Illinois Institute of Technology

MATH 564 Final Project

NYC AirBnb Statistical Analysis

Shrey Jaradi (sjaradi@hawk.iit.edu) – 33.33%
Sohaib Jawad (sjawad@hawk.iit.edu) – 33.33%
Mohammed Wasim R D (mrafeeqahameddilshad@hawk.iit.edu) – 33.33%

Prof. Lulu Kang

December 2, 2022

TABLE OF CONTENTS

Abstract

Statistical analysis is the process of gathering and analyzing data in order to find patterns and trends, remove bias, and assist decision-making. This portion of business intelligence includes the collection, analysis, and reporting of business data as well as trend reporting. Mean, standard deviation, regression, testing of hypotheses, and determining sample size are the five fundamental techniques in statistical analysis.

In this project, we are conducting statistical analysis for Airbnb data from New York. We used data from the free Kaggle platform to undertake exploratory data analysis in order to accomplish this. Our data analysis was based on the costliest neighborhood, the highest, lowest, and average listing prices, among other factors. Data that was categorical in nature was deleted, and the correlation was computed to discover which attributes are significantly connected. The Regression model was applied in the next stage, and its matching R squared value was assessed.

Due to the low R squared value, we scaled the latitude and longitude variables and applied the logarithmic transformation to the pricing. We investigate employing Random Forest and Decision Tree classifiers to increase the accuracy of price prediction. Although the precision was slightly improved by that. Good feature selection and some tunings are definitely required.

Introduction

Airbnb is an internet-based marketing company that connects people looking for lodging (Airbnb guests) with people looking to rent out their properties (Airbnb hosts) on a short-term or long-term basis. Although apartments are the most popular kind of rental property, there are still many other options available, including homes and boats. Airbnb makes money by charging hosts and guests fees for making bookings: hosts pay 3% of the overall booking value to Airbnb, while guests pay 6%–12%, depending on the type of booking. As a rental ecosystem, Airbnb generates a ton of data, including but not limited to rental density across areas (cities and neighborhoods), pricing differences across rentals, host-guest interactions in the form of reviews, and so on.

NYC Airbnb Market

New York City (NYC) has a very active Airbnb market, with more than 48,000 listings as of August of the 2019 calendar year (this corresponds to a rental density of 48000 rentals per 468 square miles, which equates to 102 rentals per square mile). The main objective of this project is to compile statistics and other useful information on NYC's Airbnb listings. The most important finding from the data is that while an Airbnb host may have multiple properties listed in a neighborhood group (the boroughs of NYC) under different host-ids, a host who has a specific property or listing in a specific neighborhood of a neighborhood group maintains the same host-id (although this is not always the case; there are exceptions where a small number of hosts maintain different ids for each listing or property in a neighborhood).

Problem Statement and Data Sources

For our project, we're aiming to provide answers to the following questions:

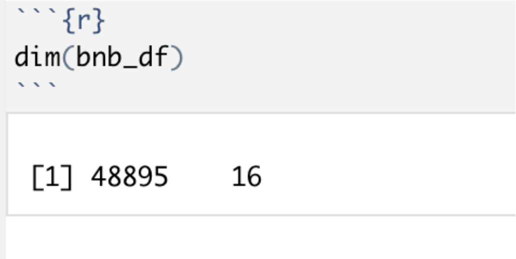
- Can we predict the accuracy or use other features to anticipate prices more accurately?
- What features are crucial for price prediction?
- What can we infer from predictions? (prices)
- Which neighborhood is the most expensive, and how much do other neighborhoods cost?

The dataset was downloaded from Kaggle, a part of Google LLC and a reputable online community for data scientists and machine learning. Data that had been downloaded was then easily accessible by uploading it to the Project GitHub Repository.

Data on listing metrics and activity in New York City for 2019 are included in the dataset. This information includes listing prices for different neighborhood groups in different local cities. It also includes additional information that may influence the listing price, such as information on the different types of properties, customer feedback, and listing availability. The dataset consists of 16 attributes and 48895 observations.

Data dimensions and summary are given below:

48895 Rows, 16 Columns



~~~
```{r}
dim(bnb_df)
```
~~~

```
[1] 48895    16
```

Feature Name	Data Types	Description
id	int	unique id
name	chr	Property name
host_id	int	unique id for each host
host_name	chr	Owner of the property
neighborhood_group	chr	Group of Counties
neighborhood	chr	Group of Particular area
latitude	num	Latitude loc of property
longitude	num	Longitude loc of property
room_type	chr	Types of room
price	int	Price of property
minimum_nights	int	Duration of night stay at property
number_of_reviews	int	Total reviews of each property
last_review	chr	Last Review date
reviews_per_month	num	Total reviews per month
calculated_host_listings_count	int	Number of properties owned by the host
availability_365	int	Availability of the property in a year

Proposed Methodology

To estimate the price, we considered using the variable of the price as our response variable and using the variable of other aspects as our predictor variable. We want to use the Regression as our basic model to predict the price, and then we want to do further regression activities to see if we can improve model performance.

