Group 1 Capstone Draft Report - Airbnb NYC Open Data

ALY 6140

Presented to: Prof. Richard He

Prepared by: Shivam Sinha

December 10th, 2024

Introduction

We explore New York City Airbnb data to draw conclusions about housing prices and review scores, among other things. Although none of us live in New York City, we thought that analyzing Airbnb and building models to draw conclusions about it would be quite interesting and potentially have results that hold value in cities across the country.

Through our analysis, our main goal was to run different regression models to help learn about which factors contribute to price and review scores. We used a variety of different regression models including linear regression, random forest regression, logistic regression, as well as point biserial correlation. Along with our main research question of learning about what drives high prices and review scores, we answered a few questions to help guide us throughout the project, such as construction year's impact on rental prices across different neighborhoods, as well as whether the host identity being verified is associated with higher review scores; the full list of questions can be found below in our exploratory data analysis section.

Exploratory Data Analysis

Analysis Description and Data Extraction

Our data can be downloaded on Kaggle using the link in our sources section below. Originally uploaded from the Inside Airbnb website, our dataset contains approximately 100,000 total records and 16 fields that we can use following our data cleanup, such as price, review score, construction year, and neighborhood.

Our goal throughout the project is to determine the factors that contribute to Airbnb housing records with the highest reviews, as well as determine what fields drive price. In learning about these different categories, we can determine what to focus on should we ever rent a space on Airbnb while in New York City or elsewhere. Listed below are our guiding questions, answered in our exploratory data analysis and predictive models sections.

Guiding Questions

- How does the year of construction impact rental prices across different neighborhoods?
- Which listing characteristics within each neighborhood are associated with high review scores?
- Is there a correlation between the host identity being verified and the review score?
- Which room types have the highest occupancy rate, and in which neighborhoods?
- How does the age of the listing (how long the property has been on Airbnb) affect review scores or occupancy rates?

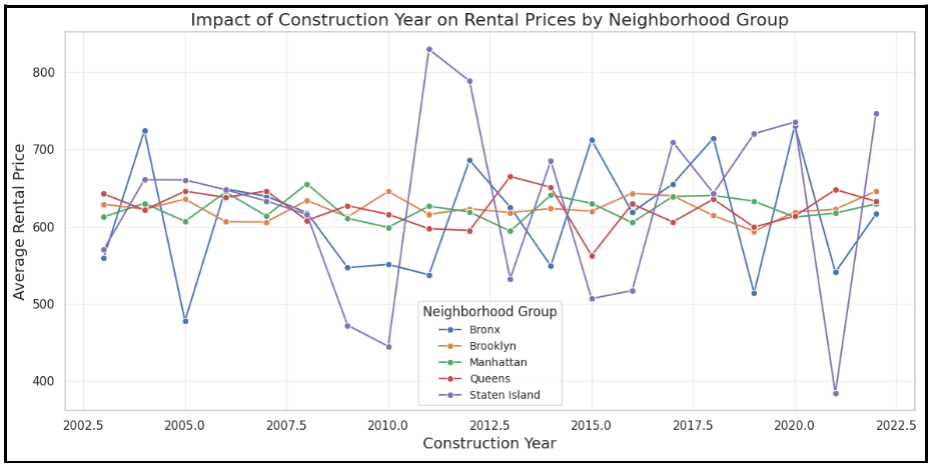
Data Cleanup

In our data cleanup, we had to modify the data in many different ways to make it usable for our goals. The first thing that we determined we needed to do was to drop different columns that we were not interested in and were providing bloat to the dataset. Next, we noticed that we should set proper types for numeric fields, as well as assign categorical variables to proper

categories in the Pandas DataFrame. We continued by removing some of the data that contained missing values in relevant columns, as well as filling in our reviews per month column to contain a zero value instead of a null value when such occurred. Lastly, we created new columns, such as occupancy rate, and cleaned our string price values into numerical data, after slicing away the dollar sign. Overall, at this point, we felt confident in our data and decided to move forward with the exploratory data analysis.

