Predictive Model for Unit Price of Properties on Airbnb

AGENDA

Motivation for dataset(business) chosen

Workflow

Insights/Conclusions

Challenges

Next Steps

Industry research

- Airbnb is a US based tech company that provides a platform for matching hosts(people) with guests who are looking for a short term rental property.

- Airbnb kickstarted a form of hospitality industry in cities across the world. Users use website to upload details of there property.

- Short term rentals includes different types of property such as cabins, apartments, lakefront, castles, countryside, skiing chalets, tiny homes

- Statistics for Airbnb : 150 million users across the world. Airbnb generates approximately $5.9 billion in revenue and about 300 million bookings made (2021)

- Chosen City for Analysis : **New York City**

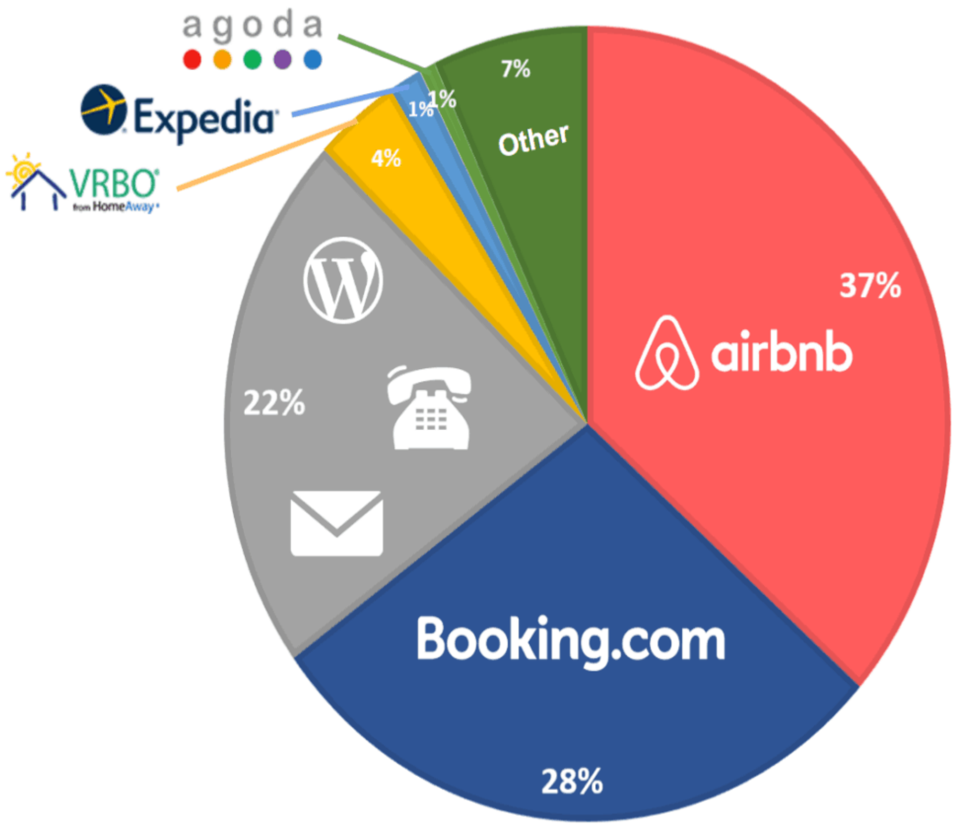


Figure 1. Booking Breakdown Based on Industry
Source: https://www.hosthub.com/

Objective

Predictive Model for Unit Price of Properties on Airbnb – NYC

- Interested in New York’s hospitality industry from the perspective of an Airbnb host
- In-depth analysis of data to develop a machine learning model.
- Optimizing the machine learning model to improve model performance
- Various factors within the dataset will be considered that will help the Airbnb host to set a price point for there property.

