# Airbnb Pricing Predictions in Amsterdam, The Netherlands Using Multiple Linear Regression

## Motivation

- Airbnb has become an increasingly popular option for travellers as an alternative to hotels.
- Knowing what influences pricing helps hosts optimize pricing.
- Informing travellers of what factors influence pricing ensures they can find the best value.

**Research Question:**

- How does the age, accommodations, and reviews of an Airbnb in Amsterdam influence its price?

## Data Collection

### Data Source

- Sourced from data.world, collected by Philip E. Cannata[1].
- Method of collection is unknown.
- Cannata is a retired professor from the University of Texas with 18 years of experience in teaching data science[2], giving credibility to the source.

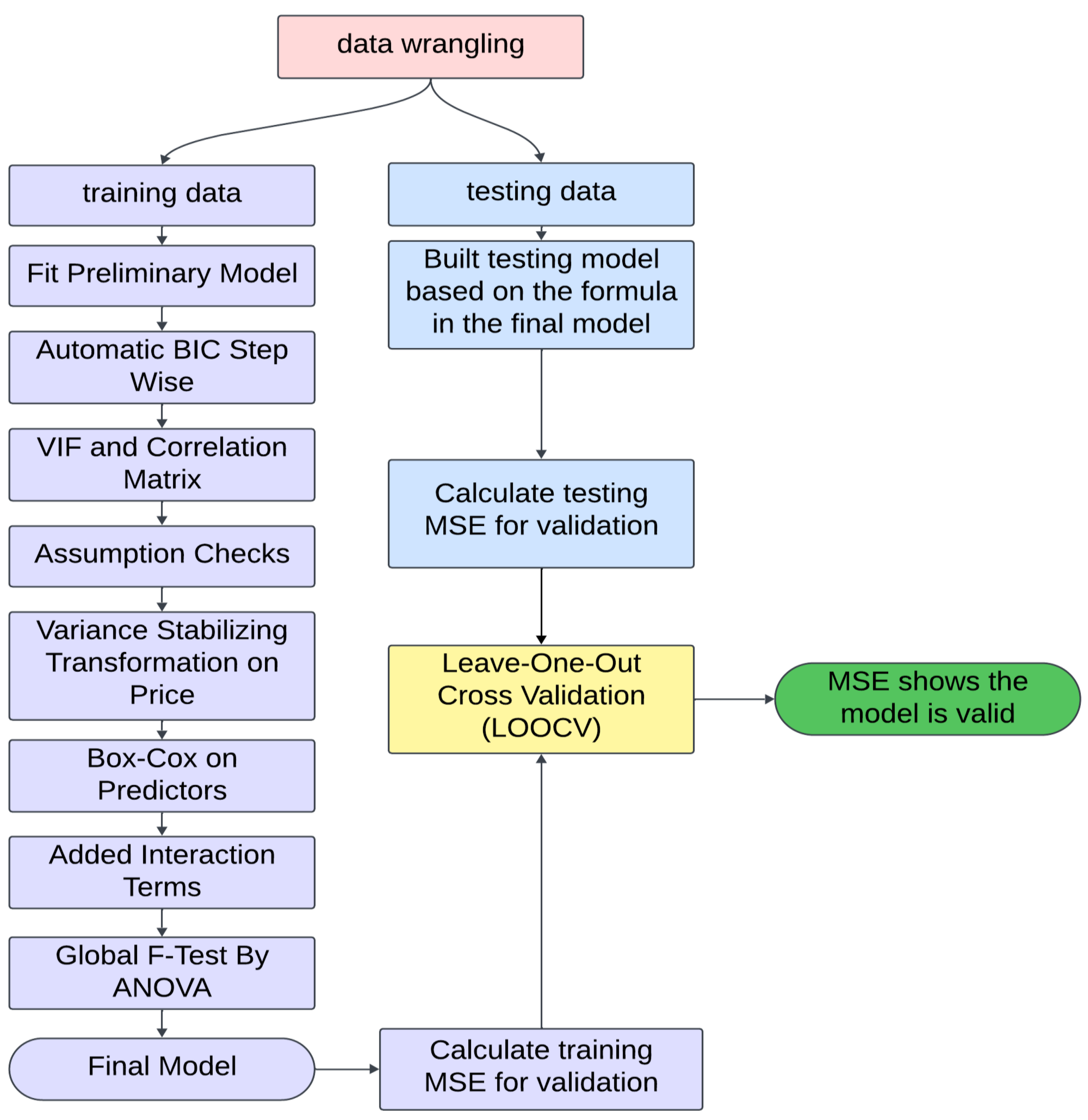
### Data Relevance

- The data focused on Airbnb listings in The Netherlands, with the majority being in Amsterdam.
