Airbnb Pricing

Matthew Hui, Tian Qi, Tiance Tan

1. Description of the dataset

Data

The data was accessed via Kaggle. It contains information on Airbnb listings in New York, NY during 2019 including price, rental attributes, and location.

Introduction

Airbnb is an online marketplace for arranging lodging, primarily homestays. The company does not own any of the real estate listings. It acts as a broker, receiving commissions from each booking. Since the number of users continues to grow, the pricing of Airbnb is an essential topic that everyone values. The users could learn about the important factors that affect the pricing and make their choices based on the demand. And the hosts can also reasonably set their price for Airbnb to attract more guests.

The data we analyzed has 48,895 observations and 16 variables, including one response variable price. After discussion, we decided to remove record id and categorical variables which have too many levels (name, host id, host name, and neighbourhood). Finally, we only looked at 7 variables, which includes the neighborhood group, room type, minimum nights, number of reviews, reviews per month, host listing count and availability in 365 days. The purpose of this analysis is to predict the pricing based on these features, which can help hosts and guests better understand the pricing information and help them make better decisions. Statistical learning techniques were applied to a sample of data of Airbnb pricing from Brooklyn.

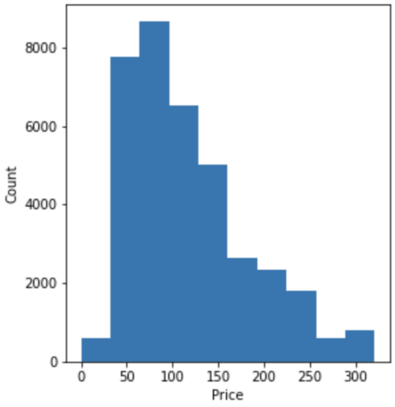
Statistically, this model is extremely limited in its application due to the data is only based on the New York area. Hence it is unreasonable if we want to conclude Airbnb pricing in other regions/countries. For future studies, we can include more variables in different regions in a larger sample size.

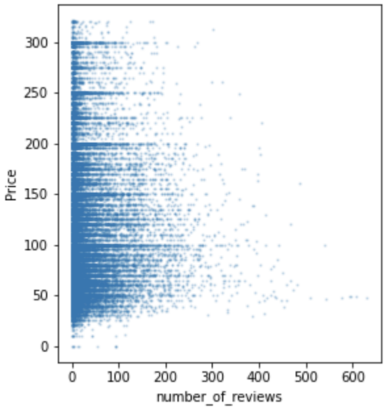
Variable description:
price - price in dollars
neighbourhood_group - group of neighborhood
latitude - latitude coordinates of the listing
longitude - longitude coordinates of the listing
room_type - listing space type
minimum_nights - amount of nights minimum
number_of_reviews - number of reviews
last_review - the date of last 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
id - id of the listing
name - name of the listing
host_id - host id
host_name - host name
neighbourhood - neighborhoods

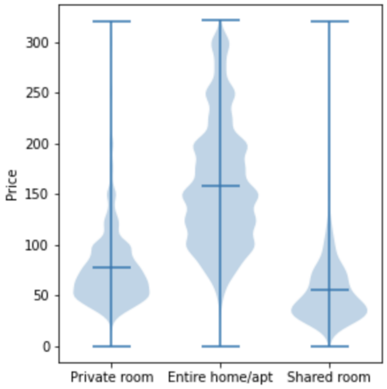
2. Summary of methods

We first cleaned our data in order to prepare it for analysis. The data had missing values as well as values that did not make sense. Once the data was cleaned, exploratory data analysis was performed in order to understand the data better before creating models. Linear models with and without log transformed responses were considered. We produced all possible models based on the variables we decided to look at and then performed the model diagnosis that includes: assumption validation, influential points, check for heteroscedasticity and multicollinearity. With our biggest problem being heteroscedasticity, we used a robust standard error method in order to test our variables' significance.

