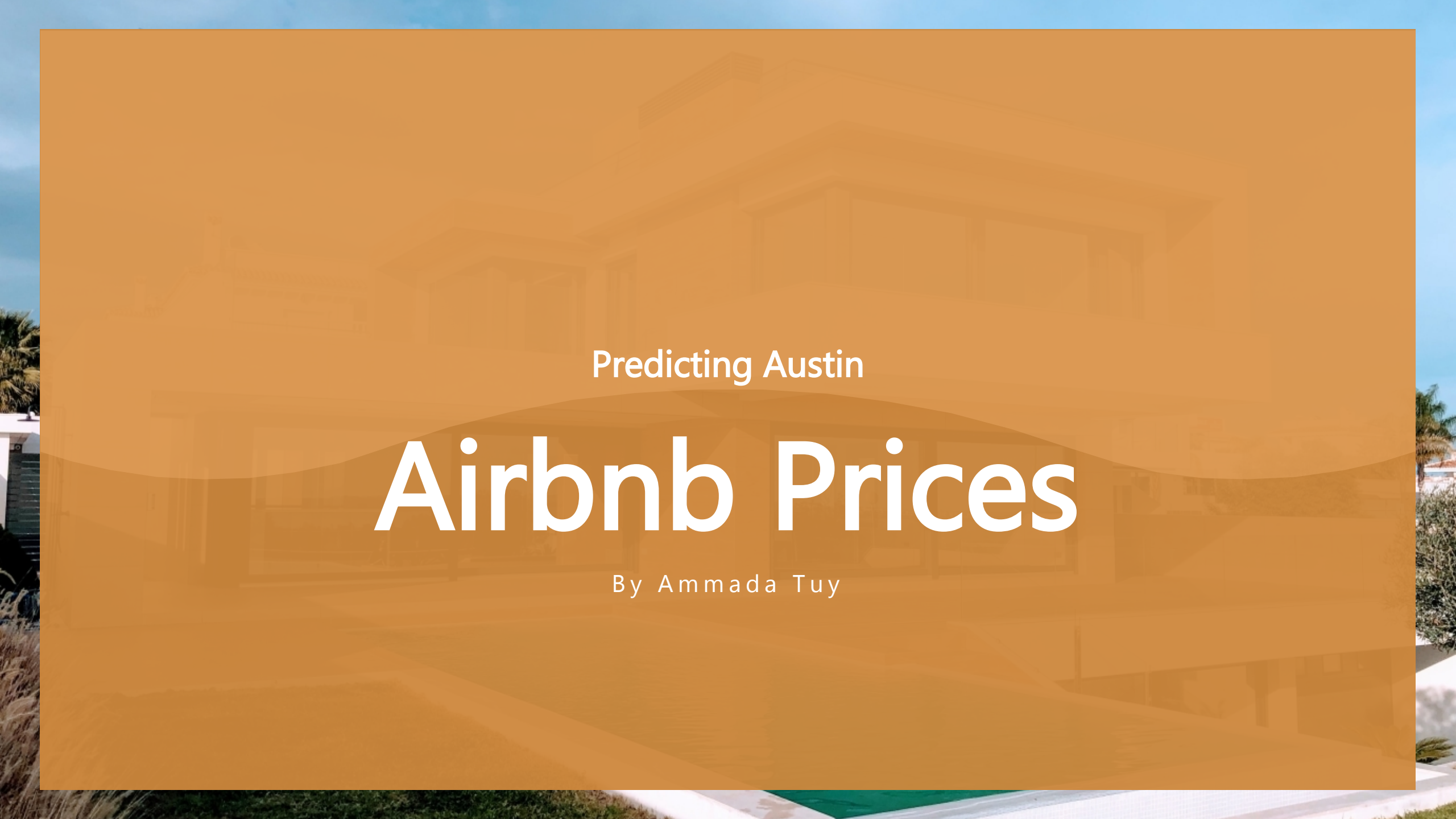


Predicting Austin

Airbnb Prices

By Ammada Tuy

THE PROBLEM

Hosts wants to know the best price to charge and attract customers, while customers want to find the best deals.

?

MY GOAL

Use data science to predict price for Airbnb listings based on various features.

DATASET USED

- Airbnb data scraped regularly from Inside Airbnb for most major cities
- Listing details which included number of bedrooms, number of bathrooms, average price, etc.
- Dataset had 11339 records and 106 features

