Analytical Report:

Understanding and predicting Airbnb prices in New York City

Executive Summary

This report aims to investigate characteristics influencing Airbnb listing prices in New York City (NYC) and to develop a predictive model. Prices are heavily influenced by location and room type (private/shared/entire apartment). Minimum stay per booking, number of listings per host, and total availability over the year has a weak positive relationship with prices, whilst total and average number of reviews per month show a weak negative relationship. Model performance is modest (explaining only 57% of price variability), but it provides useful information for both Airbnb management and hosts to guide pricing and identify ways to improve revenue.

Table of contents

1. Background and Objectives

Airbnb is a major player in the online marketplace for short- and long-term homestays and experiences. Understanding and predicting listing prices can add significant value to the business, namely by:

- advising hosts on competitive pricing;
- implementing dynamic pricing strategies to maximize revenue;
- ensuring fair pricing in line with market;
- benchmarking listings against competitors and adjusting strategy to increase market share.

The objective of this analysis is to identify factors that influence price and provide Airbnb with a price prediction model to support the strategic and operational goals mentioned. The dataset used was accessed through Kaggle and describes Airbnb listings in NYC (Gomodonov, 2019). The complete statistical code and report which supports this analysis is provided as an annex (Laudriec et al., 2024).

2. Exploratory data analysis and data processing

An initial exploratory data analysis (EDA) process was employed to gather a general grasp of data structure and contents, to investigate distributions of individual variables, and to assess relationships among different variables.

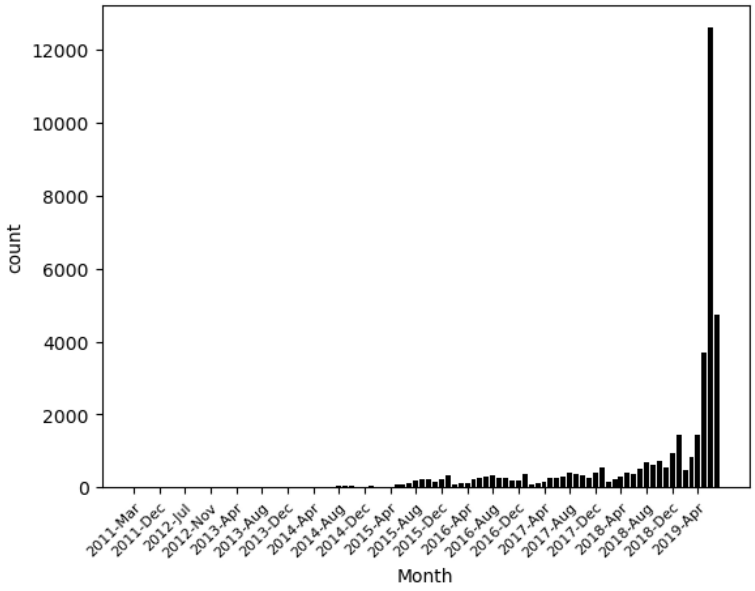
2.1 General data exploration

Raw data was well structured and labelled, and generally clean, facilitating subsequent exploration and transformations. There were 48,895 observations and 16 columns (Table 1), including listings from March 2011 to April 2019 (based on the date of last review; Figure 1).

Table 1 - Summary of variables included in the dataset

Variable	Description	Type
id	Individual listing id	Integer
name	Individual listing name	String
host_id	Individual host id	Integer
host_name	Individual host name	String
neighbourhood_group	High-level neighbourhood group	String
neighbourhood	Specific neighbourhood	String
latitude	Latitude	Numeric
longitude	Longitude	Numeric
room_type	Room type	String
price	Listing price in dollars	Integer
minimum_nights	Minimum number of nights per booking	Integer
number_of_reviews	Total number of reviews for specific listing	Integer
last_review	Date of last review for specific listing	String
reviews_per_month	Average number of monthly reviews for specific listing	Numeric
calculated_host_listings_count	Total number of listings per host	Integer
availability_365	Number of days listing is available over the year	Integer

Figure 1 - Monthly counts for date of last review



Most variables do not have any missing values, except name (0.03% missing values), host_name (0.04% missing values), reviews_per_month (20.6% missing values), and last_review (20.6% missing values) - Figure 2.

