|  |  |
| --- | --- |
| A sunset over a body of water with a city in the background  Description automatically generated  nyc Airbnb  An analysis to determine the best predictor variables of NYC Airbnb prices. | Jaclyn Coate, Laura Lazarescou, Reagan Meagher  MSDS 6372 Applied Statistics: Project 1 |

Table of Contents

[1. Introduction 3](#_Toc32780327)

[2. Data description 3](#_Toc32780331)

[3. Exploratory data analysis (EDA) 4](#_Toc32780367)

[4. Objective 1: Intuitive multilinear regression model 5](#_Toc32780377)

[A. Question of Interest & Methodology 5](#_Toc32780378)

[B. Chosen Models: Interpretation and Prediction 6](#_Toc32780382)

[C. Comparing Models 6](#_Toc32780386)

[i. Summary Chart of all models that were developed in Figure 8 6](#_Toc32780389)

[ii. Linear and Log-Linear Intuitive Model: Figures 9 and 10 for model output in R 6](#_Toc32780390)

[iii. Linear and Log-Linear Forward, Backward, and Seqrep (All Produce the Same Plots): Figures 11 and 12 for model output in R 6](#_Toc32780391)

[iv. Linear and Log-Linear Ridge: Figures 13 and 14 for model output in R 6](#_Toc32780392)

[v. Linear and Log-Linear LASSO: Figures 15 and 16 for model output in R 7](#_Toc32780393)

[vi. Linear and Log-Linear Elastic Net: Figures 17 and 18 for model output in R 7](#_Toc32780394)

[D. Assumptions: Intuitive Model 7](#_Toc32780395)

[i. Normality of Residuals 7](#_Toc32780396)

[ii. Constant Variance. 7](#_Toc32780397)

[iii. Independence 7](#_Toc32780398)

[iv. Multicollinearity 7](#_Toc32780399)

[v. Outliers: Leverage / Cook’s D 7](#_Toc32780400)

[E. Cross Validation 7](#_Toc32780401)

[F. Interpretation: Intuitive Model 7](#_Toc32780403)

[i. *minimum\_nights* 7](#_Toc32780405)

[ii. *number\_of\_reviews* 7](#_Toc32780406)

[iii. *tsquare\_distance* 7](#_Toc32780407)

[iv. *neighbourhood\_group*, *room\_type*, and *neighbourhood\_group*:*room\_type* 7](#_Toc32780408)

[G. Conclusions 8](#_Toc32780410)

[5. Objective 2: Two-way analysis of variance 8](#_Toc32780414)

[A. Problem Statement & Methodology 8](#_Toc32780415)

[B. Analysis 8](#_Toc32780417)

[i. Two Factor, Two-Way Analysis of Variance (ANOVA) 8](#_Toc32780418)

[ii. Mean Profile Plotting 8](#_Toc32780420)

[iii. Diagnostics 9](#_Toc32780422)

[iv. Testing 9](#_Toc32780427)

[C. Interpretation 10](#_Toc32780431)

[6. Appendix 11](#_Toc32780434)

[A. Figure 1 11](#_Toc32780435)

[B. Figure 2 11](#_Toc32780437)

[C. Figure 3 12](#_Toc32780439)

[D. Figure 4 12](#_Toc32780441)

[E. Figure 5 13](#_Toc32780443)

[F. Figure 6 13](#_Toc32780445)

[G. Figure 7 14](#_Toc32780447)

[H. Figure 8 14](#_Toc32780449)

[I. Figure 9 – Linear Intuitive Model 15](#_Toc32780450)

[J. Figure 10 – Log – Linear Intuitive Model 16](#_Toc32780453)

[K. Figure 11 – Linear Forward, Backward, and Seqrep Selection Models (All Produce Same Plots) 18](#_Toc32780457)

[L. Figure 12 – Log – Linear Forward, Backward, and Seqrep Selection Models (All Produce Same Plots) 20](#_Toc32780462)

[M. Figure 13 - Linear Ridge Model, glmnet() alpha=1 21](#_Toc32780467)

