– DATA ANALYSIS –

AIRBNB IN NEW YORK CITY

Group 1
Thomas Kantaros,
Gregor Shaw, Shivam Sinha

ALY 6140

Prof. Richard He

12/09/24

INTRODUCTION TO DATASET

- New York City Airbnb Open Dataset from Kaggle
- Contains ~100,000 entries and 16 relevant fields

Variable	Description
host_identity_verified	Whether the host's identity is verified
neighbourhood_group	The borough where the listing is located
neighbourhood	The neighborhood that the property is in
instant_bookable	Where the listing can be booked instantly (Yes/No)
cancellation_policy	The type of cancellation policy offered
room_type	The type of room offered (e.g., Entire Home, Private Room)
construction_year	Year when the property was constructed
price	Cost per night for the listing
service_fee	Additional fee for services per booking
minimum_nights	Minimum number of nights required for booking
number_of_reviews	Total number of reviews received by the listing
last_review	Date of the most recent review
reviews_per_month	Average number of reviews per month
review_rate_number	Average review score given by guest
availability_365	Whether the property is listed all year round (Yes/No)
house_rule	Whether the listing includes house rules for guests

Goal: Draw insights into the factors that contribute to the price and review scores of Airbnb listings in New York City

BRONX
NEW YORK
Nassau
Essex
Hudson River
MANHATTAN
East River
La Guardia Airport
NEW JERSEY
Hudson
Newark International Airport
QUEENS
Union
Upper Bay
BROOKLYN
John F. Kennedy International Airport
Jamaica Bay
STATEN ISLAND
Middle-esex
Lower Bay
ATLANTIC

Source: https://www.loumovesyou.com/blog/boroughs-of-new-york/

EXPLORATORY DATA ANALYSIS

Guiding Questions:

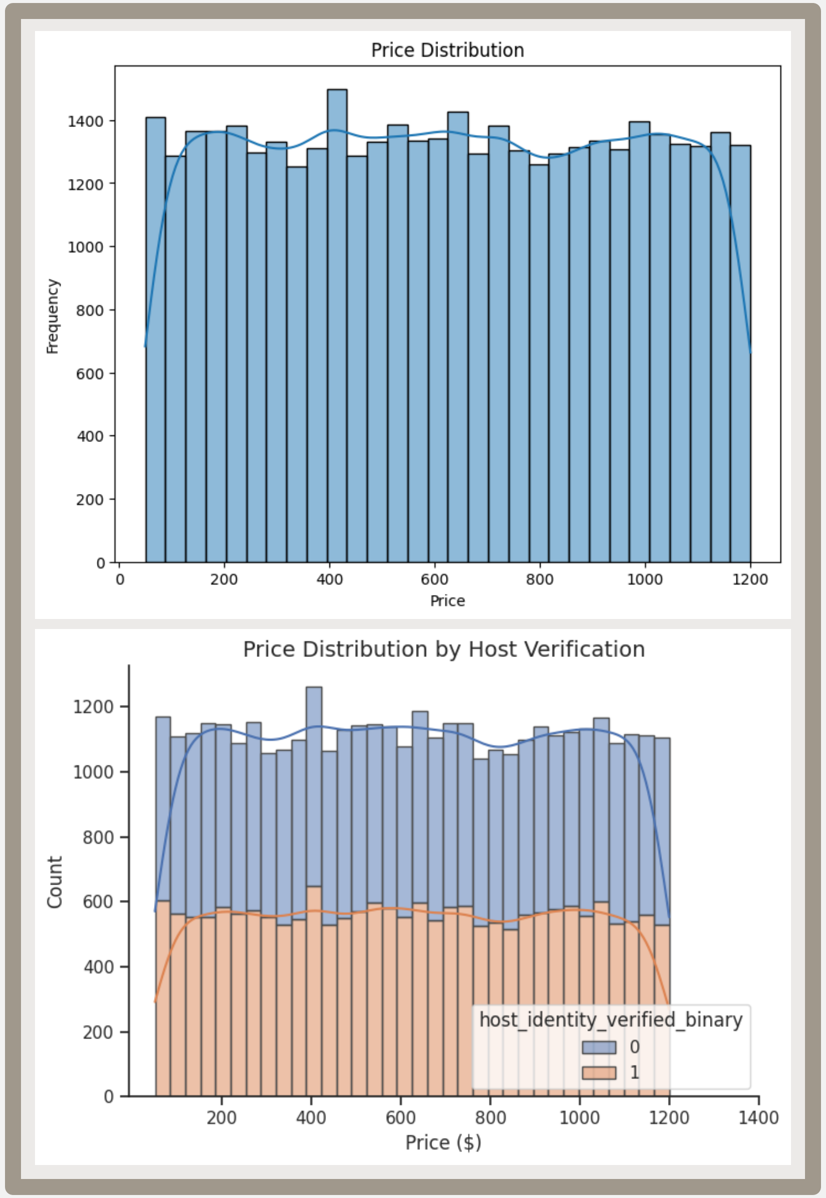
- How does occupancy rate vary between room types and neighborhoods?
- Are rental prices associated with the property's year of construction?
- Which characteristics are associated with high review scores?
- Does host verification impact price and review score?
- How does the age of the listing (how long the property has been on Airbnb) affect review scores or occupancy rates?

DATA CLEANING AND PREPARATION

- **Column Reduction**: Removed unnecessary columns to streamline the dataset and reduce bloat.
- **Data Type Optimization**:Assigned proper numeric and categorical data types in the Pandas DataFrame.
- **Handling Missing Data**: Removed rows with missing values in relevant columns and replaced null values in the "reviews per month" column with zeros.
- **Data Transformation**: Created new variables (e.g., occupancy rate) and converted string price values to numeric by removing dollar signs.

