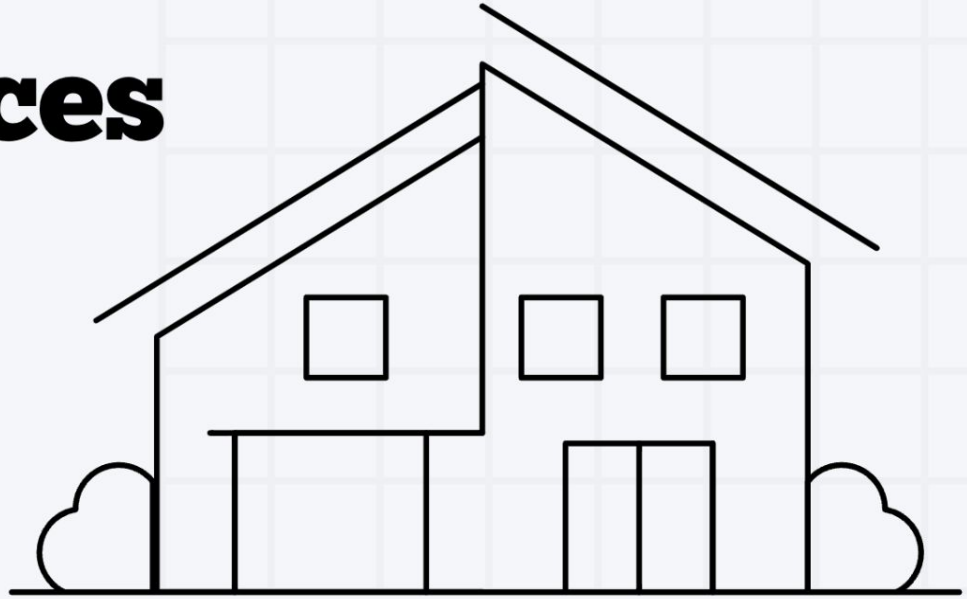
Optimizing Airbnb Listing Prices

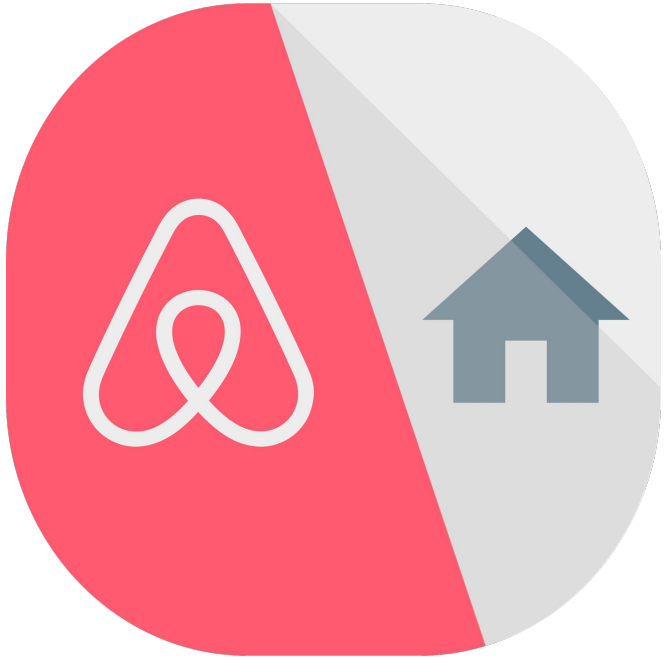
MIS S381N

Manasa Maganti, Téa McCormack, Vishwa Patel,
Alexander Schmelzeis, Gaurav Shukla, Mallika Singh



Why Airbnb prices...?

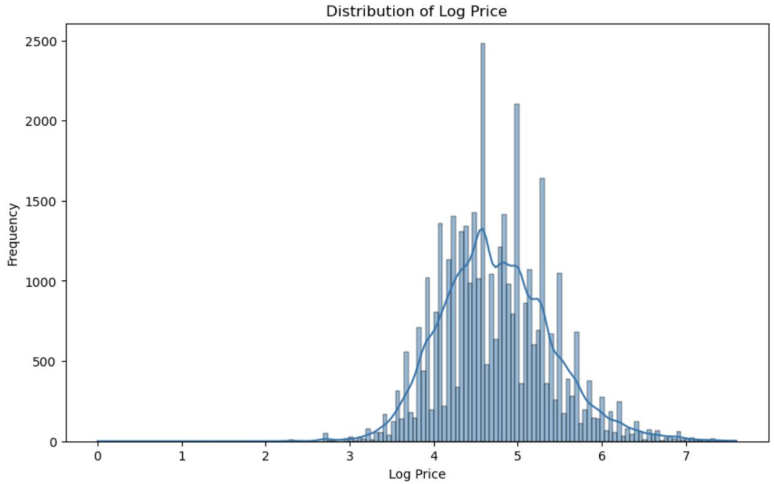
Airbnb serves various purposes—vacations, short-term stays, and even real estate investment. However, its **popularity** is **waning** due to rising prices and hidden fees. We want to **help assist hosts** in setting **competitive prices** while helping **guests** assess the **fairness of their stay**.