[N. Figure 14 - Log-Linear Ridge Model, glmnet() alpha=1 22](#_Toc32780470)

[O. Figure 15 - Linear LASSO Model, glmnet() alpha = 0 22](#_Toc32780473)

[P. Figure 16 - Log-Linear LASSO Model, glmnet() alpha = 0 22](#_Toc32780476)

[Q. Figure 17 - Linear Elastic Net Model, glmnet() chose alpha=.1 23](#_Toc32780479)

[R. Figure 18 - Log-Linear Elastic Net Model, chose alpha=.1 23](#_Toc32780482)

[S. Figure 19 23](#_Toc32780485)

[T. Figure 20 24](#_Toc32780487)

[U. Figure 21 25](#_Toc32780490)

[V. Figure 22 25](#_Toc32780492)

[W. Figure 23 27](#_Toc32780495)

[X. Figure 24 27](#_Toc32780496)

[Y. Figure 25 29](#_Toc32780501)

[Z. Figure 26 30](#_Toc32780503)

[AA. Figure 27 30](#_Toc32780505)

[BB. Figure 28 30](#_Toc32780507)

[CC. Figure 29 33](#_Toc32780513)

[DD. Figure 30 33](#_Toc32780515)

NYC Airbnb

# Introduction

## Who doesn’t love a vacation? I think it is safe to say, no one. In a recent [report by Eventbrite](http://eventbrite-s3.s3.amazonaws.com/marketing/Millennials_Research/Gen_PR_Final.pdf) “Millennials just can’t get enough. 72% say they’d like to increase their spending on experiences rather than on material things […], pointing to a move away from materialism and a growing appetite for real-life experiences.” We believe that the biggest thing people are looking to do is capitalize on cheap costs to stay places in order to have more money to spend on experiences while they visit their destination. Airbnb has captured this market.

## New York City is the city that never sleeps. The big apple. [Destination New Yorker](https://nycfuture.org/research/destination-new-york) has recorded the number of tourists tops around 60 million annually and hit a record high in 2017 with 62.8 million. With this being one of the top destinations in the world we focused our analysis on visiting this beautiful and eclectic destination.

## Contained within this report is an effort to predict the price of a traveler’s Airbnb when NYC is their destination. We utilize multiple linear regression techniques to try and fit the best model: intuitive, forward, backward, seqreq, Ridge, LASSO and ElasticNet selection methods. We complement these efforts with a two-way analysis of variance methodology and our findings explained for all of these models. This report is meant to provide our readers with a way to help predict their costs for staying in NYC Airbnb in order to help them leverage their vacation budget towards more experiences while they are there.

# Data description

## The data we used was provided by a Kaggle competition but originates from [Airbnb’s owned data](http://insideairbnb.com/). This data set in its raw state contains over 48,000 observations and 16 variables. Our dependent variable is *price* in US dollars for Airbnb rental properties across the major NYC boroughs. Please reference the below table for a full list of the variables and their descriptions.

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| id | Listing ID |
| name | Name of the listing |
| host\_id | Host ID |
| host\_name | Name of the host |
| neighbourhood\_group | Location |
| neighbourhood | Area |
| latitude | Latitude coordinates |
| longitude | Longitude coordinates |
| room\_type | Listing space type |
| price | Price in dollars |
| minimum\_nights | Amount of nights minimum |
| number\_of\_reviews | Number of reviews |
| last\_review | Latest review |
| reviews\_per\_month | Number of reviews per month |
| calculated\_host­­\_listings\_count | Amount of listing per host |
| availability\_365 | Number of days when listing is available for booking |

# Exploratory data analysis (EDA)

## Upon first look we reviewed the date columns. Since real estate in New York City is known for its high cost and constant fluctuation with growing and changing neighborhoods; we concluded that the most recent data would be the best in modeling something like the cost of renting real estate. We chose to leverage the *latest\_review* variable and only build our model from listings that have reviews posted between 2017-2019.

## Next we decided to drop logically irrelevant variables. Things like unique identifiers (*id, host\_name, host\_id, name*) were the first we chose to remove. Unique identifiers like names and IDs would not be useful in predicting something like price. These are number and categorical labels that would only be considered white noise and if there is a correlation it would be at random and not have any practical significance in our model.