Data Preparation

- Dataset is taken from Kaggle

- Shape of Data: (48895,16)

- Features
- Host Name, Neighbourhood, Room type, Number of reviews

- Target – Price

Field	Description
id	Airbnb's unique identifier for the listing
name	Name of the listing
host_id	Airbnb's unique identifier for the host/user
host_name	Name of the host. Usually just the first name(s).
host_total_listings_count	The number of listings the host has (per Airbnb calculations)
neighbourhood	Neighbouhoods within the city
latitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
longitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
room_type	[Entire home/apt|Private room|Shared room|Hotel] All homes are grouped into the following three room types: Entire place Private room Shared room
price	daily price in local currency
minimum_nights	minimum number of night stay for the listing (calendar rules may be different)
availability_365	avaliability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
number_of_reviews	The number of reviews the listing has
calculated_host_listings_count	The number of listings the host has in the current scrape, in the city/region geography.
reviews_per_month	The number of reviews the listing has over the lifetime of the listing

Data Preparation

Data Cleaning

Checking Stage

- Checking for Data types
    i) Integers, floats, Categorical data

- Checking for Missing values
    i) NAN values

- Checking for Duplicates
    i) No Duplicates

Replacing Stage

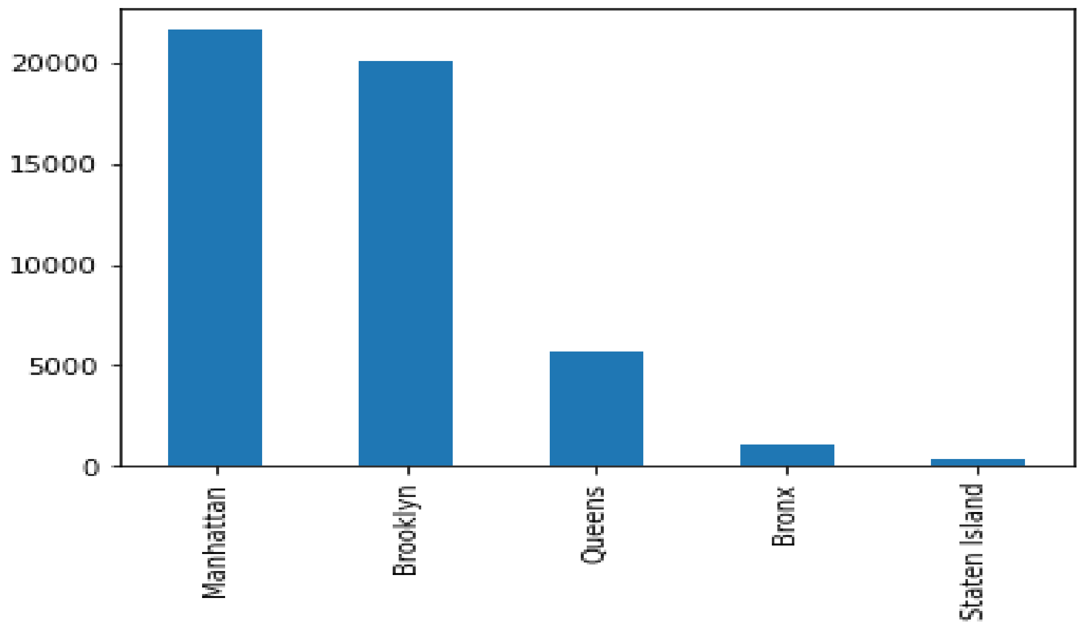
- Replace missing values with zero

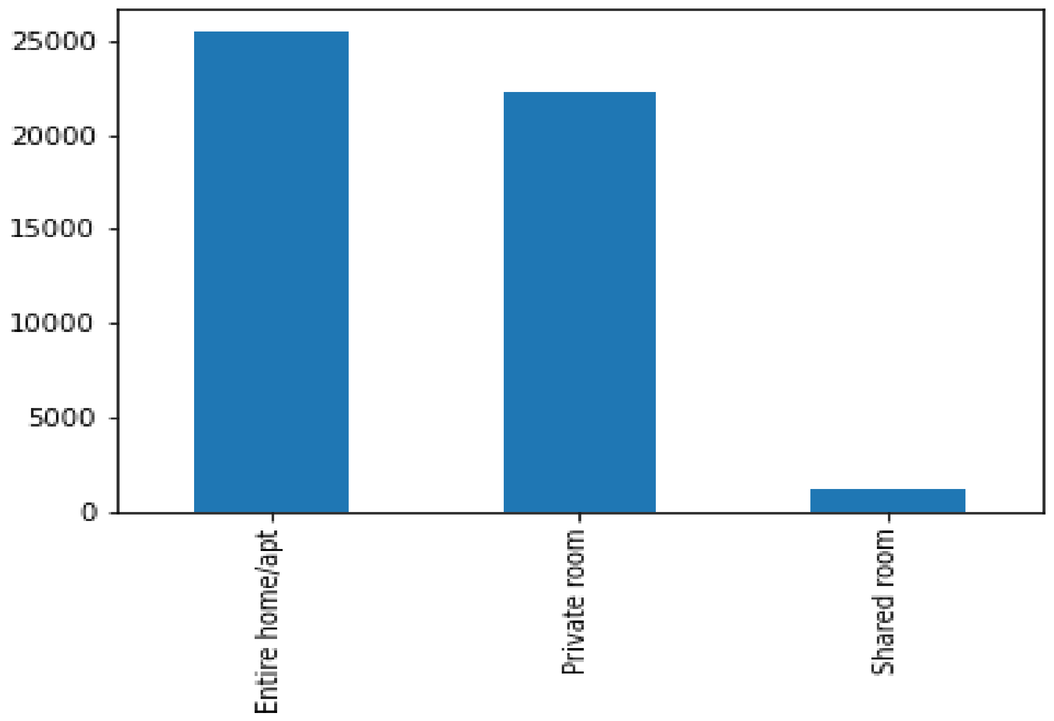
- Changed datatype for last review to datetime

