Predicting Airbnb Prices

By: Vishnu Kodicherla

Problem Statement

- With the growing number of hosts and customers for Airbnb in the past decade, I was assigned to help them in operational matters which include helping new hosts in being able to gauge the price they should be selling and help with them to establish a base price. Due to New York City having a lot of competition, it is necessary to provide a baseline understanding of the data.

Dataset

- Insider Airbnb- September 7, 2022, New York City

Data Cleaning

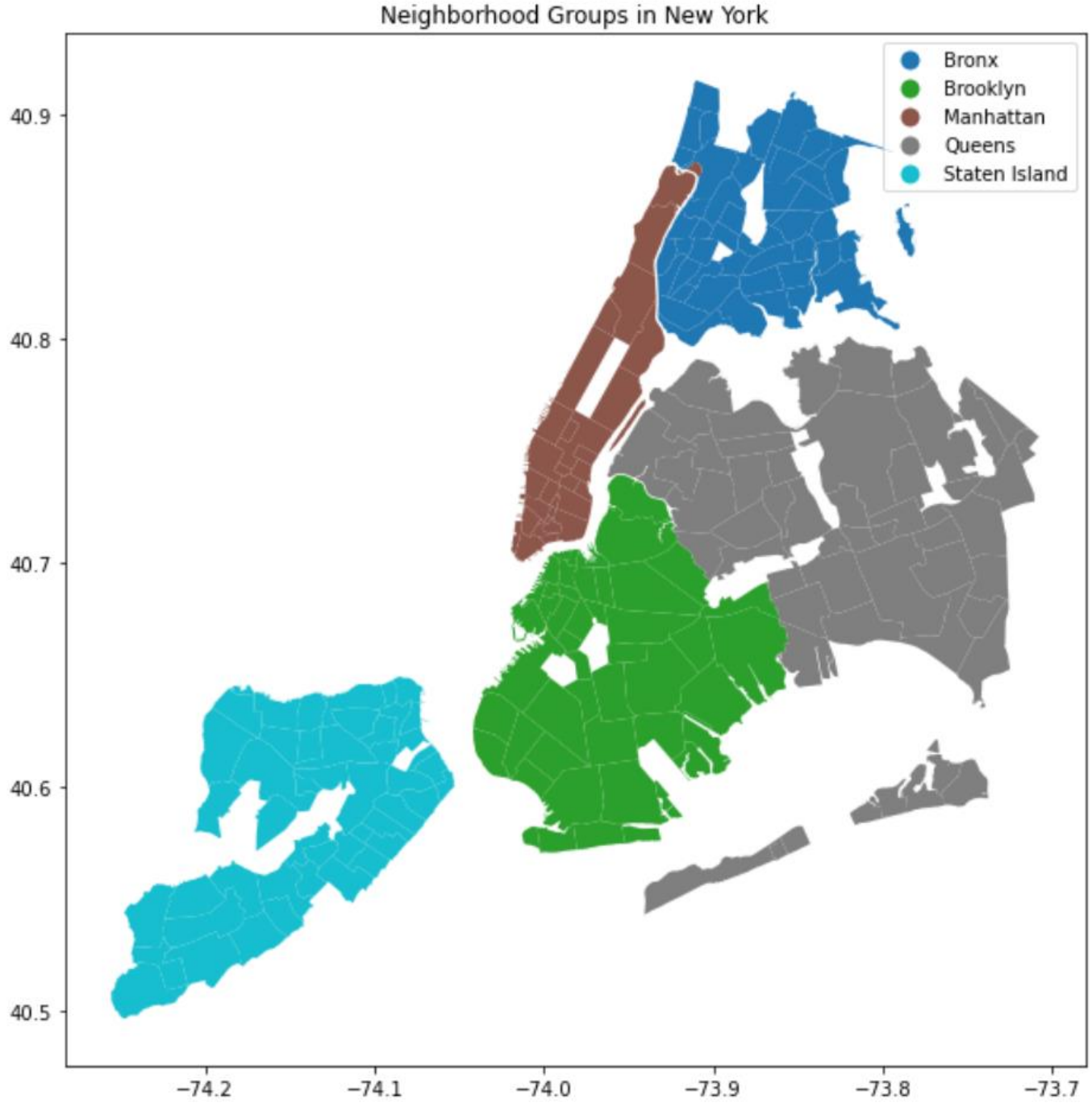
Got rid of any null values

Simplified the amenities columns by categorizing similar amenities together as their own separate column

