Airbnb Price Prediction Strategy

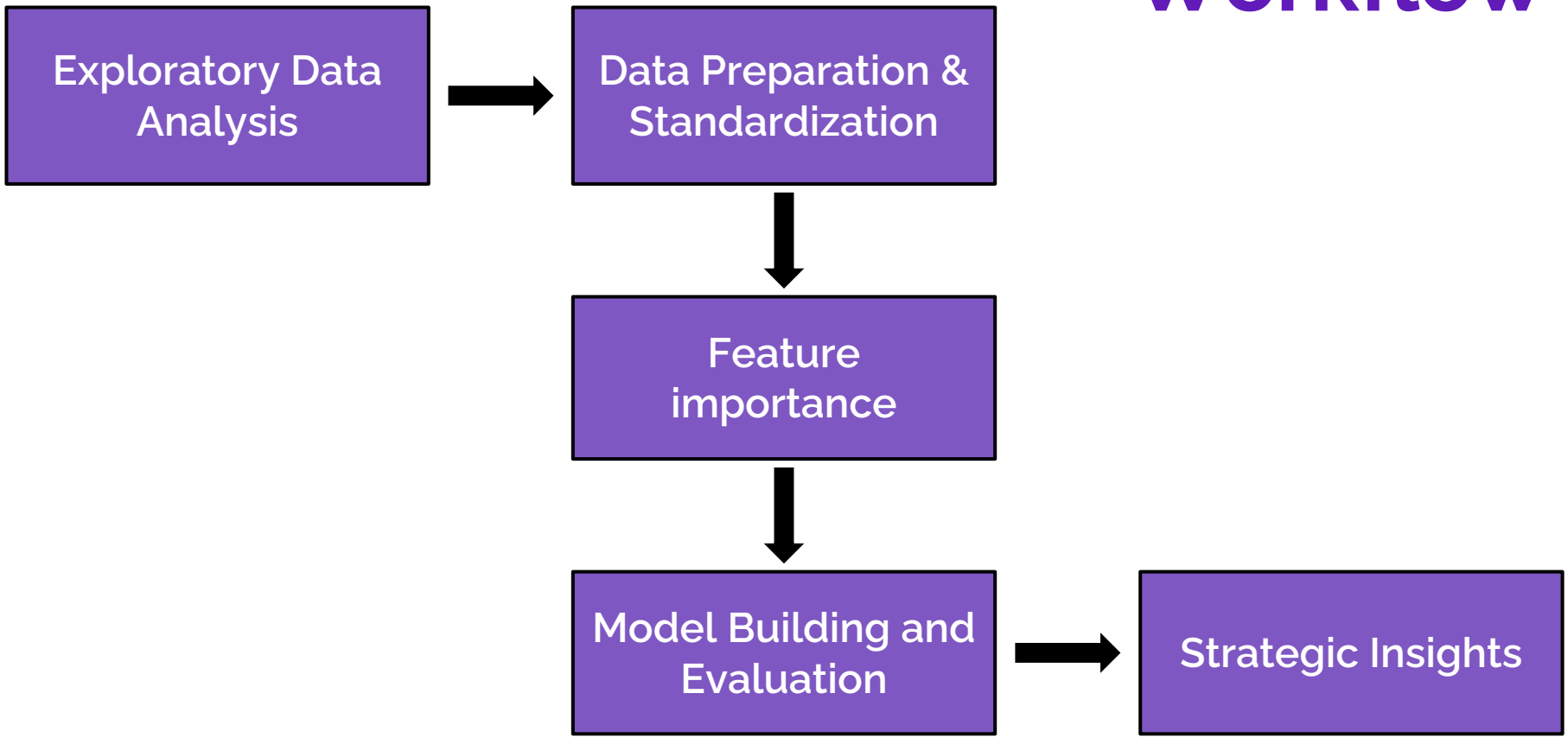
Objectives

- Explore the impact of Airbnb on global hospitality with over 6 million listings worldwide
- Discuss the role of data-driven pricing strategies for maximizing host earnings
- Develop a predictive model that optimizes Airbnb listing prices