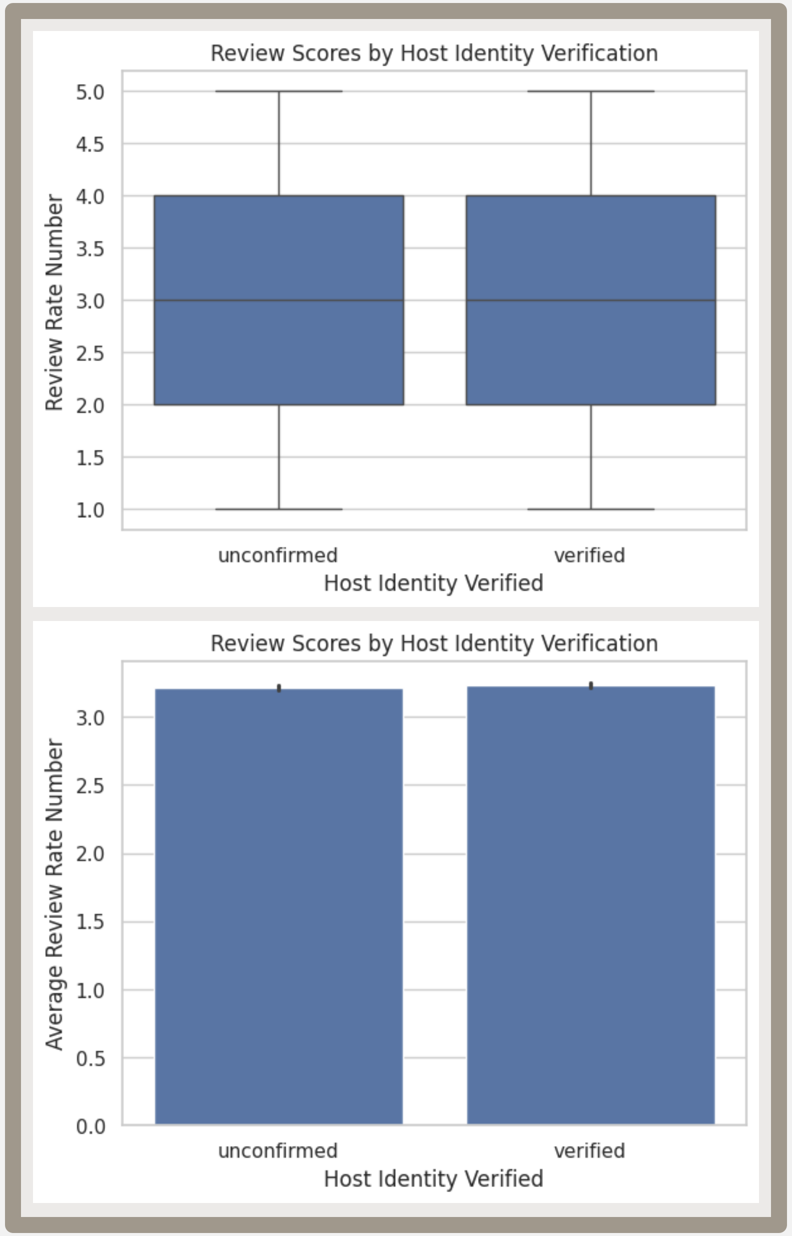
PRICE DISTRIBUTION

- Price is distributed uniformly across its range. This was surprising – we expected a normal distribution
- No obvious sign that verified hosts tend to charge higher prices. Again – somewhat surprising.



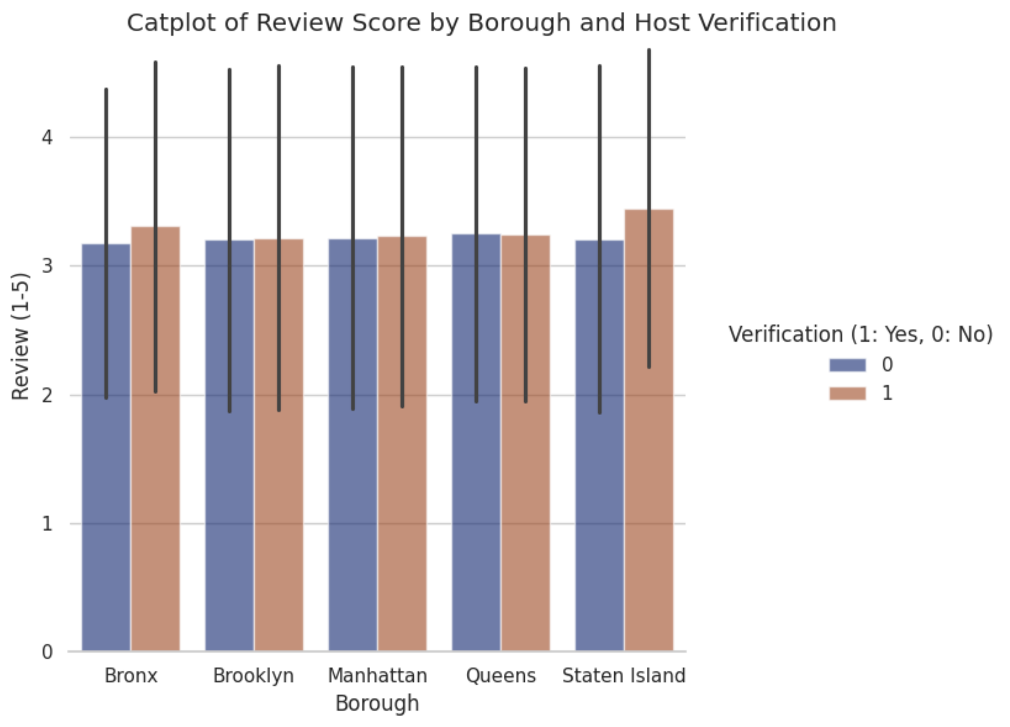
VERIFICATION AND REVIEW SCORE

- Review scores are identically distributed between verified and unverified hosts
- More evidence that whether a host is verified is unimportant



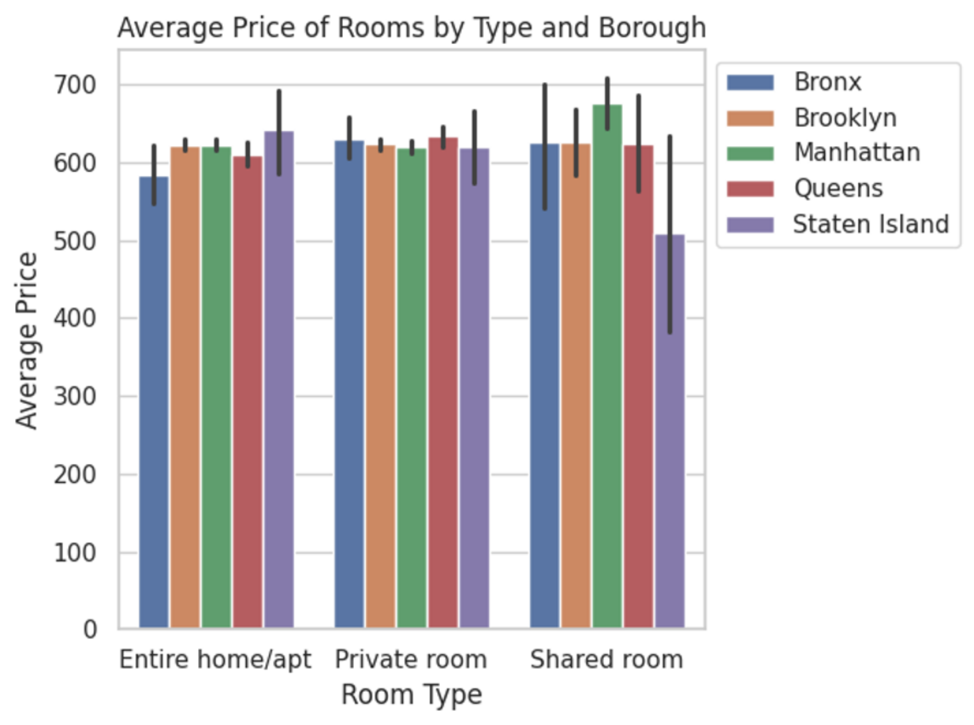
REVIEW SCORE

- No borough yields far higher reviews than the others
- Again, verification plays little role, although there is a marked difference between scores of verified and non-verified hosts in the Bronx and Staten Island



