**Milestone #5**

We can propose a number of different techniques and methodologies to apply to our data pipeline.

**Data Cleaning**

*Handling of Imperfections in Data*

1. The larger the number of missing values in a column, the lower the accuracy when imputing values for these features. We should test different imputation cutoffs to remove features (ie. remove all columns that have over 50% of values missing).
2. We should also test different imputation methods depending on variables. Perhaps ‘mode’ for categorical variables, ‘median’ for quantitative; or maybe KNN for everything.
3. After selecting our features, and introducing new features from seasonal and reviewer data, we should examine for multicollinearity between features. Remove problematic features.

*Clustering of Data (Multilevel Model)*

1. There are certain features, such as ‘accommodates’ or ‘bedrooms’, or perhaps ‘neighborhoods’, that will have a large impact on price. We can test clustering data based on these features, and using clustered training data to predict clustered testing data (ie. use a training set with only 1-bedroom listings to predict only 1-bedroom listings).

**Data Exploration (and Analysis)**

*Exploring Secondary Data Layers*

1. The seasonality data compared to pricing could help us better predict pricing during holiday seasons. We can impute missing data values in ‘calendar.csv’ using KNN as the only missing data is the quantitative response ‘pricing’. We could also choose to remove these rows depending on how much data is missing. We could then examine the frequency of prices over dates using a histogram - this will reveal months where there is a surge in pricing. If we find that seasonality has a notable impact on pricing, we can introduce a binary predictor (or *n-*label predictor), ‘holiday surge’, to indicate if a listing is impacted by seasonal pricing.
2. The reviewer data on a listing could also be transformed from categorical bag-of-words data into a binary predictor to indicate a positive or negative review. We could use sentiment analysis to classify these reviews - getting rid of less necessary words, or using a dimension reduction technique to minimize the number of features (PCR). This allows us to incorporate the reviewer data (along with the seasonality data) into our model as features.

*Visualizing Data*

1. A number of visualizations aid us in further exploring data. Correlations matrices and heat maps help us pinpoint multicollinearity. Graphs of response variables and MSE reveal the best combination of predictors. Colored scatterplots of individual ‘cluster variables’ (like the ones described in the *Multilevel Model*) vs. response are used to examine different distributions of clusters.

**Modeling (and Analysis)**

*Testing Regression Methods*

We our trying to predict a quantitative response, so classification methods (ie. logistic regression) will be less useful to us. On all regression methods, we use K-Fold cross validation the minimize overfitting of data. We tune for the best *k* parameter.

1. Ridge Regression and Lasso Regression: These will allow us to create models with lower bias to minimize or perform subset-selection on our features. We can test different tuning parameters to find the model with the lowest MSE. This will also reveal which features are essential to the model. Although Lasso regression is capable of performing subset selection, it is susceptible to high correlation.
2. Polynomial Regression with PCR: After standardizing data, polynomial regression would be used on particular features (ie. creating quadratic, cubic, etc. features of the ‘bedrooms’ to see their predictive ability). If this introduces multicollinearity, or we are still dealing with previous collinearity among explanatory variables, we can conduct PCR to see if we can reduce dimensionality and increase efficiency of prediction in the model.
3. ElasticNet Regression. This is a combination of Ridge and Lasso created to address the following:
   1. *n > p*, where *n* is sample size and *p* ispredictors, but covariates are strongly correlated
      1. This could be an alternative to PCR - without losing our features we add the bias of the penalty term of Ridge Regression, and minimize the effect of multicollinear terms. This reduces the effect of multicollinearity (present in the Lasso model).
4. Random Forest Regressor. This is another regression that allows us to make a lower-bias model. This ensemble method should be used over decision trees to combat overfitting, but also requires tuning of a large number of parameters. Gradient boosting will help us optimize the performance of the random forest.