|  |  |
| --- | --- |
| A sunset over a body of water with a city in the background  Description automatically generated  nyc Airbnbs  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 20](#_Toc32155455)

[Who doesn’t love a vacation? I think it is safe to say, no one. In a recent report by Eventbrite “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. 20](#_Toc32155456)

[New York City is the city that never sleeps. The big apple. Destination New Yorker 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. 20](#_Toc32155457)

[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, backwards, seqreq, and XXXXXX 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. 20](#_Toc32155458)

[2. Data description 20](#_Toc32155459)

[The data we used was provided by a Kaggle competition but originates from Airbnb’s owned data. 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’s available across the major NYC borrows. Please reference the below table for a full list of the variables and their description. 20](#_Toc32155460)

[Variable Name 20](#_Toc32155461)

[Variable Description 20](#_Toc32155462)

[id 20](#_Toc32155463)

[Listing ID 20](#_Toc32155464)

[name 20](#_Toc32155465)

[Name of the listing 20](#_Toc32155466)

[host\_id 20](#_Toc32155467)

[Host ID 20](#_Toc32155468)

[host\_name 20](#_Toc32155469)

[Name of the host 20](#_Toc32155470)

[neighbourhood\_group 20](#_Toc32155471)

[Location 20](#_Toc32155472)

[neighbourhood 20](#_Toc32155473)

[Area 20](#_Toc32155474)

[latitude 20](#_Toc32155475)

[Latitude coordinates 20](#_Toc32155476)

[longitude 20](#_Toc32155477)

[Longitude coordinates 20](#_Toc32155478)

[room\_type 21](#_Toc32155479)

[Listing space type 21](#_Toc32155480)

[price 21](#_Toc32155481)

[Price in dollars 21](#_Toc32155482)

[minimum\_nights 21](#_Toc32155483)

[Amount of nights minimum 21](#_Toc32155484)

[number\_of\_reviews 21](#_Toc32155485)

[Number of reviews 21](#_Toc32155486)

[last\_review 21](#_Toc32155487)

[Latest review 21](#_Toc32155488)

[reviews\_per\_month 21](#_Toc32155489)

[Number of reviews per month 21](#_Toc32155490)

[calculated\_host\_listings\_count 21](#_Toc32155491)

[Amount of listing per host 21](#_Toc32155492)

[availability\_365 21](#_Toc32155493)

[Number of days when listing is available for booking 21](#_Toc32155494)

[3. Exploratory data analysis (EDA) 21](#_Toc32155495)

[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. 21](#_Toc32155496)

[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. 21](#_Toc32155497)

[With further review we had a unique thought on a variable we could create distance to time 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. 21](#_Toc32155498)

[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. 21](#_Toc32155499)

[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. 22](#_Toc32155500)

[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. 22](#_Toc32155501)

[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 borrow 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. 22](#_Toc32155502)

[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 borrows 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. 22](#_Toc32155503)

[Our EDA complete we begin our model building for a multilinear regression. 23](#_Toc32155504)

[4. Objective 1: inuitive multilinear regression model 23](#_Toc32155505)

[A. Problem Statement 23](#_Toc32155506)

[XXXXXXXXXXXXXXXXX 23](#_Toc32155507)

[B. Model Building 23](#_Toc32155508)

[C. Comparing Models 23](#_Toc32155509)

[We worked on comparing a large group of different models using intuitiveness and multiple automatic selection methodologies offered through software packages such as: forward, backward, seqrep, and LASSO/LARS. The strength of these different models can be seen below in our adjusted r2 23](#_Toc32155510)

[i. Forward (Figure XX) 23](#_Toc32155511)

[ii. Backward (Figure XX) 23](#_Toc32155512)

[iii. Seqrep (Figure XX) 23](#_Toc32155513)

[iv. LASSO 23](#_Toc32155514)

[v. LARS 23](#_Toc32155515)

[vi. Adj R2 Chart between tried models (Figure XX) 23](#_Toc32155516)

[Based on our above adjusted r2 chart we can conclude that the strongest model is formed by our intuitive model. We acknowledge the adjusted r2 for our intuitive model may be larger than those others shown, however it is still not as strong as we would like. This again supports the possible conclusion that a multiple linear regression model may not be the best fit for our data. 23](#_Toc32155517)