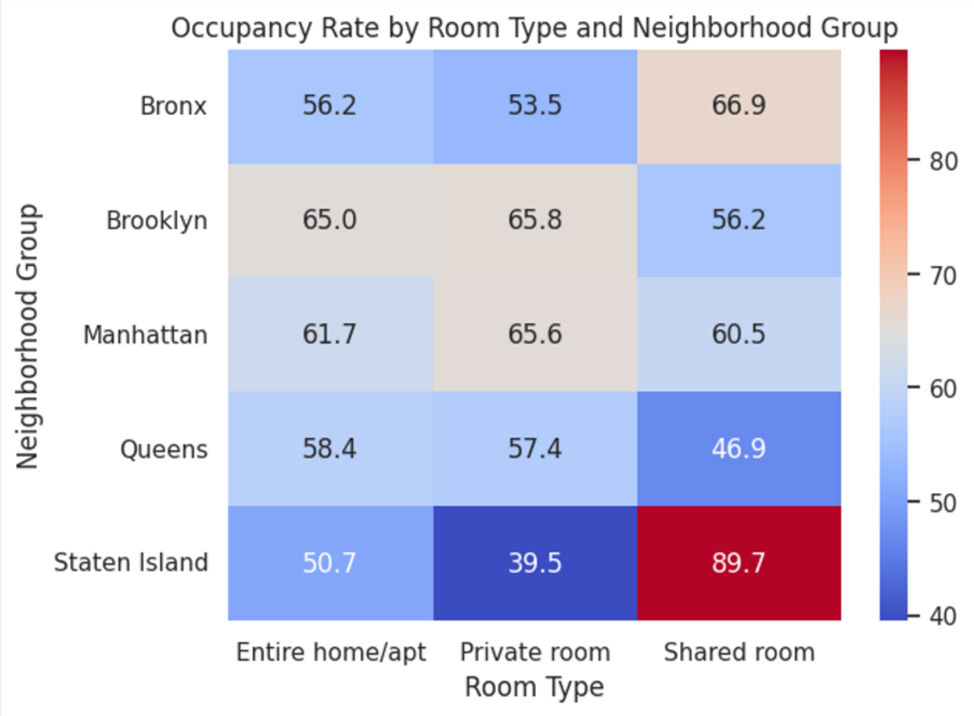
ROOM TYPE

- Generally modest differences in average price neighborhoods and room types
- The biggest exception to this is Staten Island shared rooms
  - ➤ Could be why they are so frequently occupied
- Most expensive are shared rooms in Manhattan
  - ➤ Highest real estate prices in all of USA



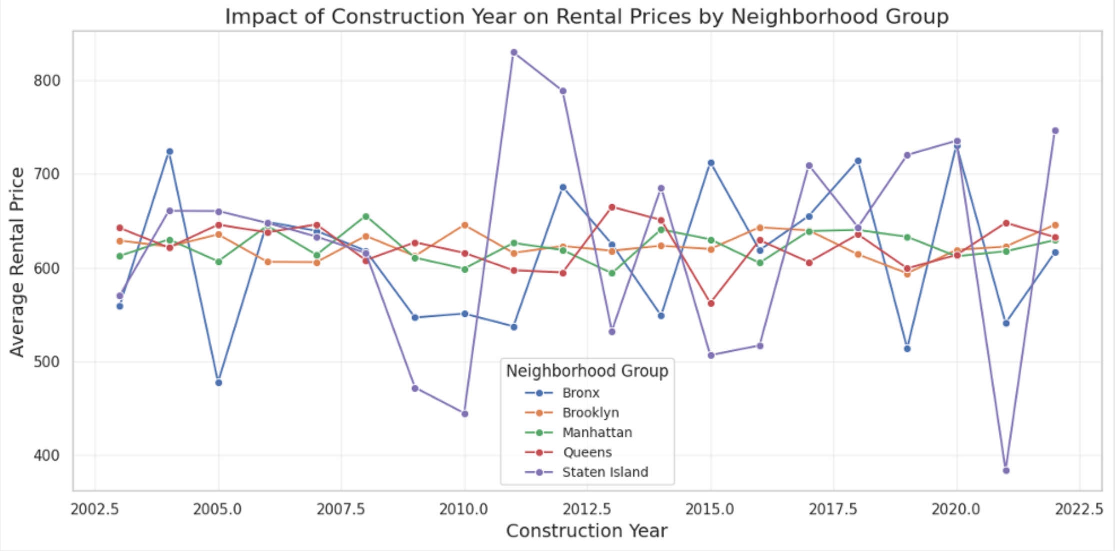
NEIGHBORHOOD GROUP

- Overall occupancy sits at around 60%, but very significant variation in occupancy rate both between neighborhoods and room types
- Staten Island has both the least occupied room type (Private) and the most occupied (Shared)



CONSTRUCTION YEAR

- Staten Island and the Bronx show the largest variance in average price year to year
- Possible small positive trend



EXPLORATORY DATA ANALYSIS: KEY INSIGHTS

Verification Impact: Host verification has minimal influence on prices or review scores, showing near-identical distributions for verified and unverified hosts.

Staten Island Trends: Staten Island stands out with the highest shared-room occupancy, the lowest private-room occupancy, and volatile year-to-year rental prices.

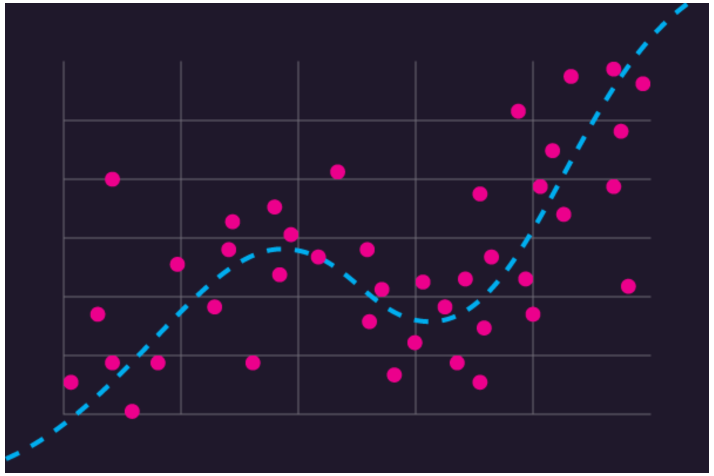
Room Type & Prices: Shared rooms in Manhattan are the most expensive across all boroughs, reflecting the city's high real estate costs.

Occupancy Variance: Occupancy rates vary significantly by borough and room type, with Staten Island’s shared rooms having the highest rates.

MODELS

Research goal: run different regression models to help learn about which factors contribute to price and review scores.

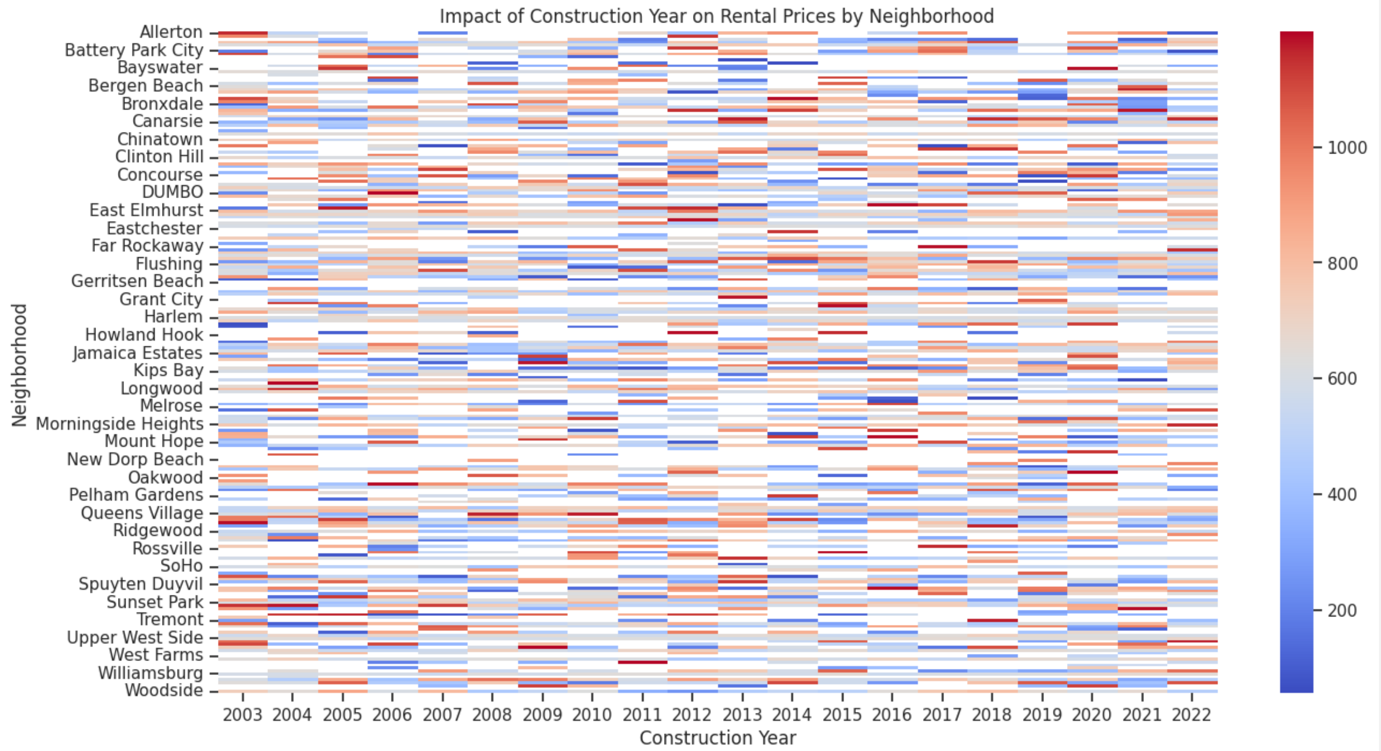
- Linear Regression: Construction year's impact on price
- Random Forest Regression: Characteristics of high review scores
- Point Biserial and Logistic Regression: Host verification impact on review score
- Linear Regression: Age of listing impact on review scores and occupancy rates
- Random Forest Regression: Price Prediction Model