EDA

Univariate Analysis

- **Count of Neighbourhood group**
- **Count of Room type**

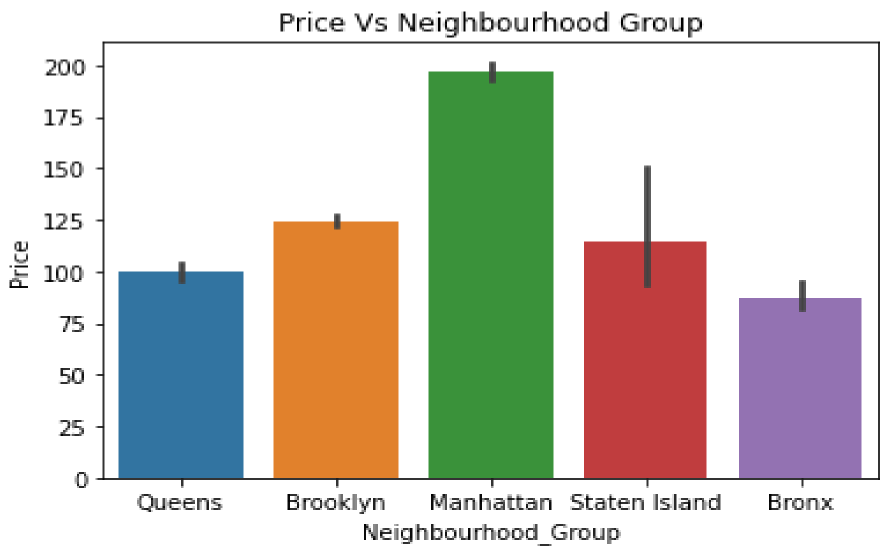


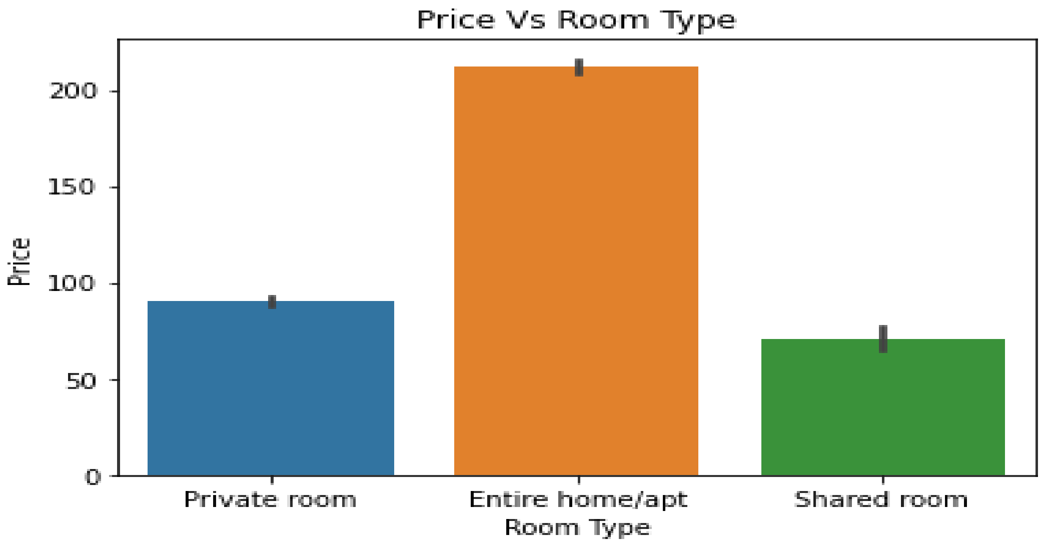


