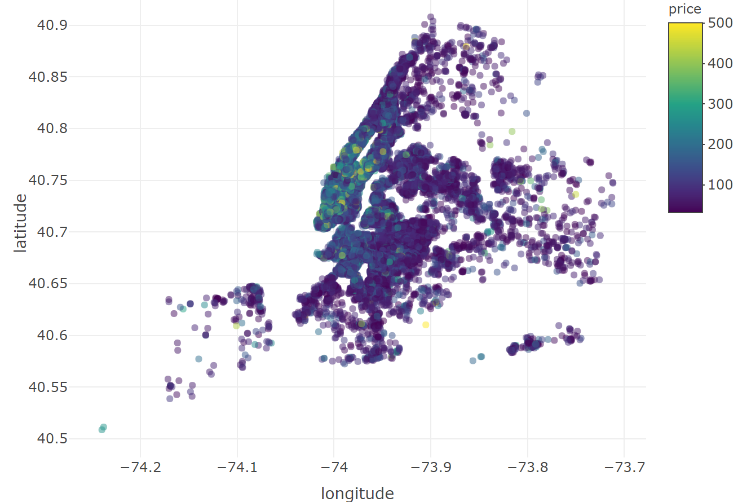
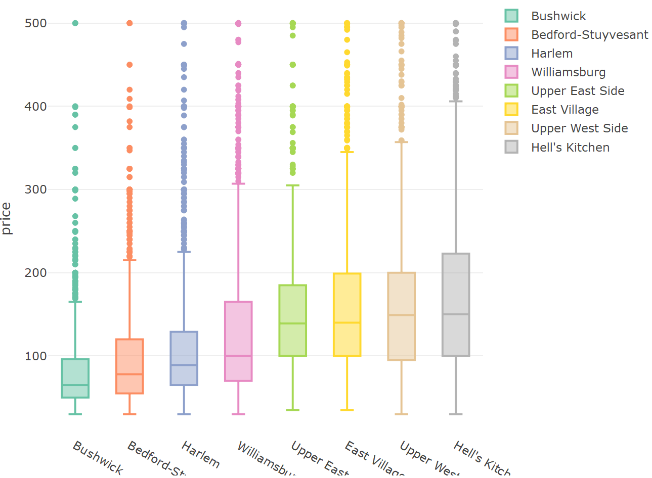
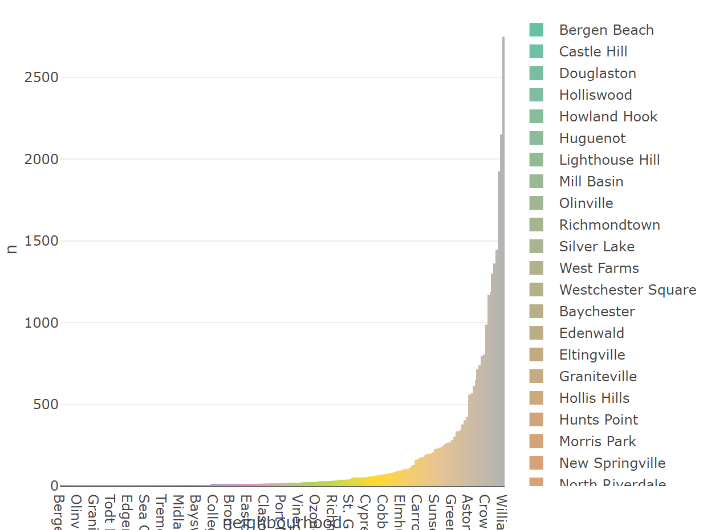
**Price Analysis and Borough Prediction of Airbnb housing in New York City**

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. To evaluate the relationship between the variables of the dataset, a correlation plot is created, which shows little important correlation between variables, which is good for modeling. Multiple methods are used to investigate the relationship between price and the other covariates.

**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.

Although PLS is the best fit model of the shrinkage methods, the test error estimate is still rather large. In addition, regression analysis was completed using polynomial, spline and gam methodologies.

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 LDA, QDA, classification tree, bagging, random forest and boosting.

Figures:

