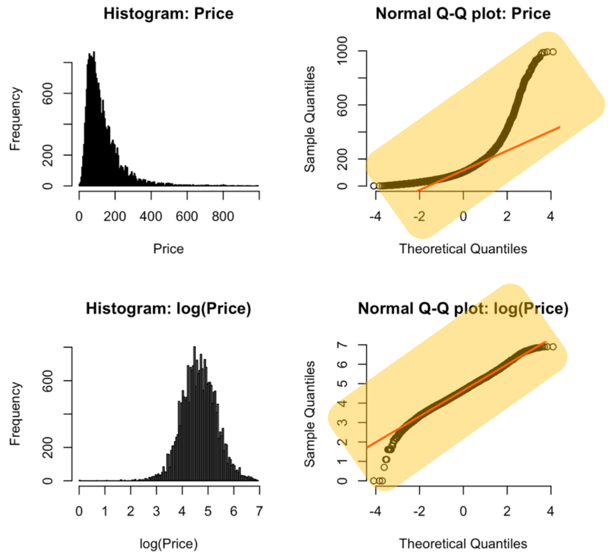
Predicting Airbnb rental price in NYC

Jisu Baek

Overview of Dataset and Preprocessing & EDA



1 **Goal**: To predict the Airbnb rental prices in NYC

Dataset: Train (N=23,313 with 96 variables)
Test (N=5,829 with 95 variables)

2 Preprocessing and Exploratory Data Analysis (EDA)

- Variables with >20% NA values were excluded from the analysis
- Categorical variables with too small observations were recategorized (e.g. `property_type: Apartment vs. Non-apartment`)
- Generated `amenity_num` that counts the number of amenities for each listing
- Outcome variable was log-transformed as the original was too skewed.
- As an EDA, visually checked the relationship among variables using `GGally::ggpairs()` function.
- After removing unnecessary variables, 42 variables were left in the dataset.

Feature Selection and Modeling

3 Feature Selection: Univariate Analysis & LASSO

- Univariate analysis was performed: log(Price) ~ each variable
    - Only significantly associated with the outcome will be left in the analysis.
- LASSO regression was also used to select the meaningful features.
- After those two feature selection process, **23 variables** were kept in the dataset.

4 Modeling: Random Forest is a winner!

- Four modeling methods became the finalists: Multiple Linear Regression, Decision Tree with Tuning, Random Forest, and Boosting model
- Top 3 most important variables through RF and Boosting Model: `room_type`, `longitude`, `accommodates`

Models	RMSE	R2
Multiple Linear Regression	0.415	0.622
Decision Tree with Tuning	0.418	0.616
Random Forest	0.175	0.942
Boosting Model	0.414	0.633

Variables	pval	correlation
accommodates	<0.001	0.538
beds	<0.001	0.424
guests_included	<0.001	0.357
bedrooms	<0.001	0.345
review_scores_location	<0.001	0.202
amenities_num	<0.001	0.182
bathrooms	<0.001	0.137
extra_people	<0.001	0.126
review_scores_cleanliness	<0.001	0.094
review_scores_rating	<0.001	0.078
latitude	<0.001	0.073
review_scores_accuracy	<0.001	0.050
review_scores_communication	<0.001	0.042
review_scores_checkin	<0.001	0.031
availability_365	<0.001	0.024
number_of_reviews	0.003	0.019
availability_30	<0.001	-0.026
reviews_per_month	<0.001	-0.044
availability_60	<0.001	-0.048
availability_90	<0.001	-0.057
calculated_host_listings_count	<0.001	-0.157
longitude	<0.001	-0.339
host_is_superhost	0.028	
host_identity_verified	<0.001	
neighbourhood_group_cleansed	<0.001	
is_location_exact	<0.001	
property_type	<0.001	
room_type	<0.001	
bed_type	<0.001	
instant_bookable	<0.001	
is_business_travel_ready	<0.001	
cancellation_policy	<0.001	
require_guest_phone_verification	0.013	

Discussion and Further Improvements

5 Review of My Analysis

- **Good**: Feature selection process was thoroughly designed and filtered using multiple methods: univariate analysis & LASSO.
- **Difficult**: Extracting meaningful information from the variable `amenities`

6 Future Improvements

- Handling outliers and imputing missing values could be better.
  - Check if there's any patterns in missing values
  - Impute proper values into the missing values that minimizes the bias

7 Takeaways

- Top 3 most important variables through RF and Boosting Model: `room_type`, `longitude`, `accommodates`
- `Longitude` was much more important than expected
  - Why? (Inference): Manhattan looks vertically longer rectangle and many tourist attractions were located in the mid-town areas, below Central Park.
  - Negative correlation: As the longitude decreases, the rental price increases.