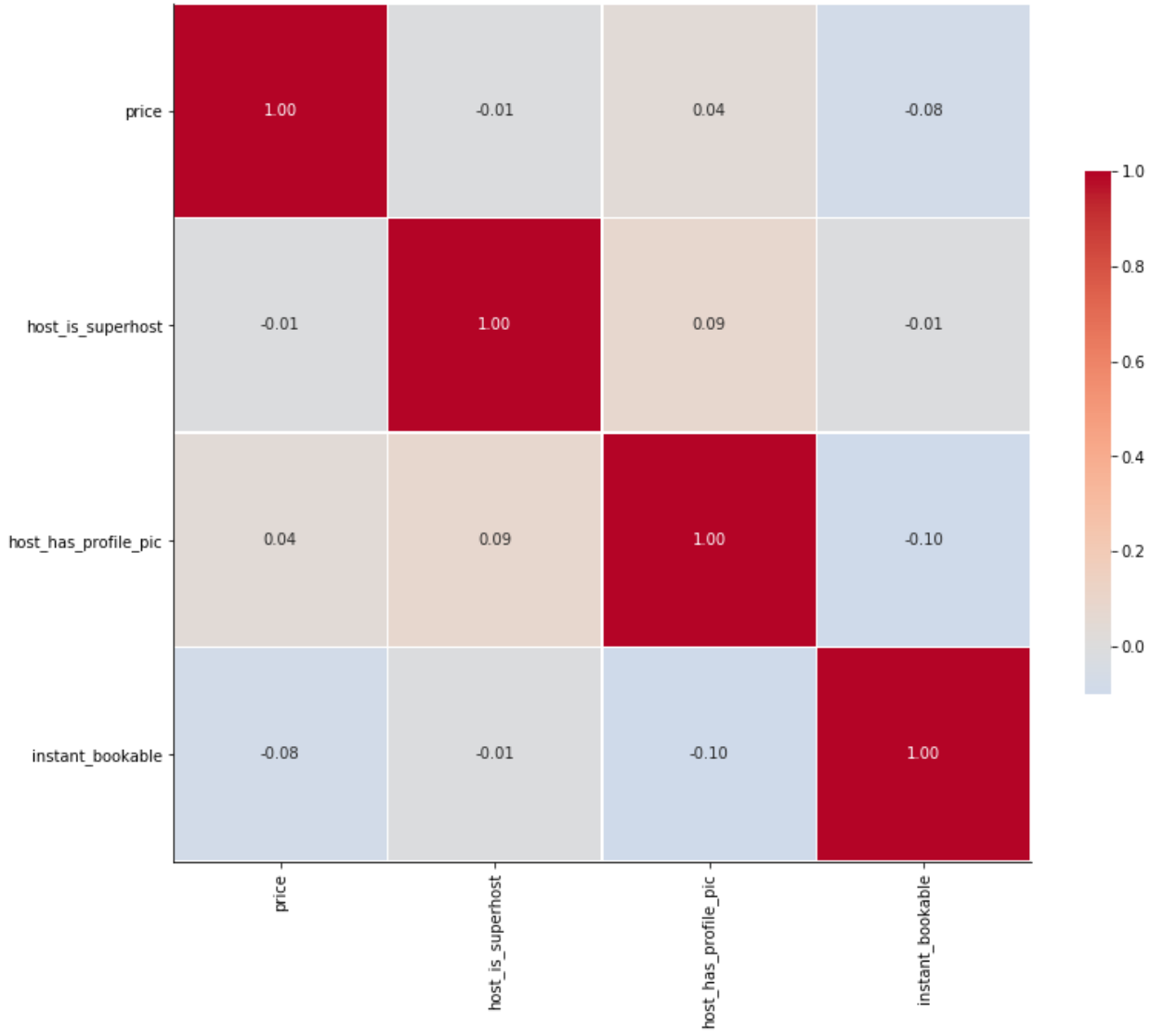


Cancellation Policy

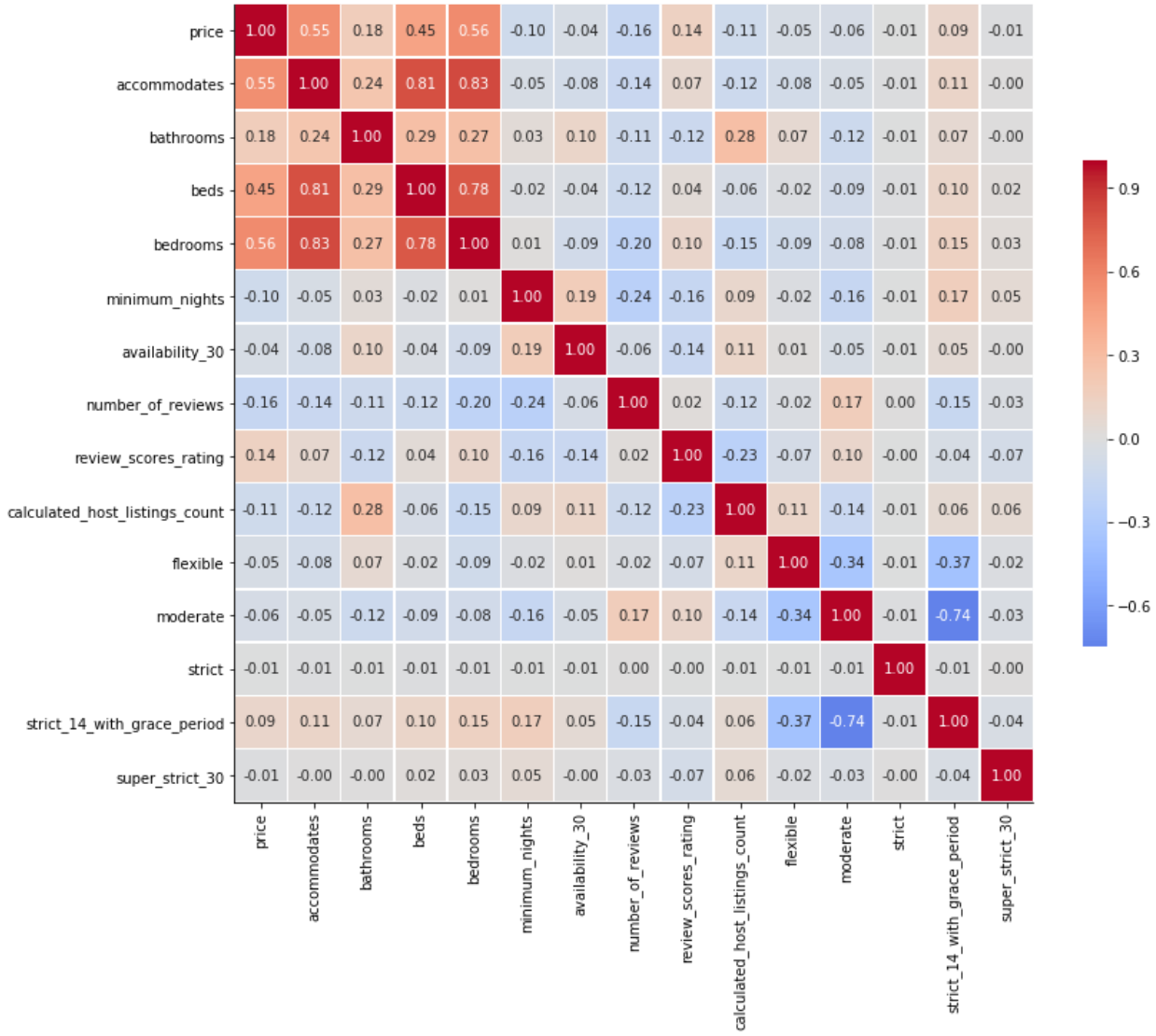


There appears to be a direct correlation between the price and level of cancellation policy, with prices increasing with a stricter policy. There is also a steadily increasing IQR as the cancellation policy becomes more strict.