[D. Assumptions: Intuitive 23](#_Toc32155518)

[i. Residuals 23](#_Toc32155519)

[ii. Constant Variance 23](#_Toc32155520)

[iii. Independence 23](#_Toc32155521)

[iv. Multicollinearity 23](#_Toc32155522)

[v. Outliers: Leverage / Cook’s D 24](#_Toc32155523)

[E. Interpretation: Intuitive Model 24](#_Toc32155524)

[Interpretation REQUIRED 24](#_Toc32155525)

[Confidence Intervals REQUIRED 24](#_Toc32155526)

[Multiple linear regression is just one option in building a predictive model for a continuous response. In our analysis we are see that a MLR 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. 24](#_Toc32155527)

[Since we surfaced the above facts about out data and our data is large, we think that other methods such as Random Forest of K-NN would perform better. These options are less time consuming because the models’ complexity is built into software algorithms. We would also not have to specify how a relationship exists ahead of time, like we have for MLR. 24](#_Toc32155528)

[This being state we have seen a strong relationship between the borrow 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 9). In our next section we explore what we surface with our Two-Way ANOVA. 24](#_Toc32155529)

[Notes from meeting with Turner 24](#_Toc32155530)

[Might need to bin (r bin a continuous variable)? 24](#_Toc32155531)

[Unit 2 slides – building blocks slide – add interactions slide 24](#_Toc32155532)

[Limitations due to variables, what are some other variables that could be collected and include those in the conclusion 24](#_Toc32155533)

[With a lot of predictors, you can go through feature selection 24](#_Toc32155534)

[Since there aren’t many, it might be necessary to include interaction terms in our MLR 24](#_Toc32155535)

[Include interactions between continuous because if you don’t include complexity, then you will not predict as well 24](#_Toc32155536)

[You have to add your own complexity by adding quadratic, or interaction terms. 24](#_Toc32155537)

[5. Objective 2: two-way analysis of variance 24](#_Toc32155538)

[A. Problem Statement & Methodology 24](#_Toc32155539)

[We are testing the ability to predict the prices of Airbnb listings in NYC based on the variables provided per unit by Airbnb. After completing a full exploratory data analysis (EDA) and then taking into account practically useful information we determined the two best factors to determine the price were: *neighbourhood\_group, and room\_type* (reference code in Figure 10). 25](#_Toc32155540)

[B. Analysis 25](#_Toc32155541)

[i. Two Factor, Two-Way Analysis of Variance (ANOVA) 25](#_Toc32155542)

[With our two-way ANOVA, we are in a two independent categorical variables scenario. With our cleaned data we created a summary containing: *mean, standard deviation, standard error, minimum, maximum, and inner quartile range* of our dependent variable: *price* (reference code in Figure 10). 25](#_Toc32155543)

[ii. Mean Profile Plotting 25](#_Toc32155544)

[Once this was completed, we used the table to create a mean profile plot (Figure 11). In seeing this we can determine this will be a nonadditive model. Nonadditive means we can see interactions between our independent variables. We can see interactions in our mean profile graph (Figure 11) by seeing lines start to converge into a single point, implying these would eventually cross and therefore representing an interaction. These interactions are distinct because the different represented *neighbourhood\_group* graph lines cross each other in the graph displayed. 25](#_Toc32155545)

[Our data set has unequal in groups, 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. If the data set were more balanced with an equal number of samples in each of the independent variables, we would have been able to leverage the *standard error* instead of the *standard deviation*. 25](#_Toc32155546)

[With the two observations, there are two categorical variables available and they contain an interaction; we will be moving forward modeling a nonadditive two-way ANOVA. 25](#_Toc32155547)

[iii. Diagnostics 25](#_Toc32155548)

[With the above results we can fit a full nonadditive model (Figure 12). There is a high concern for the violation of constant variance, as well as normality. In Figure 12 (right graph), we can see that the data is skewed right. There is a long tail leaving the right side of the graph, showing the majority of the data bunched towards the left of the histogram. To further support our violation of the normality assumption the QQ plot is showing a curved hook, Figure 12 (middle graph), we would expect to see this as close to a straight line as possible. 25](#_Toc32155549)

