Data Analysis on Airbnb Dataset

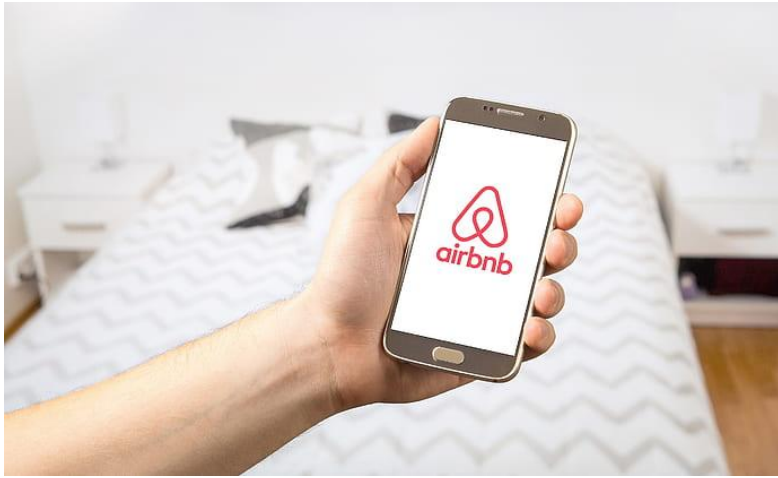
in Los Angeles Area

COMM 557 Midterm Presentation

Kelly Choy
Shih-Min (Julia) Huang
Tongxin (Shirley) Ye

Project Details

- Analyze short-term rental markets in LA
- Dataset:
  - Provided by Inside Airbnb (http://insideairbnb.com/get-the-data/)
  - Dataset time frame: 9/4/2023



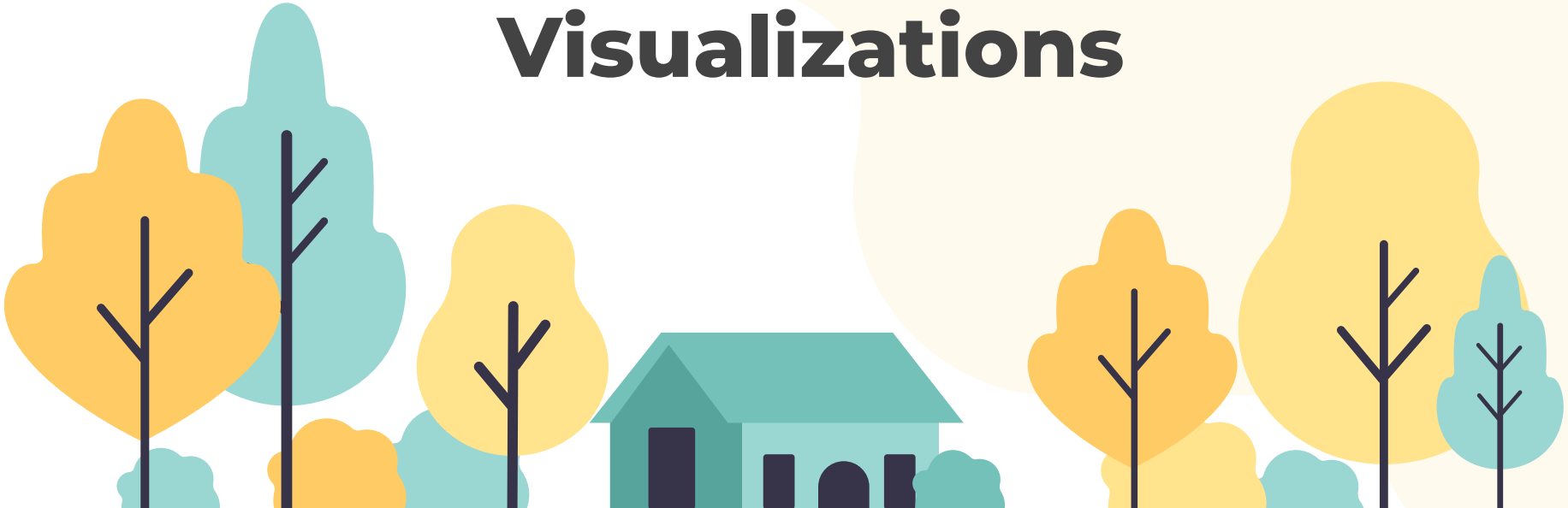
Problem

- Individual look at these features when selecting which Airbnb to book:
  - price, location, reviews, ratings of hosts, etc
- One main factor that people look at is whether the homeowner is a Superhost
  - How do homeowners receive the “superhost” title?
    - If a new homeowner wants a higher chance of people selecting their home to stay in, they need to achieve that title and reputation
- Goal → to identify what makes a homeowner a Superhost

