Capstone Project Report

Price Optimization Analysis for AirBnB

By,
Naveen Sheeba

Scope of Project

1. **Introduction**
    1. Overview of the project and its objectives.
    2. Brief description of the dataset and its context.
2. **Business Understanding**
    1. Define the business objective: To create a pricing strategy for Airbnb listings in NYC, NY for 2019.
    2. Identify the type of problem: Regression Analysis.
    3. Define key metrics: RMSE, Cross Validation Score, R^2, Adjusted R^2.
3. **Data Acquisition and Understanding**
    1. Raw to Relevant Data: Identify and extract relevant data for analysis.
    2. Data Type Inspection and Conversion: Ensure data types are appropriate for analysis.
    3. Address Dirty Data due to Constraints: Clean data to remove inconsistencies.
    4. Outlier Detection and Treatment: Identify and handle outliers in the dataset.
    5. Handling Missing Values: Impute or remove missing values appropriately.
    6. Exploratory Data Analysis: Conduct correlation analysis, visualizations, and statistical analysis to gain insights into the data.
    7. Handling Categorical Variables: Encode categorical variables for machine learning modeling.
    8. Train and Test Split: Divide the dataset into training and testing sets for model evaluation.
    9. Data Scaling/Feature Scaling: Scale numerical features as necessary.
4. **Modeling**
    1. Select appropriate regression models for analysis (e.g., linear regression, random forest regression, etc.).
    2. Perform hyperparameter tuning using cross-validation to optimize model performance.
    3. Train and evaluate models using relevant metrics (RMSE, R^2, etc.).
    4. Select the best-performing model for deployment.
5. **Deployment**
    1. Deploy the selected model for use in pricing strategy optimization.
    2. Ensure the model is accessible and user-friendly for Airbnb hosts.
    3. Provide necessary documentation and support for model usage.
6. **Customer Acceptance**
    1. Gather feedback from users (Airbnb hosts) regarding the effectiveness and usability of the pricing strategy tool.
    2. Make necessary adjustments based on user feedback to improve customer acceptance.
7. **Conclusion**
    1. Summarize key findings and outcomes of the price optimization analysis.
    2. Discuss potential future enhancements or extensions to the project.

- This scope of project outlines the steps and activities involved in conducting a price optimization analysis for Airbnb listings in NYC, NY for 2019, from data acquisition to model deployment and customer acceptance.

Introduction and Data understanding

The listings were scrapped on July 8th 2019 and are specific to NYC, NY.

The objective of the analysis is to:

- estimate listing price based on provided information
- derive additional useful and interesting insights

Part 1 - Deals with taking the existing dataset, performing data cleaning, feature engineering and running preliminary analysis.
The product of this part is a data file for Machine Learning analysis.

Part 2 - The Machine Learning part of the project applies machine learning algorithms to predict price of listings based on various input variables.

Business Understanding

- **Business Objective -** To create a pricing strategy for Airbnb, which helps Airbnb hosts set the right price for their Airbnb listing and provides customers the benefit of cost.
- **Type of Problem -** Regression Analysis (Dependent Variable is 'Airbnb listing price per night (in USD)' which is regressed against a bunch of independent variables (listing attributes - accommodates, bedrooms, bathrooms, beds, amenities provided, neighborhood and the room type).
- **Metrics -** RMSE (Root Mean Squared Error) to check which ML model provides accurate predictions, Cross Validation Score for hyperparameter tuning in certain ML models, R^2 and Adjusted R^2 for explainability power and to check model fit.
- **Data Science Lifecycle-**

    Business Understanding

    Data Acquisition and Understanding

    2.1 Raw to Relevant Data

    2.2 Data Type Inspection and Conversion

    2.3 Dirty Data due to Constraints

    2.4 Outlier Detection and Treatment

    2.5 Handling Missing Values

    2.6 Exploratory Data Analysis (Correlation Analysis, Visualizations & Statistical Analysis)

    2.7 Handling Categorical Variables for ML Modeling (Text Encoding)

    2.8 Train and Test Split and Data Scaling/Feature Scaling

    Modeling

    Deployment

Dataset

Following are the datasets columns, grouped together for better understanding

Host Descriptors:

- **host_id:** host ID
- **host_name:** name of the host
- **calculated_host_listings_count:** amount of listing per host