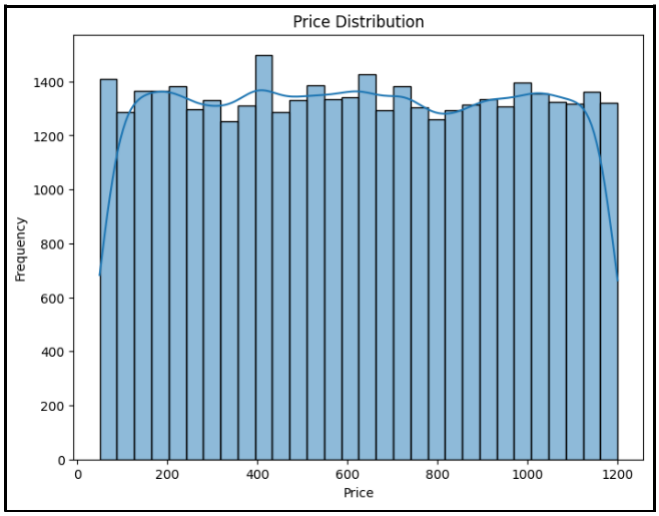
Data Visualization

Construction year

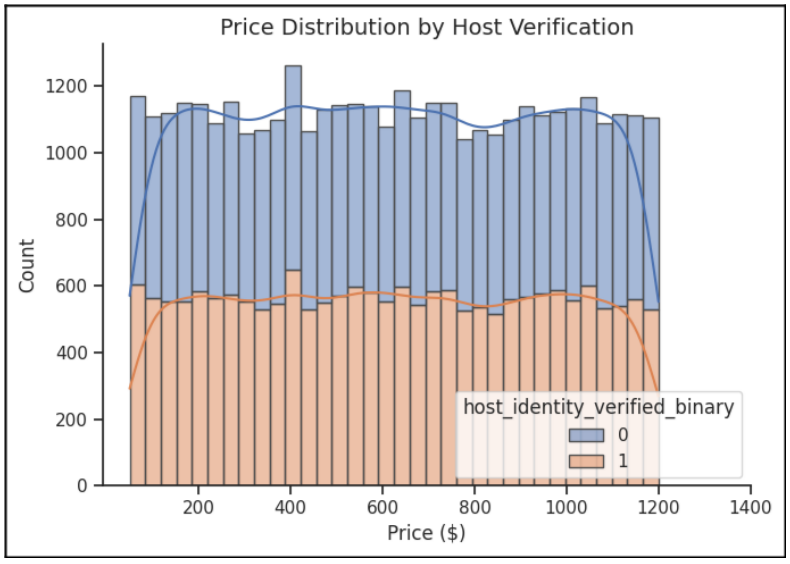


Overall, there seems to be a slight positive trend, with the construction year not playing a significant role in rental price. We noted that the construction years following the 2008 financial crisis and 2020 pandemic yielded lower rental prices in Staten Island and the Bronx, but given that it didn't directly relate to our research question, we decided not to pursue it further.

Price

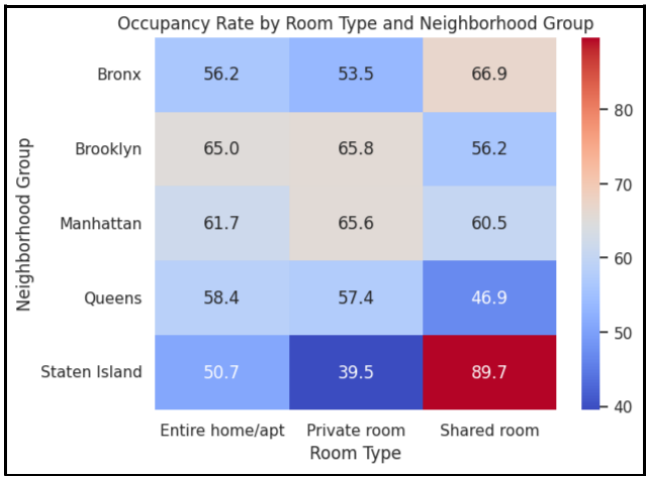


This visualization shows that prices in our dataset are uniformly distributed; this surprised us, but tells us that our dataset doesn't have any large outliers.



