**Final Class Project**

*Open Data Hacking:*

Predicting San Francisco Airbnb Pricing

Brittany Katsch

UCLA Extension, COM SCI-X 450.1

Introduction to Data Science

Spring 2020

Table of Contents

[The Data 3](#_Toc43470796)

[Definition of the variables 3](#_Toc43470797)

[The Project 5](#_Toc43470798)

[Data Munging 6](#_Toc43470799)

[Exploratory Data Analysis 8](#_Toc43470800)

[Multiple Regression 16](#_Toc43470801)

[Training the Algorithm 16](#_Toc43470802)

[Determine Model Accuracy 17](#_Toc43470803)

[Testing the Algorithm 19](#_Toc43470804)

[Determine Model Performance 19](#_Toc43470805)

[Conclusions 20](#_Toc43470806)

# The Data

The data used in this project was collected from Kaggle, a website with downloadable open datasets an projects. The data was downloaded from <https://www.kaggle.com/jeploretizo/san-francisco-airbnb-listings>.

This dataset was scraped from the Airbnb website (Airbnb.com) on October 14, 2019 and includes 8,111 unique observations of 106 variables. Each observation is a separate Airbnb listings in the San Francisco, CA geographical area. The variables include information about the host (are they a “super host?”, Do they have an image available? Can they be contacted by phone or email?, etc.), about the listing itself (Is it a room for rent or entire house? How many people does it accommodate? What types of beds are in the rental?, etc.), and geographical location (Latitude and Longitude information, zip codes, neighborhoods, etc.).

With 106 variables, there are some that will be unnecessary to this project (for instance, listing URL is redundant has no place in the predictive model). There will be 22 variables that are deemed important to the project. They are:

## Definition of the variables

"host\_is\_superhost" – A “superhost” has multiple exemplary reviews & a badge on their profile.

"neighbourhood\_cleansed" - What neighborhood is the listing located in?

"zipcode" – What zip code is the listing located in?

"latitude" – What is the latitude of the listing?

"longitude" - What is the latitude of the listing?

"property\_type" - What is the property? (ex. House, Apartment, Condominium)

"room\_type" – What is the room type? (ex. Full House, Private Room, Shared room)

"accommodates" – How many people does the property accommodate?

"bathrooms" – How many bathrooms does the listing have?

"bedrooms" – How many bedrooms does the listing have?

"beds" – How many beds does the listings have?

"bed\_type" - What type of beds does the listing have?

"price" - What is the price per night of the listing?

"security\_deposit" – How much is the security deposit per stay?

"cleaning\_fee" - How much is the cleaning fee per stay?

"guests\_included" - How many guests are included in the price?

"extra\_people" – What is the charge for additional guests per night?

"minimum\_nights" - What is the minimum number of nights that must be booked?

"maximum\_nights" - What is the minimum number of nights that must be booked?

"number\_of\_reviews" - How many reviews by previous renters does this listing have?

"instant\_bookable" – Instant bookable’s do not require approval from the host before booking.

"cancellation\_policy" – How strict is the cancellation policy?

A screenshot of a newspaper

Description automatically generated

A screenshot of text

Description automatically generated

A screenshot of text

Description automatically generated

# The Project

Airbnb has quickly become a popular tool for booking travel accommodations. Airbnb allows “hosts” to advertise their accommodations on the website where vacationers can browse and book based on numerous factors such as location, number of bedrooms, and price, just to name a few. Hosts are able to list accommodations from single rooms-for-rent, to full apartments and houses, to non-traditional dwellings such as RV’s and Teepeees. Though not common, hotel rooms can also be found on Airbnb’s website.

San Francisco has always been a popular destination for both tourists and business travel alike. In the dataset used in this project, prices of accommodations range low (<$100 per night) to extremely high (>$2,000 per night). The goal of this project is to determine whether price per night of an Airbnb listing can be predicted. A successful model could potentially help tourists choose their accommodations, and help hosts determine a fair, reasonable price to list their properties. Tourists would be able to know if they’re paying a fair value for a property based on a series of factors, and hosts would not have to worry about charging too little or too much for their rental.

With this project I hope to prove that given a specific set of details about a vacation rental in San Francisco, we can predict the Airbnb listing price per night.

# Data Munging

The first step in the data cleaning process was to get rid of the variables we would not be needing. The dataset was paired down to just 22 variables that have the potential to be significant in predicting rental price.

Some variables I removed from the dataset included "listings\_url" "scrape\_id" "last\_scraped" "name" "summary". I removed variables based on what I considered to be irrelevant information. For example, the listings\_url (web url of the property) will not be a valid predictor of the listing’s price on the website. I also removed variables where all information was the same, or missing as it was with “experiences\_offered”; there was no data in this column for any of the observations.

At this point I narrowed the dataset to 32 variables (31 relevant predictors and 1 price column), I took an in-depth look at the individual variables using summary(). From summary I found some immediate issues.