Review Descriptors:

- **number_of_reviews:** number of reviews
- **last_review:** latest review
- **reviews_per_month:** number of reviews per month

Listing Descriptors:

- **id:** listing ID
- **name:** name of the listing
- **room_type:** listing space type
- **minimum_nights:** amount of nights minimum
- **availability_365:** number of days when listing is available for booking
- **price:** price in dollars

Location Descriptors:

- **neighbourhood_group:** location
- **neighbourhood:** area
- **latitude:** latitude coordinates
- **longitude:** longitude coordinates

Data Cleaning

Replace missing values

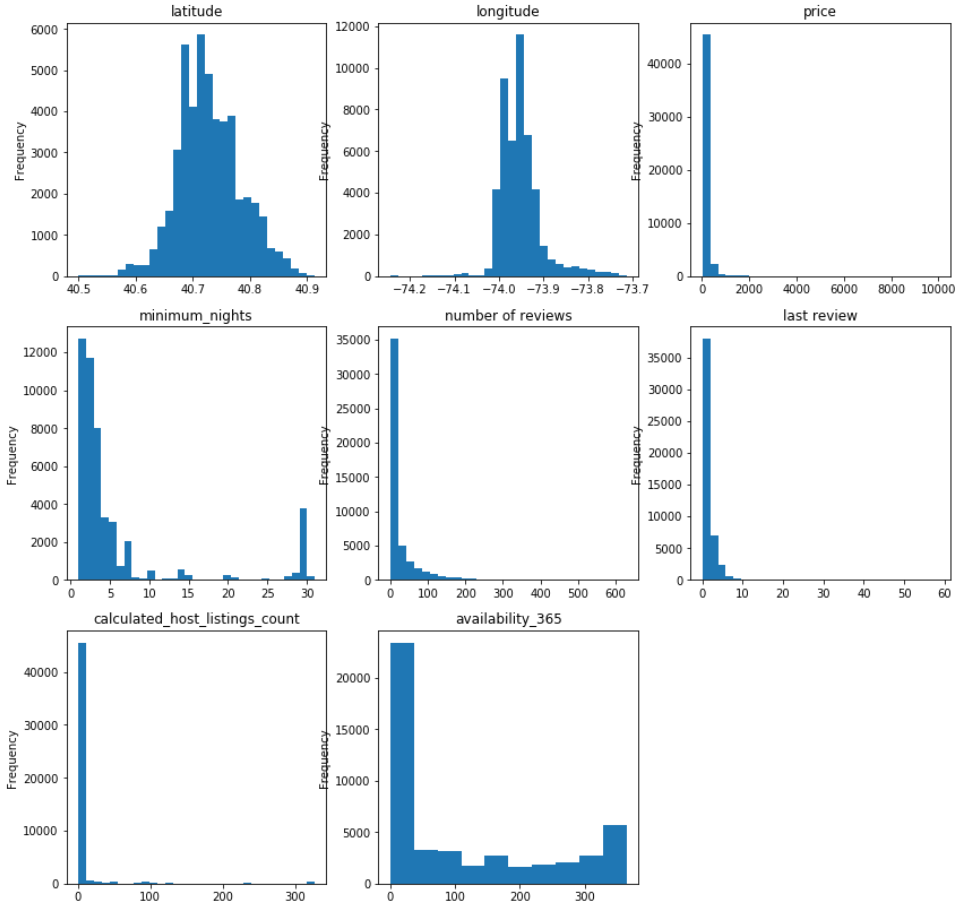
- Listings without reviews have missing values for “last_review” and “reviews_per_month”.
  For these lisitngs the missing values will be replaced by “0”

- Some listings are missing a name or the host name is missing.
  These will be replaced with "None"

Data inconsistencies

- The initial dataset contained 11 listings with price of 0 USD, that had been removed.
- In order to narrow the scope of this analysis,
  we will focus on a potential tourist market, assuming that tourists will not want to stay longer than 31 days.
  Hence, we can remove any listings with required minimum nights greater than 31.