Prior to modeling, there are a number of other tasks that must be completed, including data cleaning, finding the relevant feature, outlier detection and removal, checking for categorical variables and responding appropriately (using one-hot encoding, for example), checking for multicollinearity if there are any significantly correlated factors (finding the VIF value), and conducting additional analysis on the data to uncover more information that will help us answer our research questions. After doing all of this, prepare the data for modeling and run a test train split to do the modeling.

We have created 5 models for our project, including Linear Regression, Ridge Regression, Lasso Regression, and Random Forest Regressor Classifier, in order to predict the price.

The machine learning techniques that we considered include in our project are listed below.

Linear Regression:

In the first model, we will create a linear regression model to predict the price, which will be our response variable, or dependent variable. Other features will be our predictor variable, or independent variable, for the model.
And will determine which characteristics are crucial by evaluating the model's performance using the R-Square value, MSE, and significant beta values for each predictor.

In our second model, we tried to transform our price variable into a logarithmic price predictor to scale down the price and see if the logarithmic transformation will improve the performance of our model.

The new model we want to apply is a basic linear regression to see if we can predict the price using any one predictor, for instance, by predicting the price range in various neighborhoods or room types.

To see how our training data is performing on the model we developed, plot the residuals and QQ-plots for both models. Finally, we will use the testing data to evaluate the model's performance.

Decision Tree Regressor:

Techniques like multiple linear regression can perform better when there is a linear relationship between the predictor and the response variable. But nonlinear approaches perform better when the connections between them are complicated.

Like the non-linear Classification and Regression Trees (CART) approach, one illustration of CART is the use of decision trees. It is a non-parametric supervised learning technique for regression and classification.

Decision Trees represent the outcome in the form of a tree-like structure by using decisions as the characteristics.
It is a typical tool for representing the algorithm's choice visually.
If we divide the dataset in half and run a decision tree on each half, the outcomes can be very different.

The advantages of employing a decision tree regression are that they are simple to understand and display, but the disadvantage is that they frequently have large variance.

Random Forest Regressor:

To estimate the price, we intend to apply the Random Forest Regressor classifier, which is based on random forest. Since this model is non-linear, it will improvise to map the link between the predictor variable and the answer.
Using the original dataset's n bootstrapped samples, we use Random Forest as the regressor.

- Create decision trees for every sample that was bootstrapped.
  - When constructing the decision tree, only a random sample of m predictors—a subset of the whole set of p predictors—is considered as split candidates.
- To create a final model, take the average of each tree's predictions.

As a result, it tends to have less variability and produces a lower test error rate because it takes the average of the predictors of each tree.

Analysis and Results

Data Cleaning:

Except for the reviews per month column, which has 10052 N/A values, the data set is essentially clean. Due of the large number of observatories, removing them all at once would result in inaccurate results. Apart from that, the variables were all appropriately labeled and in the proper case, so they didn't need to be renamed. Each of the character variables was changed to a factor variable.

We do an investigation and make the following observations to identify the kind of listings that are typical of a particular community.

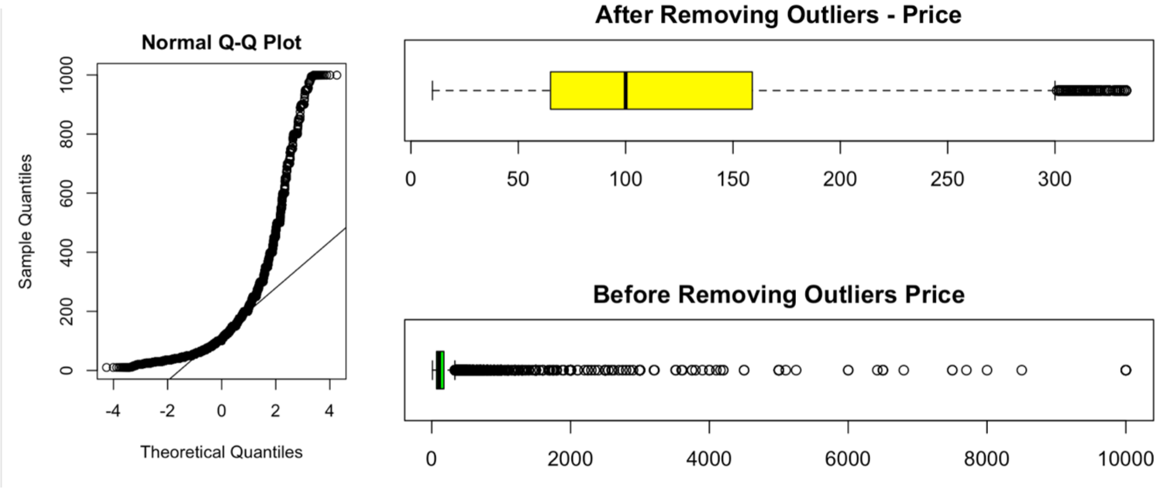
1. Private rooms are the most common sort of listing generally, with Manhattan being the exception, where full homes/apartments are the most common type of offering.

