Project 1 Writeup

Jenna Tan

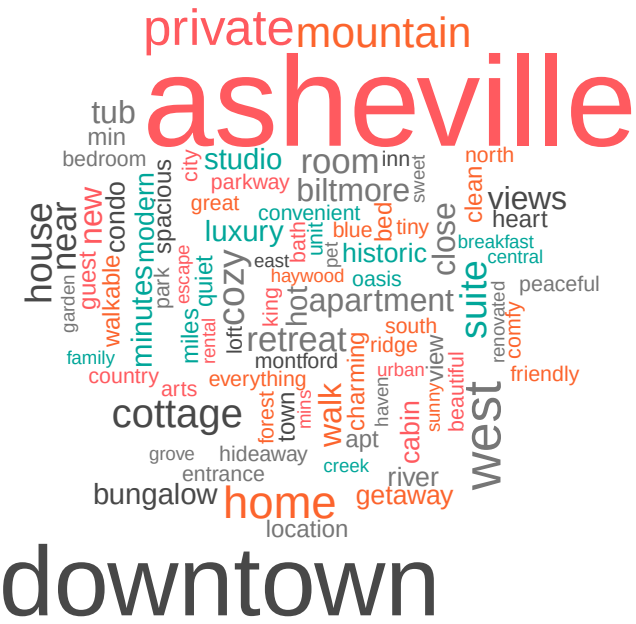
Introduction

Booking travel accomodations is one of the most important part of planning a vacation and oftentimes, it is the most expensive part of your travel budget.

In this analysis, I break down a dataset containing listings of Airbnb properties in Asheville, NC to determine what characteristics predict the price of a listing. I will then create a predictive model to determine if I'm being overcharged for the listing I chose to use for my honeymoon.

Dataset Description

This dataset was sourced from InsideAirbnb.com and was last updated on June 25, 2020. It contains 2,407 different listings with 106 points of data on each property. Examples of included properties are superhost status, price, bedrooms, zipcode, and rating. Below, I've included a word cloud of the name of Airbnb listings in Asheville to illustrate the common characteristics advertised in listings.



We can see that popular words include Asheville (obviously), downtown, private, mountain, home and even Biltmore. More obscure words include arts, sunny, haven and haywood.

Exploration

I first explore my outcome variable, price. Price is represented in multiple ways in the dataset including nightly price, weekly price and monthly price. I decided to look at the price for a one-night stay, including

the cleaning fee. Thus, the usage of "price" in this analysis represents the sum of rent for one night and the cleaning fee. I chose to exclude taxes and other fees since it would vary from property to property.

Since the goal of my analysis is to create a model that can be used to predict prices as I browse Airbnb, I am limited to predictors that can be determined based on the listing page alone. Predictors that are available in the dataset and the listing page is the property type, the room type, the number of bathrooms, the number of bedrooms, the number of beds, the average score of reviews, and the superhost status of the host.

