

Capstone 1: Milestone Report

Problem

With over 4 million listings in 191 different countries, Airbnb has become an amazing investment opportunity for people all around the world. While it is a remarkable achievement, the supply side continues to grow, which can make being an Airbnb host difficult. What can a host do to allow their listing to stand out and ultimately increase their revenue? This analysis aims to understand the most important factors for predicting the prices and ratings for every listing in San Francisco.

Client

Airbnb would benefit from using this analysis by being able to advise their hosts on the most important factors that lead to higher ratings and higher prices. By understanding the most important factors for price, a host could optimize their listing to increase their revenue. By understanding the factors for ratings, Airbnb's host would be able to create a better experience for Airbnb customers, who will be more willing to use the service again, while fostering positive communication around booking with Airbnb.

The Data - Inside AirBnB

The data was obtained from independent, non-commercial organization that utilizes public information compiled from the Airbnb web-site and is available to download on <http://insideairbnb.com/get-the-data.html>. The original dataset is a static 365-day snapshot from October 2, 2017 and includes 96 features on 8933 listings in San Francisco. These 96 features include 6 date/time features, 33 numerical features, 32 categorical features, 21 textual features, and 4 id keys.

Process

We first started out by exploring each column to remove the unneeded columns. Information we got rid of included:

- **Meta Information:**
 - Unused keys:(last_scraped, scrape_id)
- **URLs:**
 - Links to pictures or the listings that will not be used in the analysis: listing_url, thumbnail_url, medium_url, picture_url, xl_picture_url, host_url, host_thumbnail_url, host_picture_url
- **Redundant Information:**
 - Since our analysis takes places in San Francisco, CA the following features all have the same value for each listings: city, state, market, smart_location, country_code, country, jurisdiction_names
 - In our analysis we use the features "zipcode" and "neighbourhood_cleansed" for understanding the impact location has on *Price* and *Ratings*. We don't need the following features, since the previously mentioned two sufficiently capture the information needed: latitude, longitude, neighborhood
 - is_location_exact refers to if the longitude and latitude are exact or within a mile radius of the listings true location. Since we are not using latitude and longitude in our analysis we can drop this one as well.
 - calculated_host_listings_count, host_total_listings_count, and host_listings_count all have the same values, so we chose to just keep calculated_host_listings_count for our analysis
- **Irrelevant Information About Host:**
 - These are features are text columns that most likely wouldn't have a strong impact on the model if we were to attempt to transform them: license, host_name
 - I doubt that a customer bases their ratings or how much they are willing to pay off of a host's location or their verifications, so we can drop: host_location, host_neighbourhood, and host_verifications
- **Calendar Information:**
 - The calendar is not very accurate, because it cannot tell the difference between a listing where a host has blacked out their own dates because they aren't renting or a listing with its dates blacked out because it has already been booked by customers. A lot of host will blackout dates until the time is closer, so the availability metrics don't necessarily represent how many days they have left to rent: availability_30, availability_60, availability_90, availability_365
- or Any column with 1 or less unique values

From there we had 56 columns remaining. These included our Response Variables (Price & Ratings), 24 Numerical columns, 11 Text columns, 16 categorical columns, and 4 date/time columns.

Numerical: Of the 24 numerical columns (26 with the response), all were loaded in as a string, and 7 of those were loaded in with special characters (Currency or Percentage). We used a loop to replace these special characters with an empty string, and then converted all 26 to a float.

- Now that the data has been transformed into a numerical type, we can find and remove any listings that are no longer active. The reason to get rid of these listing is that our goal is to find ways to increase the ratings and price of listings and if a listing is inactive then they cannot benefit from our analysis anyway. They are not a good representation of the population we are trying to help, so they may negatively affect our predictive performance. Using the column '*number_of_reviews*' we removed any rows that had less than 1 more review.
- Next we looked at the Null Values for each numerical column, so that we could either impute the missing values, transform the feature into something else, or know if we needed to remove it:
- **Square Feet:** The square feet column has over 97% of its rows missing. Instead of just removing the column, we transformed it to a column that states whether a listings has included square feet.
- **Weekly & Monthly Price:** These columns have a high amount of missing values. These columns represent a separate price (A discount) if you are to book by the over a Week / Month. We transformed these to binary values that say if there is a discount or not for a longer stay
- **Bedrooms / Bathrooms / Beds:** Since the column accommodates had no missing value and is highly correlated with all these columns, we imputed the missing values based on the median value of the respective columns that matched with the value for accommodates.
- **Review Scores:** review_scores_rating (One of our response variables) had no missing value, so we can use that to impute the missing values for the other review scores, since they are all highly correlated with ratings
- **Security Deposit / Cleaning Fee:** Security Deposit is most correlated with bedrooms(.32) and price(.32) and Cleaning Fee is most correlated with accommodates(.56), bedrooms(.59), beds(.56), and price(.51). We imputed cleaning fee first with median value grouped by bedrooms, so that we could impute the security deposit with the cleaning fee.

