# Airbnb Suggested Pricing Project

#### Milestone #5 Team: Stephen Camera-Murray, Himani Garg, and Vijay Thangella TF: Christine Hwang

## Proposal of Future Work

### Project Definition

Our goal remains the same as originally stated: provide guidance to New York City property owners listing their properties on Airbnb. Using the information they provide as predictors, our model will suggest a rental price range based on what we learned from past rentals, taking seasonality into account. For price categories we’ve settled on the following:

* Low: Up to $125
* Mid: $125 - $250
* High: Over $250

As a stretch goal, we’ll analyze text descriptions of the properties to recommend words and/or phrases that result in higher rental prices. This data is available at the Inside Airbnb website: <http://data.insideairbnb.com/united-states/ny/new-york-city/2016-07-02/visualisations/listings.csv> and <http://data.insideairbnb.com/new-york-city/2015-01-01/data/calendar.csv.gz>

### Approach for Price Category Prediction

#### Further Exploration of Seasonality

In milestone #3, we didn’t find any significant seasonality effect on price when looking strictly at the calendar seasons. But we did see price fluctuations and will explore further what might be causing them by looking at other seasonal effects such as holidays, back-to-school, or summer to see if there's a real effect. We'll delve deeper into the potential for seasonal effects from a holiday perspective and/or through interactions (e.g. higher prices around the holidays to accommodate friends and relatives).

#### Check for Collinearity

Next, we’ll create scatter matrices for any variables we might suspect to have some collinearity since doing them all would be unrealistic. If we see any very obvious relationships, we’ll drop one of the variables.

#### Build an Improved Model

Next, we’ll try a few different models with their default parameters to see if any of them look promising to improve the accuracy over the baseline using F-Score as the benchmark. We’ll start with: Logistic Regression, LDA, QDA, SVM, and Random Forest. From here, we’ll select a promising model and perform grid search cross-validation to tune the models hyper-parameters.

#### Analyze and Present the Results

We’ll conclude with a visualization of the feature importance (from the random forest model assuming we have a good fit to the data) to highlight which factors contribute to price. We will also include a visualization demonstrating goodness of fit (e.g. an ROC curve). For our website we also plan to include a static or somewhat staged demonstration of how our model could be used in the Airbnb UI.

### Approach for Text Description Recommendations (Stretch Goal)

#### Feature Extraction

We’ll start by creating a bag of words by stemming words from the listing descriptions and choose the most frequent stems. From this list, all stop words and neighborhood words will be removed. Frequencies for these words within each property listing will be created and the sum of the features will be normalized to one.

Since we’re trying to recommend words associated with higher prices, we’ll create a new binary target: high\_price. We’ll set the value to 0 for listings in the “low” price category and 1 for those in the “mid” and “high” price category.

#### Baseline Model

Before creating a new model, we’ll build a baseline models as we did earlier: all high price, all low price, and random. We’ll use these baseline models to determine whether or not we can make predictions on price that beat our baselines.

#### Build an Improved Model

Since our goal is to suggest words associated with higher prices, we’ll build a random forest model so we can then look at feature importance to determine which words in the description are the most important when it comes to predicting prices. We’ll then perform some moderate tuning of the model hyper-parameters with grid search cross-validation to get it to a point where it’s more accurate than our baselines.

#### Analyze and Present the Results

We’ll show a visualization of feature importance for the random forest model and select around the top fifty words which have a high importance. We’ll then look at the frequency with which they’re associated with high or low prices and create word cloud visualizations, one containing words associated with low prices (words to avoid) and another for words associated with high prices (recommended words).