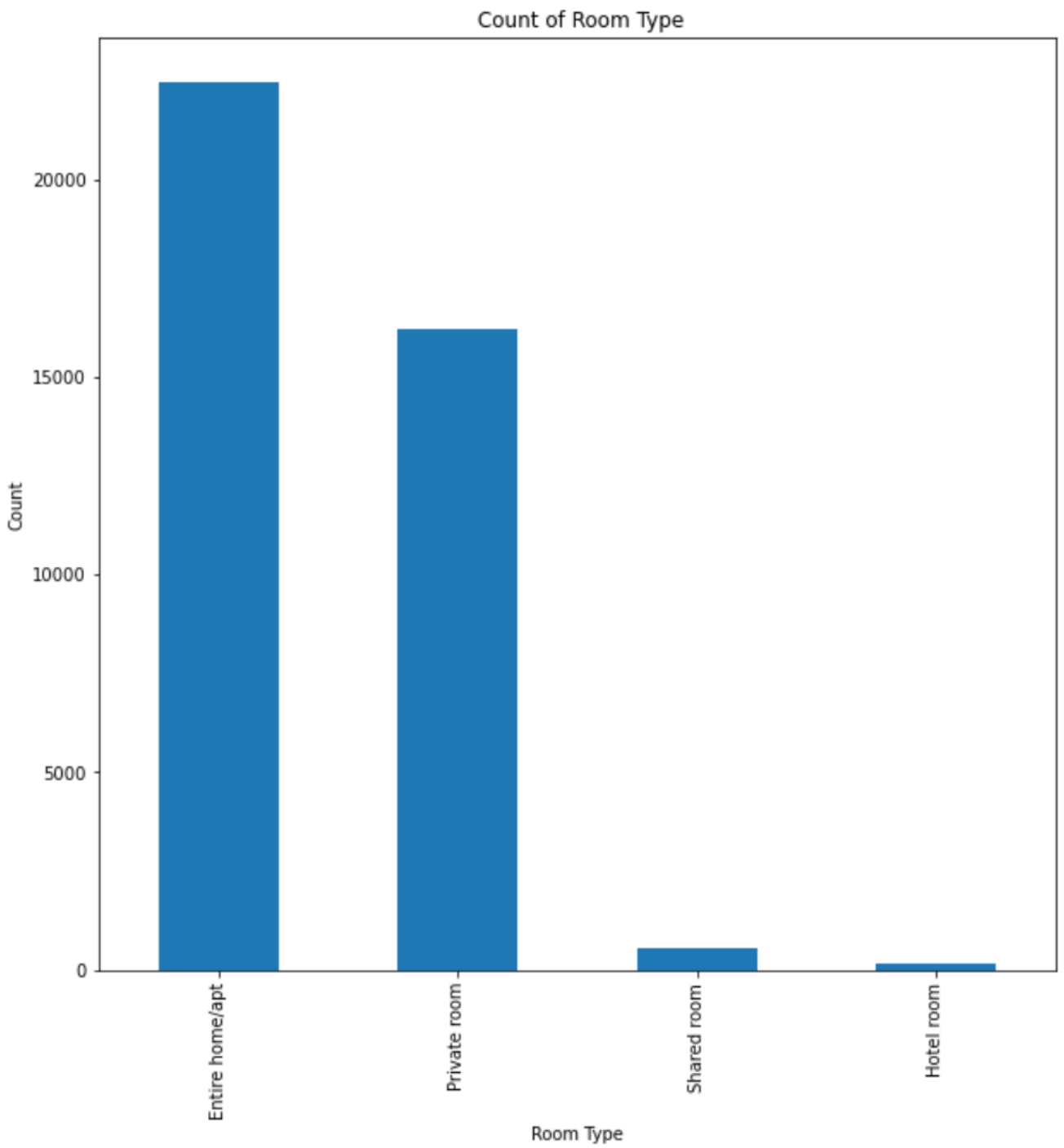
Unusual property types outside the most commonly used properties for AirBnB were made into a new column called other

A lot of the columns in the dataset had dtypes that were disoriented and so I made the appropriate changes to these columns