2. Across all neighborhoods', shared housing is the least prevalent.

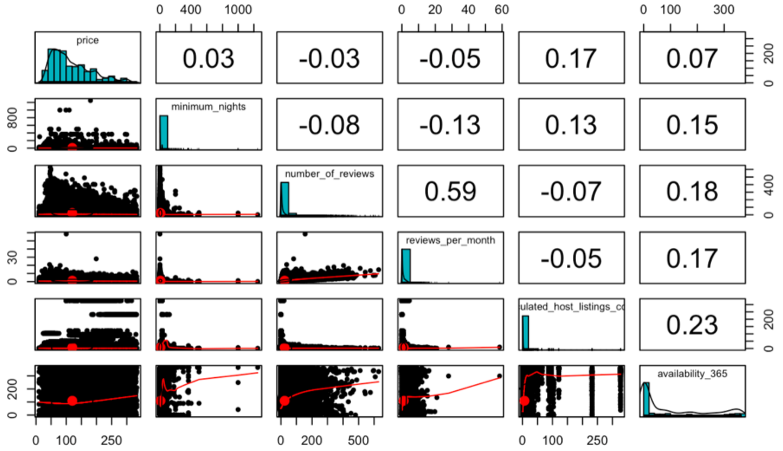
Data Preparation and Visualization:

Since our first objective is to predict the price using the predictor factors and identify certain key features, we did study our data and discovered some insightful information.

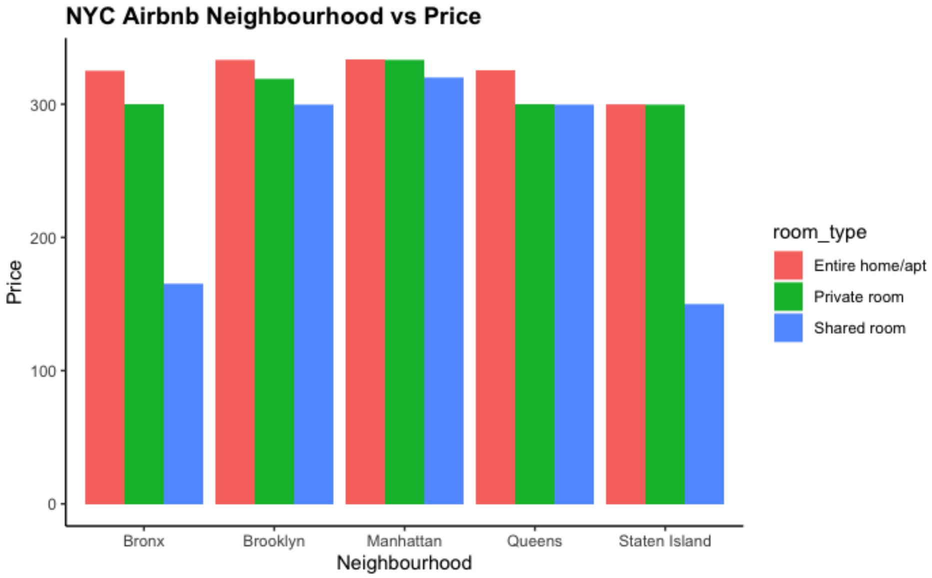
Our response variable pricing did exhibit skewness, which we discovered and eliminated.



We did plot the correlation matrix to determine how our attributes were correlated. We've just selected a select few features, including price, minimum nights, number of reviews, reviews per month, calculated host listings count, and availability 365. Because there are some more category factors that we have not considered



The housing costs for the Neighborhood Group were also compared. Three different housing options (entire home/apartment, private room, and shared room) are available in New York's several neighborhood groups, including the Bronx, Brooklyn, Manhattan, Queens, and Staten Island. The neighborhood group's price comparison for various types of lodging is shown in the chart below. As can be observed, the cost of lodging for private rooms and entire homes/apartments is almost the same throughout all neighborhood groups, while the cost of shared rooms is much lower in the Bronx and Staten Island when compared to other groups.

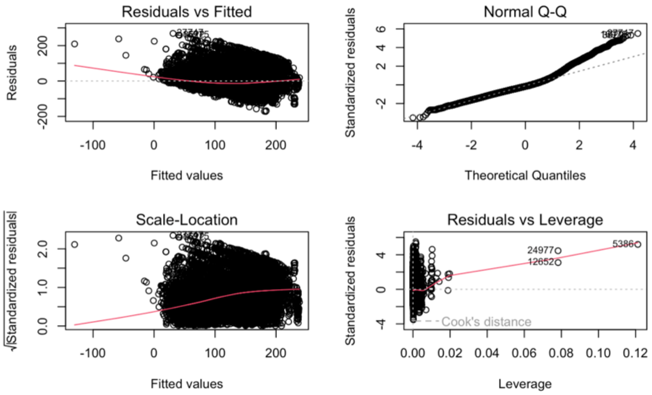


Refer to the appendix section for further plots pertaining to housing neighborhoods.