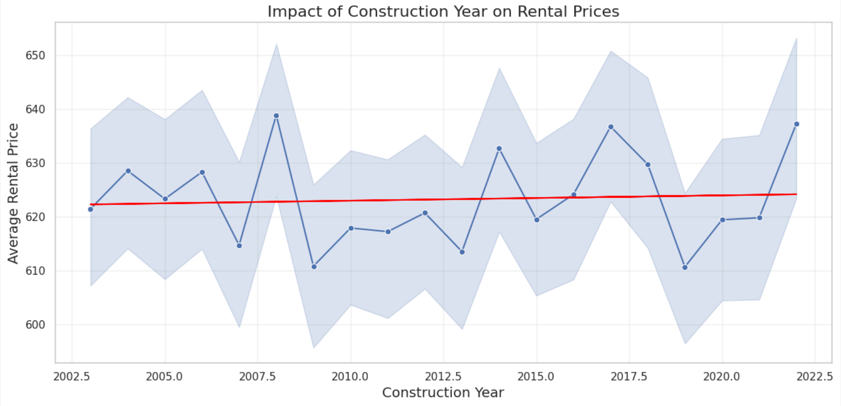


Source: https://www.imsl.com/blog/what-is-regression-model

PREDICTIVE MODEL 1: LINEAR REGRESSION PRICE VS. CONSTRUCTION YEAR

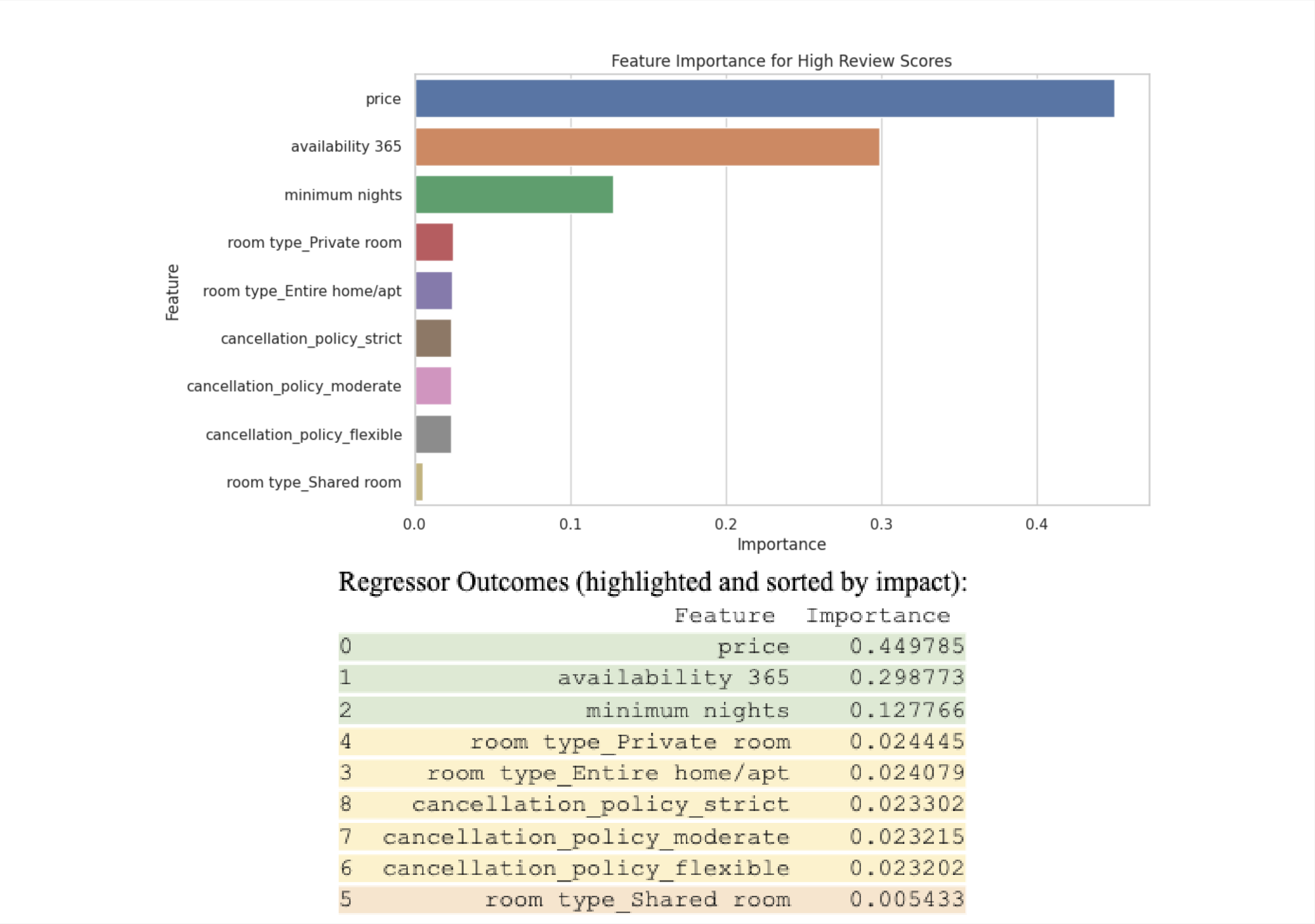
- Regression Coefficient: 0.0987 (Impact of Construction Year - small)
- Intercept: 424.63
- Visuals support findings: very small positive correlation





PREDICTIVE MODEL 2: RANDOM FOREST REGRESSION REVIEW SCORES

- Price, availability, and minimum nights have the most impact on high review scores
- When searching for places to stay, look based on price and availability
- Linear regression on price and review score: -4.59e-05 regression coefficient
  - Very small negative correlation



