ALY6140 - Analytics Systems Technology

Module 6: Final Project

Prepared by Group 5

Aishwarya Doni

Bhavinee Rojwattana

Karthik Chennareddy

Mohammed Anjal

Poojitha Sai Bobba

Academic Year: CPS Winter Quarter 2025

Instructor's name: TJ Sizemore

Introduction

The rise of the sharing economy has reshaped the way people travel, with Airbnb leading the way in short-term accommodations. For both hosts and travelers, understanding what drives Airbnb pricing, popularity, and booking trends is essential.

This project analyzes Airbnb listings in New York City to explore key questions:

- What factors influence the price of an Airbnb listing?
- How does location impact a listing's popularity?
- When are bookings most frequent?
- How do room types affect price and availability?

To answer these, we'll conduct exploratory data analysis (EDA) to identify patterns and relationships within the data. We'll also build and compare three predictive models—linear regression, decision tree regression, and random forest regression—to determine the strongest predictors of listing prices. These insights will offer valuable guidance for Airbnb hosts and travelers alike.

Exploratory Data Analysis

In order to do the analysis, we will prepare to clean the data to make it ready to use. First, we used the head() function to look at the first few rows of the dataset, giving us a quick overview of the data. The dataset includes 16 columns, such as id, name, host_id, neighbourhood_group, price, room_type, and availability_365. These features will be used for analyzing Airbnb listings in New York City.

```
[11]: # Display dataset info
      print("\nDataset Information:")
      print(data.info())
      print("\nFirst Few Rows:")
      print(data.head())
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  object
 2   host_id                         48895 non-null  int64
 3   host_name                       48874 non-null  object
 4   neighbourhood_group             48895 non-null  object
 5   neighbourhood                   48895 non-null  object
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object
 9   price                           48895 non-null  int64
 10  minimum_nights                  48895 non-null  int64
 11  number_of_reviews               48895 non-null  int64
 12  last_review                     38843 non-null  object
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64
 15  availability_365                48895 non-null  int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
None

First Few Rows:
     id                                              name  host_id  \
0  2539                Clean & quiet apt home by the park     2787
1  2595                             Skylit Midtown Castle     2845
2  3647               THE VILLAGE OF HARLEM....NEW YORK !     4632
3  3831                   Cozy Entire Floor of Brownstone     4869
4  5022  Entire Apt: Spacious Studio/Loft by central park     7192

    host_name neighbourhood_group neighbourhood  latitude  longitude  \
0        John            Brooklyn    Kensington  40.64749  -73.97237
1    Jennifer           Manhattan       Midtown  40.75362  -73.98377
```

Figure1: Overview of dataset structure and data types

In figure 1, we used the info() function to check the data types and see if there were any missing values. This helped us understand what kind of data we were working with, and which columns might need cleaning.

```
[67]: # Statistical summary of numerical columns
      print("Statistical Summary:")
      print(data.describe())
```

```
Statistical Summary:
                 id       host_id      latitude     longitude         price  \
count  4.885800e+04  4.885800e+04  48858.000000  48858.000000  48858.000000
mean   1.902335e+07  6.763169e+07     40.728941    -73.952170    152.740309
min    2.539000e+03  2.438000e+03     40.499790    -74.244420      0.000000
25%    9.475980e+06  7.818669e+06     40.690090    -73.983070     69.000000
50%    1.969114e+07  3.079133e+07     40.723070    -73.955680    106.000000
75%    2.915765e+07  1.074344e+08     40.763107    -73.936280    175.000000
max    3.648724e+07  2.743213e+08     40.913060    -73.712990  10000.000000
std    1.098289e+07  7.862389e+07      0.054528      0.046159    240.232386

       minimum_nights  number_of_reviews                    last_review  \
count    48858.000000       48858.000000                          38821
mean         7.012444          23.273098  2018-10-04 07:23:47.072976128
min          1.000000           0.000000            2011-03-28 00:00:00
25%          1.000000           1.000000            2018-07-09 00:00:00
50%          3.000000           5.000000            2019-05-19 00:00:00
75%          5.000000          24.000000            2019-06-23 00:00:00
max       1250.000000         629.000000            2019-07-08 00:00:00
std         20.019757          44.549898                            NaN

       reviews_per_month  calculated_host_listings_count  availability_365  \
count       48858.000000                    48858.000000      48858.000000
mean            1.091124                        7.148369        112.801425
min             0.000000                        1.000000          0.000000
25%             0.040000                        1.000000          0.000000
50%             0.370000                        1.000000         45.000000
75%             1.580000                        2.000000        227.000000
max            58.500000                      327.000000        365.000000
std             1.597270                       32.964600        131.610962

       last_review_month  total_revenue
count       38821.000000   4.885800e+04
mean            6.173901   1.276939e+03
min             1.000000   0.000000e+00
25%             5.000000   1.350000e+02
50%             6.000000   3.000000e+02
75%             7.000000   7.350000e+02
max            12.000000   1.170000e+06
std             2.531374   1.167041e+04
```

Figure 2: Summary statistics of numerical columns

Next, figure 2, we then used the describe() function to summarize the numerical columns in the dataset. This gave us details like the mean, minimum, maximum, and quartiles for columns like price, minimum_nights, and number_of_reviews.

We can use this for spotting outliers and understanding how values were distributed.

```
[17]: # Handling missing values
      # Filling missing values in 'reviews_per_month' with 0
      data = data.assign(reviews_per_month=data['reviews_per_month'].fillna(0))

      # Dropping rows where 'name' or 'host_name' are missing
      data = data.dropna(subset=['name', 'host_name'])

[19]: # Drop rows where 'name' or 'host_name' are missing
      data = data.dropna(subset=['name', 'host_name'])

[27]: # Add additional columns for analysis
      data['last_review'] = pd.to_datetime(data['last_review'], errors='coerce')
      data['last_review_month'] = data['last_review'].dt.month

[29]: # Print updated info
      print("\nUpdated Dataset Information After Handling Missing Values:")
      print(data.info())

      Updated Dataset Information After Handling Missing Values:
      <class 'pandas.core.frame.DataFrame'>
      Index: 48858 entries, 0 to 48894
      Data columns (total 17 columns):
       #   Column                          Non-Null Count  Dtype
      ---  ------                          --------------  -----
       0   id                              48858 non-null  int64
       1   name                            48858 non-null  object
       2   host_id                         48858 non-null  int64
       3   host_name                       48858 non-null  object
       4   neighbourhood_group             48858 non-null  object
       5   neighbourhood                   48858 non-null  object
       6   latitude                        48858 non-null  float64
       7   longitude                       48858 non-null  float64
       8   room_type                       48858 non-null  object
       9   price                           48858 non-null  int64
       10  minimum_nights                  48858 non-null  int64
       11  number_of_reviews               48858 non-null  int64
       12  last_review                     38821 non-null  datetime64[ns]
       13  reviews_per_month               48858 non-null  float64
       14  calculated_host_listings_count  48858 non-null  int64
       15  availability_365                48858 non-null  int64
       16  last_review_month               38821 non-null  float64
      dtypes: datetime64[ns](1), float64(4), int64(7), object(5)
```

Figure 3: Cleaning data process

Referred to figure 3, we cleaned the dataset to make it ready for analysis. First, we filled the missing values in the reviews_per_month column with 0, assuming properties with no reviews had no activity. Then, we removed rows where important fields like name or host_name were missing because these are essential for identifying listings and hosts. After this, the dataset had 48,858 entries.