Figure 2 - Missing values distribution

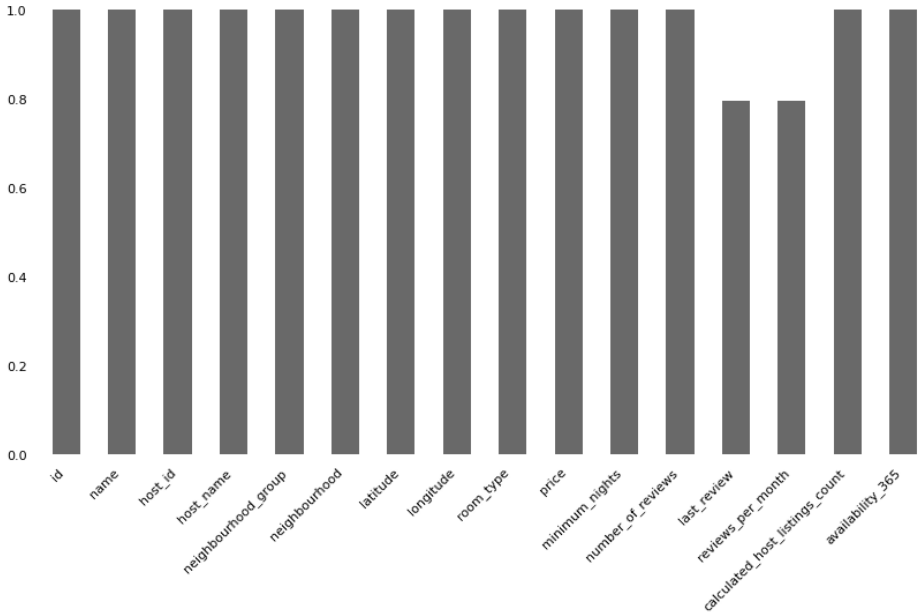


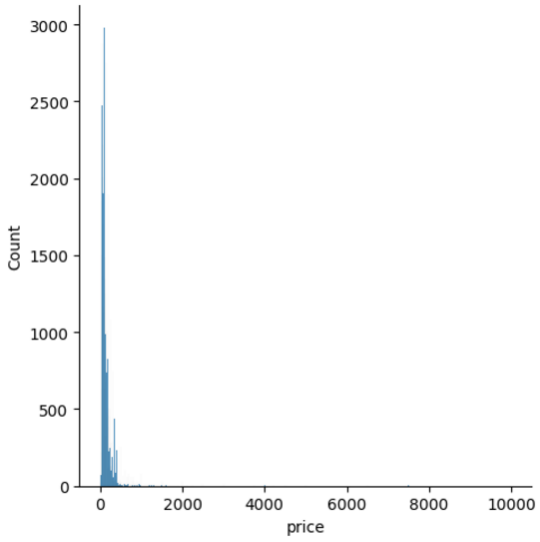
Figure legend: Vertical axis shows the proportion (0-1) of rows with complete values for each variable (depicted along the horizontal axis).

Table 2 presents basic descriptive statistics calculated for numerical variables (excluding latitude/longitude). Many of these exhibited heavily right-skewed distributions, including the target variable price - Figure 3.

Table 2 - Descriptive statistics for numerical variables

	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
Mean	152.7	7.0	23.3	1.4	7.1	112.8
Standard deviation	240.2	20.5	44.6	1.7	33.0	131.6
Minimum	0	1	0	0	1	0
Quartile 1	69	1	1	0.2	1	0
Median	106	3	5	0.7	1	45
Quartile 3	175	5	24	2	2	227
Maximum	10000	1250	629	58.5	327.0	365

Figure 3 - Histogram of price



Based on this initial exploration, candidate variables for the model were selected as detailed in Table 3.