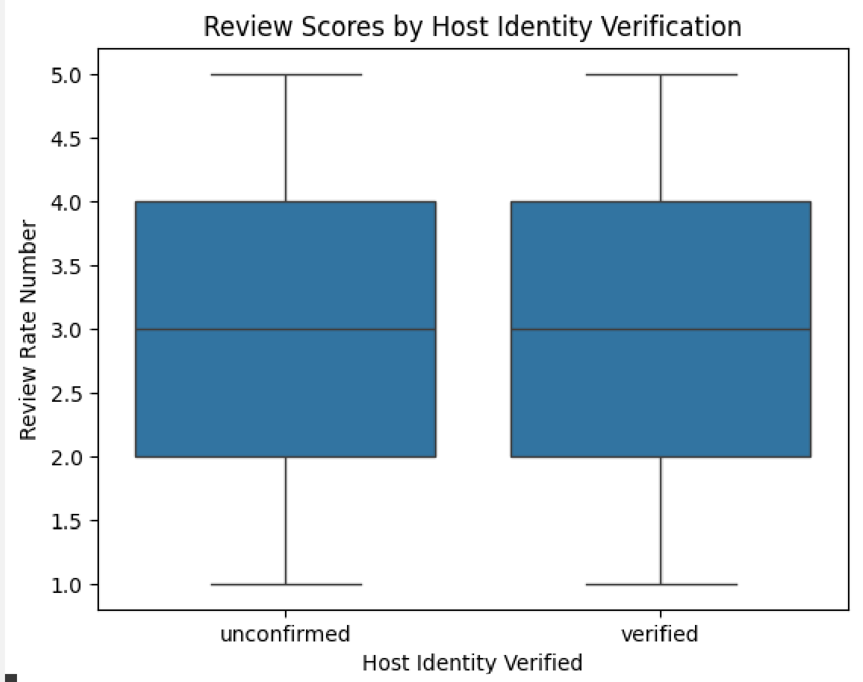
Regressor Outcomes (highlighted and sorted by impact):

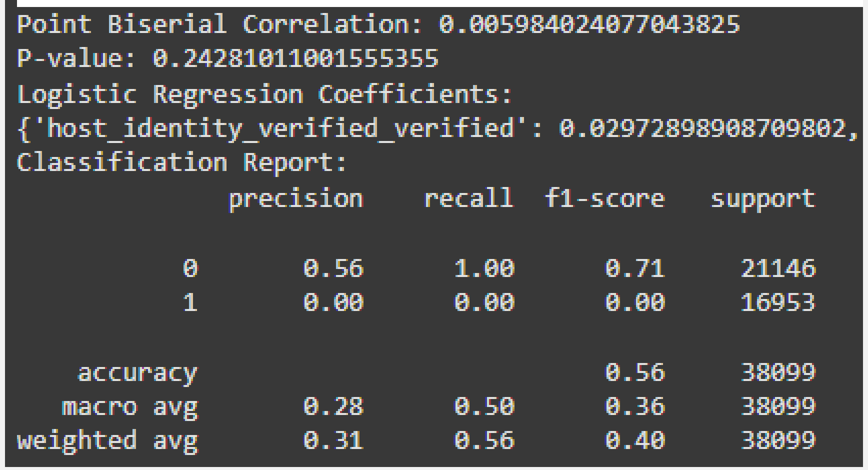
```
                        Feature  Importance
0                         price    0.449785
1              availability 365    0.298773
2                minimum nights    0.127766
4        room type_Private room    0.024445
3     room type_Entire home/apt    0.024079
8    cancellation_policy_strict    0.023302
7  cancellation_policy_moderate    0.023215
6  cancellation_policy_flexible    0.023202
5         room type_Shared room    0.005433
```

CORRELATION : HOST IDENTITY AND REVIEW SCORE
POINT BISERIAL

•**Point Biserial Correlation**: The correlation between host identity verification and review scores is weak and insignificant.
Correlation: 0.00598
P-value: 0.2428

•**Logistic Regression**: `host_identity_verified` has a negligible positive coefficient. The model performs poorly, likely due to weak relationships between features and the target and imbalanced data

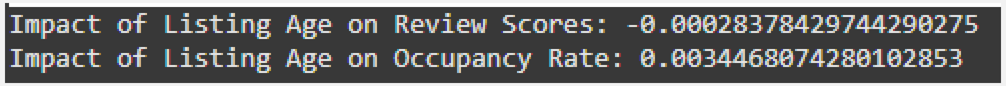




```
Point Biserial Correlation: 0.005984024077043825
P-value: 0.24281011001555355
Logistic Regression Coefficients:
{'host_identity_verified_verified': 0.029728989087098 02,
Classification Report:
              precision    recall  f1-score   support

           0       0.56      1.00      0.71     21146
           1       0.00      0.00      0.00     16953

    accuracy                           0.56     38099
   macro avg       0.28      0.50      0.36     38099
weighted avg       0.31      0.56      0.40     38099
```

AGE OF THE LISTING VS REVIEW SCORES

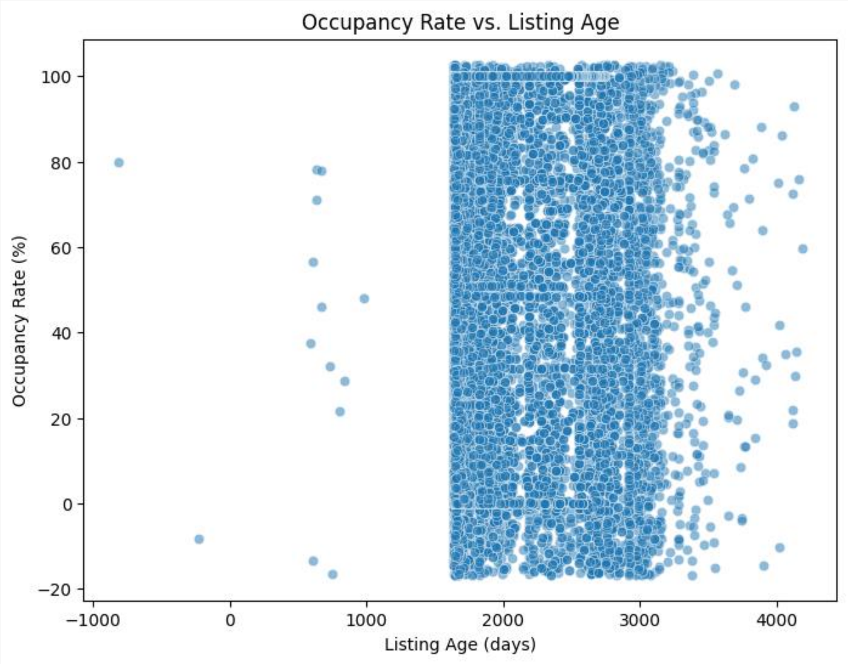
LINEAR REGRESSION 2

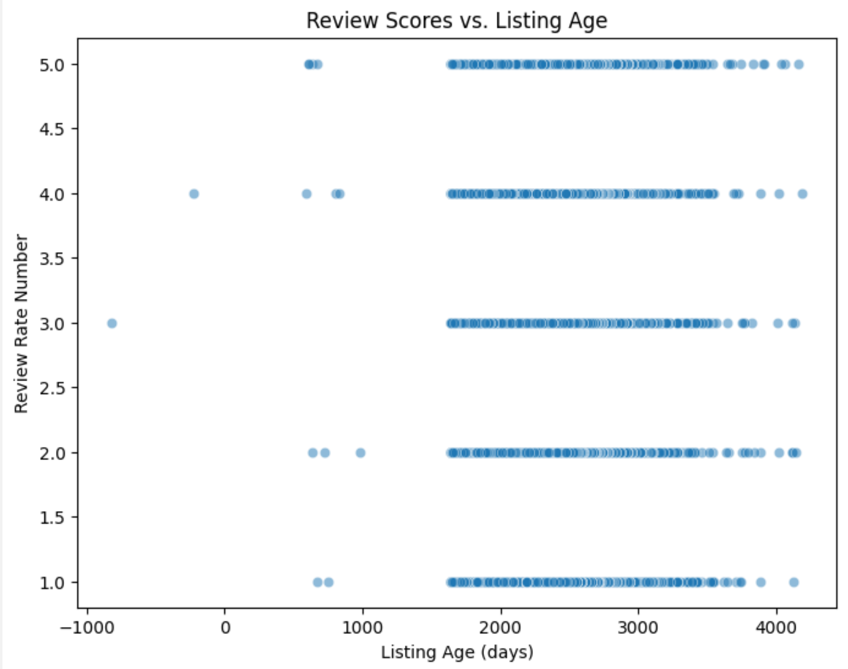


```
Impact of Listing Age on Review Scores: -0.00028378429744290275
Impact of Listing Age on Occupancy Rate: 0.0034468074280102853
```

- The negative impact of listing age on review scores suggests that older listings tend to have slightly lower review scores.
- The positive impact of listing age on occupancy rates suggests that older listings are somewhat more likely to have higher occupancy rates, but the effect is very small.