Today's Presentation

1. Data Cleaning / Preprocessing
2. Exploratory Data Analysis
3. Models
4. Conclusion

Data Cleaning and Preprocessing



Target Variable: log_price

- This was already logged in the original dataset
- Likely logged due to the skew of actual (exponentiated) price

- Treated outliers: removed log_price < 2 to improve model performance and prevent overfitting

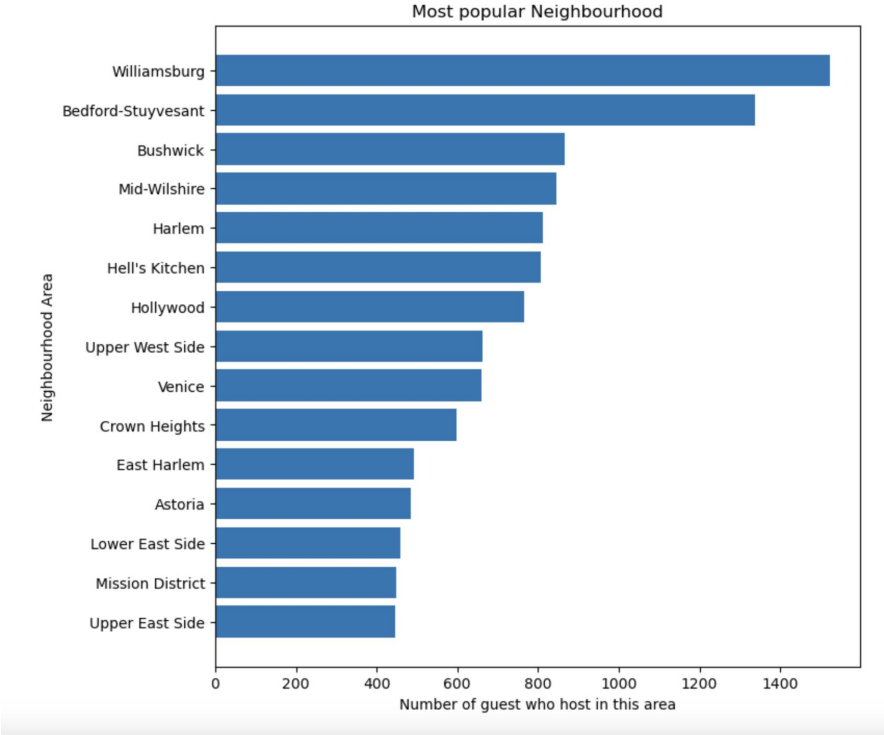
Data Cleaning and Preprocessing

amenities
{"Wireless Internet","Air conditioning",Kitche...
{"Wireless Internet","Air conditioning",Kitche...
{TV,"Cable TV","Wireless Internet","Air condit...
{TV,"Cable TV",Internet,"Wireless Internet",Ki...

Refrigerator	Pool	Breakfast	Firm matress	Dryer	Room-darkening shades	Carbon monoxide detector

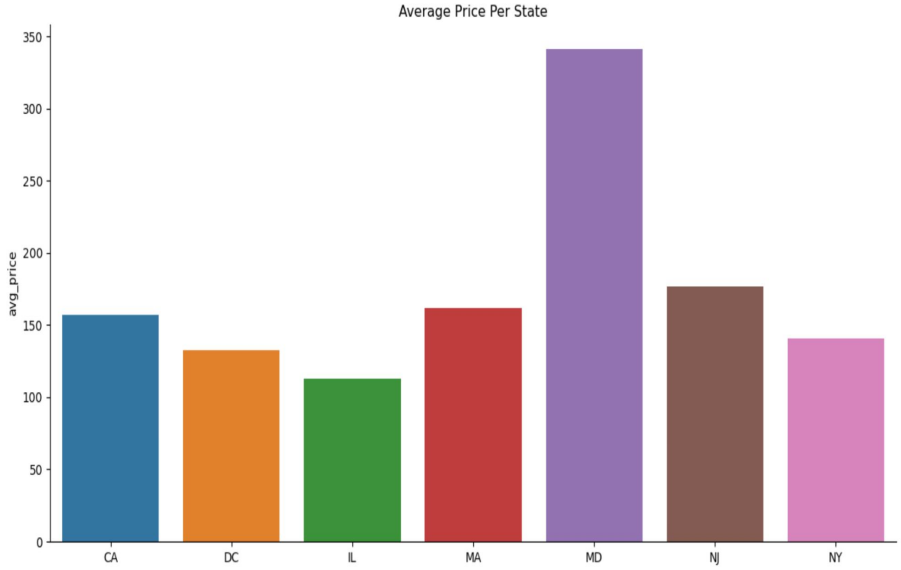
We added over **100 columns** when **dummy coding** as well as **extracting valuable data** from variables like ‘amenities’, ‘accommodates’, ‘property_type’, and so on. Our final dataset went from 74,111 rows and 30 columns to **38,166 rows** and **151 columns**.