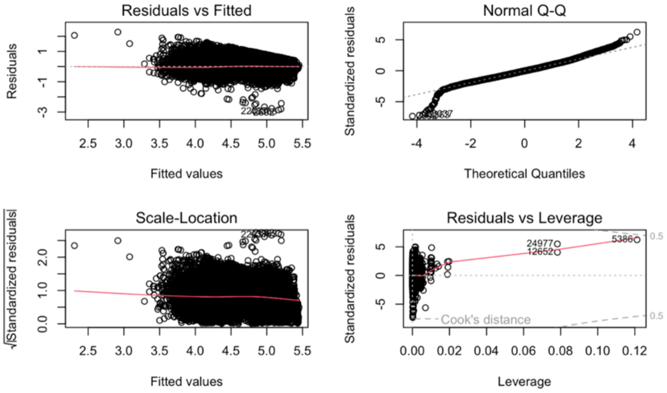
Modeling:

In the modeling process, we trained various models to test their accuracy, plotted various plots, such as the Q-Q plot and residual plot, to see how our model is behaving on the training dataset, and, in the end, used the Random Forest regressor classifier to identify the key features needed to predict the price.

When we looked at the Q-Q plot for the first multiple linear regression model, we saw that the residuals varied greatly, indicating that our data had many extreme values. And the residual vs. fitted graph demonstrates the nonlinear nature of the connection in our data.

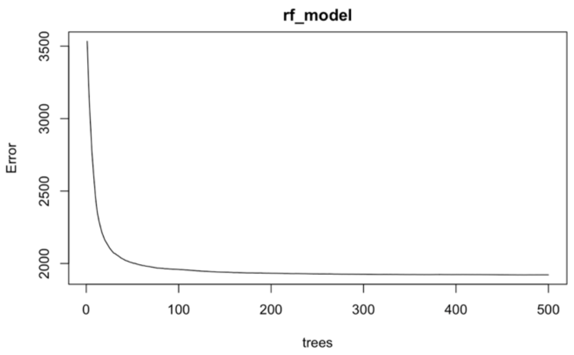


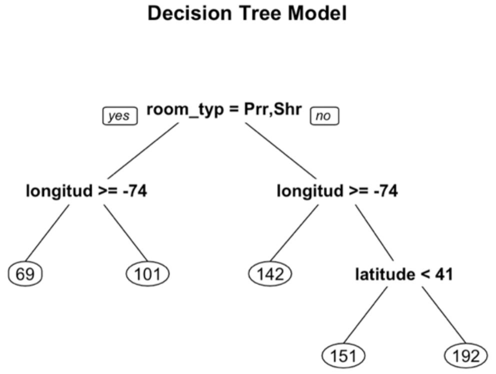
We performed the transformation and developed a different model. We changed the price response variable's logarithmic transformation, which enabled us to improve the Q-Q plot. Although there is little variation in our data, there are still some high values. Additionally, the residual vs. fitted plot demonstrates that there is a linear relationship between our data and the two variables, which is beneficial because it increased the model's accuracy.



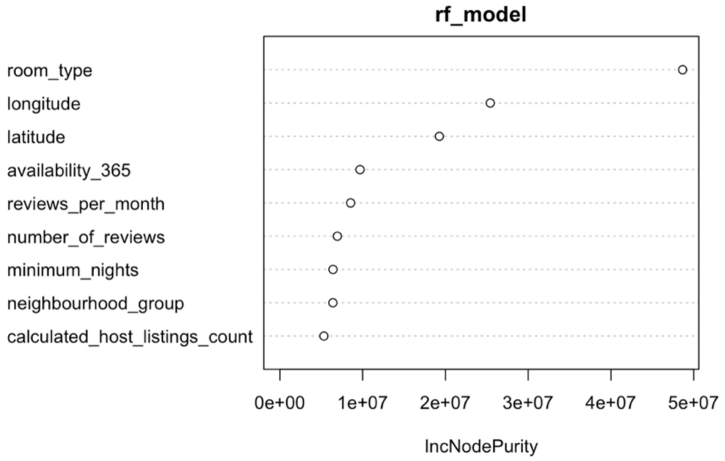
In order to test for a relationship between price and neighborhood, price and neighborhood group, and finally price and room type, we then created a variety of simple linear regression models. However, these models did not significantly improve the accuracy of the linear relationship with the price variable.

At first, we used the Decision Tree and Random Forest classifiers to try to solve problem. The data was divided up using the decision tree classifier using room type, longitude, latitude, and so on. We obtained the best R-square result with the Random Forest Classifier (0.58). However, as the plot demonstrates, the error rate lowers as more trees are added to the data.





Finally, we plot the critical variables required to estimate the price, demonstrating the significance of the features room type, longitude, latitude, and neighborhood group



Conclusion

After comparing several models, such as decision trees, random forests, decision trees, simple linear regression, and multiple linear regression.