[Overall, we can see the residuals assumption is violated. The residuals versus fitted graph, Figure 12 (left graph), suggests that the variance trends to increase. 25](#_Toc32155550)

[With the listed assumptions violated we move forward with a transformation to try and reach data normality. 26](#_Toc32155551)

[iv. Log-Linear Transformation 26](#_Toc32155552)

[Due to the above assumptions being violated we decided to move forward with a log transformation on our dependent variable. This is changing our model to a log-level two-way analysis of variance regression. 26](#_Toc32155553)

[[Please reference Figure 13 for code changing *price* to a logged *price*.] 26](#_Toc32155554)

[v. Diagnostics 26](#_Toc32155555)

[With the log transformation complete we refitted the nonadditive model (Figure 14). Overall, we can see the residual diagnostics show a more even distribution, Figure 14 (left graph). We noted a near random cloud of residuals around the 0 x-axis. There is a slight unevenness in the difference of distribution (+4, -2), however we move forward with the assumption met as this is much close to normality and allows for modeling. 26](#_Toc32155556)

[The normality of data assumption is distributed near evenly in the histogram, Figure 14 (right graph). With this new distribution of data, we will move forward with the normality assumption met. To further support the assumptions met, we see a QQ plot also showing a near straight line, post log transformation, Figure 14 (middle graph). With these small deviations from expected results there is lower concern with the constant variance assumption, and we move forward with the assumption met. 26](#_Toc32155557)

[There are no extreme outliers present in the residuals graph. However, we did note a grouping of larger values, since this isn’t a singular point (or singular few points), we know the higher values are valid measurements and we would leave it to form a proper model. 26](#_Toc32155558)

[Lastly, to address the independence assumption, it is important to note that New York City is a place of expensive real estate. The borrows (neighborhoods) contained within the city contribute to how expensive real estate may or may not be. This would in sense, seem to violate our assumption of independence, 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 in order to model for price. Therefore, we move forward cautiously with the assumption of independence. 26](#_Toc32155559)

[vi. Testing 26](#_Toc32155560)

[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; our results show that both our individual metrics *neighbourhood\_group,* *room\_type*, as well as our interaction *neighbourhood\_group\* room\_type* are overwhelmingly statistically significant at the alpha 0.05 level. The F-values and p-values are: [635.165, <.0001]; [239.342, <.0001], and [12.493, <.0001] respectively (Figure 15). 26](#_Toc32155561)

[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. 27](#_Toc32155562)

[Examining the adjusted p-values and confidence intervals, there are multiple (if not near all) comparison that yields overwhelmingly statistically significant results with p-values < 0.05 alpha level (at a 95% confidence level). These groups can be found in Figure 16. There are so many groups we recommend referencing the plot found in Figure 17, any confidence internal shown that does not cross the zero threshold is showing a statistically significant p-value at the 0.05 alpha level. All non-significant pairs have been called out in Figure 18. 27](#_Toc32155563)

[C. Interpretation 27](#_Toc32155564)

[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 have used two categorical variables to predict the price of an Airbnb in NYC based on neighborhood and type of rental. Our immediate conclusions tell us yes, with over whelming evidence, the price of your Airbnb will be determined by neighborhood and the type of rental you choose. Almost every pairwise comparison of the different neighborhoods and room types are significant with p-values < 0.0001 which is less than the alpha level of 0.05 (Figure 17). 27](#_Toc32155565)

[In order to bring some practical significant to this data, we are going to highlight the highest difference paired groups and help you know what that means. As you can see in Figure 19, we have taken the back-log transformation of the *diff* column provided by our XXXXXX 27](#_Toc32155566)

[i. Notes from meeting with Turner 27](#_Toc32155567)

[ii. There’s so much sample size that everything is significant. 27](#_Toc32155568)

[iii. \*\*Practical Significance 27](#_Toc32155569)

[6. Appendix 29](#_Toc32155570)

[A. Figure 1 29](#_Toc32155571)