Exploratory Data Analysis: Variable Specific

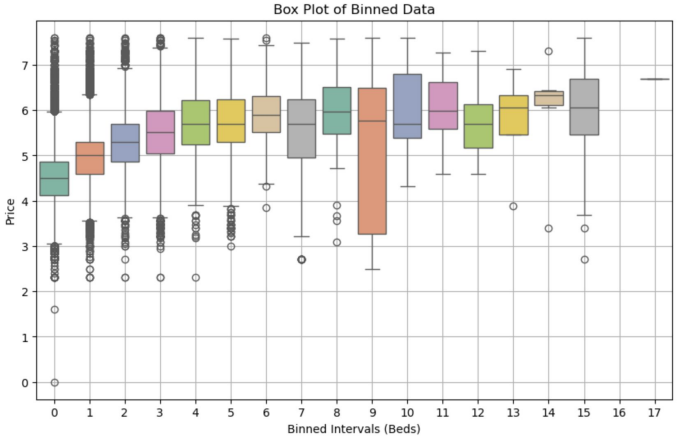


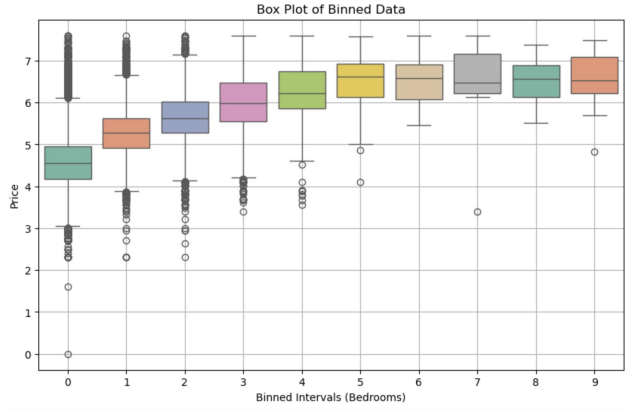
The most popular neighborhoods were primarily in New York, with the exception of Venice Beach and Hollywood in California.

The average price per state showed that Maryland had the highest average price. This led us to realize that Maryland had a significantly lower number of observations (11) when compared to the other states. The average was being skewed here.