Distribution Charts

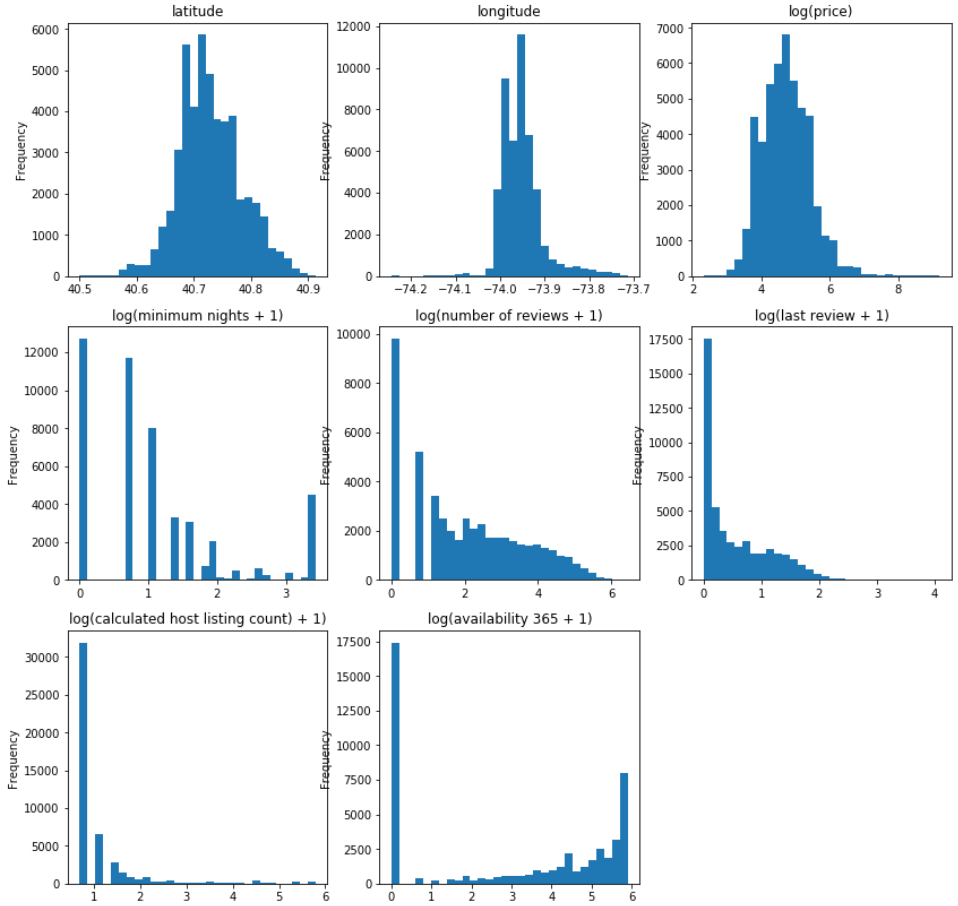


Looking at the distributions,
clearly the following are heavily right skewed:

- price
- minimum_nights
- number of reviews
- last review
- calculated_host_lising_count