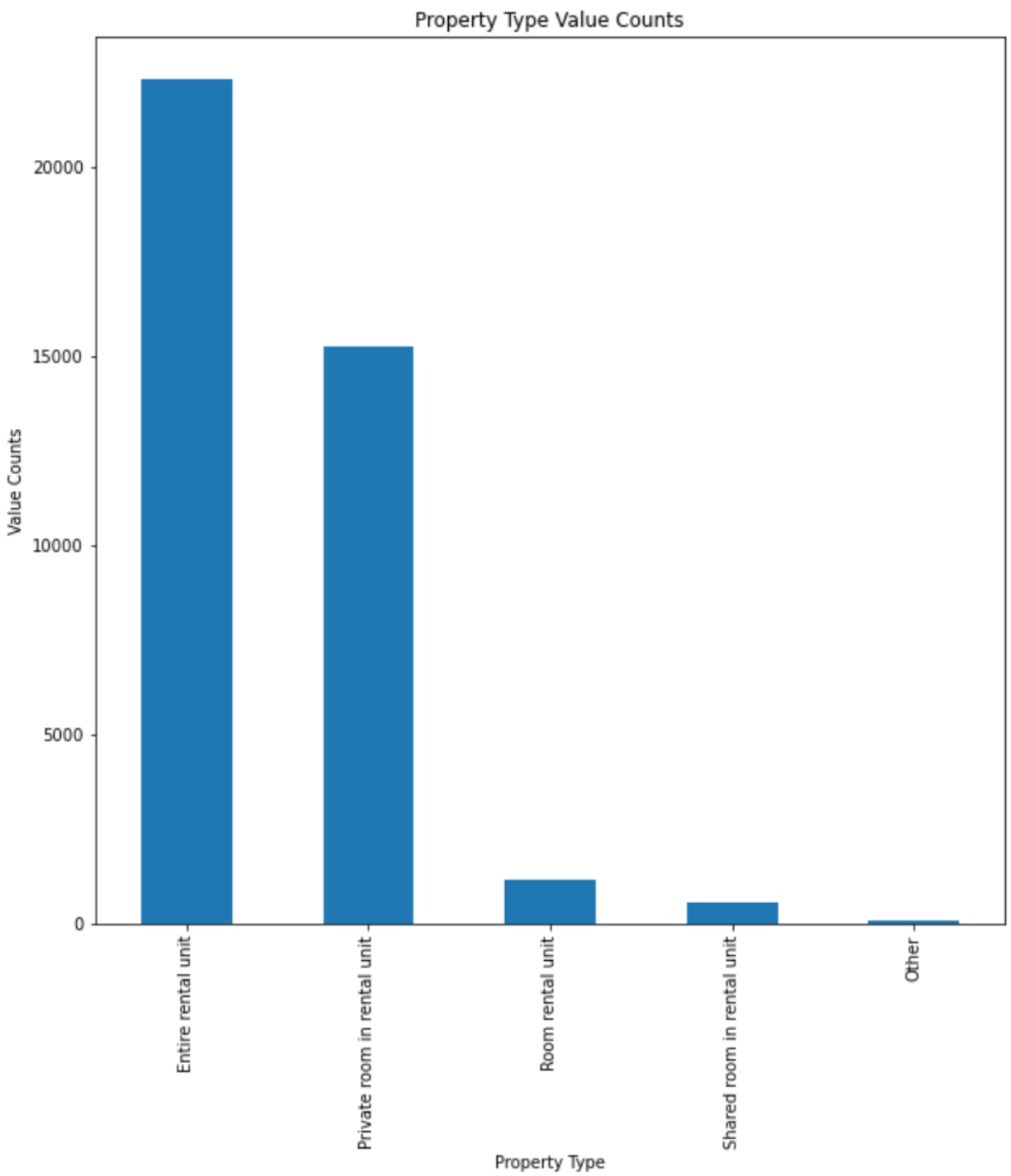
EDA


Neighborhood Groups in New York
Bronx
Brooklyn
Manhattan
Queens
Staten Island
40.9
40.8
40.7
40.6
40.5
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7