Exploratory Data Analysis: Variable Specific

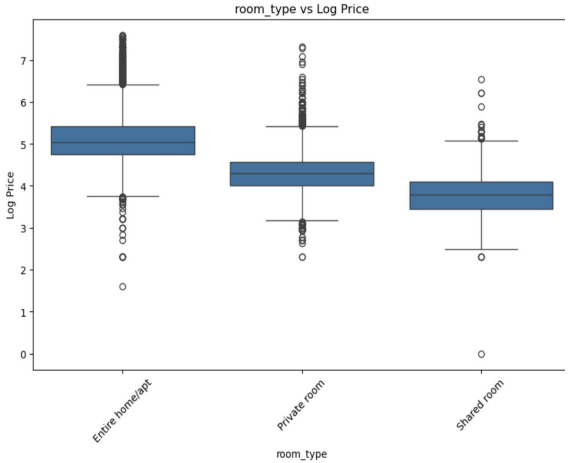


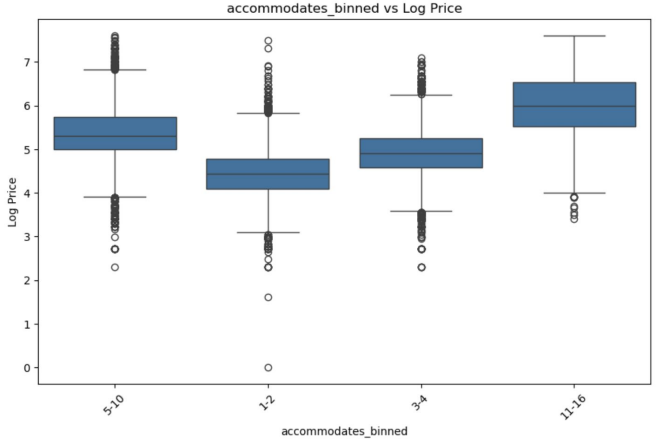


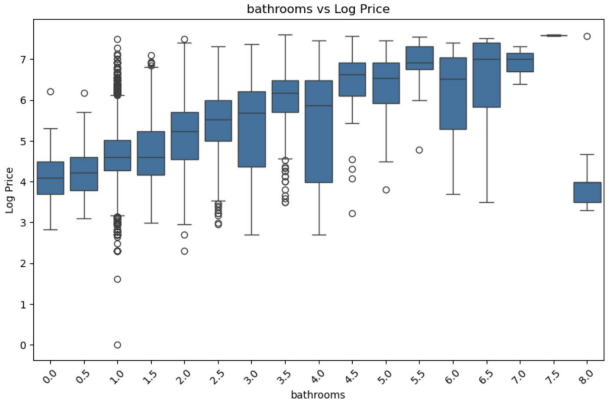
Beds and Bedrooms did not have much variation with price. The **correlation** between the predictors and log_price was around **0.45**. However, we start to see a **slight trend** that the **more beds/bedrooms present**, the **price range stretches** as well.

Our expectations were confirmed when looking at other variables, like **property type**, **number of bathrooms** and how many **people** can be **accommodated** in the Airbnb.

Exploratory Data Analysis

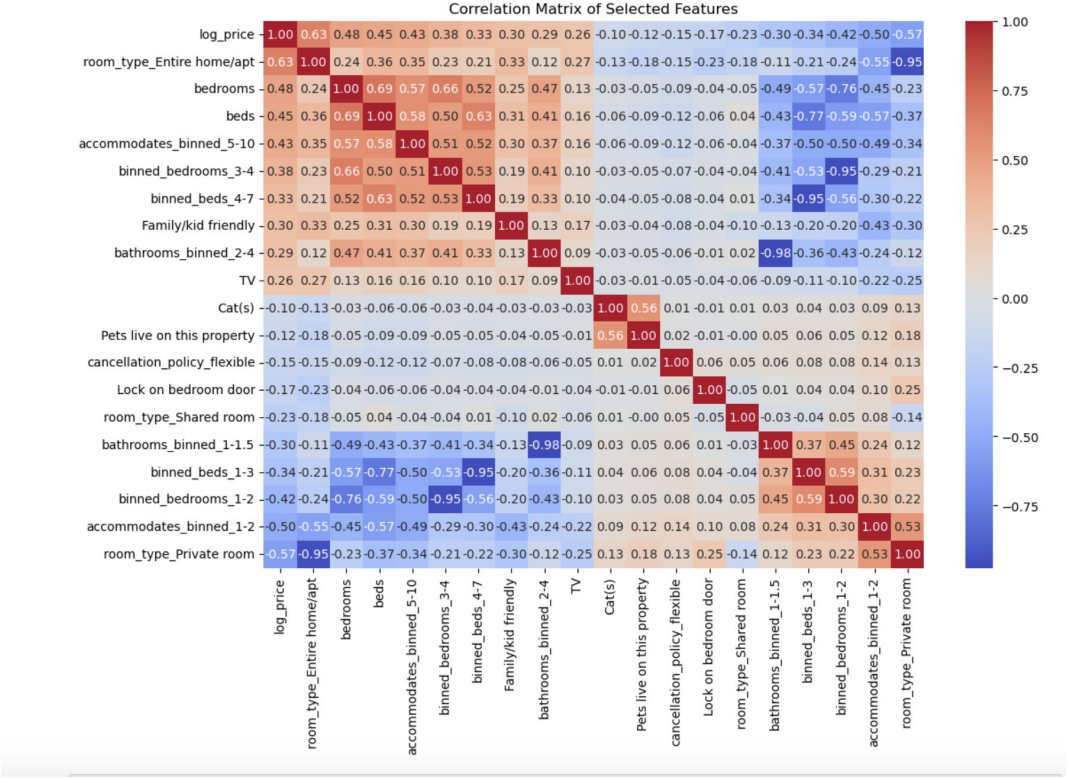






Log_price is on the higher end for entire homes that accommodate larger numbers of people, with more bathrooms.