## With further review we had a unique thought on a variable we could create distance to Times Square. We were provided with the longitude and latitude of the different Airbnb listings. Since we are not creating a geographical analysis, we thought a good way to utilize these metrics was to create a new one called *tsquare\_distance* to represent that Airbnb’s listing to Times Square, one of the most popular tourist destinations in NYC.

## Our next round of variables that were removed based on usability were: *last\_review*, latitude, longitude, and *neighbourhood\_group*. While the *last\_review* variable was useful from filtering the data perspective, we did not see any practical significant to use this in our modeling since it does not come with any sort of sentiment indicator, which we do believe if something like review levels were: good, okay, and bad; that would affect the price of the Airbnb. Next, we chose to drop latitude and longitude. This is because, while positions of the Airbnb we do believe matter in price, a more accurate way to interpret this that will have practical meaning to our readers is through the names of the burrows (or *neighbourhood*). Then we dropped *neighborhood\_group* this is because we found two variables that were measuring the same thing: *neighbourhood* and *neighbourhood\_group*. We found the *neighbourhood* variable being the most granular variable and therefore potentially the most accurate.

## The next few steps in our EDA were to ensure things like a clean data set, outlier detection and multicollinearity. We did a ‘zero’ value check on variables where that does not make sense (e.g. any rows with 0 listed in *price* or *availability\_365* were removed). Then we checked for NAs in our remaining data set. Where we found none. We also decided to target our model in order to only model those Airbnb listings that are between $25 - $400 USD. This is because we are looking to model some of the best deals for those individuals who are interested in their experiences once they arrive in NYC and not necessarily wanting to spend all their vacation dollars on the location where they store their luggage. Then we took a look at the *minimum\_nights* variable, and removed anything over 365 days, thinking that a contract might be for a year but something that ranges in the ranges of thousands doesn’t make for our purposes.

## After producing a cleaned data set, we worked through non-linear trends using transformations. We initially check for linearity with our numeric variables against our dependent variable (*price*). We do not see any linearity (Figure 1). Note that all of the numeric variables show random clouds of data this indicating no real sign of a linear trend. Next we tried a log-linear transformation to see if we could surface one. Again, we’ve surfaced no direct linearity with the independent and depend variables (Figure 2) due to random clouds of data and in some cases, what appear to be a mountain like formations or ‘peaks’. Next in our attempt to surface linearity and performed a log-log transformation. In order to see that this has create an even stronger random cloud effect (Figure 3) and has not increased linearity, but in reality, has actually taken us farther away from creating an MLR relationship. In our last attempt to surface a plausible linearity with our numeric variables we have binned the data (Figure 4). Binning our continuous variables allows us to see if there is a large linear relationship from a high perspective by taking away the granularity in the data set, which can be helpful in a set as large as areas. Unfortunately, we are again met with random clouds of data and/or what appears to be ‘peaks’ in the data itself. After these exhaustive manipulations we are determining that this points us towards a more complicated relationship, likely not linear.

## NOTE: Within all of these linearity checks we included the graphical display by the *neighbourhood* in color. This was our check to see if even though there wasn’t a high linearity, we could derive any sort of borough effect on the relationship, which would have been displayed by obvious groups. Unfortunately, these efforts also did not produce the ability to discern any type of relationship between the numeric variables and the independent numeric variables.

## Lastly, we checked for signs of linearity between our remaining two categorical variables and our dependent variable (*price*). We built boxplot graphs in order to investigate any differences in price based on the categories *neighbourhood* and *room\_type*. In Figure 5 we can see that there are many boroughs that are listed in our *neighbourhood* variable. While it is difficult to view, we can closely examine and see that there are multiple neighborhoods that seem to spike higher than others in our graph. In an attempt to confirm this, we are going to bin our neighborhoods to confirm our conclusion that there is a linear relationship to surface with this categorical variable and our dependent variable. In Figure 6 we can see that we were right, with varying levels in our point graph we are able to move forward with our conclusion of a significant linear relationship to surface between *neighbourhood* and *price*. In Figure 7 you can see that the room type greatly changes the price of an Airbnb. Therefore, it is likely we will surface a significant relationship in an MLR model. So, we will retain this variable as well.

