Airbnb Optimizer

Smart Pricing and Investment Advisory Platform
based on the Madrid Airbnb Dataset

Agenda

1. Introduction
2. Motivation
3. Dataset overview
4. Feature visualisation
5. Predictive models
6. Classification models

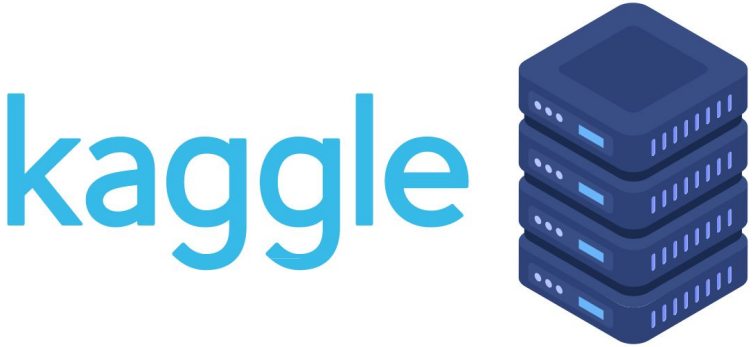
Introduction

Problem & Value Proposition:

- Development of an AI-powered platform that helps property owners and investors maximize their Airbnb rental potential in Madrid
- **Provided data-driven insights for:**
    1. Optimal pricing strategies
    2. Neighborhood performance analysis
    3. Potential return on investment
    4. Competitive positioning

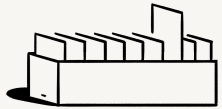
Dataset overview

- **Madrid Airbnb Data**
- Information about Airbnb listings in Madrid, Spain.
- Relatively up-to-date data: 11 September, 2024
- source: Kaggle platform
- It consists of 6 csv files that contain data about airbnb places in the city

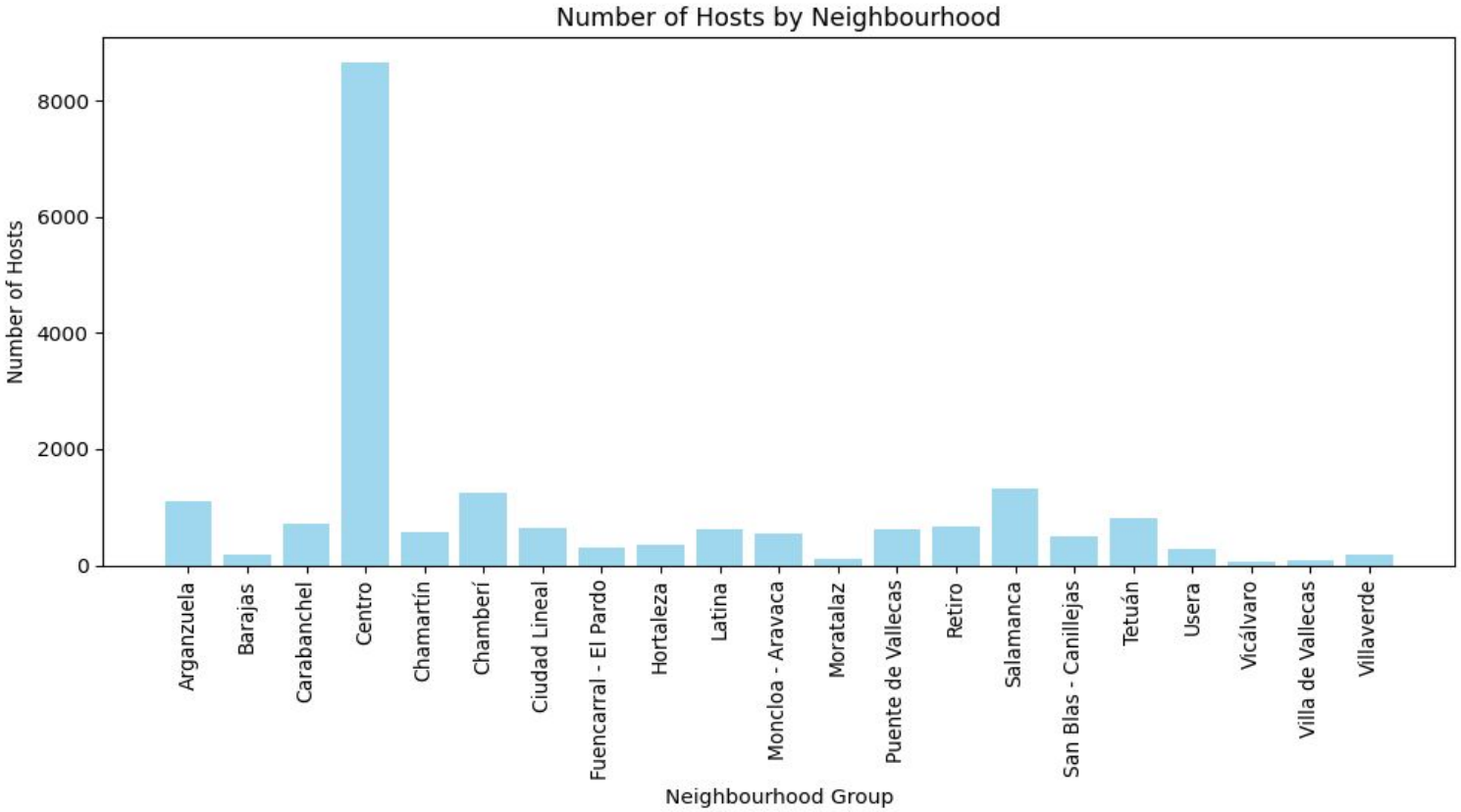


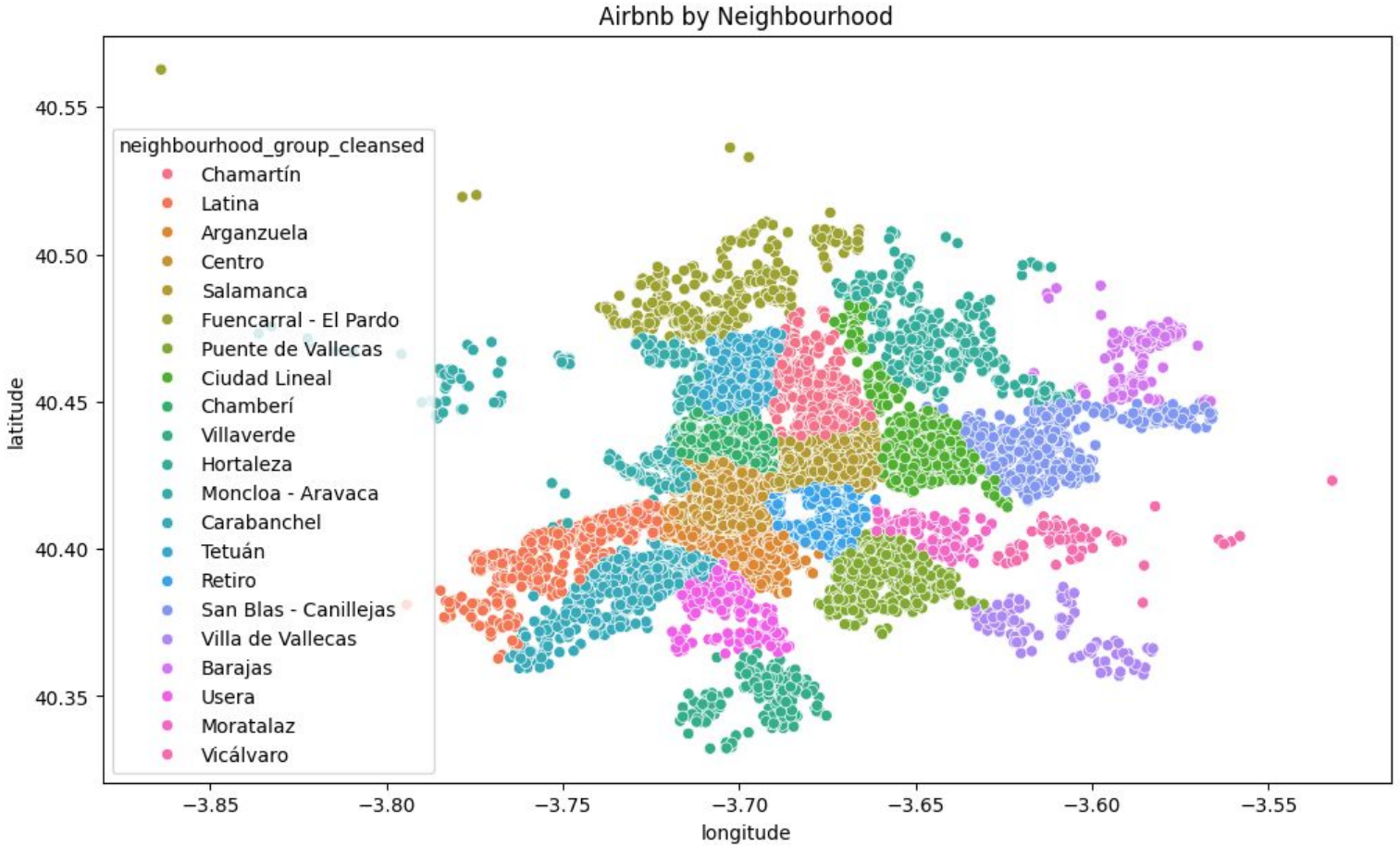
Contents – the same, probably not enough time

