Final Project - Airbnb Listing

Tyler Cheung

Introduction and data

Airbnb has emerged as a highly successful company since its establishment in 2008, with over 7 million listings in 191 countries and regions, operating in more than 100,000 cities. New York City is one of the most popular cities in the world and has become a thriving market for Airbnb, with nearly 50,000 listings in the city. Airbnb has seamlessly integrated into the rental landscape of the city in just a decade. Analyzing this dataset can provide valuable insights into key factors that can predict pricing for house owners to understand the market better. These analyses can uncover patterns and relationships within the data, leading to a deeper understanding of the Airbnb market in NYC and providing useful information for house owners in the industry to learn how to better their businesses.

The aim of the project is to perform analyses on New York City Airbnb dataset and uncover insights into the sharing economy in one of the biggest cities of the world. My research question is "How can Airbnb's listing information potentially predict an Airbnb's pricing? I predict that factors such as the type of living conditions, location, and the number of reviews will help increase an Airbnb's value and predict its price.

Variable	Class	Description
ID	integer	Airbnb listing's ID
Name	character	Listing's description
Host ID	integer	Airbnb Host ID
Host Name	character	Name of the Host
Borough	character	Borough Location in NYC of the Airbnb (BK, MAN, BX, Queens)
Neighborhood	character	Specific Neighborhood Names
Latitude	integer	Numerical location of Airbnb in Latitude
Longitude	integer	Numerical location of Airbnb in Longitude
Room_Type	character	Type of room of the Airbnb
Price	integer	The Amount it costs per night for the Airbnb

Variable	Class	Description
Minimum_Nights	integer	Minimum amount of nights a person can stay at the Airbnb
Number_of_Reviews	integer	Amount of reviews posted in 2019 for the Airbnb
Last_Review_Date	integer	Last Review posted for the Airbnb
Reviews_per_Month	integer	Amount of reviews posted per month on Average
Availability_365	integer	How many days out of the year is the Airbnb available for

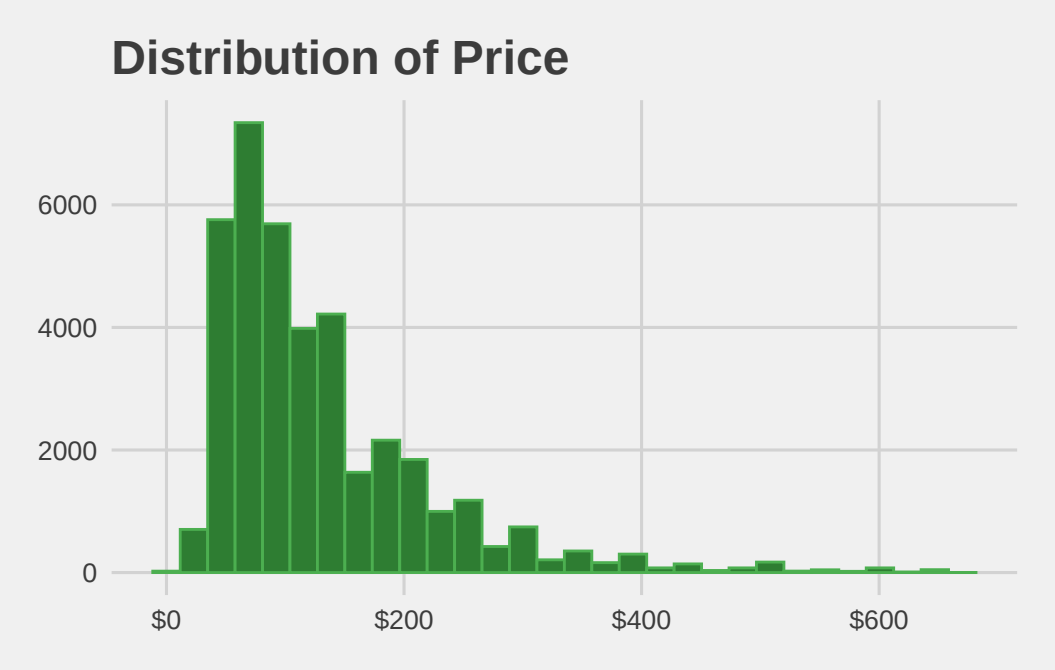
Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019. This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions. This public dataset is part of Airbnb, and the original source can be found on this website.