Correlation Matrix of Selected Features



Features Most Highly Correlated with log_price:

- Room Type - Entire Home/Apt and Private Room
- # of Bedrooms and Bathrooms

Multiple Linear Regression

Variable Selection:

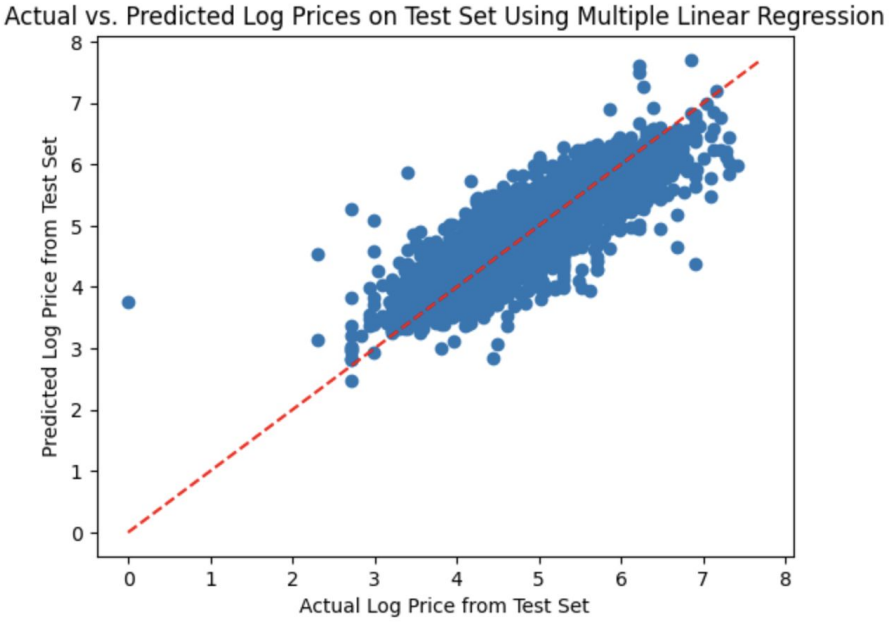
- Attempted to use forward stepwise subset selection, but too computationally expensive due to large number of categorical variables with unique values
- Model with all variables ended up producing lowest test error

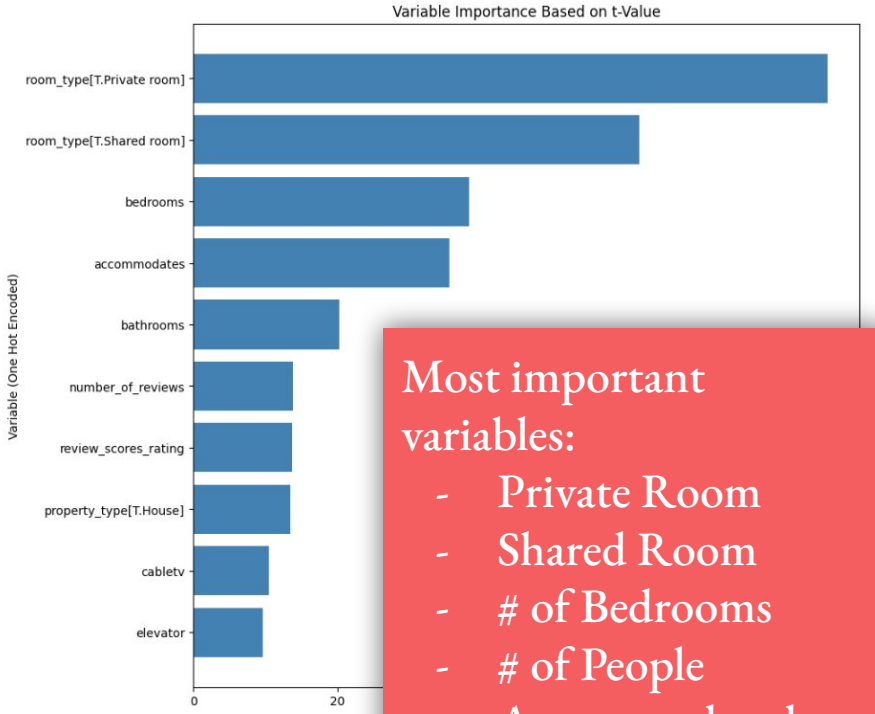
A few of the most important features (p-value < 0.05): property type, room type, city

79.9% of variation in log_price can be explained by predictors (the R^2 value from final model).

	In Sample (Train) RMSE	**Out of Sample (Test) RMSE**
log_price	0.295	0.343
price (exponentiated)	$68.29	$80.46

Multiple Linear Regression (cont.)