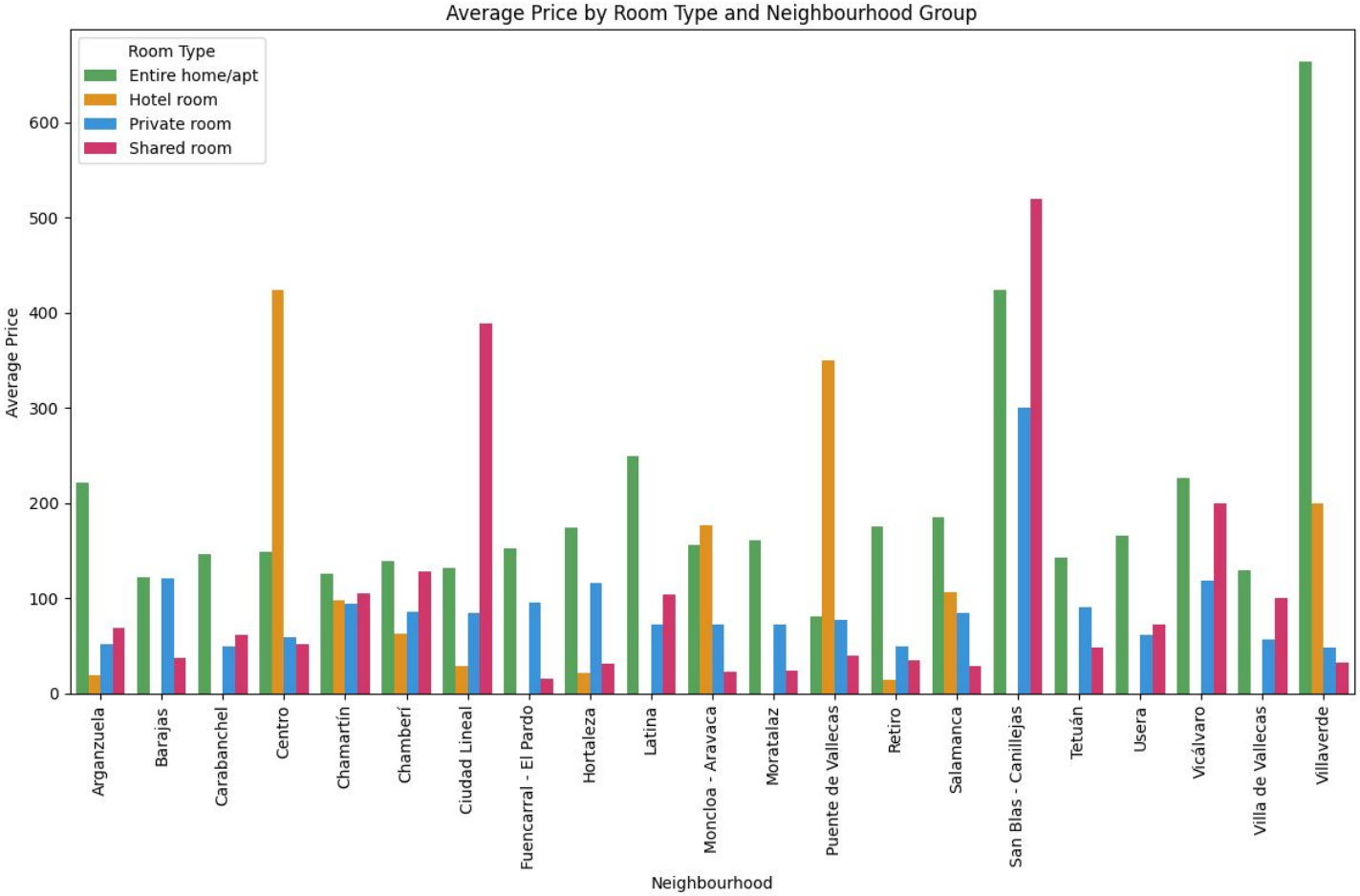
- **calendar.csv** - availability of places depending on the dates of arrival, as well as the prices and number of min and max nights that are possible to book.
- **reviews.csv** - dates of the collected reviews
- **reviews_detailed.csv** - extension of the previous file, with added information about the author of the review and the review content
- **neighbourhoods.csv** - assembly of the particular neighbourhoods
- **listings.csv** - information such as name, hostname, neighbourhood, location, type of room/flat, price, min nights, availability and reviews about a particular place
- **listings_detailed.csv** - extension of the previous file with added links to the websites of such place and its description. On top of that we can obtain some statistical data describing number of reviews or room scores.