Key Takeaways

- Best model: Gradient Boosting Regressor
- Feature Importance: What influence pricing strategies the most? Housing hosts can focus on improving them
- Suggestion: Dynamic pricing strategies, considering seasonality and booking frequencies to maximize revenue

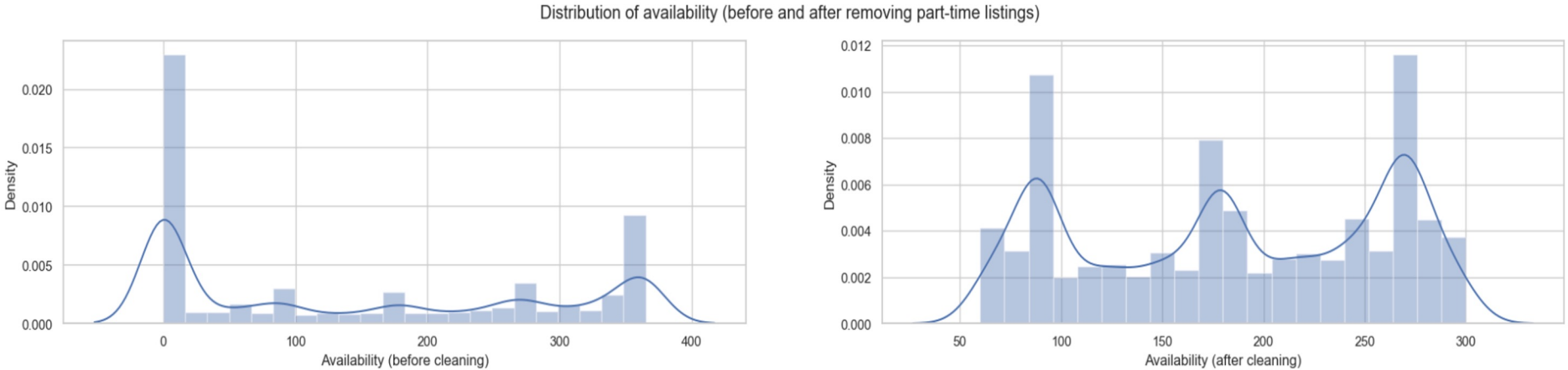
Workflow


Exploratory Data Analysis
Data Preparation & Standardization
Feature importance
Model Building and Evaluation
Strategic Insights