3. Exploratory Data Analysis

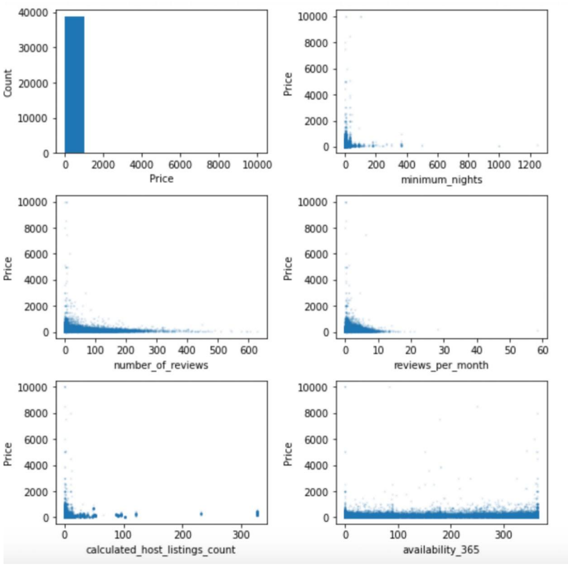


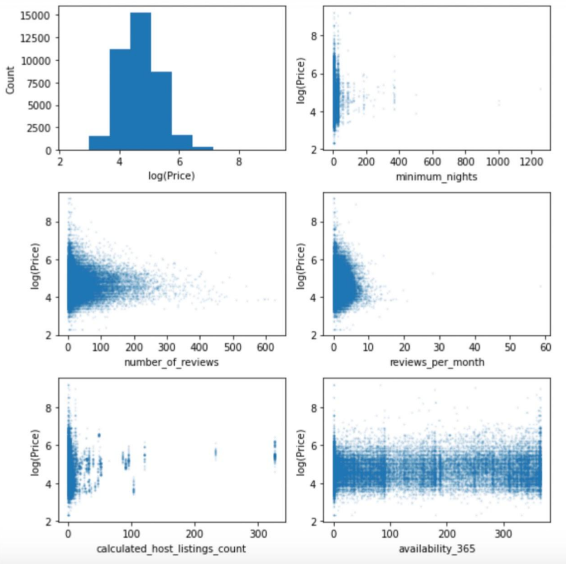




neighborhood_group	Average Price
Bronx	79.644571
Brooklyn	121.515209
Manhattan	180.052489
Queens	95.762571
Staten Island	89.964968

When exploring the data, we remove some outliers in order to perform better graphs and give us ideas about the dataset. Our data had many outliers, so it made it difficult to get a good sense of our data when we looked at the various graphs. However, once we removed outliers and graphed it again, we could see the shape of the data a lot better. Based on the graphs and table, we find that there are large differences in average listing price between each neighborhood. Listings with lower prices happen to have more reviews. Room types also have an impact on the prices. Entire homes or apartments have the highest average listing price. It is harder to see a trend in the quantitative variables when comparing them to price, but when we did initial regression tests between price and various variables, they all turned out significant.





Plot: Price against Numerical Variables **Plot: log(Price) against Numerical Variables**

Based on the graphs, we can see that there are some influential points and outliers which need to be considered in the model diagnosis. Immediately, we noticed that equal variance may be an issue within the graph, so we graphed quantitative variables against both price and log transformed price. A log transformation in y makes the graphs look better and it gives us hints to consider log transformation in our future models as well. Even when transforming price, we still see potential issues with our data and model assumptions.

4. Regression Analysis and Model Selection

Since price is a quantitative variable, we used multiple linear regression to build our model for predicting price. We chose 7 of 15 variables to consider in our model. Most of the variables we took out of consideration were categorical variables that had too many levels to even consider. We also removed longitude and latitude because interpretation of the variables did not really make sense even though they might have been significant. Since all 7 variables we chose to include were significant, we fit a full model to do some initial checks. After making changes with our initial checks, we were ready for model selection.

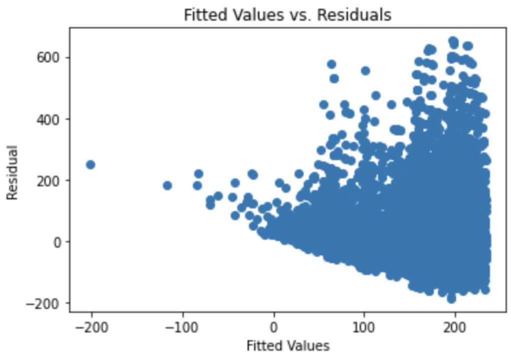
When doing model selection, we used the best subsets method in order to figure out what model would be best based on our criteria. With seven possible variables, we had to test 127 models (did not test the null model with 0 predictors). The criteria we looked at were R^2-adjusted, Mallows' Cp, AIC, and BIC.

After comparing all the models, we decided on the model that included the variables: neighbourhood_group, room_type, minimum_nights, number_of_reviews, and availability_365. This model produced the second highest R^2-adjusted and lowest Mallows' Cp, AIC, and BIC. We chose this model because out of the four criteria, it was the best in three of them. It also had five variables instead of six which gives benefits in interpretation and reduces risk of overfitting.

5. Model Diagnosis

We did an initial model diagnosis on the full model of seven variables we produced in order to look for any initial fixes before we start to do model selection. We first looked at and removed influential points. We found influential points to remove by looking at each observation's Cook's distance. We found that 429 points were influential to our model so we removed them. Before removing the influential points we obtained an R^2-adjusted value of 0.112 and after it improved to 0.367. This indicates that the influential points both affected our models coefficients and also were large sources of error likely being large residuals.