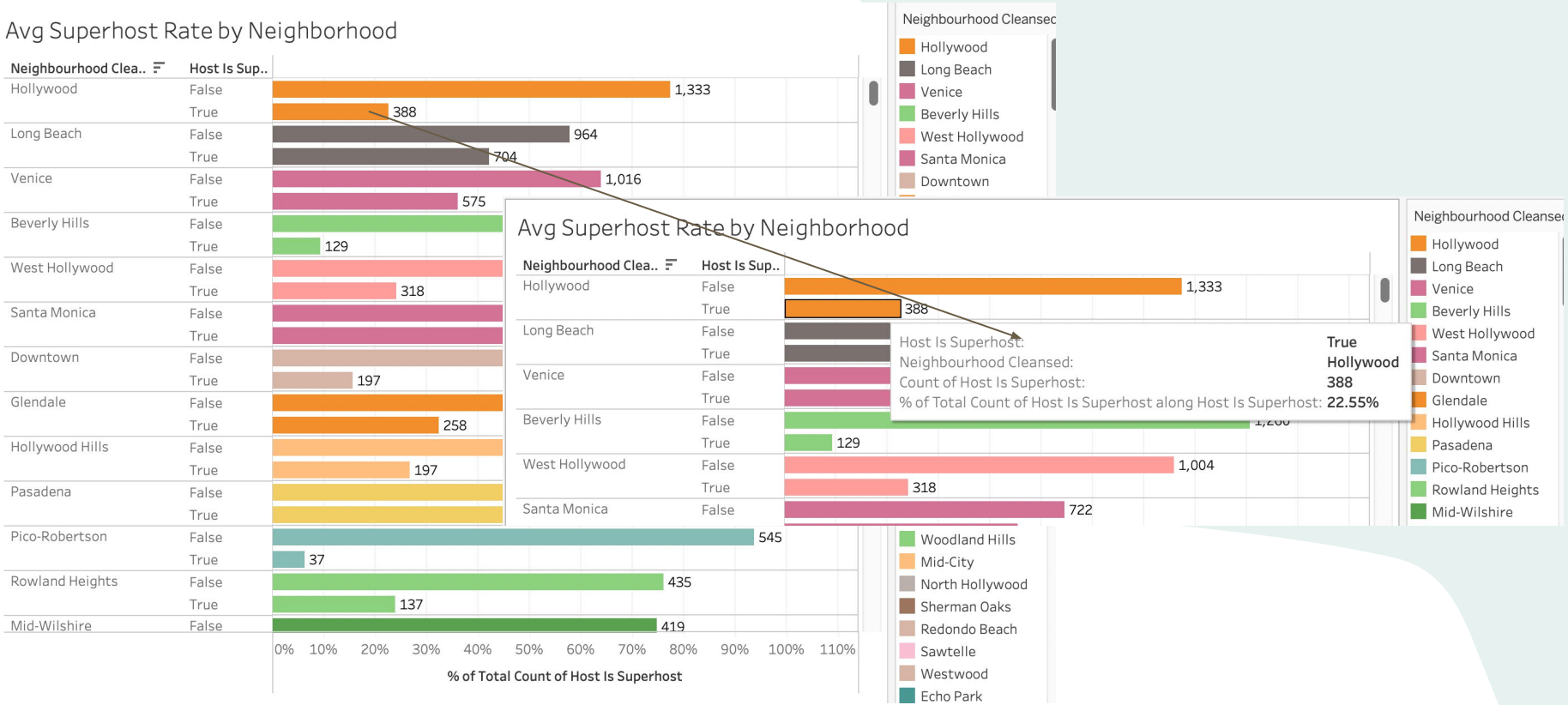
Methods

- **Mapping Spatial Distribution: Visualizing Geographic Patterns**
  - Utilize geographic visualizations to highlight high-density areas and identify hotspots effectively
  - Analyzing and visualizing data pertaining to pricing, superhost rate, etc, across diverse neighborhoods
- **Utilizing Machine Learning (classification) for Superhost Prediction:**
  - Identifying the distinguishing factors that set Superhosts apart
    - Utilize tree-based methods, e.g. boosting
  - Make predictions on whether a host will attain Superhost status or not

