# Exploratory Data Analysis and Price Prediction on Airbnb in NYC (2019)

Citina Liang

May 10 2020

## 1 Abstract

Airbnb, founded August, 2008 in San Francisco, is the most popular online marketplace for renting and booking primarily home-stays and tourism experiences nowadays; and New York City as the most populous city in the United States, also leads the pack as the most expensive city. Smart Pricing is one of the machine learning models of Airbnb, which suggests an appropriate price after a client has entered the rental details, and it can also adjust the price automatically based on the changes in demand for similar listings. The goal of our project is to do an exploratory data analysis on the Airbnb in New York City (2019) and to explore metrics that build Smart Pricing Model by multiple linear regression.

## 2 Introduction

This dataset is posted on Kaggle and was originally collected from here. The dataset contains 48,895 observations and 16 variables which are information about hosts, stays' geography, availability, listings' description and price (dollars per night). Our target variable is **Price**, and we will use variables after data cleansing to build a multiple linear regression to predict price.

The following report is established in four sections. The third section shows procedures of data cleaning and details of the remaining variables. The fourth section takes a profound explorations in the relations between variables of the dataset. The fifth section provides a smart pricing model using multiple linear regression and model diagnostic. The last section concludes the result of the project, the limitations and future works can be done.

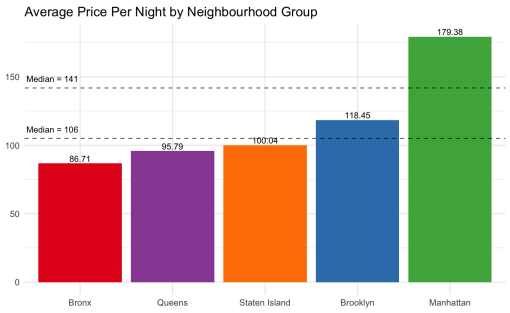


Figure 1: Average Price per Night by Neighbourhood Groups

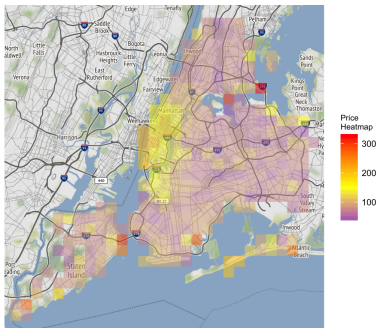


Figure 2: Average Price per Night on Maps

# 3 Data Overview

## 3.1 Data Cleansing

The data contains several redundant and uninformative variables as well as few missing values, there are three steps in total to construct the clean dataset. First, check for missing values and remove observations with missing prices. Second, remove observations with price outside 0.5 to 99.5 percentiles. Third, remove observations with min_nights out of 0.1 to 99.9 percentiles. Fourth, remove uninformative variables which will not help the prediction for price, for example, id (id of the listing), host_name (name of the host) etc. After data cleansing, there are 47,709 observations and 12 variables left.

## 3.2 Remaining Variables

**name**: name of the listing, **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, **reviews_per_month**: number of reviews per month, **host_listings**: amount of listing per host, **availability_365**: number of days when listing is available for booking

# 4 Exploratory Data Analysis

Figure 1 shows the average price per night by neighbourhood groups. The average price in Manhattan is the most expensive, 179.38 dollars per night, then follows by Brooklyn, Staten Island, Queens and Bronx. The average price of Bronx is 86.71 dollars per night, which is nearly half of the price in Manhattan. The median is 141 dollars per night, and mean is 106 dollars per night. Figure

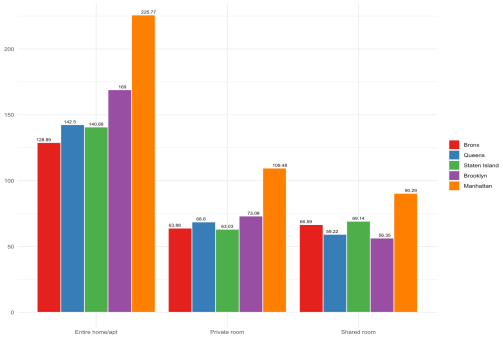


Figure 3: Average Price by Neighbourhood and Room Type

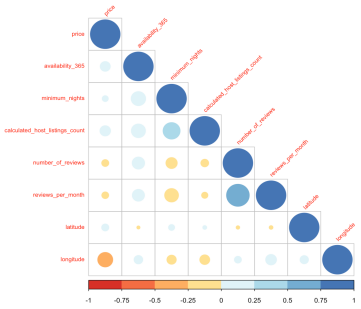


Figure 4: Correlation

2 shows similar information on a map, and we can better visualize the price difference geographically.

It is obvious that entire homes are more expensive than private rooms and then than shared rooms on average, but when combining with neighbourhood the situation is different. From Figure 3 we have some interesting findings. On average, a shared room in Manhattan is more expensive than a private room in any other neighbourhoods; a private room on Staten Island is cheaper than a shared room.