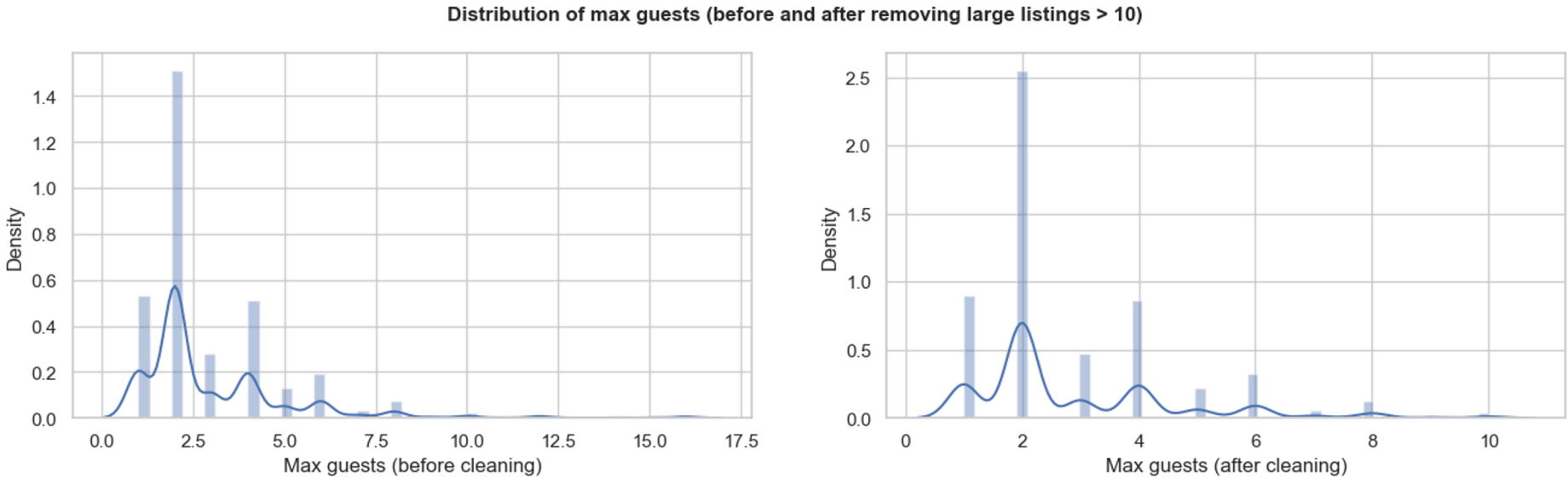
Step 1: Exploratory Data Analysis

- Availability Distribution: 75% of listings are available 150-200 days/year, highlighting seasonal booking trends.
- Guest Capacity: Most listings accommodate 1-4 guests, median capacity at 4, reflecting market demand for small group/family accommodations.
- Price Distribution: Prices predominantly range from $50 to $200 per night, with a median at $120, indicating a competitive market for affordable listings.

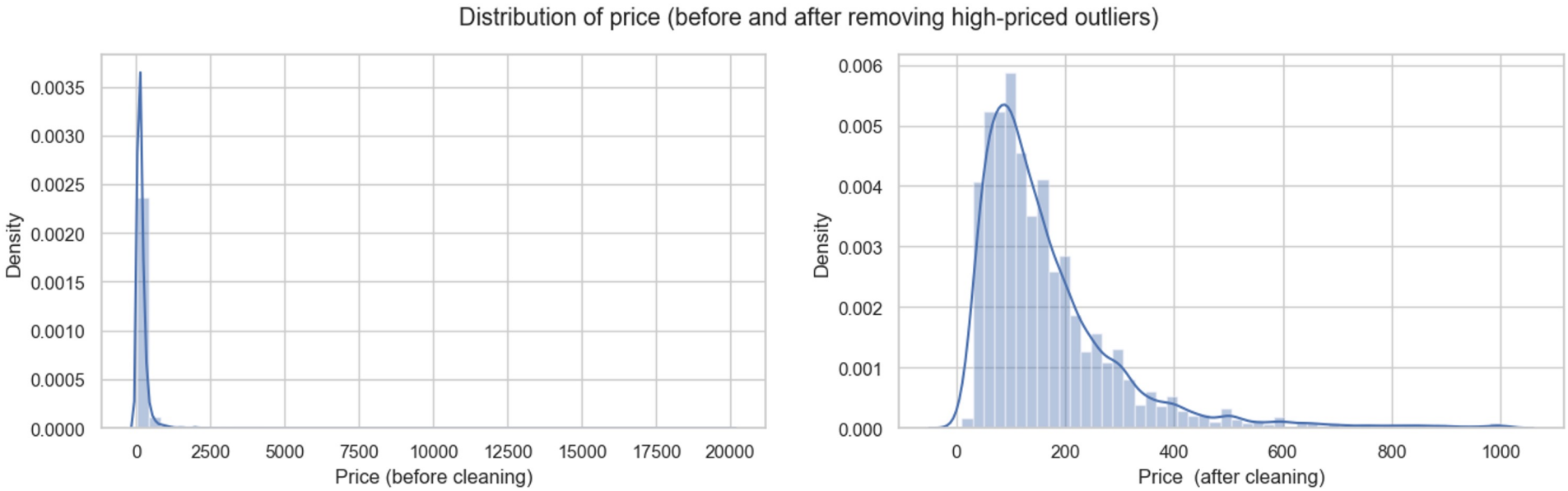
EDA: Availability



EDA: Max guests



EDA: Price

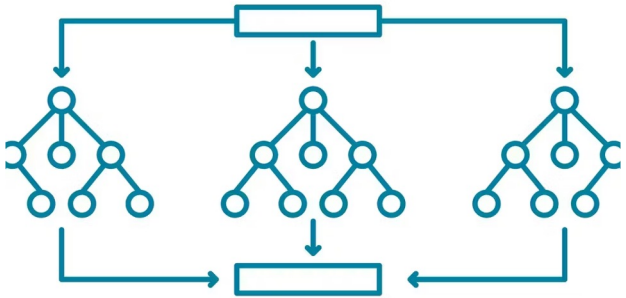

Distribution of price (before and after removing high-priced outliers)
Density
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
2500
5000
7500
10000
12500
15000
17500
20000
Price (before cleaning)
Density
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
200
400
600
800
1000
Price (after cleaning)

