# Seattle AirBnB Market Inefficiencies

## 1. Introduction

### a. Problem Statement

What exactly is the right price for an AirBnB home? Each city and neighbourhood is a unique market, with hosts needing to decide their own pricing based on their own market research and gut instinct. Some hosts may be doing very well, with 100% booking rates, driven by below market pricing. Other hosts may be over priced, or taking actions that are not helping them charge higher prices. Analysis around this subject may provide some insight that would be of use to both groups of hosts.

The main question is how would a seller know what the right price is? There are theoretically six different factors that could influence the price of a listing:

1. Size of home
2. Location
3. Host actions, such as cleaning, responsiveness, communication, etc.
4. Quality of listing, as determined by photographs
5. Quality of listing, as determined by the host's written description.
6. Host bias in pricing

Hosts need to use heuristics to determine pricing, such as good pictures are a worthwhile investment, and quick response times are important. But how important are they really?

### b. Background

AirBnB hosts listings across the world, and is a favorite alternative to hotels for uniqueness of experience, cost, convenience and home types. Hosts will rent out their units, from a single room in a home to entire homes, with some hosts acting as mini-hoteliers and managing multiple homes (also known as listings). AirBnB is the listing platform, while hosts are responsible for entering listing information, descriptions and photographs.

Following a stay of one or more nights, renters can leave reviews of the property, rating such items as cleanliness, hosts, location and overall satisfaction. Other renters can see these reviews, as well as the info provided by the host, and decide if they wish to rent the unit.

### c. Goals

The goal of this analysis is to analyze the first three groups of features (size, location, host reviews) and determine how important they are in setting prices.  This will require regression models that focus on $R^2$ and MAE to determine predictive accuracy and feature importance.

This analysis will not look at the last three feature groups (listing quality by photos, quality by description, host bias) due to their much larger complexity and subjectiveness.  Each of these feature groups would be much higher in complexity than the first three.

There are two immediate goals of the model to be created:

1. Can listings with high occupancy rates (greater than 90%) be shown to consistently underprice relative to the market?
2. Are there actions which hosts can take to improve listing prices, based on the importance of feature reviews?

## 2. Datasets

Data about listings have been acquired from the Kaggle website for Seattle, with a snapshot of listings from early 2016. There are two datasets: one for reviews, and one for listings. The listing dataset has around 2700 records and 92 fields.

These fields can broadly be broken down into seven groups:

1. Neighbourhood
2. Listing Description
3. Photograph URLs
4. Data Scraping Info
5. Host Info
6. Home Characteristics
7. Review Scores

Many of the fields overlap or can immediately be dismissed as being irrelevant, such as info on when the data was scraped.

## 3. Data Cleansing

The combined data set, called listings, was investigated for the normal data quality issues.

- Missing data resulted in either fields being removed or reasonable values being entered (such as setting the number of bathrooms to one when no data existed).
- No duplicate records were found
- Outliers were not a problem but high priced properties are clearly unique and so no listings above $400 were included. Furthermore, odd home types (such as a treehouse and a yurt) were removed as they were not conducive to model creation
- Formatting was straightforward, with some fields being changed from object files to integers

Fields that were clearly not needed were removed. This included fields that gave information on the date the data was scraped, or host information, or descriptive text or picture URLs. Also, superfluous fields were removed, such as the two extra neighbourhood fields.