[A screenshot of a cell phone

Description automatically generated 29](#_Toc32155572)

[B. Figure 2 29](#_Toc32155573)

[A picture containing screenshot

Description automatically generated 29](#_Toc32155574)

[C. Figure 3 30](#_Toc32155575)

[A picture containing screenshot

Description automatically generated 30](#_Toc32155576)

[D. Figure 4 30](#_Toc32155577)

[A picture containing screenshot

Description automatically generated 30](#_Toc32155578)

[E. Figure 5 31](#_Toc32155579)

[A screenshot of a social media post

Description automatically generated 31](#_Toc32155580)

[F. Figure 6 31](#_Toc32155581)

[A screenshot of a social media post

Description automatically generated 31](#_Toc32155582)

[G. Figure 7 32](#_Toc32155583)

[A screenshot of a social media post

Description automatically generated 32](#_Toc32155584)

[H. Figure 8 32](#_Toc32155585)

[I. Figure 9 32](#_Toc32155586)

[J. Figure 10 33](#_Toc32155587)

[K. Figure 11 33](#_Toc32155588)

[L. Figure 12 33](#_Toc32155589)

[M. Figure 13 33](#_Toc32155590)

[N. Figure 14 33](#_Toc32155591)

[O. Figure 15 33](#_Toc32155592)

[P. Figure 16 33](#_Toc32155593)

[Q. Figure XX A screenshot of a cell phone

Description automatically generated 33](#_Toc32155594)

[R. Figure XX 33](#_Toc32155595)

[A screenshot of a cell phone

Description automatically generated 33](#_Toc32155596)

[A screenshot of a cell phone

Description automatically generated 34](#_Toc32155597)

[S. Figure XX 34](#_Toc32155598)