EDA

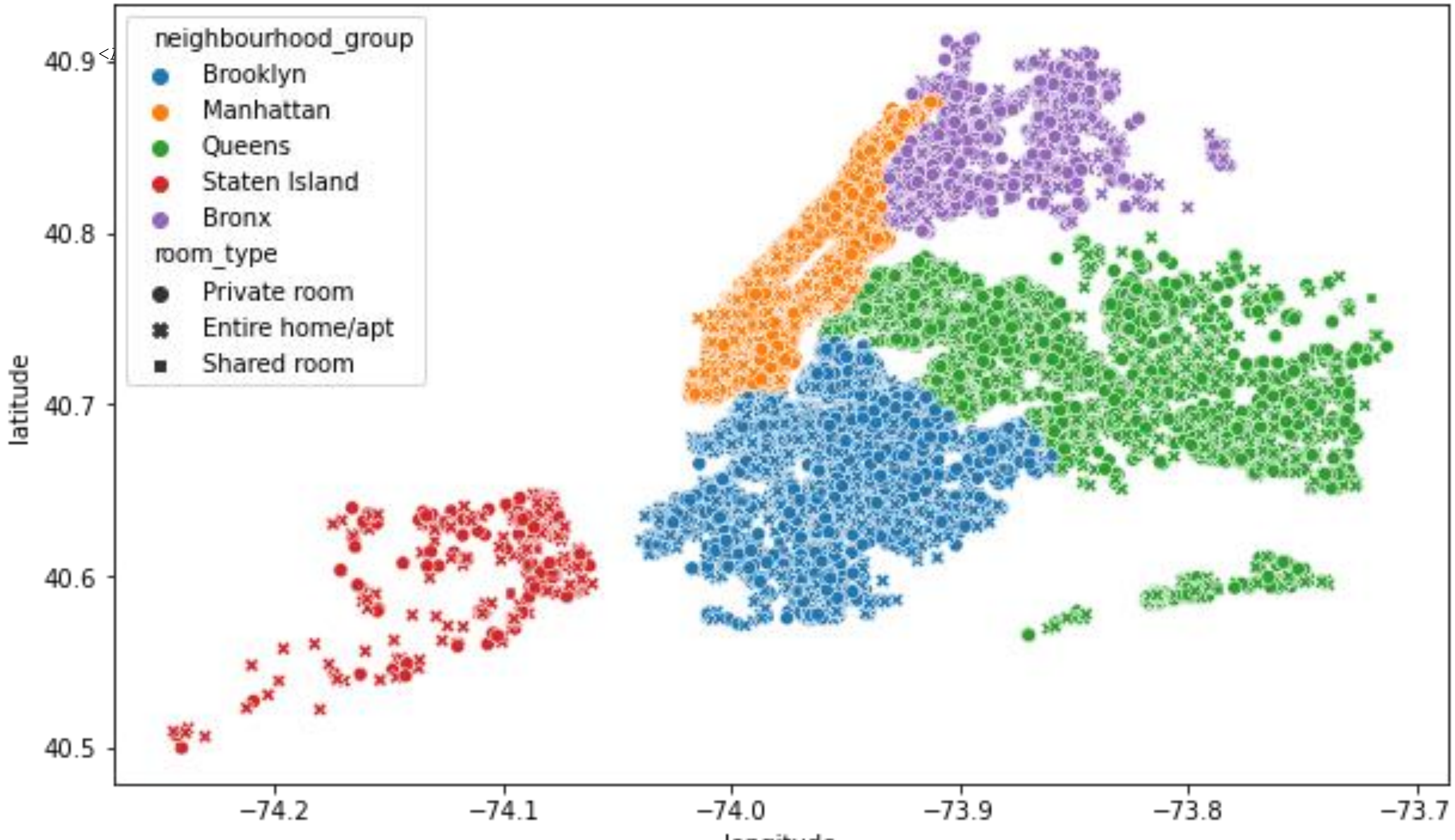
Bivariate Analysis

- **Bar plot – Price Vs Neighbourhood Group**
- **Bar plot – Price vs Room Type**
- **Scatterplot**