DATA WRANGLING

FEATURE SELECTION

Dropped several features I didn't need such as host_name

Created new columns bedroom_bath_ratio and booking percentage

STRUCTURAL ISSUES

Converted price object type to float type

Converted room_type and neighbourhood object type to categorical type

MISSING VALUES

Imputed missing values of bathrooms, bedrooms and beds

Flagged missing review_scores_rating with -1 values

HANDLING OUTLIERS

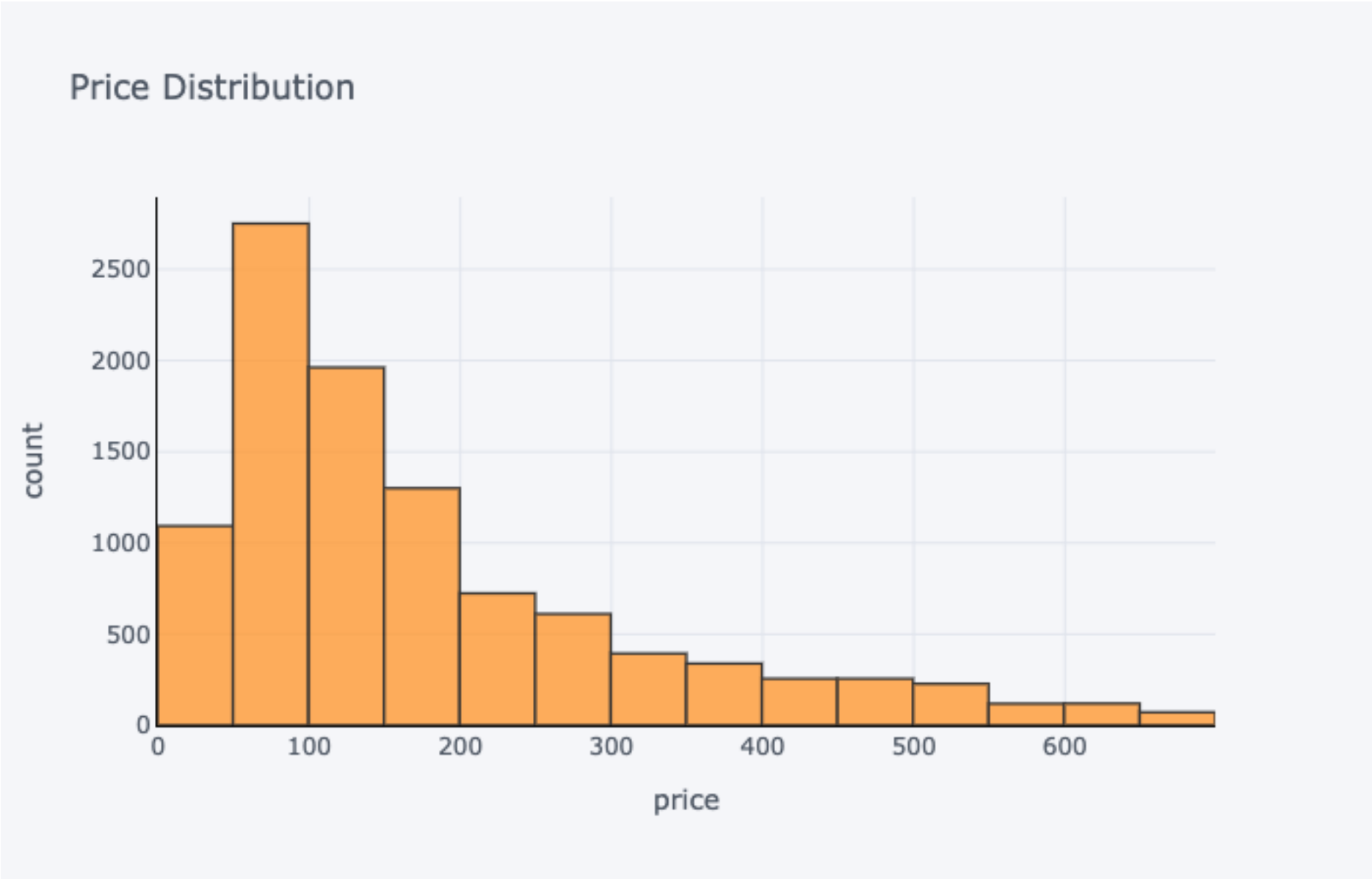
There were several outliers but I did not remove them