- Able to generalize pricing factors within Amsterdam due to large dataset ($n = 5572$).
- Data contained variables relevant to research question.

## Methods of Analysis

- A MLR model was suitable because price is a continuous response variable with multiple factors influencing its value.

### Decision Making Flowchart




## Analysis and Results

### Preliminary Model Predictors

- Property Type
- Property Age
- # of Bedrooms
- # of Bathrooms
- Number of Reviews
- Review Rating
- Beds
- Minimum Nights
- Neighbourhoods
- Host Response Rate

## Analysis and Results

### Automatic Step-Wise BIC Selection

- Removed minimum nights, property age, and host response rate predictors due to relatively low BIC value.

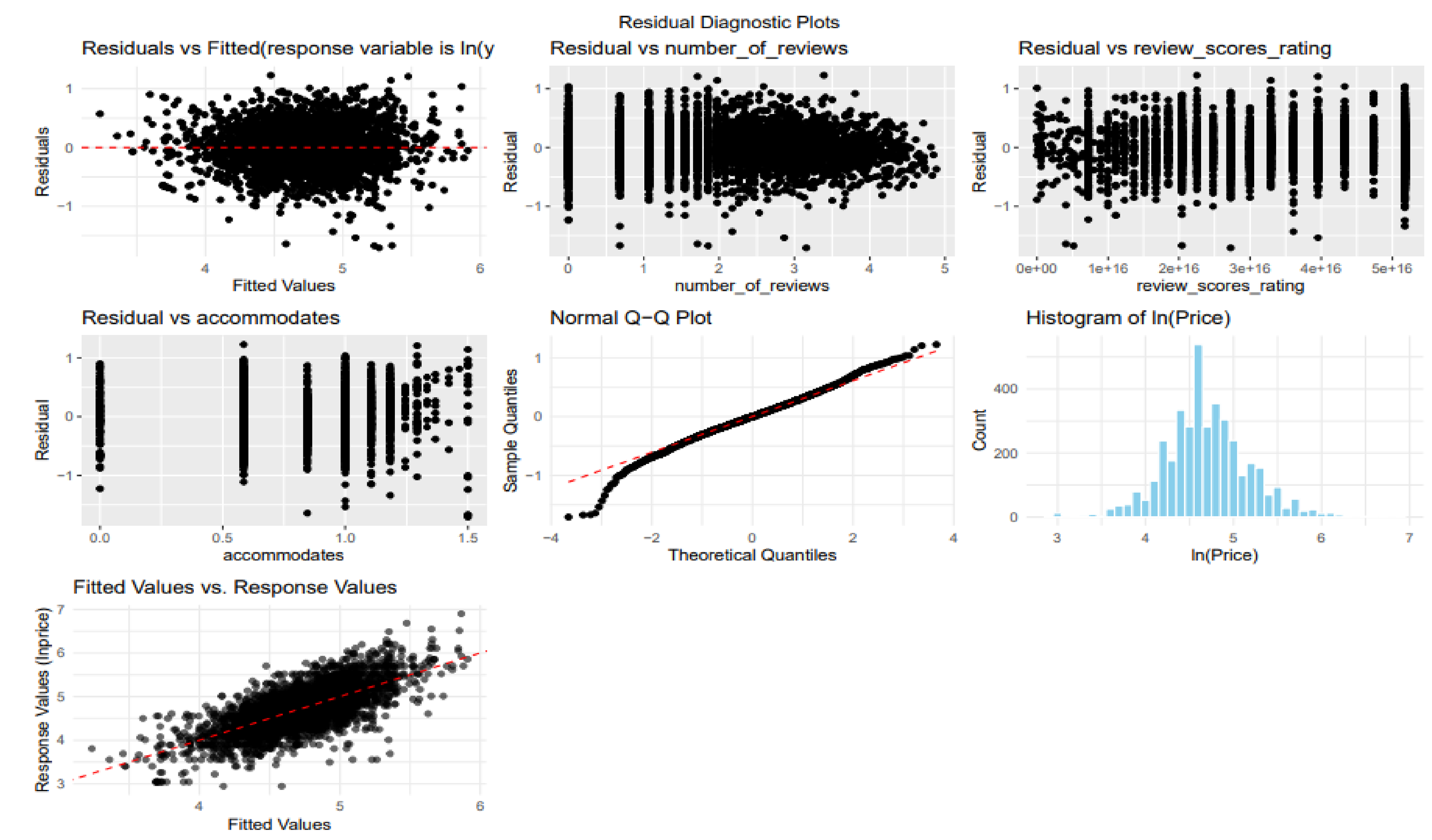
### Variance Inflation Factor (VIF) and Correlation Matrix

