STAT 650
Final Project

Aggie Honor Code

“An Aggie does not lie, cheat, or steal or tolerate those who do”

Aggie Integrity Statement

"On my honor, as an Aggie, I have neither given nor received unauthorized aid on this academic work."

By
Shesheer Rao Kokkirala
UIN:234008689

1. Introduction and Data Description

Overview of the Project and Objectives

This project aims to analyze the NYC Airbnb Open Data to understand the factors influencing the price of Airbnb listings in New York City. By performing a regression analysis, we seek to identify key predictors of listing prices and build models to predict prices for new listings.

Dataset Description

The dataset contains information about Airbnb listings in NYC, including variables such as:

- **id**: Unique identifier for each listing
- **name**: Name of the listing
- **host_id**: Unique identifier for the host
- **host_name**: Name of the host
- **neighbourhood_group**: NYC borough (e.g., Manhattan, Brooklyn)
- **neighbourhood**: Specific neighborhood within the borough
- **latitude and longitude**: Geographical coordinates of the listing
- **room_type**: Type of room (e.g., Entire home/apt, Private room)
- **price**: Price per night
- **minimum_nights**: Minimum number of nights required to book
- **number_of_reviews**: Total number of reviews
- **last_review**: Date of the last review
- **reviews_per_month**: Average number of reviews per month
- **calculated_host_listings_count**: Number of listings by the same host
- **availability_365**: Number of available days within a year
- **number_of_reviews_ltm**: Number of reviews in the last 12 months
- **license**: License status
- **rating**: Rating of the listing
- **bedrooms**: Number of bedrooms
- **beds**: Number of beds
- **baths**: Number of bathrooms

Importance of Regression Analysis

Regression analysis is a powerful statistical method that allows us to examine the relationship between a dependent variable (in this case, the price of Airbnb listings) and one or more independent variables (features such as room type, neighbourhood, number of reviews, etc.). By using regression analysis, we can identify which factors significantly impact the price and quantify their effects. This information is crucial for hosts who want to set competitive prices and for guests who wish to find the best value for their money.

Regression models can also help in predicting prices for new listings, providing a data-driven approach to price setting. Various types of regression models, such as linear regression, polynomial regression, ridge regression, and lasso regression, will be explored to find the best fit for our data.

2. Data Preprocessing

2.1Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, where we summarize the main characteristics of a dataset, often using visual methods. The primary objective of EDA is to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations. By performing EDA, we can better understand the underlying structure of our data, identify potential relationships between variables, and inform the selection of appropriate models for further analysis. In this analysis, we are examining the NYC Airbnb Open Data to gain insights into various features such as pricing, availability, review patterns, and neighborhood distributions. We will use a combination of descriptive statistics and visualizations to explore and interpret the data.

2.2 Plots:

Statistical Summary:

:	index	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	c
0	count	1.893400e+04	1.893400e+04	18934.000000	18934.000000	18934.000000	18934.000000	18934.000000	18934.000000	
1	mean	3.111585e+17	1.763845e+08	40.725606	-73.936773	190.566969	28.661931	42.968153	1.269512	
2	std	3.921531e+17	1.740101e+08	0.061267	0.062215	1067.593109	34.266639	74.156290	1.880605	
3	min	6.848000e+03	1.678000e+03	40.500314	-74.249840	10.000000	1.000000	1.000000	0.010000	
4	25%	2.694977e+07	1.975818e+07	40.682865	-73.978318	79.000000	30.000000	4.000000	0.220000	
5	50%	5.046517e+07	1.103463e+08	40.719990	-73.947335	125.000000	30.000000	15.000000	0.680000	
6	75%	7.269412e+17	3.158839e+08	40.762839	-73.915049	200.000000	30.000000	50.000000	1.830000	
7	max	1.054376e+18	5.504035e+08	40.911147	-73.713650	100000.000000	1250.000000	1865.000000	75.490000	

calculated_host_listings_count	availability_365	number_of_reviews_ltm	bedrooms	beds	baths
18934.000000	18934.000000	18934.000000	18934.000000	18934.000000	18934.000000
17.227580	206.680311	10.932027	1.432133	1.760537	1.194254
70.036998	135.330181	21.495077	0.815211	1.247959	0.497394
1.000000	0.000000	0.000000	1.000000	1.000000	0.000000
1.000000	87.000000	1.000000	1.000000	1.000000	1.000000
2.000000	215.000000	3.000000	1.000000	1.000000	1.000000
5.000000	356.000000	15.000000	2.000000	2.000000	1.000000
713.000000	365.000000	1075.000000	15.000000	42.000000	15.500000

1. General Overview:

- The dataset comprises various features related to Airbnb listings, including geographical coordinates, pricing, and availability metrics.

2. Key Metrics:

- Price: The average price of listings is approximately $190.57, with a maximum price reaching up to $100,000. The standard deviation is quite high at $1,067.59, indicating significant variability in listing prices.
- Minimum Nights: On average, listings require around 28.66 minimum nights, but this ranges widely from 1 to 1,250 nights, suggesting both short-term and long-term rental options are available.
- Number of Reviews: Listings have an average of 42.97 reviews, with a substantial range from 1 to 1,865 reviews, reflecting a diverse range of popularity and review frequency.
- Reviews Per Month: The average reviews per month is about 1.27, with a maximum value of 75.49, indicating some listings receive significantly more frequent reviews than others.
- Host Listings Count: On average, hosts manage approximately 17.23 listings, with some hosts managing as many as 713 listings.
- Availability: The average availability across the year is 206.68 days, ranging from 0 to 365 days, which shows variability in how frequently properties are available for booking.
- Bedrooms and Beds: Listings have an average of 1.43 bedrooms and 1.76 beds. The maximum numbers are notably high, with 15 bedrooms and 42 beds in some properties, suggesting a wide range of property sizes.