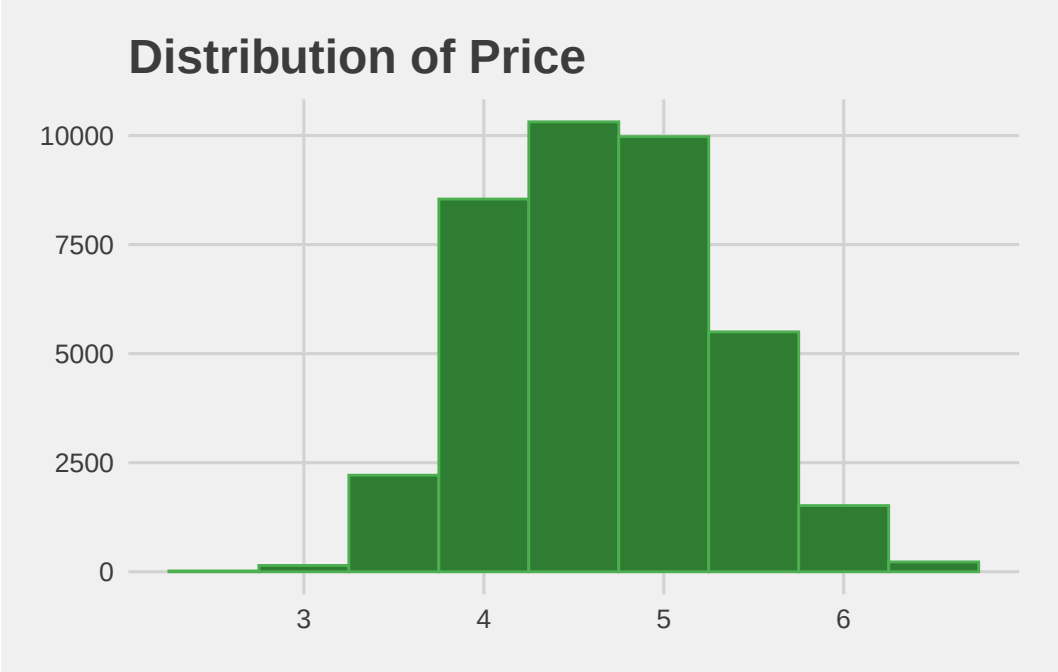
https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Cleaning up the Dataset

Exploratory Data Analysis - EDA

Distribution of Response Variable: Pricing of Airbnbs





The initial examination of the distribution of Airbnbs' average pricing in NYC shows a right-skewed, non-normal distribution with a range of 0 to 400 dollars. However, upon taking the logarithm of the response variable, the distribution becomes roughly normal, unimodal, with a mean around 4.5, and minimal skewness, ranging from 3 to 6.5. While we may not necessarily utilize the log-transformed response variable in our regression, it is important to consider its impact on the distribution.

Upon analyzing the distribution of Airbnbs' pricing, my focus was drawn towards the average prices of Airbnb listings in the five distinct boroughs of New York City. Notably, Brooklyn and Manhattan emerged as the two most sought-after locations, boasting the highest number of listings per year and commanding the highest average prices for Airbnb accommodations in NYC. These intriguing findings have piqued my curiosity and have led me to delve deeper into investigating the potential variables that may influence the pricing of Airbnbs in these popular boroughs. I broke my research down into two sub-section by first looking into the effects of my quantitative variables and following that looking the effects of categorical variables. By conducting a thorough analysis, I hope to gain a comprehensive understanding of the underlying factors that contribute to the observed price differentials and shed light on the complex dynamics that drive the Airbnb market in Brooklyn and Manhattan.