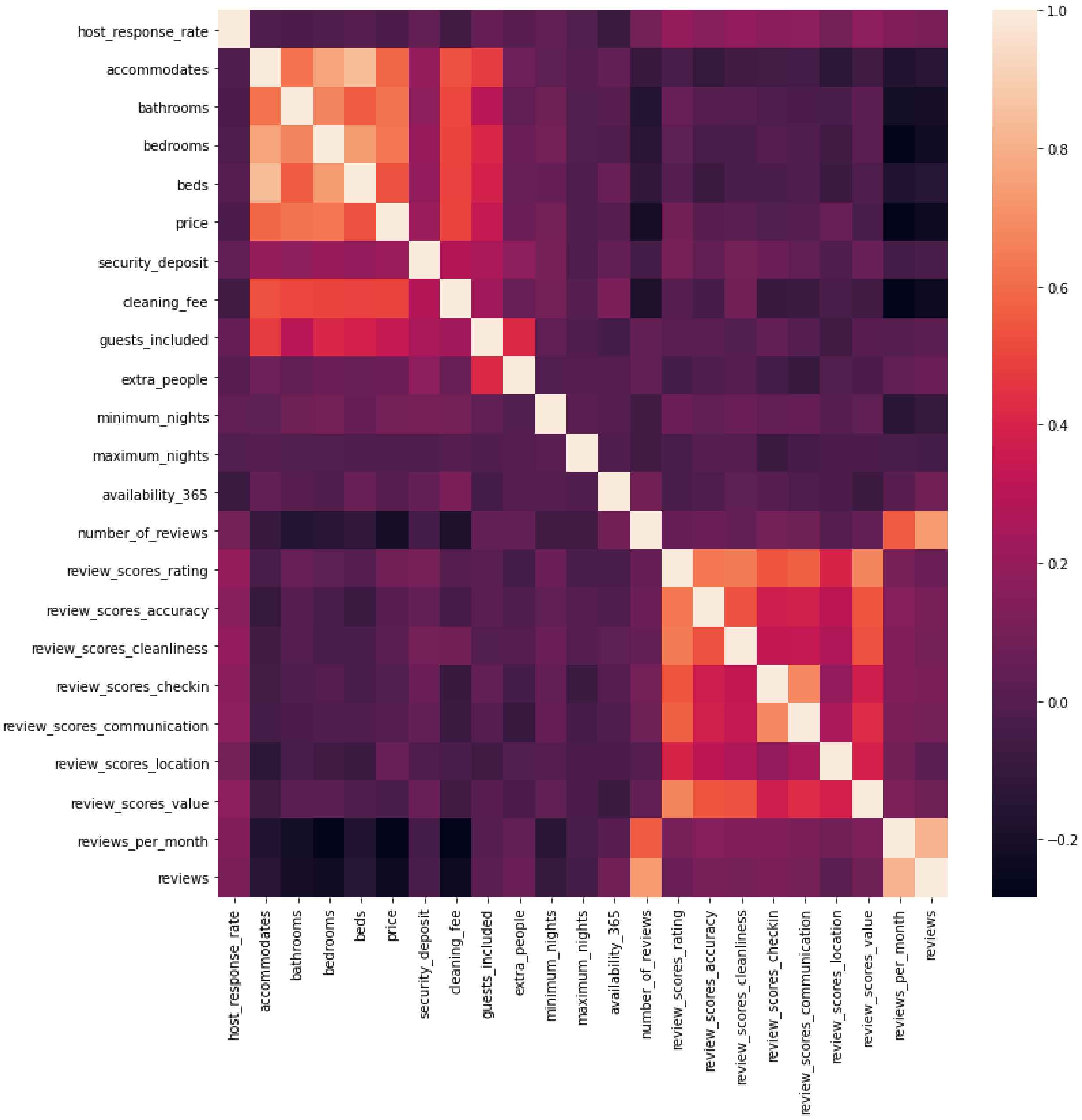
Some categorical data needed to be rolled up. An apartment and a condominium are really the same thing to a renter, so they were rolled up. Same with house and townhome.

All these steps resulted in a dataset with 1,714 records in 33 fields. This was a vast reduction from the original 3,818 records and 92 fields, but now represents Seattle data for Homes or Apartments below $400 a night.

Finally, categorical fields were one hot encoded, so that a field like neighbourhood was split up into several new binary fields.

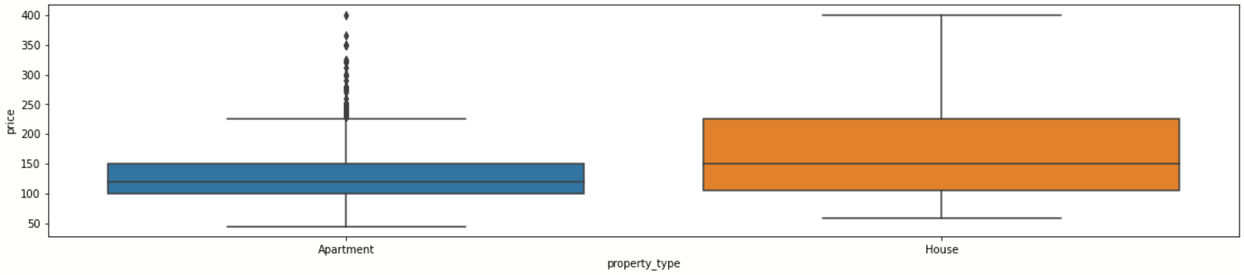
## 4. Exploratory Data Analysis

The major goal of my exploratory data analysis phase was to understand correlation between fields as well as potential drivers of higher prices. To that end, my first step was to create a heatmap to map correlation between fields.



