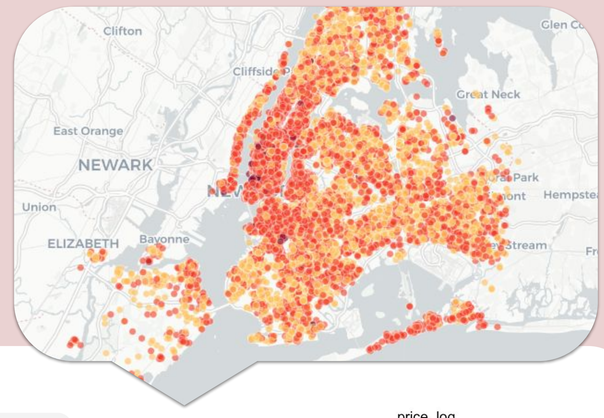
NYC Airbnb Listing Price Prediction

DATA2020-Spring2023 Final Project
Juefei Chen, Tongzhao Liu, Zhangqi Liu, Chuning Xiao, Cynthia Zhang
GitHub: https://github.com/CynthiaCZ/DATA2020-final-project.git

Data Science Initiative
BROWN

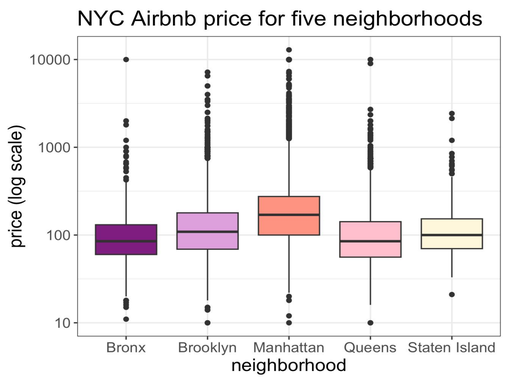
Introduction

In this project, we explore the NYC Airbnb market, analyzing its listing prices and predicting the prices of future listings using statistical learning techniques. We will utilize various data analysis and machine learning algorithms to build a predictive model that can estimate the price of an Airbnb listing based on various factors such as location, number of bedrooms, amenities, and reviews. Our aim is to provide insights that can be useful for both Airbnb hosts and guests, as well as to gain a better understanding of the NYC Airbnb market.

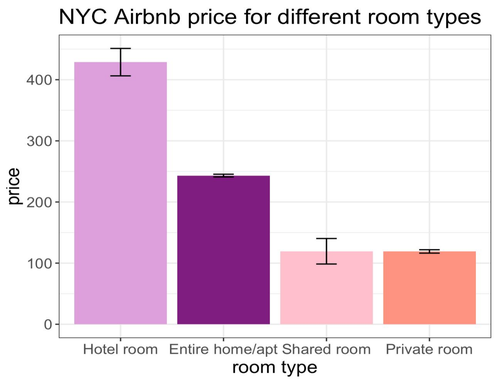


EDA

- In log scale, the price approximately follows a normal distribution with a slightly negative skew, which means the original price is strongly negatively skewed.
- The median is around $150.

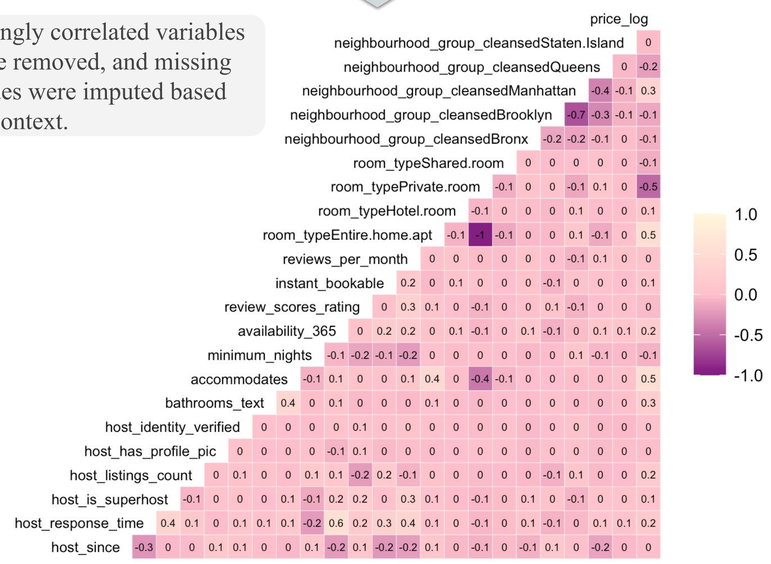


- The price distributions of all 5 neighborhoods are similar.
- Manhattan has the highest median price and the Bronx has the lowest.
- The outliers of the Bronx and Queens deviate the most.



- Hotel rooms have the highest price, and shared rooms or private rooms have the lowest price
- Hotel room prices have the highest standard deviation, whereas entire home/apartment prices have the lowest.

Strongly correlated variables were removed, and missing values were imputed based on context.