## Our EDA complete we begin our model building for a multilinear regression.

# Objective 1: Intuitive Multilinear Regression

## Question of Interest & Methodology

## Our question of interest in this objective was to predict the price of an Airbnb listing in New York City using linear regression techniques and the variables remaining from our exploratory data analysis. These variables are price (our dependent variable), *neighbourhood\_group* (categorical), room\_type (categorical), *minimum\_nights* (continuous), *number\_of\_reviews* (continuous), *host\_listings\_count* (continuous), *availability\_365* (continuous), and our new, self-created variable, *tsquare\_distance* (continuous).

## It was clear from our EDA that we may have to do transformations on our data to satisfy the assumptions of multiple linear regression during model building. Therefore, we have two data sets that we are working with for each of our model selection techniques: linear.nyc and log.nyc.

## Overall, we did seven types of feature selection: forward, backward, seqrep, Ridge, LASSO, Elastic Net, and what we are calling an intuitive approach. After utilizing these seven methods, we feel that we have the best possible regression model to predict the price of an Airbnb listing in New York City, while accounting for the need of interpretability and limiting overcomplexity.

## Chosen Models: Interpretation and Prediction

## Ultimately, we chose two models: the intuitive model ran on linear data for interpretation and the LASSO model ran on log - linear data for prediction.

## The intuitive model was built by logically adding and subtracting variables based on how we thought they would impact price. This was a very practical approach based on our deep knowledge of the data and firsthand knowledge of what impacts the list price of an Airbnb. Our intuitive model contains the following variables to predict price: *minimum\_nights, number\_of\_reviews, tsquare\_distance, neighbourhood\_group,* and *room\_type.* It also includes an interaction term between *neighbourhood\_group* and *room\_type*. This intuitive model did not have the highest adjusted R-squared, but it is the best model for interpretation because it does not have log transformed data.

## For the predictive model, we used the LASSO regularization and factor selection technique, where the extreme coefficients are dampened, improving predictive performance.  This predictive LASSO model includes many of the same factors as our intuitive model, but it is run on log-linear data to improve linearity and provides a higher adjusted R-squared RMSE. The log-linear LASSO model had the highest adjusted R-squared out of all of our models and is the best model for prediction.

## Comparing Models

## During the feature selection portion of our analysis we went through seven different techniques to try and generate the best model. From this process we were able to generate metrics on each of the models selected from the various feature selection techniques. These metrics for each of the models we generated can be seen in Figure XX.

## When looking at all of these models to determine which would be best, we took three things into account: the metrics, the need for easy interpretation, and limiting overcomplexity. From the metrics we can see that all of the linear and log models actually achieved very similar results compared to each other, so there was not a good way to discern which was best. The forward, backward, and seqrep feature selection techniques also did not select an interaction term in their model, which we felt was important for the model. We did include an interaction term in our intuitive model between *neighbourhood\_group* and *room\_type*. The Ridge, LASSO, and Elastic Net all produced models with better metrics. Since the LASSO model on log-linear data had the best metrics, we felt it would be the best model for prediction. It was also clear that none of the models we derived from feature selection were great models, with all having an adjusted R Squared between .4 and .6. This is consistent with our EDA in showing us that linear regression is likely not the best way to model our data. Therefore, we concluded that the intuitive model, when run on linear data, would be the best regression option for interpretation due to its similar metrics to other models, ease of interpretation, and lack of overcomplexity. The LASSO model, when run on log - linear data, is the best regression option for prediction due to it having the highest adjusted R-squared.