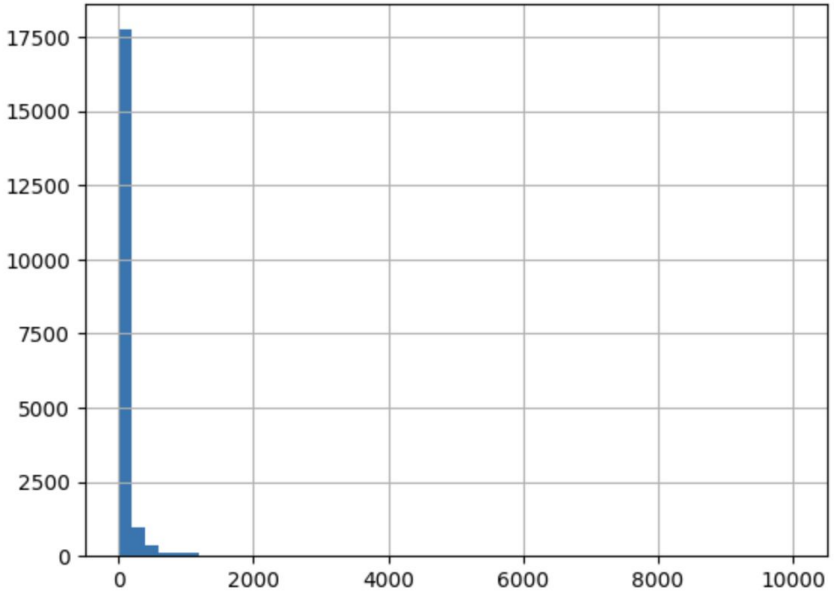
Data Visualization

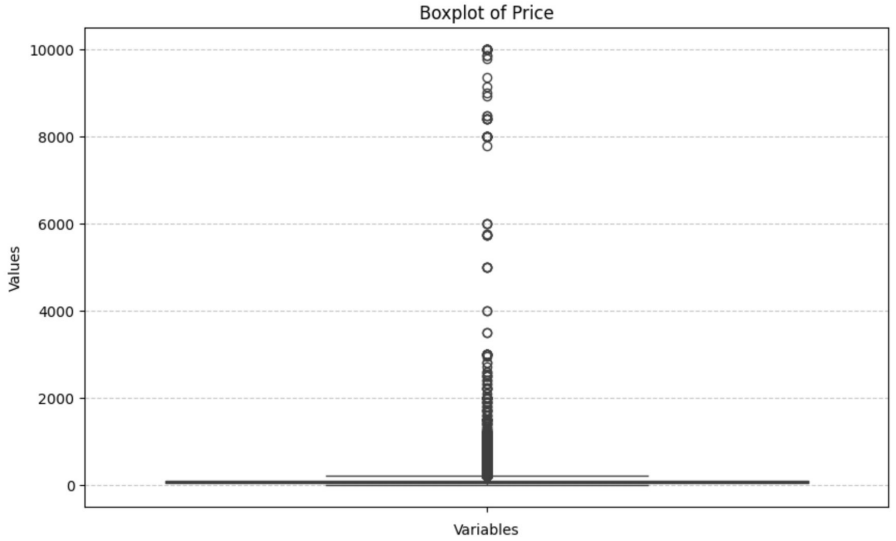

Number of Hosts by Neighbourhood
Number of Hosts
0
2000
4000
6000
8000
Arganzuela
Barajas
Carabanchel
Centro
Chamartín
Chamberí
Ciudad Lineal
Fuencarral - El Pardo
Hortaleza
Latina
Moncloa - Aravaca
Moratalaz
Puente de Vallecas
Retiro
Salamanca
San Blas - Canillejas
Tetuán
Usera
Vicálvaro
Villa de Vallecas
Villaverde
Neighbourhood Group


Airbnb by Neighbourhood
neighbourhood_group_cleansed
Chamartín
Latina
Arganzuela
Centro
Salamanca
Fuencarral - El Pardo
Puente de Vallecas
Ciudad Lineal
Chamberí
Villaverde
Hortaleza
Moncloa - Aravaca
Carabanchel
Tetuán
Retiro
San Blas - Canillejas
Villa de Vallecas
Barajas
Usera
Moratalaz
Vicálvaro
latitude
40.55
40.50
40.45
40.40
40.35
longitude
−3.85
−3.80
−3.75
−3.70
−3.65
−3.60
−3.55

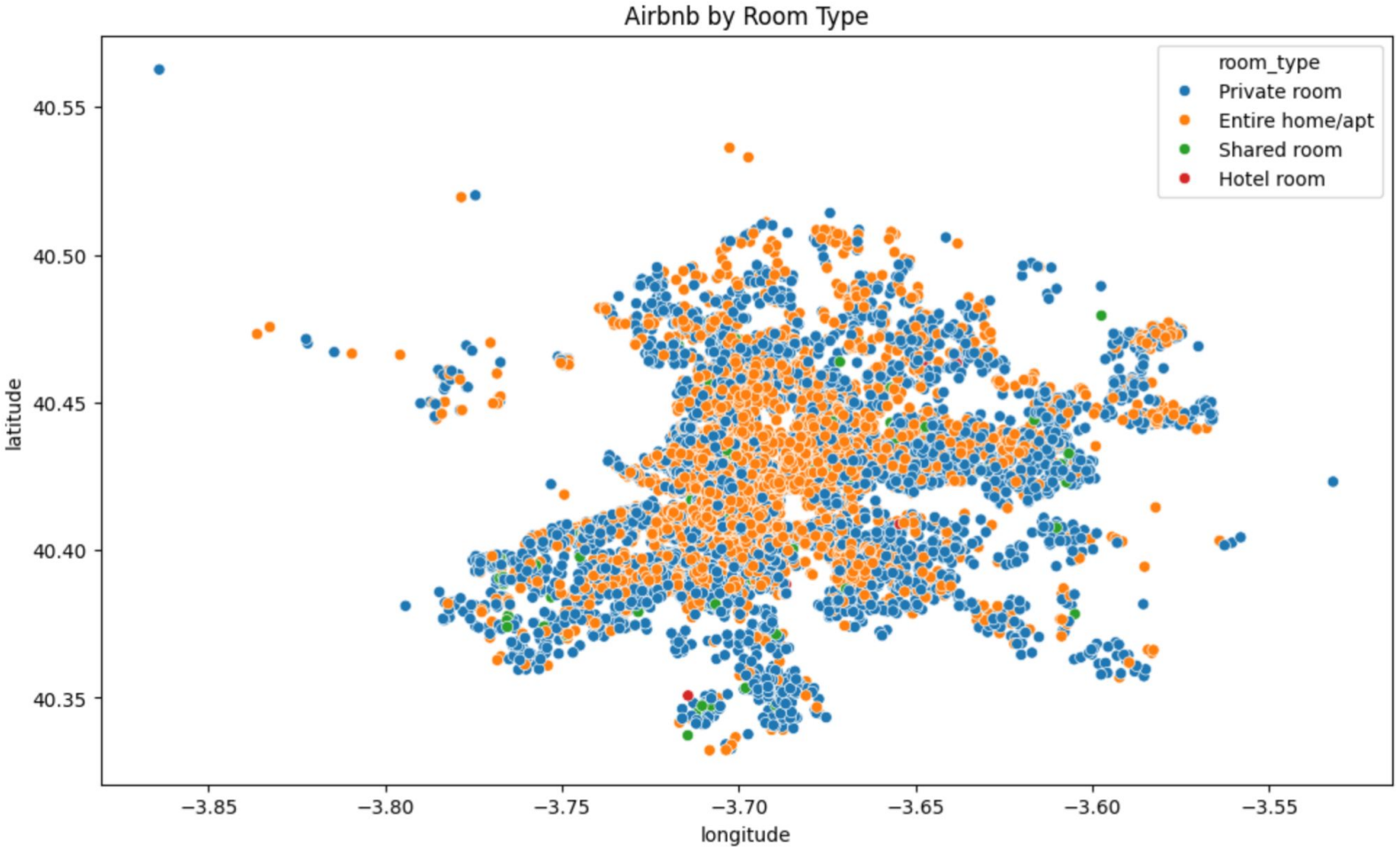

Average Price by Room Type and Neighbourhood Group
Room Type
Entire home/apt
Hotel room
Private room
Shared room
Average Price
0
100
200
300
400
500
600
Arganzuela
Barajas
Carabanchel
Centro
Chamartín
Chamberí
Ciudad Lineal
Fuencarral - El Pardo
Hortaleza
Latina
Moncloa - Aravaca
Moratalaz
Puente de Vallecas
Retiro
Salamanca
San Blas - Canillejas
Tetuán
Usera
Vicálvaro
Villa de Vallecas
Villaverde
Neighbourhood

