**Price Analysis and Borough Prediction of Airbnb housing in New York City**

Kwangmin Kim (kk3225), Leiyu Yue (ly2428), and Kathryn Addabbo (kma2162)

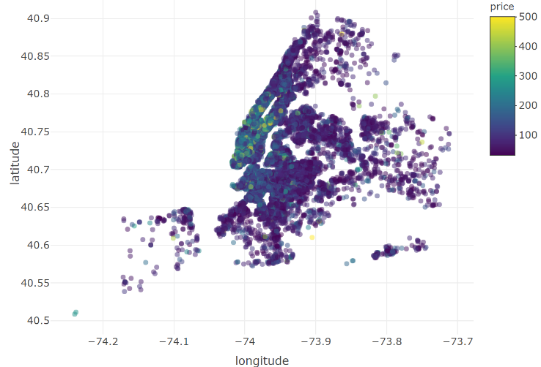
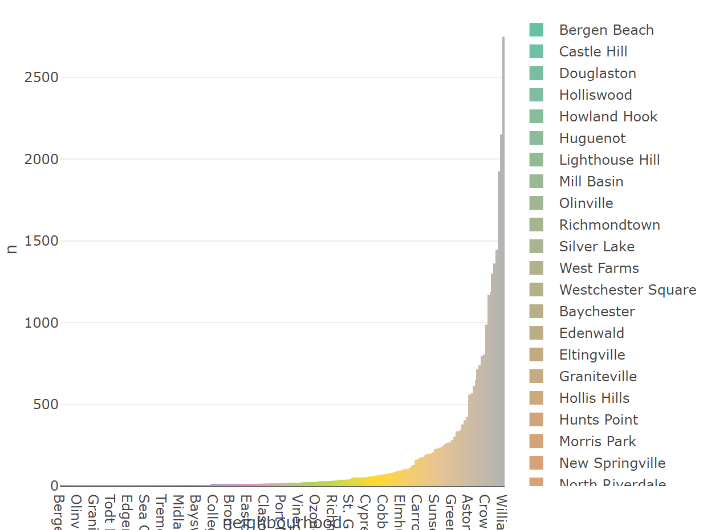
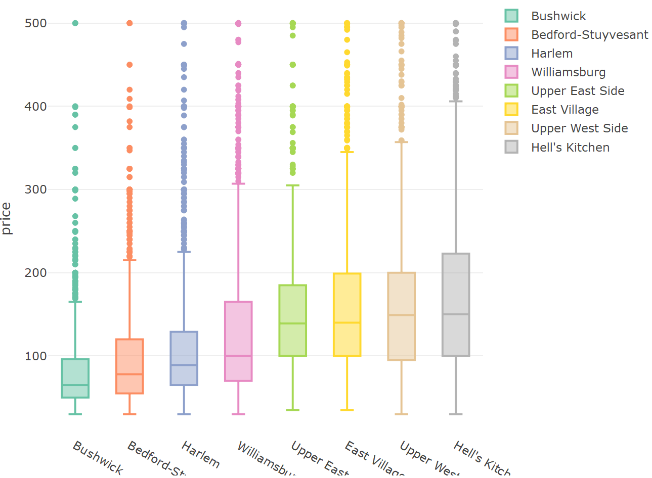
**1. Introduction**

Countless tourists and students visit to explore all the attractions NYC has to offer as one of the largest cities in the world. Being a dynamic environment, the overall population of the city fluctuates daily. To handle the influx of individuals, information on available housing is extremely important. One of the most well-known housing companies is Airbnb. In this report, we analyze the Airbnb housing affordability and availability in each of New York City’s boroughs using the company’s released dataset in September 2017. The aim of this report is to provide clear visualization and meaningful interpretation of the data for convenience of visitors and all interested in the service of Airbnb.

This data is the dataset that has been used in the data science I course referenced by http://jeffgoldsmith.com/DSI/dataset\_airbnb.html. This dataset is the result of  merging “listings.csv.gz” with “listings.csv” from Inside Airbnb using the code from the referenced website. It contains a single data frame with 40,753 rows of data on 17 variables: id, review\_scores\_location, name, host\_id , host\_name, neighbourhood\_group, neighbourhood, latitude, longitude , room\_type, price , minimum\_nights, number\_of\_reviews, last\_review, reviews per month, calculated\_host\_listings\_count, availability\_365.