The output for the intuitive, forward, backward, seqrep, Ridge, LASSO, and Elastic Net models can be seen in the following Figures:

### Summary Chart of all models that were developed in Figure 8

### Linear and Log-Linear Intuitive Model: Figures 9 and 10 for model output in R

### Linear and Log-Linear Forward, Backward, and Seqrep (All Produce the Same Plots): Figures 11 and 12 for model output in R

### Linear and Log-Linear Ridge: Figures 13 and 14 for model output in R

### Linear and Log-Linear LASSO: Figures 15 and 16 for model output in R

### Linear and Log-Linear Elastic Net: Figures 17 and 18 for model output in R

## Assumptions: Intuitive Model

### Normality of Residuals: to check the normality assumption we need to check “Normal Q-Q” residual plot. For the intuitive model we can see there is not ideal normality as the end of the plot tails up, but it is near normal and we move forward with the assumption met (Figure 9)

### Constant Variance: to check the constant variance assumption we need to check the “Residual vs Fitted” residual plot. For the intuitive model again there appears to be a slight departure from constant variance showing a little bit of a higher end on the y axis above zero. However, we move forward cautiously with the assumption met (Figure 9).

### Independence: For our analysis we are assuming that each Airbnb listing was made independently and that for both the linear and log-linear data sets there is independence.

### Multicollinearity: To check for multicollinearity within our model we need to look at the VIF chart. For the intuitive model the only value in the VIF causing concern is for *room\_type* with a VIF value of 5.23. However, the rest of the values are between 1 and 2, so we are not concerned about multicollinearity in our model (Figure 9).

### Outliers: Leverage / Cook’s D: We were able to eliminate a lot of outliers from the data in our EDA, but to check for outliers in our intuitive model we need to check the “Residuals vs Leverage” residual plot. Based on the plot and the fact that we removed outliers in our EDA there is no concern for outliers in our intuitive model (Figure 9).

## Cross Validation

## To address the need for cross validation in our analysis, we used a 70/30 train/test split for each of the two data sets used in our analysis (linear – linear and log – linear).   The Ridge, LASSO and Elastic Net methods applied internal cross-validation to determine the best lambda before determining final coefficients. All models were subsequently cross validated by a 70/30 train/test split for predicting on the test data.

## Interpretation: Intuitive Model

## The model that we are using for interpretation is the intuitive model run on a linear data set. An output from R with the coefficient values from this model and their respective confidence intervals can be seen in Figure 9. The interpretation of each of the coefficients in this model are:

### *minimum\_nights*: All else held constant, for every 1-night increase in the minimum number of nights required by the listing, the predicted price of the listing decreases by $0.56 (A 95% Confidence Interval ($0.51, $0.61)).

### *number\_of\_reviews*: All else held constant, for every additional review that someone leaves on the Airbnb listing, the price of the listing decreases by $0.078 (A 95% Confidence Interval ($0.065, $0.091)).

### *tsquare\_distance*: All else held constant, for every mile away the listing is from Times Square, the price of the listing decreases by $4.71 (A 95% Confidence Interval ($4.34, $5.09)). Clearly, our renters will pay more for proximity to the center of Manhattan and touristic sites.

### *neighbourhood\_group*, *room\_type*, and *neighbourhood\_group*:*room\_type*: These are all categorical variables/an interaction term and so the specific increase/decrease in price associated with these coefficients depend on which categories of the variable the listing falls in. All possible coefficients for these variables and their associated confidence intervals can be seen in Figure 9.

## In order to make the model better, we would have liked to consider additional factors in our original data set. The dataset from the Kaggle site was limited, so when we looked at New York City Airbnb listings on Airbnb’s app, some of the factors we would have used in our analysis are: number of guests, number of bedrooms, number of beds, number of bathrooms, private bathroom (Y/N), Super-host (Y/N), and amenities.

## Conclusions

## Ultimately, we feel that for multiple linear regression, our intuitive model is best for predicting the price of an Airbnb listing. However, multiple linear regression is just one option in building a predictive model for a continuous response. In our EDA we saw that a Multi Linear Regression analysis was likely not the best way to create our prediction model. We found it difficult to surface any linearity between the independent continuous variables and the dependent continuous variable. We saw shapes in Figure 1 through Figure 4 that likely demonstrate a more complex relationship than just linear. Any sort of extreme transformation actually results in the data moving farther away from linearity.