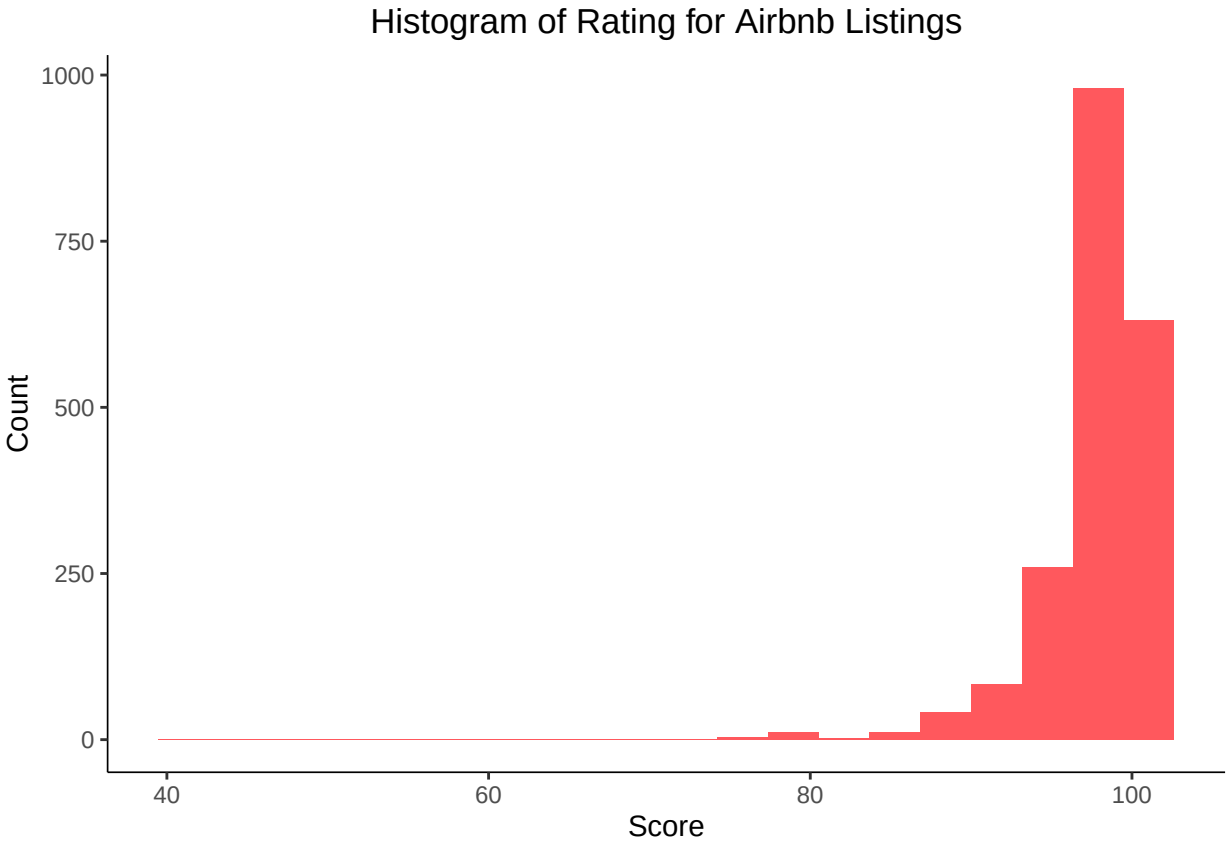


Figure 1: The ratings for properties are skewed to the left and have a very narrow distribution.

With a mean of 97.72 and standard deviation of 3.40, the distribution is very narrow and skewed to the left. This makes it difficult to distinguish listings based on rating alone. Thus, I decided to eliminate rating as a potential predictor.

Figure 2: The figure shows the spread of price by property type. Note that houses have a large number of outliers.

Due to the great variation in price by property type demonstrated in the boxplots, I decided to narrow my analysis to a smaller scope and only examined types of properties that we would potentially book: condominiums, townhomes, houses, and apartments.

Investigation of Outliers in House

I attempt to investigate the reason for the large number of outliers for houses and assess the relationship between price and beds as well as the relationship between price and bathrooms.