- **Host Response Rate:** We removed this column, because there was extremely small variation in the listings. 95% of listings had 100% response rate.

Date/Time Columns: We turned all four columns to the numerical type, so we can add them to our model. The data was scrapped on October 2, 2017, so all dates are relative to this date. We used the `datetime.datetime` and `datetime.timedelta` package to convert each column into days since the time scrapped.

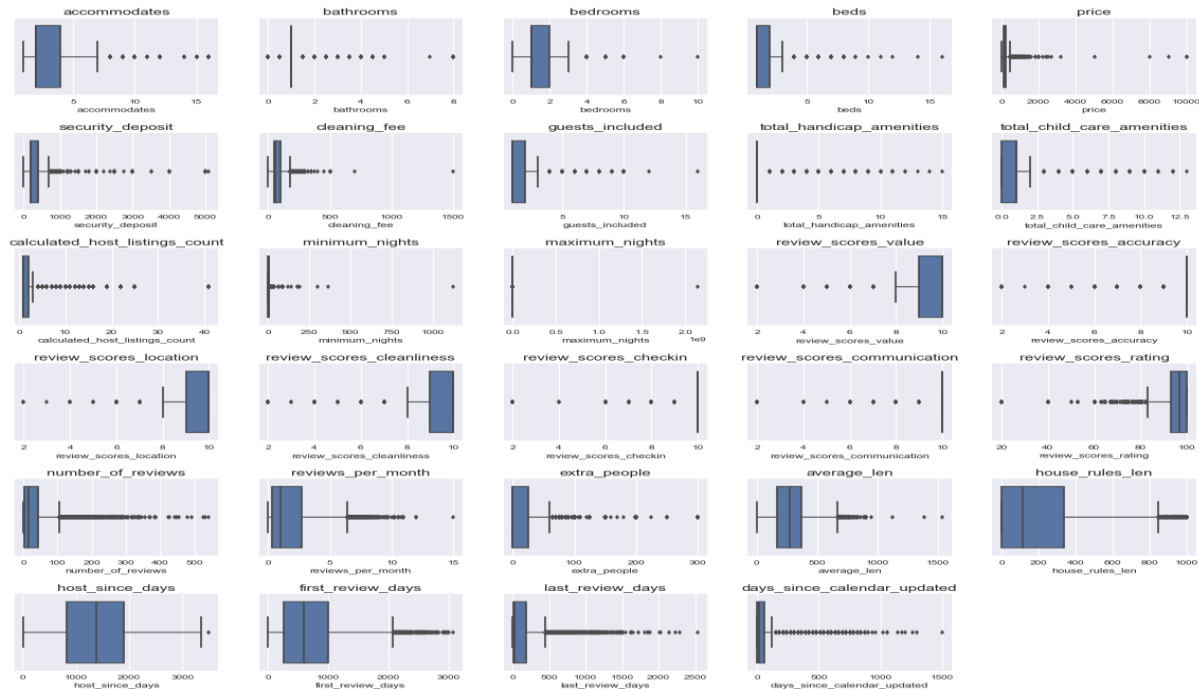
Text Columns: Since we can't impute for missing values in text columns, we just filled missing values with blanks ". Then we calculated the average of the length of all the text columns, because there may be a relationship between the effort put into the listing summary, description, etc. with the ratings or price. We kept the length of the column `house_rules` and deleted the rest. The reason we kept this column is because if the list of rules is too long, then that may have a negative relationship with reviews. No one wants to have to worry the whole time about a bunch of rules while on vacation.

Categorical Columns: Our categorical data contains Binary columns, Multi-level columns, and a list within each row. We need to turn our categorical data from text into the categorical or boolean type.