We used all the features in our initial effort at multiple linear regression, except for neighborhood, id, name, host id, host name, and last review, and obtained an R-Square value of.4826. Although there were other statistical criteria, such as the p-value and F-statistic, that we needed to consider. Our "price" response variable was then converted into a logarithmic variable. In order to make our price variable appear more continuous, we can scale it down. This results in an R-square value of 0.5467, which is significantly higher than that of our prior model.

Using the predictor variables neighborhood, room type, and neighborhood group, we created three simple linear regression models, with respective R-square values of 0.2798, 0.4442, and 0.1338.

Then, using a decision tree, we created a model whose R-square value on the test dataset was 0.46, which isn't much better than the multiple linear regression model alone, but we thought we could increase the R-square and accuracy by using the Random Forest Regressor classifier, which gave the model its highest accuracy on the test dataset to date, with an R-square value of 0.58654. The Random Forest model might be optimized if it is tuned.

We learned about the MSE, SSE, and SSR values in addition to the R-square values because we shouldn't always rely on R-square numbers.

We sought to identify the key characteristics of our model from the Random Forest model, and it suggests that room type, longitude and latitude, and neighborhood groups are some of the key characteristics.

This leads us to the conclusion that the Airbnb dataset to predict pricing does not function well using regression since the characteristics are too complicated to obtain decent estimated predictor values. We demonstrate that we can anticipate prices using classifiers like Random Forest by utilizing the Random Forest model. Therefore, in order to accurately anticipate the price, we also need to change some data.

Different variables, such as room type, location, and neighborhood, can be used to categorize pricing, but they cannot provide an exact estimate. In order to anticipate price, we should have at least some other features connected to the property.

All the questions we posed in the problem statement were addressed in the conclusion and results, and several of them were also addressed in the analysis section.

Future Work

This project has several limitations. In our upcoming work, we should examine the last review data as well in order to break it down to the month and observe how things change with the seasons and weather. If we had a bit more knowledge about the property, we could be able to estimate the price more precisely.

With more advanced predictive modeling approaches like Support Vector Machine, tuning Random Forest Classifier, and employing some ensemble techniques, we can analyze this data more in the future.

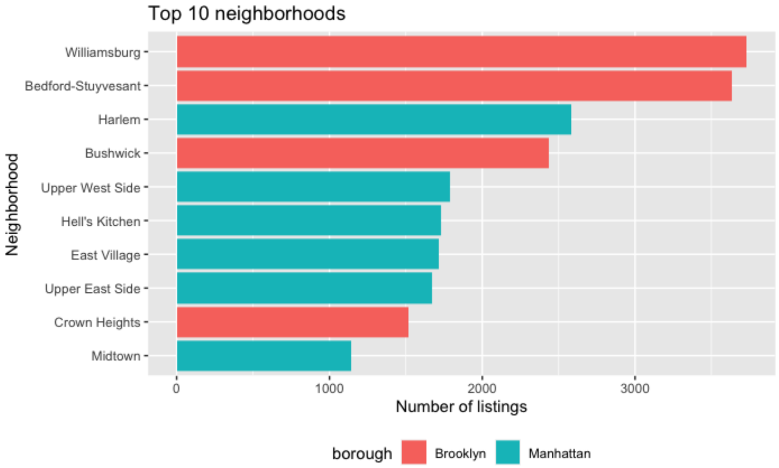
References

1. **https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data?select=AB_NYC_2019.csv**
2. **https://medium.com/almabetter/exploratory-data-analysis-on-nyc-airbnb-2019-dataset-fe908c2accaa**
3. **https://medium.com/analytics-vidhya/python-exploratory-data-analysis-eda-on-nyc-airbnb-cbeabd622e30**

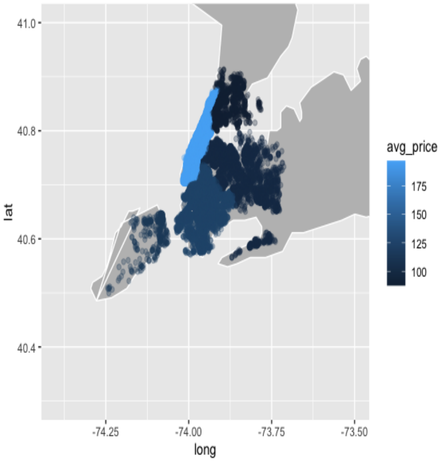
Appendix

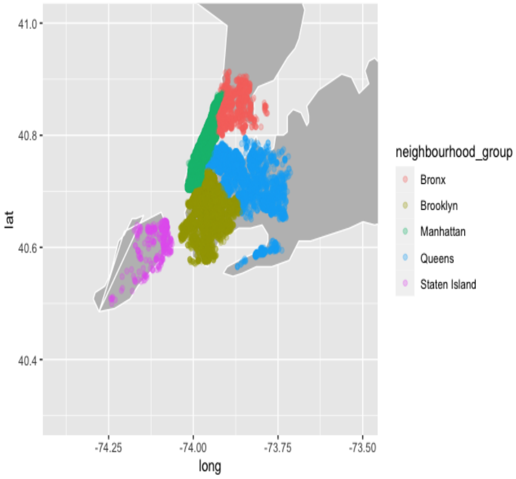
EDA:

We looked at the most popular New York neighborhoods based on the number of homes listed on Airbnb. The top 10 neighborhoods are shown in the chart below, with Williamsburg in Brooklyn being the most popular in terms of the number of listings. As can be seen, Brooklyn and Manhattan are home to the most popular neighborhoods. Brooklyn is New York City's most popular neighborhood because, while being the city's center and drawing the most visitors, Manhattan is only a short drive away and offers rentals that are comparably less expensive.