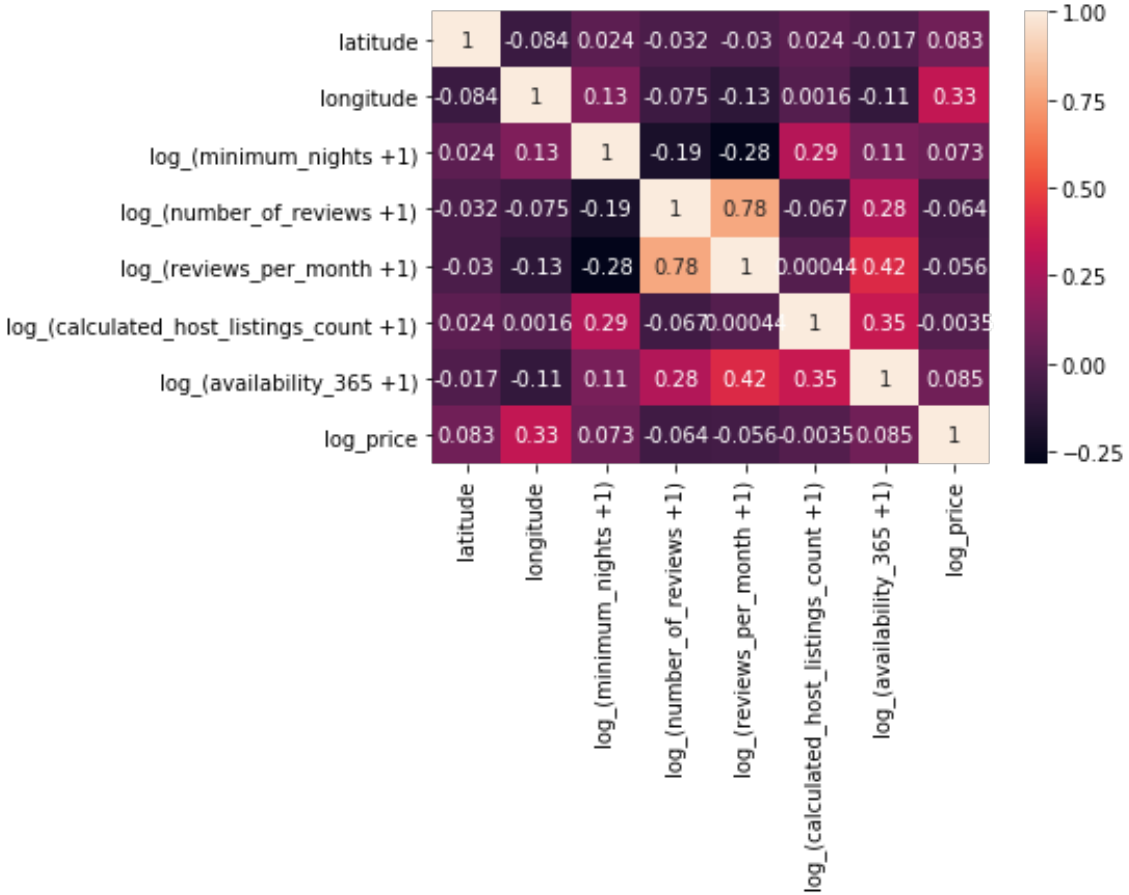
One way to reduce the skewness is to logarithmically transform the distributions

Distribution Charts



Here, logarithmic data transformation was used to smooth out the distribution

Correlation between variables



Observations on price:

• Price seems to be positively correlated with longitude meaning that one can expect higher prices as position in NYC moves West. This is expected because Manhattan, which is the most expensive borough of the city, is located on the west side of the city

• Latitude seems to have lesser effect on the price. However, there is a slight indication of higher prices located in the northern parts of the city