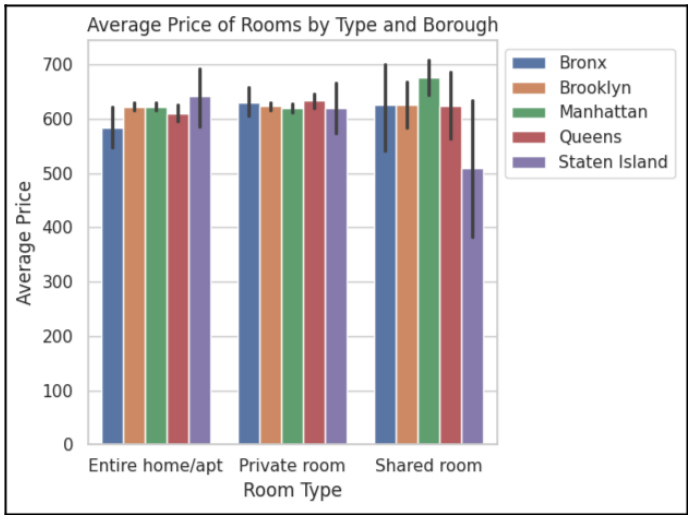
As can be seen here, the host identity also does not seem to play a significant role in prices, given the data is uniformly distributed. Initially, one might have thought that Airbnb homes with higher prices would have more verified hosts, but this chart shows that is not the case.

Neighborhood group



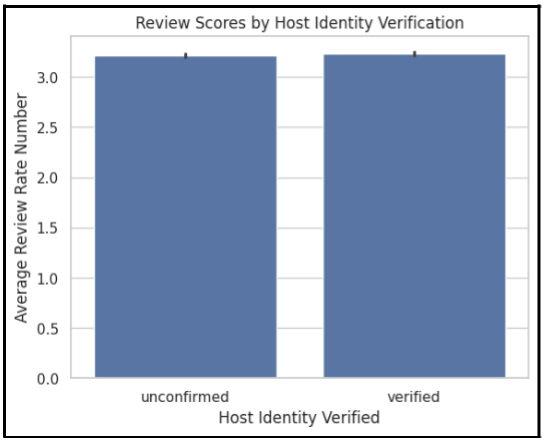
As can be seen from the heatmap, shared rooms are quite occupied on Staten Island, while private rooms on Staten Island are only occupied less than 40% of the time. Besides Staten Island, occupancy rate by room type for the other boroughs hovers around 60% with Queens rooms a little bit lower.

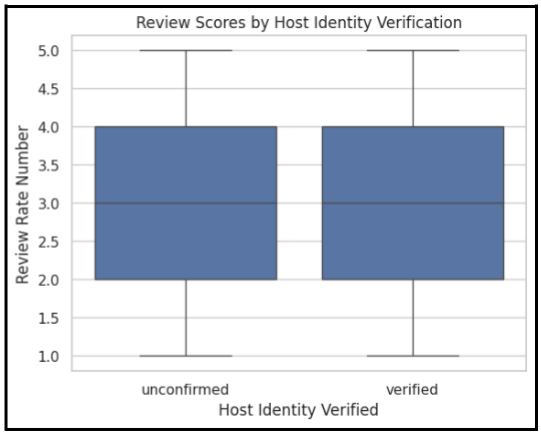
Room Type



The graph shows that the lowest cost is for a shared room on Staten Island, which may be why they are so frequently occupied. The most expensive average prices come from shared rooms in Manhattan, which makes sense, as Manhattan real estate prices are known as some of the highest in all of the United States. In our price model below, keeping an eye on shared rooms on Staten Island for potential bargain prices will be important.

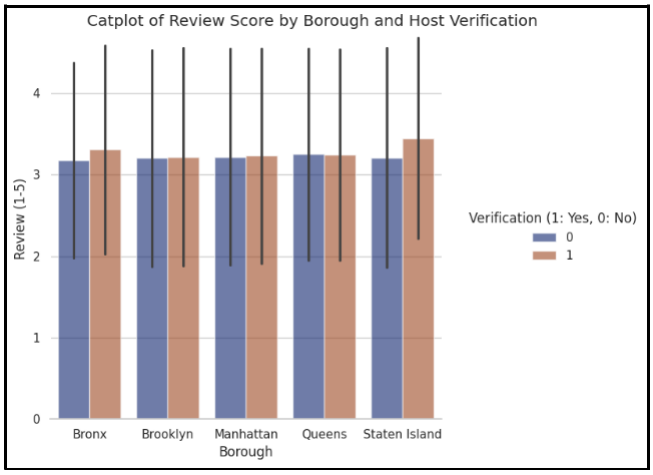
Host Verified





The two graphs above both show that review scores are not significantly different based on host verification status. These results will be examined further below, by drawing in the different associated boroughs.