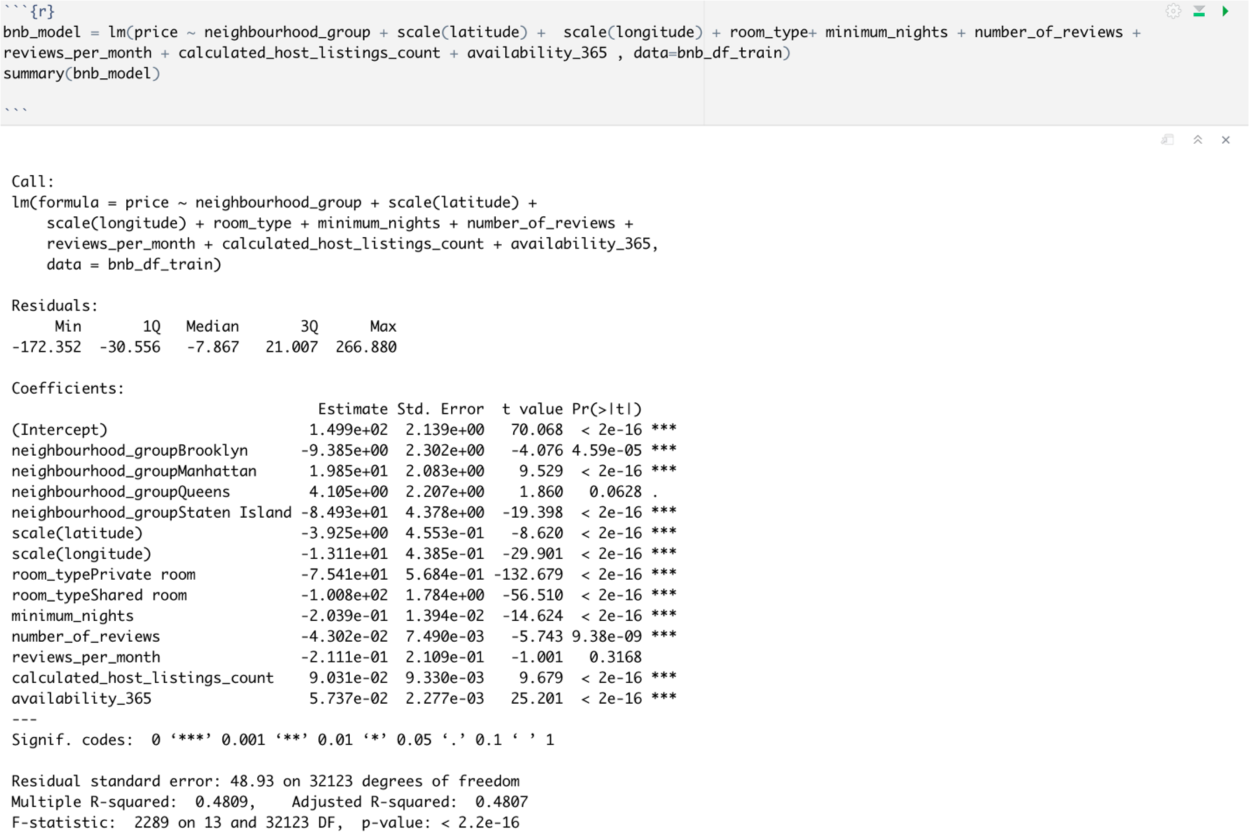
Look at the following plot of New York's prices to see how the light shade indicates high average prices and the dark shade indicates lower average prices to support the pricing argument made in the previous paragraph. We've also linked another map to the right of it so you can see where neighborhoods are located geographically in New York.