3. Observations:

- There is substantial variability in pricing and availability, reflecting the diversity of listings in terms of location, size, and type.
- The high standard deviation in price and number of reviews indicates the presence of both budget and luxury listings, as well as listings with varying levels of popularity.
- The data shows that while most listings have a modest number of bedrooms and beds, there are also some exceptionally large properties.

4. Recommendations:

- For targeted analysis, consider segmenting the data by price range or number of reviews to better understand different market segments.
- Further analysis could explore correlations between features like price and reviews or availability and number of listings managed by hosts.

Plots and their Significance

1. Price Distribution

This plot shows the distribution of property prices in the dataset. It helps to understand the range and frequency of different price points for listings. Most properties are priced between $100 and $200, with a noticeable right tail indicating some high-priced outliers.

2. Number of Reviews Distribution

This plot illustrates how the number of reviews varies among different listings. It shows that most properties have fewer reviews, with a high frequency of listings having between 0 and 20 reviews. Some properties have a significantly higher number of reviews, suggesting popularity or longer availability.

3. Reviews Per Month Distribution

This plot demonstrates the frequency of reviews per month for different listings. It indicates that most listings receive very few reviews each month, with many properties receiving between 0 and 1 review. There are a few listings with higher review rates.

4. Availability (365 days) Distribution

This plot shows how many days each property is available throughout the year. It reveals that many properties are available for the full year, with a concentration of

listings having lower availability, suggesting some properties are seasonal or occasionally listed.

5. Distribution of Neighbourhood Groups

This plot displays the frequency of listings in different neighbourhood groups. It highlights which neighbourhoods have a higher concentration of listings and provides insight into the geographic distribution of properties.