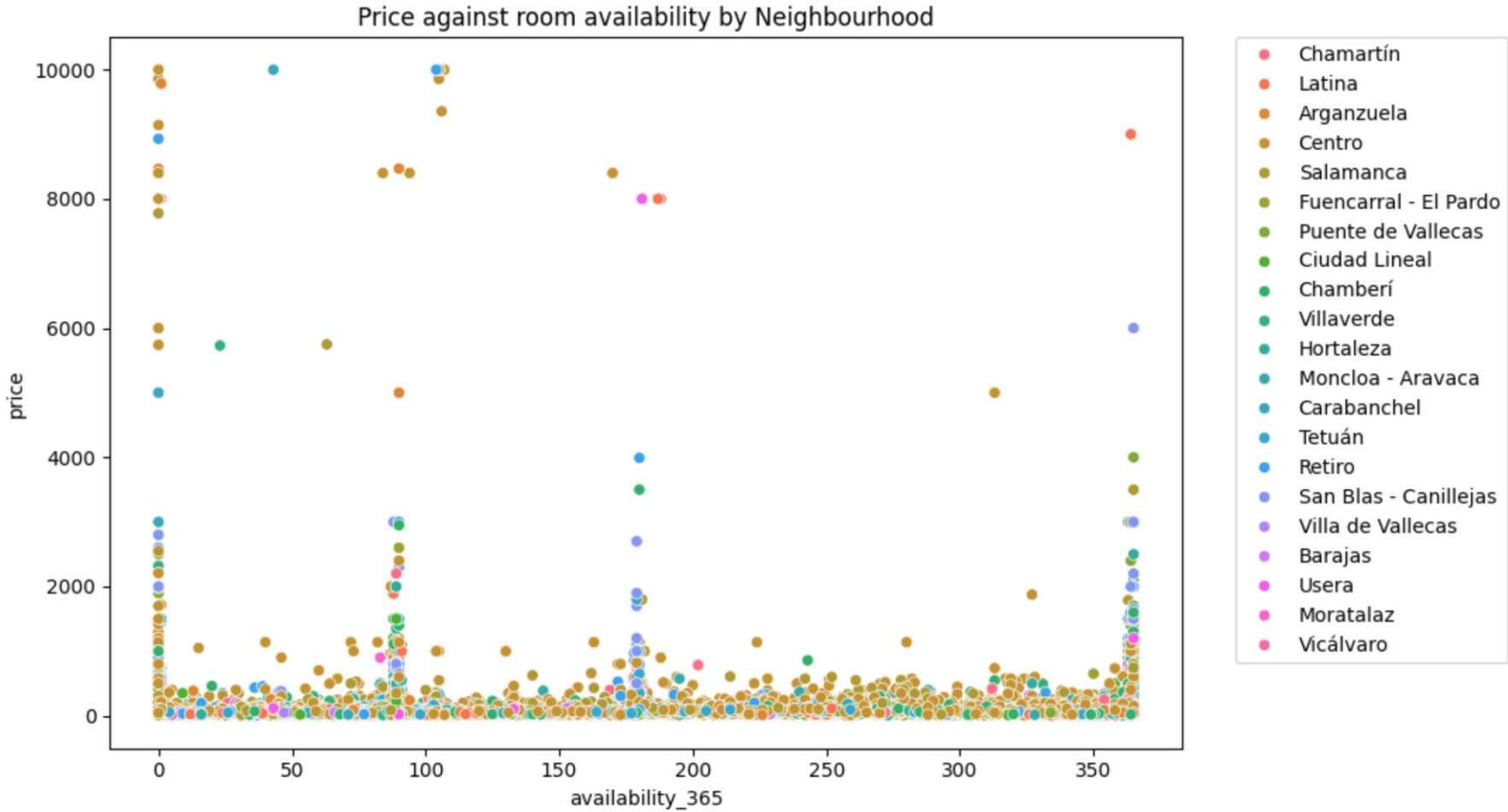
Distribution of the target feature – price (raw data)