Based on the “city” variable I saw that some non-San Francisco data had snuck into the data set, so I removed all observations that were not from San Francisco. I fixed NAs in "host\_is\_superhost", "security\_deposit", and "cleaning\_fee". I assumed since there was no data here that hosts was not a superhost, deposit and fees were $0.

The “city” variable had 235 “NA” values. To fix this, I grouped the data by neighborhood and populated the NA’s based on the zipcode for the neighborhood groups. I used the median() function to specify which zip code to use if there were multiple zip codes for one neighborhood.

Monetary variables “price”, “security\_deposit”, “cleaning\_fee”, and “extra\_people” were all stored at characters strings, so I removed the ‘$’ and ‘.00’ from each and coerced to numeric.

I made other minor changes to variables including changing “host\_is\_superhost” to logical TRUE/FALSE, “zipcode” to factor, etc.

I removed any final unnecessary variables (such as ‘city’ which was now “San Francisco” across all observations) and any other variables that had escaped my initial screening process. I now have a cleaned data frame with 8010 observations of 22 variables.

A screenshot of a newspaper

Description automatically generated

# Exploratory Data Analysis

To explore the data, I looked at each variable individually. Starting with Price, I was able to see that there were some obvious outliers; some prices were $0 which is not possible for a rental, and some were $10,000 per night. This is important to note for the future.

A screenshot of a cell phone

Description automatically generated

From the plot above, it looks like anything above about $1,500 per night is an outlier. Looking now at a histogram of price, the data is too spread to see with the outliers included. So we will view a histogram of price less than $2,000 per night.

A screenshot of a cell phone

Description automatically generated

Again, the majority of the prices are below $1,500 per night. Looking at summary statistics for price, 25% of prices are below $100 per night while 75% of prices are below $240 per night.

It may be helpful to look at price by room type:

A screenshot of a cell phone

Description automatically generated

Seeing the mass of green at the bottom of the plot indicates that perhaps Hotel Rooms are the least expensive options. The mid and higher range of pricing is dominated by Entire Home/Apartment rentals.

Next, we’ll explore average price by neighborhood: Presidio heights seems to be the most expensive, while Treasure Island is the least expensive neighborhood to stay in.

A close up of text on a white background

Description automatically generated A close up of text on a white background

Description automatically generated

Next, let’s take a deeper look into neighborhood.

A close up of a piece of paper

Description automatically generated

This bar plot shows that the majority of listings are in the Western Addition, South of Market, Mission, and Downtown neighborhoods. There are very few listings in Presidio, Treasure Island, and Golden Gate Park. This sligns with Zip Code data below.

Looking at room type, there are 4 levels: Shared, Private, Hotel, and Entire Home/Apartment.

A picture containing microwave, drawing

Description automatically generated

The vast majority of the listings in the data set are Entire Home/Apartment rentals or Private rooms. There are fewer listings for shared rooms and private rooms.

Next, looking at property type:

A screenshot of a computer

Description automatically generated

There are 26 different types of listings. Houses, Apartments, Guest Suites and Condominiums are the most listed while there are very few listings with in-law, huts, dome houses, and camper/RV property types, which is important to note.

Let’s see the distribution of Property type by room type:

A screenshot of a computer

Description automatically generated

Some observations here: “House” listing types are nearly split by the Entire home and Private Rooms. Apartment listings mostly rent the entire apartment while some are private rooms and a smaller few are shared rooms.

Digging deeper into amenities of the listings, summary statistics of bedrooms and bathrooms show outliers with.a max of 14 for each. The 3rd quartile for bedrooms is only 2 while the third quartile for bathrooms is 1.5. 14 as the max is clearly an outlier for each of these. Upon deeper inspection, these both are referring to observation 2386, listed as a Boutique Hotel, which explains why there are so many bedrooms and bathrooms. This may be important to note. We can see this visually as well:

A screenshot of a cell phone

Description automatically generated

Lastly, we’ll look at cancellation policy, a factor variable with 6 factors.

A screenshot of a cell phone

Description automatically generated

The majority of listings have a strict, 2 week cancellation policy. There does not appear to be an apparent correlation between the policy and the property type.

# Multiple Regression

Since there are 22 variables in this dataset that may (or may not) be relevant to predicting price per night of Airbnb listings in San Francisco, the best supervised machine learning model to use is Multiple Regression since it will allow me to train the model with more than one input. Ultimately, I trained and retrained the model 6 times to refine the model and get the final predictive equation.

## Training the Algorithm

For the first model (lm1), I created a training data set with a randomly chosen 70% of the observations from the dataframe, and a testing set with the remaing 30%. For this model, I used all 21 predictors, and all 8010 observations. There were significant p-values of the f-statistics in zipcode, property\_type, room\_type, accommodates, bedrooms, beds, cleaning\_fee, guests\_included, extra\_people, and number of reviews. Adjusted R2 was very low (0.1395) and Residual Standard Error very high (370.9).