## Since we surfaced the above facts about our data and our dataset is large, we think that other methods such as Random Forest or K-NN would perform better. These options are less time consuming because the models’ complexities are built into software algorithms. We would also not have to specify how a relationship exists ahead of time, like we have for MLR.

## This being stated we have seen a strong relationship between the borough of the city as well as the room type being rented. This leads us into a Two-Way Analysis of Variance (ANOVA) model as it meets all the requirements to use this unique method (Figure 19). In our next section we explore what we surface with our Two-Way ANOVA.

# Objective 2: two-way analysis of variance

## Problem Statement & Methodology

## We analyzed the ability to predict the prices of Airbnb listings in NYC based on the variables provided per unit by Airbnb with the above MLR. In this process we completed a full exploratory data analysis (EDA) and took into account the practicality of the information we determined the two best categorical factors to that contribute to the price are: *neighbourhood\_group, and room\_type*.

## Analysis

### Two Factor, Two-Way Analysis of Variance (ANOVA)

### With our two-way ANOVA, we were in a two independent categorical variables scenario. With our cleaned data we created a summary table containing: *mean, standard deviation, standard error, minimum, maximum, and inner quartile range* of our dependent variable: *price* (Figure 20).

### Mean Profile Plotting

### Once this was completed, we used the table to create a mean profile plot (Figure 21). We determined our Two-Way ANOVA was displaying nonadditive model characteristics; in other words, our graph, Figure 21, indicated that the changes in the mean price of an Airbnb in a specific neighborhood depends on what type of room you are renting. Our data set had an unequal number of observations between groups in the independent variable: *room\_type*, this means our data is unbalanced. Therefore, the means plot we have displayed is using the standard deviation instead of our standard errors from the summary table. This is because the plot using standard deviations allows us to gain better insight into the assumptions of equal variance. This is because the standard error is not a measure of variability of the data itself.

### Diagnostics

### With the above results we can fit a full nonadditive model (Figure 21). Overall, we can see the residuals show close to an even distribution, Figure 22 (left graph). We identified a near random cloud of residuals around the 0 x-axis. There is a slight unevenness in the difference of distribution, however we move forward with the assumption met as this is close to normality and while a logged transformation may bring it closer to a random cloud, it is also not perfect, Figure 24 (left graph). Since the log-linear model is more towards normality, we will build this model for prediction and reference. However, since the non-transformed data allows for a more practical interpretation, we will be building this model as well for our conclusions.

### There is a small concern for the violation of normality. In Figure 22 (right graph), we can see that the data is slightly skewed right. There is tail reaching to the right side of the histogram, however the majority of the data is distributed around zero. In reviewing this further we can see the same results in our QQ plot, Figure 22 (middle graph), while we would prefer a perfectly straight line, this is close with only a we would expect to see this as close to a straight line as possible. We moved forward with the assumption passed after invoking the Central Limit Theorem due to the large sample size and moved forward with the analysis containing this slight departure from normality. With the log-linear model (Figure 24) we can see a near perfect distribution in both the histogram and QQ plot. Again, for predicting we would use the log-linear model, but for interpretation we will use the non-transformed data. We will continue to build these models side by side.

### There are no extreme outliers present in our Cook’s D / Leverage graph (Figure 23). All of the points on the far right still fall below our dashed line and also fall well below our level of leverage of 1. The same results can be said for the log-linear model (Figure 24).

### Lastly, to address the independence assumption, it is important to note that New York City is a place of expensive real estate. The boroughs (neighborhoods) contained within the city contribute to how expensive real estate may or may not be. This would seem to violate our assumption of independence; however, since we are saying the neighborhood a house is in would often determine the price and therefore likely the rental price. But since all real estate is going to be dependent on the real estate around it, all modeling for pricing would have to have assumed independence Therefore, we move forward cautiously with the assumption of independence met.