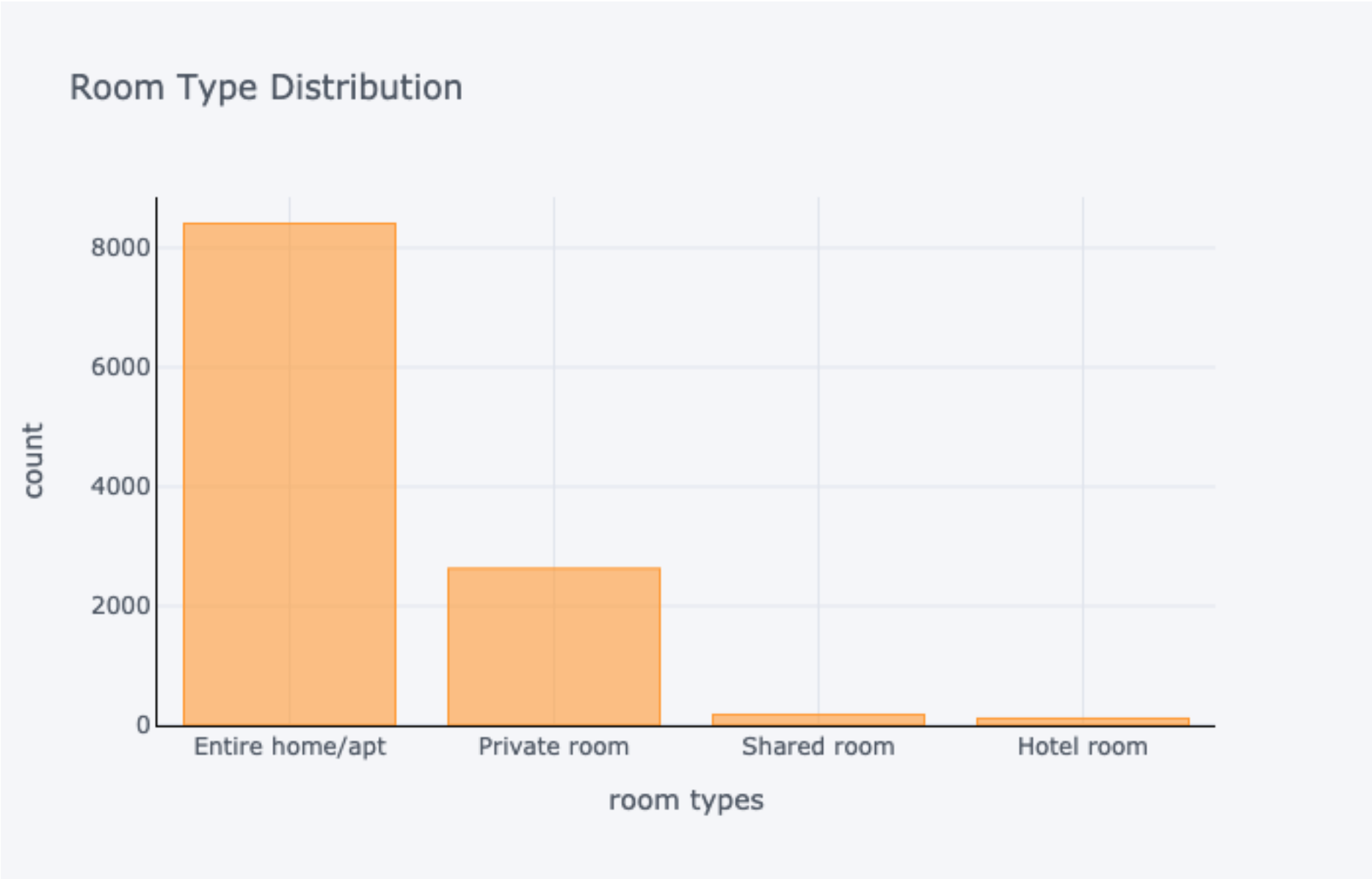
Removed a handful of records that were priced at zero

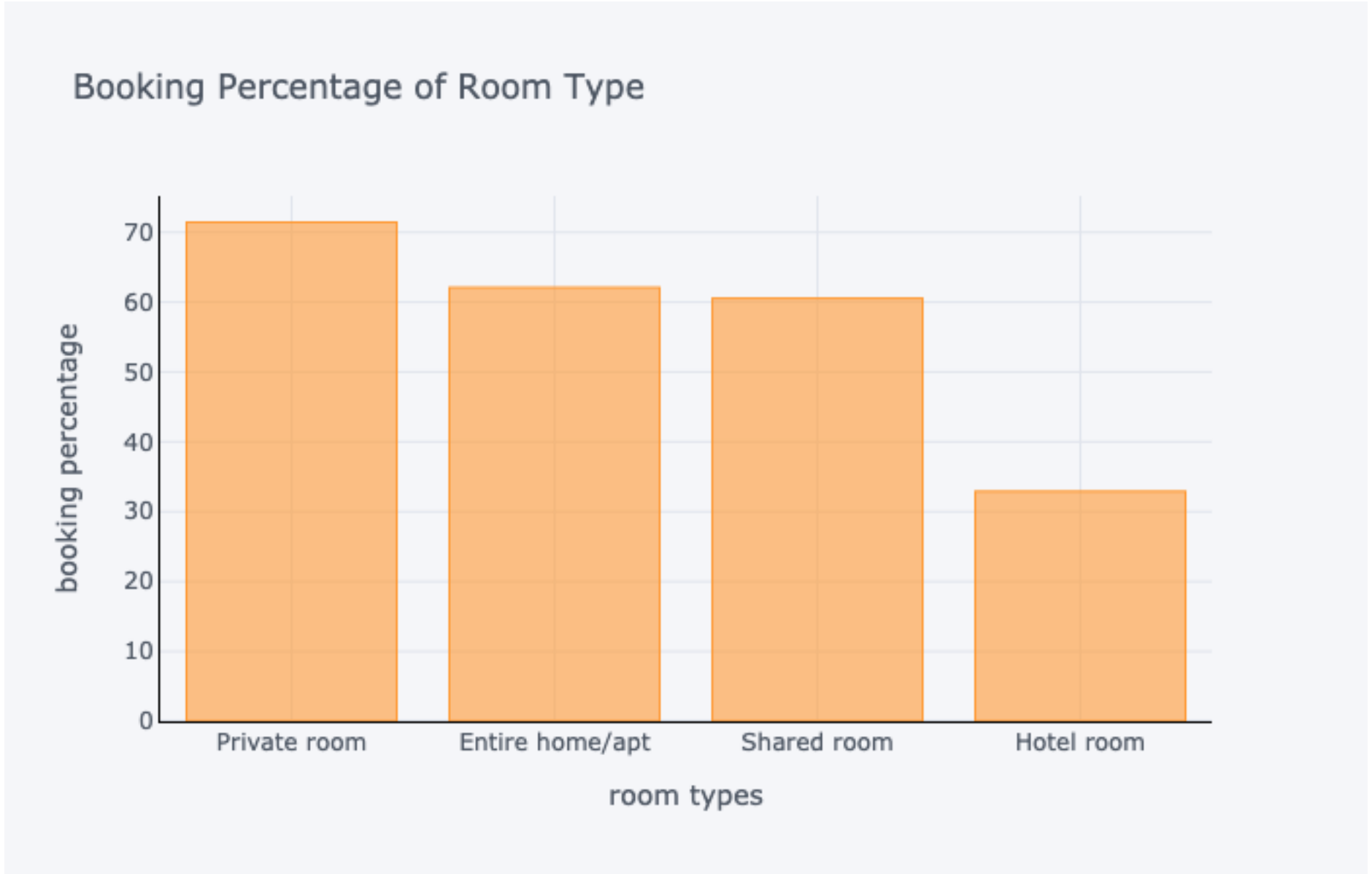
11,333 records and 16 features remained after cleanup