Table 3 - Candidate dependent and independent model variables

Dependent variable
Price
Independent variables
host_id
neighbourhood_group
neighbourhood
room_type
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_365

2.2 Relationships between price and candidate predictor variables

Associations between price and candidate numerical variables were generally weak (Figures 4-5).

Figure 4 - Correlation matrix for numerical variables

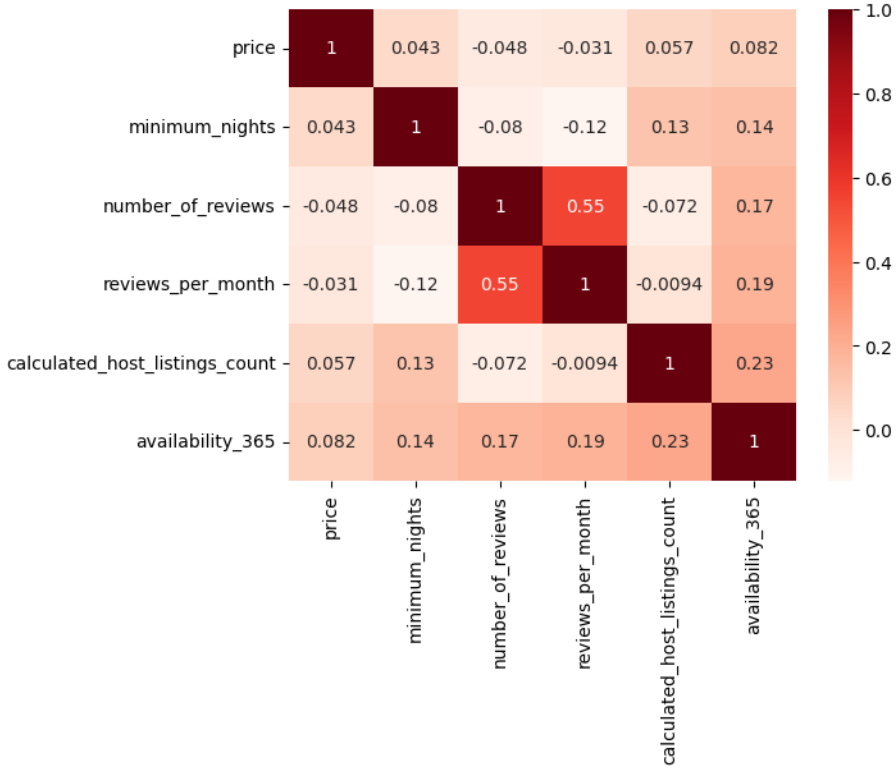
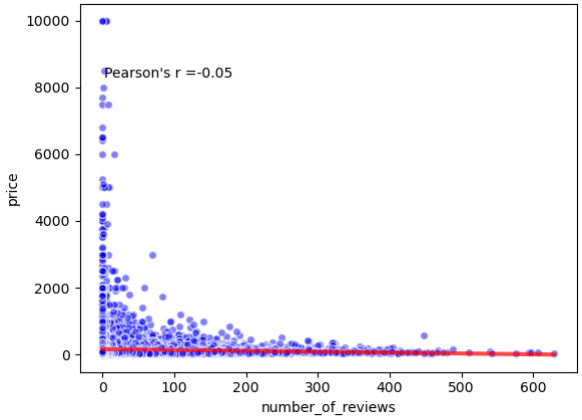


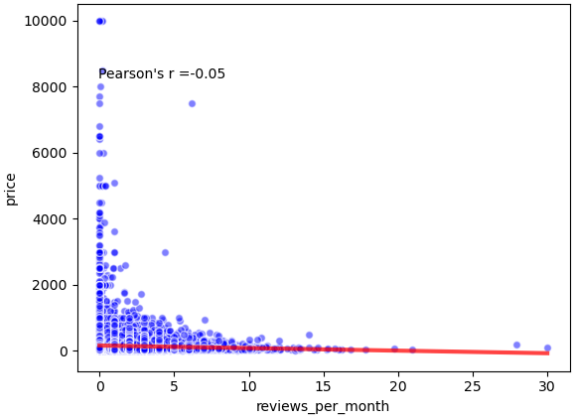
Figure legend: Numerical estimates shown are Pearson correlation coefficients.