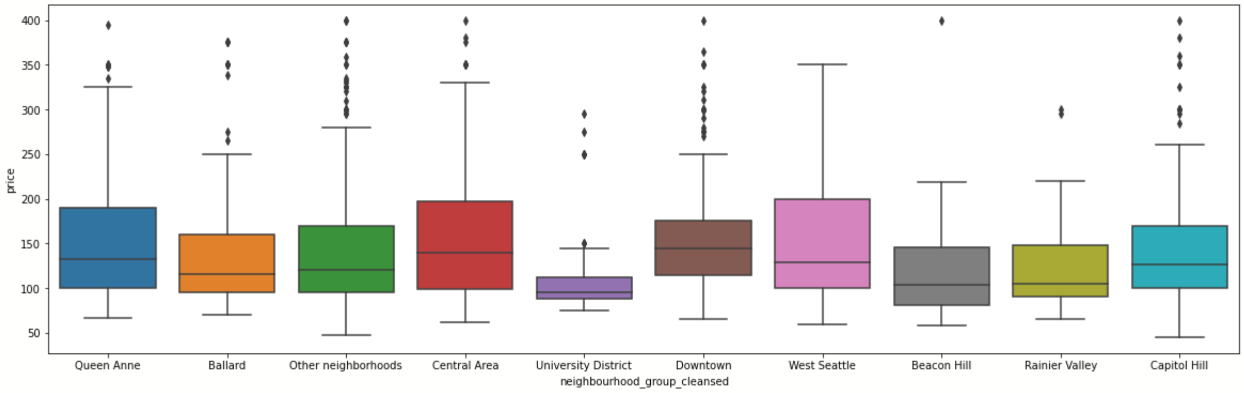
There are a lot of fields, but some things are apparent immediately.  There is a high degree of correlation between accommodates, bathrooms, bedrooms, beds and guests_included.  As there should be, as these all indicate the size of the listing.  These are all highly correlated with price as well.Furthermore, review scores are also all highly correlated, but not with price.

The next area of interest was checking prices against unit type (apartment or house).  Not surprisingly, a box plot showed that the range of prices was higher for house rentals.



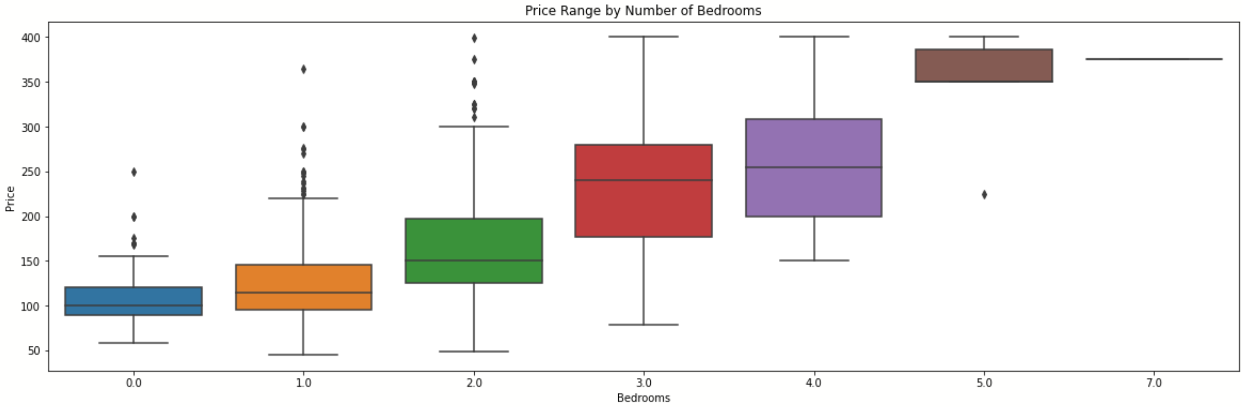
This makes sense, as a house would more likely contain more bedrooms and accommodate more people than an apartment.

Next was an investigation into price ranges by neighborhood.



I would expect downtown to be noticeably higher than others, but at the same time other neighborhoods would have homes that would drive prices up in relation to the downtown apartments.  A clear relationship is not visible in this graph.

Finally, I checked the price ranges based on the number of bedrooms, from zero bedrooms (a studio apartment) up to 7.  There was a clear visual relationship between bedrooms and price.