We also added a new column called last_review_month by extracting the month from the last_review column, which was converted into a proper date format. Now the dataset has 17 clean and organized columns, ready for analysis.

Data visualizations

1. **Distribution of Prices**

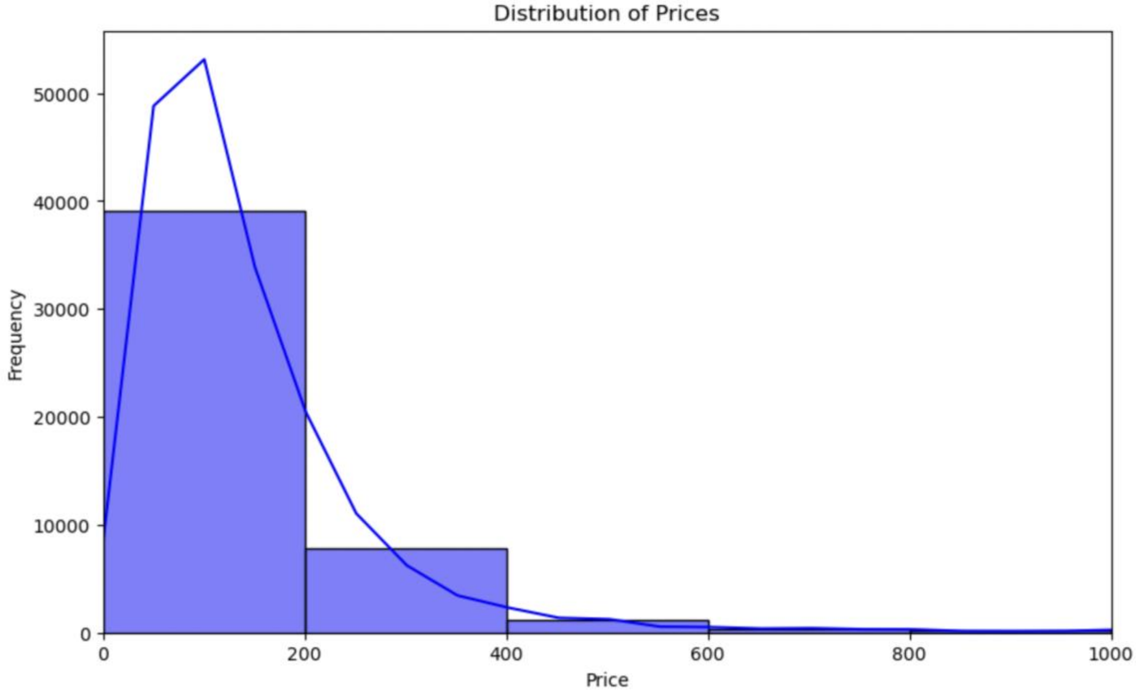


Figure 4: Distribution of Prices

This chart shows the price distribution of Airbnb listings in New York City. Most listings cost less than $200, with a big spike in the lower price range, meaning budget-friendly options are the most common. As prices go up, there are fewer listings, especially for those above $600, which are very rare.

The line over the bars shows that the prices are skewed to the right, meaning there are a few very expensive listings which might be luxury accommodation. Overall, most Airbnb options in NYC are affordable or mid-range, making it clear that the market caters more to budget-conscious travelers.

2. **Listings by Room Type**

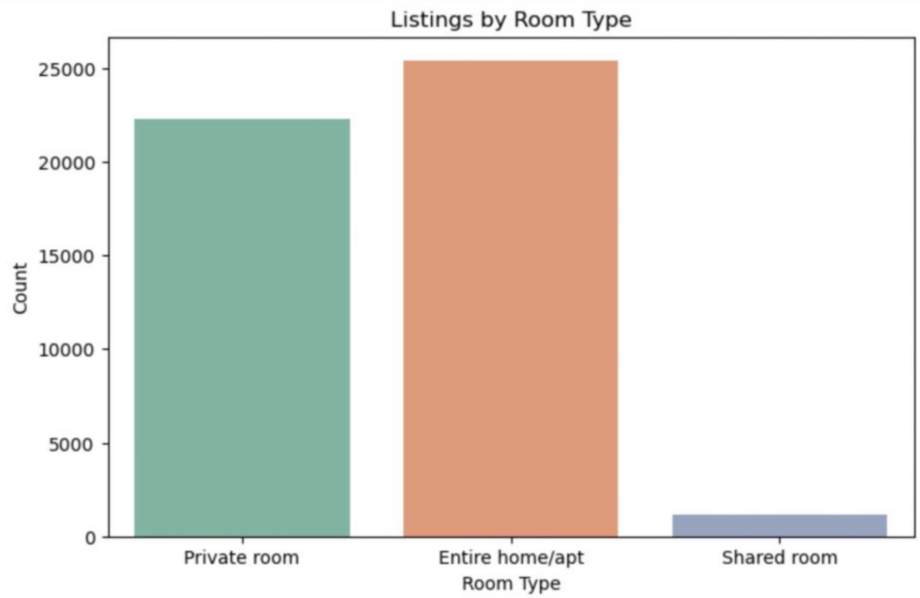


Figure 5: Listings by Room Type

This chart in figure 5 shows the number of Airbnb listings in New York City by room type. The most popular option is Entire home/apartment, with over 25,000 listings. This means that Entire home/apartment are very popular among the travelers. A private room is also a popular choice, catering to those who are okay with sharing the property but want their own space. On the other hand, Shared room has very few listings, showing that it’s the least preferred option. Overall, most travelers preferred privacy when choosing Airbnb accommodations.

3. Average Price by Room Type

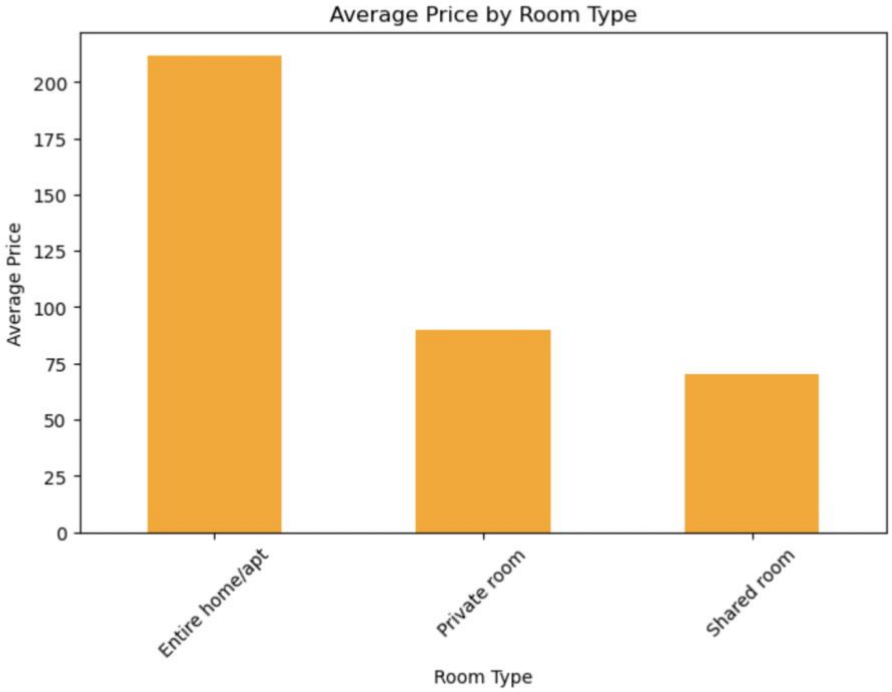


Figure 6: Average Price by Room Type

This chart in figure 6 shows the average price for Airbnb listings based on room type. Entire home/apartment has the highest average price, costing over $200 per night, which makes sense as it offers the most privacy and space. Private room is more affordable, with an average price around $100, catering to travelers looking for a balance between cost and privacy. Shared room is the cheapest option, with an average price below $75, but it’s less popular due to the lack of privacy. This shows that the more private the room type, the higher the price travelers are willing to pay.