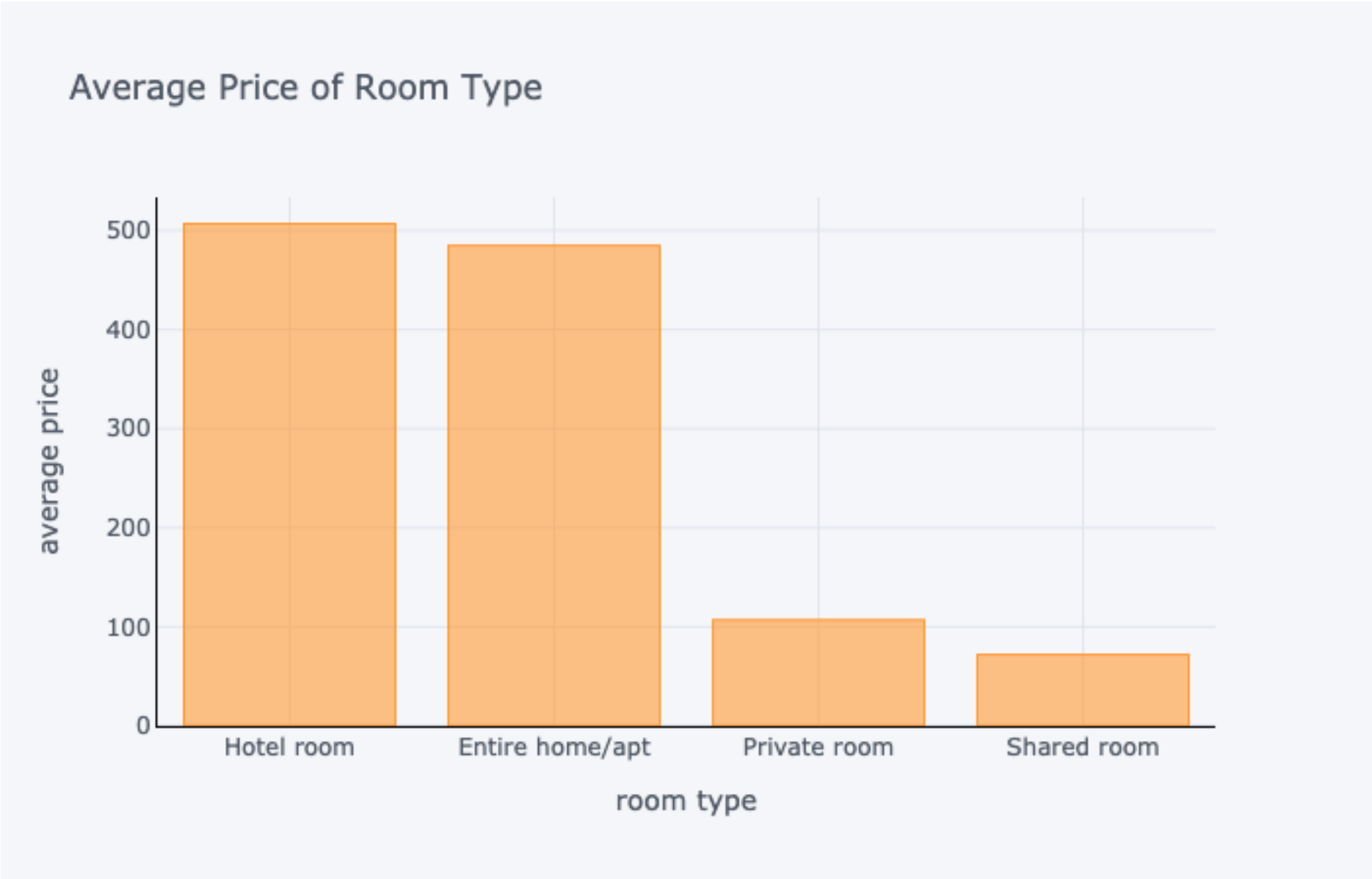
EXPLORATORY
DATA ANALYSIS

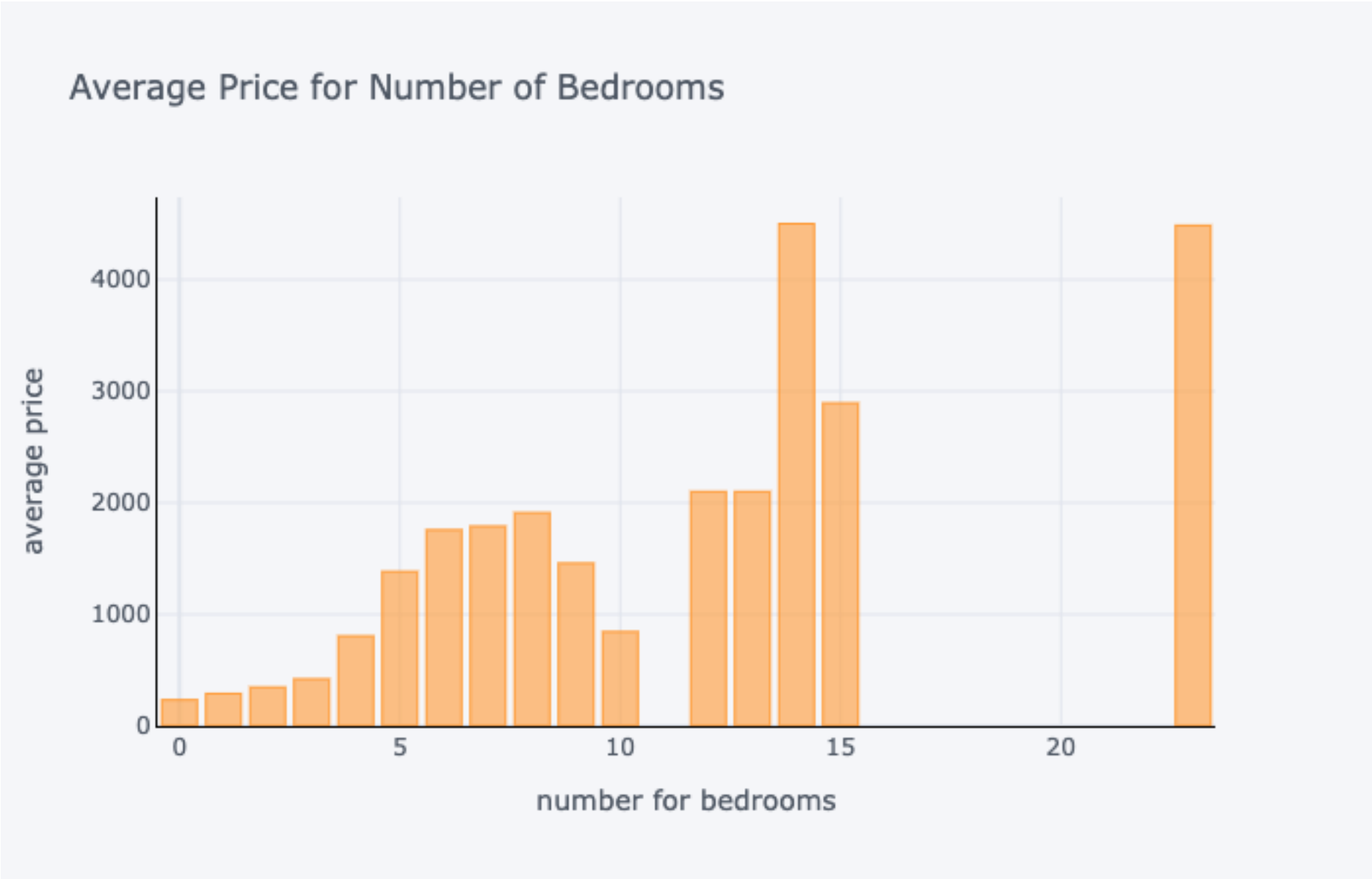
Discovered patterns, spotted anomalies and checked assumptions with the help of summary statistics and graphical representations.