Figure 5 - Association of numerical predictor variables with price

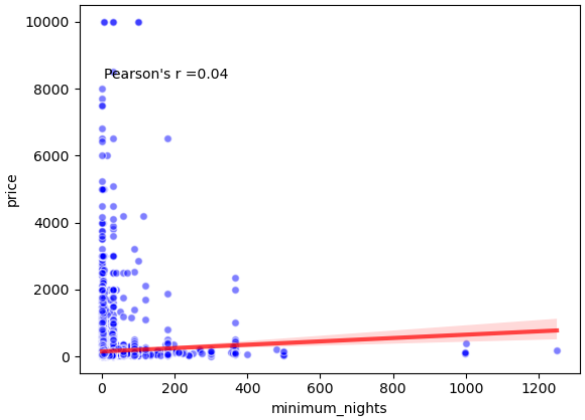
A - Number of reviews



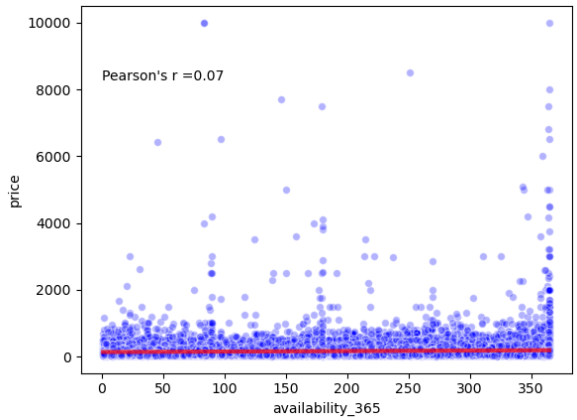
B - Reviews per month



C - Minimum nights per booking



D - Availability over 365 days



E - Total listings per host

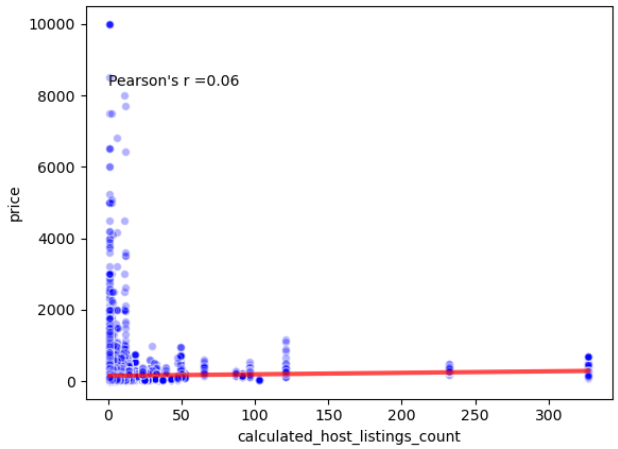


Figure legend: Each panel shows a scatter plot for a different numerical variable (horizontal axis) versus price (vertical axis). Pearson's correlation coefficients are presented, along with fitted linear regression lines and associated 95% confidence intervals.

Categorical variables showed more noticeable impacts on price, with entire homes generally listed for higher prices than private or shared rooms, and the latter having cheaper prices on average (Figure 6).