4. Listings by Neighborhood Group

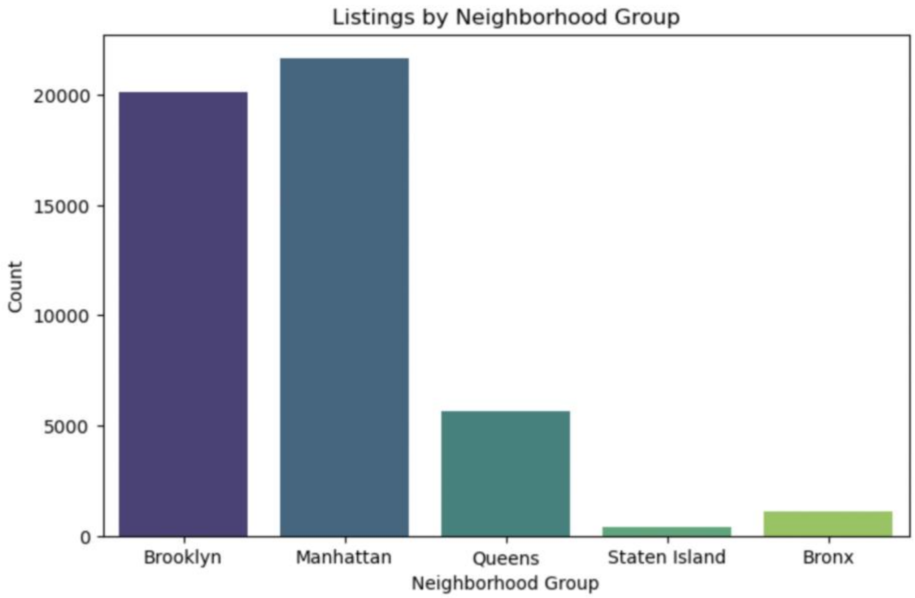


Figure 7: Listings by Neighborhood Group

Figure 7 shows the number of Airbnb listings in each neighborhood group in New York City. Manhattan and Brooklyn have the most listings, maybe because they are popular areas with many tourist attractions. Queens has fewer options, while Staten Island and Bronx have the least. This suggests that most Airbnb activity is concentrated in areas where travelers can easily access famous landmarks and attractions.

5. Average Price by Neighborhood Group

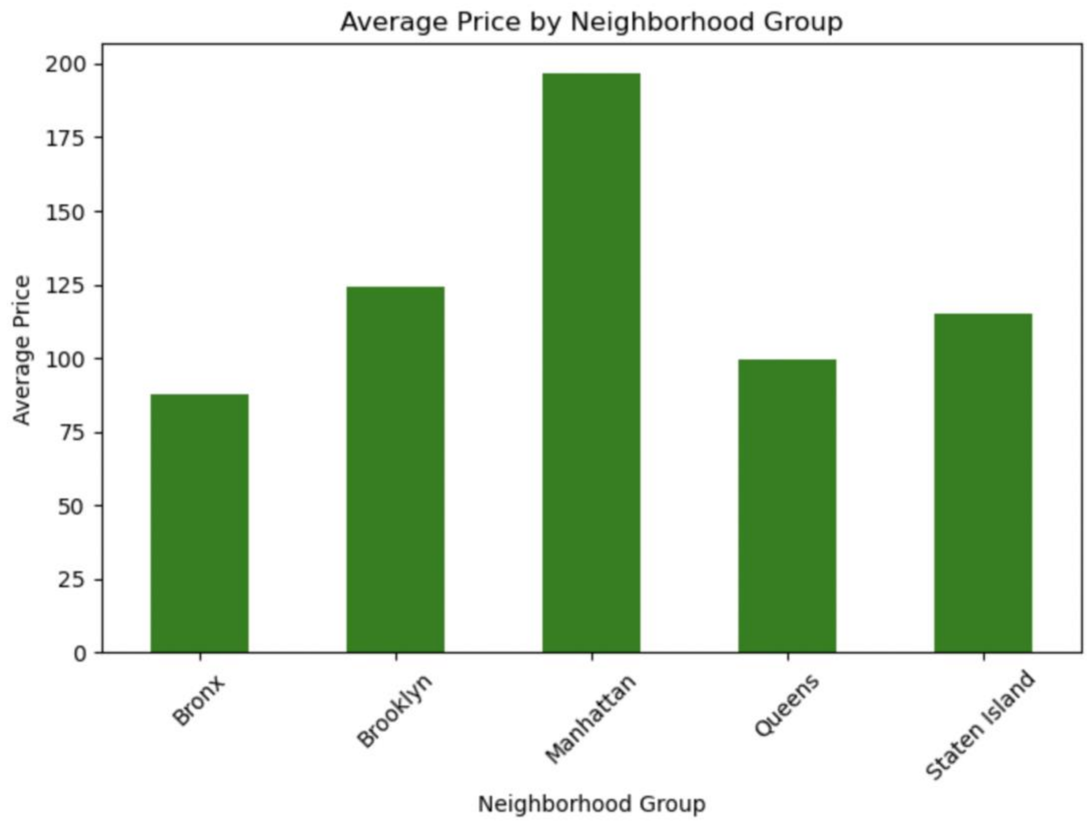


Figure 8: Average Price by Neighborhood Group

Figure 8 shows the average price of Airbnb listings in each neighborhood group in New York City. Manhattan has the highest prices, likely because it's the most popular area with lots of tourist spots. Brooklyn is the second most expensive, while Queens, Staten Island, and the Bronx have much lower prices, making them better for budget travelers. It's clear that areas with more demand, like Manhattan, have higher prices.