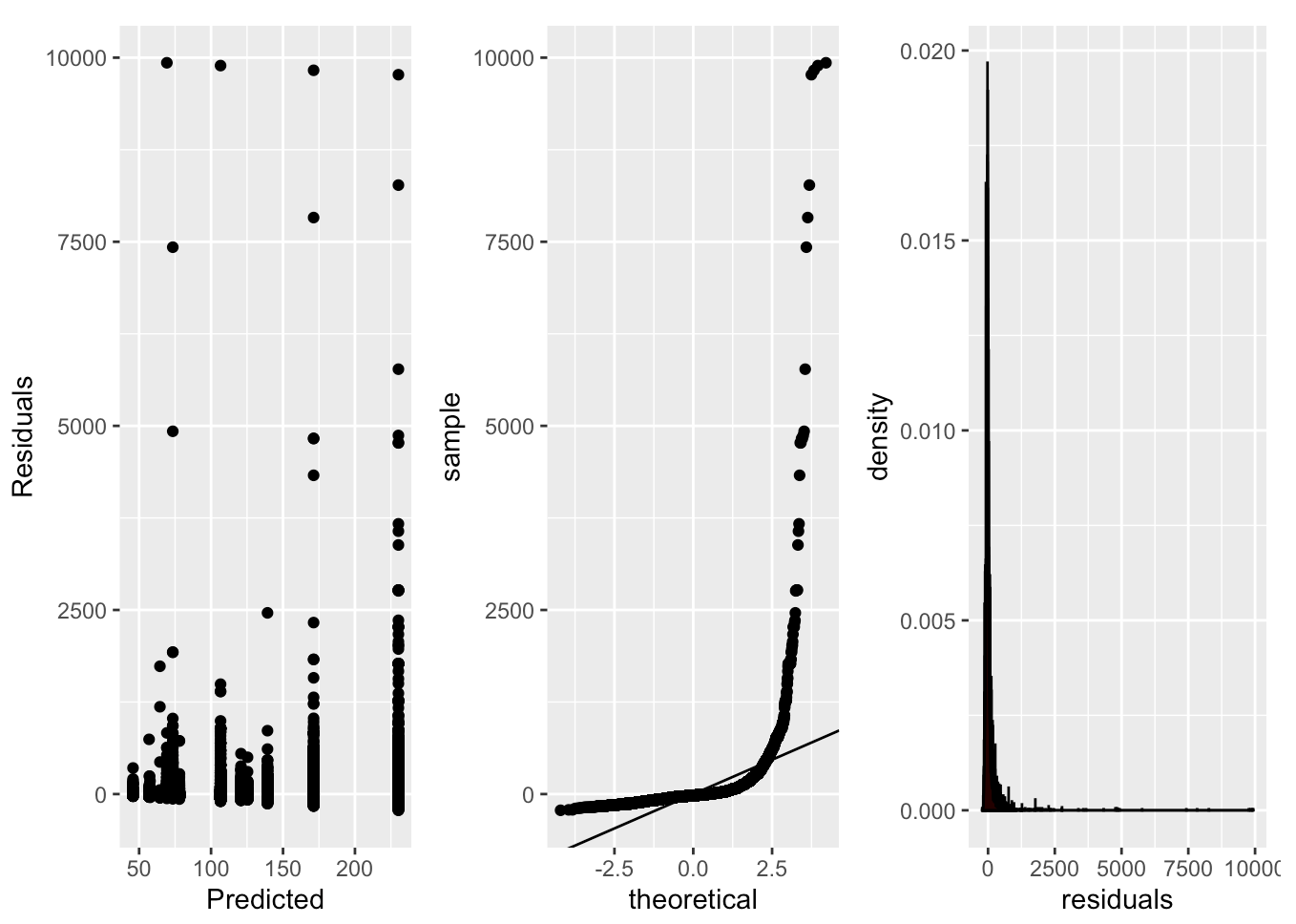
[A screenshot of a map

Description automatically generated 34](#_Toc32155599)

[T. Figure XX 34](#_Toc32155600)

[A screenshot of a social media post

Description automatically generated 35](#_Toc32155601)

[ 35](#_Toc32155602)

[U. Figure XX 35](#_Toc32155603)