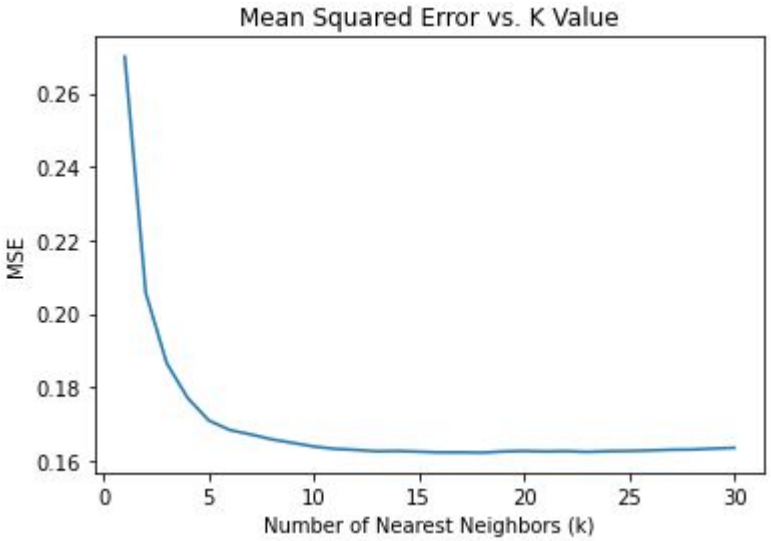




Most important variables:

- Private Room
- Shared Room
- # of Bedrooms
- # of People Accommodated
- # of Bathrooms

K-Nearest Neighbors (Regression)

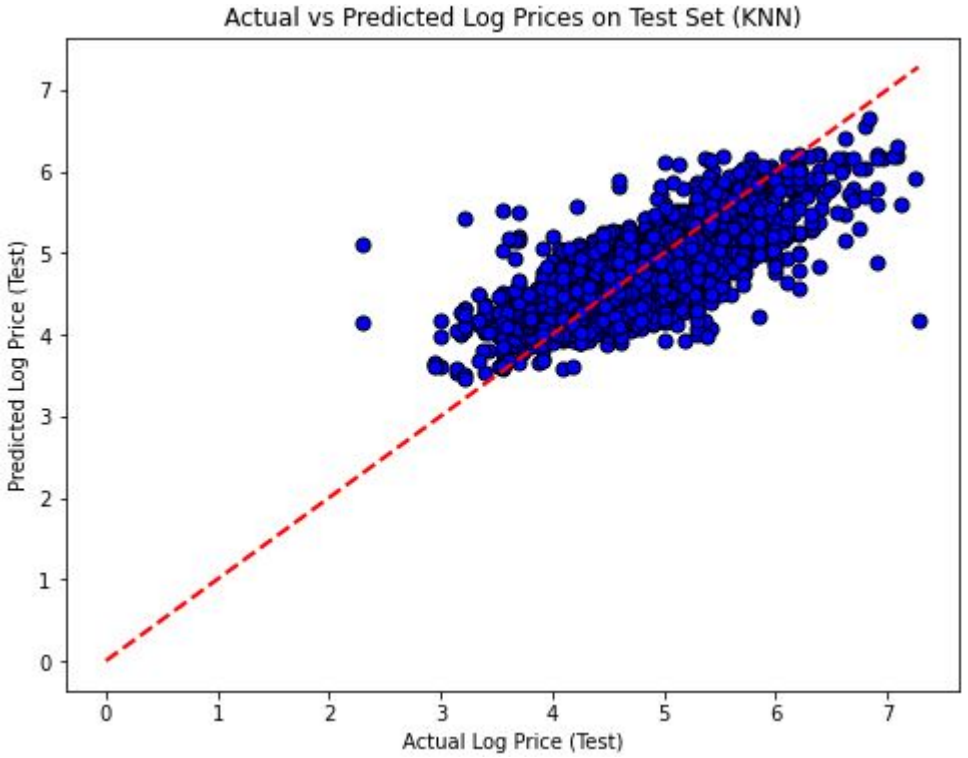


	log_price	price (exponentiated)
Train MSE	0.143	7,355.53
Test MSE	0.162	7,093.99
Train RMSE	0.379	$85.76
Test RMSE	0.403	$84.23

- Selected an optimal k value of **18** (done through cross-validation)
- Resulted in the lowest test MSE

- Only **62.30%** of the variation in log_price can be explained by predictors (R-squared)
- Tried performing permutation variable importance, but too computationally expensive