Step 2: Data Preparation & Standardization

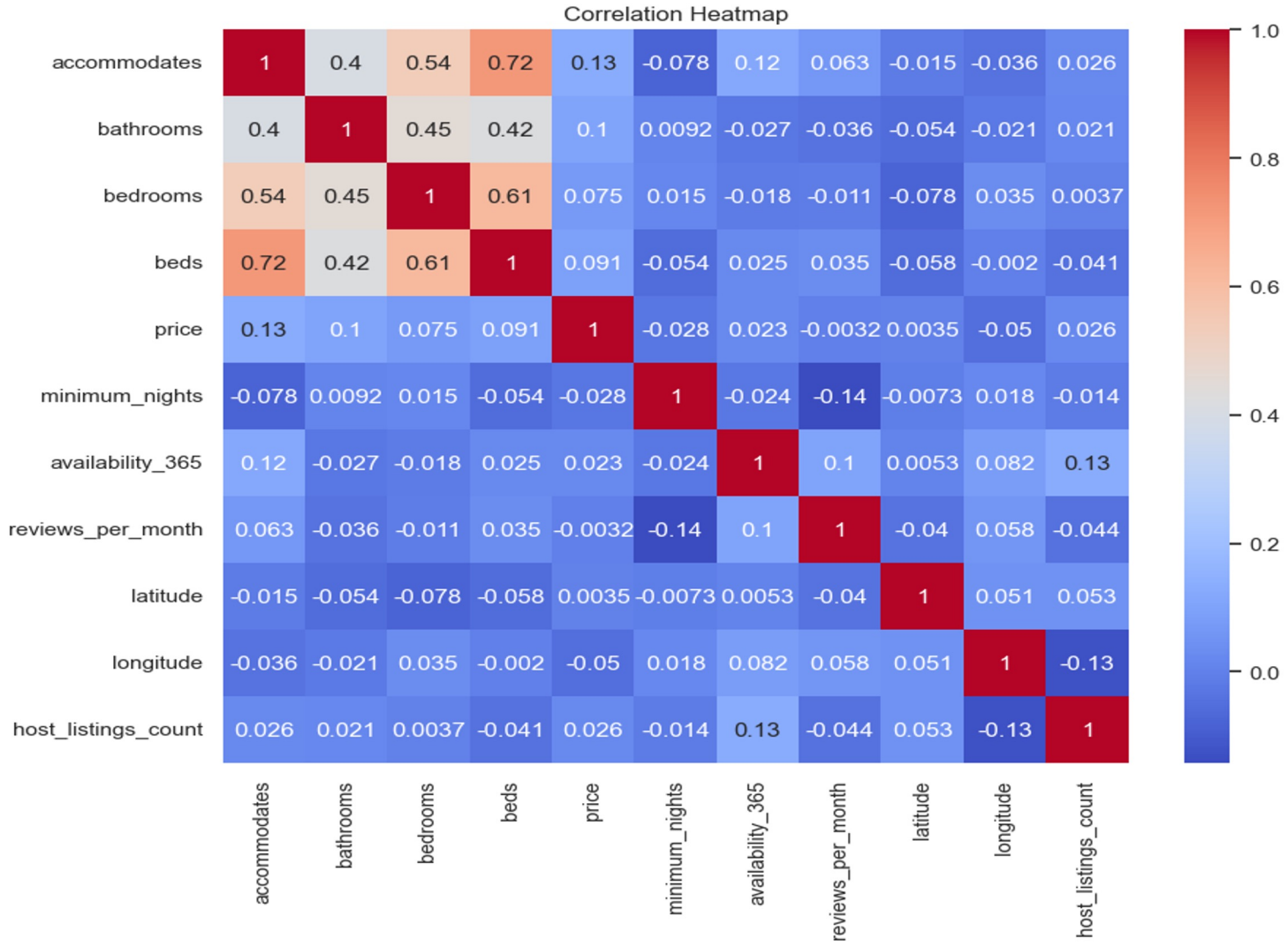
- Scaling Features: “MinMaxScaler” to ensure all features are on a 0-1 scale, enhancing model fairness by preventing scale dominance.
- Handling Missing Values: Missing values imputed based on median (continuous variables) and mode (categorical variables) to preserve data integrity.
- Feature Engineering: Created composite features such as 'yield' introduced as a composite metric to assess profitability more holistically than price alone