Figure 6 - Price distributions by room type

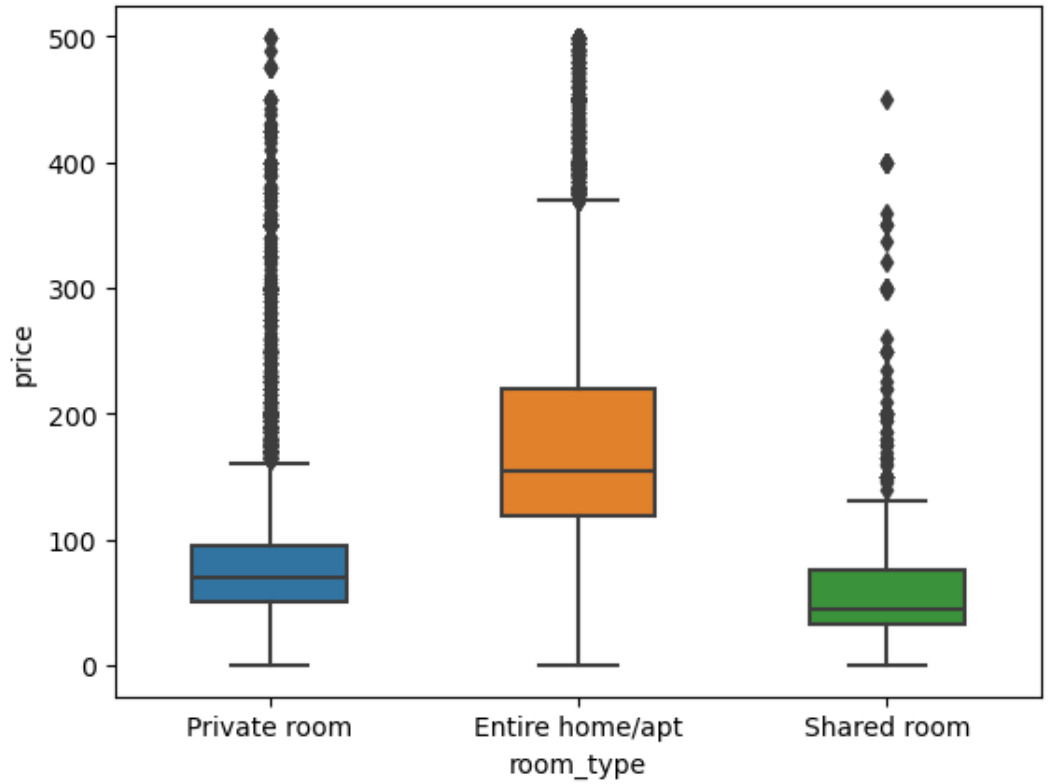


Figure legend: Price distributions shown using boxplots. Listing prices restricted to values <500$ to facilitate visualisation. Each box extends between quartile 1 (Q1) and quartile 3 (Q3), with the horizontal bar in the middle showing median. Box whiskers span to 1.5 times the interquartile range (Q1-Q3) below Q1 and above Q3. Data points outside these margins are shown as the individual dots.

Prices also varied substantially by location, as captured by individual neighbourhood and neighbourhood groups. Listings in Brooklyn and Manhattan were generally more expensive than those in Queens, Staten Island, or Bronx (all with similar distributions; Figure 7), with large heterogeneity across different neighbourhoods within the same group (Figure 8).