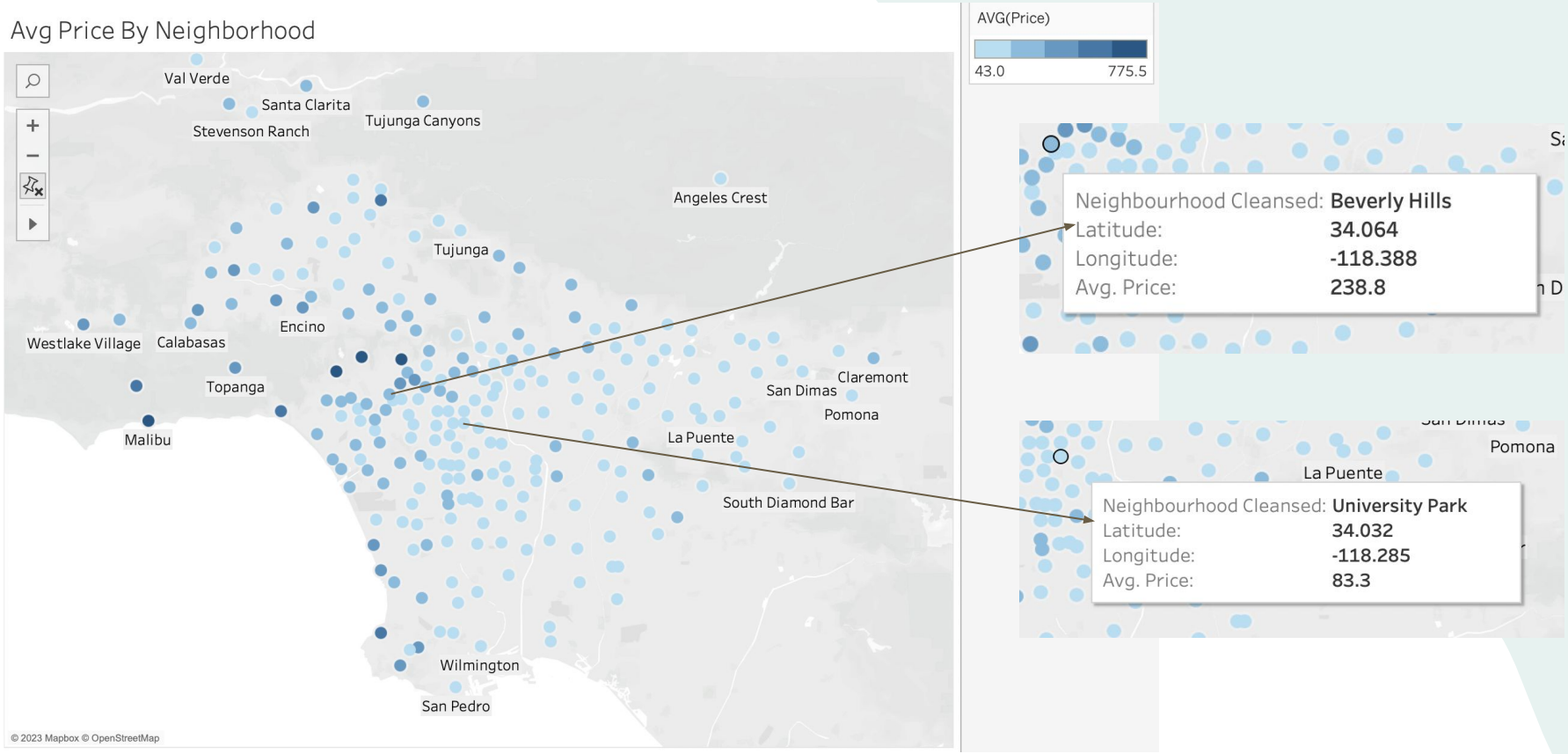
Exploratory Data Analysis and Visualizations