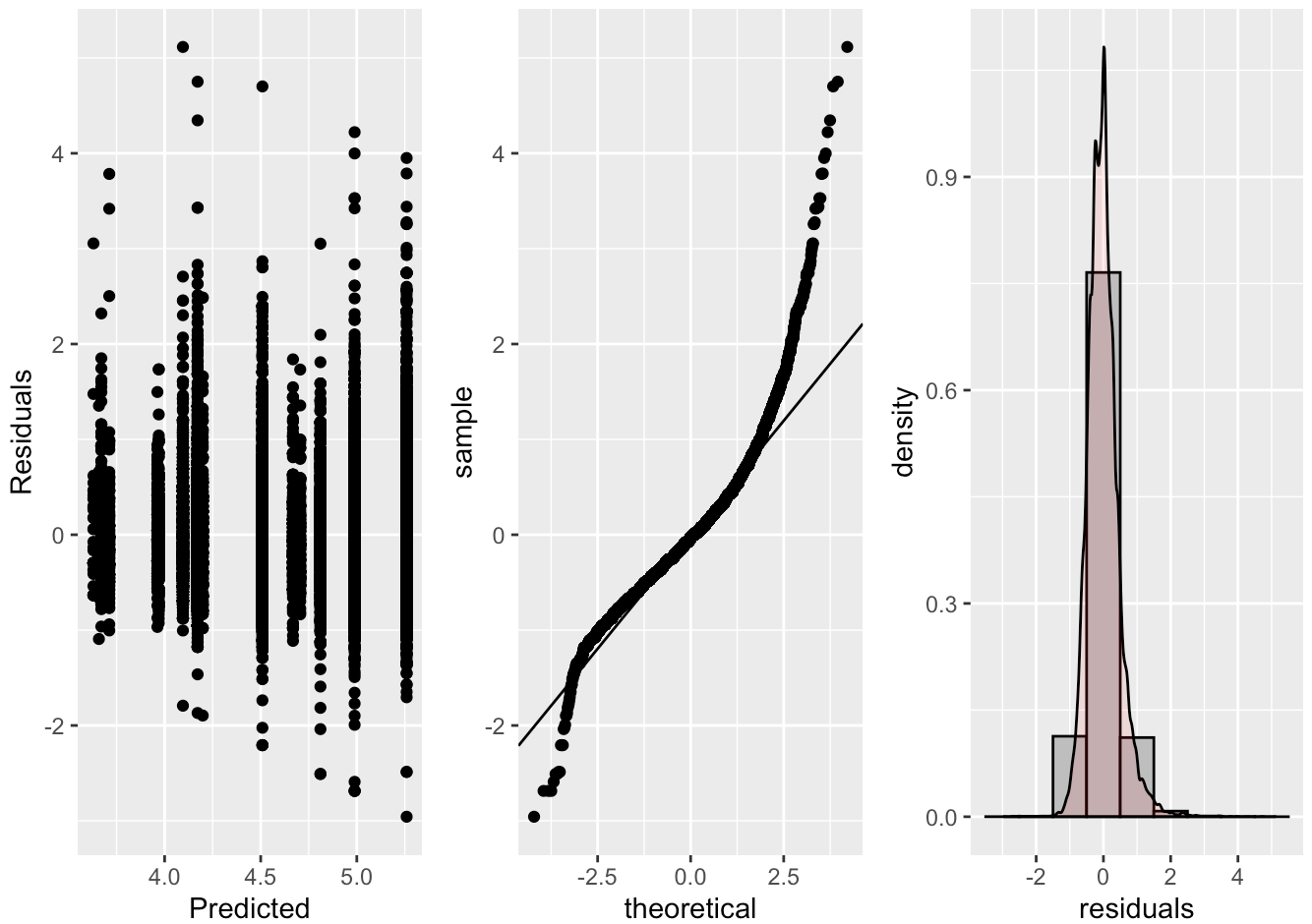
[A close up of a logo

Description automatically generated 35](#_Toc32155604)

[V. Figure XX 35](#_Toc32155605)

[A screenshot of a social media post

Description automatically generated 36](#_Toc32155606)

[ 36](#_Toc32155607)

[W. Figure XX 36](#_Toc32155608)

[A screenshot of a cell phone

Description automatically generated 36](#_Toc32155609)

[X. Figure XX 37](#_Toc32155610)

[A screenshot of a social media post

Description automatically generated 37](#_Toc32155611)

[A close up of a piece of paper

Description automatically generated 37](#_Toc32155612)

[A close up of a newspaper

Description automatically generated 37](#_Toc32155613)

[Y. Figure XX 38](#_Toc32155614)

[A screenshot of a cell phone

Description automatically generated 38](#_Toc32155615)

[Z. Figure XX 38](#_Toc32155616)

[neighborhood w room type 38](#_Toc32155617)

[diff 38](#_Toc32155618)

[lower 38](#_Toc32155619)

[upper 38](#_Toc32155620)

[padj 38](#_Toc32155621)

[Staten Island:Privateroom v Bronx:Privateroom 38](#_Toc32155622)

[-0.006676979 38](#_Toc32155623)

[-0.14234533 38](#_Toc32155624)

[0.128991371 38](#_Toc32155625)

[1.0000 38](#_Toc32155626)

[Staten Island:Entirehome apt Bronx:Entirehome apt 38](#_Toc32155627)

[0.009552579 38](#_Toc32155628)

[-0.14087822 38](#_Toc32155629)

[0.159983374 38](#_Toc32155630)

[1.0000 38](#_Toc32155631)

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, backwards, seqreq, and XXXXXX 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’s available across the major NYC borrows. Please reference the below table for a full list of the variables and their description.

|  |  |
| --- | --- |
| **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 time 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 borrow 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 borrows 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: inuitive multilinear regression model

## Problem Statement

## XXXXXXXXXXXXXXXXX

## Model Building

## Comparing Models

## We worked on comparing a large group of different models using intuitiveness and multiple automatic selection methodologies offered through software packages such as: forward, backward, seqrep, and LASSO/LARS. The strength of these different models can be seen below in our adjusted r2

### Forward (Figure XX)

### Backward (Figure XX)

### Seqrep (Figure XX)

### LASSO

### LARS

### Adj R2 Chart between tried models (Figure XX)

## Based on our above adjusted r2 chart we can conclude that the strongest model is formed by our intuitive model. We acknowledge the adjusted r2 for our intuitive model may be larger than those others shown, however it is still not as strong as we would like. This again supports the possible conclusion that a multiple linear regression model may not be the best fit for our data.