**2. Exploratory Data Analysis**

**Figure 1: Price and Location in NYC**

One of the covariates of interest (price) has values ranging from $10 to $10,000 per night. We restricted the analysis to values between $30 and $500 because the average consumer does not have the appropriate funds to spend thousands of dollars a night on a room for a night, and we do not want too much variation in our analysis. The rating of each property was on a scale of 1 to 10, we changed the rating to a 5 – star scale and assigned zeros to the properties with no ranking. We also assigned zeros to the rating and reviews per month column where there were NAs present.

**Figure 3: Top Neighborhood Prices**

**Figure 2: Property Count**

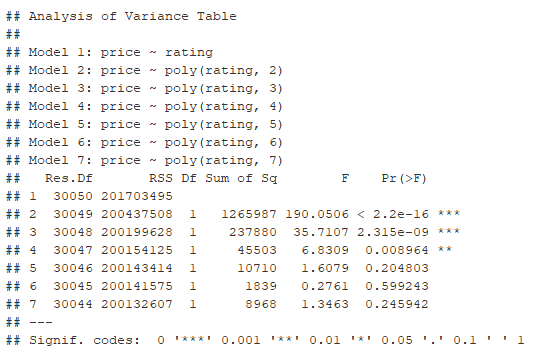
Housing is widely available through all five of the boroughs, however generally the closer to downtown Manhattan the property is, the more expensive the price per night as seen in Figure 1. Of the properties, the majority are in Williamsburg and Bedford – Stuyvesant, both of which are in Brooklyn, which are displayed in the bar plot of Figure 2. To explore further, Figure 3 displays the price range of the top eight most populated neighborhoods, which range of average from $50 to $150 and maximums of $500 and presumably over. To obtain an awareness of the relationships between the variables a correlation plot is created, which is viewed below in Figure 4. Outside of the obvious correlation between number of reviews and reviews per month in addition to id, the important correlation to take into note is that price is correlated with room type. Customers are willing to pay more money to have an entire house to themselves for their stay as opposed to just renting out a room. Outside of this observation, there is very little correlation in the dataset, which is ideal for analysis.

**Figure 4: Correlation Plot of Covariates**

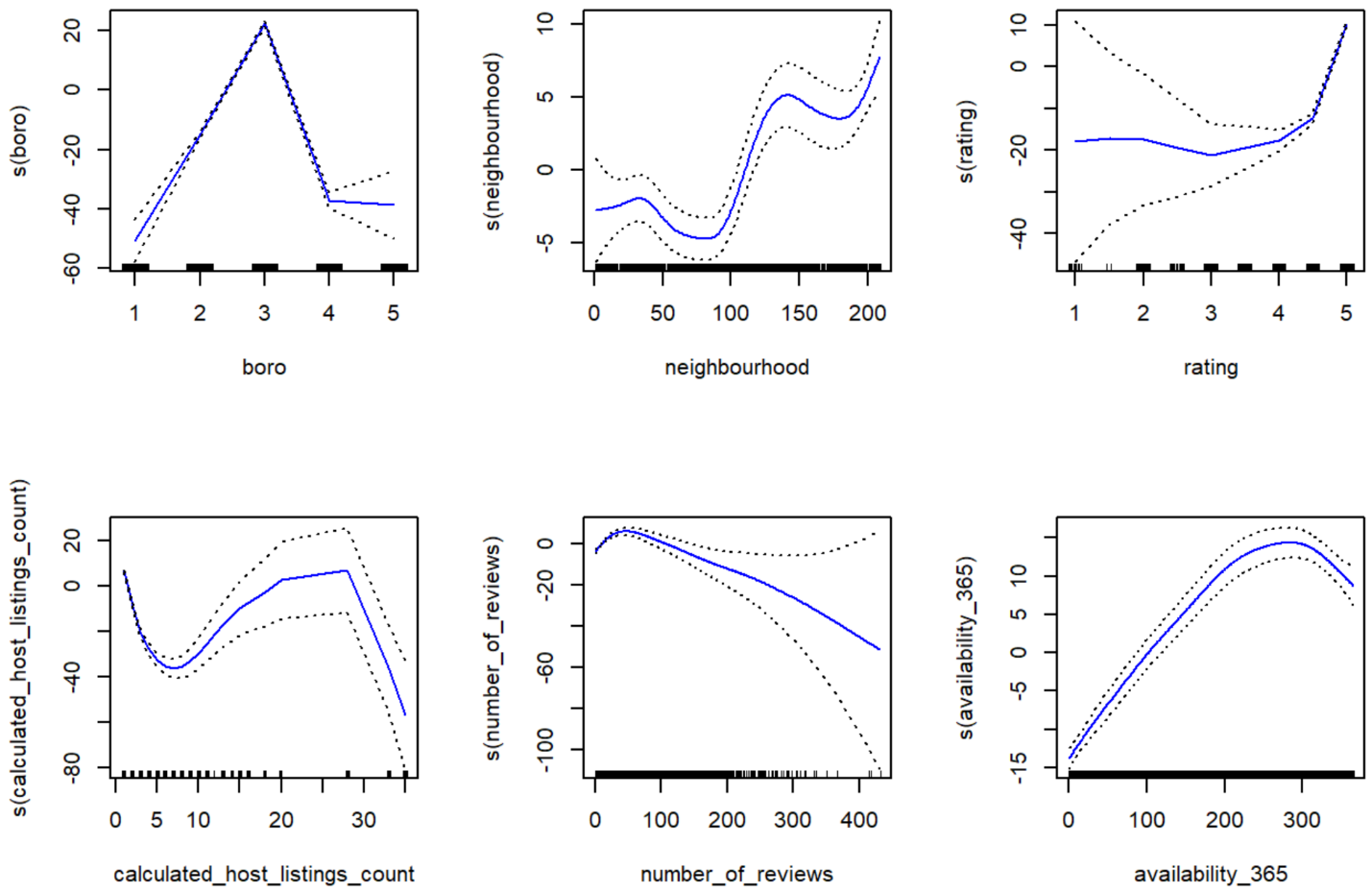
Multiple methods are used to investigate the relationship between price and the other covariates. Latitude, longitude, id and neighborhood and reviews per month are excluded from this analysis due to unique values and correlation between neighborhood and borough.

