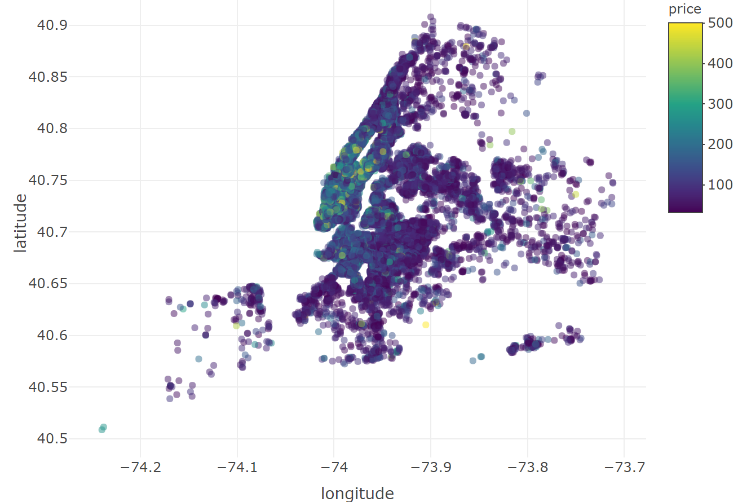
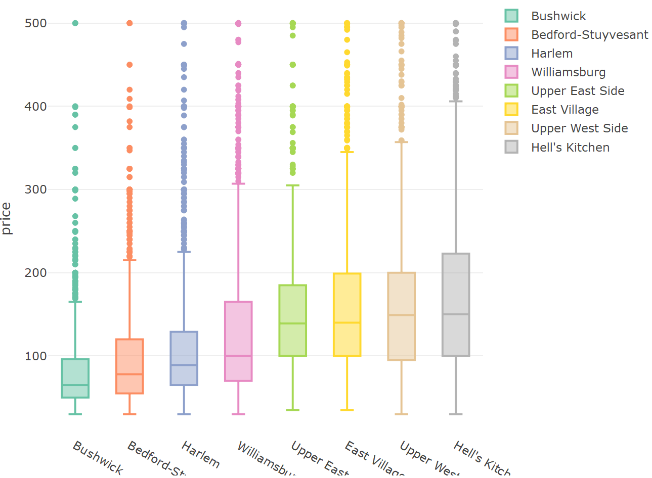
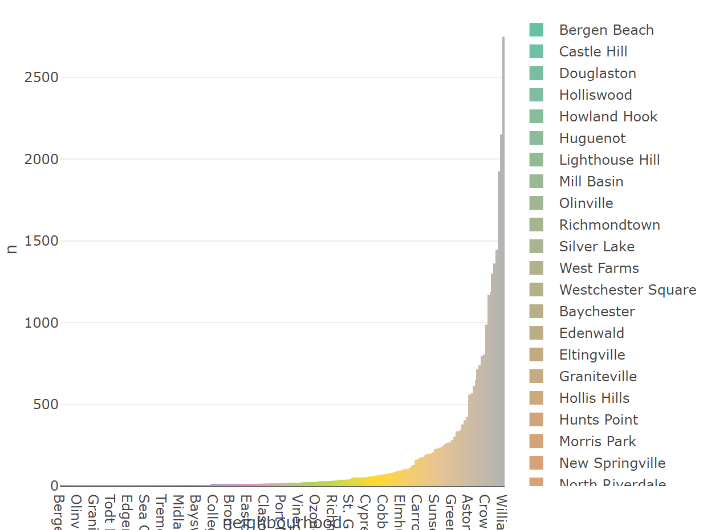
**Price Analysis and Borough Prediction of Airbnb housing in New York City**

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

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.

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. 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 1: Price and Location in NYC**

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. 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.

**Figure 3: Top Neighborhood Prices**

**Figure 2: Property Count**

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.

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. 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.

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. 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. 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, SVM, LDA and QDA.

Fitting a classification tree yields an optimal size of 3 with a test error of 0.447. Bagging yields a test error of 0.4083 and indicates that price and availability are the most important variables in the analysis. Boosting gives a calculated test error of 0.3156 with the most important variable being price followed by room type, rating and availability. Support vector machine yields a test error of 0.4257 while LDA and QDA achieve measures of 0.5699 and \_\_\_\_ respectively. Of all the results investigating variation of price, boosting achieves the best fit of the data. As the data has multiple predictors, boosting is easier to tune and unlikely to overfit the data compared to the other methods.