- **Amenities:** The amenities column contains a list of all the amenities that a listing has. We then created dummy variables without removing the first column, since not having all the others doesn't imply you would have the removed amenity. In total there were 124 different amenities.
- **Binary Categorical Columns:** These columns were all *True/False* columns but loaded in as 't' or 'f'. We converted the text to 'True' or 'False', and then transformed it into boolean type.
- **Multi-Level Categorical Columns:** We changed these to categorical type and then checked for null values. Zip code and Host_response_time were the only ones to contain null values:
- **Host Response Time:** There were no easy ways to impute this column, so we used a naive method of backfilling the missing values.
- **Zipcode:** The neighbourhood_cleaned column had no missing values, so we created a dictionary of each neighborhood's most frequent value and then imputed the missing values by matching the dictionary neighborhood and its most frequent zipcode

Initial Findings

Numerical Features



We first constructed boxplots for all the numerical features, so that we can examine the outliers and skewness of each variable. Some things notice:

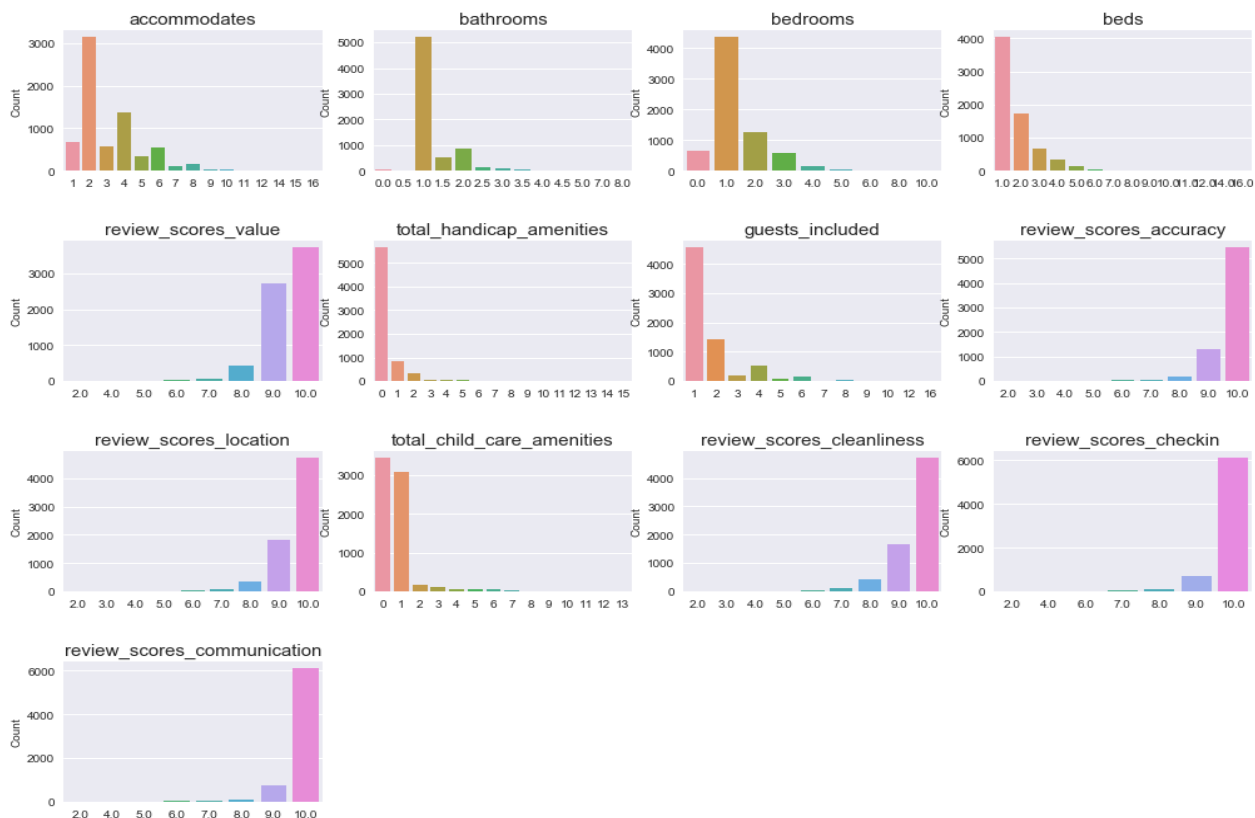
- All values are skewed, and most of them are skewed heavily right. This tell us that the majority of the listings in San Francisco are smaller listings, with some extreme unusual cases (such as 10 bedrooms).
 - Minimum nights and maximum nights have some extreme points, that appear to be errors, or the host chosen an arbitrary high number. (Maximum nights is over 2 billion). We removed these as the will only be harmful to our prediction model.
 - Price is also heavily skewed to the right, and since it is one of our response variables we will examine this more closely later.
- The features that are skewed left are the review scores. Most of them have an extremely small IQR as well. This is a common occurrence in ratings, where most

people give a perfect score unless there was some unusual negative event that occurred.