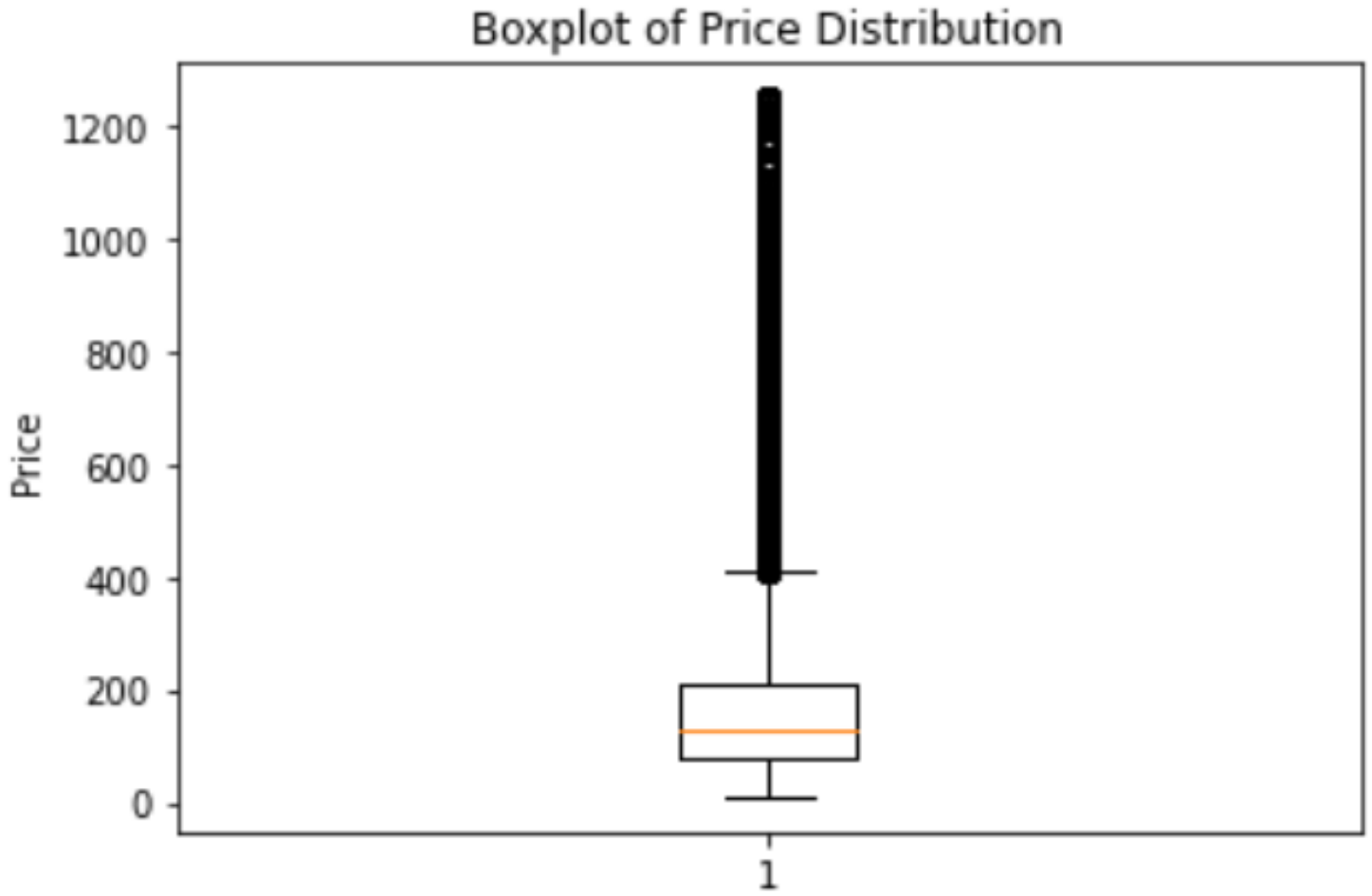
EDA


Count of Room Type
Count
20000
15000
10000
5000
0
Entire home/apt
Private room
Shared room
Hotel room
Room Type