## Assumptions: Intuitive

### Residuals

### Constant Variance

### Independence

### Multicollinearity

### Outliers: Leverage / Cook’s D

## Interpretation: Intuitive Model

## Interpretation REQUIRED

## Confidence Intervals REQUIRED

## Multiple linear regression is just one option in building a predictive model for a continuous response. In our analysis we are see that a MLR 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 out data and our data is large, we think that other methods such as Random Forest of K-NN would perform better. These options are less time consuming because the models’ complexity is 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 state we have seen a strong relationship between the borrow 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 9). In our next section we explore what we surface with our Two-Way ANOVA.

## Notes from meeting with Turner

## Might need to bin (r bin a continuous variable)?

## Unit 2 slides – building blocks slide – add interactions slide

## Limitations due to variables, what are some other variables that could be collected and include those in the conclusion

## With a lot of predictors, you can go through feature selection

## Since there aren’t many, it might be necessary to include interaction terms in our MLR

## Include interactions between continuous because if you don’t include complexity, then you will not predict as well

## You have to add your own complexity by adding quadratic, or interaction terms.

# Objective 2: two-way analysis of variance

Perform a secondary analysis using tools from Unit 3, Unit 4, or Unit 5.

## Problem Statement & Methodology

## We are testing the ability to predict the prices of Airbnb listings in NYC based on the variables provided per unit by Airbnb. After completing a full exploratory data analysis (EDA) and then taking into account practically useful information we determined the two best factors to determine the price were: *neighbourhood\_group, and room\_type* (reference code in Figure 10).

## Analysis

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

### With our two-way ANOVA, we are in a two independent categorical variables scenario. With our cleaned data we created a summary containing: *mean, standard deviation, standard error, minimum, maximum, and inner quartile range* of our dependent variable: *price* (reference code in Figure 10).

### Mean Profile Plotting

### Once this was completed, we used the table to create a mean profile plot (Figure 11). In seeing this we can determine this will be a nonadditive model. Nonadditive means we can see interactions between our independent variables. We can see interactions in our mean profile graph (Figure 11) by seeing lines start to converge into a single point, implying these would eventually cross and therefore representing an interaction. These interactions are distinct because the different represented *neighbourhood\_group* graph lines cross each other in the graph displayed.

### Our data set has unequal in groups, 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. If the data set were more balanced with an equal number of samples in each of the independent variables, we would have been able to leverage the *standard error* instead of the *standard deviation*.

### With the two observations, there are two categorical variables available and they contain an interaction; we will be moving forward modeling a nonadditive two-way ANOVA.

### Diagnostics

### With the above results we can fit a full nonadditive model (Figure 12). There is a high concern for the violation of constant variance, as well as normality. In Figure 12 (right graph), we can see that the data is skewed right. There is a long tail leaving the right side of the graph, showing the majority of the data bunched towards the left of the histogram. To further support our violation of the normality assumption the QQ plot is showing a curved hook, Figure 12 (middle graph), we would expect to see this as close to a straight line as possible.

### Overall, we can see the residuals assumption is violated. The residuals versus fitted graph, Figure 12 (left graph), suggests that the variance trends to increase.

### With the listed assumptions violated we move forward with a transformation to try and reach data normality.

### Log-Linear Transformation

### Due to the above assumptions being violated we decided to move forward with a log transformation on our dependent variable. This is changing our model to a log-level two-way analysis of variance regression.

### [Please reference Figure 13 for code changing *price* to a logged *price*.]

### Diagnostics

### With the log transformation complete we refitted the nonadditive model (Figure 14). Overall, we can see the residual diagnostics show a more even distribution, Figure 14 (left graph). We noted a near random cloud of residuals around the 0 x-axis. There is a slight unevenness in the difference of distribution (+4, -2), however we move forward with the assumption met as this is much close to normality and allows for modeling.