Observations



Observations

Categorical Variables of Importance

- Room Types
- Neighbourhood
- Neighbourhoods group
- Host name

Numerical Variables of Importance

- Latitude
- Longitude
- Reviews
- Minimum nights

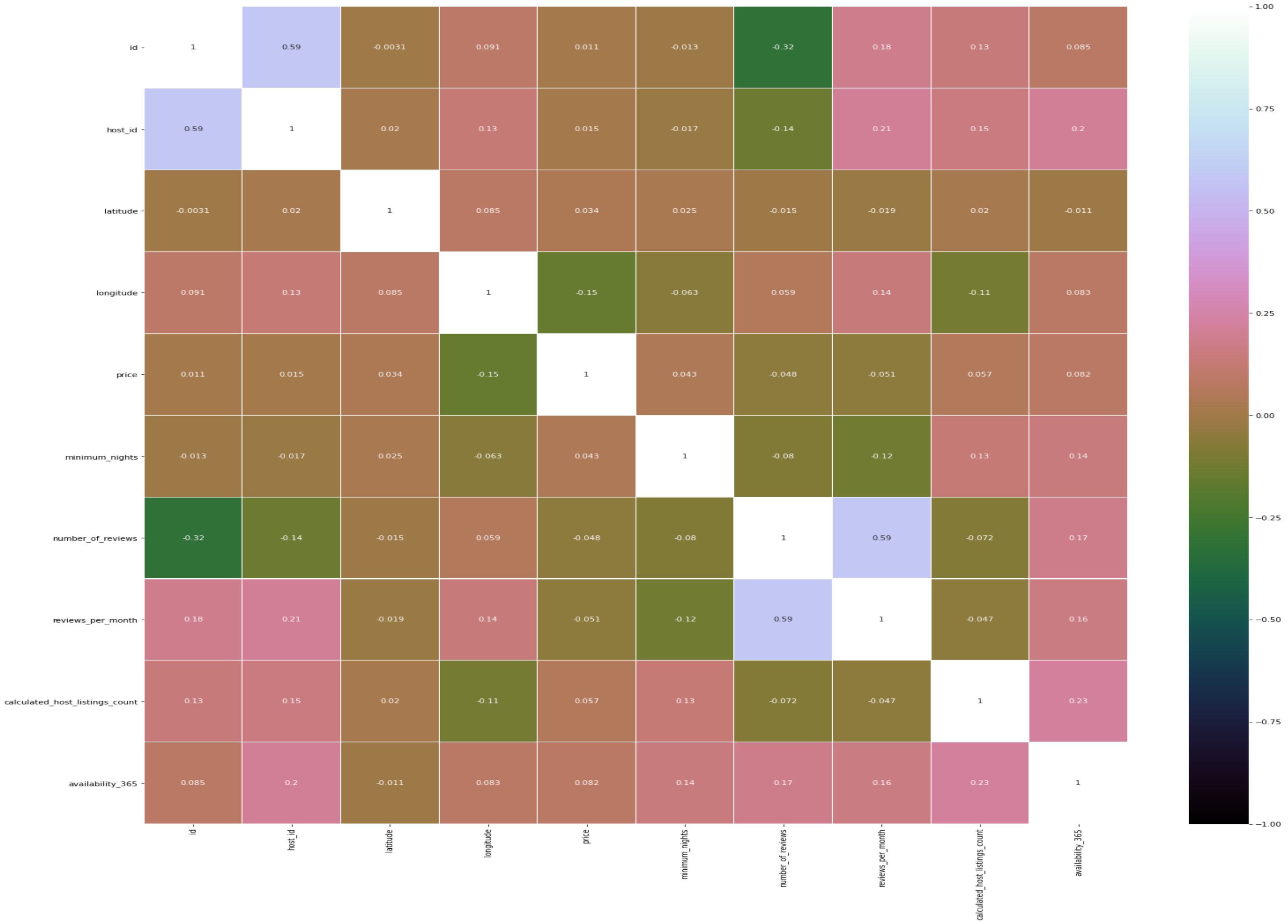
Feature Selection

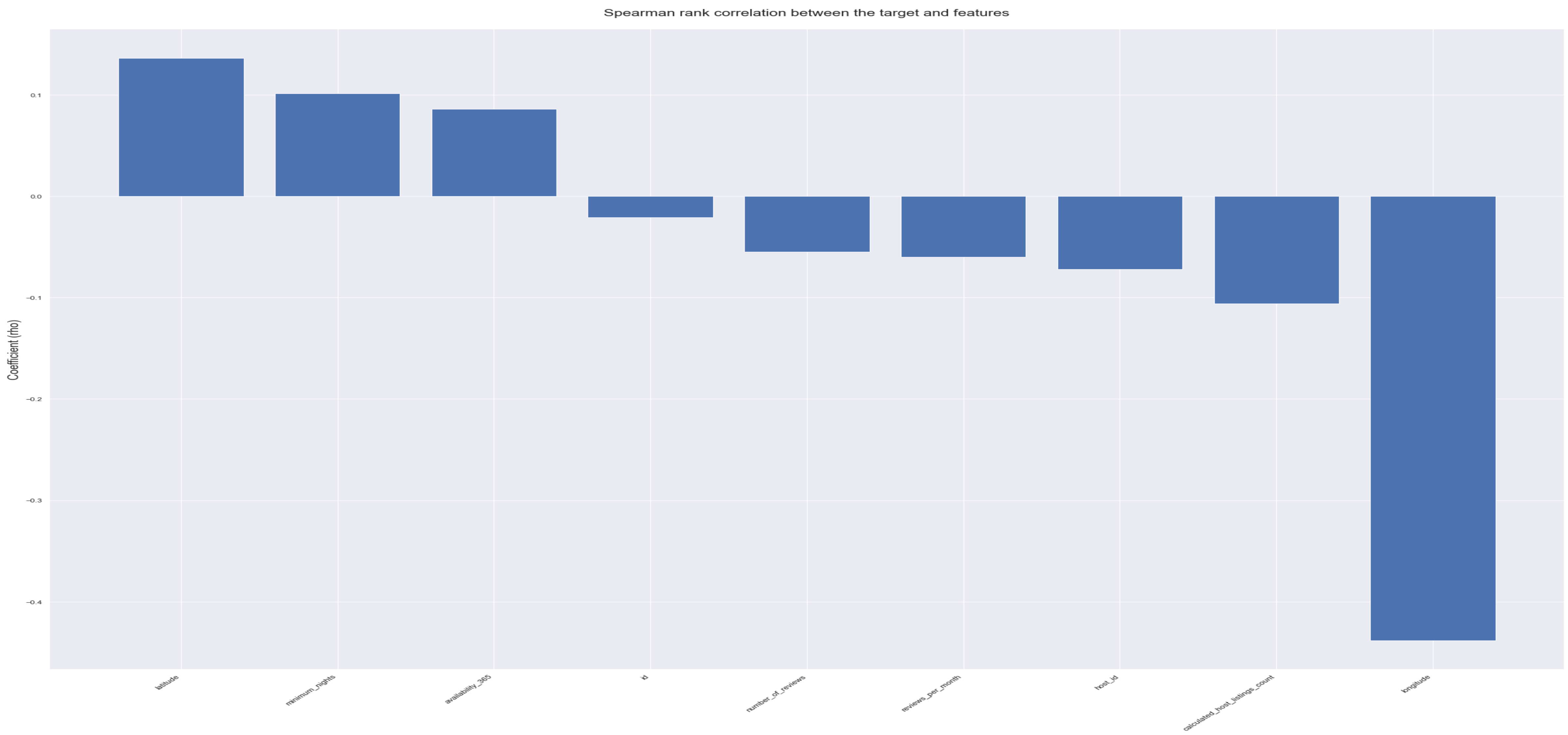
Pearson's Correlations
- **Heat map**

High Correlation Analysis
- **Spearman Correlation**

Multicollinearity Analysis
- **Variance Inflation Factor**

Dimensionality Reduction Analysis
- **PCA**





Spearman rank correlation between the target and features
Coefficient (rho)
0.1
0.0
-0.1
-0.2
-0.3
-0.4
latitude
minimum_nights
availability_365
id
number_of_reviews
reviews_per_month
host_id
calculated_host_listings_count
longitude

Model Building

Dropping features

Categorical Variables to Numerical Variables

- pd get dummies
- Shape: 48895 rows × 237 columns

Train/Test Split