### Testing

### For our testing we performed a high-level ANOVA that tested to see if the individual metrics, as well as their interaction show a significant p-value at the 0.05 alpha level. To do this we completed a Type-III Sums of Squares F-test on both the unlogged (Figure 25) and logged data (Figure 26); our results show that both our independent metrics *neighbourhood\_group,* *room\_type*, as well as our interaction between them *neighbourhood\_group\* room\_type* are overwhelmingly statistically significant at the alpha 0.05 level. The F-values and p-values for unlogged data are: [622.677, <.0001]; [98.739, <.0001], and [20.121, <.0001] respectively (Figure 25). Since the interaction term is significant at the 0.05 alpha level, we have confirmed our assumption that this model is nonadditive and reject our H0, determining solidifying that the neighbourhood depends on room type (and vice versa) when it comes to predicting price of an Airbnb in NYC. While the F-values for the logged model are higher, we can see there is no discernable difference for the p-values (Figure 25 versus Figure 26).

### Since our interactions have proven significant, we moved forward with Tukey-Kramer (since all assumptions have been met in our model) pairwise comparisons of every variation of the *neighbourhood\_group* and *room\_type* variables paired. This will tell us which pairs of the *neighbourhood\_group* and *room\_type* are significant since the Type III SS test only tells us that they are significant in general.

### Examining the adjusted p-values and confidence intervals, there are multiple comparisons that yields overwhelmingly statistically significant results with p-values < 0.05 alpha level (at a 95% confidence level). These comparisons can be found in Figure 28. There are so many groups we recommend referencing the plot found in Figure 29, any confidence internal shown that does not cross the zero threshold is showing a statistically significant p-value at the 0.05 alpha level. The major difference in the Tukey Kramer adjustment is that even more pairs are found significant with the logged data (Figure 26 and Figure 27). Keeping our data unlogged allows us to be a little more cautious with how many significant pairs there are and below we move into the interpretation of our unlogged data model.

## Interpretation

## When we ask ourselves how much something is going to cost, we often research those similar items on our own to have an educated guess on the amount of money we are going to spend. We have done just this with our two-way ANOVA for NYC Airbnb customers. In this analysis we used two highly significant categorical variables that affect the dependent variable: price of an Airbnb in NYC based on neighborhood and type of rental. Our immediate conclusions tell us yes, with overwhelming statistical evidence, the price of your Airbnb will be determined by neighborhood and the type of rental you choose. Out of 105 comparisons 69 (Figure 28) were deemed statistically significant in our pairwise comparison of the different neighborhoods and room types with p-values <= 0.01 which is less than the alpha level of 0.05 (Figure 29).

## Differently stated, the practical significant in taking into account the borough and room type you rent will highly affect the amount of money you spend on your Airbnb in NYC. For example, in Figure 30 we have ordered the difference in price mean from lowest to highest. This is showing us that the difference in money you can save by renting a shared room in the Bronx as compared to an entire home/apt in Manhattan can be upwards of $147.56 dollars a night! This is huge savings when you want to spend more money on your experiences and not where you might be storing your luggage. Overall, you can see that any shared room in a neighborhood other than Manhattan is showing us a great saving in dollars spent. I would note that if a shared room is something you can handle (Figure 30); this is how you would save the most money when renting and Airbnb in NYC. Secondly, you can see that shared and private rooms in any neighborhood other than Manhattan are showing a significant difference in mean price of the Airbnb. This means that in general, if you want to save money while staying at an Airbnb in NYC, you might want to consider staying outside of the city center and commuting into Manhattan for all your tourist needs.