Review Score



No borough yields significantly higher reviews than the others, as they all hover a little above 3. Also, the verification of Airbnb owners does not seem to play a significant part for borough review scores, besides the Bronx and Staten Island: in those two boroughs, hosts who were verified on average yielded a slightly higher review.

Exploratory Data Analysis Conclusion

Through the use of many different graphs sliced by different dimensions of data, we were not able to see any clear factors that contributed to higher prices or reviews. One smaller takeaway that we saw was that for the average price of shared rooms by borough, prices in Manhattan were the highest, and Staten Island were the lowest. We will look towards the descriptive and predictive models below to aid in determining what factors contribute to the highest review scores, as well as attempt to predict the price of each apartment and determine whether or not the price is accurate.

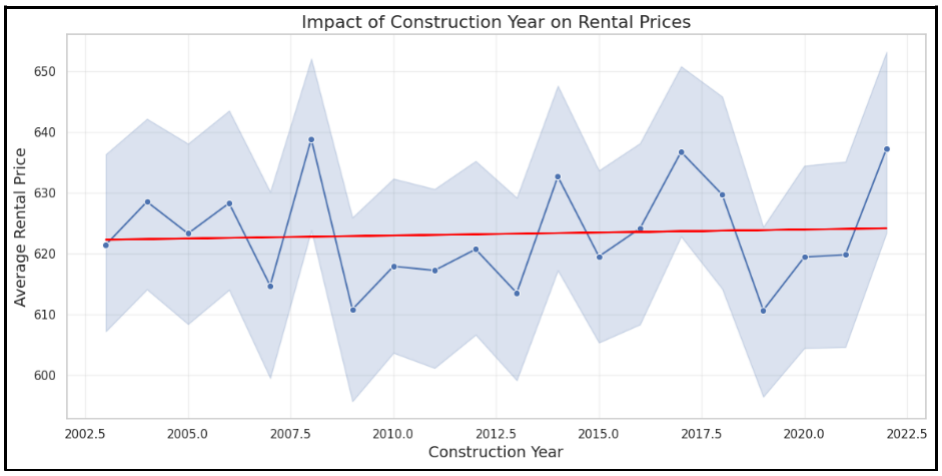
Predictive Models

Linear Regression 1: Price and Construction Year

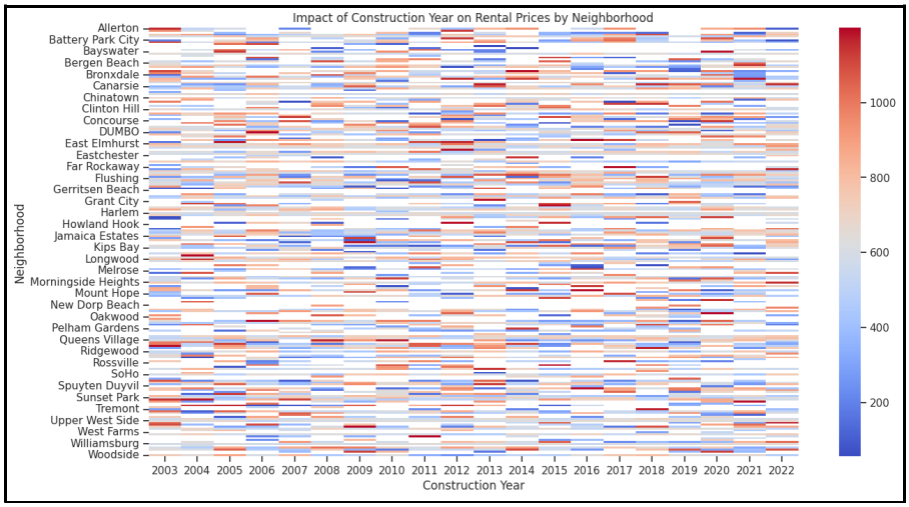
Regression Coefficient: 0.0987 (Impact of Construction Year - small)

Intercept: 424.63

R2: 2.9e-06 (very weak)



The visualization above is meant to act as a support visual to the findings of our linear regression model, which states that there is a very weak positive relationship between construction year and price of rental. The R-squared value is near zero, meaning our model is not responsible for explaining the variation in the dependent variable.