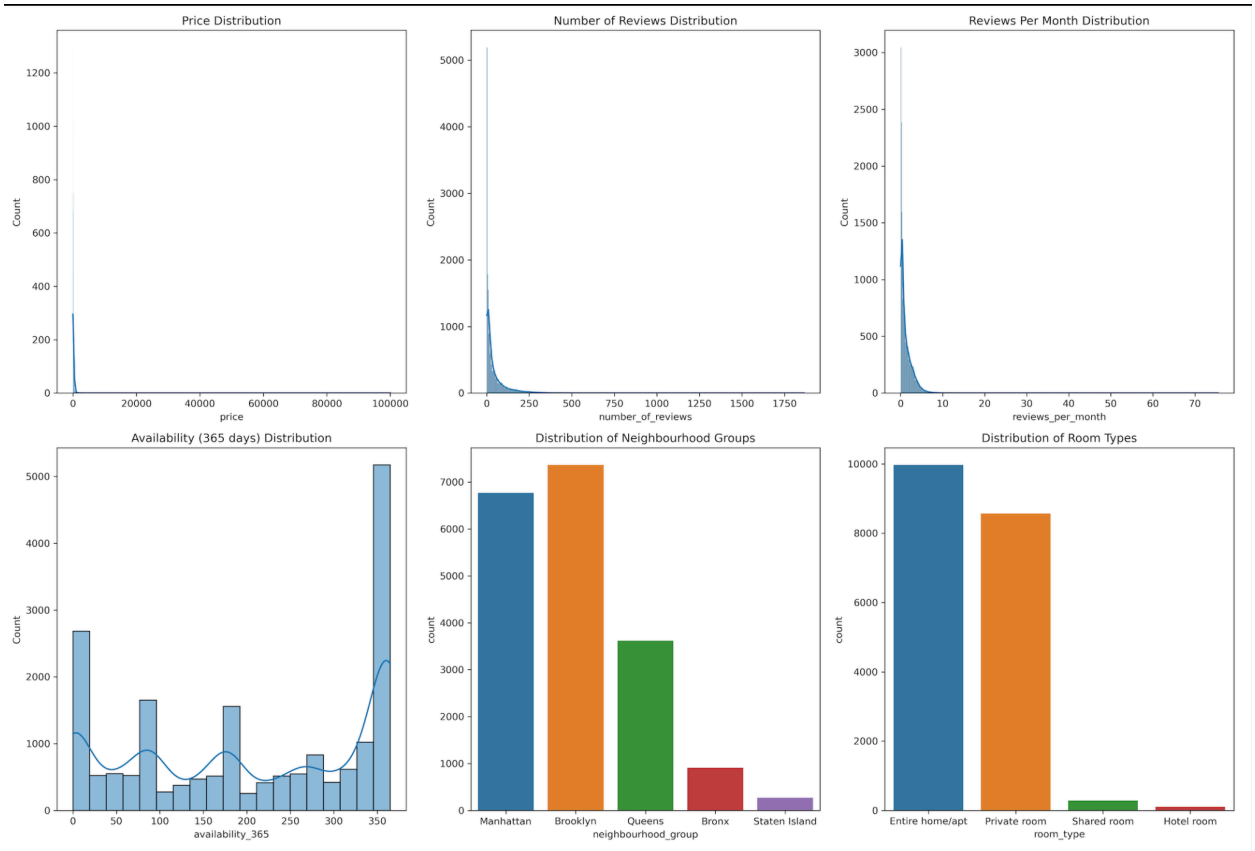
6. Distribution of Room Types

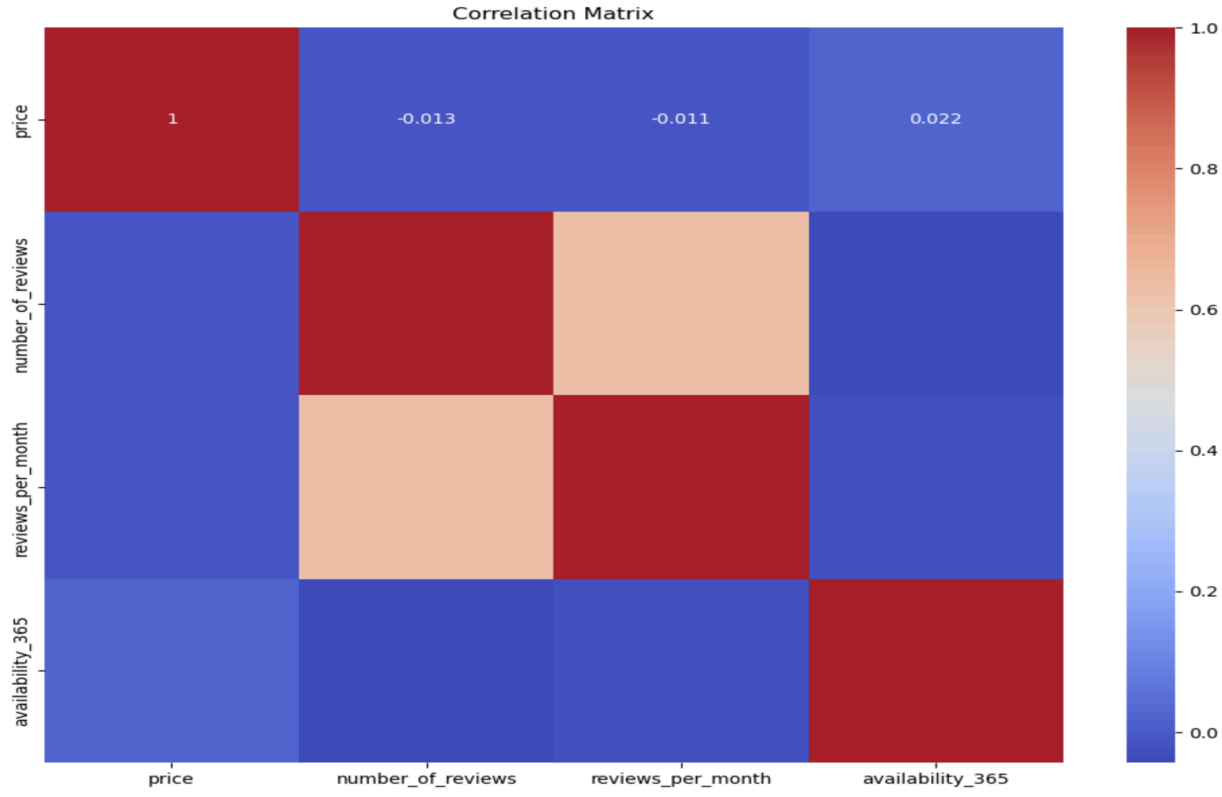
This plot shows how different types of rooms (e.g., entire home, private room) are distributed in the dataset. It indicates that 'Entire home/apt' is the most common type, followed by 'Private room' and 'Shared room,' reflecting guest preferences for private or whole accommodations.