**3. Supervised Analysis**

Simple linear regression yields a test error of 4777.552, an enormous number indicating poor fit. Our team moved to utilize shrinkage modeling methods such as ridge, lasso, principle component regression and partial least squares with the goal of identifying the lowest test error of those analyses. Each of the methods were assigned specific tuning parameters. Ridge regression with the tuning parameter 0.178 had a test error of 4859.771. This estimate is close to simple linear regression estimate due the ridge tuning parameter’s proximity to zero, however it is slightly larger and thus a worse fit.

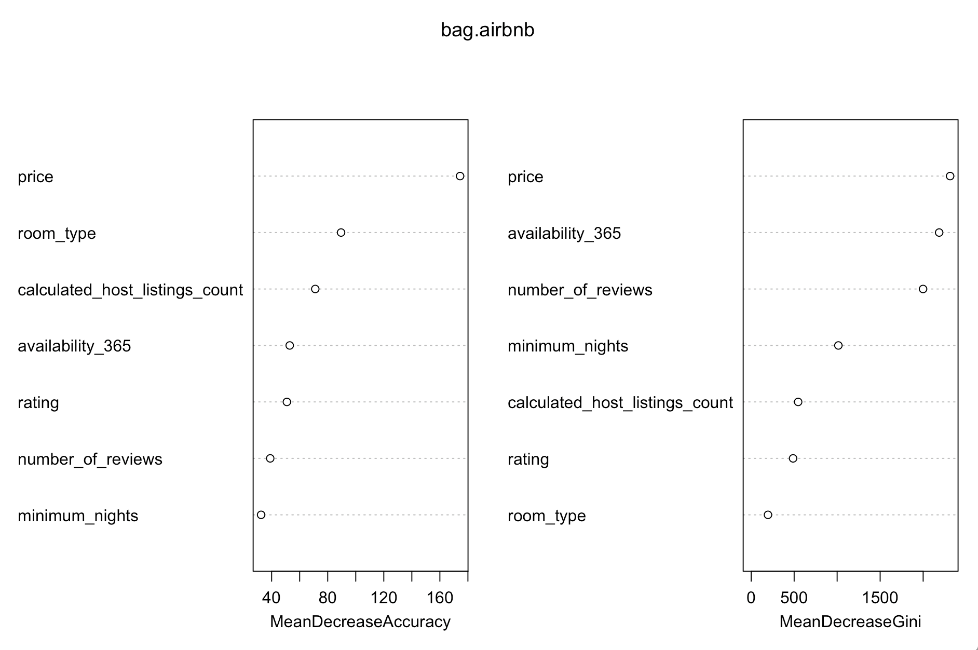
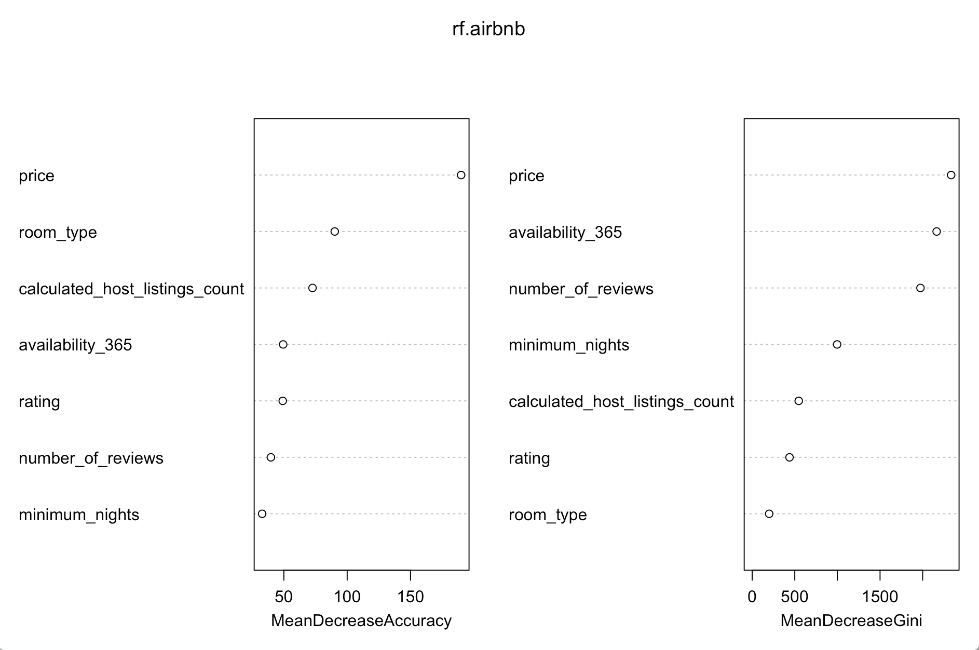
Lasso was the next method observed; although the method is ideal for sparse models it compensates for variable selections unlike ridge regression. This methodology often reduces the variance at the cost of increased bias, which we see in the results. With a best lambda of 0.28 the test error is 4860.522, this method is the worst of the three utilized thus far.