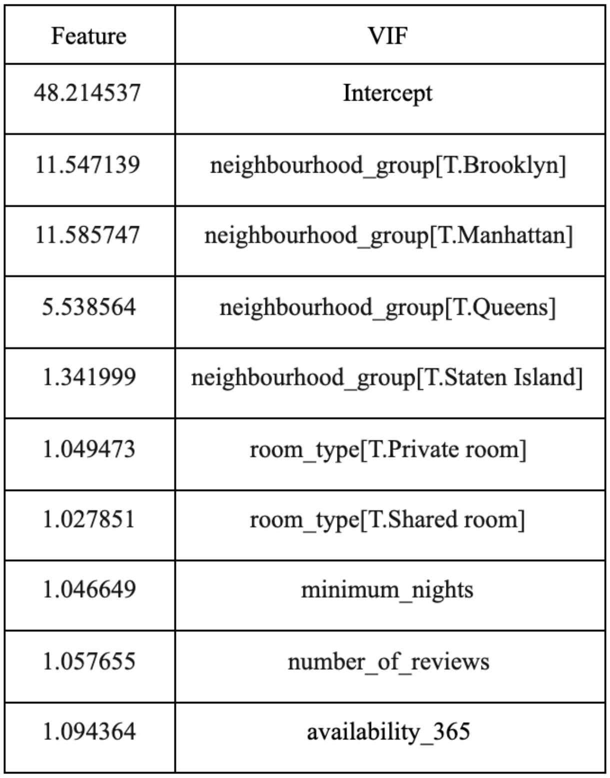
Model	R^2-adjusted
Full Model with Influential Points	0.112
Full Model No Influential Points	0.367



We then looked at our assumption and as expected from our exploratory data analysis, the equal variance assumption was violated. We see an obvious cone shaped pattern in the Fitted Values versus

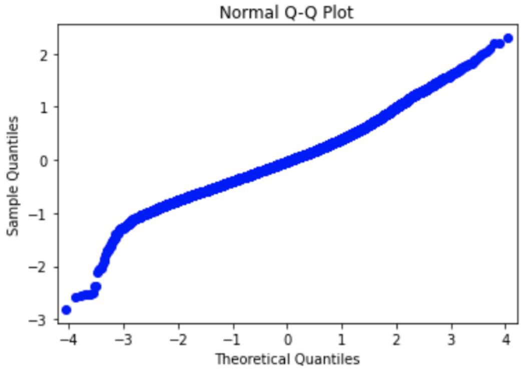
Residual plot. To fix this, we took the natural logarithm of price as our response in hopes to help remedy the equal variance assumption. We were then ready to find the best subset of our candidate variables. While our other assumptions had no clear violations, we checked them again once we created our best subsets model.

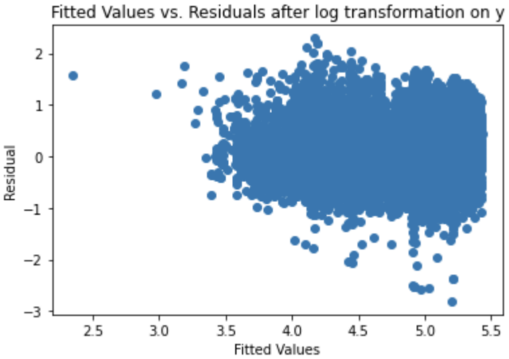
After finding our best model, we checked for multicollinearity as well as revalidating our assumptions with the new model. When checking for multicollinearity, we saw that the variance inflation factors (VIF) of our predictors were all relatively low. The parameters with high VIFs were intercept and two of the neighbourhood group predictors. Categorical variables tend to have multicollinearity so this is not as big of an issue. We can determine that multicollinearity is not a serious issue within our model.



Feature	VIF
48.214537	Intercept
11.547139	neighbourhood_group[T.Brooklyn]
11.585747	neighbourhood_group[T.Manhattan]
5.538564	neighbourhood_group[T.Queens]
1.341999	neighbourhood_group[T.Staten Island]
1.049473	room_type[T.Private room]
1.027851	room_type[T.Shared room]
1.046649	minimum_nights
1.057655	number_of_reviews
1.094364	availability_365

Since multicollinearity was not an issue, we checked our assumptions again. When looking at the Normal Q-Q plot, we see that our residuals are slightly left skewed, but since the sample size is very large, the normality assumption is still valid. When looking at the fitted values versus residuals, we do not see a curve in the data, but there is still a fan shape in the plot. Although the residual plot looks slightly better than before the transformation, the equal variance assumption is still violated and we see that heteroscedasticity exists. This is confirmed with the Breusch-Pagan test, where we got an extremely low p-value. The independence assumption is not violated since there should be no pattern in residuals for an observational study.