- Review scores for *cleanliness*, *value*, and *location* have the largest IQR range, which may indicate that people put more thought into these reviews and they may be stronger predictors.

We then separate the discrete numerical values and the continuous numerical values. For the *Discrete* numerical values, we created count plots to see how the values were distributed.

- It's clear that the columns beds, bathrooms, accommodates, and bedrooms have very extreme values. 99% of listings fall within half of the max beds, bathrooms, and bedrooms.



For the *Continuous* variables we constructed histograms to examine their distribution. We really just see how strongly everything is skewed. We should consider dropping more outliers, since they are so extreme.

Next we created a correlation heatmap of all the variables.

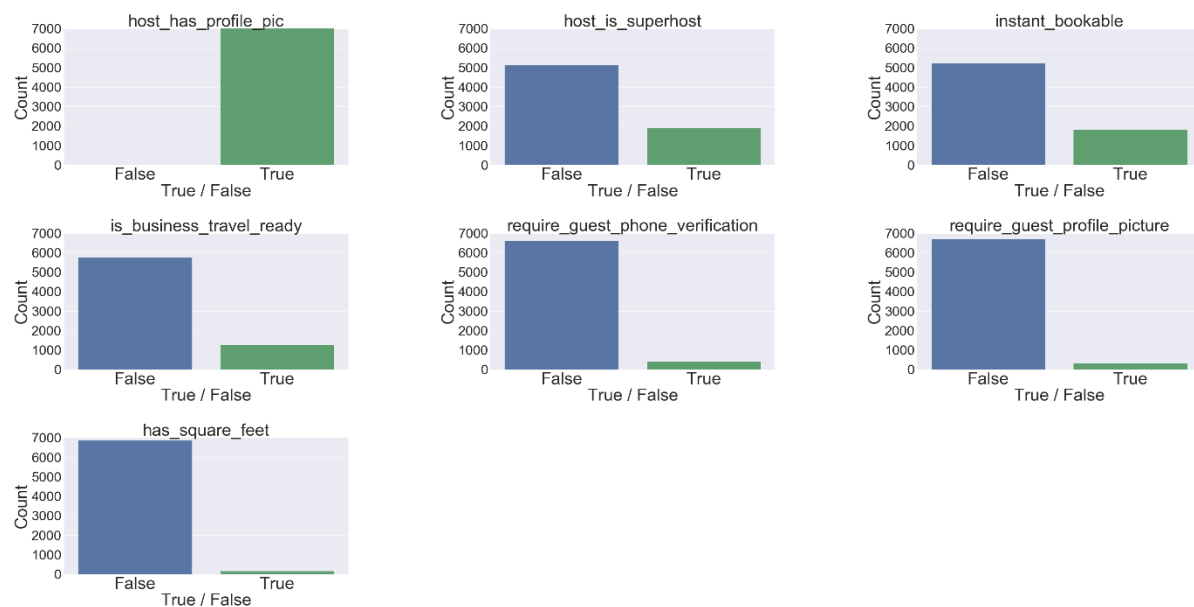
- We can see that Accommodates, Beds, Bathrooms, Cleaning Fee, and Guest included are all highly correlated with each other. May be useful to drop a few of these to improve our model's prediction.
- The review scores are all highly correlated with each other as well. We should consider dropping a few of these too. I think there is reason to keep them in as well. These may bring insights into what is the most important review a host can focus on to improve their overall rating.
- Days since first review and number of reviews are correlated, which makes sense since the longer a host has been renting out the more reviews they would have. I assume that days since host isn't correlated because there are a lot of inactive listings from host that started and stopped Airbnb a long time ago. This could be because they chose not to comply with San Francisco's Airbnb regulations.

Categorical Features

We created bar charts for all 110 amenities. Only 56 of the 110 amenities applied to even 5% of the listings. These should be considered outliers it may be useful to drop these.

We then may a similar plot with all the True / False categories in our data. Has_square_feet, host_has_profile_pic, require_guest_profile_picture, and require_guest_phone_verification are very skewed and have a heavy majority of listings either True or False.