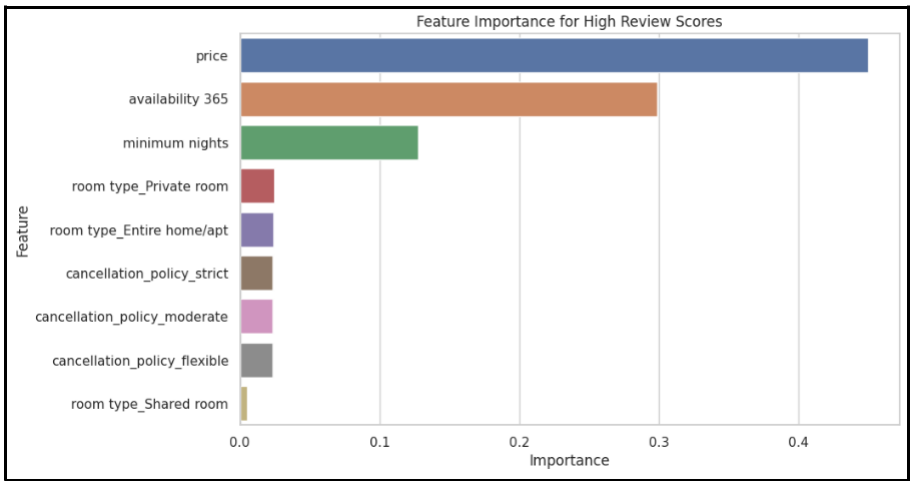
The heatmap here also visualizes what our linear regression results are showing, that there is little to no positive correlation between construction year and rental price.

Random Forest Regression: Importance for High review Scores

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

Regressor Outcomes Visualized



As can be seen from the visualization and tabular data above, the price, availability, and minimum nights of the Airbnb property are by far the most impactful fields when predicting high review scores. Therefore, we as guests know that when finding a place to stay on Airbnb, prioritizing the price and minimum nights will potentially yield the greatest satisfaction.

Point Biserial Correlation: Review Scores by Host Identity Verification

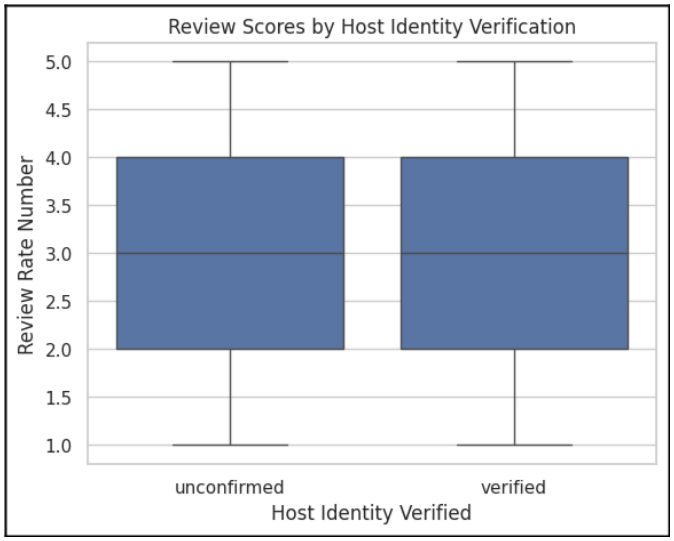
```
Point Biserial Correlation: 0.0061
```

This indicates a very weak positive correlation between the host identity being verified and the review rate number. The correlation suggests that there is almost no linear relationship between whether the host identity is verified and the review score.

```
P-value: 0.2344
```

Our null hypothesis states that there is no correlation between the host identity being verified and the review rate. Our alternative hypothesis, H1, is that there is a correlation between the review rate and the host identity being verified.