- Removed minimum nights, property age, and host response rate predictors due to high VIF and covariance values.

### Assumption Checks

- The residual plots showed violation of constant variable. To correct this, the natural log on the response variable was taken and Box-Cox was performed on the predictors.

### Residuals Plots After Transformations




### Model Improvements From Preliminary to Final

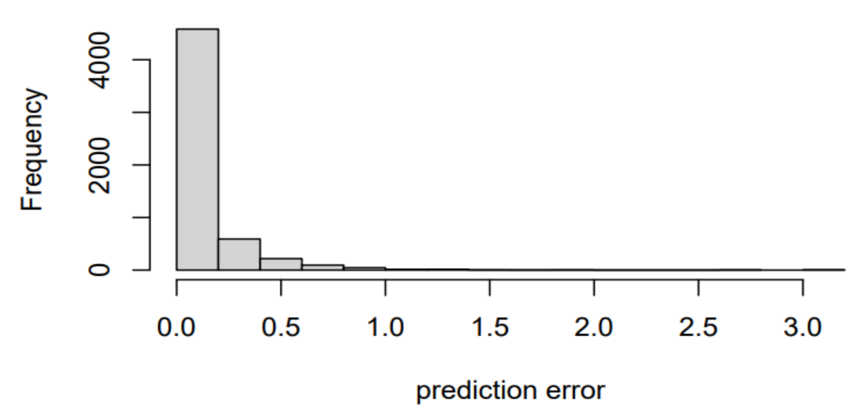
| Model | R_Squared | Adj_R_Squared | AIC | BIC |
|---|---|---|---|---|
| Model 1 | 0.5229616 | 0.5192617 | 41251.252 | 41451.844 |
| Model 4 | 0.4822821 | 0.4785364 | 2770.514 | 2958.568 |
| Final Model | 0.5020004 | 0.4924960 | 2709.111 | 3179.247 |

### Train-Test Validation

| Model | MSE | MAE |
|---|---|---|
| Train Model on Test Data | 0.1289892 | 0.2765025 |
| Train Model on Train Data | 0.1128691 | 0.2587314 |
| Test Model on Test Data | 0.1197326 | 0.2655872 |

- MSE & MAE values were similar, model was not underfitting/overfitting.

### Leave-One-Out Cross Validation Prediction Error




- MSE of residuals on testing data is similar to training data MSE.

## Conclusions

- Property type, number of reviews, review scores, number of people accommodated, and the neighbourhood influenced pricing the most.
- Property age did not have a large influence on price.

## Limitations

- Data did not include seasonality, limiting accuracy of findings.
- Proximity to central business districts impacted pricing in previous studies[3], but analyzing each Airbnb was beyond the project's scope.

## References


1. P. E. Cannata, "GAAirbnb," Data.world, 2017. [Online]. Available: https://data.world/cannata/gaairbnb.
2. P. E. Cannata, "Curriculum Vitae," Self-published, Austin, Texas, 2024.
3. A. Lawani, M. R. Reed, T. Mark, and Y. Zheng, "Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston," Regional Science and Urban Economics, vol. 75, pp. 22–34, 2019. [Online]. Available: https://doi.org/10.1016/j.regsciurbeco.2018.11.003