### The normality of data assumption is distributed near evenly in the histogram, Figure 14 (right graph). With this new distribution of data, we will move forward with the normality assumption met. To further support the assumptions met, we see a QQ plot also showing a near straight line, post log transformation, Figure 14 (middle graph). With these small deviations from expected results there is lower concern with the constant variance assumption, and we move forward with the assumption met.

### There are no extreme outliers present in the residuals graph. However, we did note a grouping of larger values, since this isn’t a singular point (or singular few points), we know the higher values are valid measurements and we would leave it to form a proper model.

### Lastly, to address the independence assumption, it is important to note that New York City is a place of expensive real estate. The borrows (neighborhoods) contained within the city contribute to how expensive real estate may or may not be. This would in sense, seem to violate our assumption of independence, 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 in order to model for price. Therefore, we move forward cautiously with the assumption of independence.

### 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; our results show that both our individual metrics *neighbourhood\_group,* *room\_type*, as well as our interaction *neighbourhood\_group\* room\_type* are overwhelmingly statistically significant at the alpha 0.05 level. The F-values and p-values are: [635.165, <.0001]; [239.342, <.0001], and [12.493, <.0001] respectively (Figure 15).

### 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 (if not near all) comparison that yields overwhelmingly statistically significant results with p-values < 0.05 alpha level (at a 95% confidence level). These groups can be found in Figure 16. There are so many groups we recommend referencing the plot found in Figure 17, any confidence internal shown that does not cross the zero threshold is showing a statistically significant p-value at the 0.05 alpha level. All non-significant pairs have been called out in Figure 18.

## 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 have used two categorical variables to predict the price of an Airbnb in NYC based on neighborhood and type of rental. Our immediate conclusions tell us yes, with over whelming evidence, the price of your Airbnb will be determined by neighborhood and the type of rental you choose. Almost every pairwise comparison of the different neighborhoods and room types are significant with p-values < 0.0001 which is less than the alpha level of 0.05 (Figure 17).

## In order to bring some practical significant to this data, we are going to highlight the highest difference paired groups and help you know what that means. As you can see in Figure 19, we have taken the back-log transformation of the *diff* column provided by our XXXXXX

### Notes from meeting with Turner

### There’s so much sample size that everything is significant.

### \*\*Practical Significance

#### Back transform the differences

#### Take the table of difference out and rank from highest difference to lowest different and talk about the highest and cut it off for practical significance of the information

#### Do not forget to mention the drop of the ‘shared rooms’ level and how that changes the model to only be used against thec XX XX levels

#### Look at the differences for the different neighborhoods, if they are all the same then the practical importance of the information

#### Another way to tackle this would be to make the call would be to base off the graph, summary stats, and comparison are the same then we could conclude this is falling more like an additive model. Move forward with an additive model to simplify the results and interpretation.

#### Interpret a single or cherry pick one and explain how to understand it, then address/highlight the top differences and talk about their practical significance as a whole

#### \*\*\*Students tend to not do enough EDA, show graph how to illustrate those relationships

#### \*\*\*Not enough talking, the graphs/tables do not tell the story. But you do, exhibit you know what you are doing

# 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

## Figure 9

## Figure 10

## Figure 11

## Figure 12

## Figure 13

## Figure 14

## Figure 15

## Figure 16

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

## Figure XX

## A screenshot of a cell phone Description automatically generated

## A screenshot of a cell phone Description automatically generated

## Figure XX

## A screenshot of a map Description automatically generated

## Figure XX

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

## 

## Figure XX

## A close up of a logo Description automatically generated

## Figure XX

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

## 

## Figure XX

## A screenshot of a cell phone Description automatically generated

## Figure XX

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

## A close up of a piece of paper Description automatically generated

## A close up of a newspaper Description automatically generated

## Figure XX

## A screenshot of a cell phone Description automatically generated

## Figure XX

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| neighborhood w room type | diff | lower | upper | padj |
| Staten Island:Privateroom v Bronx:Privateroom | -0.006676979 | -0.14234533 | 0.128991371 | 1.0000 |
| Staten Island:Entirehome apt Bronx:Entirehome apt | 0.009552579 | -0.14087822 | 0.159983374 | 1.0000 |