Model Codes:

Multiple Linear Regression Model



~~~
```{r}
bnb_model = lm(price ~ neighbourhood_group + scale(latitude) +  scale(longitude) + room_type+ minimum_nights + number_of_reviews +
reviews_per_month + calculated_host_listings_count + availability_365 , data=bnb_df_train)
summary(bnb_model)

```
~~~

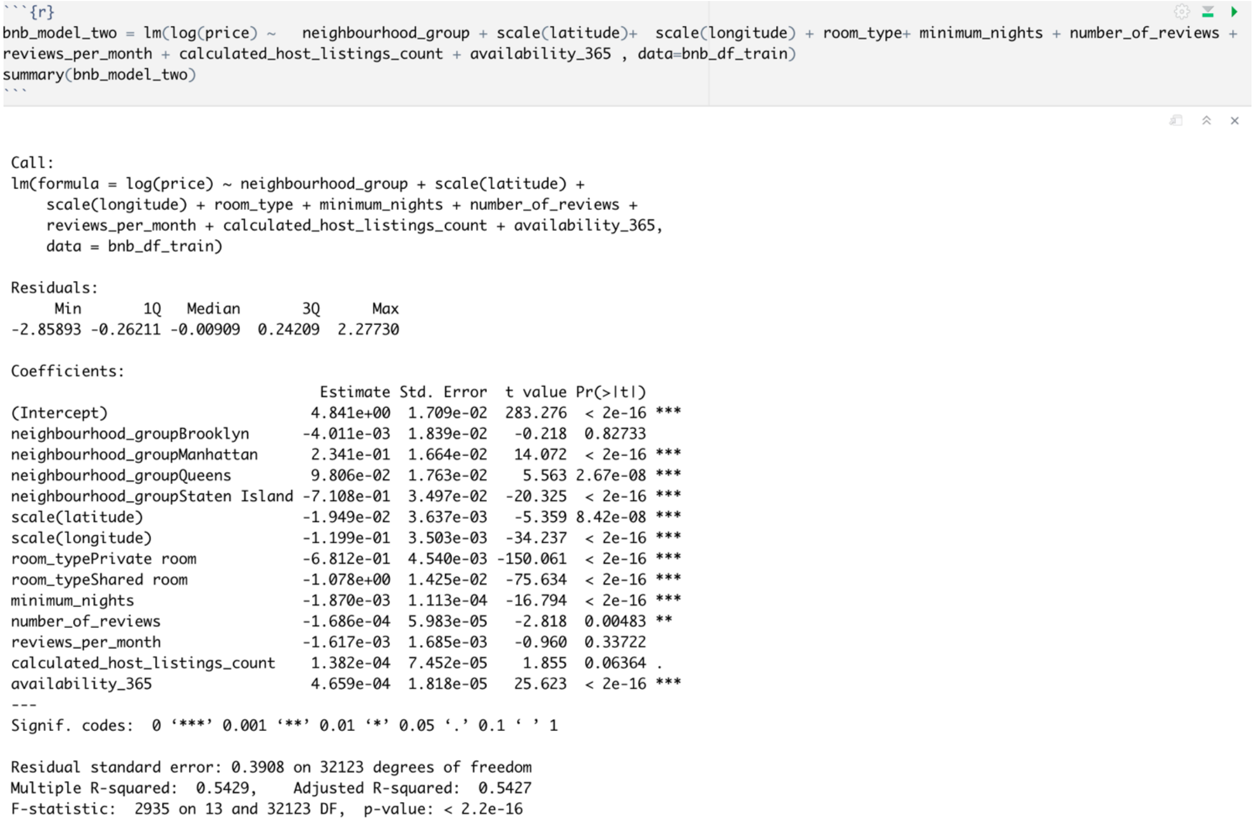
```
Call:
lm(formula = price ~ neighbourhood_group + scale(latitude) + 
    scale(longitude) + room_type + minimum_nights + number_of_reviews + 
    reviews_per_month + calculated_host_listings_count + availability_365, 
    data = bnb_df_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-172.352  -30.556   -7.867   21.007  266.880 

Coefficients:
                                   Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                       1.499e+02  2.139e+00   70.068  < 2e-16 ***
neighbourhood_groupBrooklyn      -9.385e+00  2.302e+00   -4.076 4.59e-05 ***
neighbourhood_groupManhattan      1.985e+01  2.083e+00    9.529  < 2e-16 ***
neighbourhood_groupQueens         4.105e+00  2.207e+00    1.860   0.0628 .  
neighbourhood_groupStaten Island -8.493e+01  4.378e+00  -19.398  < 2e-16 ***
scale(latitude)                  -3.925e+00  4.553e-01   -8.620  < 2e-16 ***
scale(longitude)                 -1.311e+01  4.385e-01  -29.901  < 2e-16 ***
room_typePrivate room            -7.541e+01  5.684e-01 -132.679  < 2e-16 ***
room_typeShared room             -1.008e+02  1.784e+00  -56.510  < 2e-16 ***
minimum_nights                   -2.039e-01  1.394e-02  -14.624  < 2e-16 ***
number_of_reviews                -4.302e-02  7.490e-03   -5.743 9.38e-09 ***
reviews_per_month                -2.111e-01  2.109e-01   -1.001   0.3168    
calculated_host_listings_count    9.031e-02  9.330e-03    9.679  < 2e-16 ***
availability_365                  5.737e-02  2.277e-03   25.201  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.93 on 32123 degrees of freedom
Multiple R-squared:  0.4809,	Adjusted R-squared:  0.4807 
F-statistic:  2289 on 13 and 32123 DF,  p-value: < 2.2e-16
```