Step 3: Feature Importance

- Top Predictors: Number of bedrooms, instant bookability, and reviews per month significantly impact pricing, with bedrooms increasing price by approximately 15% per additional room.
- Feature Selection: Lasso Regression highlighted that removing non-impactful features (like certain amenities) simplifies the model without losing predictive power.

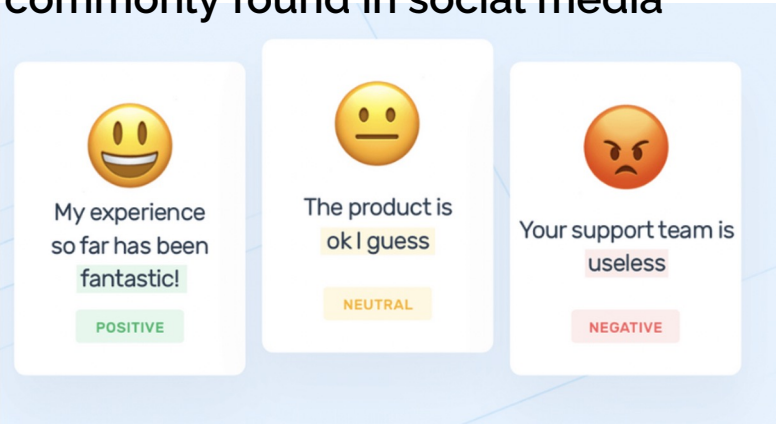


Feature Importances
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
reviews_per_month
price
instant_bookable
minimum_nights
latitude
room_type_Private room
accommodates
beds
host_listings_count
room_type_Entire home/apt
property_type_Room in hotel
bathrooms
property_type_Entire rental unit
bedrooms
host_is_superhost
property_type_Entire loft
longitude
property_type_Entire condo
property_type_Entire home
property_type_Private room in home
property_type_Private room in rental unit
positive_sentiment
property_type_Entire townhouse
property_type_Private room in condo
property_type_Entire guesthouse
property_type_Room in boutique hotel
neutral_sentiment
property_type_Private room in townhouse
property_type_Entire serviced apartment
property_type_Entire place
negative_sentiment
property_type_Entire guest suite
property_type_Private room in loft
property_type_Shared room in home
room_type_Shared room
property_type_Unique
property_type_Room in aparthotel
room_type_Hotel room
property_type_Private room
property_type_Shared room in rental unit
property_type_Private room in guesthouse
property_type_Private room in casa particular
property_type_Private room in bed and breakfast
property_type_Private room in serviced apartment
property_type_Entire vacation home


Correlation Heatmap
accommodates 1 0.4 0.54 0.72 0.13 -0.078 0.12 0.063 -0.015 -0.036 0.026
bathrooms 0.4 1 0.45 0.42 0.1 0.0092 -0.027 -0.036 -0.054 -0.021 0.021
bedrooms 0.54 0.45 1 0.61 0.075 0.015 -0.018 -0.011 -0.078 0.035 0.0037
beds 0.72 0.42 0.61 1 0.091 -0.054 0.025 0.035 -0.058 -0.002 -0.041
price 0.13 0.1 0.075 0.091 1 -0.028 0.023 -0.0032 0.0035 -0.05 0.026
minimum_nights -0.078 0.0092 0.015 -0.054 -0.028 1 -0.024 -0.14 -0.0073 0.018 -0.014
availability_365 0.12 -0.027 -0.018 0.025 0.023 -0.024 1 0.1 0.0053 0.082 0.13
reviews_per_month 0.063 -0.036 -0.011 0.035 -0.0032 -0.14 0.1 1 -0.04 0.058 -0.044
latitude -0.015 -0.054 -0.078 -0.058 0.0035 -0.0073 0.0053 -0.04 1 0.051 0.053
longitude -0.036 -0.021 0.035 -0.002 -0.05 0.018 0.082 0.058 0.051 1 -0.13
host_listings_count 0.026 0.021 0.0037 -0.041 0.026 -0.014 0.13 -0.044 0.053 -0.13 1
accommodates
bathrooms
bedrooms
beds
price
minimum_nights
availability_365
reviews_per_month
latitude
longitude
host_listings_count
1.0
0.8
0.6
0.4
0.2
0.0