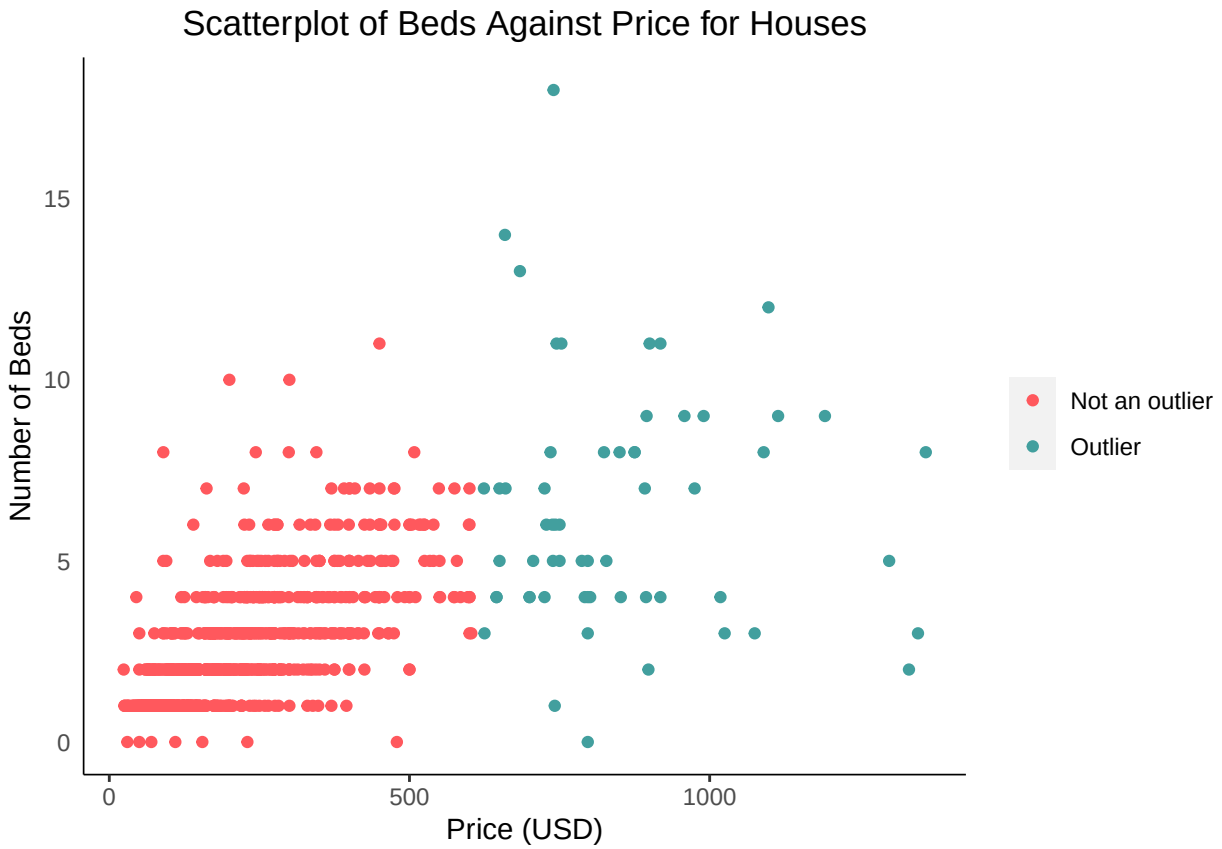


Figure 3: Out of the listings of houses, the outliers include listings with a wide spread of beds, suggesting there isn't a strong relationship between beds and price.