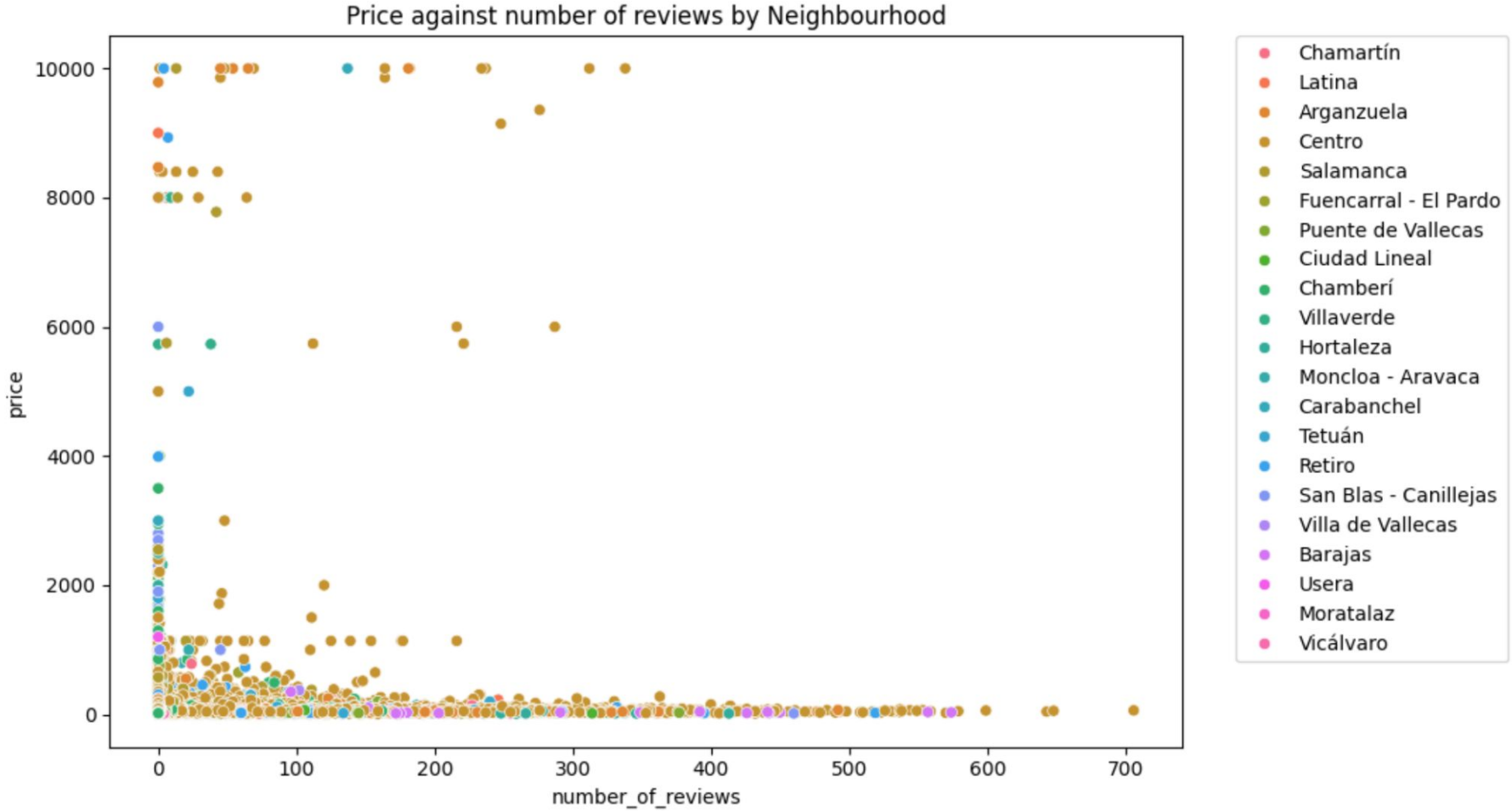


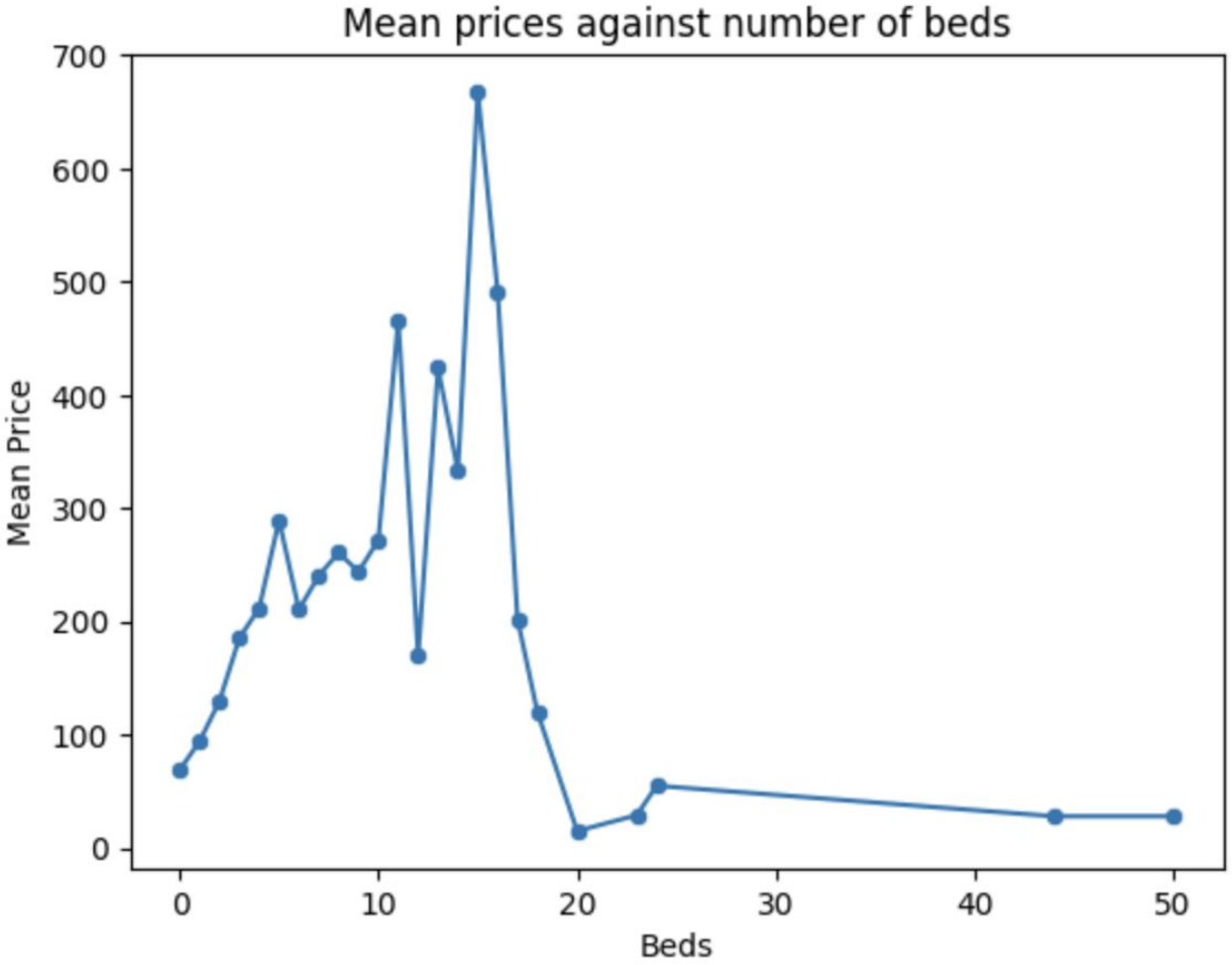


min	0.000000
25%	35.000000
50%	58.000000
75%	100.000000
max	9999.000000


Airbnb by Room Type
room_type
Private room
Entire home/apt
Shared room
Hotel room
40.55
40.50
40.45
40.40
40.35
latitude
−3.85
−3.80
−3.75
−3.70
−3.65
−3.60
−3.55
longitude


Price against room availability by Neighbourhood
price
availability_365
0
2000
4000
6000
8000
10000
0
50
100
150
200
250
300
350
Chamartín
Latina
Arganzuela
Centro
Salamanca
Fuencarral - El Pardo
Puente de Vallecas
Ciudad Lineal
Chamberí
Villaverde
Hortaleza
Moncloa - Aravaca
Carabanchel
Tetuán
Retiro
San Blas - Canillejas
Villa de Vallecas
Barajas
Usera
Moratalaz
Vicálvaro


Price against number of reviews by Neighbourhood
price
number_of_reviews
0
2000
4000
6000
8000
10000
0
100
200
300
400
500
600
700
Chamartín
Latina
Arganzuela
Centro
Salamanca
Fuencarral - El Pardo
Puente de Vallecas
Ciudad Lineal
Chamberí
Villaverde
Hortaleza
Moncloa - Aravaca
Carabanchel
Tetuán
Retiro
San Blas - Canillejas
Villa de Vallecas
Barajas
Usera
Moratalaz
Vicálvaro


Mean prices against number of beds
Mean Price
Beds
0
100
200
300
400
500
600
700
0
10
20
30
40
50

Preprocessing

Preprocessing steps

1- Feature Engineering: Remove redundant columns (id, host_id, neighbourhood_group and neighbourhood [as the location given as longitude and latitude]) & Add new helpful columns (host_ist_superhost, host_acceptance_rate merged from the other dataset on column “id”)

2- Handling missing values: fill with False & 0

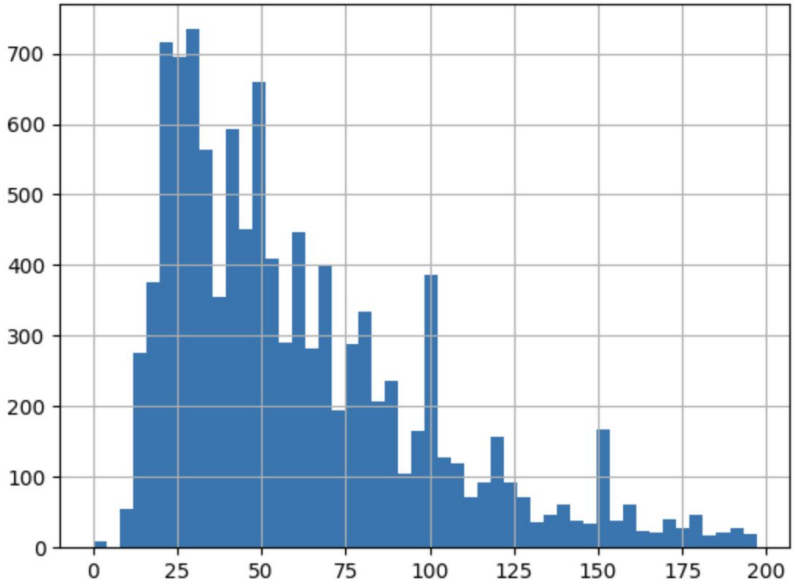
3- Handling outliers: Identify and remove outliers beyond the Interquartile Range threshold

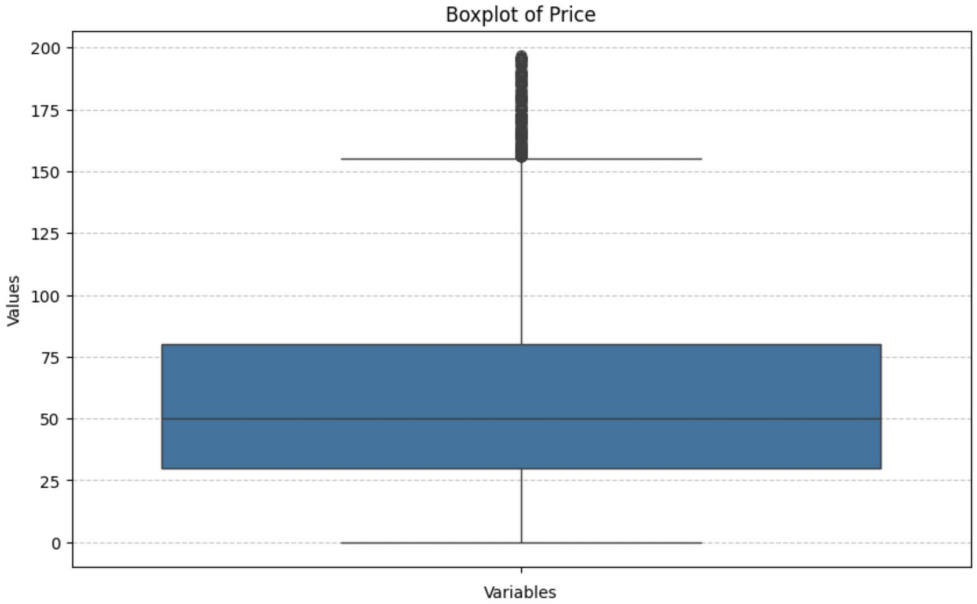
4- Dealing with categorical data: Apply one hot encoding