So far, the EDA has pointed me towards a feeling that listing size (bedrooms, bathrooms, accommodates, etc.) will be strongly predictive of price. Location does not appear at first glance to be a great predictor, but this could be because of the home type mix. Finally, review scores do not seem to predict prices.

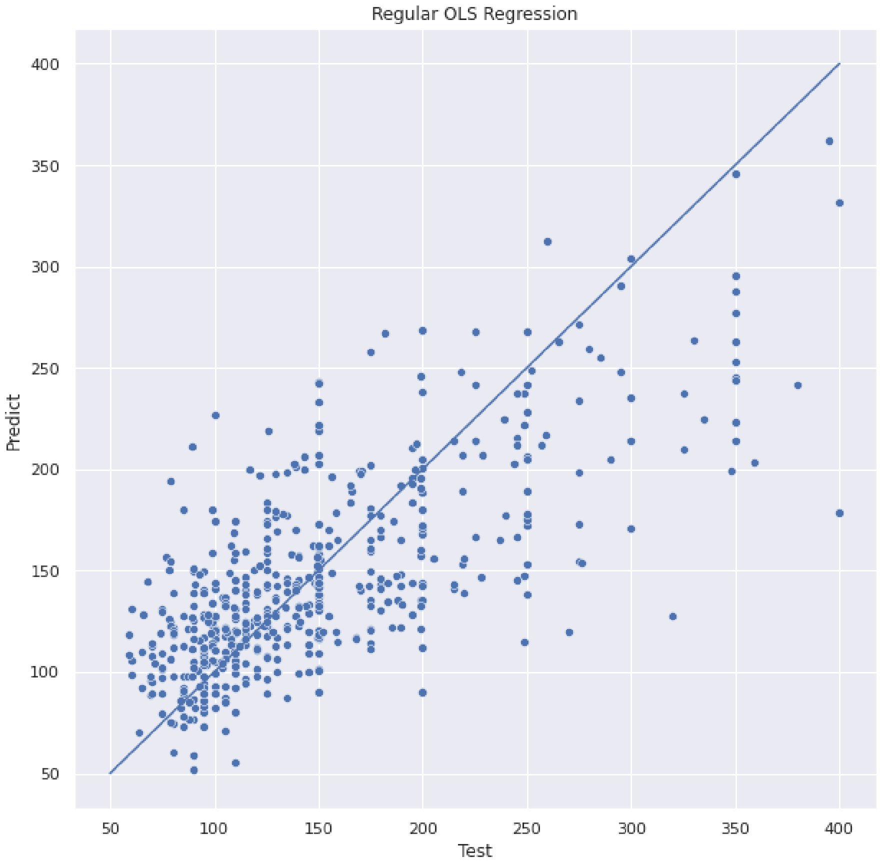
## 5. Modeling

Since the goal of this project is to develop a model to predict prices, three types of models came to mind: Regression, Random Forests and XGBoost. Since I'm interested in explainability, $R^2$ was a major quality metric to determine model suitability. Secondly, I used Mean Average Error (MAE) to determine the degree of the error rate.