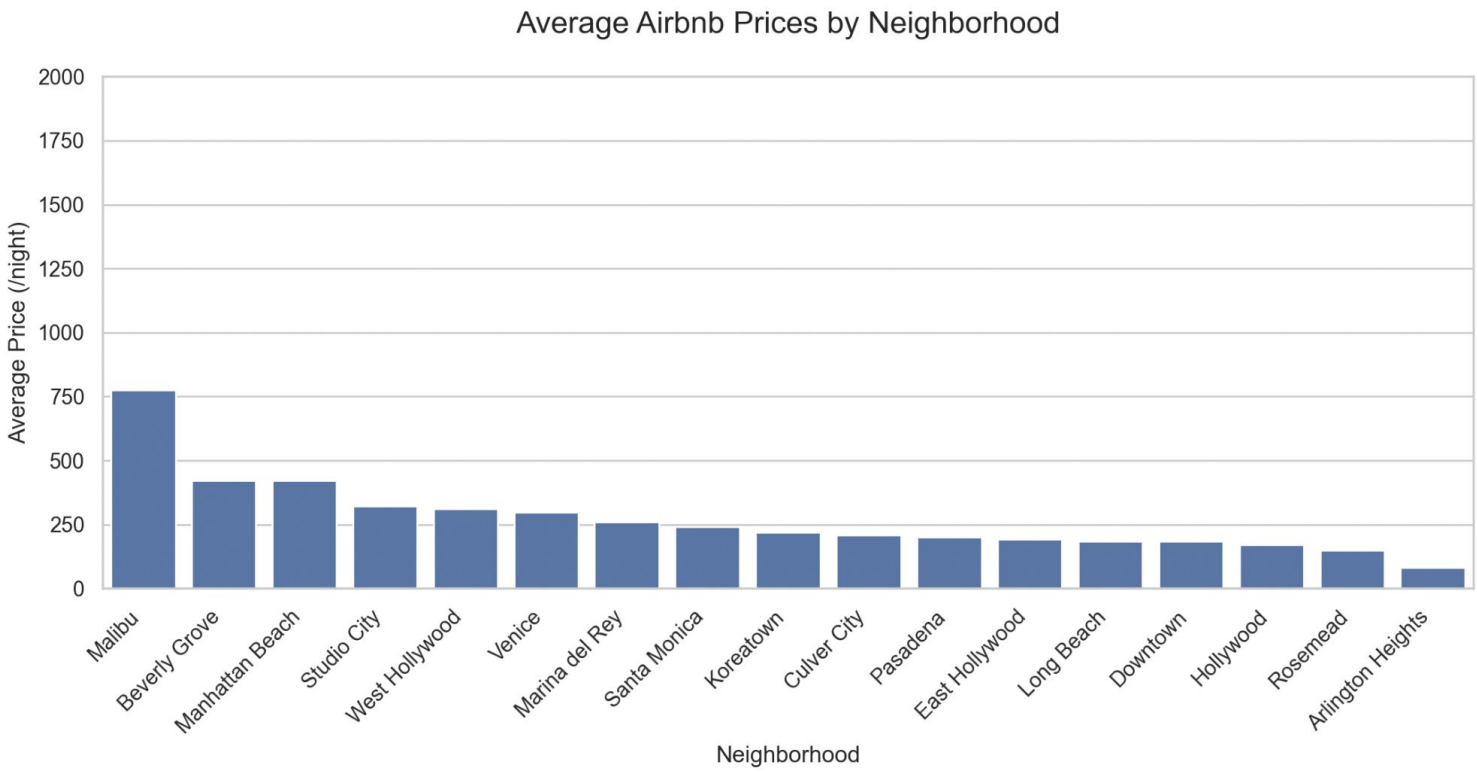
Average Airbnb Superhost Rates by Neighborhood