Linear Regression – Logarithmic Transformation



~~~
```{r}
bnb_model_two = lm(log(price) ~   neighbourhood_group + scale(latitude)+  scale(longitude) + room_type+ minimum_nights + number_of_reviews +
reviews_per_month + calculated_host_listings_count + availability_365 , data=bnb_df_train)
summary(bnb_model_two)
```
~~~

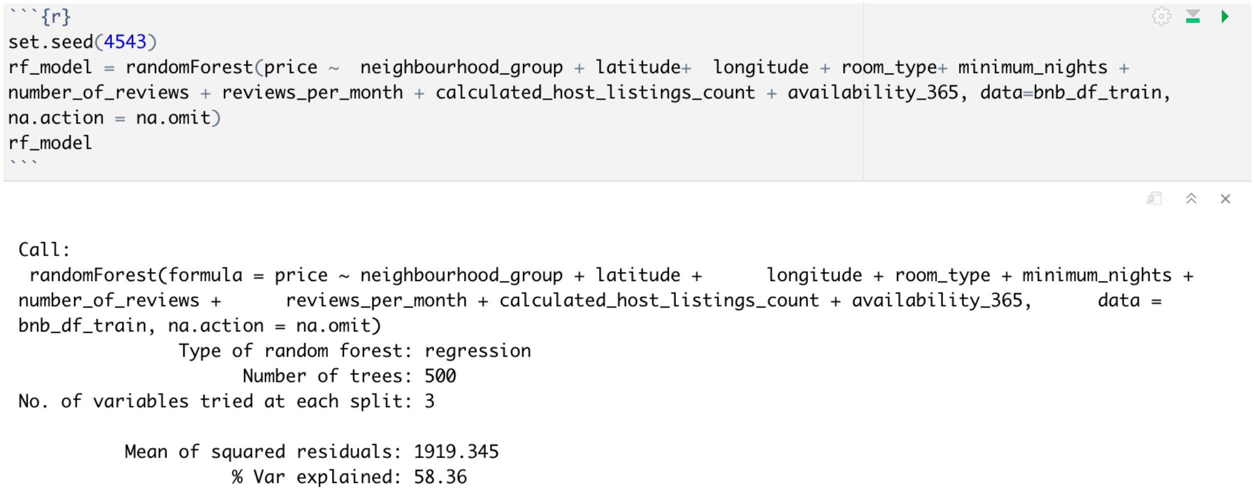
```
Call:
lm(formula = log(price) ~ neighbourhood_group + scale(latitude) + 
    scale(longitude) + room_type + minimum_nights + number_of_reviews + 
    reviews_per_month + calculated_host_listings_count + availability_365, 
    data = bnb_df_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.85893 -0.26211 -0.00909  0.24209  2.27730 

Coefficients:
                                   Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                       4.841e+00  1.709e-02  283.276  < 2e-16 ***
neighbourhood_groupBrooklyn      -4.011e-03  1.839e-02   -0.218  0.82733    
neighbourhood_groupManhattan      2.341e-01  1.664e-02   14.072  < 2e-16 ***
neighbourhood_groupQueens         9.806e-02  1.763e-02    5.563 2.67e-08 ***
neighbourhood_groupStaten Island -7.108e-01  3.497e-02  -20.325  < 2e-16 ***
scale(latitude)                  -1.949e-02  3.637e-03   -5.359 8.42e-08 ***
scale(longitude)                 -1.199e-01  3.503e-03  -34.237  < 2e-16 ***
room_typePrivate room            -6.812e-01  4.540e-03 -150.061  < 2e-16 ***
room_typeShared room             -1.078e+00  1.425e-02  -75.634  < 2e-16 ***
minimum_nights                   -1.870e-03  1.113e-04  -16.794  < 2e-16 ***
number_of_reviews                -1.686e-04  5.983e-05   -2.818  0.00483 ** 
reviews_per_month                -1.617e-03  1.685e-03   -0.960  0.33722    
calculated_host_listings_count    1.382e-04  7.452e-05    1.855  0.06364 .  
availability_365                  4.659e-04  1.818e-05   25.623  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3908 on 32123 degrees of freedom
Multiple R-squared:  0.5429,	Adjusted R-squared:  0.5427 
F-statistic:  2935 on 13 and 32123 DF,  p-value: < 2.2e-16
```

RandomForest



~~~
```{r}
set.seed(4543)
rf_model = randomForest(price ~   neighbourhood_group + latitude+  longitude + room_type+ minimum_nights +
number_of_reviews + reviews_per_month + calculated_host_listings_count + availability_365, data=bnb_df_train,
na.action = na.omit)
rf_model
```
~~~

```
Call:
 randomForest(formula = price ~ neighbourhood_group + latitude +      longitude + room_type + minimum_nights +
number_of_reviews +      reviews_per_month + calculated_host_listings_count + availability_365,      data =
bnb_df_train, na.action = na.omit) 
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 1919.345
                    % Var explained: 58.36
```