- Linear regression is useful for quantifying the relationship between listing age and review scores or occupancy rate, allowing for clear insights into how these factors are influenced by the age of the listing


Occupancy Rate vs. Listing Age


Review Scores vs. Listing Age

PRICE PREDICTION MODEL

- Our objective is to create a model predict the **price** of a listing based on various features such as review counts, availability, construction year, and others. It's an application of **supervised machine learning** where the model learns from historical data to predict the price based on input features. The dataset includes various categorical and numerical features, which are processed and used to train the model
- Our model learns from the data provided and after predicting an estimated price, classifies the listing as over-price, Good-deal or bargain. This helps in real world implementation and visualization of our output.

MODEL WE USE

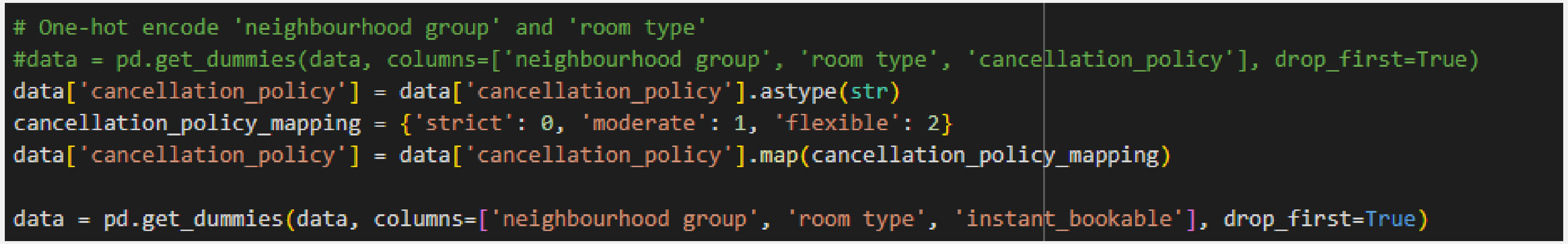
The **Random Forest Regressor** is the machine learning model used in this code

- **Handling Non-linear Relationships:** Random Forests are powerful when it comes to handling complex, non-linear relationships between features and the target variable (price). In this dataset, price is likely influenced by several factors in non-linear ways (e.g., a large number of reviews might affect the price differently depending on the room type).
- **Feature Importance Evaluation:** Random Forest can also help identify the importance of different features (e.g., days since last review, number of reviews) in predicting the target variable.
- **Robust to Overfitting:** Random Forest models reduce the risk of overfitting by averaging the results of multiple decision trees, making them more stable and accurate.
- **Handles Mixed Data Types:** Random Forest works well with both numerical and categorical data, which is present in this datase

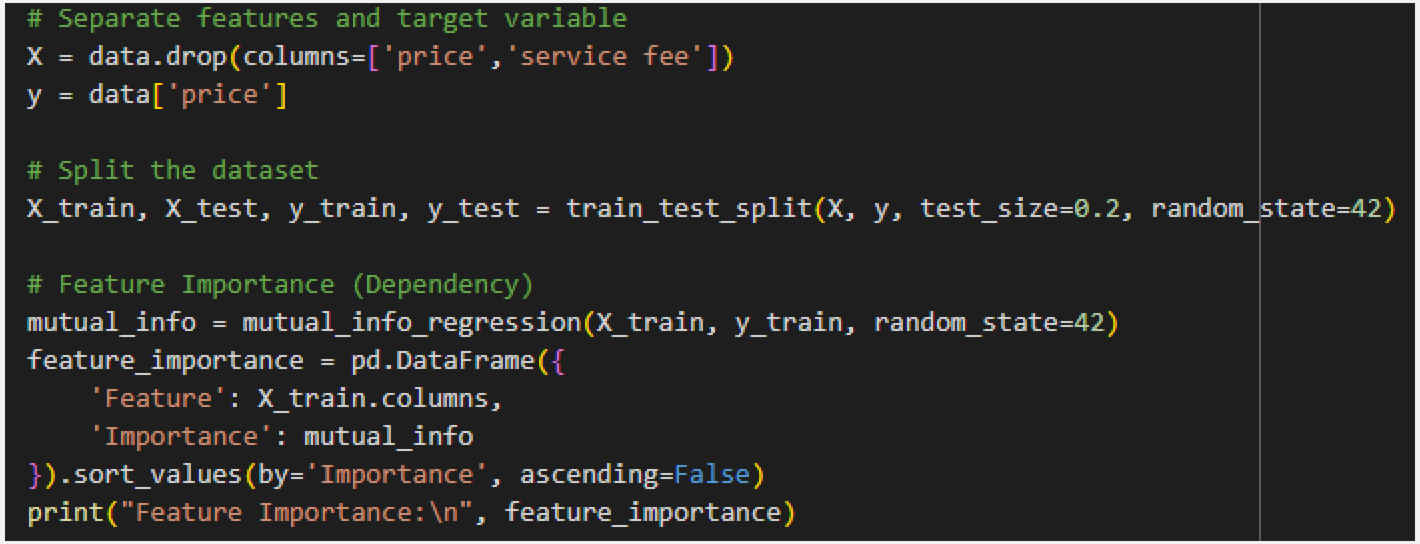
STEPS TO CREATE
PRICE PREDICTION MODEL

1. Import Dataset
2. Clean Dataset
3. One-Hot Encode Columns
4. Feature Extraction
5. Train Model
6. Predict and classify prices