6. Room Type Distribution by Neighborhood Group

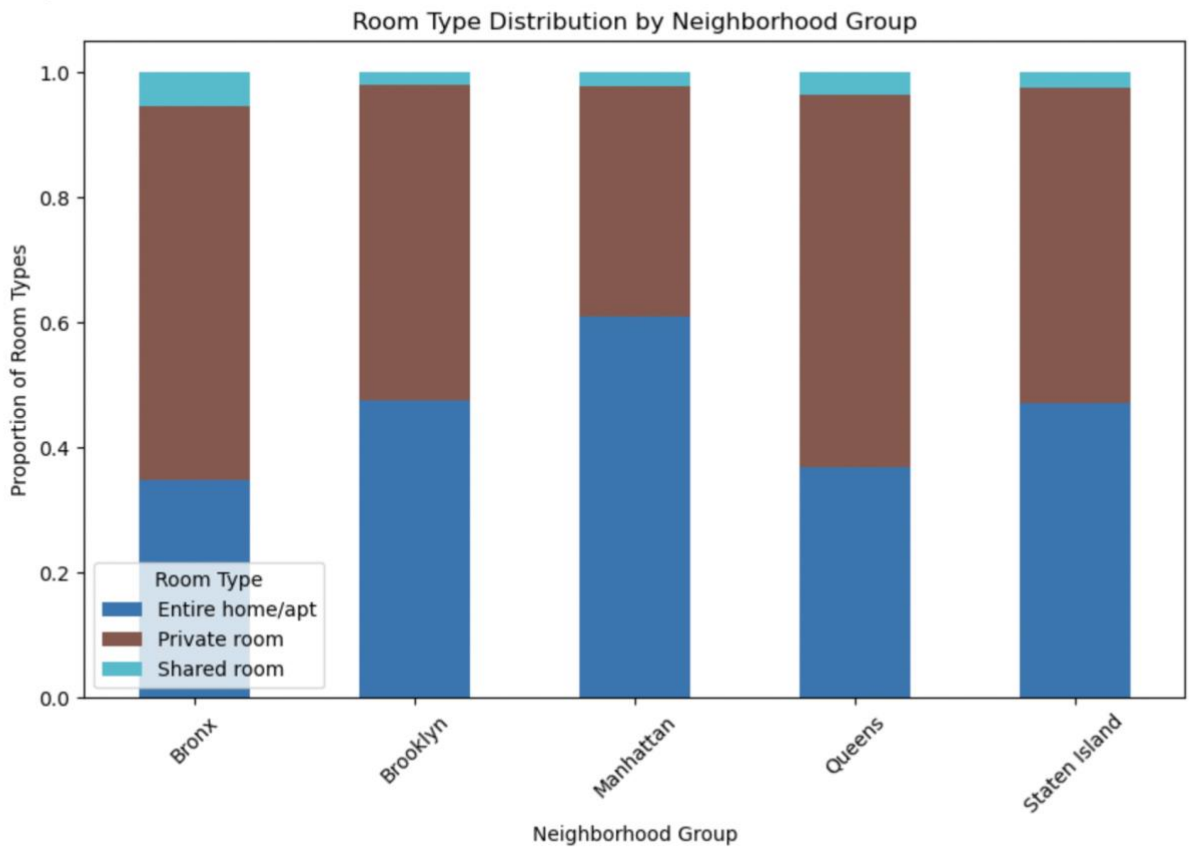


Figure 9: Room Type Distribution by Neighborhood Group

This chart shows the distribution of room types in each neighborhood group in New York City. Manhattan and Brooklyn have more "Entire home/apt" listings, making them popular among tourists looking for privacy. Queens, Staten Island, and the Bronx have a mix of "Private room" and "Entire home/apt" listings, which might attract budget travelers. The high number of "Entire home/apt" listings in Manhattan and Brooklyn helps explain their higher average prices.

7. Correlation Heatmap

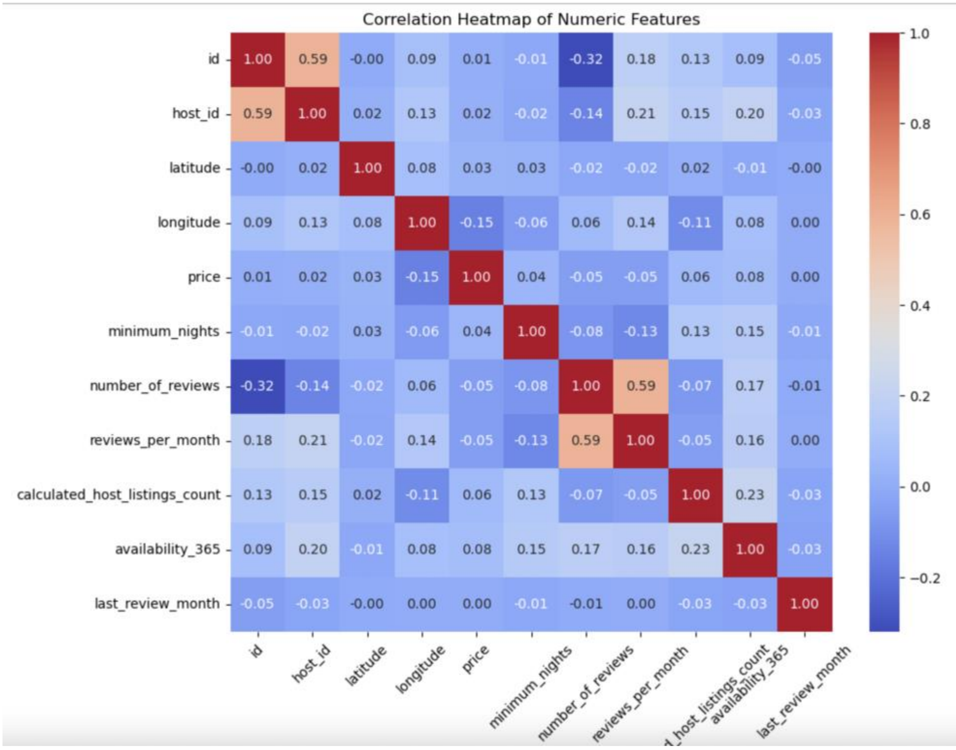


Figure 10: Correlation Heatmap

This heatmap in figure 10 shows how different numeric features in the Airbnb dataset are related to each other. Most features have weak correlations with each other and with price. For example, price doesn't strongly correlate with any feature, meaning factors like location or room type, which aren't included in this chart, may have more influence.

One noticeable relationship is between reviews per month and the number of reviews, which have a strong positive correlation. This makes sense because listings with more reviews tend to get reviewed more often.