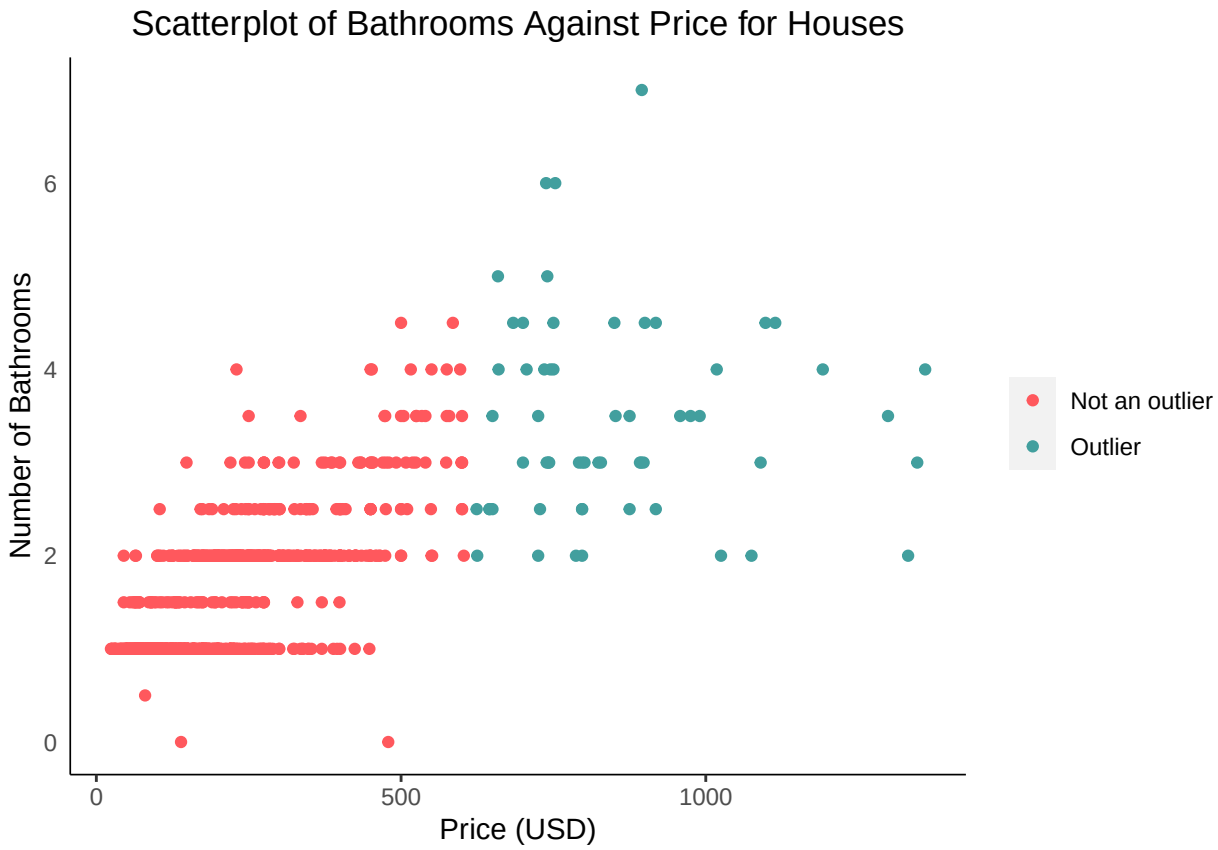


Figure 4: Out of the listings of houses, the outliers include only listings with 2 or more bathrooms, suggesting the number of bathrooms may be associated with price.

Based off of these two scatterplots, it seems that the outliers included listings with 0 to 18 beds and only listings with two or more bathrooms. For this analysis, I've decided to leave in the outliers, but hope that future investigations will take note of these observations.

Analysis Dataset

After limiting the dataset to relevant property types and to the predictors we are interested in, the data now has 1,419 complete cases for us to predict and test on.

We split the data about 50/50 to create testing and training datasets, using a seed of 1107.

Best Subset Selection

With the method of best subset selection, the predictors selected included whether a listing was a condominium, whether the room was a private room, the number of bathrooms, the number of beds, and whether the host was a superhost. Surprisingly, if the host was a superhost, the total price was lower by $27.21 on average. This is the opposite of what I would anticipate.

The model we trained has a testing error of 14052.95.

Forward and Backward Selection

Forward and backward selection produce the same model, including the following as predictors of price: whether the property was a condominium, whether the property was a house, whether the room was private,

the number of bathrooms, the number of bedrooms, the number of beds, and whether the host was a superhost.

The model we trained has a testing error of 13393.39.

Since the error rate is lower for this model compared to the model selected by best subset selection, we shall use this as our predictive model.

Prediction

Using the model trained by forward/backward selection, I predict the price of the listing linked below. The price was predicted to be $66.22 which is much lower than the price which is $175. This suggests that the price of Airbnbs is associated with predictors beyond what I included in my analysis dataset.

Listing used for prediction: https://www.airbnb.com/rooms/34151081

Further Exploration

To further expand on my analysis, I would incorporate the presence of specific amenities in my model training. Due to the format of the dataset which included the presence of all amenities within one variable, I was unable to invest time in deriving appropriate variables. Thus, if given more time or a second attempt at this investigation, it would be interesting to explore if the presence of amenities, such as a full kitchen or a hot tub, would lead to more accurate predictive model for pricing.