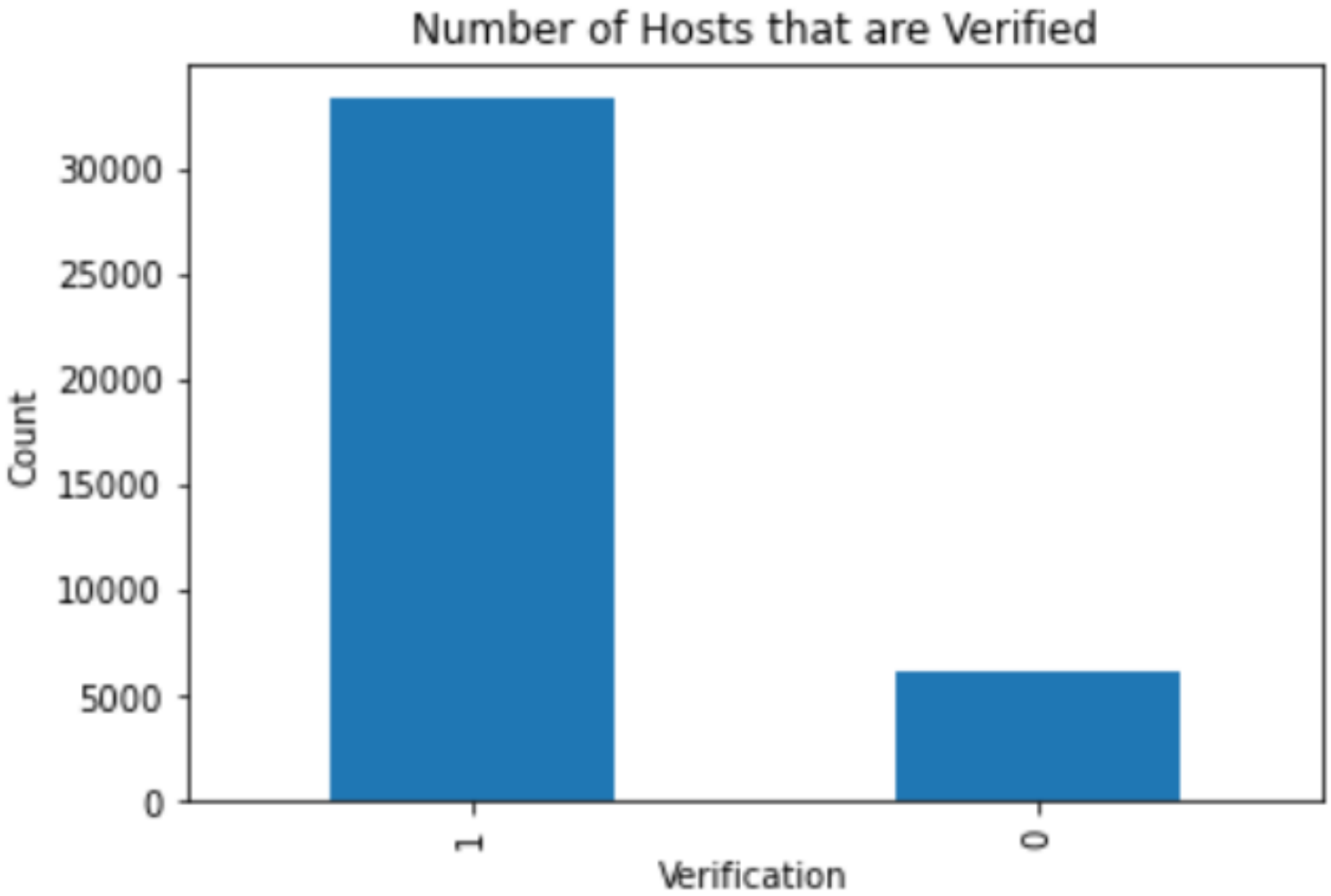
EDA

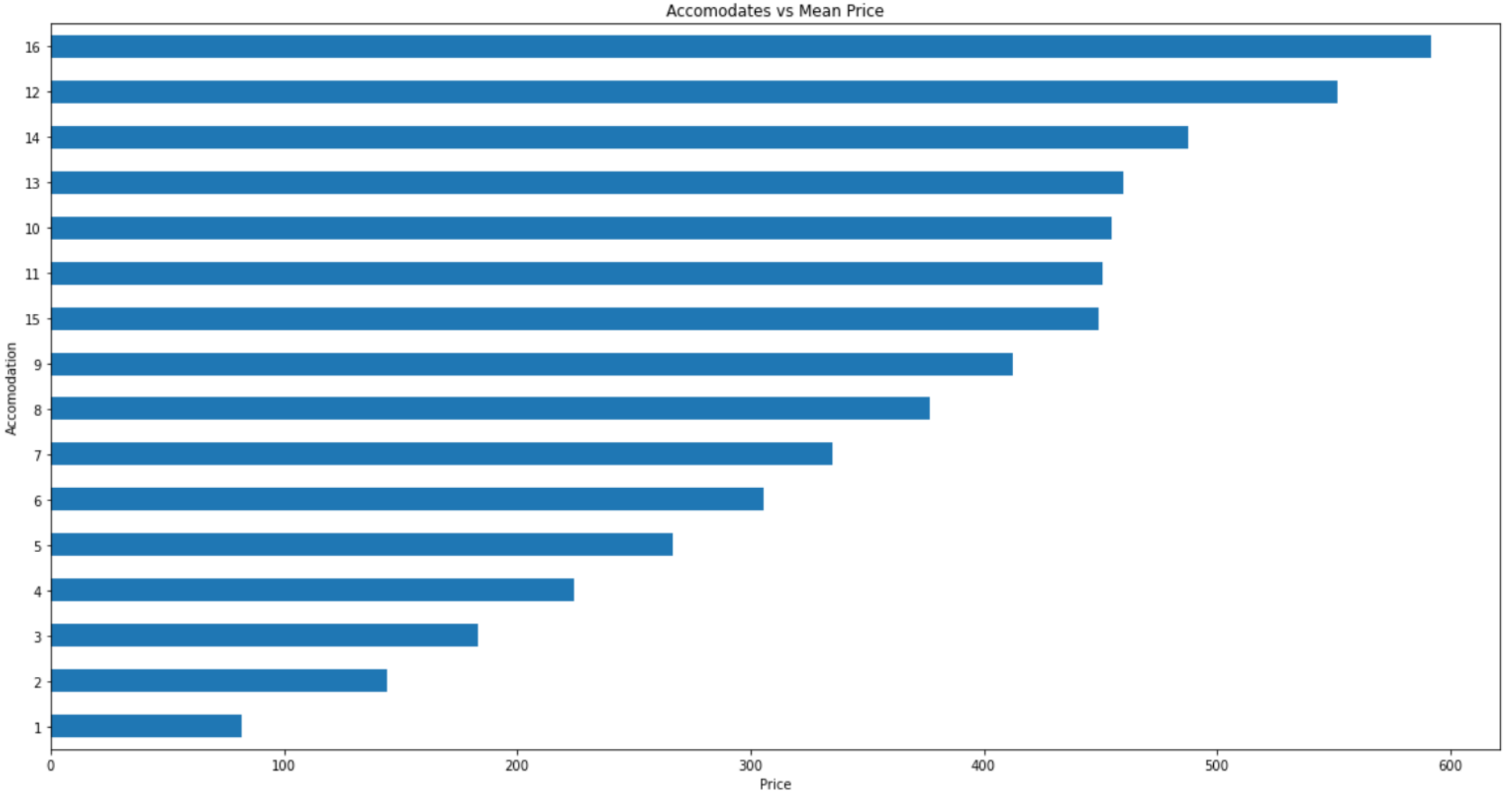

Property Type Value Counts
Value Counts
20000
15000
10000
5000
0
Entire rental unit
Private room in rental unit
Room rental unit
Shared room in rental unit
Other
Property Type