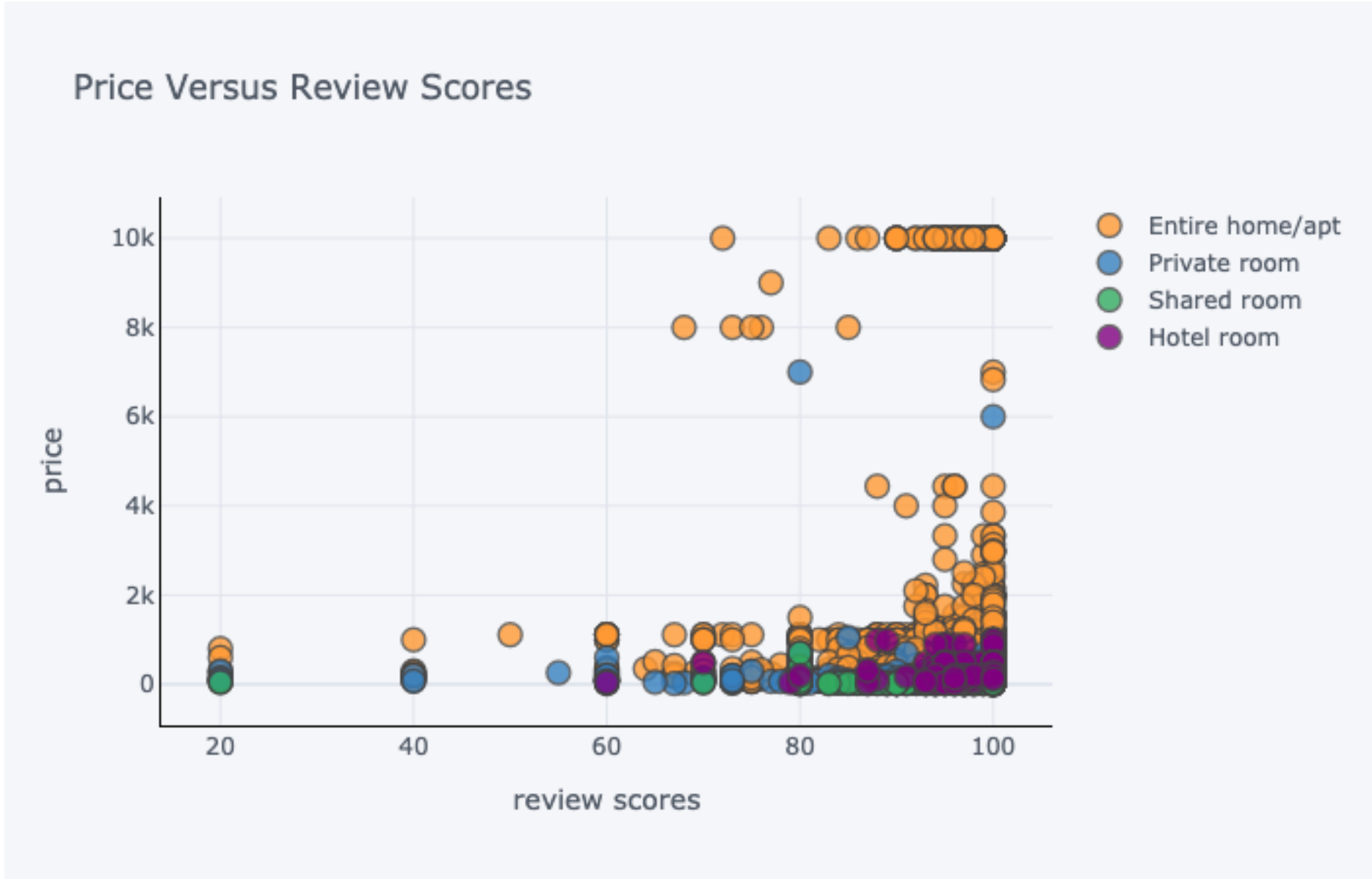

Price Distribution
count
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
price


Room Type Distribution
8000
6000
4000
2000
0
count
Entire home/apt
Private room
Shared room
Hotel room
room types


Booking Percentage of Room Type
booking percentage
0
10
20
30
40
50
60
70
Private room
Entire home/apt
Shared room
Hotel room
room types


Average Price of Room Type
average price
500
400
300
200
100
0
Hotel room
Entire home/apt
Private room
Shared room
room type


Average Price for Number of Bedrooms
average price
4000
3000
2000
1000
0
0
5
10
15
20
number for bedrooms


Price Versus Review Scores
price
10k
8k
6k
4k
2k
0
20
40
60
80
100
review scores
Entire home/apt
Private room
Shared room
Hotel room

OTHER FINDINGS

- **Most expensive neighborhoods were Westlake Hills, Barton Creek, Travis Heights, East Downtown and Old West Austin – many listings over $2,500 to $5,000 per night**