For the second model (lm1B) I decided to include only the significant factors. This model is no good either; R2 increased slightly to 0.1442 meaning only 14.4% of the output is explainable by the predictors. The residual standard error is also still very high at 371.9 The diagnostic plots were also not very good.

For the third model (lm2), I thought back to the EDA process where we found outliers, particularly those listings that were $0 per night, or the few that were greater than $1500 per night. I decided to remove those For this model, I subsetted a new data set “final\_listings\_no\_ouliers” that included only observations where Price per night is >0 and <1500. With this new outlier-less dataframe, I created another randomly chosen training and testing data set with 70% and 30% of the observations, respectively. I trained a model using all 21 predictors. R2 increased to 0.468 by removing the outliers, and residual standard error dropped to 127.6. Significant predictors at the 0.05 level are all but longitude, bathrooms, bed\_type, cleaning fee, and instant bookable

For the fourth model, I dropped the insignificant predictors and retrained the model (lm2B). Residual Standard Error increased to 130.2 and R2 decreased to 0.4418. When looking at the diagnostic plots, this is still not a great fit. Residual vs. Fitted diagnostic plot doesn't show a large visible nonlinear pattern, but the spread of the residuals is more above the line than below; the data is not equally spread. Normal Q-Q plot still shows vast deviance from the line beyond the 2nd theoretical quantile. Scale location does not have a linear line: residuals are not spread equally. In the Residuals vs. Leverage plot there are 2 concerning observations, #1901 and #571

For the fifth model, I thought back to the EDA process. The price data was highly skewed to the right; 75% of the price data was $240 or below. For this model (lm3) I tried using Base R’s log() function to logarithmically transform the “price” output to a normal distribution. I used all 21 predictors for this model. There was significance in superhost, neighborhood\_cleansed, latitude, property type, room type, accommodates, bathrooms, bedrooms, beds, guests\_included, extra people, max and min nights, reviews, instant\_bookable, and cancellation policy. Standard error decreased to 0.4275 and multiple R2 increased to 0.6118.

For the last training of the model (lm3A) I removed the insignificant predictors found in model (lm3). Standard error is 0.4276 and R2 is 0.614, p-value of the f statistic for the model is < 2.2e-16 which is highly significant.

### Determine Model Accuracy

To determine model accuracy, I used RSE, Multiple R2 and the F-Statistic of the model.

A picture containing holding

Description automatically generated

I also plotted the diagnostics for this model (lm3A) to see how accurate it is for the testing set of data.

A screenshot of a cell phone

Description automatically generated

Residuals vs. Fitted plots shows a linear relationship with a linear red line, and residuals plotted fairly equally along the line except for a few more spread points toward the top left and lower right.

A screenshot of a cell phone

Description automatically generated

Normal Q-Q looks much better than previous diagnostics for other models, though there is still some deviance from the line at the lower left and upper right.

A close up of a map

Description automatically generated

Scale location plot looks good, save for a few outliers 571 and 5169 Other than those observations, the residuals seem to be equally distributed along the visibly linear line.

A close up of a map

Description automatically generated

The residuals vs leverage plot shows 2 influential outlier cases out of Cooks distance, 571 (which was also an outlier in the scale location plot) and 1901.

## Testing the Algorithm

Using the model (lm3A) I, used the testing set data to get predicted outputs of price so I can compare them to the actual price values for that data set and see how well this model performs as a predictor of Airbnb prices per night in San Francisco. I made sure to use the test data set that did not have any outliers.

Upon analyzing the correlation between the model and the actual prices, I got a correlation of 0.617 or 61.7%.

### Determine Model Performance

To determine the model performance, I found the RMSE using the model predicted (fitted) output values, and the actual output values from the test set. RMSE is 263.58, which is still a high value, meaning the model does have substantial errors.

I also computer R2 for the test set model, which will tell how much of the output (price) can be predicted by the predictors. R2 is 0.4042, or 40.42%, which means that 40% of price in the test set is determined accurately by model lm3A. and about 60% is explained by other factors.

# Conclusions

Using data collected from the Airbnb listings, namely whether or not the host is a “superhost” on the website, the neighborhood the listing is in, the latitude of the listings, the property type, the room type, how many people the listings can accommodate, how many bathrooms and bedrooms the property has, how many guests are included in the price, the charge for extra people, the minimum and maximum number of nights, the number of reviews, whether or not the property is instantly bookable, and the cancellation policy, we are able to determine the price per night of an Airbnb listing per night about 40% of the time.

A close up of a map

Description automatically generated

Plotting the Fitted values vs. Actual price values above (training data set is on the left and testing data set is on the right) shows visually how far off some of the predictions are. A perfect model would show all datapoints on the red line or very close to it. We can see that is not the case.

The predictive model is not perfect, which leads me to believe that there are other factors not included in the model that have impact on the price of the rental.