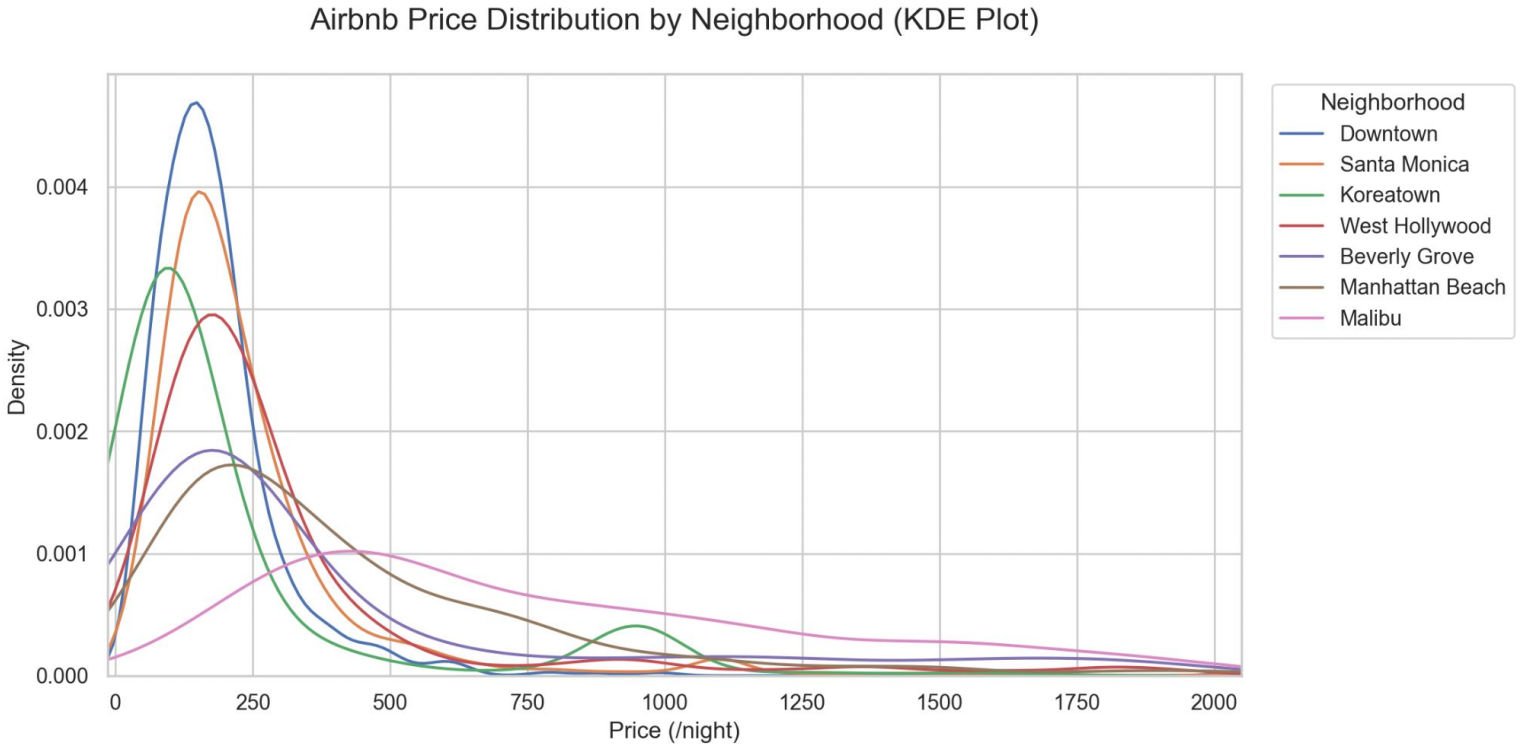
Average Airbnb Prices by Neighborhood - Geographic Mapping