- **Median price for all listings was $145 and the average was $391**
- **Most expensive listing was $17,999 but the majority were under $300**
- **Listings with 1 bedroom, half bathroom, 1 bed or the possibility to accommodate 3 people were booked the most**

MODELING METHODS

Linear Regression (baseline model)
Ridge Regression
Lasso Regression
Random Forest Regression
Gradient Boosting
Support Vector Machine Regression

METHODS

RIDGE REGRESSION
LASSO REGRESSION

Regularization techniques to penalize magnitude of coefficients.

They differ by how penalty is assigned. Lasso can shrink them to zero.

RANDOM FOREST REGRESSION
GRADIENT BOOSTING

Ensemble methods that aggregates many decision trees.

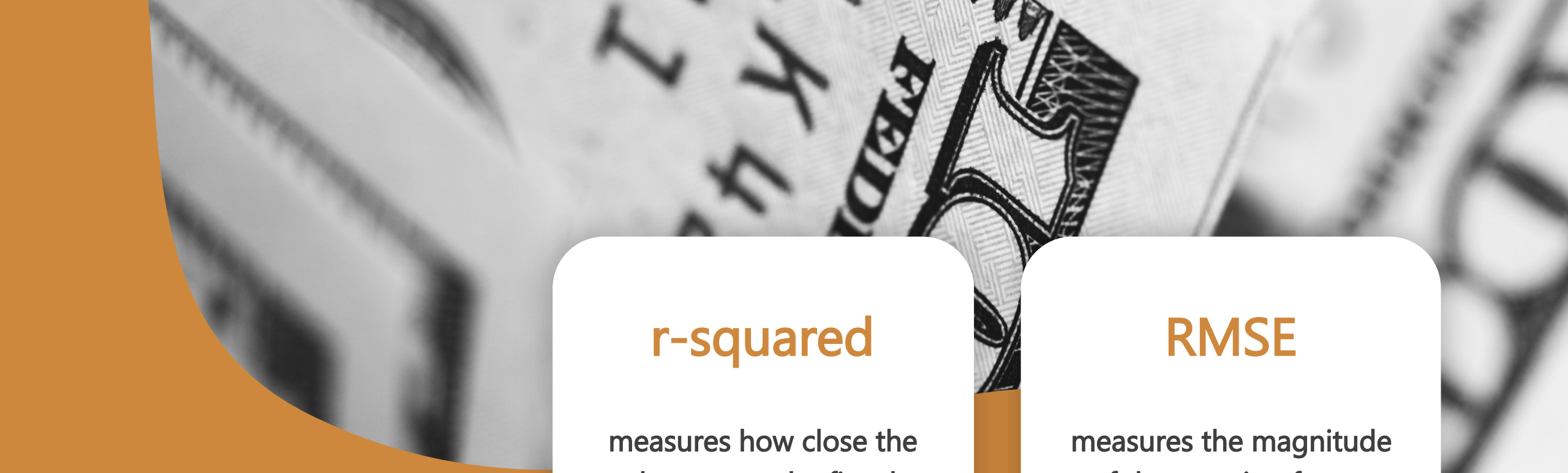
Random forest builds trees in parallel while gradient boosting does it serially.