Min Max Scaling

Baseline Model – Linear Regression

Results:

predicted 1-5: [225.9024561   88.82045955 97.73996587 248.37350158 228.9605378]
actual 1-5: [89, 30, 120, 470, 199]
MAE: 996089683.1034601
MSE: 2.912849858153375e+21
RMSE: 53970824138.17094
R-Squared: -7.212290253922323e+16
EVS: -7.209833555103493e+16

Model Building

Regression Analysis

- **Random Forest Regressor: RMSE:186   R-Squared:0.143**
- **Decision Tree Regressor : RMSE: 186   R-Squared:0.143**
- **KNN: RMSE: 200   R-Squared: 0.009**

Regression Analysis – Hyperparameter tuning

- **Random Forest Regressor**
- **Cross Validation and GridSearchCV**
- **Parameter Tuned: Estimator, Maxdepth, Maxfeatures, Max sample leaf**
- **Results:**

predicted 1-5: [154.74 115.88 144.88 224.19 168.76]
actual 1-5: [89, 30, 120, 470, 199]
MAE: 63.31
MSE: 34600.82
RMSE: 186.01
R-Squared: 0.14327
EVS: 0.14439

Model Building

- **Decision Tree Regressor: RMSE: 189 R-squared: 0.108**
- **ExtraTrees Regressor: RMSE: 190 R-squared: 0.143**
- **Gradient Boosting Regressor: RMSE: 439 R-squared:-3.79**
- **AdaBoosting Regressor: RMSE: 188 R-squared: 0.116**

- **Observations:**
- **Random Forest Regressor performed best after hyperparameter tuning Low RMSE, High R-square in comparison to the other models tuned.**
- **More robust hyperparameter tuning is required to get better results**

Challenges and Next Steps

Challenges Faced

- Categorial Features – Correlation analysis
- Robust outlier analysis is required
- Regression Analysis results were on the lower side, might not be a good predictor for price
- Feature Selection

Next Steps

- Classification problem with price splitting into 3 categories
- Include crime data and distance to transportation such as subway system
- Deep learning
- Build Pipelines
- Auto-Sklearn

Thank you!!!