### a. Regression

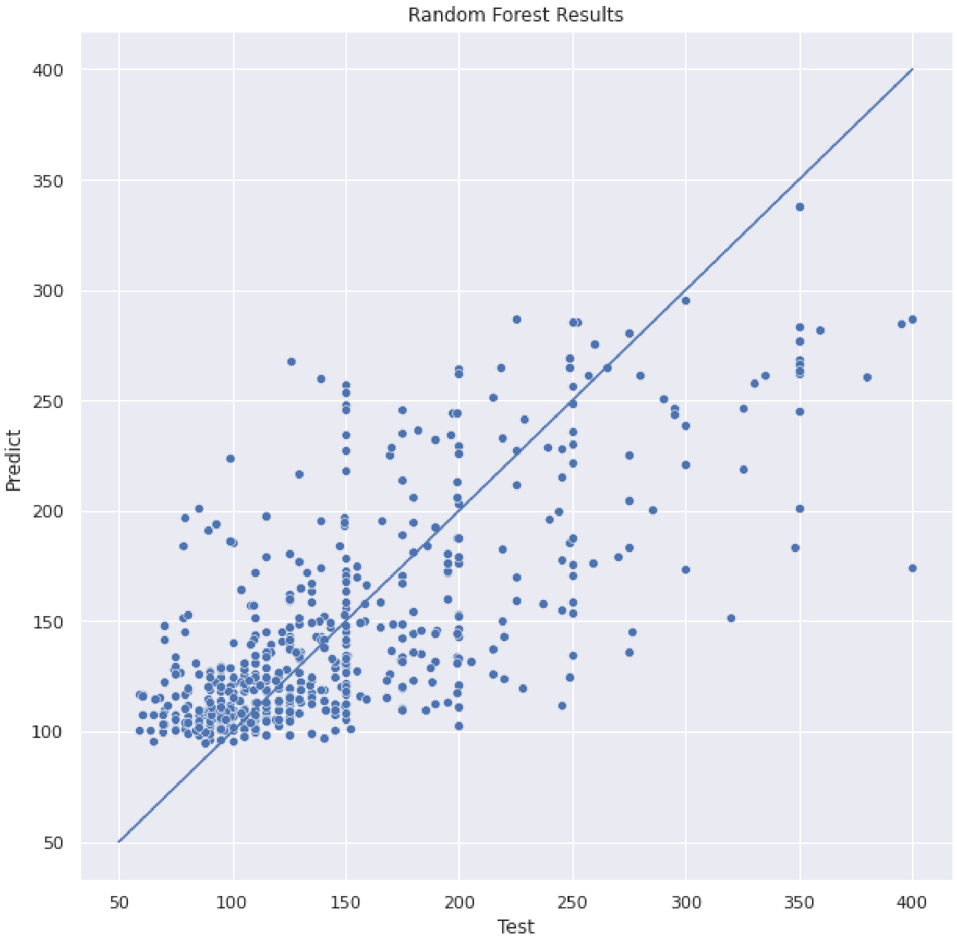
I tried a regular OLS regression model, as well as Lasso and Ridge regressions. Depending on the random split between train and test data, the $R^2$ was usually around .53 or a little higher. A scatterplot of predictions for actuals shows a generally correct trend, but not a lot of predictive accuracy. Also, predictions tended to be higher than actual at the low end, and lower than actual at the high end.

There was no real difference between the OLS, Lasso or Ridge regressions.



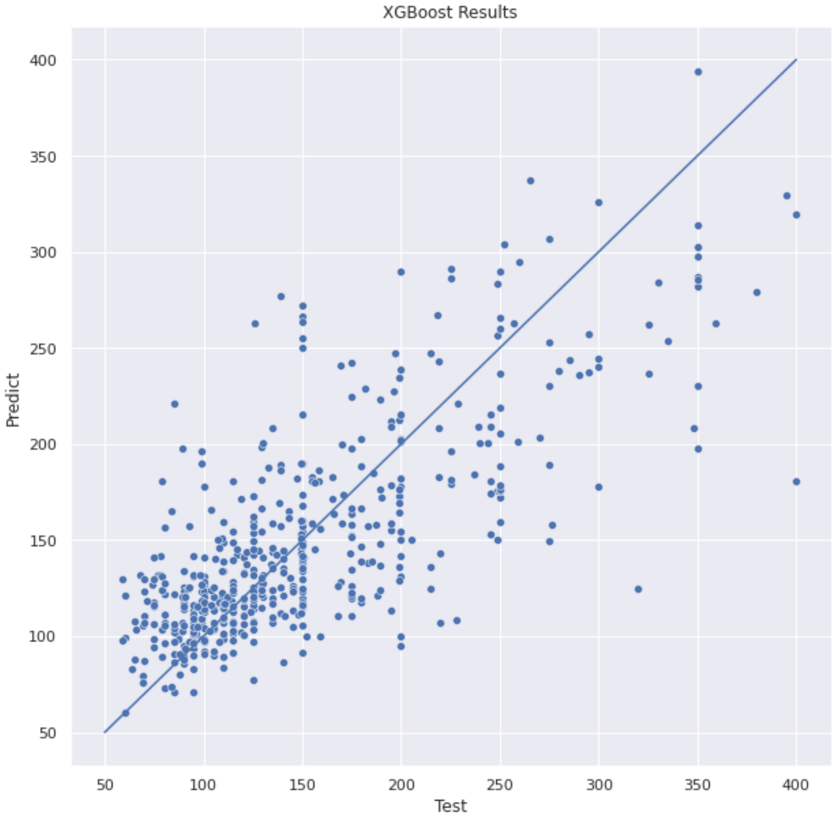
### b. Random Forests

The scatter plot also indicated the same issues at the low end and high end of the markets. Depending on the random state, results were the same as regression models or occasionally much better.