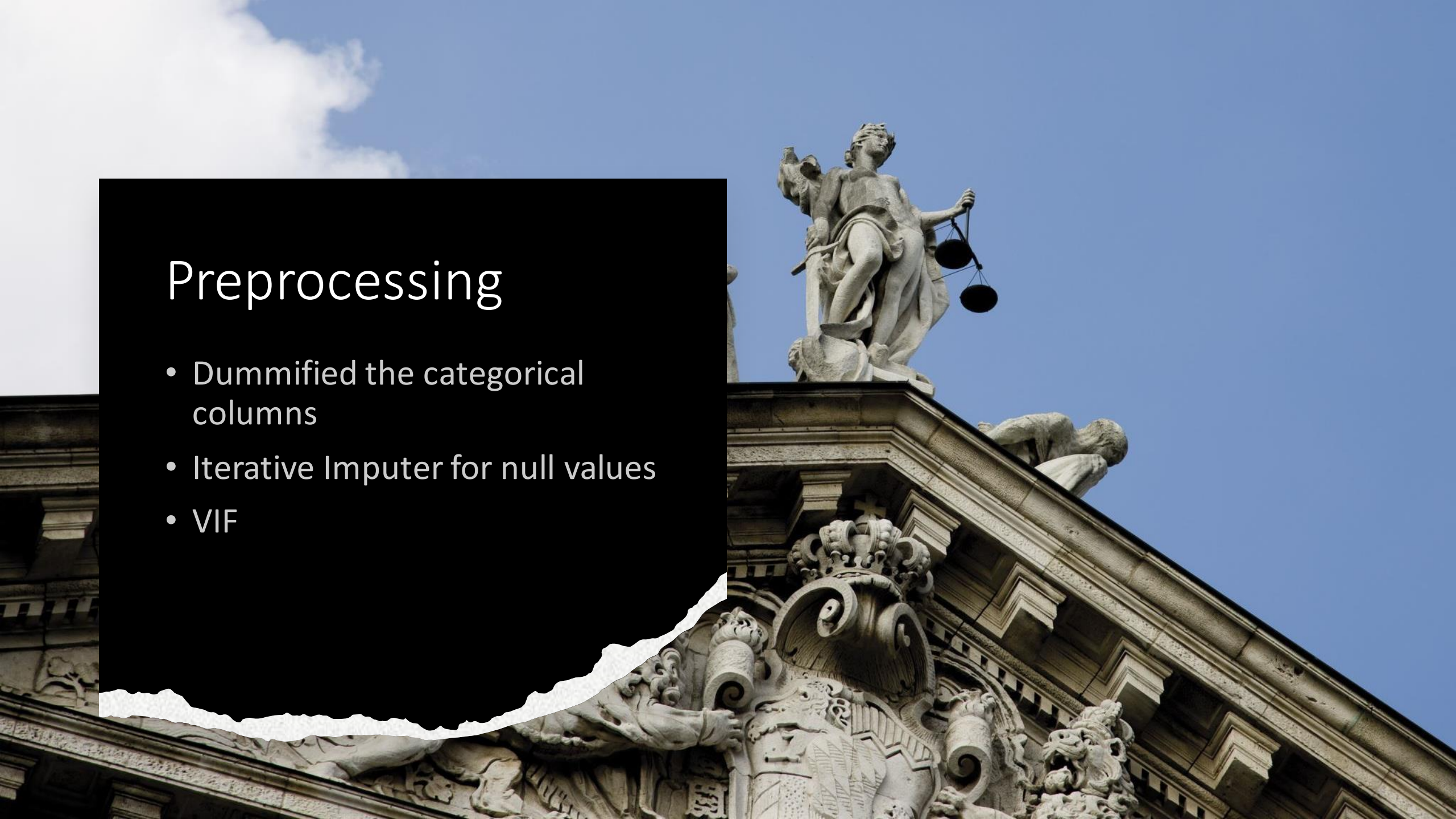
EDA


Boxplot of Price Distribution
Price
1200
1000
800
600
400
200
0
1

EDA


Number of Hosts that are Verified
Count
30000
25000
20000
15000
10000
5000
0
1
0
Verification


Accomodates vs Mean Price