The p-value of 0.23 is larger than the typical alpha value of 0.05, which tells us that the correlation is not statistically significant, and we fail to reject the null hypothesis. These results can be depicted in the visual below.



Logistic Regression: Review Scores by Host Identity Verification

```
Logistic Regression Coefficients:
{'host_identity_verified_verified': 0.030112869007587053,
'room type_Private room': 0.005551321304304382,
'room type_Shared room': -0.029216627674819765,
'neighborhood group_Brooklyn': 0.03564743329095475,
'neighborhood group_Manhattan': 0.040324648668406536,
'neighborhood group_Queens': 0.03944581471128395,
'neighborhood group_Staten Island': 0.1179921383201146}
```

This logistic regression model tries to predict whether a review is high (1) or low (0) based on the various features. As can be seen in the coefficients above, the coefficient for the host identity being verified is 0.030, meaning that the odds of receiving a high review very slightly increase when the host identity is verified. It is interesting to note that the coefficient for Staten Island is far larger than other boroughs.

Classification Report(prediction):

	precision	recall	f1-score	support
0	0.56	1.00	0.71	21142
1	0.00	0.00	0.00	16938
accuracy			0.56	38080
macro avg	0.28	0.50	0.36	38080
weighted avg	0.31	0.56	0.40	38080

Class 0 - Low Review Score

Precision: 0.56: Among all instances predicted as a low review score, 56% were actually low review scores.

Recall: 1.00: The model identified 100% of actual low review scores correctly.

F1-score: 0.71: A balanced score that takes both precision and recall into account for class 0.

Class 1 - High Review Score

Precision: 0.00: The model predicted 0% of the actual high review scores as high.

Recall: 0.00: The model failed to identify any high review scores (all high review scores were missed).

F1-score: 0.00: The model's performance for predicting high review scores is extremely poor, as indicated by the F1-score of 0.00.

Overall F1-scores

Accuracy: 56%: This is the percentage of correct predictions, but it is not a reliable metric here because of the class imbalance (much more data on low review scores than high review scores).

Macro avg: The average performance across both classes, which is 0.28 for precision, 0.50 for recall, and 0.36 for F1.

Weighted avg: This takes into account the size of each class, and you see similar performance with 0.31 precision, 0.56 recall, and 0.40 F1.

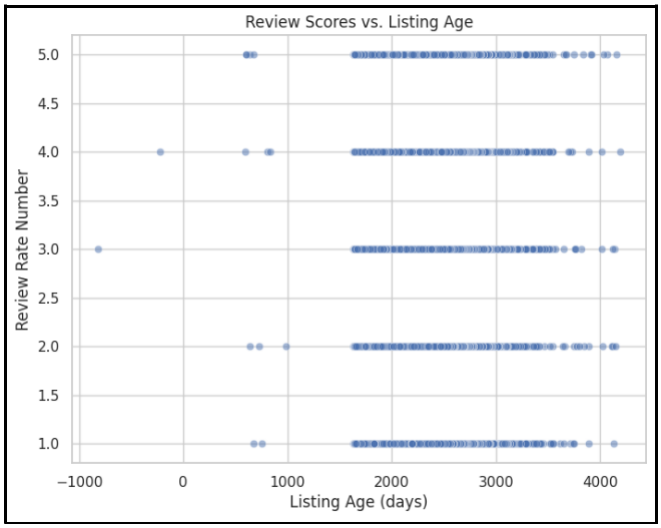
Key Takeaways

Logistic Regression has a poor ability to predict high vs. low review scores. While it performs well at predicting low review scores (class 0), it fails to predict high review scores (class 1). This is likely due to an imbalance in the classes (more low scores than high).