Distribution of Key Predictors (Numeric):

Key predictors that are considered for this potential model are the numerical stats for each listing's hosting criteria and user feedback in order to examine the predictors effect on our

response variable (Pricing). Given that the distribution of Number of Reviews, Minimum Nights, and Availability within the year are right-skewed we wanted to continue to consider the log transformation of these predictor values.

Upon examining the correlation matrix, we can observe that there is no significant correlation between the variables. This implies that if we were to use interaction terms in our model we would be able to avoid problems like multicollinearity, which can make it difficult to interpret the effects of individual variables on the response variable. This is because the assumption of correlation between the variables is not fully met.

Methodology

Observation of Quantitative & Categorical Variables

```
# A tibble: 10 x 5
   term                          estimate std.error statistic       p.value
   <chr>                            <dbl>     <dbl>     <dbl>         <dbl>
 1 (Intercept)                      121.       19.6      6.17  0.00000000186
 2 Minimum_Nights                    -5.59      4.80    -1.16  0.245
 3 Number_of_Reviews                -19.9       4.85    -4.11  0.0000492
 4 Availability_365                   5.37      4.87     1.10  0.271
 5 Neighborhood_East.Village         72.8      23.3      3.13  0.00192
 6 Neighborhood_Greenpoint           83.9      26.4      3.18  0.00161
 7 Neighborhood_Harlem               17.2      23.9      0.719 0.473
 8 Neighborhood_Upper.West.Side      42.6      26.7      1.60  0.111
 9 Neighborhood_Williamsburg         62.1      22.2      2.79  0.00550
10 Neighborhood_other                48.5      19.1      2.55  0.0113
```

```
# A tibble: 10 x 5
   term                          estimate std.error statistic   p.value
   <chr>                            <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)                     4.62      0.102     45.1   2.26e-150
 2 Minimum_Nights                 -0.0157    0.0276    -0.568 5.70e-  1
 3 Number_of_Reviews              -0.121     0.0279    -4.33  1.93e-  5
 4 Availability_365                0.0267    0.0279     0.955 3.40e-  1
 5 Neighborhood_East.Village       0.422     0.134      3.15  1.75e-  3
 6 Neighborhood_Greenpoint         0.430     0.152      2.84  4.80e-  3
 7 Neighborhood_Harlem             0.0516    0.137      0.375 7.08e-  1
 8 Neighborhood_Upper.West.Side    0.324     0.153      2.11  3.52e-  2
 9 Neighborhood_Williamsburg       0.371     0.128      2.91  3.86e-  3
10 Neighborhood_other              0.261     0.109      2.39  1.74e-  2
```

Model	RMSE	Adj. R Sq.	AIC	BIC
preprocess_1	90.297	0.063	3970.072	4011.610
logModel	0.524	0.070	524.324	565.862

The objective of our analysis was to identify the optimal model for predicting the predicted price of Airbnb listings based on their various listing criteria (Borough, Room Type, Neighborhood, Minimum Nights, Availability) and customer feedback (Number of Reviews). To capture the variability in Airbnb pricing, we utilized a multiple linear regression framework. We evaluated two potential multiple regression models, both of which predicted the Airbnb price. The first model directly predicted the price, while the second model predicted the logarithm of the price. The models incorporated various predictors such as Borough, Neighborhood, Number of Reviews, Availability during the year, Minimum nights, and Room Type.

Through cross-validation techniques, we compared two models to determine the optimal one for predicting an Airbnb's potential listing price. The table shows that the 'preprocess_1' model, which directly predicts the price, produces significantly different results from the 'logModel', which predicts the logarithmic transformation of the price. Our primary focus was on the RMSE and BIC metrics. RMSE, which measures the average error within a model, should be minimized, and it is the most important value for prediction purposes. After considering the metrics, we concluded that 'logModel' is the better model to use as it has a much lower RMSE of 0.524 compared to 'preprocess_1', which has an RMSE of 90.297. Moreover, the BIC value of 'logModel' is 565.862, which is much lower than 'preprocess_1''s BIC of 4011.610.