### c. XGBoost

This was the most computationally intensive of the methods, in terms of finding optimal parameters and requiring hours to run.  As with the other models, there was a clear bias evident at both the low and high end.



### d. Model Comparison

Comparing the five models against a dummy regressor (every listing having the same price) indicated an improvement on random guessing, and some utility to the model.  XGBoost did perform best, but at the same time is only marginally better than other models and not much better at predictive power.

| | **Dummy Reg.** | **Linear Reg.** | **Lasso Reg.** | **Ridge Reg.** | **Random Forest** | **XGBoost** |
|---|---|---|---|---|---|---|
| **$R^2$** | -0.01 | 0.53 | 0.54 | 0.53 | 0.52 | 0.57 |
| **MAE** | 51.12 | 33.96 | 33.89 | 33.95 | 34.79 | 32.86 |

A comparison of the top five features for each model indicates a large degree of consistency.

| **Linear Regression** | **Lasso Regression** | **Ridge Regression** | **Random Forest** | **XGBoost** |
|---|---|---|---|---|
| bedrooms | bedrooms | bedrooms | bedrooms | bathrooms |
| bathrooms | bathrooms | host_response_time_within an hour | bathrooms | bedrooms |
| neighbourhood_group_cleansed_Downtown | cleaning_fee | bathrooms | cleaning_fee | cleaning_fee |
| cleaning_fee | accommodates | host_response_time_within a few hours | availability_365 | neighbourhood_group_cleansed_Downtown |
| accommodates | neighbourhood_group_cleansed_Downtown | host_response_time_within a day | number_of_reviews | guests_included |

Ridge regression was the only model where host response times were important, but due more to negative coefficients.  Bedrooms, bathrooms and cleaning_fee were consistently important features, while several models also indicated the downtown location as being important as well.

For consistently best results and future analysis, the XGBoost model would be best to use.

## 6. Findings and recommendations

The original goal was to create a model that could determine the importance of home size, location and host actions (determined by reviews) in the setting of price. The best model has an $R^2$ that could range from .55 to .60, which is directionally decent but can in no way be used to provide firm price guidance.

What we can conclude, however, is that home size and location (downtown) do matter in setting the price, while the hosts' actions in terms of cleaning, responsiveness, etc., do not greatly impact the price.

Also, all the models over-predicted at the low end, and underpredicted at the high end. It seems there is some bias in the data which was not unearthed in the chosen features.

My recommendations to AirBnB on using this model:

1. Create a quality score, akin to a hotel's star rating. There are clear differences between listings based on uncaptured quality metrics. The model showed bias towards the topic and bottom price points, most likely driven by perceived quality.
2. Investigate supply and demand by size segment and neighborhood, and focus efforts on increased supply of tight demand unit types.
3. Focus effort on a greater understanding of price drivers through photographs and descriptions. Features may be uncovered which are important to renters that can actually be actioned by hosts.

## 7. Areas for future analysis

The previous section basically said that the best model is only a directional tool and discloses the most important features for price. However, further research could be undertaken to create new features, based on existing data. This analysis would potentially lead to more features that could be used to improve the previous models.

### a. Analyze Market Supply and Demand

The effects of market forces were not analyzed, but may shed some insight on pricing. Exactly how many one bedroom homes are in Downtown, and how occupied are they? This may shed light into supply issues and may help explain the bias found in the models.

Analysis of this would create models based on very distinct units. A person interested in renting a one bedroom unit would have no interest in a two bedroom unit, for example.

### b. Determine the effect of photographs on price

I have personally chosen AirBnB listings based on what I see in pictures, and it is possible that high quality, or informative, or quantity of pictures may help. This would be an exercise in creating a neural network to analyze pictures in order to guess prices, and would definitely be a much larger project in the realm of computer vision.

### c. Analyze the effect of descriptions

Hosts took the time to write descriptions of their homes. Could these ‘sales jobs’ have actually driven the price up or down? Perhaps they identified interesting aspects of the home, or pointed out neighborhood characteristics.  It is an unknown variable that would be worth investigating using NLP techniques.

### d. Identify host pricing biases

Many hosts have more than one listing, and from this it would be possible to infer if they have a pricing bias, either pricing too high or too low.  A clustering algorithm would be of interest here.

### e. Better feature selection

There were many features in the final dataset, and it was clear that many of them were correlated.  Initial filtering removed obviously unnecessary features, but others could have been removed based on multi-collinearity.  This may have reduced a model’s ability to explain features, but would have surfaced more unique features and definitely identified unimportant features.