True / False Count Plot



We then explore the Multi-Level Categorical features with count plots. It was clear the majority of listings were a house, apartments, or condominium with a real bed. The neighborhoods and zip codes were very well dispersed.

Price

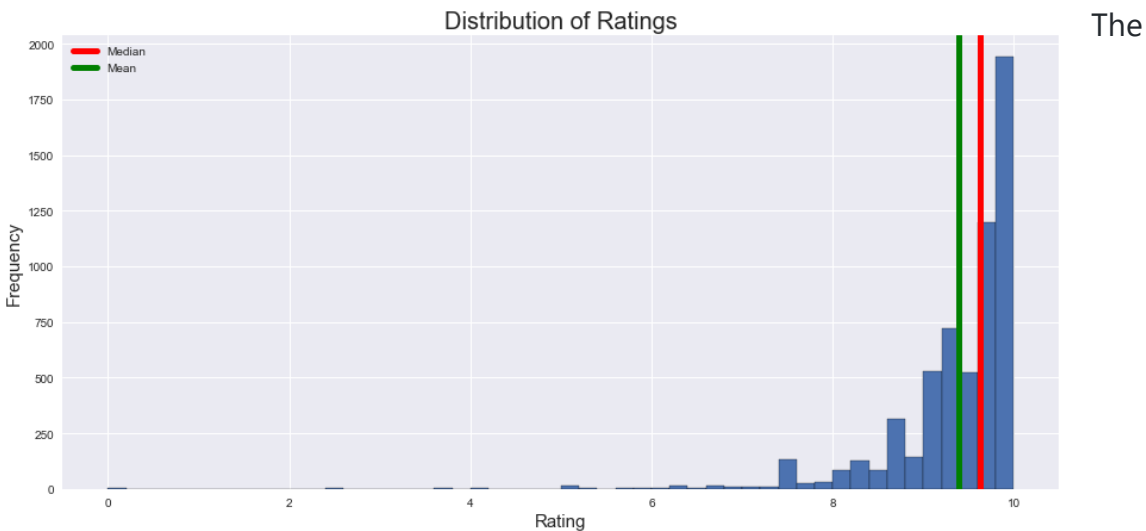


From there we explore **Price**. We constructed a distribution plot that showed us the center of the distribution falls at \$153 and the mean is at \$210 with a right skew with a long tail. The Inter-Quartile Range for price was \$105 and \$250, yet the prices extend to around \$10,000. The IQR rule for outliers shows us that there are 347 outliers for price and some of those outliers have a large influence on the entire dataset and need to be addressed. Over half of the listings in San Francisco are 1 bedroom, 1 bathroom, 1 bed, and accommodates 2-3 people, so it isn't surprising why the prices are right skewed. Some values seem to be outliers and should be removed from the dataset, so that they don't negatively affect the prediction models.

The characteristics of higher priced listings:

- The more a listings can accommodate / bedrooms / bathrooms / beds
- There are 37 different neighborhoods with a range in medians of \$80 to \$217. The Marina has the highest median price and seems to one of the most expensive places to rent an Airbnb
- Business Travel Ready listings are statistically significantly higher in price
- Non-Instant Bookable listings are statistically significantly higher in price

Ratings



distribution and boxplot for listings ratings showed that there was a strong grouping above a rating of 9 or higher and that the data was heavily left skewed. People tend to give high ratings as long as their experience was positive. Of the 5000+ listings only 296 were considered outliers on the lower boundary IQR outlier rule.

Although the rating is one of the variables we are trying to predict, there are 7 total review scores. These are more specific ratings such as communication, location, etc. Looking at the correlation table of ratings, we see that the only significant relationship ratings has are the other reviews. Although they are correlated with each other I think it may be important to keep these as these could be a useful way to understand what customers appreciate more compared to the other review scores.

The characteristics of higher rated listings:

- The more a listings can accommodate / bedrooms
- 8 of the top 10 rated property types have less than 30 listings and 6 of those have less than 10 total listings. Also, the 132 listings that don't supply a couch, air mattress, pull-out sofa, or futon are higher rated on average than those that supply a bed. The unique property types and bed types have higher ratings.
- Having an Entire House / Apartment is higher rated than having a private room or having to share a room.
- The Business Travel Ready listings are higher rated
- The listings that are not Instant Bookable are higher rated