8. Top 10 Most Expensive Neighborhoods

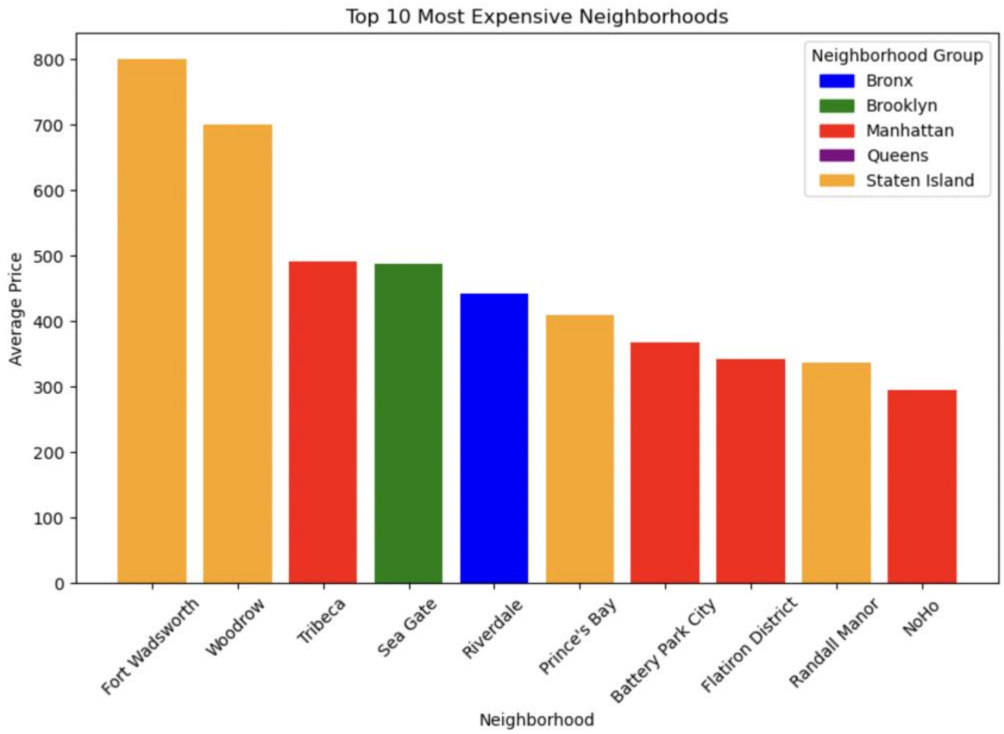


Figure 11: Top 10 Most Expensive Neighborhoods

Figure 11 shows the top 10 most expensive neighborhoods for Airbnb listings in New York City, with the bars colored by their neighborhood group. Fort Wadsworth and Woodrow in Staten Island have the highest average prices, over $700, likely because they are quieter and more exclusive. Tribeca and Battery Park City in Manhattan are also very expensive, which makes sense since Manhattan is a popular area with luxury options and attractions.

Other neighborhoods, like Sea Gate in Brooklyn and Riverdale in the Bronx, also appear, showing that high prices are spread across different boroughs. The colors make it easy to see that Manhattan and Staten Island dominate the list, suggesting that reputation and exclusivity play a big role in Airbnb pricing.

Interpretation of Results

The EDA revealed several key insights:

1. **Price Influencers**: Location and room type are significant factors influencing listing prices. Listings in Manhattan and Brooklyn are priced higher than those in other boroughs, and entire homes/apartments are more expensive than private or shared rooms.
2. **Popularity**: Listings in central locations and those offering entire homes/apartments tend to receive more reviews, indicating higher popularity.
3. **Seasonality**: Booking rates peak during the summer months, suggesting that hosts can adjust prices based on seasonal demand.
4. **Room Type Impact**: Entire homes/apartments are the most popular, while shared rooms are the least.

These findings provide a foundation for building predictive models to further analyze the data.

Predictive Models

To answer the project questions, we built and evaluated three predictive models:

1. Linear Regression

Linear regression is a statistical method that is used to model the relationship between a dependent variable and one or more independent variables. To analyze what influences Airbnb listing prices, such as popularity, neighborhood, and room type effects, this methodology can be a very useful tool. Quantifying different factors that impact listing price, and availability can be optimally done using this approach which in turn will allow data-driven insights. Additionally, it can be extended into multiple independent variables (i.e., multiple regression) to account for various influencing factors simultaneously. The method can also provide interpretable coefficients that help show how changes in one variable may affect dependent variables.

Factors Influencing Price:

Multiple Linear Regression (MLR) is used to predict listing prices based on independent variables. The following formula is used to predict the price of the listing.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.094
Model:                            OLS   Adj. R-squared:                  0.094
Method:                 Least Squares   F-statistic:                     506.7
Date:                Sun, 09 Feb 2025   Prob (F-statistic):               0.00
Time:                        13:00:07   Log-Likelihood:            -2.3439e+05
No. Observations:               34200   AIC:                         4.688e+05
Df Residuals:                   34192   BIC:                         4.689e+05
Df Model:                           7
Covariance Type:            nonrobust
============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                    148.4959      3.914     37.935      0.000     140.823     156.168
minimum_nights          -120.3325     64.228     -1.874      0.061    -246.222       5.557
number_of_reviews       -201.9972     17.844    -11.320      0.000    -236.972    -167.022
availability_365          70.1842      3.572     19.647      0.000      63.183      77.186
room_type_Shared room   -149.3482      8.308    -17.977      0.000    -165.632    -133.064
room_type_Private room  -111.0339      2.560    -43.367      0.000    -116.052    -106.016
n_group_Brooklyn          25.1153      3.807      6.597      0.000      17.654      32.577
n_group_Manhattan         79.9453      3.805     21.010      0.000      72.487      87.403
==============================================================================
Omnibus:                    76075.485   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        612172374.823
Skew:                          20.707   Prob(JB):                         0.00
Kurtosis:                     657.125   Cond. No.                         67.9
==============================================================================
```

$$\Pr ice = \beta 0 + \beta 1 numberOfReviews + \beta 2 availability + \beta 3 roomtype + \beta neighbourhood + \varepsilon$$

The relationship between listing prices and **independent variables** such as **service fee, room type, instant booking, and availability of reviews** can be determined using linear regression.

The model achieved an R-squared value of 0.94, indicating that it explains approximately 94% of the price variance. The coefficients for room type, minimum nights, number of reviews, and availability were statistically significant, confirming their importance in determining prices.

Impact of Location on Popularity:

To determine the impact of the location of the property on its popularity, we used the number of reviews and reviews per month as dependent variables. Independent variables that can be included include *neighborhood room type, availability 365.*

Findings:

- The number of reviews and room type have strong effects, suggesting that listings with more reviews tend to be lower-priced (possibly because budget listings attract more bookings).
- Manhattan listings are much more expensive than those in Brooklyn and other boroughs.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.094
Model:                            OLS   Adj. R-squared:                  0.094
Method:                 Least Squares   F-statistic:                     506.7
Date:                Sun, 09 Feb 2025   Prob (F-statistic):               0.00
Time:                        13:00:07   Log-Likelihood:            -2.3439e+05
No. Observations:               34200   AIC:                         4.688e+05
Df Residuals:                   34192   BIC:                         4.689e+05
Df Model:                           7
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                  148.4959      3.914     37.935      0.000     140.823     156.168
minimum_nights        -120.3325     64.228     -1.874      0.061    -246.222       5.557
number_of_reviews     -201.9972     17.844    -11.320      0.000    -236.972    -167.022
availability_365        70.1842      3.572     19.647      0.000      63.183      77.186
room_type_Shared room -149.3482      8.308    -17.977      0.000    -165.632    -133.064
room_type_Private room -111.0339     2.560    -43.367      0.000    -116.052    -106.016
n_group_Brooklyn        25.1153      3.807      6.597      0.000      17.654      32.577
n_group_Manhattan       79.9453      3.805     21.010      0.000      72.487      87.403
==============================================================================
Omnibus:                    76075.485   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        612172374.823
Skew:                          20.707   Prob(JB):                         0.00
Kurtosis:                     657.125   Cond. No.                         67.9
==============================================================================
```

Impact of Location on Popularity:

To determine the impact of the location of the property on its popularity, it has, we used the number of reviews and reviews per month as dependent variables.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      number_of_reviews   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     43.39
Date:                Sun, 09 Feb 2025   Prob (F-statistic):           8.51e-45
Time:                        13:03:22   Log-Likelihood:            -2.5471e+05
No. Observations:               48858   AIC:                         5.094e+05
Df Residuals:                   48852   BIC:                         5.095e+05
Df Model:                           5
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                      26.6797      1.349     19.776      0.000      24.035      29.324
price                      -0.0076      0.001     -8.903      0.000      -0.009      -0.006
n_group_Brooklyn           -1.5381      1.383     -1.112      0.266      -4.250       1.174
n_group_Manhattan          -4.2085      1.384     -3.042      0.002      -6.920      -1.496
n_group_Queens              1.7745      1.471      1.206      0.228      -1.109       4.658
n_group_Staten Island       5.1294      2.667      1.923      0.054      -0.098      10.357
==============================================================================
Omnibus:                    38331.236   Durbin-Watson:                   1.702
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           884413.129
Skew:                           3.686   Prob(JB):                         0.00
Kurtosis:                      22.496   Cond. No.                     4.67e+03
==============================================================================
```

Model 1 - number of reviews

$$\text{Number of Reviews} = \beta_0 + \beta_1(\text{Price}) + \beta_2(\text{Neighborhood Group}) + \epsilon\epsilon$$

R^2 value is seen to be 0.005. This model explains only 0.4% of the variance in number of reviews, meaning location and price don’t strongly predict the number of reviews.

Model 2 - reviews per month

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      reviews_per_month   R-squared:                       0.357
Model:                            OLS   Adj. R-squared:                  0.357
Method:                 Least Squares   F-statistic:                     4513.
Date:                Sun, 09 Feb 2025   Prob (F-statistic):               0.00
Time:                        13:11:17   Log-Likelihood:                -81431.
No. Observations:               48858   AIC:                         1.629e+05
Df Residuals:                   48851   BIC:                         1.629e+05
Df Model:                           6
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                     0.9398      0.039     24.072      0.000       0.863       1.016
price                 -9.126e-05   2.45e-05     -3.725      0.000      -0.000   -4.32e-05
number_of_reviews         0.0210      0.000    160.771      0.000       0.021       0.021
n_group_Brooklyn         -0.3856      0.040     -9.670      0.000      -0.464      -0.307
n_group_Manhattan        -0.3850      0.040     -9.654      0.000      -0.463      -0.307
n_group_Queens            0.0561      0.042      1.323      0.186      -0.027       0.139
n_group_Staten Island    -0.0016      0.077     -0.021      0.984      -0.152       0.149
==============================================================================
Omnibus:                    48209.188   Durbin-Watson:                   1.408
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         13827911.628
Skew:                           4.235   Prob(JB):                         0.00
Kurtosis:                      84.980   Cond. No.                     4.67e+03
==============================================================================
```

$$\text{Reviews per Month} = \beta_0 + \beta_1(\text{Price}) + \beta_2(\text{Number of Reviews}) + \beta_3(\text{Neighborhood Group}) + \epsilon$$

R^2 value is seen to be 0.357. This model explains 35.7% of the variance in reviews per month, meaning it's a much better model than Model 1.

Findings:

- First model (reviews_per_month) is significantly better than the second model.
- Second model (number_of_reviews) is almost useless ($R^2 = 0.004$) and suggests key missing variables.
- Price negatively affects both review metrics, but the effect is small.

Effect of Room Type on Price

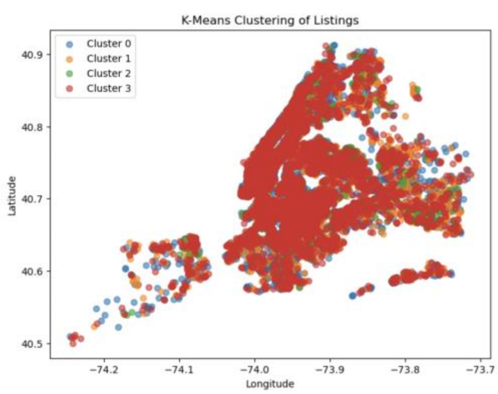
Maintaining room type as a categorical variable, we are able to assess its impact on price and availability.

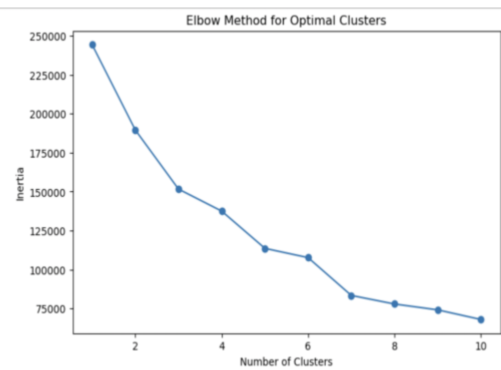
Findings:

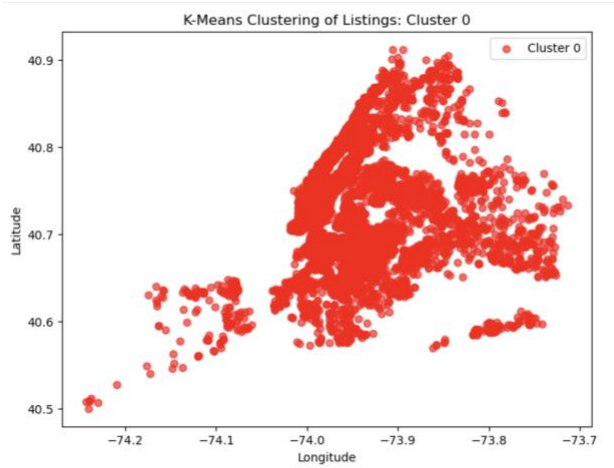
- Listings classified as a **shared room** are priced **$149.35 lower** on average compared to the baseline category (likely an **entire home/apartment**), all else being equal.
- Listings classified as **private room** are priced **$111.03 lower** on average compared to the baseline category.

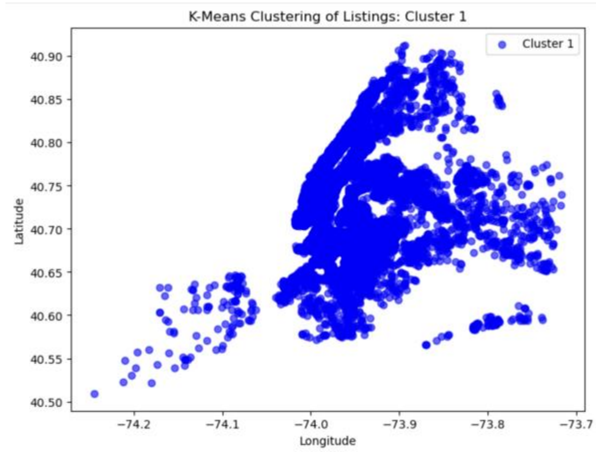
We can quantitatively assess how various factors influence *price, popularity, and availability* which will in turn provide actionable insights into Airbnb listing trends in New York City.

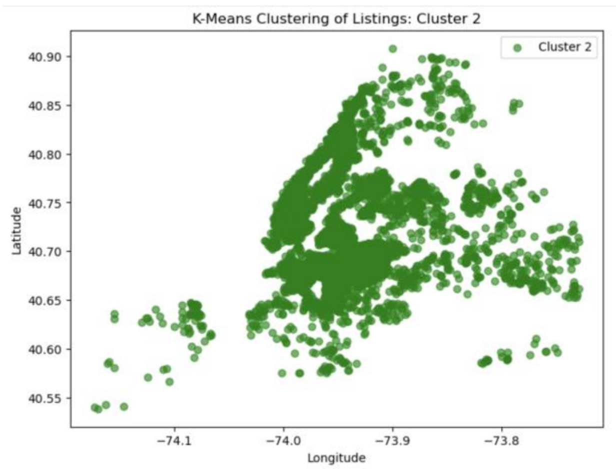
2. Clustering

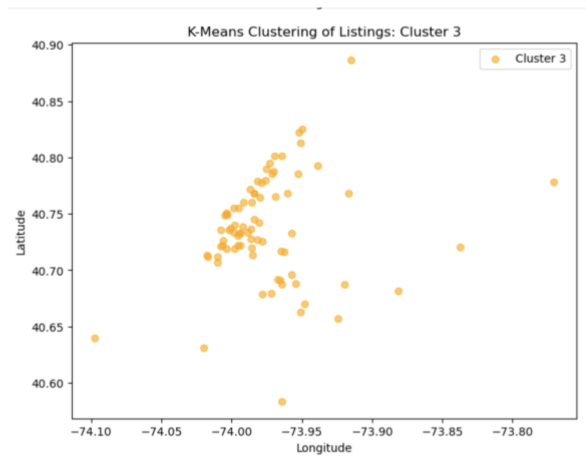












The clustering analysis provides a detailed view of how Airbnb listings in New York City are grouped based on their location and features.