K-Nearest Neighbors - Actual vs. Predicted



Random Forest

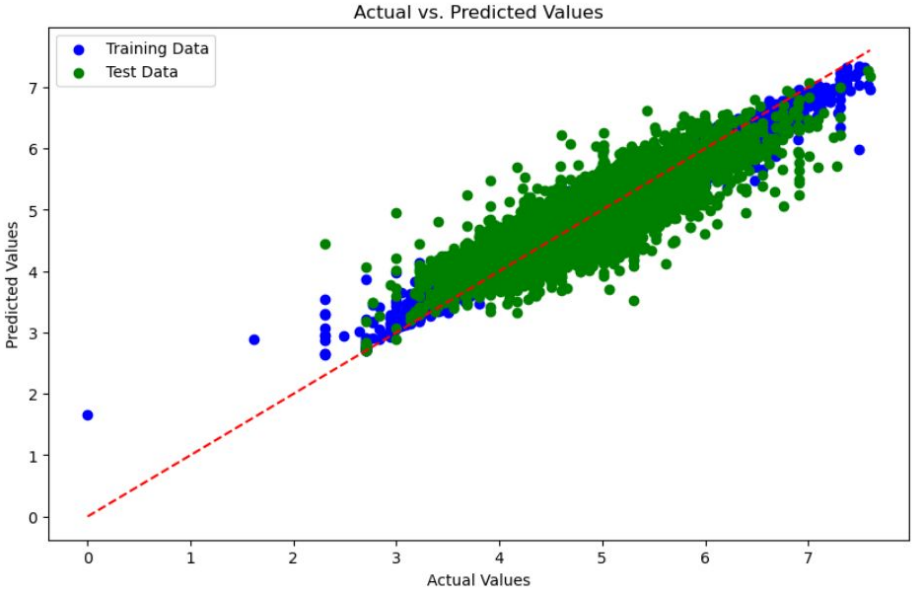
- Trained on **13 features** - Performed Lasso regularization method to remove less important predictors from the model
- Conducted 5-fold cross-validation on the training set which yielded the following values:

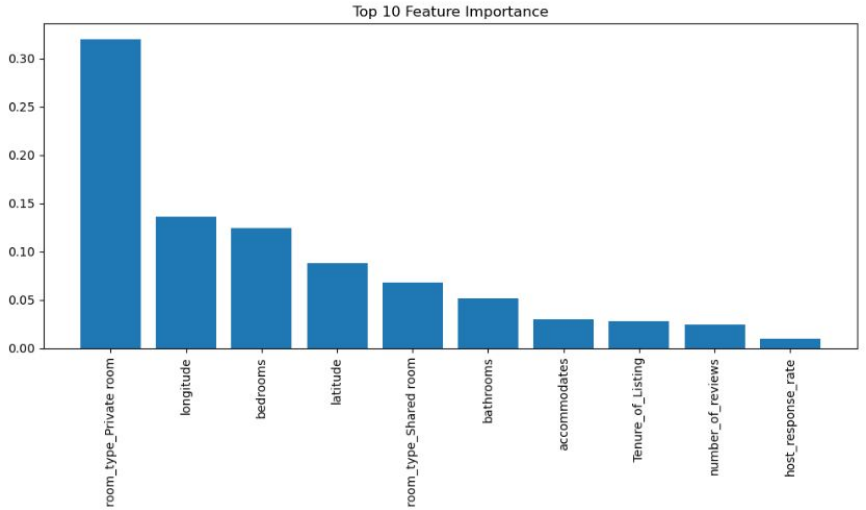
	log_price	**price (exponentiated)**
Train MSE	0.105	1225.17
Test MSE	0.108	5215.08
Train RMSE	0.324	$35.00
Test RMSE	0.330	$72.21