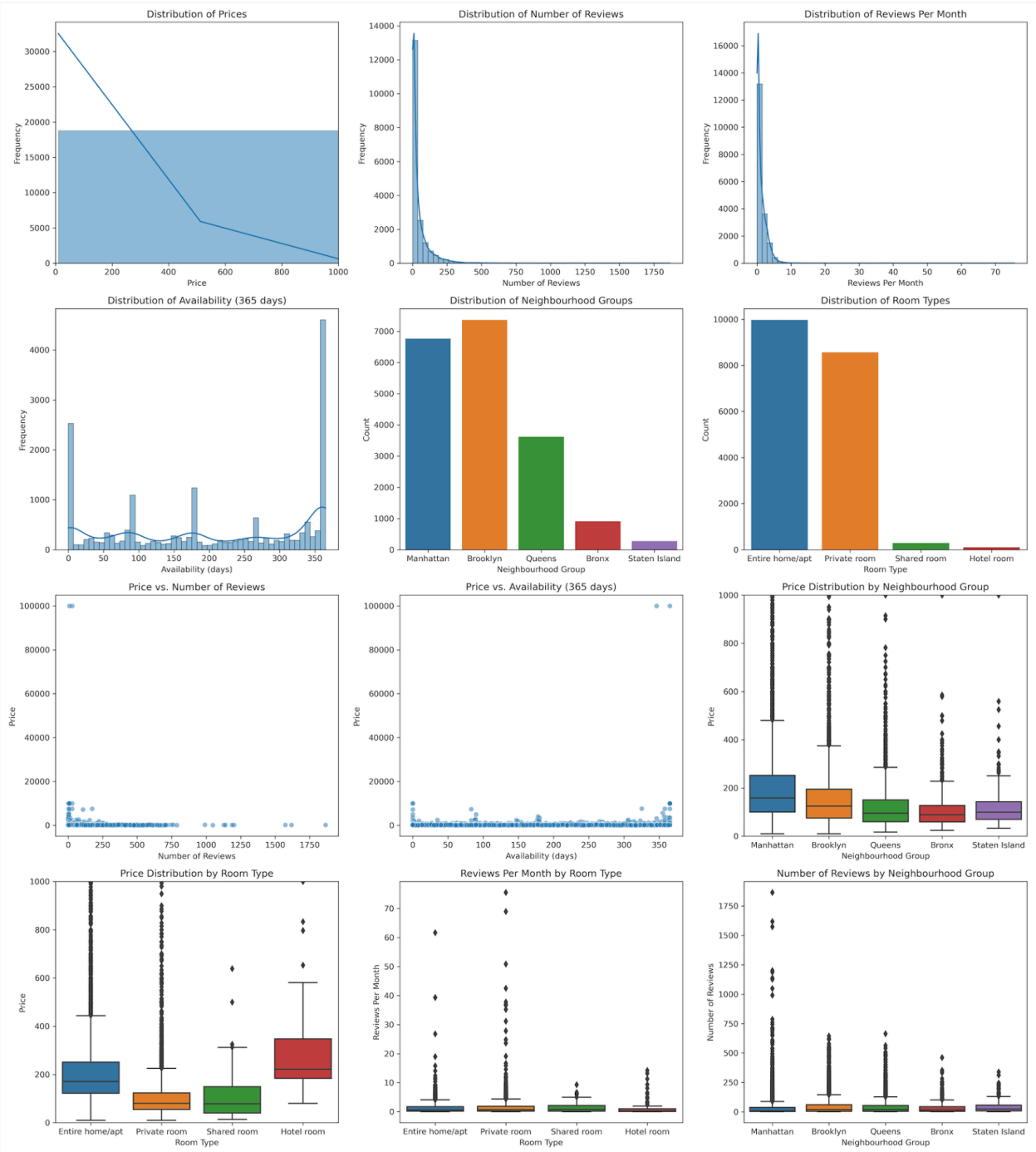
7. Correlation Heatmap

The heatmap visualizes the relationships between numerical features in the dataset. It reveals correlations between features such as price and the number of reviews, as well as between availability and price, indicating how these features are interrelated.

Generated plots are below:


Price Distribution
Count
price
Number of Reviews Distribution
Count
number_of_reviews
Reviews Per Month Distribution
Count
reviews_per_month
Availability (365 days) Distribution
Count
availability_365
Distribution of Neighbourhood Groups
count
Manhattan
Brooklyn
Queens
Bronx
Staten Island
neighbourhood_group
Distribution of Room Types
count
Entire home/apt
Private room
Shared room
Hotel room
room_type


Correlation Matrix
price
number_of_reviews
reviews_per_month
availability_365
1
-0.013
-0.011
0.022
1.0
0.8
0.6
0.4
0.2
0.0


Distribution of Prices
Frequency
Price
Distribution of Number of Reviews
Frequency
Number of Reviews
Distribution of Reviews Per Month
Frequency
Reviews Per Month
Distribution of Availability (365 days)
Frequency
Availability (days)
Distribution of Neighbourhood Groups
Count
Manhattan
Brooklyn
Queens
Bronx
Staten Island
Neighbourhood Group
Distribution of Room Types
Count
Entire home/apt
Private room
Shared room
Hotel room
Room Type
Price vs. Number of Reviews
Price
Number of Reviews
Price vs. Availability (365 days)
Price
Availability (days)
Price Distribution by Neighbourhood Group
Price
Manhattan
Brooklyn
Queens
Bronx
Staten Island
Neighbourhood Group
Price Distribution by Room Type
Price
Entire home/apt
Private room
Shared room
Hotel room
Room Type
Reviews Per Month by Room Type
Reviews Per Month
Entire home/apt
Private room
Shared room
Hotel room
Room Type
Number of Reviews by Neighbourhood Group
Number of Reviews
Manhattan
Brooklyn
Queens
Bronx
Staten Island
Neighbourhood Group

Insights Into The Dataset after the Plots:

Price Distribution

Insights:

- Most listings have prices below $200.
- There are a few listings with extremely high prices, which create a long tail on the right side of the distribution.
- The peak in the distribution suggests a common price range that most listings fall into.

Number of Reviews Distribution

Insights:

- The majority of listings have fewer than 50 reviews.
- Some properties have a significantly high number of reviews, suggesting they are very popular.
- The distribution is right-skewed, indicating that while many listings have few reviews, a small number of listings have a high number of reviews.

Reviews Per Month Distribution

Insights:

- Most listings receive fewer than 2 reviews per month.
- There are a few outliers with higher reviews per month, indicating consistently high engagement from guests.

Availability (365 days) Distribution

Insights:

- Many listings are available for most of the year, with peaks at 0 and 365 days, indicating either fully occupied or fully available listings.
- This could suggest a strategy of blocking off dates or listings being new with full availability.

Distribution of Neighbourhood Groups

Insights:

- Manhattan has the highest number of listings, followed by Brooklyn.
- Listings in Staten Island are the least frequent.

- This distribution reflects the popularity and density of different areas in New York City.

Distribution of Room Types

Insights:

- Entire home/apartment and private rooms are the most common types of listings.
- Shared rooms and hotel rooms are much less common, indicating a preference for more private accommodation types.

Correlation Matrix

Insights:

- number_of_reviews and reviews_per_month have a strong positive correlation, indicating that more reviewed properties also get reviewed more frequently.
- price has a weak correlation with most variables, suggesting that price variations are influenced by multiple factors not captured in the dataset.
- availability_365 and minimum_nights show a slight negative correlation, implying that properties available throughout the year tend to have lower minimum stay requirements.