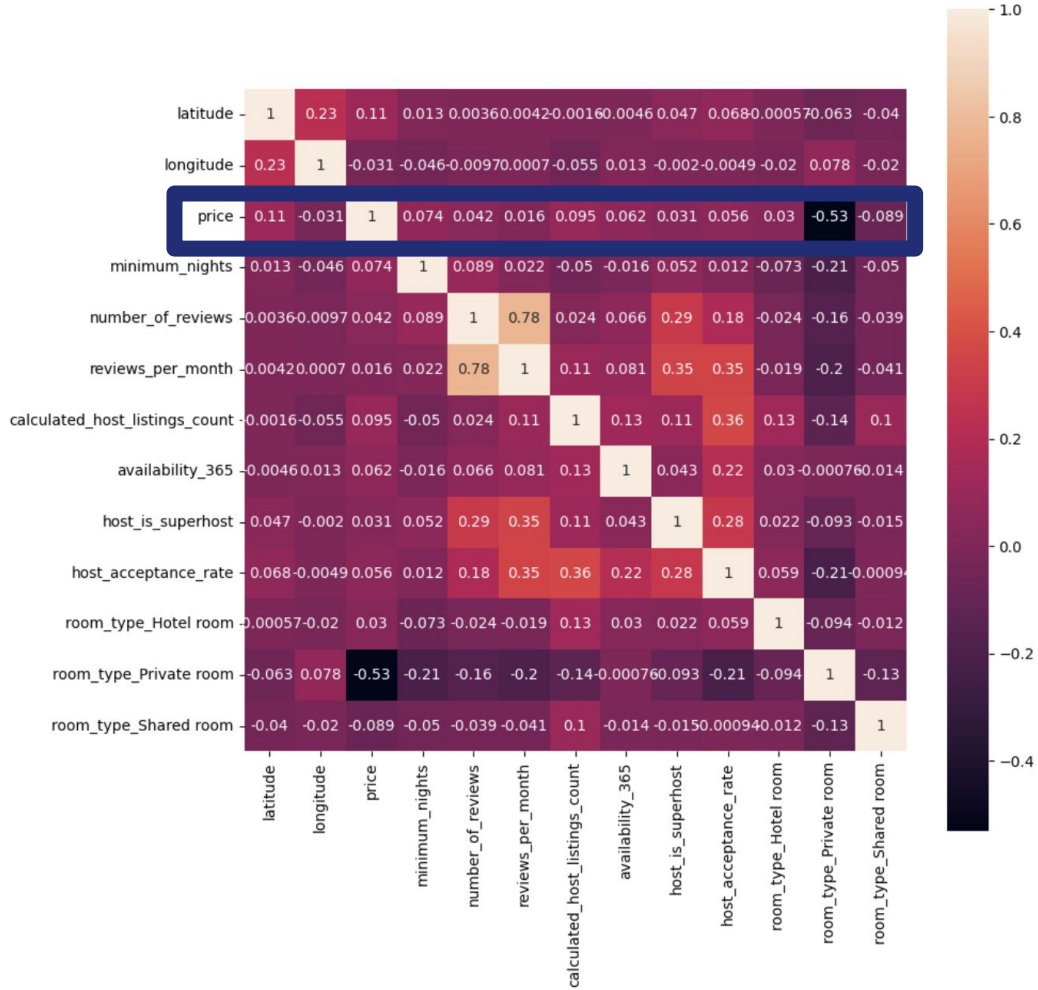
5- Preparing and splitting data: Organize data into features and targets and split the data into training and test sets

6- Normalization: Power Transformer to minimize the skewness of the data and improve the performance of the models

Distribution of the target feature – price (after preprocessing)







Correlations

Predictive Modeling – Regression

Price Prediction

Models:

1. Linear Regression
2. Decision Tree (CART)
3. Random Forest
4. Gradient Boosting
5. Multilayer Perceptron

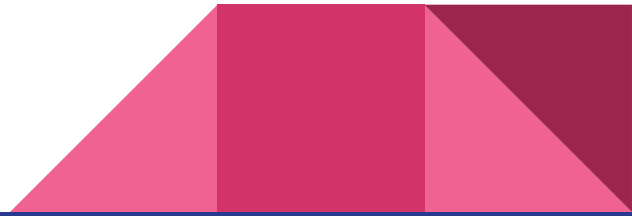
Price Prediction

Models:

1. Linear Regression
2. Decision Tree (CART)
3. Random Forest
4. Gradient Boosting
5. Multilayer Perceptron

Training and Evaluation (optimized with Gridsearch using 5-fold cross validation)

- performance optimization
- avoiding overfitting



Evaluation Metrics

Mean Absolute Error (MAE)

Root Mean Squared Error (RMSE)