Figure 7 - Price distributions by neighbourhood group

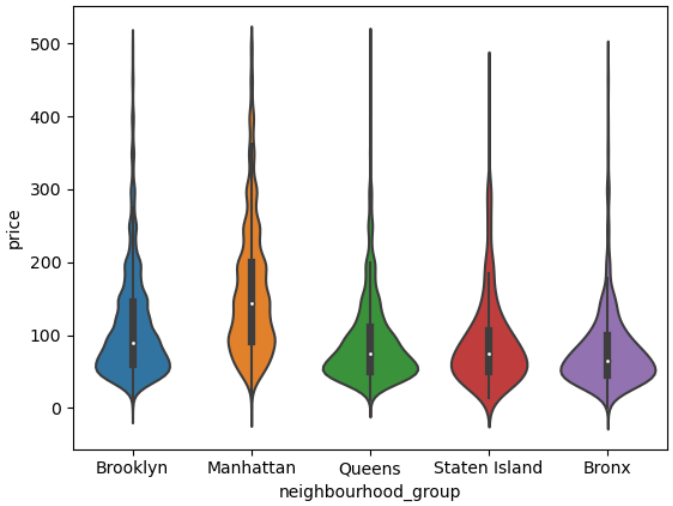


Figure legend: Price distributions shown using violin plots. Each violin shows a rotated and mirrored probability distribution (generated from kernel density estimates). A boxplot is included within each violin, with each box extending between quartile 1 (Q1) and quartile 3 (Q3), and the horizontal bar in the middle showing median. Box whiskers span to 1.5 times the interquartile range (Q1-Q3) below Q1 and above Q3.

Figure 8 - Price distributions by neighbourhood

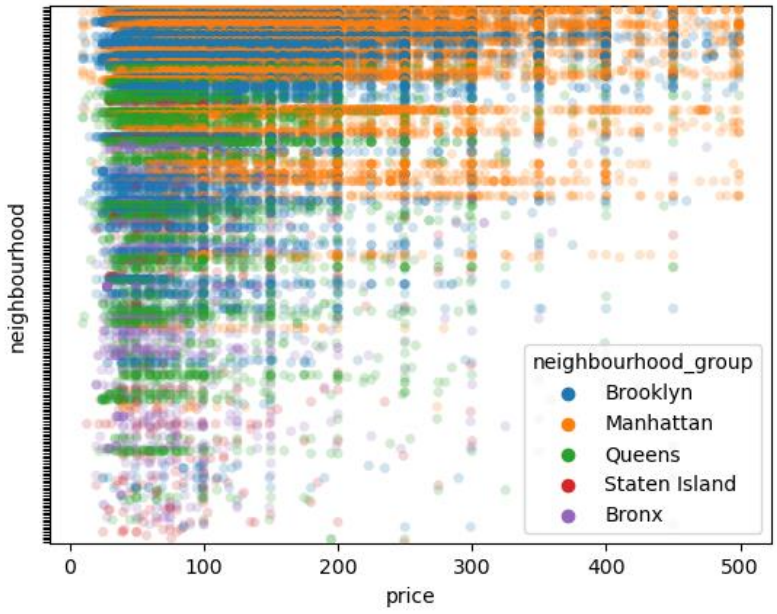


Figure legend: Each listing is shown with a single dot, with the vertical axis showing individual neighbourhoods (names not listed) and the horizontal axis showing price. Dots are coloured by neighbourhood groups.

2.3 Data processing

Following from the EDA, several processing steps were then applied to prepare data for modelling, as detailed in Table 4. These included removing or transforming values likely to be incorrect, identifying and handling outliers (Dash et al., 2023), and deriving a new variable.

Table 4 - Data processing steps

Variable	Processing steps	Justification
price	Remove observations with price 0	Listings with price 0 are likely to be incorrect; including these observations in the model was likely to hurt performance, and being able to predict price of 0 has little business value
	Apply 1st/99th percentile approach to identify outliers	Outlier detection approach which yielded the best balance between improved distribution discrimination and proportion of observations removed
	Winsorising outliers	Outliers kept in the model (versus trimming), but price values for those observations imputed with either 1st percentile (if below) or 99th percentile (if above)
Minimum nights	Apply 1/99 percentile approach to identify outliers, but only with 99 percentile (values below 1st percentile are reasonable ones)	1st percentile included minimum 1 night per booking, which is a reasonable and very frequent value; therefore considering this an outlier would become problematic.
	Winsorising outliers	Outliers kept in the model (versus trimming), but price values for those observations imputed with 99th percentile.
reviews_per_month	Assign one observation with average ~57 to 30	Average reviews per month >30 would indicate an average of more than one distinct review per day, which seemed implausible.
	Replace missing values with 0	0 reviews per month captures the underlying information (missing data) and allows these observations to be included in the model
availability_binary	Created from availability_365 (0 if availability_365 = 0, 1 if availability_365 >0)	Many listings have availability 0, a value which contains fundamentally different information from all others. New variable created to capture this information.