Accomodation
16
12
14
13
10
11
15
9
8
7
6
5
4
3
2
1
0
100
200
300
400
500
600
Price



Preprocessing

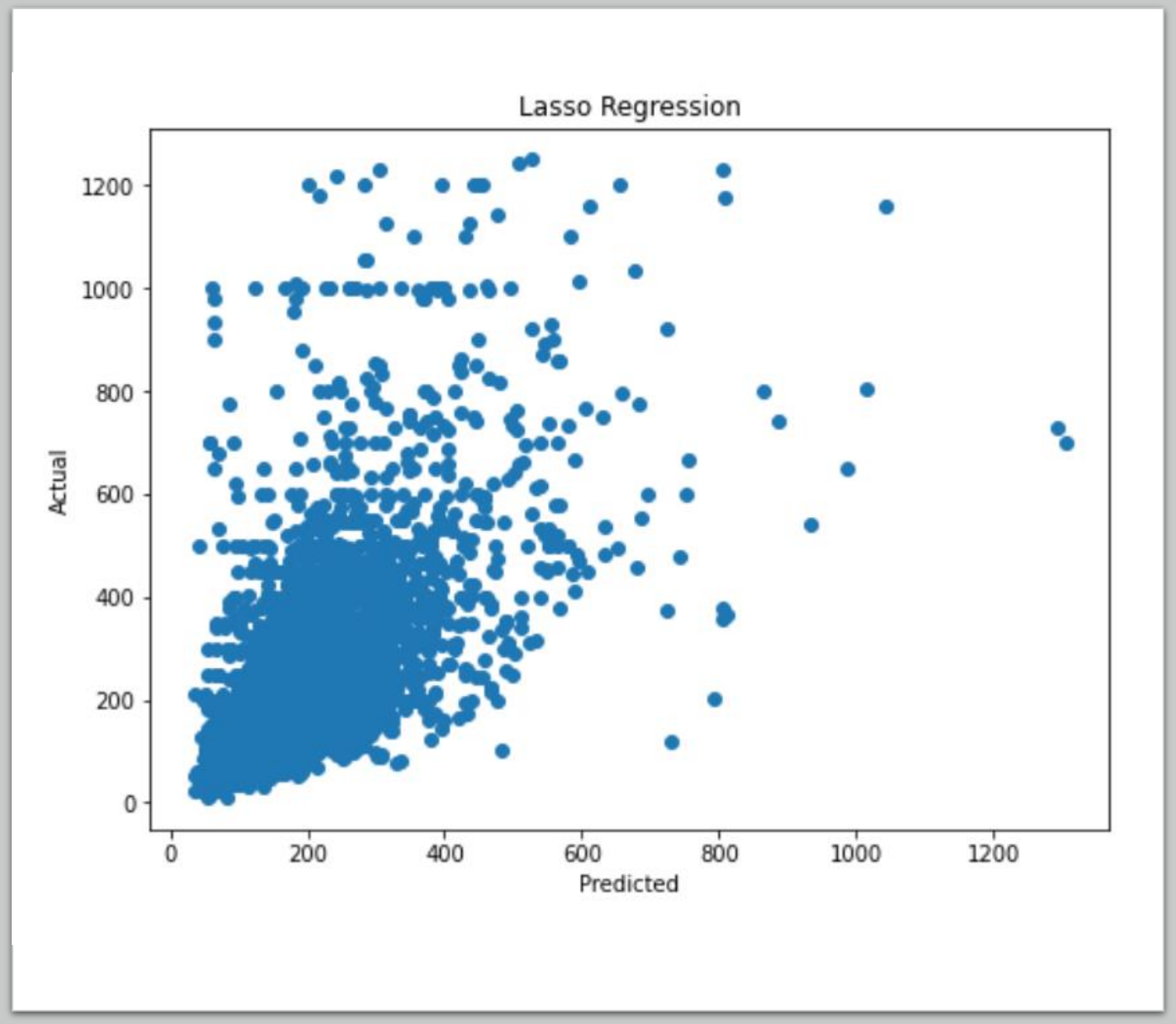
- Dummified the categorical columns
- Iterative Imputer for null values
- VIF