Correlation Matrix between Price and Features:

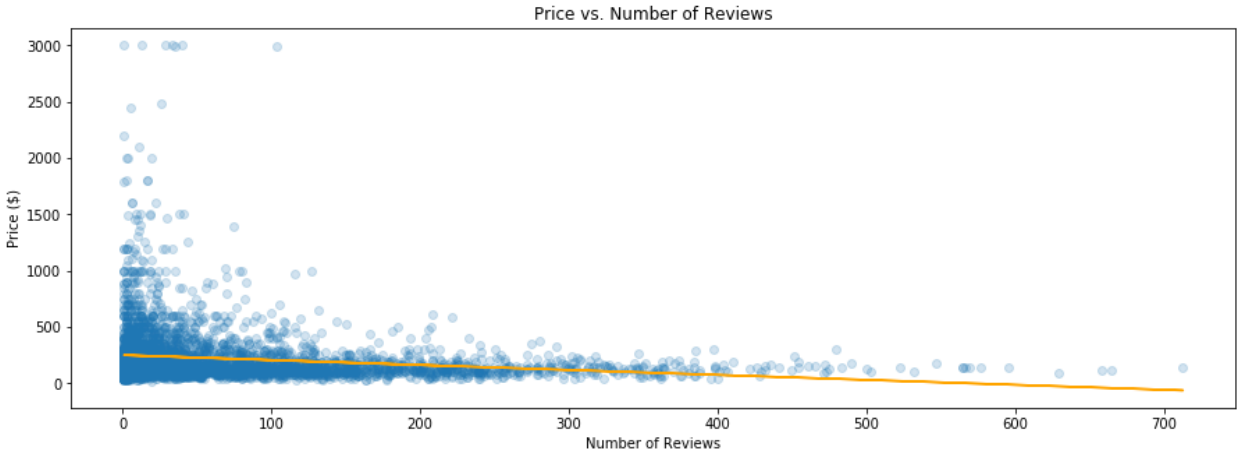


The features in the correlation matrix above do not show a high correlation between each other.

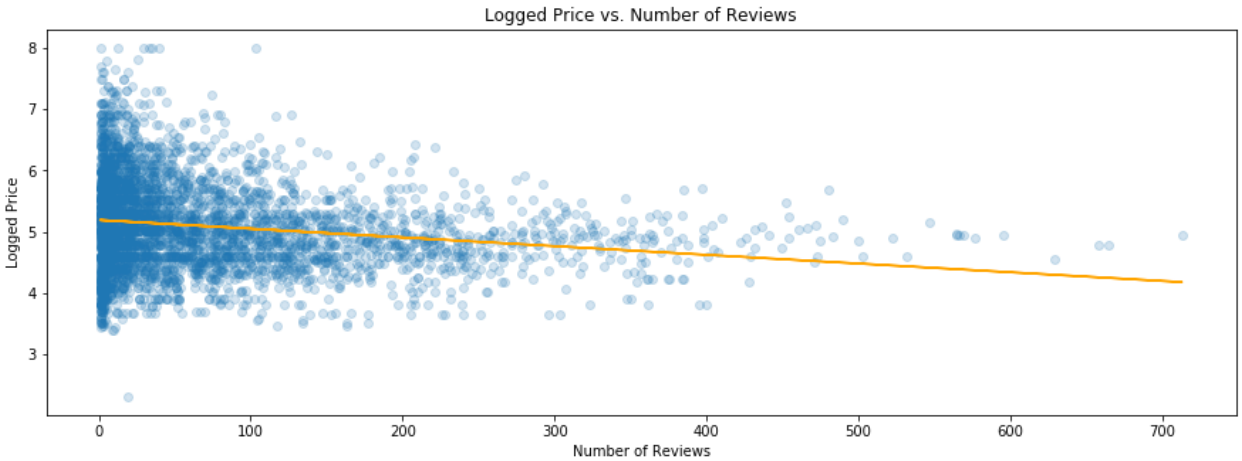


Accommodates, bathrooms, beds and bedrooms are all highly correlated with each other. The cancellation policies (flexible, moderate, strict, strict_14_with_grace_period and super_strict_30) are categorical data (hence correlation isn't technically correct), but included in the heatmap to get an overall sense of the differences between features and price.

Number of Reviews



The scatter plot above shows the number of reviews are highest for listings lowest in price.



Here is the same data depicted with logged pricing.

Conclusion

All of the features discussed above will be explored more using machine learning to predict the prices of Airbnb rentals.