Sentiment Intensity Analyzer

- Original score ranging from -1 (negative) to 1 (positive) → 3 column ranging 0 to 1
- It's especially good at handling texts with emoticons, slang, and abbreviations commonly found in social media



🥰 loved staying geoff place absolutely brillia...

yield	negative_sentiment	neutral_sentiment	positive_sentiment
24174.00	0.000	0.778	0.222
23924.16	0.000	0.819	0.181
35305.92	0.000	0.608	0.392
11016.00	0.000	0.622	0.378

Step 4: Model Building and Evaluation

- **Regression models** are a staple in prediction, can quantify relationships between multiple variables
- Evaluation: We used metrics like R^2 and MSE
  - R^2 can tell the proportion of variance in data that is explained by this model
  - MSE can tell average squared difference between the actual and predicted values.

	Model	R^2 Score	MSE
0	Linear Regression	0.915270	2.233330e+08
1	Lasso Regression	0.915288	2.232860e+08
2	Ridge Regression	0.915285	2.232922e+08
3	SVR	0.840608	4.201286e+08
4	MLP Regressor	0.917127	2.184394e+08
5	Gradient Boosting Regressor	0.997167	7.466529e+06

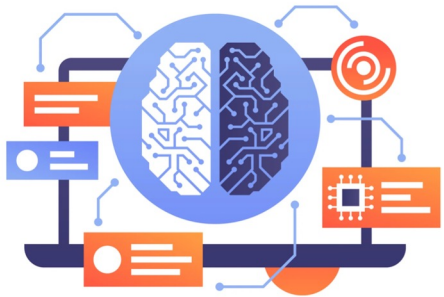
Step 4: Model Building and Evaluation

- Model Performance: **Gradient Boosting Regressor** performed best with an R^2 of 0.9971, and MSE of 7.46 million
- Ridge and Lasso Regression models also performed well with an R^2 of 0.9153, significantly reducing MSE to 223 million from a baseline of 232 million in Linear Regression
- Validation: We employed 5-fold cross-validation in model tuning
    - ensure robust against overfitting, more generalization power

	Model	R^2 Score	MSE
0	Linear Regression	0.915270	2.233330e+08
1	Lasso Regression	0.915288	2.232860e+08
2	Ridge Regression	0.915285	2.232922e+08
3	SVR	0.840608	4.201286e+08
4	MLP Regressor	0.917127	2.184394e+08
5	Gradient Boosting Regressor	0.997167	7.466529e+06

Step 5: Strategic Insights

- Dynamic Pricing: Analysis suggests adjusting prices based on seasonal trends and booking frequencies can optimize revenue—e.g., increasing rates during peak tourist seasons.
- Marketing Focus: Enhancing features that significantly impact pricing, such as adding instant bookability, can increase listing attractiveness and competitive edge in the market.



Conclusions

- Employed **Gradient Boosting Regressor** performed best with an R^2 of 0.9971 - strong predictive accuracy for Airbnb pricing
- Identified critical pricing factors like bedrooms and bookability, influencing listing prices up to 15%
- We enable dynamic pricing strategies, optimizing revenue potential during peak seasons
- This application can give data-driven recommendations for hosts to enhance listings and increase competitiveness.

Future Work

- We have tried other tools to convert comments into sentiment score, such as TextBlob, but the performance was not as good as Sentiment Intensity Analyzer.
    - We would like to find out why this happens or how to tune this tool to better suit for our data
- We can other two models: 1) Elastic Net 2) K means cluster with Ridge Regression
- We can develop a user interface to enable them enter housing conditions and get the predicted ‘yield’ data