3. Model development, evaluation, and interpretation

A log10 transformation was applied to price due to non-linear relationships between the predictor variables and price, as indicated by Figure 5. This transformation helps to linearize the relationship, stabilize variance, and address skewness in the data, which improves model conformity to linear regression assumptions (Kirkwood and Sterne, 2003; Roustaei, 2024).

The model was fit using a train-test split (70% of data used for training and 30% for testing), and evaluated using R-squared and Root Mean Squared Error (RMSE). The model showed moderate predictive performance, evidenced by an R-squared of 0.568 (i.e. the model accounts for 57% of the variance within actual prices), and RMSE of 1.55$ (i.e. average error between predicted and actual price). The model generally underperformed for listings with higher prices (Figure 9), suggesting that the price of those listings is determined to a significant extent by information not included in the model.

Figure 9 - Actual vs predicted price (log10-transformed)

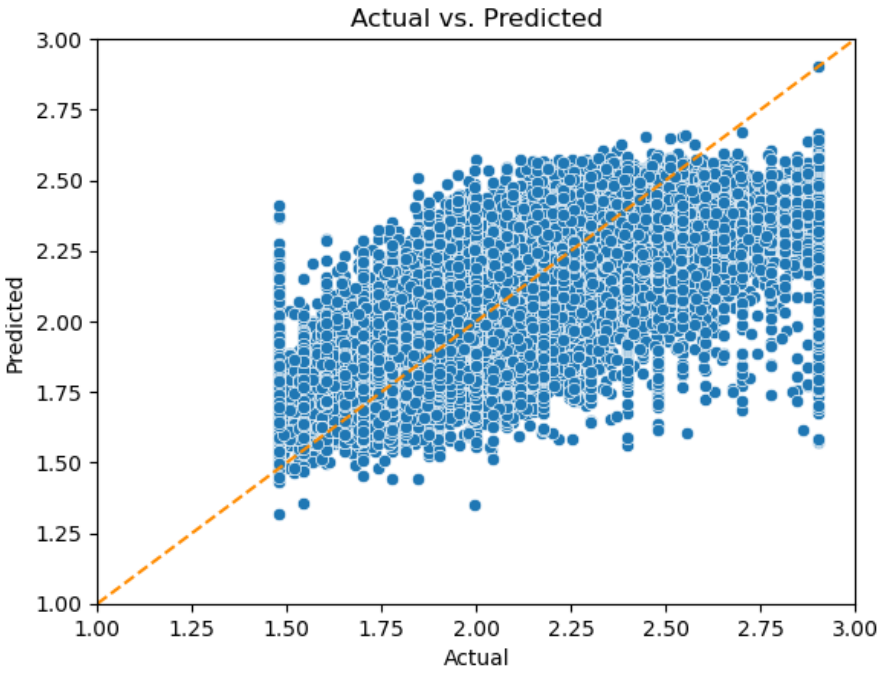


Figure legend: Actual prices are shown on the horizontal axis, and predicted prices on the vertical axis (both log10-transformed, i.e. a value of 2 represents 10^2 = 100). Observations along the diagonal line indicate perfect fit. Note the increasing number of observations lying far from the diagonal line as actual price increases, indicating model underperformance for higher price values.

The most important predictors seemed to be those related to location (both broad neighbourhood groups and individual neighbourhoods) and room type, as identified by the regression coefficients (Table 5).