R-Squared

Results

Linear Regression:
MAE: 23.04€
RMSE: 31.82€
R^2: 0.34

Decision Tree
MAE: 23.15€
RMSE: 31.97€
R^2: 0.34

Random Forest
MAE: 21.68€
RMSE: 30.31€
R^2: 0.40

Gradient Boosting
MAE: 21.71€
RMSE: 30.43€
R^2: 0.40

MLP
MAE: 22.52€
RMSE: 31.22€
R^2: 0.37

Results

Linear Regression:
MAE: 23.04€
RMSE: 31.82€
R^2: 0.34

Decision Tree
MAE: 23.15€
RMSE: 31.97€
R^2: 0.34

Random Forest
MAE: 21.68€
RMSE: 30.31€
R^2: 0.40

Gradient Boosting
MAE: 21.71€
RMSE: 30.43€
R^2: 0.40

MLP
MAE: 22.52€
RMSE: 31.22€
R^2: 0.37

Results

Linear Regression:
MAE: 23.04€
RMSE: 31.82€
R^2: 0.34

Decision Tree
MAE: 23.15€
RMSE: 31.97€
R^2: 0.34

Random Forest
MAE: 21.68€
RMSE: 30.31€
R^2: 0.40

Gradient Boosting
MAE: 21.71€
RMSE: 30.43€
R^2: 0.40

MLP
MAE: 22.52€
RMSE: 31.22€
R^2: 0.37

Results

Linear Regression:
MAE: 23.04€
RMSE: 31.82€
R^2: 0.34

Decision Tree
MAE: 23.15€
RMSE: 31.97€
R^2: 0.34

Random Forest
MAE: 21.68€
RMSE: 30.31€
R^2: 0.40

Gradient Boosting
MAE: 21.71€
RMSE: 30.43€
R^2: 0.40

MLP
MAE: 22.52€
RMSE: 31.22€
R^2: 0.37

Results

4

Linear Regression:
MAE: 23.04€
RMSE: 31.82€
R^2: 0.34

Decision Tree
MAE: 23.15€
RMSE: 31.97€
R^2: 0.34

1

Random Forest
MAE: 21.68€
RMSE: 30.31€
R^2: 0.40

2

Gradient Boosting
MAE: 21.71€
RMSE: 30.43€
R^2: 0.40

3

MLP
MAE: 22.52€
RMSE: 31.22€
R^2: 0.37

Results

4

Linear Regression:
MAE: 23.04€
RMSE: 31.82€
R^2: 0.34

5

Decision Tree
MAE: 23.15€
RMSE: 31.97€
R^2: 0.34

1

Random Forest
MAE: 21.68€
RMSE: 30.31€
R^2: 0.40

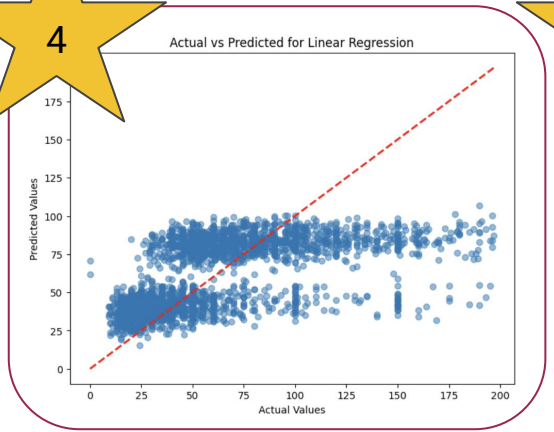
2

Gradient Boosting
MAE: 21.71€
RMSE: 30.43€
R^2: 0.40

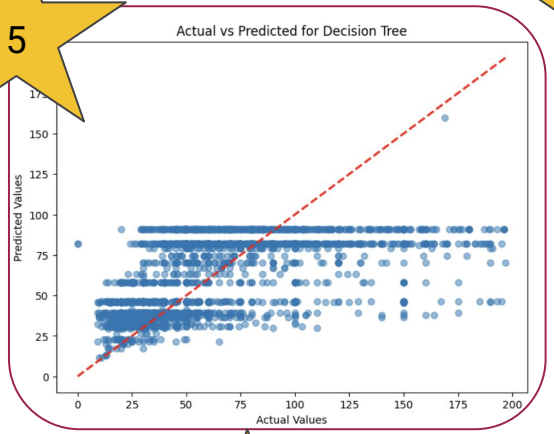
3

MLP
MAE: 22.52€
RMSE: 31.22€
R^2: 0.37

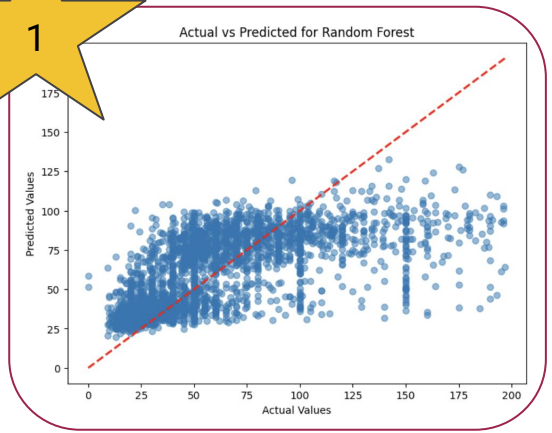
4


Actual vs Predicted for Linear Regression
Predicted Values
Actual Values

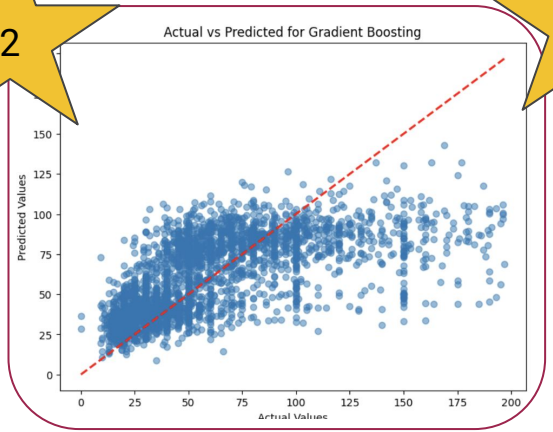
5


Actual vs Predicted for Decision Tree
Predicted Values
Actual Values

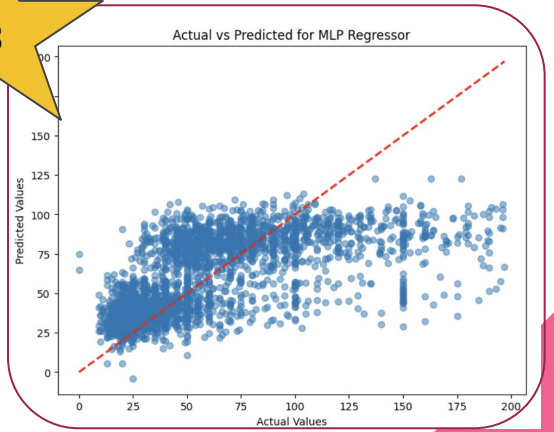
1


Actual vs Predicted for Random Forest
Predicted Values
Actual Values

2


Actual vs Predicted for Gradient Boosting
Predicted Values
Actual Values

3

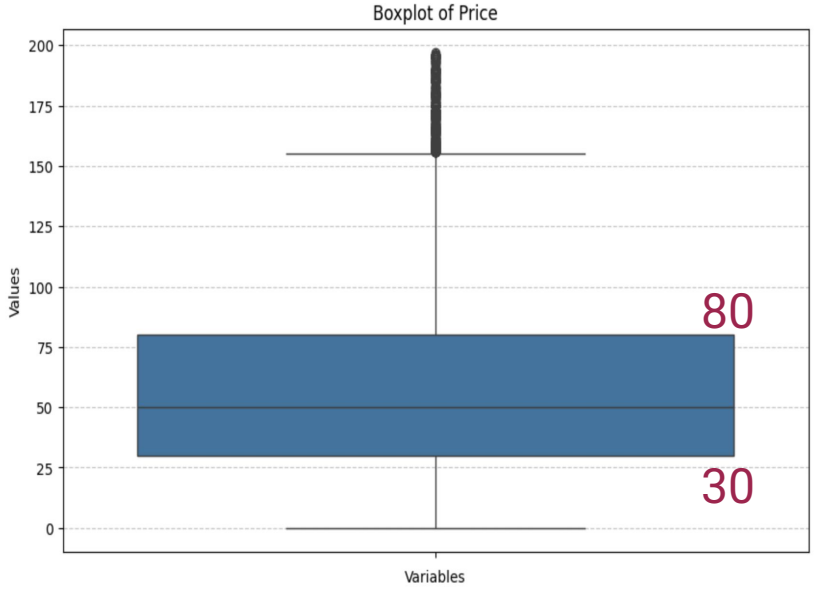

Actual vs Predicted for MLP Regressor
Predicted Values
Actual Values

Predictive Modeling – Classification

Price Tier Prediction

Defining Price tiers according to the distribution of the data

- Low (under 30€)
- Medium (between 30€ and 80€)
- High (above 80€)



Price Tier Prediction

Defining Price tiers according to the distribution of the data

- Low (under 30€)
- Medium (between 30€ and 80€)
- High (above 80€)

Models:

- Naive Bayes
- Decision Tree
- Neural Network (MLP classifier)
- Support Vector Machine (SVM)
- k-nearest neighbours (kNN)

Price Tier Prediction

Defining Price tiers according to the distribution of the data

- Low (under 30€)
- Medium (between 30€ and 80€)
- High (above 80€)

Models:

- Naive Bayes
- Decision Tree
- Neural Network (MLP classifier)
- Support Vector Machine (SVM)
- k-nearest neighbours (kNN)