**Figure 6: ANOVA tests**

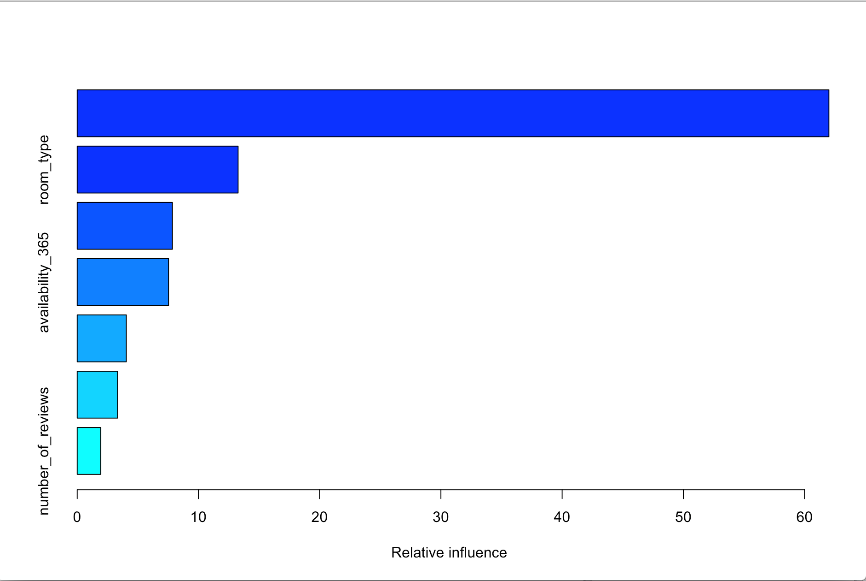
PCR is a technique which derives a low dimensional set of features from a large set of variables, in which directionality of the data indicates which observations vary the most. PCR yields a test error of 4564.748. The final method used is partial least squares which is the supervised alternative of PCR. PLS yields a test error of 4209.287, which is by far the best estimate of all the methods, therefore PLS delivers the best fitting model of the shrinkage methods as seen in Figure 5. However, the test error estimate is still rather large. To remedy this, polynomial, spline, GAM and regression tree methodologies were explored to see if it was possible to obtain a better fit.

**Figure 7: GAM results**

The optimal polynomial degree when selected by cross validation is 5 however when evaluating by p value and RSS, the optimal degree is 4. This is decided by evaluating the best degrees of freedom using the ANOVA test, which shows a substantial decrease in RSS, as seen in Figure 6. Spline also shows the best fit for a degree of 4. The data was fit using GAM with six predictors, price is shown to have nearly linear positive associations with availability and negative association with the number of reviews. Rating also appears to have a linearly increasing relationship with price, with Manhattan being the borough with the highest prices by far. These relationships can be viewed in Figure 7. The regression tree was pruned with an optimal value of 3 and with a cp of 0.0132 the test error was calculated to be 4503.777.