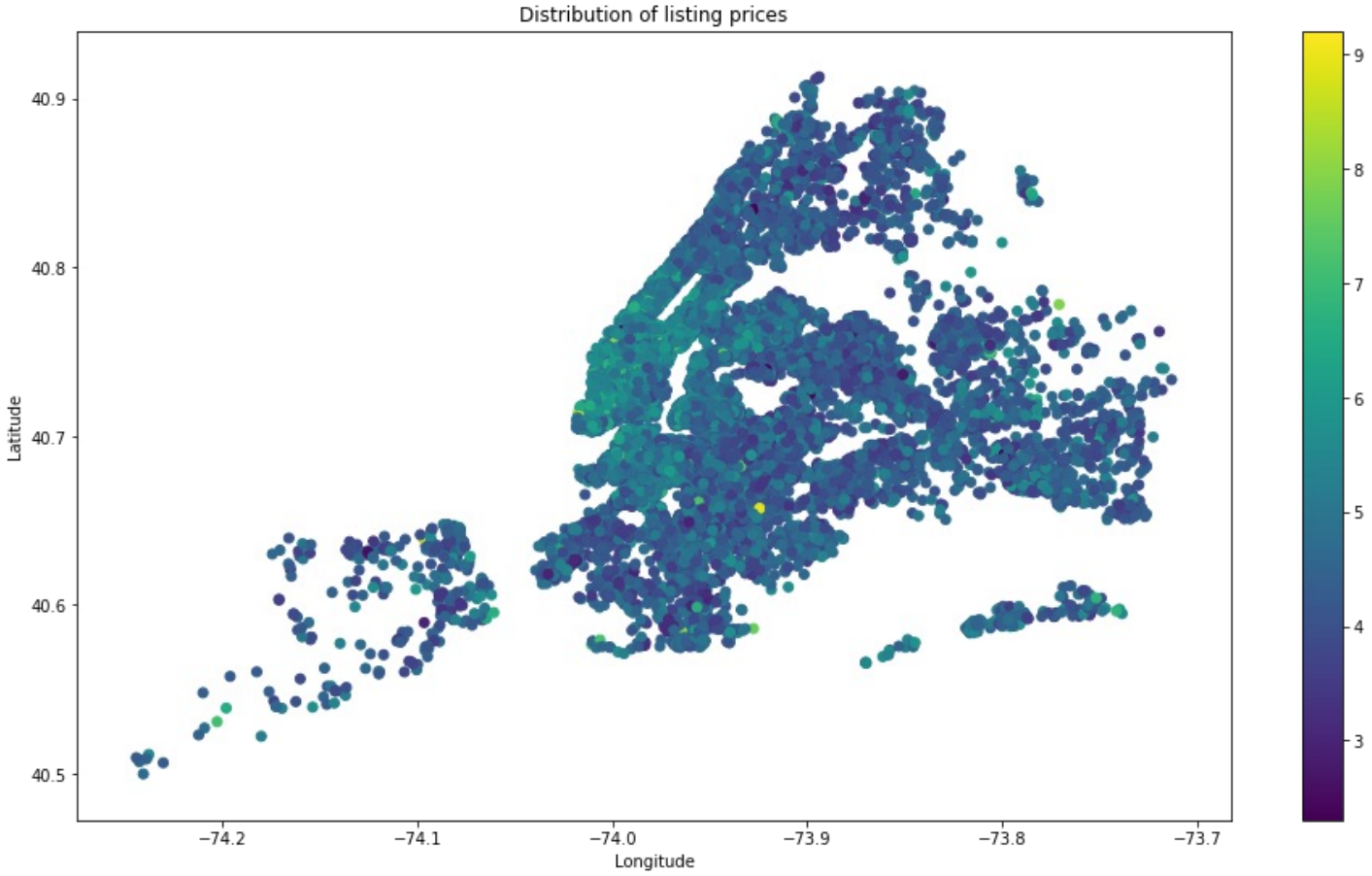
• Price is also positively correlated with: increasing availability, the fact that the property is rented by a host who lists other properties, and increasing number of minimum nights

• Price is negatively correlated with number of reviews and reviews per month, indicating that it is possible that the prior reviews could depress the prices to some extent

Other interesting observations:

Calculated host listing count is positively correlated with minimum nights and availability_365 indicating that hosts who list more than one property may be more strategic rather than opportunistic about their rentals. That may attempt to maximize the amount of time a single renter stays at their property to minimize turnover cost. They also tend to maximize the amount of time the property is being rented.