Models

- Linear Regression
- Lasso
- Ridge
- Bagging Regressor
- Random Forest
- 2 Neural Networks

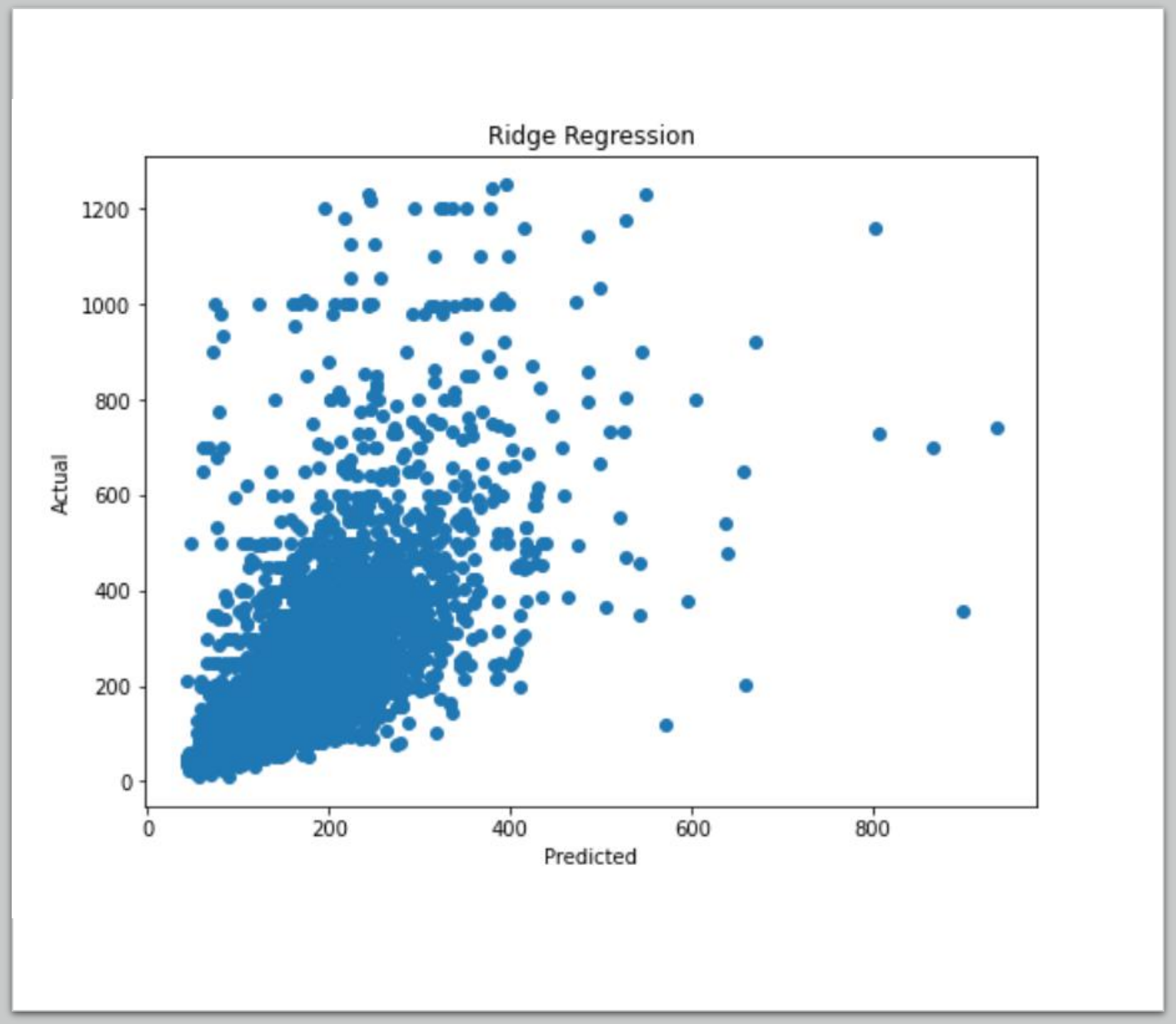
Best Model

- Best Model was Lasso Regression with a high R squared value of 0.65



Ridge

- Ridge had an R squared score of 0.62 and 0.63 for testing set



Conclusion

I would focus more into the data and see how using the log on the features that had skewed distributions would affect the models

Make more neural networks with denser layers to see how that would influence the models as well

NLP modeling with the predictions would help on accurately predicting AirBnB prices