# 5 Price Prediction

After doing Exploratory Data Analysis, we remove variable **name** and **neighbourhood** since the first one can not be fit in a multiple linear regression without text mining and the second one is highly correlated to **neighbourhood_group**. We also generate a correlation plot (Figure 4) to visualize the correlation between quantitative variables in the dataset. We will use this figure to select variables later.

Before building the model, we make a density plot of the target variable **price**, and it is right-skewed as shown in Figure 4. In order to meet the normality assumption of the multiple linear regression, we transform **price** by applying a natural log, and as we can see in Figure 5, **price** becomes approximately normal now. And we split data in 80% for training and 20% for testing.

We first fit a full model with training data. The full model includes nine predictors. Next I apply Stepwise selection to the full model from both directions, it results that the full model has the smallest AIC. From the Correlation plot (Figure 4), we can see **price** is not so correlated with **minimum_nights**, so I fit a model excluding **minimum_nights**, and use goodness of fit test to compare the two model, the result shows the full model is better than the reduced model, and then I try the same steps for other numerical variables, and get the same result, so I choose the full model as my final model.

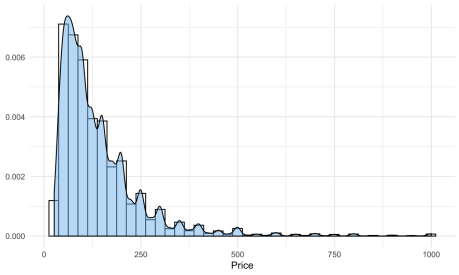


Figure 5: Density Plot of Price

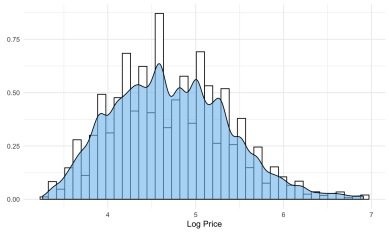


Figure 6: Density Plot of Price after Transformation

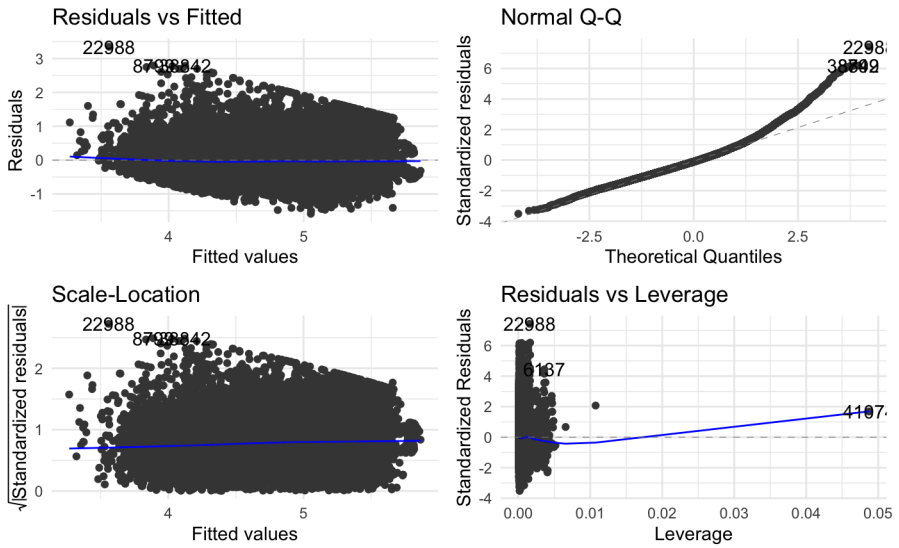


Figure 7: Diagnostic Plots

**Final model:** log_price = - 191.2 - 0.526 latitude - 2.94 longitude - 0.76 room_typePrivateroom - 1.09 room_typeSharedroom - 0.012 minimum_nights + 0.00084availability_365 - 0.00065 number_of_reviews - 0.015 reviews_per_month + 0.00045 calculated_host_listings_count - 0.0063 neighbourhood_groupBrooklyn + 0.29 neighbourhood_groupManhattan + 0.12 neighbourhood_groupQueens - 0.77 neighbourhood_groupStaten Island

# 6 Conclusion

## 6.1 Limitations

In conclusion, the adjusted $R^2$ of training data is 0.5328 and for the testing dataset is 0.5316, which is not bad since we are only using nine predictors, and


Figure 8: Wordcloud of Listings' names

Smart Pricing algorithm of Airbnb takes into account over 70 different factors. The model summary of the model shows all predictors are significant expect neighbourhood group Brooklyn. The diagnostic plots (Figure 7) also look fine, but still some potential influential points exist, also the normal Q-Q plot is a little bit off on the tail.

## 6.2 Future Work

This project only try using multiple linear regression to predict the price, and there are more work can be done to get a better prediction.

- Consider interaction terms
- Try other models (ridge, lasso, knn, etc.)
- Try time series analysis with more data from more years
- Do a text mining analysis (for example, topic modeling to create new predictors)
- Access external data (area criminal rates, transportation, etc.)