After identifying the potential model, we tested the conditions for linear regression. The first condition is independence, which we can assume to hold true since each Airbnb listing may not be affected by others. Although, we may consider that the location of two Airbnbs could potentially affect pricing. However, we can assume that each listing doesn't affect the other and independence is satisfied. Linearity was tested from the model, and normality was seen to be satisfied from the QQ plot. Finally, we tested constant variance from the residual plot, which showed an even distribution along the line of our data points, allowing us to assume constant variance is satisfied.

Results

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.61630	0.10238	45.08936	0.00000	4.41496	4.81764
Minimum_Nights	-0.01567	0.02758	-0.56832	0.57017	-0.06990	0.03856
Number_of_Reviews	-0.12068	0.02787	-4.33075	0.00002	-0.17547	-0.06588
Availability_365	0.02667	0.02794	0.95455	0.34044	-0.02828	0.08162
Neighborhood_East.Village	0.42206	0.13382	3.15386	0.00175	0.15889	0.68523
Neighborhood_Greenpoint	0.42999	0.15154	2.83757	0.00480	0.13199	0.72800
Neighborhood_Harlem	0.05159	0.13742	0.37543	0.70756	-0.21865	0.32184
Neighborhood_Upper.West.Side	0.32392	0.15327	2.11346	0.03525	0.02251	0.62533
Neighborhood_Williamsburg	0.37102	0.12757	2.90832	0.00386	0.12014	0.62190
Neighborhood_other	0.26139	0.10940	2.38940	0.01739	0.04626	0.47653

$$\widehat{Price} = e^{4.616} \times e^{-0.0156 \times Minimum\_Nights} \times e^{-0.12068 \times Number\_of\_Reviews} \times e^{0.0267 \times Availability\_365} \times e^{0.42206 \times Neighborhood\_East.Village} \times e^{0.430 \times Neighborhood\_Greenpoint} \times e^{0.052 \times Neighborhood\_Harlem} \times e^{0.324 \times Neighborhood\_Upper.West.Side} \times e^{0.37102 \times Neighborhood\_Williamsburg} \times e^{0.261 \times Neighborhood\_Other}$$

By analyzing the slopes/effects of the predictors in our model, we were able to test the hypothesis that factors such as living conditions, location, and the number of reviews influence an Airbnb's value and predict its price. We found that neighborhoods in Brooklyn have a greater impact on the price of an Airbnb than those in Manhattan, assuming all other factors remain constant. This is not surprising given that Brooklyn has become a popular travel destination in recent years due to its gentrification and the addition of tourist attractions. For instance, Greenpoint was found to have a higher predictive value than other neighborhoods, possibly due to its popularity among young people in 2019.

Our analysis revealed significant relationships between the median price of an Airbnb and the minimum nights, number of reviews, and availability during the year predictors. Specifically, each additional year in these predictors is expected to multiply the median price of an Airbnb by a factor of -0.0157, -0.1207, and 0.0267, respectively, on average, while holding all other factors constant.

Discussion

This research project aimed to predict an Airbnb's listing price in 2019 based on potential predictors, such as Borough, Neighborhood, Room Type, Minimum Nights, Number of Reviews, and Availability during the year. Our analysis revealed that certain neighborhoods, including Greenpoint, East Village, Upper West Side, Williamsburg, and Other neighborhoods, had a significant positive impact on an Airbnb's price, as reflected by p-values less than 0.05. This finding supports our hypothesis that factors such as location and amenities would increase an Airbnb's value. However, we also acknowledged that the variables were skewed, and thus we

used a multiple linear regression with log-transformed response to address this issue. Despite this adjustment, the model did not meet some of the required statistical assumptions, which limits the validity of our findings. There was definitely difficulty in finding an accurate model to use during my research. I initially wanted to use LASSO regression to help me find which predictor variables to use, but after realizing that majority of my datapoints were regressed to coefficient 0 leaving me with very few variables to observe I decided not to use LASSO as it would limit the scope of my research.

Appendix

Apendix 1.2: Linearity and Constant Variance

Appendix 1.3 : Normality