The other investigation being completed with this dataset is whether we can predict the borough the property is located using the other covariates of the dataset. Methods used for this analysis include classification tree, bagging, random forest, boosting, LDA and SVM. The goal of random forest is to improve the performance of decision trees by averaging the variance of the trees. This method creates a strong model by balancing the bias variance tradeoff and yields a test error of 0.42932.

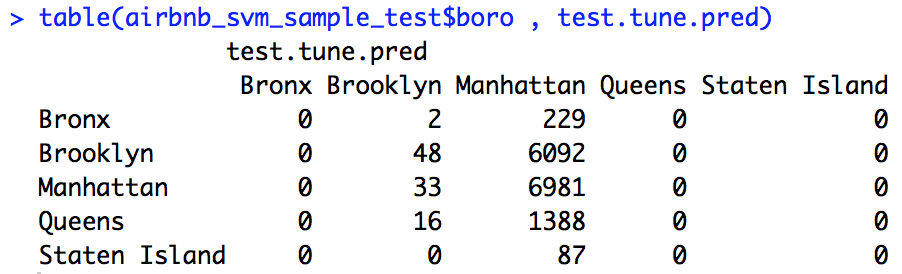
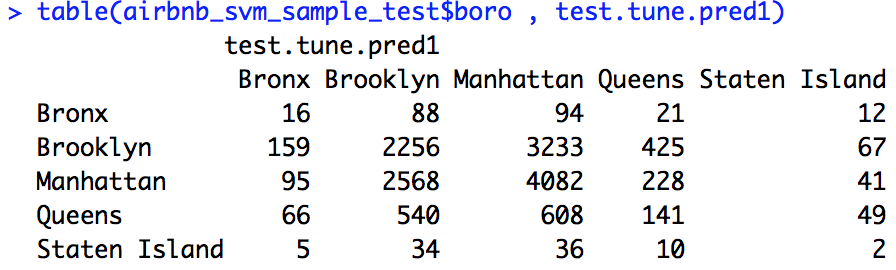
**Figure 8: Bagging & Random Forest Covariate Results**

Fitting a classification tree with a cross validated pruning parameter of tree size 3 yields a test error of 0.43445. Because this is a simple elementary analysis of the relationship between the borough and the covariates, bagging is used to fit several models on repeatedly sampled different subsets of the data with replacement. The corresponding models are then averaged and give a calculated test error of 0.4083. The analysis also indicates that price and availability are the most important variables in the analysis. Bagging and random forest summary plots of the accuracy and significance of the covariates can be viewed in Figure 8.

**Figure 9: Boosting Covariate Influence**

When using the sequential model building method of boosting, the calculated test error is 0.3156 with the most important variable being price followed by room type, rating and availability. Figure 9 displays a bar plot of the influence of the covariates on the response of borough.

**Figure 10: SVM Prediction Results**

An attempt at support vector machine was made using a sample of 200 observations from the original dataset, which separates the sample using hyperplanes. This method is very useful when the distribution of the data is unknown. To increase the accuracy of this method, the svm model was tuned with the parameter for linear type with an optimal cost of 0.01 and a gamma of 0.001 which calculates a test error of 0.52749. The model was then tuned with a radial type parameter, with optimal cost as 100 and gamma as 0.5. This yields a test error of 0.56326 while LDA achieves a measure of 0.5699. LDA is ideal for dimensionality reduction and viewing what variables are important to distinguish the differences in the response variable. Although a ROC curve cannot be plotted due to the multiple levels of the response variable. There is theory being investigated currently to plot categorical variables of multiple levels however nothing is proven at this time.

Of all the results investigating variation of price, boosting achieves the best fit of the data and classification tree produces the worst performance. With the multiple predictors in the data, boosting is easier to tune and unlikely to overfit the data compared to the other methods. In conclusion, the two analyses conducted on this dataset don’t reproduce ideal prediction values. This is most likely due to the skewed nature of the data and that most of the properties are in Manhattan and have the highest price overall.

**4. Conclusion**

In our project, we apply ridge regression, lasso regression, PCR and PLS for continuous variable "Price". And PLS reports the least error. And for categorical variable "borough", classification tree, bagging, random forest boost and svm are used. Classification tree has the worst performaces, which makes sense. After the reduce of bias-variance tradeoff, boost has the least test error. Price is the most important varible to predict which borough the airbnb is located, with Manhattan having the highest price. Room type and availability are as well of importance.