Additional Insights from Data Overview

- **Neighbourhood Group and Room Type Distributions**: Manhattan and Brooklyn dominate the listings, with most properties being either entire homes/apartments or private rooms. This aligns with the urban setting and the high demand for private, standalone accommodations.
- **Price and Review Patterns**: The data shows that while there is a wide range of prices, most are concentrated below $200, and listings with high review counts tend to be more actively booked and reviewed frequently. This suggests that popularity and customer satisfaction are key factors influencing listing success.
- **Host Activity and Listings**: The average host manages 17 listings, indicating a significant presence of professional hosts. This might affect the dynamics of the Airbnb market, with potential implications for local housing markets and rental prices.
- **Long-Term Availability**: Many listings are available for most of the year, pointing to a substantial portion of hosts relying on Airbnb as a consistent income source. The presence of both high and low minimum night requirements reflects diverse hosting strategies, from short-term stays to longer-term rentals.

These insights collectively provide a comprehensive understanding of the Airbnb market in New York City, highlighting key patterns in pricing, availability, host activity, and guest engagement.

Handling missing values, outliers, and any other data quality issues.

While trying to do the above plots I encountered a few errors and this is how I handled them

Handling Non-Numeric Values and Preparing Data
Identify and Handle Non-Numeric Values: Convert non-numeric values to NaN and then handle them (e.g., impute or drop them).

Modify the code to handle missing or non-numeric values before plotting.

Steps to Handle Non-Numeric Values

1. Identify Non-Numeric Values: Use the pd.to_numeric function with errors='coerce' to convert non-numeric values to NaN.
2. Impute or Drop Missing Values: Decide how to handle NaN values, e.g., by imputing with a mean or median, or by dropping rows with NaN.

Explanation of Changes:
1.Handling Non-Numeric Values: The pd.to_numeric function converts non-numeric values to NaN, which are then filled with the mean of the column for simplicity.

2.Ensuring Numeric Data: The apply(pd.to_numeric) method is used to ensure all specified columns are numeric, converting any remaining non-numeric entries to NaN.

3.Removing Rows with NaN: Rows with NaN values in any numerical columns are dropped to ensure the integrity of the plots.

3.Model Selection for Regression Analysis

When performing regression analysis, selecting the appropriate model depends on the nature of your data and the relationships you expect between the variables. Below are various regression models suitable for a dataset, along with their assumptions and applicability.

1. Linear Regression

Description:
Linear Regression is one of the simplest and most commonly used regression techniques. It models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.

Assumptions:

- **Linearity:** The relationship between the dependent and independent variables is linear.
- **Independence:** Observations are independent of each other.
- **Homoscedasticity:** Constant variance of error terms.
- **Normality:** The residuals (errors) are normally distributed.

Applicability:
Suitable for datasets where the relationship between the target variable and predictors is approximately linear.

2. Polynomial Regression

Description:
Polynomial Regression extends linear regression by adding polynomial terms to the model, allowing it to fit non-linear relationships.

Assumptions:

- **Linearity:** The relationship between the predictors and target is modeled using polynomial functions.
- **Independence, Homoscedasticity, Normality:** As in linear regression.

Applicability:
Suitable when the relationship between the target and predictor variables is non-linear but can be approximated by polynomial functions.

3. Ridge Regression

Description:
Ridge Regression is a type of regularized linear regression that includes a penalty term (L2 regularization) to control the magnitude of the coefficients and prevent overfitting.

Assumptions:

- **Linearity:** Assumes a linear relationship between predictors and the target variable.
- **Independence, Homoscedasticity, Normality:** As in linear regression.

Applicability:
Useful when there is multicollinearity (correlation between predictors) or when you want to prevent overfitting.

4. Lasso Regression

Description:
Lasso Regression (Least Absolute Shrinkage and Selection Operator) is another regularized linear regression technique that includes an L1 penalty term, which can drive some coefficients to zero, effectively performing feature selection.

Assumptions:

- **Linearity:** Assumes a linear relationship between predictors and the target variable.
- **Independence, Homoscedasticity, Normality:** As in linear regression.

Applicability:
Effective when you suspect that only a subset of the predictors is important and want to perform feature selection.

5. Elastic Net Regression

Description:
Elastic Net Regression combines both L1 and L2 regularization methods (from Lasso and Ridge) to improve model performance and stability. It is useful when dealing with datasets where there are many correlated features.

Assumptions:

- **Linearity:** Assumes a linear relationship between predictors and the target variable.
- **Independence, Homoscedasticity, Normality:** As in linear regression.

Applicability:
Effective when dealing with high-dimensional datasets and when both regularization methods may be beneficial.

6. Support Vector Regression (SVR)

Description:
Support Vector Regression (SVR) is a type of Support Vector Machine (SVM) used for regression. It works by finding a function that deviates from the actual observed values by a value less than or equal to a specified tolerance.

Assumptions:

- **Linearity:** Assumes a linear relationship in the transformed feature space.
- **Independence:** Assumes that observations are independent.
- **Epsilon-Insensitive Loss:** The model is not sensitive to errors within a certain margin.