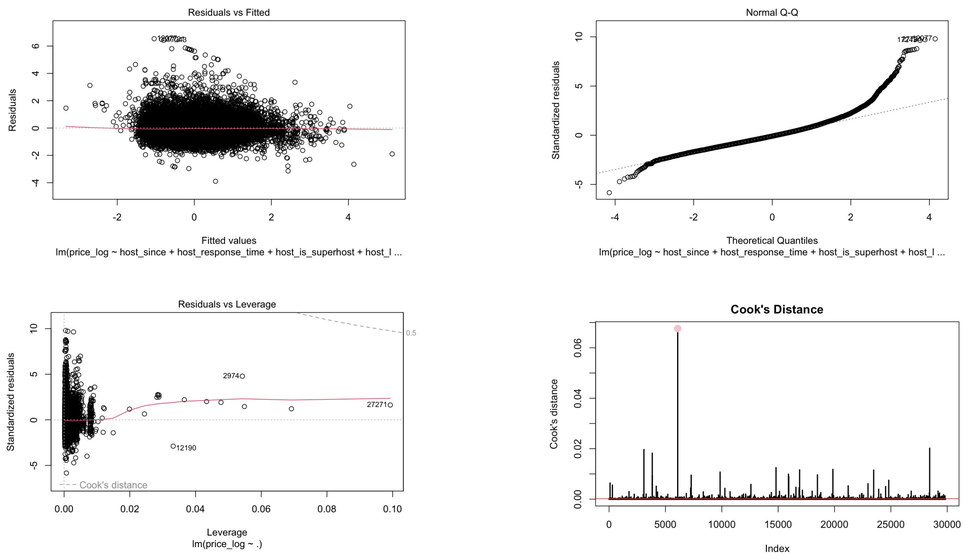
Correlation plot of selected and processed features

Methods

Simple Linear Regression

Predictor variables: host_since, host_response_time, host_is_superhost, host_listings_count, host_has_profile_pic, host_identity_verified, neighbourhood_group_cleansed, latitude, longitude, room_type, bathrooms_text, accommodates, price, minimum_nights, availability_365, review_scores_rating, instant_bookable, reviews_per_month
Target variable: price_log

Check Assumptions



Check Assumptions: According to residual vs fitted value, Cook distance, QQ plot, and correlation plot, all assumptions of the linear model are satisfied.
Five methods of predictor selection: Linear, Full, Backward, Forward, Lasso, Ridge
Metrics: RMSE & R^2
The five models were trained on training data and evaluated on testing data. After the best model is identified based on the evaluation metrics, bootstrapping was performed to measure the uncertainty of its predictions.

Multilevel Linear Regression

Multilevel models (varying intercept and varying slope & intercept) were utilized to examine the variation in the impact of the most significant predictor, type_entire, on price in different neighborhoods.

Results

Simple Linear Regression

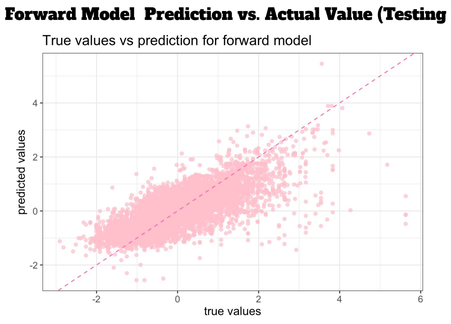
Performance Comparison of Linear Models

Model	RMSE	R-Squared
Baseline	0.9999866	0
Full	0.6655835	0.5604641
Forward	0.6655275	0.5605380
Backward	0.6655275	0.5605380
Lasso	0.6656020	0.5604397
Ridge	0.6656068	0.5604397

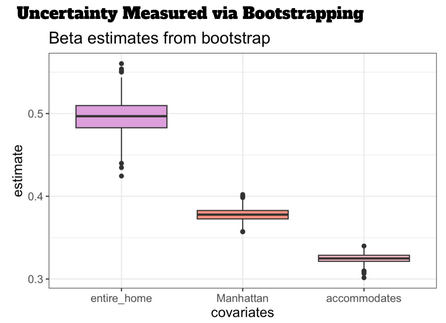
Feature Importance of the Forward Model (Top 5)

Feature	Estimate	95% CI	p-value
Entire home apt	0.496	[0.464, 0.529]	< 2e-16
Accommodates	0.324	[0.315, 0.333]	< 2e-16
Manhattan	0.373	[0.355, 0.392]	< 2e-16
Host listings count	0.127	[0.119, 0.135]	< 2e-16
Availability 365	0.086	[0.076, 0.095]	< 2e-16

Forward Model Prediction vs. Actual Value (Testing data)

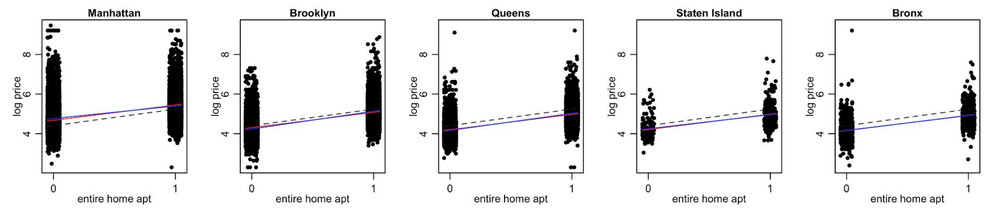


Uncertainty Measured via Bootstrapping



Although all five models demonstrated similar performance, the forward model showed the lowest RMSE and the highest R^2. In particular, the predictor variables that were deemed significant in the forward model are interpretable within the context of the problem.

Multilevel Linear Regression



Varying intercept (red): lmer(price_log ~ type_entire + (1 | neighborhood))
Varying intercept & slopes (blue): lmer(price_log ~ type_entire + (1 + type_entire | neighborhood))
These models revealed the variability in the effect of the predictor variable, type_entire, on the target variable, price_log, across different neighborhoods.

References

Data Source: https://www.kaggle.com/datasets/konradb/inside-airbnb-usa

Contributions: We have collectively worked on tasks such as data cleaning, processing, modeling, and coding. In terms of poster writing, Chuning Xiao composed the introduction section, while Juefei Chen delved into the exploratory data analysis. Tongzhao Liu, Zhangqi Liu, and Cynthia Zhang played key roles in formulating the methodology and showcasing the results. Additionally, Cynthia Zhang oversaw code management and Github operations, while Chuning Xiao handled poster formatting tasks.