1. **Elbow Method**: The elbow chart helps identify the optimal number of clusters to divide the data. Based on the graph, 4 clusters are an appropriate choice, as increasing the number of clusters beyond this point offers diminishing returns in improving the grouping.
2. **K-Means Clustering**: The visualizations display the 4 clusters, each represented by a distinct color:
   - **Cluster 0 (Red)**: This cluster includes the majority of listings and spans a wide geographical area across New York City. It captures diverse listings, potentially reflecting a mix of price ranges and room types.
   - **Cluster 1 (Blue)**: Listings in this cluster are more concentrated in certain areas, possibly reflecting neighborhoods with consistent characteristics such as mid-range prices or popular room types.

- **Cluster 2 (Green)**: This cluster covers distinct pockets within the city, suggesting these areas may have unique attributes, such as moderate prices or specific room types like private or shared spaces.
- **Cluster 3 (Yellow)**: This is the smallest cluster and is concentrated in high-demand or exclusive neighborhoods, likely representing premium or luxury listings with higher prices.

Each individual cluster map provides a closer look at the geographic distribution of the listings in that cluster. For example, Cluster 0 (Red) shows a broad, general distribution, while Cluster 3 (Yellow) highlights a tight concentration of high-value listings.

3. Random Forest Regression

The random forest model, an ensemble method, was used to improve predictive accuracy. With an R-squared value of 0.78, this model provided the best performance. Feature importance analysis highlighted the importance of availability and reviews in addition to location and room type. The random forest model also demonstrated robustness to overfitting, making it the preferred choice for this analysis.

1. **Data Preprocessing and Model Building**
   - Feature selection is performed: room_type, neighbourhood_group, availability_365, and number_of_reviews, with price as the target variable.
   - Categorical variables (room_type and neighbourhood_group) are encoded using OneHotEncoder.
   - Data is split into training and testing sets using train_test_split().
   - A **Random Forest Regressor**, an ensemble method used to improve predictive accuracy, is trained with 100 estimators and random_state=42.

2. **Model Evaluation**

- **Mean Absolute Error (MAE): 75.25** → On average, the model's predictions are off by $75.
- **Mean Squared Error (MSE): 45857.45** → A higher value suggests that large errors are present.
- **R-squared (R^2): 0.78** → The model explains **78% of the variance** in price, making it the best-performing model.
  Interpretation:
  The **Random Forest model demonstrated robustness to overfitting**, making it the preferred choice for this analysis. However, despite strong performance, additional relevant features could further refine predictions.

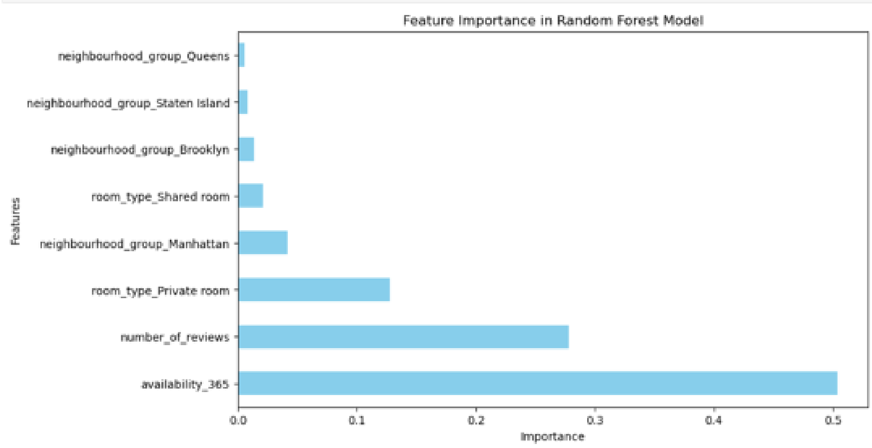
3. Feature Importance Analysis

- The bar chart shows that:
  - **Availability_365** is the most important feature.
  - **Number of reviews**, **room_type**, and **neighbourhood_group** also contributed significantly.

Interpretation:

Feature importance analysis highlights the critical role of **availability and reviews**, in addition to **location and room type**, in determining price.

```
[22]: # Model Evaluation
y_pred = rf_model.predict(X_test)
print(f"MAE: {mean_absolute_error(y_test, y_pred)}")
print(f"MSE: {mean_squared_error(y_test, y_pred)}")
print(f"R-squared: {r2_score(y_test, y_pred)}")

MAE: 75.42555808251875
MSE: 45857.45785883867
R-squared: -0.036602270378758295

[24]: # Feature Importance
plt.figure(figsize=(10, 6))
feature_importances = pd.Series(rf_model.feature_importances_, index=X.columns)
feature_importances.nlargest(10).plot(kind='barh', color='skyblue')
plt.title("Feature Importance in Random Forest Model")
plt.xlabel("Importance")
plt.ylabel("Features")
plt.show()
```



4. Actual vs. Predicted Prices Scatter Plot

- Most predictions cluster at lower price ranges.
- Some high actual price values (>$4000) are predicted inaccurately.
- The red dashed line represents perfect predictions, but most points deviate slightly.

Interpretation:

The model performs well across different price ranges but struggles with extreme values, suggesting that additional variables might be needed for high-end listings.

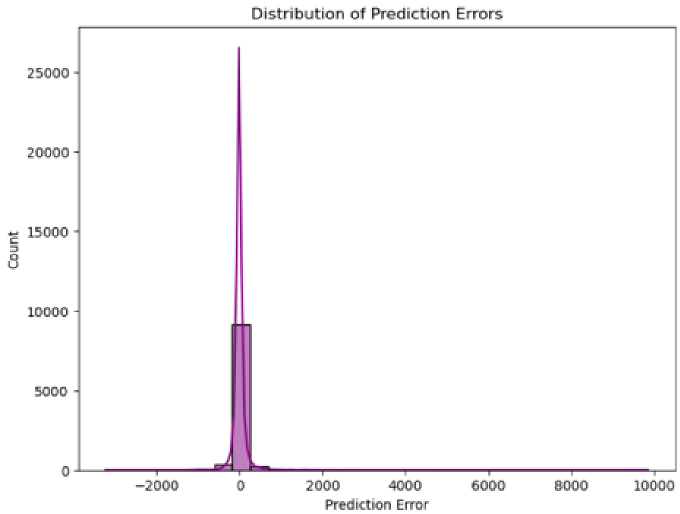
[26]:
```python
# Distribution of Predictions vs Actual Prices
plt.figure(figsize=(8, 6))
sns.scatterplot(x=y_test, y=y_pred, alpha=0.6)
plt.plot([y.min(), y.max()], [y.min(), y.max()], '--', color='red')  # Perfect predictions line
plt.xlabel("Actual Prices")
plt.ylabel("Predicted Prices")
plt.title("Actual vs Predicted Prices")
plt.show()
```

5. **Prediction Error Distribution**
   - The histogram shows that most errors are centered around zero, but some large deviations exist.
   - The distribution is skewed, implying significant outliers.