Table 5 - Regression coefficients for predictor variables

Feature	Regression coefficient	p-value
neighbourhood_group_Manhattan	0.3126	<0.001
neighbourhood_group_Brooklyn	0.1465	<0.001
neighbourhood_group_Queens	0.0981	<0.001
available_binary	0.0426	<0.001
neighbourhood_group_Staten Island	0.0420	<0.001
availability_365	0.0003	<0.001
host_id	0.0000	<0.001
calculated_host_listings_count	0.0000	<0.001
number_of_reviews	-0.0002	<0.001
minimum_nights	-0.0048	<0.001
reviews_per_month	-0.0111	<0.001
room_type_Private room	-0.3127	<0.001
room_type_Shared room	-0.4718	<0.001

Table legend: features for individual neighbourhoods removed from this table due to the larger number of groups (>200)

In particular, listings in Manhattan, Brookly, Queens, and Staten Island were on average 99.9$, 38.5$, 20$ and 11.6$ more expensive respectively than those in the Bronx, while private rooms and shared rooms were on average 94.8$ and 123.9$ cheaper respectively versus an entire apartment. Higher availability was generally associated with higher prices (average 13.8$ increase for currently available listings versus those not available, and 0.13$ increase for each additional day available). Other features were negatively associated with price, namely:

- a small (<0.01$) decrease in price for each additional total number of listings per host;
- a 0.12$ decrease for each additional total number of reviews;
- a 1.6$ decrease for each additional minimum number of nights per booking; and
- a 5.2$ decrease for each additional average number of reviews per month

4. Recommendations

This analysis indicates that potential hosts may increase revenue by:

- purchasing properties in highly sought after locations (which can be investigated in detail from our model);
- converting shared or private room listings to entire apartments;
- increasing property availability; and
- reducing the minimum stay per booking.

The model can also be used to provide an estimate of the average price for properties within the same characteristics to guide pricing. Finally, Airbnb can use this model to identify over or underperforming listings versus similar ones (based on model residuals), and explore those patterns to support underperforming hosts or further investigate characteristics of top-performers.

5. Strengths and limitations

This analysis used well-structured data from >48,000 listings spanning a large temporal window. Predictive model performance was moderate, but similar to that obtained by other analysts, including with more advanced techniques (Turgut, 2019; Patel, 2020). Model performance could be further improved with additional data collection (more features or more observations) and more robust modelling approaches. In particular, temporal data for actual booking date should allow assessments of important factors including market trends, inflation, and seasonality.

6. Conclusion

This analysis identified key factors influencing Airbnb prices in NYC, such as location and room type. Insights can guide hosts in optimizing pricing and availability, while helping Airbnb refine dynamic pricing strategies for increased revenue.

7. References

Dash, Ch.S.K., Behera, A.K., Dehuri, S. & Ghosh, A. (2023) ‘An outliers detection and elimination framework in classification task of data mining’, *Decision Analytics Journal*, 6, p. 100164. Available from: https://doi.org/10.1016/j.dajour.2023.100164.

Gomodonov, D. (2019) ‘New York City Airbnb Open Data’. Kaggle. Available from: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data (Accessed: 27 November 2024).

Kirkwood, B.R. & Sterne, J.A.C. (2003) *Essential Medical Statistics*. 2nd edn. Malden, Massachusetts, USA: Blackwell Science.

Laudriec, C., Papachristou, G., Amorim, G. & Panashe, M. (2024) *Airbnb group project*. GitHib. Available from: https://github.com/christgithub/airbnb-group-project/blob/main/EDA/data-analysis.ipynb.

Patel, Y. (2020) *Airbnb price prediction*. Kaggle. Available from: https://www.kaggle.com/code/yashvi/airbnb-price-prediction.

Roustaei, N. (2024) ‘Application and interpretation of linear-regression analysis’, *Medical Hypothesis, Discovery & Innovation Ophthalmology Journal*, 13(3), pp. 151–159. Available from: https://doi.org/10.51329/mehdiophthal1506.

Turgut, D. (2019) *Airbnb NYC Price Prediction*. Kaggle. Available from: https://www.kaggle.com/code/duygut/airbnb-nyc-price-prediction.