Price variation by neighborhood

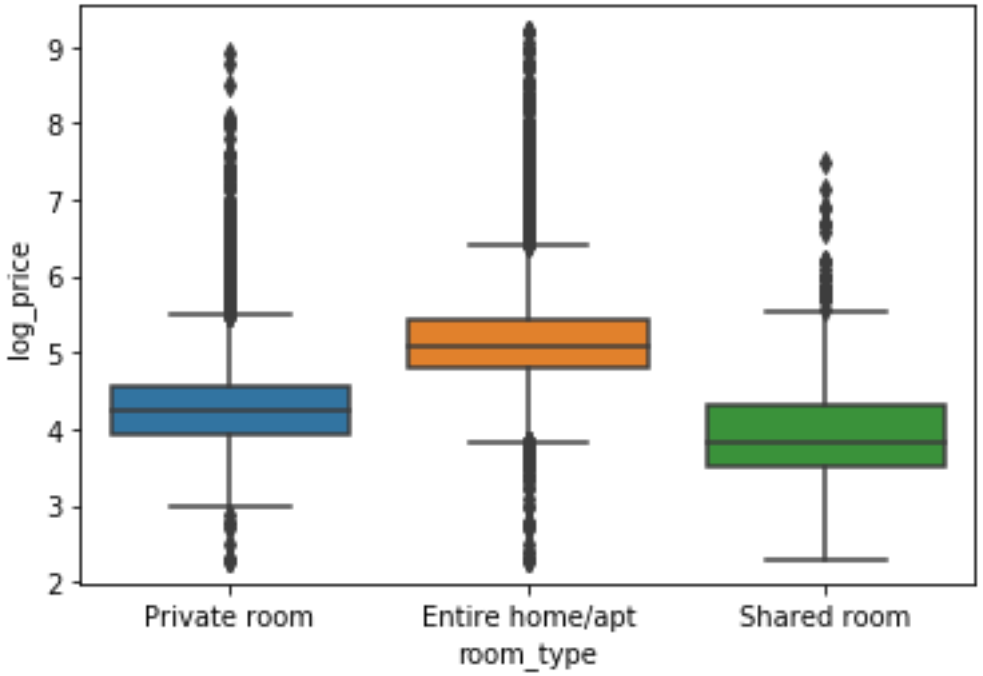


This scatter plot visualizes the geographical distribution of listings along with the relative pricing (increasing with brightening color).

High prices appear to be concentrated around Manhattan starting with neighborhoods around Central Park going south, as well as around portions of Brooklyn and Queens close to Manhattan.

Price variation by Room type

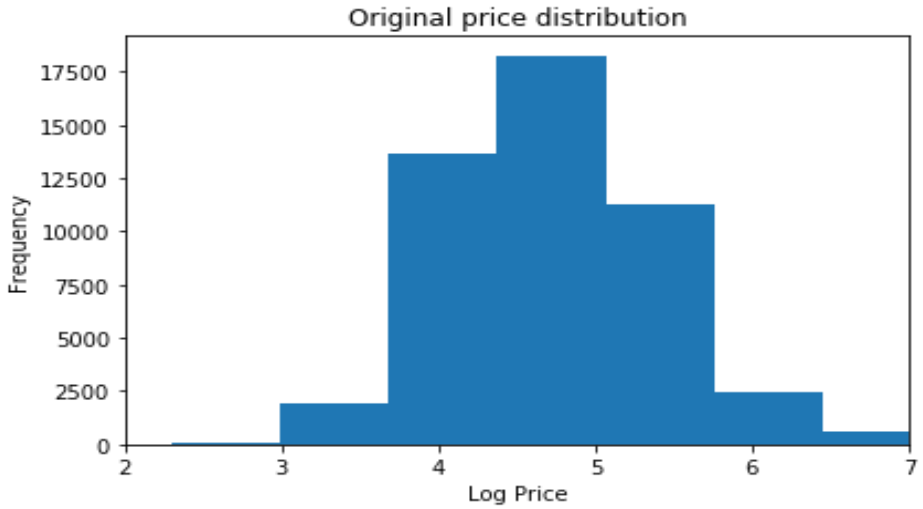
room_type	mean	median	amax	std
Shared room	3.954147	3.806662	7.495542	0.651180
Private room	4.295073	4.248495	8.922658	0.510161
Entire home/apt	5.142397	5.075174	9.210340	0.564829

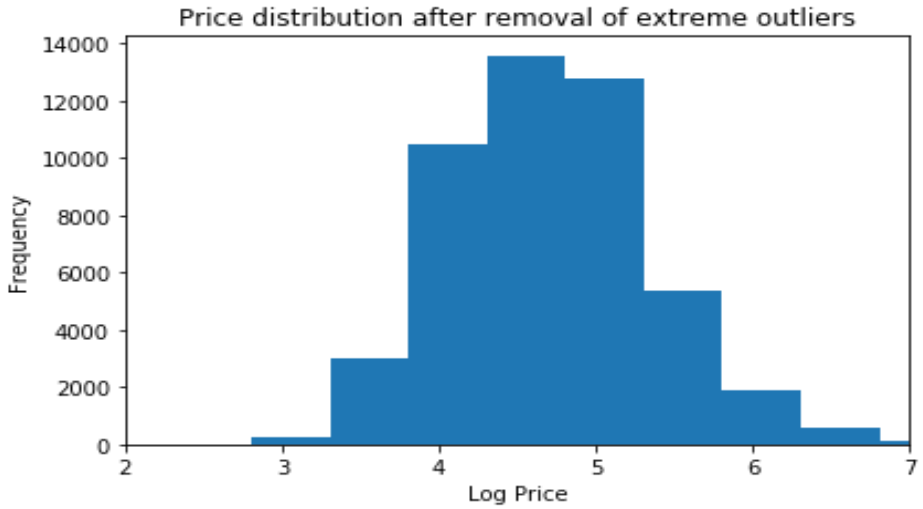


Extreme Outliers

- The data most likely includes extreme outliers, which will be difficult to model.
- An outlier is defined as 3 x IQR below 25th quantile and above 75th quantile

Price variation by Room type





As a result of outloer removal 312 rows of data were removed

room_type	mean	median	amax	std
Shared room	3.925321	3.806662	6.214608	0.592874
Private room	4.279373	4.248495	6.484635	0.467583
Entire home/apt	5.130023	5.075174	7.311218	0.527369

Machine Learning

With one hot encoding, feature engineering and natural language processing the shape of the dataframe grew substantially.