Average Airbnb Prices by Neighborhood - Bar Chart



Neighborhood Price Distribution: Exploring the Variance



Exploratory Sentiment Analysis - Word Cloud

Superhost Prediction

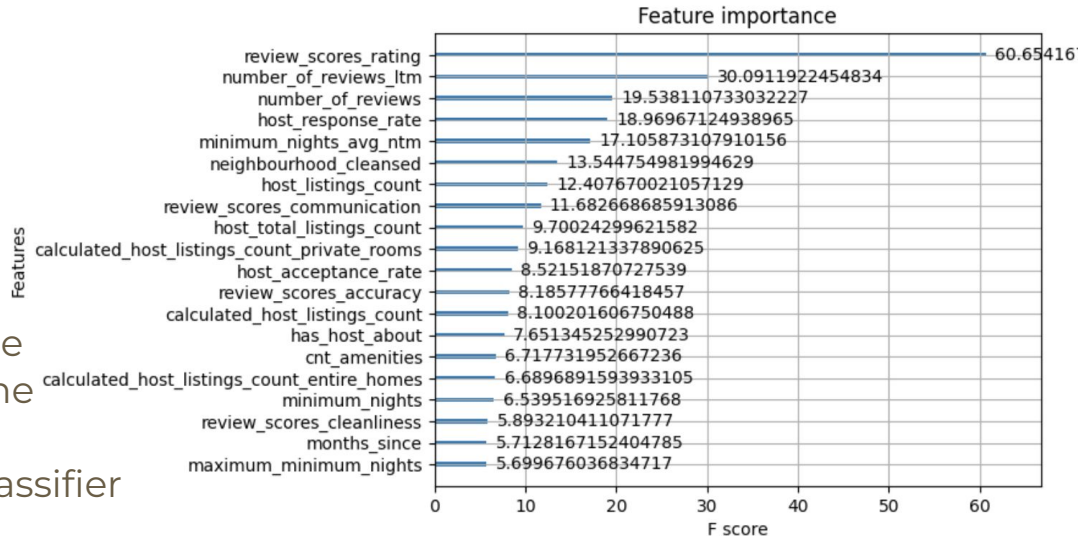
Machine learning

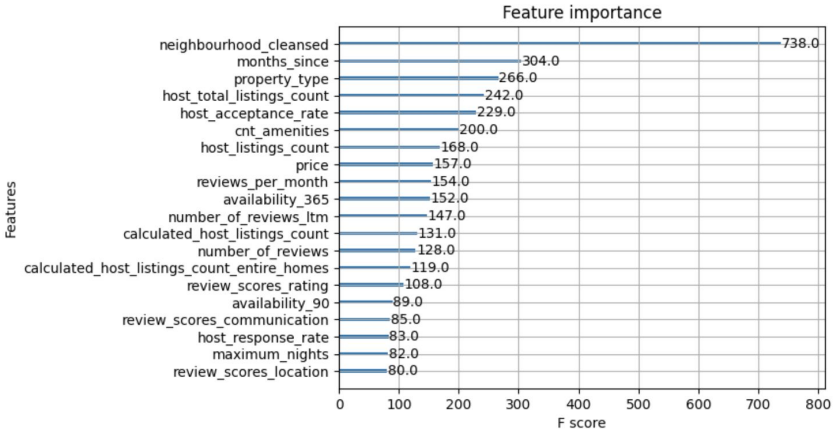
- Data Cleaning
  - Drop
    - Less-informative/duplicated features
    - Rows having null values in host_is_superhost and scores
    - Price = 999/9999
  - Reformat
    - Str to float: '$1,400' to 1400.00, '93%' to 0.9300
    - Date to date difference: 2023-06-18 to 4 (months)
  - Encode
    - Binary: t/f to 1/0
    - Category: dtype from object to category
  - Extract
    - Summarize the cnt of amenities/verifications

Training	20629
Testing	5158
Column	46
Superhost %	47.47%

Feature Selection

- XGBoost Classifier
  - Select Top20 features based on the average gain of splits which use the feature.
  - Use the 20 features to train xgb classifier with 5-fold cross validation, mean accuracy = 0.8627, std = 0.0039
- Backward Elimination
  - Try to eliminate the num of features using backward elimination based on improved accuracy.
  - In XGBoost classifier, no feature is eliminated.
  - Plan to explore other models for further feature selection and interpretation.





CHALLENGE

- **Data Cleaning**
  - Wide range of pricing, unable to determine if the prices are real or not...
- **Feature Selection Process**
  - Not sure if there's intercorrelation between the features
  - SHAP library isn't compatible with categorical variables used in XGBoost

NEXT STEPS

- **Data Visualization**
  - Try to implement dashboard with visualizations via Tableau Server
- **Machine Learning for Superhost Prediction**
  - Explore different ML models' explanatory powers
    - logistic regression
    - decision trees

THANK YOU