SUPPORT VECTOR MACHINE REGRESSION

Can transform the data into higher dimensional space, in which the data becomes linearly separable.

METHODS CONTINUED

- **Scikit-Learn library**
- **Randomized search and grid search for best optimal model parameters**
- **3-fold and 5-fold cross validation instead of hold-out method**
- **Gradient boosting took the longest to train – 416 minutes**
- **Random forest took 42 minutes, the rest took less than 10 minutes each**



r-squared

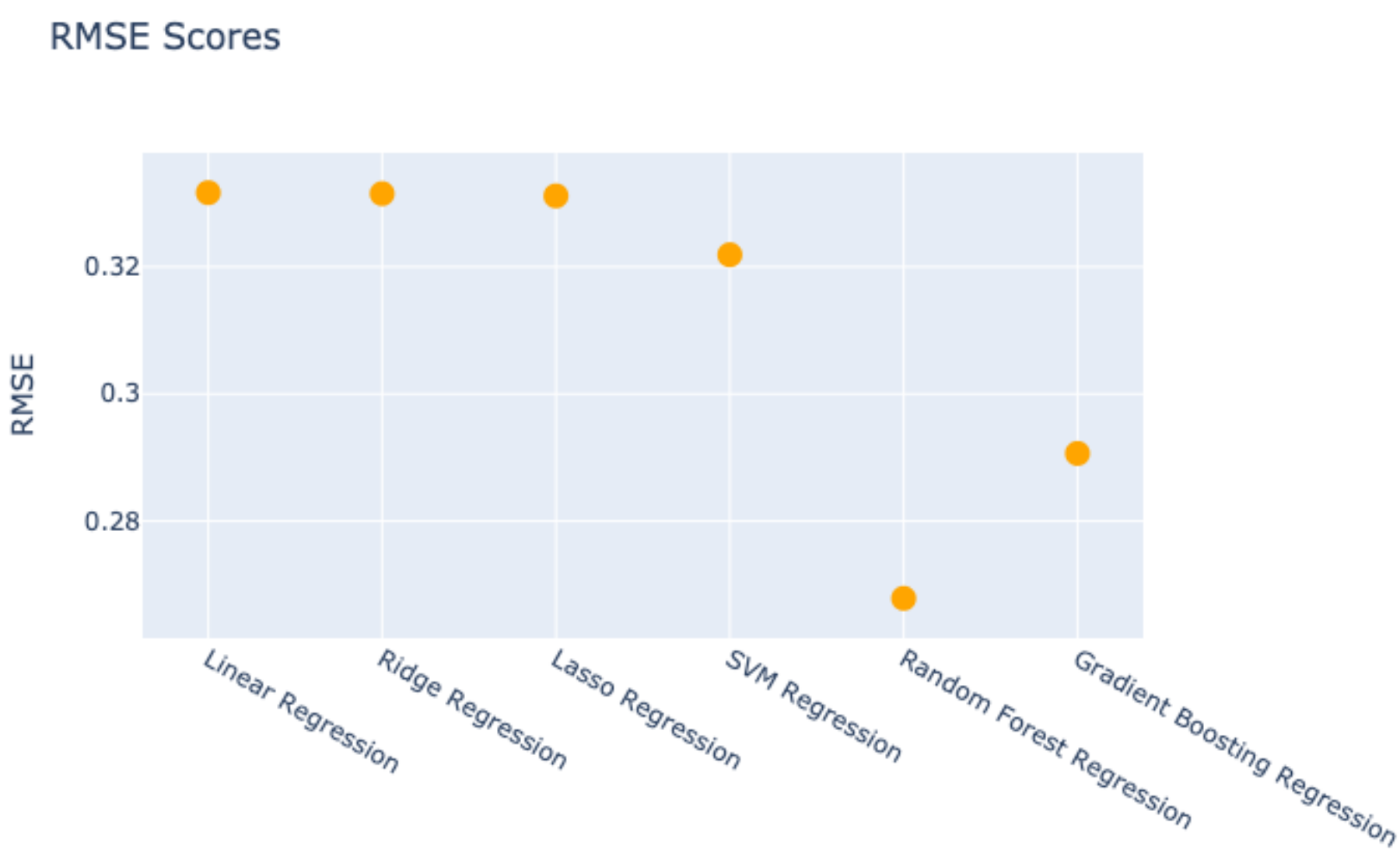
measures how close the data are to the fitted regression line