N-estimators: 100

R2 score on validation set: 68.89%

Random Forest - Actual vs Predicted



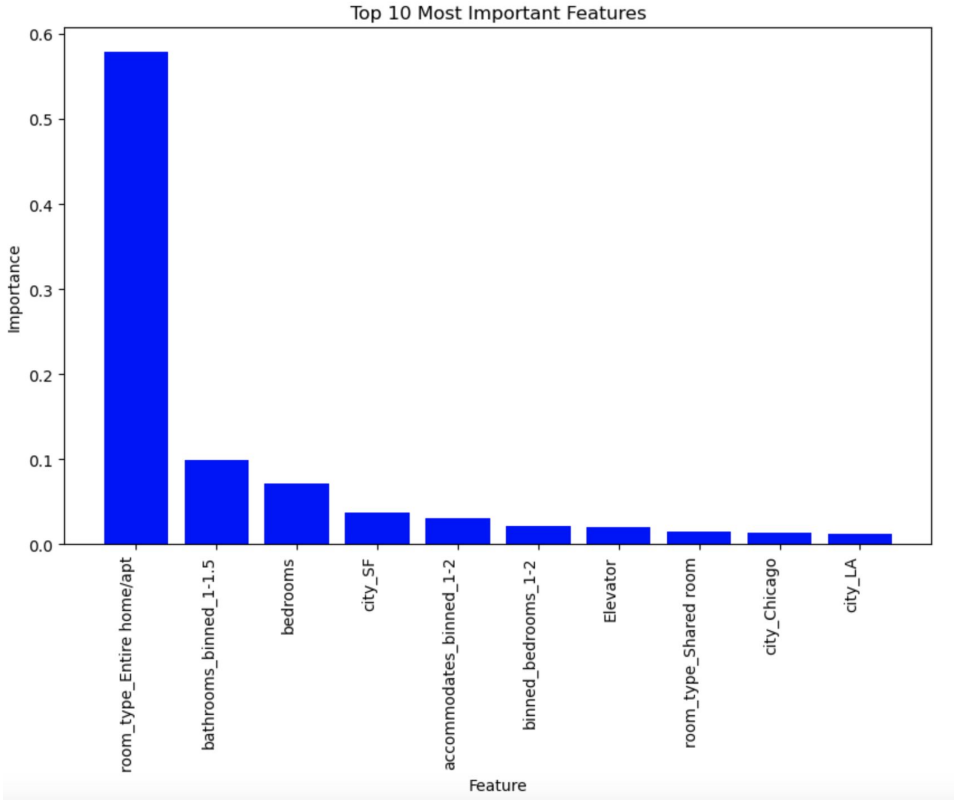


- The Random Forest model appears to overfit, as indicated by a substantial disparity between Train and Test RMSE
- Additional experimentation with variable selection is needed to improve the model's ability to generalize its predictions on the given data

Boosting

	log_price	price
In Sample MSE	0.1435	6852.84
Out of Sample MSE	0.1463	23293.78
In Sample RMSE	0.3788	$82.78
Out of Sample RMSE	0.3824	$152.62

It is important to note that over 50 variables had 0 importance

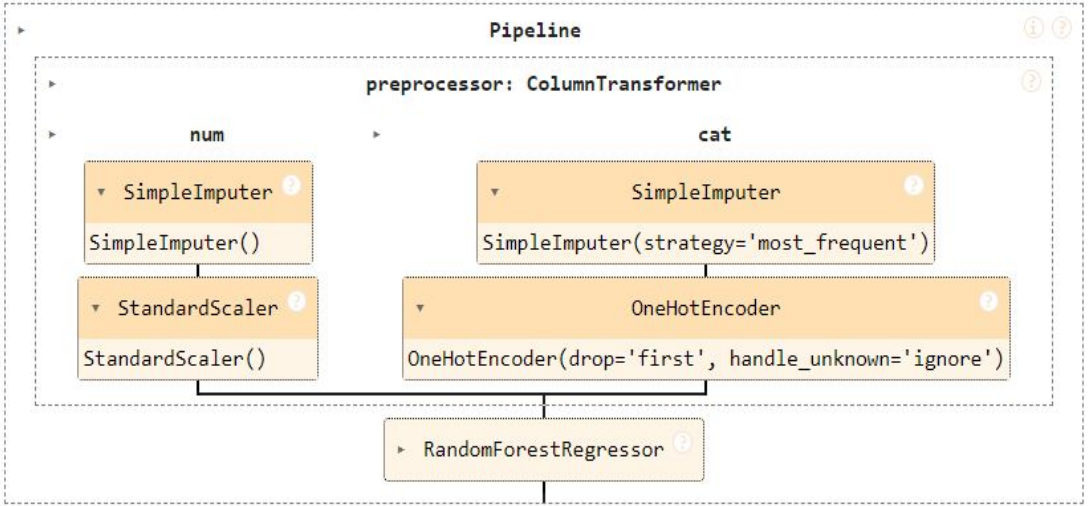


Insights to Report

- **Random Forests** achieved the **lowest Test RMSE** (0.330) but the model was **overfit**
- **Stepwise Selection** was not feasible due to **computational inefficiency**
- Recommendations for **Airbnb Hosts** to **optimize listing price** while keeping guest satisfaction in mind:
  - Focus on **Property Type**
  - Optimize **guest capacity**
  - Consider the **number of bathrooms**

Thank you

Data Cleaning and Preprocessing



- Processed numerical, binary and categorical features through different encoding methods
- These features are then passed into the model for processing

Incorporated ‘Tenure Listing’

Compute duration between the first and last review, to understand how reliable the property is:

Tenure_of_Listing = difference in days between last_review and first_review