Used grid search with cross validation for classification models too

Results – Evaluation Metric Accuracy

Naive Bayes
Accuracy: 0.48

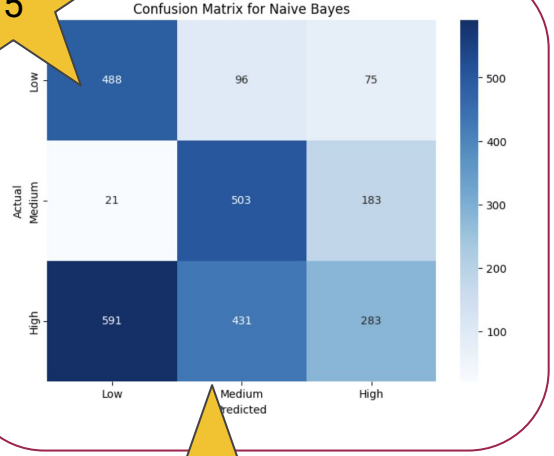
Decision Tree
Accuracy: 0.59

Neural Network
Accuracy: 0.61

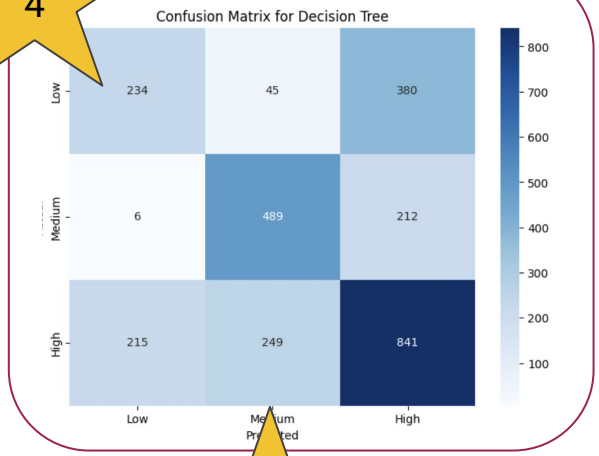
SVM
Accuracy: 0.60

kNN
Accuracy: 0.59

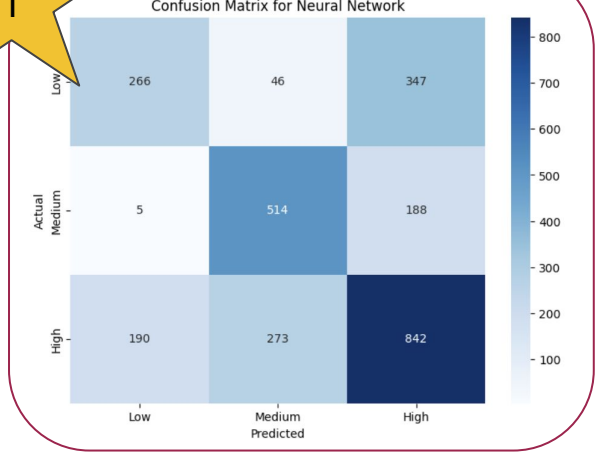
5


Confusion Matrix for Naive Bayes
Low
Medium
High
Actual
488
96
75
21
503
183
591
431
283
Low
Medium
High
Predicted
500
400
300
200
100

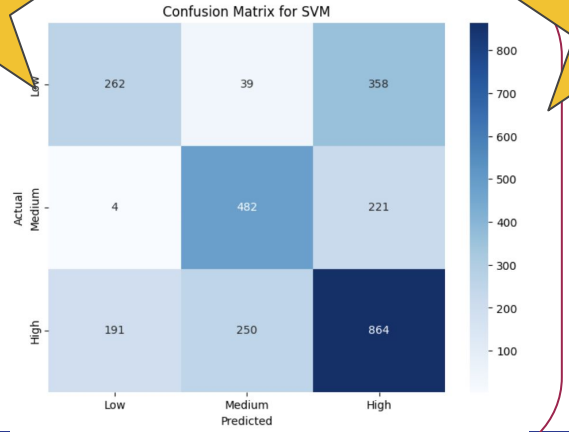
4


Confusion Matrix for Decision Tree
Low
Medium
High
234
45
380
6
489
212
215
249
841
Low
High
800
700
600
500
400
300
200
100

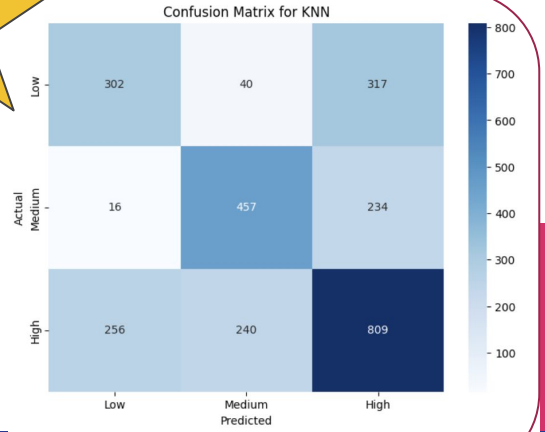
1

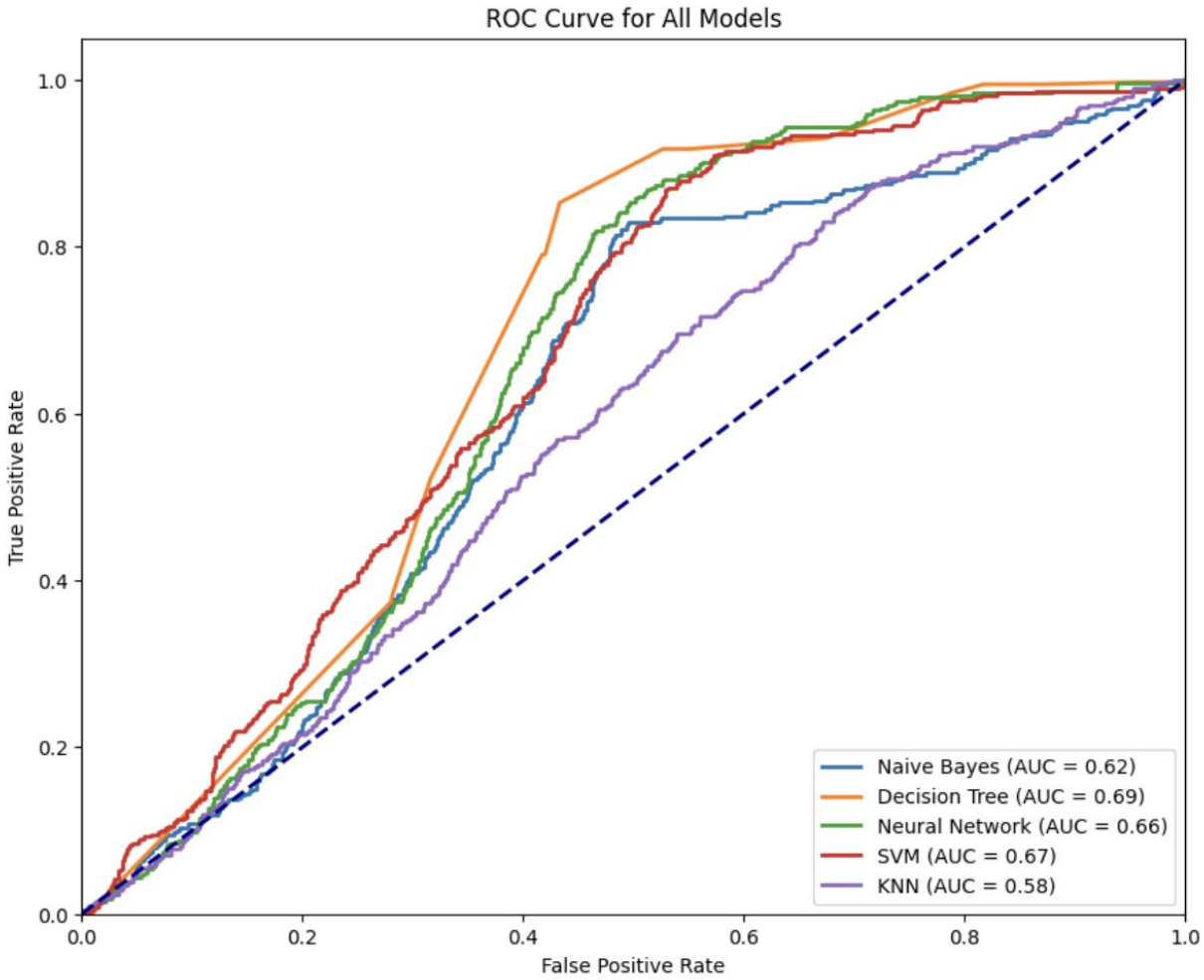

Confusion Matrix for Neural Network
Low
Medium
High
Actual
266
46
347
5
514
188
190
273
842
Low
Medium
High
Predicted
800
700
600
500
400
300
200
100

2


Confusion Matrix for SVM
Low
Medium
High
Actual
262
39
358
4
482
221
191
250
864
Low
Medium
High
Predicted
800
700
600
500
400
300
200
100

3


Confusion Matrix for KNN
Low
Medium
High
Actual
302
40
317
16
457
234
256
240
809
Low
Medium
High
Predicted
800
700
600
500
400
300
200
100


ROC Curve for All Models
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Naive Bayes (AUC = 0.62)
Decision Tree (AUC = 0.69)
Neural Network (AUC = 0.66)
SVM (AUC = 0.67)
KNN (AUC = 0.58)

Sources

- https://www.kaggle.com/datasets/rusiano/madrid-airbnb-data/data
- https://insideairbnb.com/get-the-data/
- https://insideairbnb.com/madrid/