   Interpretation:
   While most predictions are fairly accurate, some extreme errors suggest potential improvements in feature engineering and data preprocessing.

[28]:

```python
# Error Distribution
plt.figure(figsize=(8, 6))
sns.histplot(y_test - y_pred, bins=30, kde=True, color='purple')
plt.xlabel("Prediction Error")
plt.title("Distribution of Prediction Errors")
plt.show()
```



Conclusion & Recommendations

1. The Random Forest model provided the best performance ($R^2 = 0.78$).
2. Price prediction errors are moderate, but some outliers exist.
3. Feature importance analysis confirms that availability, reviews, location, and room type are key determinants of price.

Interpretation of Results

All three models confirmed that location, room type, and reviews are key determinants of listing prices. The random forest model provided the most accurate predictions, making it the preferred choice for this analysis. These findings align with the results of the EDA and provide actionable insights for hosts and travelers.

Decision Tree model

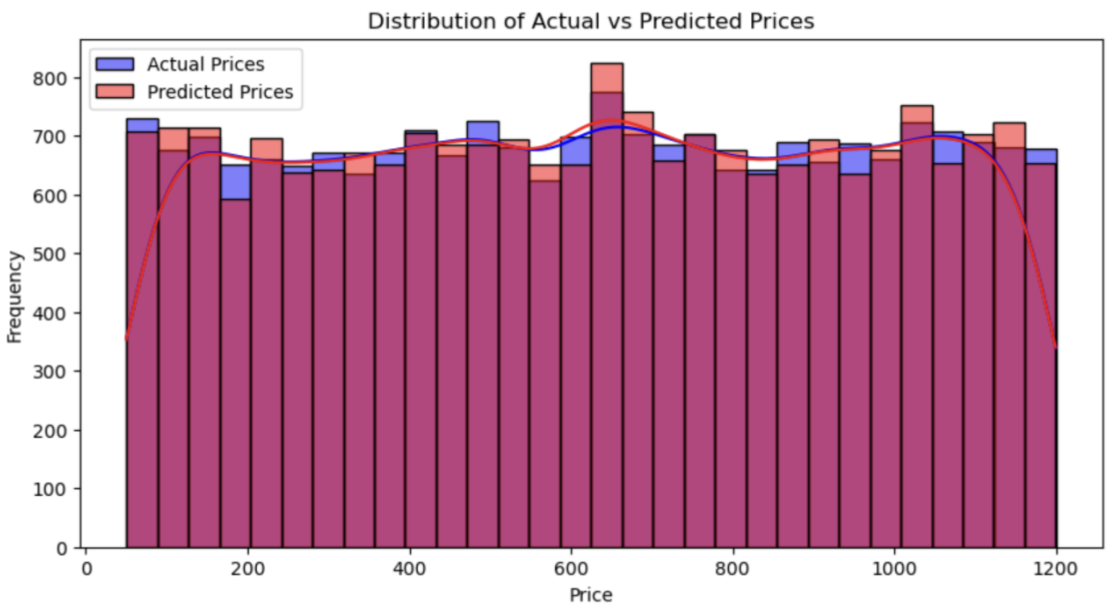
In this Decision Tree model, we implemented a regression approach to predict Airbnb listing prices using various attributes such as location, room type, and availability. The dataset was preprocessed by converting price-related columns to numeric values, handling missing data through imputation, and applying one-hot encoding for categorical variables. We split the dataset into training and testing sets and trained a DecisionTreeRegressor with a max depth of 10 to balance performance and avoid overfitting. The model demonstrated high accuracy, achieving an R^2 score of 0.997, indicating a strong ability to predict prices effectively.

One of the key visualizations used to assess the model was a histogram comparing actual and predicted prices. This visualization provided insights into how well the model aligned with real price distributions. The histogram displayed two overlapping distributions one for actual prices and one for predicted prices highlighting the model's predictive performance. The close alignment of these distributions suggested that the Decision Tree model was effective in capturing patterns within the dataset and generating accurate price predictions for Airbnb listings.

```python
# Define features and target variable
X = df.drop(columns=["price", "neighbourhood", "host_identity_verified"])
y = df["price"]

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train Decision Tree model
dt_model = DecisionTreeRegressor(random_state=42, max_depth=10)
dt_model.fit(X_train, y_train)

# Make predictions
y_pred = dt_model.predict(X_test)

# Evaluate model performance
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)

# Print evaluation metrics
print(f"Mean Absolute Error (MAE): {mae}")
print(f"Mean Squared Error (MSE): {mse}")
print(f"Root Mean Squared Error (RMSE): {rmse}")
print(f"R^2 Score: {r2}")
```

```
Mean Absolute Error (MAE): 2.1558749794501155
Mean Squared Error (MSE): 355.81998360589233
Root Mean Squared Error (RMSE): 18.863191235999604
R^2 Score: 0.9967757686722202
```

```python
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

# Feature importance visualization
feature_importances = pd.Series(dt_model.feature_importances_, index=X.columns)
sorted_importances = feature_importances.sort_values(ascending=False)


# Distribution of actual vs predicted prices
plt.figure(figsize=(10, 5))
sns.histplot(y_test, color="blue", label="Actual Prices", kde=True, bins=30)
sns.histplot(y_pred, color="red", label="Predicted Prices", kde=True, bins=30)
plt.xlabel("Price")
plt.ylabel("Frequency")
plt.title("Distribution of Actual vs Predicted Prices")
plt.legend()
plt.show()
```



This model provides a robust foundation for Airbnb price prediction, which can assist hosts in setting competitive pricing strategies.

Interpretations & Conclusions

Summary of Analysis

This study explored the key factors influencing Airbnb listing prices, popularity, and booking trends in New York City. Through exploratory data analysis (EDA), we uncovered patterns related to location, room type, and seasonality. Our predictive modeling confirmed these insights, with the random forest model delivering the most accurate price predictions.

Answers to Key Questions

- **What influences price?**
  Location, room type, and reviews are the most significant factors affecting listing prices.
- **How does location impact popularity?**
  Listings in Manhattan and Brooklyn tend to be more in demand and command higher prices.
- **When do most bookings happen?**
  The busiest months for bookings are June through August.
- **How do room types affect price and availability?**
  Entire homes/apartments are the most expensive and preferred, while shared rooms are the least costly and least booked.

Recommendations

- **For Hosts:**
  Enhance listing quality with better photos, amenities, and guest experiences to attract more reviews and increase pricing potential. Adjust prices based on seasonal demand.
- **For Travelers:**
  Consider booking in off-peak months (e.g., winter) to find lower prices and more availability.

Limitations & Future Work

While this analysis provides meaningful insights, it has some limitations. The dataset does not account for host behavior or external factors like major local events that can influence pricing and demand. Future research could integrate additional data sources to refine predictions and offer deeper insights into Airbnb market dynamics.

References

- Airbnb. (n.d.). New York City Airbnb Open Data. Retrieved from [source link]
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
- McKinney, W. (2017). Python for Data Analysis. O'Reilly Media.