Linear Regression 2: Review Scores vs. Listing Age

```
Impact of Listing Age on Review Scores: -0.00028
```

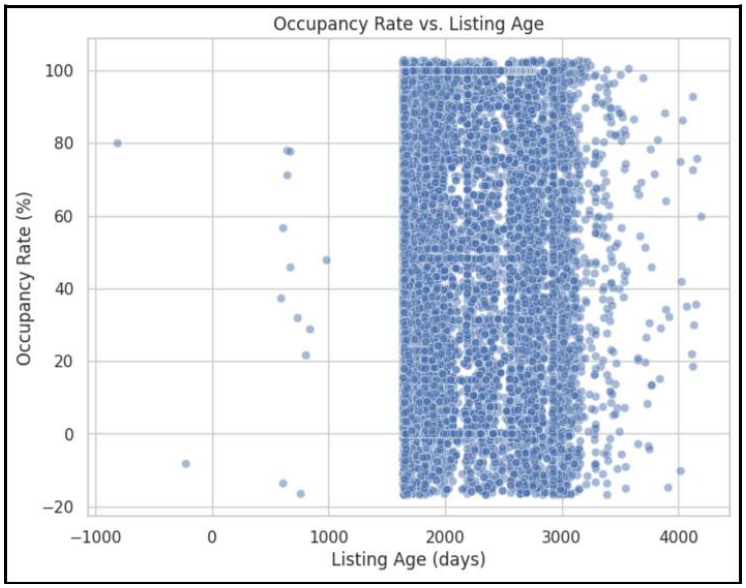
This linear regression coefficient tells us that there is a very small negative correlation between the listing age and review rate. The graph below confirms that, as we can't see any meaningful correlation between the listing age and the review score.



Linear Regression 3: Occupancy Rate vs. Listing Age

```
Impact of Listing Age on Occupancy Rate: 0.0034
```

Similar to the results from above, the listing age does not have a statistically significant relationship with the occupancy rate. We can tell this based on our low positive correlation of 0.0034 found in our linear regression, as well as the results of our graph below.



Final Model: Price Prediction

*Introduction***:**

This model predicts and classifies prices by analyzing historical and current data, providing insights into how much a particular price deviates from the market average. It takes into account various features such as product attributes, market conditions, and pricing trends to generate a prediction for a target price. By comparing the predicted price with actual market prices, the model enables users to assess whether they are paying above or below market value, helping them make more informed purchasing or investment decisions.

The model incorporates three main components: data cleaning, linear regression, and random forest modeling, allowing for a robust analysis and evaluation of price prediction methods.

Data Cleaning:

The clean_data function handles essential preprocessing steps:

- Removal of missing or inconsistent values
- Encoding categorical variables (e.g., one-hot encoding)
- Normalization of numerical features (e.g., Min-Max scaling or Z-score normalization)

This ensures the dataset is optimized for machine learning models.

```
/usr/local/lib/python3.10/dist-packages/dask/dataframe/__init__.py:42: FutureWarning:
Dask dataframe query planning is disabled because dask-expr is not installed.

You can install it with `pip install dask[dataframe]` or `conda install dask`.
This will raise in a future version.

  warnings.warn(msg, FutureWarning)
Non-numeric columns: []
Feature Importance:
                              Feature  Importance
20             days_since_last_review    0.400284
5                   reviews per month    0.380897
7                    availability 365    0.265864
4                   number of reviews    0.259994
2                   Construction year    0.129753
19                       review_month    0.069062
3                      minimum nights    0.063357
6                  review rate number    0.034114
18                        review_year    0.032630
8        neighbourhood group_Brooklyn    0.010526
16             room type_Shared room    0.008886
10         neighbourhood group_Queens    0.007815
11  neighbourhood group_Staten Island    0.006888
9       neighbourhood group_Manhattan    0.004385
15            room type_Private room    0.003123
12        neighbourhood group_brookln    0.000000
13       neighbourhood group_manhatan    0.000000
14              room type_Hotel room    0.000000
17             instant_bookable_True    0.000000
1                cancellation_policy    0.000000
0              host_identity_verified    0.000000
```

Linear Regression Model:

Linear regression serves as a baseline model, assuming a linear relationship between features and the target variable. Key metrics for evaluation include:

- MAE: 287.97
- MSE: 110470.03
- RMSE: 332.37
- R^2: -0.0006 (poor fit)