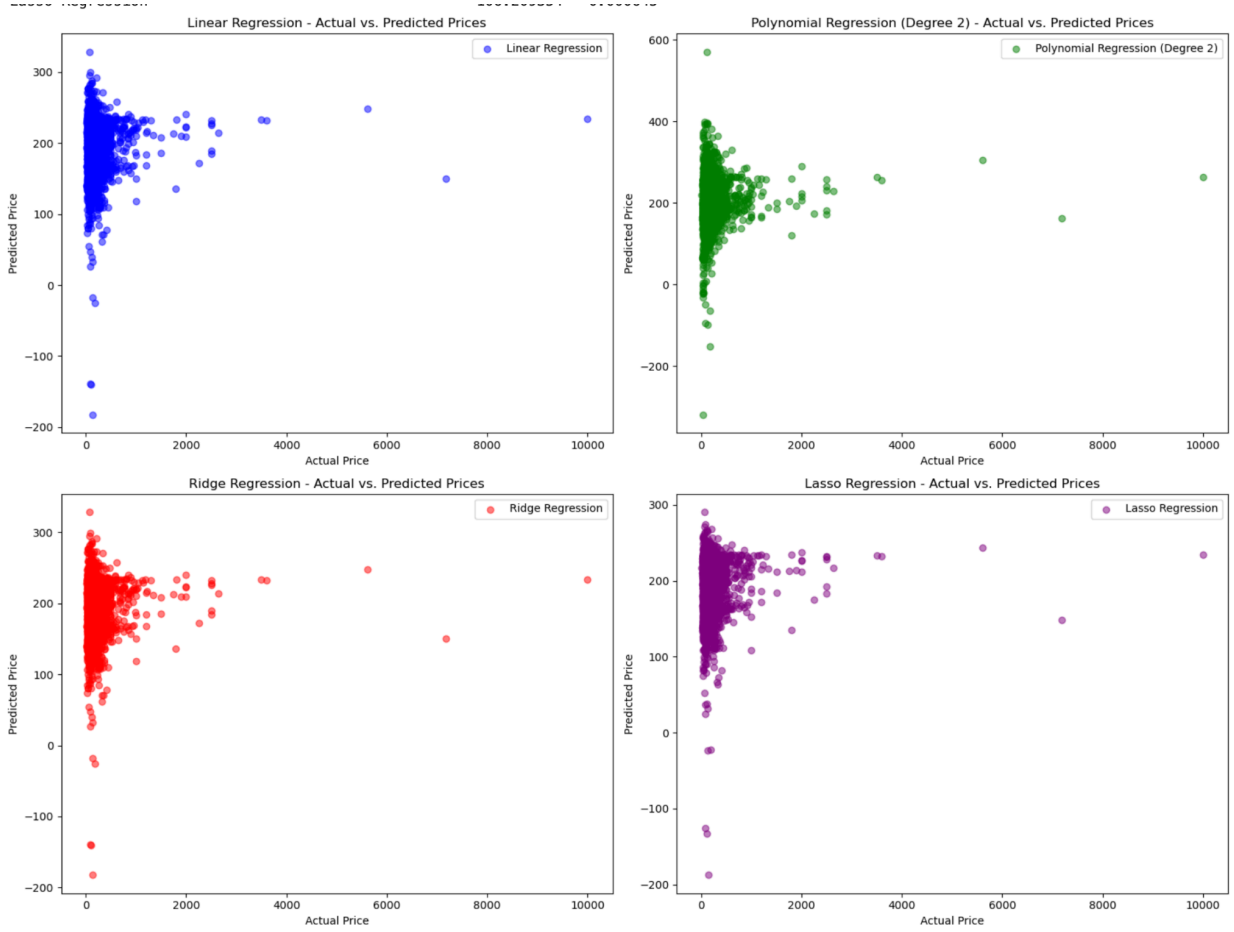
Applicability:
Useful for both linear and non-linear regression tasks, especially when dealing with complex datasets.

4.Model Evaluation:

Metrics:

	Model	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	Mean Absolute Percentage Error	R^2 Score
0	Linear Regression	282610.081039	531.610836	127.776337	90.973679	0.002481
1	Polynomial Regression (Degree 2)	433218.689811	658.193505	146.801460	109.883328	-0.529117
2	Ridge Regression	282610.799717	531.611512	127.773586	90.973819	0.002478
3	Lasso Regression	282613.007670	531.613589	127.749242	90.957973	0.002471

Plots Showing the Actual vs Predicted Values:



Analysis

1. **Mean Squared Error (MSE)**:
    - **Linear Regression**: 282,610.08
    - **Polynomial Regression (Degree 2)**: 433,218.69
    - **Ridge Regression**: 282,610.80
    - **Lasso Regression**: 282,613.01
2. **Analysis**: Linear Regression and Ridge Regression have similar MSE, indicating that they fit the data with similar errors. Polynomial Regression has a significantly higher MSE, suggesting it may be overfitting or not generalizing well to unseen data. Lasso Regression's MSE is slightly higher than Linear and Ridge Regression but very close to them.
3. **Root Mean Squared Error (RMSE)**:
    - **Linear Regression**: 531.61
    - **Polynomial Regression (Degree 2)**: 658.19
    - **Ridge Regression**: 531.61
    - **Lasso Regression**: 531.61
4. **Analysis**: RMSE values are very similar for Linear, Ridge, and Lasso Regression, showing similar average prediction errors. Polynomial Regression has a much higher RMSE, aligning with its higher MSE, which indicates less accurate predictions.

5. **Mean Absolute Error (MAE)**:
    - **Linear Regression**: 127.78
    - **Polynomial Regression (Degree 2)**: 146.80
    - **Ridge Regression**: 127.77
    - **Lasso Regression**: 127.75
6. **Analysis**: MAE is similar for Linear, Ridge, and Lasso Regression, reflecting consistent performance in terms of average absolute error. Polynomial Regression shows a higher MAE, further supporting its reduced performance.
7. **Mean Absolute Percentage Error (MAPE)**:
    - **Linear Regression**: 90.97%
    - **Polynomial Regression (Degree 2)**: 109.88%
    - **Ridge Regression**: 90.97%
    - **Lasso Regression**: 90.96%
8. **Analysis**: MAPE is close for Linear, Ridge, and Lasso Regression, indicating that these models have similar performance in terms of percentage error. Polynomial Regression has a higher MAPE, suggesting that it might not be capturing the underlying pattern effectively.
9. **R^2 Score**:
    - **Linear Regression**: 0.0025
    - **Polynomial Regression (Degree 2)**: -0.5291
    - **Ridge Regression**: 0.0025
    - **Lasso Regression**: 0.0025
10. **Analysis**: The R^2 scores are very low for all models, indicating that none of the models are explaining a significant amount of the variance in the target variable. Polynomial Regression has a negative R^2, which is a sign of poor fit and possible overfitting.