RMSE

measures the magnitude of the margin of error

EVALUATION METRICS


R-squared Scores
r-squared
0.65
0.6
0.55
0.5
Linear Regression
Ridge Regression
Lasso Regression
SVM Regression
Random Forest Regression
Gradient Boosting Regression

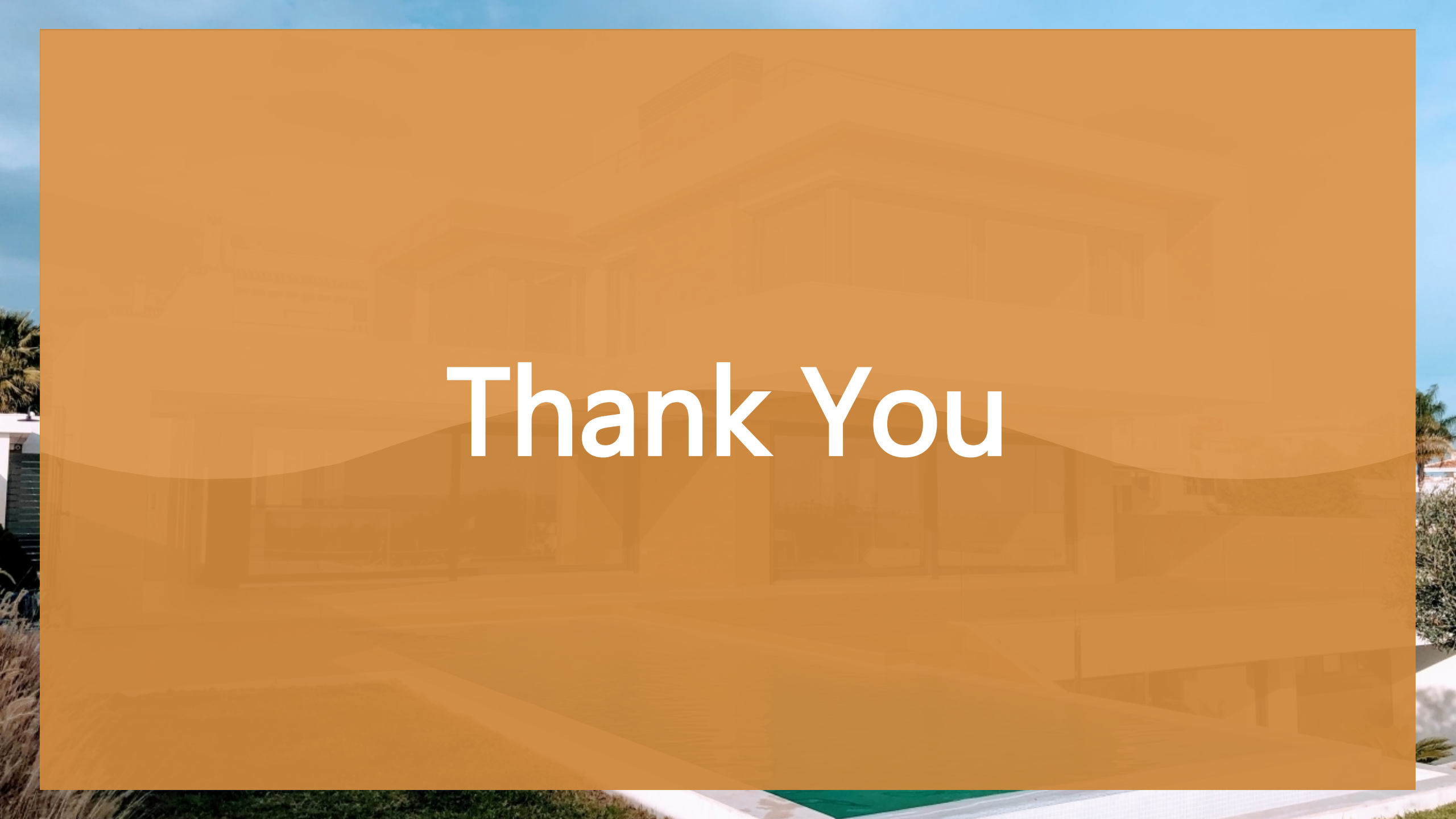

RMSE Scores
RMSE
0.32
0.3
0.28
Linear Regression
Ridge Regression
Lasso Regression
SVM Regression
Random Forest Regression
Gradient Boosting Regression

RECOMMENDATIONS

- **More bedrooms, bathrooms, beds and accommodating more guests = higher premiums**
- **Some neighborhoods can charge more than others – Westlake Hills and Barton Creek are the most expensive**
- **Entire home/apartment or private room = greater chance of getting booked**
- **Private room and shared room are significantly cheaper than entire home/apartment and hotels**
- **Listings at higher prices generally do not have scores under 70**

FUTURE EXPLORATION

- More feature engineering
- Other machine learning models
- Incorporating calendar data
- Apply model to other cities


Thank You