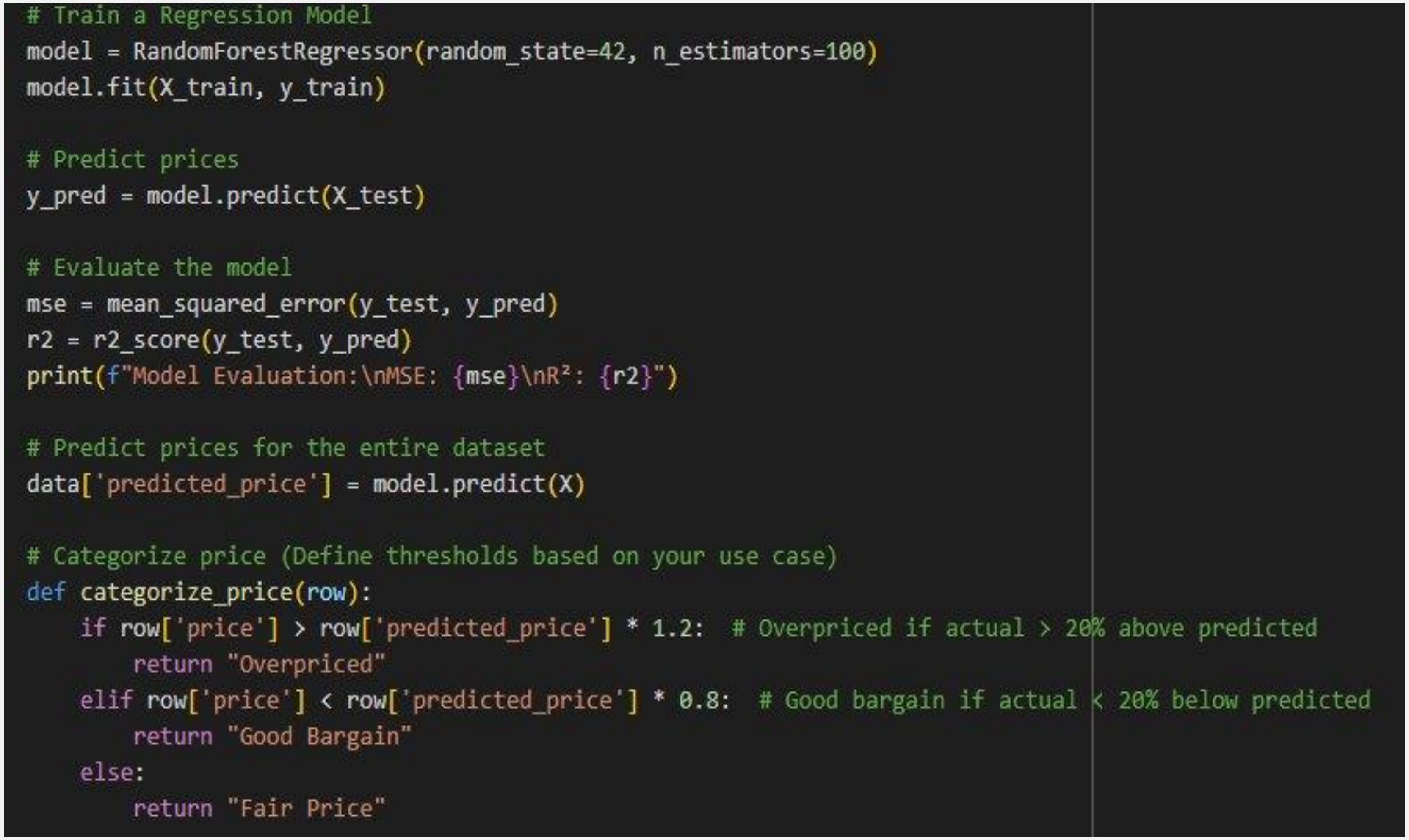
CODE SNIPPETS



```
# One-hot encode 'neighbourhood group' and 'room type'
#data = pd.get_dummies(data, columns=['neighbourhood group', 'room type', 'cancellation_policy'], drop_first=True)
data['cancellation_policy'] = data['cancellation_policy'].astype(str)
cancellation_policy_mapping = {'strict': 0, 'moderate': 1, 'flexible': 2}
data['cancellation_policy'] = data['cancellation_policy'].map(cancellation_policy_mapping)

data = pd.get_dummies(data, columns=['neighbourhood group', 'room type', 'instant_bookable'], drop_first=True)
```



```
# Separate features and target variable
X = data.drop(columns=['price','service fee'])
y = data['price']

# Split the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Feature Importance (Dependency)
mutual_info = mutual_info_regression(X_train, y_train, random_state=42)
feature_importance = pd.DataFrame({
    'Feature': X_train.columns,
    'Importance': mutual_info
}).sort_values(by='Importance', ascending=False)
print("Feature Importance:\n", feature_importance)
```



```python
# Train a Regression Model
model = RandomForestRegressor(random_state=42, n_estimators=100)
model.fit(X_train, y_train)

# Predict prices
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Model Evaluation:\nMSE: {mse}\nR²: {r2}")

# Predict prices for the entire dataset
data['predicted_price'] = model.predict(X)

# Categorize price (Define thresholds based on your use case)
def categorize_price(row):
    if row['price'] > row['predicted_price'] * 1.2:  # Overpriced if actual > 20% above predicted
        return "Overpriced"
    elif row['price'] < row['predicted_price'] * 0.8:  # Good bargain if actual < 20% below predicted
        return "Good Bargain"
    else:
        return "Fair Price"
```

FEATURE IMPORTANCE: PRICE

- **`days_since_last_review`** (0.400280) — The number of days since the last review is the most important feature.
- **`reviews per month`** (0.380897) — The number of reviews per month also has significant predictive power.
- **`availability 365`** (0.265864) — The listing's availability throughout the year plays an important role.
- **`number of reviews`** (0.259994) — The total number of reviews is also impactful.
- **`Construction year`** (0.129753) — The year the property was built shows some influence.
- Other features with relatively lower importance include `review_month`, `minimum nights`, `review_rate_number`, and several neighborhood and room type indicators

MODEL EVALUATION (SAMPLE PREDICTIONS):

- **Price:** 966.0 → **Predicted price:** 795.13 → **Category:** Overpriced
- **Price:** 142.0 → **Predicted price:** 319.44 → **Category:** Good Bargain
- **Price:** 368.0 → **Predicted price:** 456.19 → **Category:** Fair Price
- **Price:** 204.0 → **Predicted price:** 358.39 → **Category:** Good Bargain
- **Price:** 577.0 → **Predicted price:** 599.20 → **Category:** Fair Price

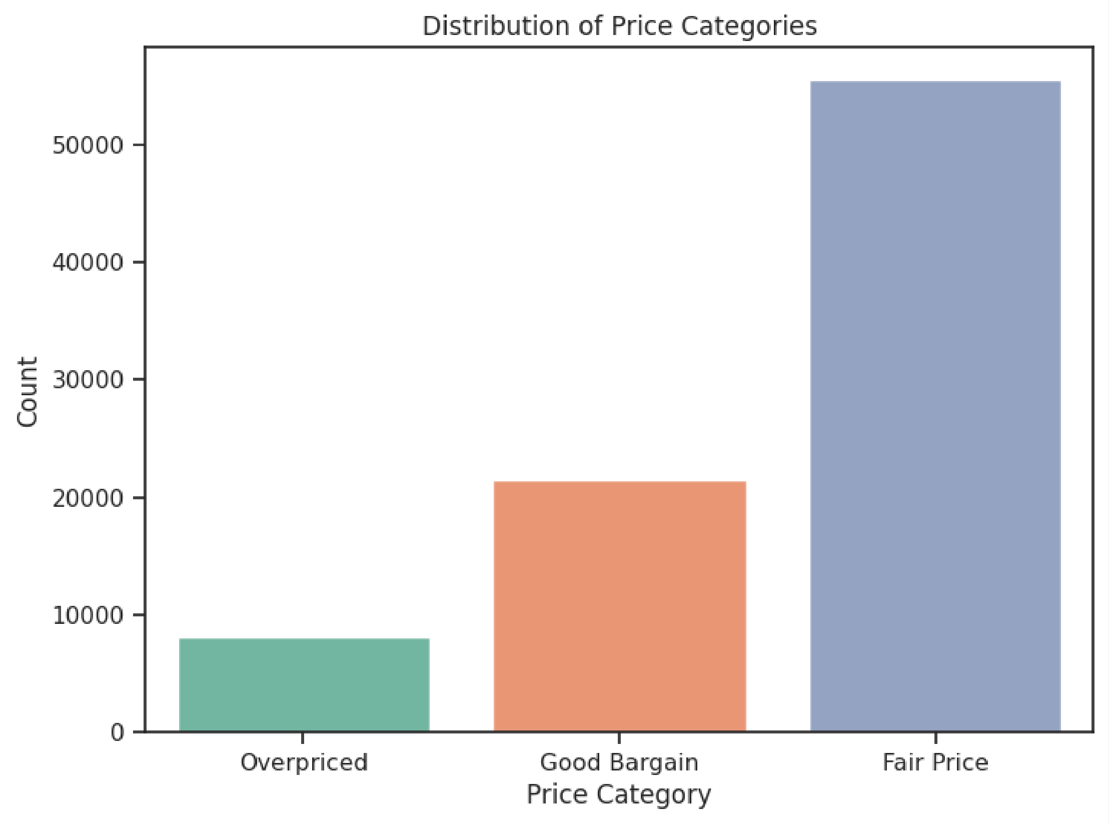
MODEL EVALUATION:

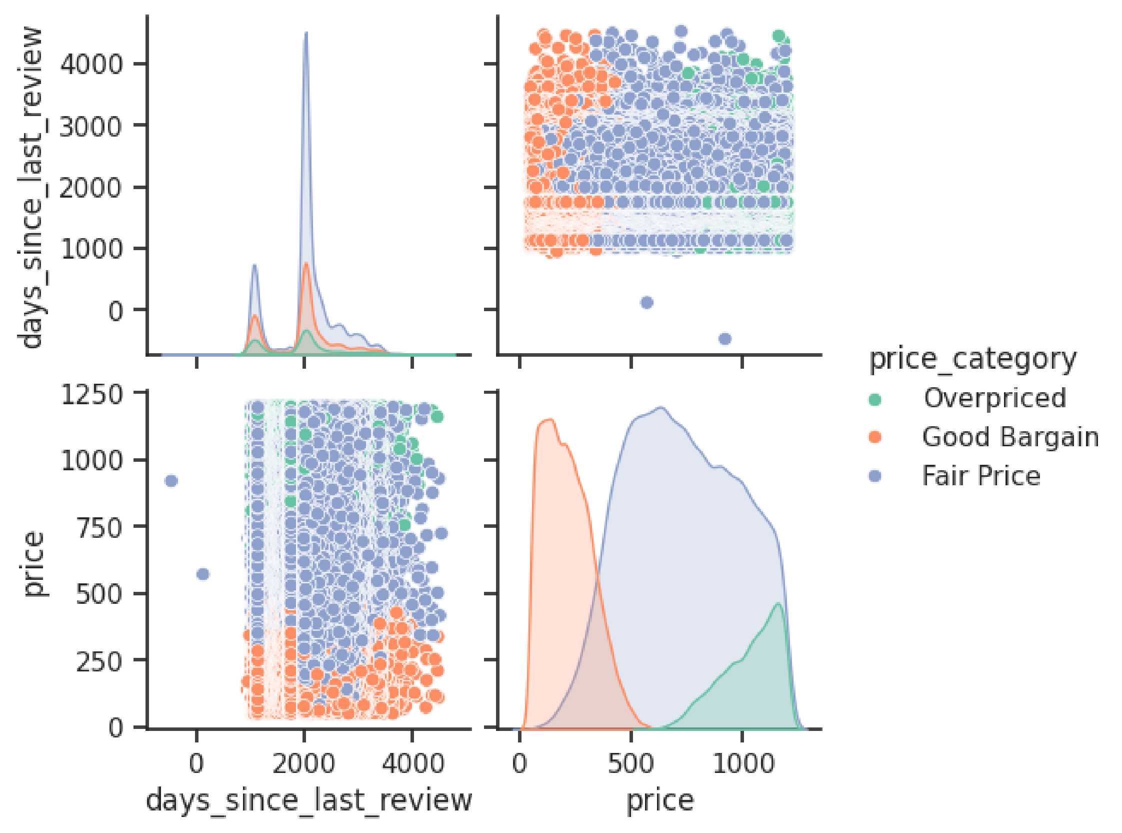
- **Mean Absolute Error (MAE):** 222.8323 — This indicates the average difference between the predicted and actual prices. A smaller MAE would suggest more accurate predictions.
- **Mean Squared Error (MSE):** 75850.9554 — MSE penalizes larger errors more heavily than MAE due to squaring the errors. This can be useful when large errors need to be minimized.
- **Root Mean Squared Error (RMSE):** 275.4105 — RMSE gives you the error in the same units as the target variable (price), which is more interpretable.
- **R-squared (R^2):** 0.3129 — This means that the model explains approximately 31.3% of the variance in the price data, which could be considered low, indicating there is room for improvement in the model.

FUTURE ENHANCEMENTS

- **R^2 (0.3129):** This value suggests that the model's performance is suboptimal, as it only explains around 31% of the variance in the target variable (price). A higher R^2 value (closer to 1) would indicate a better fit.
- **MAE, MSE, RMSE:** These metrics suggest that while the model is making reasonable predictions, there is still significant room for improvement, especially considering that the MAE and RMSE are relatively high.
- **Improvement Areas:**
- More features could be engineered or selected.
- Hyperparameter tuning could help improve model performance.
- Trying more complex models, like Gradient Boosting Machines (GBM) or XGBoost, may yield better results.

OTHER CONCLUSIONS FROM THE MODEL





OVERALL CONCLUSIONS – PREDICTION MODELS

Linear Regression

- **Best For**: Simple, linear relationships; small datasets.
- **Strengths**: Fast, interpretable, computationally efficient.
- **Limitations**: Struggles with complex or non-linear data; sensitive to outliers.

Random Forest

- **Best For**: Complex, non-linear data; large datasets.
- **Strengths**: Robust to overfitting, handles feature interactions well, high accuracy.
- **Limitations**: Less interpretable, computationally expensive.

Biserial Correlation

- **Best For**: Analyzing relationships between continuous and binary variables.
- **Strengths**: Measures strength of association.
- **Limitations**: Not used for predictive modeling; limited to binary categorical variables.

Best Choice: **Random Forest**

- Highest predictive accuracy for complex and non-linear datasets.

OVERALL CONCLUSIONS – DATASET

Price Distribution Insights:

- The distribution of prices is highly skewed, with a few expensive listings driving the higher end of the spectrum. This suggests that most listings are reasonably priced, but there are some outliers that need special attention (e.g., luxury or unique properties).

Feature Importance:

- From the feature importance analysis, `reviews per month` and `days since last review` are critical features influencing the model's predictions. Listings with frequent reviews and more recent reviews tend to have higher prices.
- `availability 365` is also a significant factor, suggesting that properties with higher availability throughout the year generally command higher prices.

Correlation Between Features:

- There is a strong correlation between `reviews per month` and `number of reviews`, indicating that properties with more reviews are typically more active and possibly better-rated, contributing to a higher price.
- `cancellation policy` and `host_identity_verified` show a weaker correlation to price, indicating they might not be as significant predictors in this case.

Price vs. Location:

- Based on neighborhood group data, **Brooklyn** and **Manhattan** have higher average prices compared to other boroughs, as seen in the graphs.

SOURCES

- Azmoudeh, A. (2022, August 1). *Airbnb Open Data*. Kaggle. https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata
- *Example gallery*. Seaborn. (n.d.). https://seaborn.pydata.org/examples/index.html
- *Example gallery*. Matplotlib. (n.d.). https://matplotlib.org/stable/gallery/index.html

IMAGE LINKS

- https://www.loumovesyou.com/blog/boroughs-of-new-york/
- https://www.imsl.com/blog/what-is-regression-model