- Several numeric columns have been log transformed
- Individual neighborhoods have been one hot encoded along with boroughs and categorized times since the last review
- Lastly columns were created to document use of popular words in the listing name

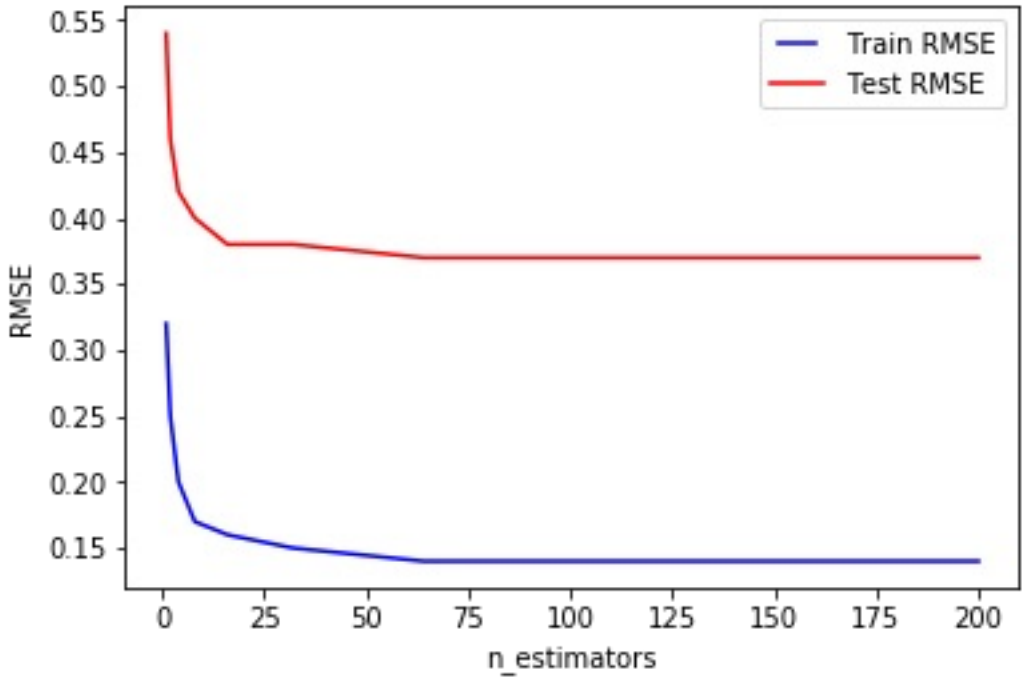
Next, let's use the entire dataset and feed it into some of the most common regression models to see what sort of root mean square error we get.

Experiment with several ML approaches

- Decision Tree Regression
  Root Mean Square error: 0.53
- XG Boost
  Root Mean Square error: 0.41
- Random Forest Regression
  Root Mean Square error: 0.378
- Neural Network
  Root Mean Square error:0.396

Random Forest model yields the lowest RMSE, followed by Neural Netork, XGBoost and Decision Tree
Since random forest yielded the best performance, let's keep optimizing it:

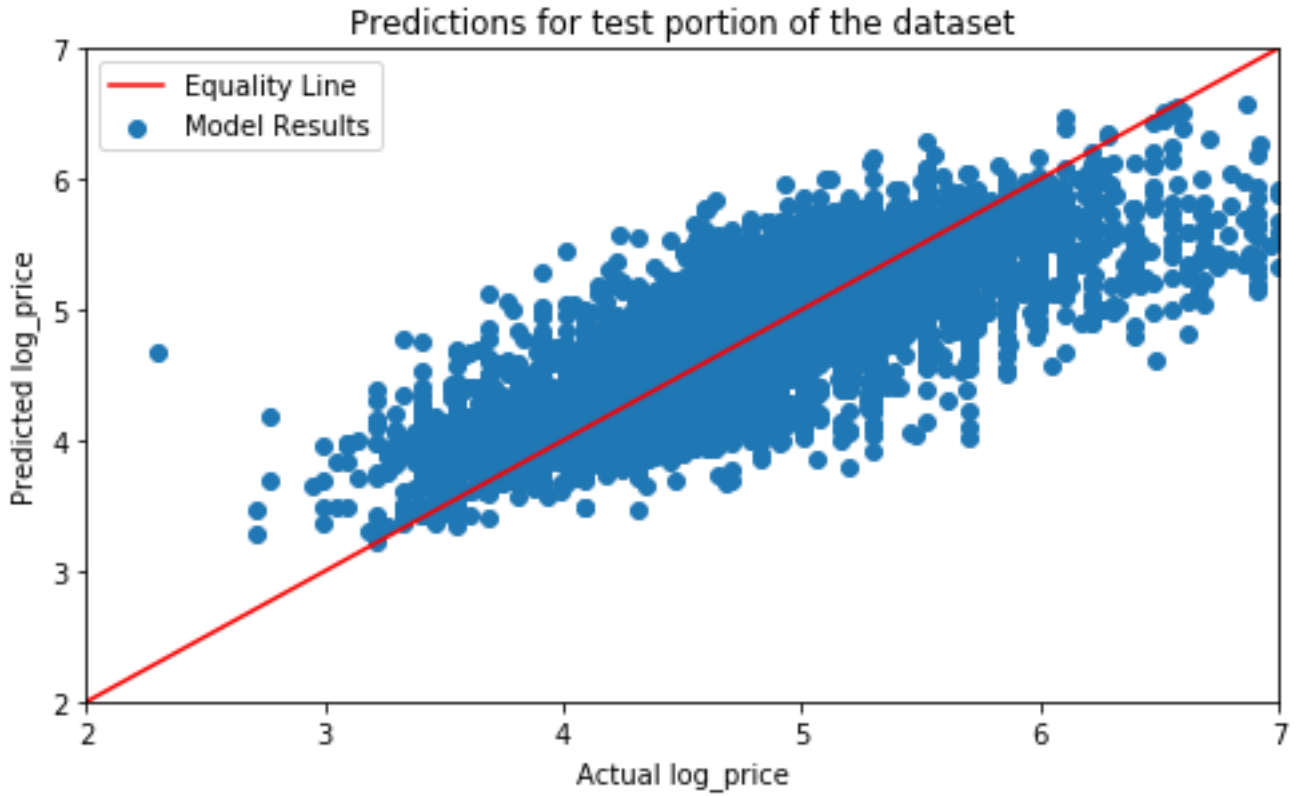
Optimizing number of estimators



•Clearly, the model overfits since the RMSE for train dataset is much lower then the test dataset regardless of the number of estimators

•For the best model performance on unseen data, a minimum of 100 estimators appears to be sufficient

Results of the model



On the low end of the *actual log_price* the results tend to cluster above the line, while on the high end they tend to cluster below the line.

This has consequences of underpredicting high prices and overpredicting low prices.

Important features in the model

	importance
Entire home/apt	0.442623
longitude	0.107530
latitude	0.098827
log_(availability_365 +1)	0.053276
Manhattan	0.035897
log_(reviews_per_month +1)	0.032671
log_(minimum_nights +1)	0.031870
log_(number_of_reviews +1)	0.028082
log_(calculated_host_listings_count +1)	0.017982
studio	0.015370
Shared room	0.007121
Midtown	0.004811
apartment	0.004306
bath	0.003757
loft	0.003705

The most important features that factor into the price of a listing are:

- Listing type (if it is a home/apartment)
- Availability and review related factors
- Certain listing descriptor words indicating the character or location of a listing
- Location, which is very intuitive considering that in real estate location is often a deciding factor for price

Conclusion

- Random Forest regression model provided best accuracy for prediction of listing price based on variables generated from the initial data
- The model as is tends to underpredict listings priced relatively high
- The model tends to underpredict listings priced relatively low
- The model importances can be used to further understand what drives the price of an Airbnb listing in NYC
- RMSE are given based on log_price values