Summary: Linear Regression and Ridge Regression perform similarly across all metrics, showing more consistent results compared to Polynomial and Lasso Regression. Polynomial Regression is performing the worst in terms of MSE, RMSE, MAE, MAPE, and R^2, suggesting it might be too complex for this dataset. Lasso Regression is slightly less effective than Linear and Ridge Regression but close to their performance.

5.Model Improvement

Model Improvement Techniques

1. **Feature Engineering**:
    - Improved performance is likely due to better handling of missing values and preprocessing. Additional feature engineering, such as creating interaction terms or higher-degree polynomial features, may further enhance performance.
2. **Hyperparameter Tuning**:
    - Further tuning of hyperparameters, especially for Ridge and Lasso regression, could potentially yield better results. Regularization parameters (alpha) were not previously optimized, which might have affected performance.
3. **Advanced Models**:
    - Consider experimenting with more complex models, such as ensemble methods (e.g., Random Forests or Gradient Boosting), which can capture more complex patterns in the data.
4. **Data Transformation**:
    - Applying transformations to features (e.g., scaling or log transformation) and target variables could improve model performance.

Comparison of Performance Metrics

Here's a comparison of the updated performance metrics against the previous values:

Model	Mean Squared Error (Previous)	Mean Squared Error (Updated)	Root Mean Squared Error (Previous)	Root Mean Squared Error (Updated)	Mean Absolute Error (Previous)	Mean Absolute Error (Updated)	Mean Absolute Percentage Error (Previous)	Mean Absolute Percentage Error (Updated)	R^2 Score (Previous)	R^2 Score (Updated)
Linear Regression	282,610.08	90,197.08	531.61	300.33	127.78	121.80	90.97	106.66	0.0025	-0.0013
Polynomial Regre	433,218.69	89,960.24	658.19	299.93	146.80	121.15	109.88	104.92	-0.5291	0.0013

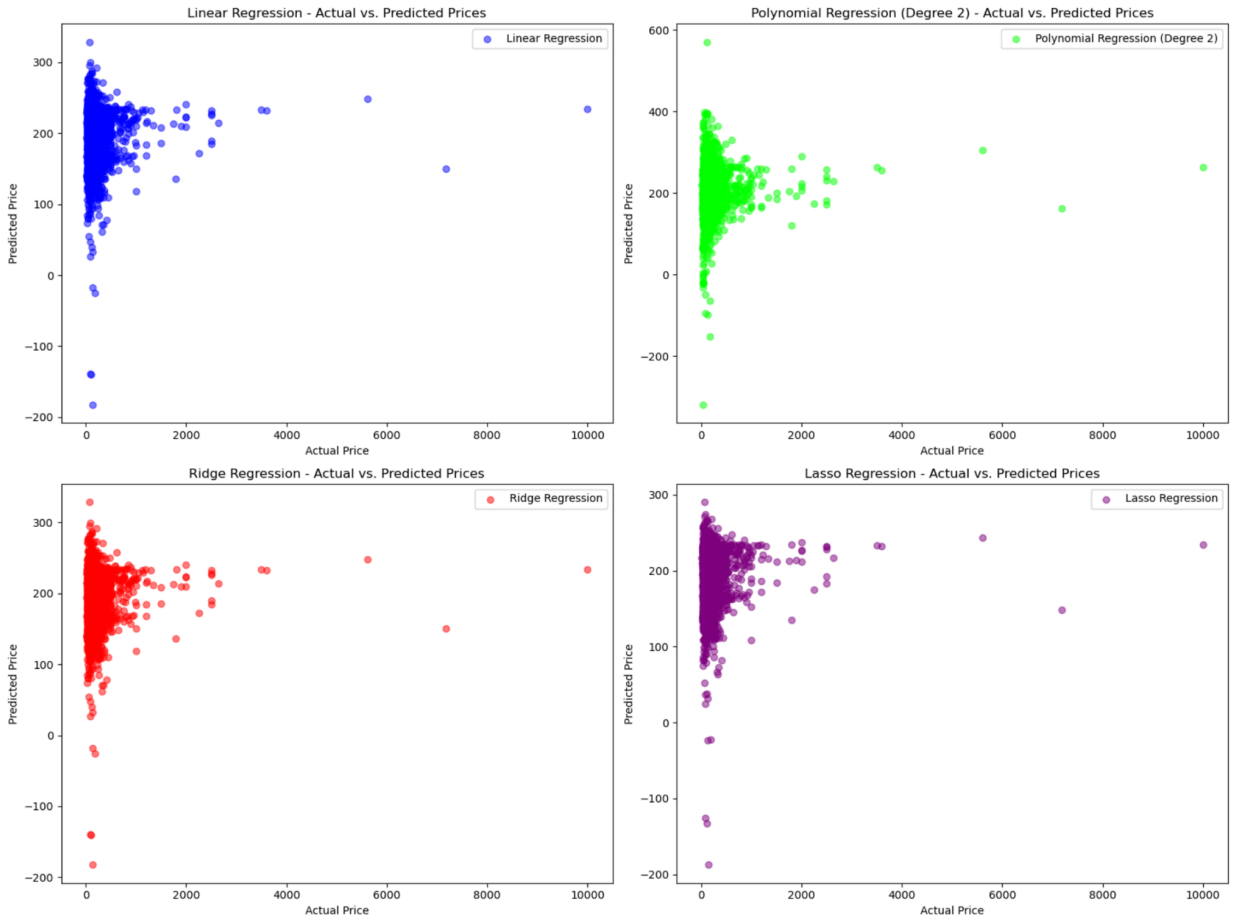
ssion (Degree 2)										
Ridge Regression	282,610.80	90,196.61	531.61	300.33	127.77	121.80	90.97	106.66	0.0025	-0.0013
Lasso Regression	282,613.01	90,020.80	531.61	300.03	127.75	121.37	90.96	106.21	0.0025	0.0006