Since heteroscedasticity exists in the model, there are problems that arise that we should pay attention to. Although our coefficients are still properly estimated, unequal variances means that both our tests for predictors as well as confidence intervals are affected since our estimates of standard deviation of each coefficient are incorrect. Because of this, before committing to a final model, we refitted the model using a robust standard error method that assumes unequal variances. With the equal variance assumption no longer needed, we get new test results. The robust standard error method makes the standard error of the coefficients bigger than originally estimated and we still see that our slope estimates are significant.

After removing points with high cook's distance, log-transforming price, and using a robust standard error method to assume unequal variance, we produce our final model.

6. Final Model

After using best subsets method and validating our model, we have our final model:

*ln(price) = 4.6378 + Brooklyn * 0.2944 + Manhattan * 0.5833 + Queens * 0.1513 + Staten Island * -0.0378 + Private room * -0.7613 + Shared room * -1.1979 + minimum_nights * -0.0040 + number_of_reviews * -0.0004 + availability_365 * 0.0006*

As stated before, this model produced the second highest R^2-adjusted out of all the possible models from our 7 variables. The reason we chose this model over the other model was because of the lower Mallows' Cp, AIC, and BIC. This model also used 5 variables while the model with the highest R^2-adjusted used 6. By using less terms, it makes interpretation easier and reduces the risk of overfitting.

7. Interpretation

Based on the final model ln(price) = 4.6378 + Brooklyn * 0.2944 + Manhattan * 0.5833 + Queens * 0.1513 + Staten Island * -0.0378 + Private room * -0.7613 + Shared room * -1.1979 + minimum_nights * -0.0040 + number_of_reviews * -0.0004 + availability_365 * 0.0006, we can predict a private room in Manhattan whose minimum nights is 0, number of reviews is 410 and availability in the next 365 days is 331 has a price about 89.5 dollars per night.

The final model contains five variables: neighbourhood group, room type, minimum nights, number of reviews and availability. The R squared is 0.531, which means that this model explains 53.1% of the variance of price.

All variables in the model have a significant impact on price. As we did log transformation on the response variable, we exponentiate the coefficients in order to interpret the results. Here are the results after exponentiation:

Parameters	Coefficient	p-value
Intercept	103.3168	0.000
neighbourhood_group[T.Brooklyn]	1.3423	0.000
neighbourhood_group[T.Manhattan]	1.7919	0.000
neighbourhood_group[T.Queens]	1.1633	0.000
neighbourhood_group[T.Staten Island]	0.9629	0.141
room_type[T.Private room]	0.4671	0.000

room_type[T.Shared room]	0.3018	0.000
minimum_nights	0.9960	0.000
number_of_reviews	0.9996	0.000
availability_365	1.0006	0.000

The baseline level for the neighbourhood group is Bronx. Compared to the price of listings in the Bronx, the price in Manhattan is on average 79.19% higher, price in Brooklyn is 34.23% higher, and price in Queens is 16.33% higher, while the price in Bronx and Staten Island do not have significant differences.

Booking a private room or shared room is cheaper than booking the entire home. The price for a private room is on average 53.29% lower than the entire home, and the price for a shared room is much cheaper, which is about 69.82% lower than the entire home .

Minimum nights and number of reviews are negatively correlated with price. When the number of minimum nights increases by 1, the average price decreases by 0.40%. When the number of reviews increases by 1, the average price decreases by 0.04%.

Availability is positively correlated with price. For every one day increase in the number of available days, the price increases by about 0.06% on average.

8. Summary of findings

From this study we conclude that the neighbourhood group of a listing and its room type has a huge impact on its price. The price of listings in Manhattan and Brooklyn is relatively high. Shared room is much cheaper than a private room and entire home, which suggests that people value privacy when choosing a listing. From a practical perspective, it seems doable to predict the pricing of Airbnb in the New York area since this data contains a few important variables. However, it still has a few limitations. First, some other factors will affect the Airbnb pricing, such as the hotel pricing in that area. If the hotel pricing is low, more people would like to choose Airbnb. Another important factor is the tourist seasonality. During holidays or weekends, Airbnb pricing would increase a lot since the demand increases.

Possible future research includes using machine learning to fit a model, doing sentiment analysis on the reviews, and adding more predictors. This model would only be applicable to listings in New York City since we use the variable neighbourhood_group. If we wanted to create a model that is more generalized, it would require more data.