While interpretable, linear regression had limited success in predicting target prices.

```
Model Evaluation - LINEAR REGRESSION:
MSE: 110470.03440223567
R²: -0.0006302935524546971
   price  predicted_price price_category
0  966.0       622.024068     Overpriced
1  142.0       620.406313   Good Bargain
3  368.0       628.157557   Good Bargain
4  204.0       618.513977   Good Bargain
5  577.0       619.170021     Fair Price
Mean Absolute Error (MAE): 287.9657
Mean Squared Error (MSE): 110470.0344
Root Mean Squared Error (RMSE): 332.3703
R-squared (R²): -0.0006
```

Linear regression is not as suitable for this price prediction task because it assumes a linear relationship between the features and the target variable, which may not capture the complex, non-linear patterns present in real-world pricing data. The model's performance, as indicated by the low R^2 value, suggests it fails to adequately account for the variability and interactions within the data. In contrast, more advanced models like random forests can handle non-linearity and feature interactions, leading to more accurate and reliable predictions.

Random Forest Model:
Random forest utilizes an ensemble of decision trees to capture non-linear relationships. Evaluation metrics include:

- MAE: 222.83
- MSE: 75850.96
- RMSE: 275.41
- R^2: 0.3129 (moderate success)

Random forest outperformed linear regression, capturing more complex patterns.

```
/usr/local/lib/python3.10/dist-packages/dask/dataframe/__init__.py:42: FutureWarning:
Dask dataframe query planning is disabled because dask-expr is not installed.

You can install it with `pip install dask[dataframe]` or `conda install dask`.
This will raise in a future version.

  warnings.warn(msg, FutureWarning)
Non-numeric columns: []
Feature Importance:
                              Feature  Importance
20            days_since_last_review    0.400284
5                  reviews per month    0.380897
7                   availability 365    0.265864
4                  number of reviews    0.259994
2                  Construction year    0.129753
19                      review_month    0.069062
3                    minimum nights    0.063357
6                 review rate number    0.034114
18                       review_year    0.032630
8        neighbourhood group_Brooklyn    0.010526
16             room type_Shared room    0.008886
10         neighbourhood group_Queens    0.007815
11  neighbourhood group_Staten Island    0.006888
9       neighbourhood group_Manhattan    0.004385
15             room type_Private room    0.003123
12        neighbourhood group_brookln    0.000000
13       neighbourhood group_manhatan    0.000000
14               room type_Hotel room    0.000000
17             instant_bookable_True    0.000000
1                cancellation_policy    0.000000
0              host_identity_verified    0.000000
Model Evaluation:
MSE: 75850.95538064688
R²: 0.31294704342705637
   price  predicted_price price_category
0  966.0           795.13     Overpriced
1  142.0           319.44   Good Bargain
3  368.0           456.19     Fair Price
4  204.0           358.39   Good Bargain
5  577.0           599.20     Fair Price
```

Results

Results and Comparison:

- Linear Regression: MAE: 287.97, MSE: 110470.03, RMSE: 332.37, R^2: -0.0006
- Random Forest: MAE: 222.83, MSE: 75850.96, RMSE: 275.41, R^2: 0.3129

Random forest showed better accuracy, though improvements are possible.

Improvements and Future Work:
Future steps to enhance model performance include:

- Hyperparameter tuning
- Feature engineering
- Exploring advanced models (e.g., GBM, XGBoost)
- Cross-validation
- Data augmentation

Interpretation & Conclusions

Overall, through our first few models, we were not able to determine direct factors that contribute to high review scores and high prices. The price prediction model yielded the highest R-squared value of 0.31, which we deemed moderately successful. The model itself deemed that the days since the last review and reviews per month were the most important factors.

The price prediction model developed in this project can also be applied across several domains where accurate pricing is crucial. In real estate, it can assist in forecasting property values based on market trends, location, and features. In retail, the model can help businesses optimize pricing strategies by predicting product demand and adjusting prices accordingly. In finance, it could support investment strategies by forecasting asset prices based on historical data and economic indicators. By utilizing advanced machine learning techniques like random forest modeling, this approach can capture complex, non-linear relationships within the data, improving prediction accuracy and enabling more informed, data-driven decision-making. Furthermore, the model's modular design ensures scalability and adaptability to different industries and datasets, making it a valuable tool for a wide range of pricing-related tasks.

Sources

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