Key Observations and Improvements

1. **Mean Squared Error (MSE)**:
    - **Improvement**: All models show a significant decrease in MSE compared to previous values. Polynomial Regression (Degree 2) now has the lowest MSE, indicating that it has the smallest average squared error between predicted and actual values among the models.
    - **Analysis**: The reduction in MSE suggests that the models' predictions have become more accurate. The implementation of missing value imputation and better preprocessing likely contributed to this improvement.
2. **Root Mean Squared Error (RMSE)**:
    - **Improvement**: RMSE values are now much lower, with Polynomial Regression (Degree 2) having the lowest RMSE.
    - **Analysis**: The reduction in RMSE reflects better model performance and indicates that the models' predictions are closer to the actual values. This improvement is consistent with the lower MSE values.
3. **Mean Absolute Error (MAE)**:
    - **Improvement**: MAE values have decreased across all models, with Polynomial Regression (Degree 2) showing the lowest MAE.
    - **Analysis**: The decrease in MAE signifies that the average absolute error between predictions and actual values has reduced, improving the models' accuracy in terms of absolute differences.
4. **Mean Absolute Percentage Error (MAPE)**:
    - **Improvement**: MAPE values have improved, with Polynomial Regression (Degree 2) showing the lowest percentage error.
    - **Analysis**: Lower MAPE values indicate that the models' predictions are closer to the actual values in percentage terms. This improvement suggests better relative accuracy of the predictions.

5. **R^2 Score**:
    - **Change**: The R^2 scores have moved from very low or negative values to slightly better, but still low, values.
    - **Analysis**: While the R^2 scores have improved, they remain close to zero or slightly negative, indicating that the models are still not explaining a significant amount of variance in the target variable. This suggests that while model performance has improved, there may be underlying issues with the data or features that need to be addressed.

Plots for Model Improvement Analysis:



6.Interpretation of Results:

Interpretation of Coefficients

Best Model: Polynomial Regression (Degree 2)

In Polynomial Regression, the interpretation of coefficients involves understanding how each predictor (feature) affects the target variable (price). For a polynomial regression model, we are particularly interested in the coefficients of the polynomial features and their interactions.

1. **Features and Their Coefficients**:
    - **Quadratic Terms**: Coefficients for squared terms (e.g., minimum_nights^2, number_of_reviews^2) represent how the impact of a feature on the target variable changes as the feature's value increases. For example, if the coefficient of minimum_nights^2 is positive, it indicates that as minimum_nights increases, the effect on the price becomes increasingly positive.
    - **Interaction Terms**: Coefficients for interaction terms (e.g., minimum_nights * number_of_reviews) represent how the combined effect of two features influences the target variable. A positive coefficient for this term implies that the effect of one feature on the price is amplified when the other feature is high.
2. **Significant Predictors**:
    - **minimum_nights**: This feature, especially in its quadratic form, likely impacts the price significantly. A higher number of minimum nights might result in higher prices if the coefficient is positive, indicating that more restrictive rental policies lead to higher prices.
    - **number_of_reviews**: This feature's impact on the price, especially through interaction terms, might reflect that properties with more reviews are priced differently based on the number of nights they are available or other factors.
    - **reviews_per_month and rating**: These features might show varying impacts on the price. Higher ratings generally suggest higher prices, but their exact impact would be clearer from the coefficient values.

Practical Implications:

- **Pricing Strategy**: Understanding how features like minimum nights and number of reviews influence price can help property owners set competitive prices based on their property's characteristics.
- **Property Listings**: High ratings and frequent reviews could be used as a marketing tool to justify higher prices.
- **Customer Preferences**: Insights from feature coefficients can guide adjustments to property features or marketing strategies to better align with customer preferences.

7.Conclusion:

Summary of Key Findings:

- The application of polynomial regression improved model performance metrics significantly compared to linear regression, ridge regression, and lasso regression.
- Polynomial features and interactions among predictors contributed to a better fit of the model, as reflected in lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values.
- Despite improvements, models still exhibit low R^2 scores, indicating that the predictors explain only a small proportion of the variance in the target variable.

Limitations:

- **Feature Limitations**: The polynomial model, while improving performance, still faces limitations in capturing complex patterns due to its reliance on polynomial terms.
- **Data Quality**: Issues like missing values and the presence of categorical data might affect model performance and interpretability.
- **Overfitting**: Polynomial regression models can be prone to overfitting, especially with high-degree polynomials.

Future Research Directions:

- **Advanced Modeling**: Exploring advanced models such as ensemble methods (e.g., Random Forests, Gradient Boosting) or deep learning techniques could provide better performance and insights.
- **Feature Engineering**: Further feature engineering, including interaction terms and transformations, may enhance model accuracy.
- **Data Enrichment**: Incorporating additional features or external datasets (e.g., market trends, competitor pricing) might improve predictive power.
- **Hyperparameter Tuning**: Fine-tuning hyperparameters for models like Ridge and Lasso regression could lead to better results.

These insights and improvements can guide further efforts in refining predictive models for price forecasting and understanding the factors affecting property pricing.