# Appendix

## Figure 1

## A screenshot of a cell phone Description automatically generated

## Figure 2

## A picture containing screenshot Description automatically generated

## Figure 3

## A picture containing screenshot Description automatically generated

## Figure 4

## A picture containing screenshot Description automatically generated

## Figure 5

## A screenshot of a social media post Description automatically generated

## Figure 6

## A screenshot of a social media post Description automatically generated

## Figure 7

## A screenshot of a social media post Description automatically generated

## Figure 8

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Test Adj. R2** | **RMSE** | **MAE** | **Error Rate** | **Full Adj. R2** | **AIC** | **BIC** | **RSS** |
| Linear intuitive | 0.4574 | 57.06 | 40.53 | 44.92% | 0.4551 | 265,318 | 265,448 | 79,902,332 |
| Log intuitive | 0.5521 | 0.3963 | 0.3093 | 8.49% | 0.5675 | 23,365 | 23,495 | 3,716 |
| Linear forward | 0.4570 | 57.08 | 40.66 | 44.94% | 0.4550 | 265,321 | 265410 | 79,941,811 |
| Log forward | 0.5518 | 0.3965 | 0.3091 | 8.49% | 0.5667 | 23,408 | 23,497 | 3,724 |
| Linear backward | 0.4570 | 57.08 | 40.66 | 44.94% | 0.4550 | 265,321 | 265410 | 79,941,811 |
| Log backward | 0.5518 | 0.3965 | 0.3091 | 8.49% | 0.5667 | 23,408 | 23,497 | 3,724 |
| Linear seqrep | 0.4570 | 57.08 | 40.66 | 44.94% | 0.4550 | 265,321 | 265410 | 79,941,811 |
| Log seqrep | 0.5518 | 0.3965 | 0.3091 | 8.49% | 0.5667 | 23,408 | 23,497 | 3,724 |
| Linear Ridge | .462518 | 56.88104 | N/A | N/A | N/A | N/A | N/A | N/A |
| Log Ridge | .557224 | .394462 | N/A | N/A | N/A | N/A | N/A | N/A |
| Linear LASSO | .463519 | 56.7406 | N/A | N/A | N/A | N/A | N/A | N/A |
| Log LASSO | .557558 | .3939339 | N/A | N/A | N/A | N/A | N/A | N/A |
| Linear Elastic Net | .463512 | 56.74125 | N/A | N/A | N/A | N/A | N/A | N/A |
| Log Elastic Net | .557551 | .3939288 | N/A | N/A | N/A | N/A | N/A | N/A |

## Figure 9 – Linear Intuitive Model

## 

## 

## Figure 10 – Log – Linear Intuitive Model

## 

## 

## 

## Figure 11 – Linear Forward, Backward, and Seqrep Selection Models (All Produce Same Plots)

## 

## 

## 

## 

## Figure 12 – Log – Linear Forward, Backward, and Seqrep Selection Models (All Produce Same Plots)

## 

## 

## 

## 

## Figure 13 - Linear Ridge Model, glmnet() alpha=1

|  |  |
| --- | --- |
|  |  |

## Figure 14 - Log-Linear Ridge Model, glmnet() alpha=1

|  |  |
| --- | --- |
|  |  |

## Figure 15 - Linear LASSO Model, glmnet() alpha = 0

|  |  |
| --- | --- |
|  |  |

## Figure 16 - Log-Linear LASSO Model, glmnet() alpha = 0

|  |  |
| --- | --- |
|  |  |

## Figure 17 - Linear Elastic Net Model, glmnet() chose alpha=.1

|  |  |
| --- | --- |
|  |  |

## Figure 18 - Log-Linear Elastic Net Model, chose alpha=.1

|  |  |
| --- | --- |
|  |  |

## Figure 19

## A screenshot of a cell phone Description automatically generated

## Figure 20

## A screenshot of a cell phone Description automatically generated

## A screenshot of a cell phone Description automatically generated

## Figure 21

## A close up of a map Description automatically generated

## Figure 22

## A screenshot of a social media post Description automatically generated

## A close up of a device Description automatically generated

## Figure 23 A screenshot of a cell phone Description automatically generated

## Figure 24

## A close up of a logo Description automatically generated

## A screenshot of a social media post Description automatically generated

## A picture containing wall, indoor, sky, next Description automatically generated

## A close up of a map Description automatically generated

## Figure 25

## A screenshot of a cell phone Description automatically generated

## Figure 26

## A screenshot of a social media post Description automatically generated

## Figure 27

## A screenshot of a cell phone Description automatically generated

## Figure 28

## A screenshot of a cell phone Description automatically generated

## A close up of text on a white background Description automatically generated

## A picture containing text, newspaper Description automatically generated

## A close up of a newspaper Description automatically generated

## A close up of a newspaper Description automatically generated

## Figure 29

## A screenshot of a social media post Description automatically generated

## Figure 30

## A screenshot of a social media post Description automatically generated