NEW YORK CITY AIRBNB: A DATA DRIVEN ANALYSIS OF PRICING LOCATION, AND MARKET SEGMENTATION

By Kanan Akparov, Ade Putra Tio Aldino,

Dwane Lay and Nelson Okolie

Machine Learning July 2025 A

Dr Steph Paladini

University of Essex

Colchester, England

September 8, 2025

New York City Airbnb Analysis

Internet services like Airbnb have disrupted the hotel market by giving users greater control over pricing, reviews, and location options (Sharma & Gupta, 2021). Today, Airbnb has become one of the most utilised accommodation services in the world and dominates the market in New York (Jiao & Bai, 2019). This report analyses pricing, location, availability, guest reviews and property types. Insights from this study could help Airbnb refine pricing strategies and improve guest relations.

Business Analytic Questions

What is the most expensive Airbnb booking in New York City, and what are the social, economic, geographical, and property-specific reasons for these high prices? Is there a link between the price of a listing and how many reviews it receives? What can we learn about how the market and users are changing from the relationship described above?

Data Analysis - Data Preprocessing

The dataset contained 48,895 rows and 16 columns. During preprocessing, we removed listings without reviews or prices, or those that included extreme outliers. This left us with 36,753 records, which we deemed the sample dataset.

Data Analysis - Exploratory Data Analysis (EDA)

Our data for several NYC neighbourhoods, along with information on real estate distribution patterns for 2023, shows Manhattan had the most available listings, followed closely by Brooklyn. Staten Island and the Bronx had fewer available listings (see Figures 1 and 2).

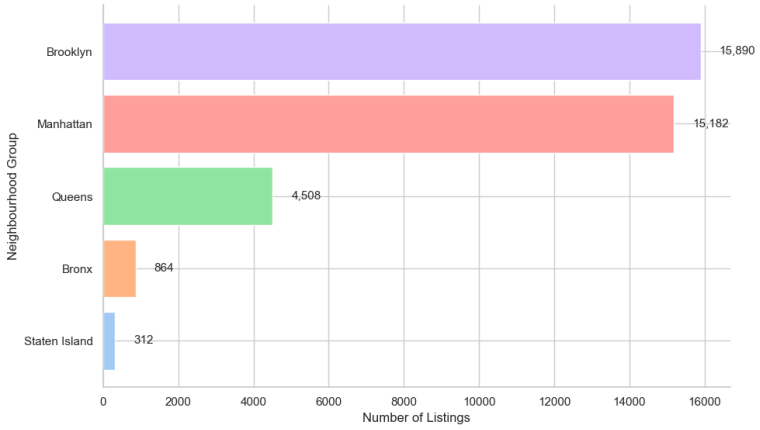

Figure 1 - Listings by Neighbourhood

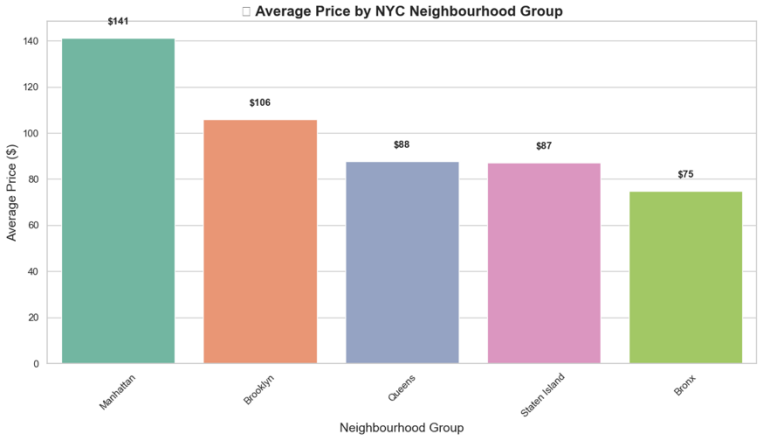

Figure 2 - Prices by Neighbourhood

Brooklyn and Manhattan had the most price variation, with Manhattan often having the most expensive listings. Staten Island had the highest variation in price, including some extreme outliers, while the Bronx and Queens had lower prices, with the Bronx being more compact. (see Figure 3).

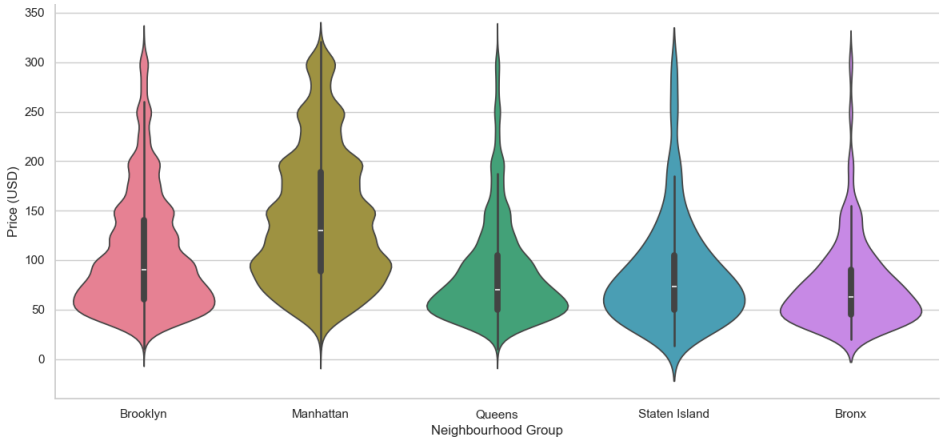

Figure 3 - Price Distribution by Neighbourhood Group

The correlation analysis also revealed some noteworthy associations. Monthly shoppers and reviewers were highly positively correlated with each other (r = 0.56), illustrating that a well-reviewed listing was likely to see repeat bookings, highly valued by hosts. Availability over the year (availability_365) showed a positive correlation with the number of reviews (r = 0.20), illustrating that properties available throughout the year garnered more reviews.

The weak negative correlation between price and longitude (r = -0.31) indicates relative location influenced pricing. Latitude, minimum nights, and calculated host listings count may also influence price, but correlations were weak, indicating minimal relationship (see Figure 4).

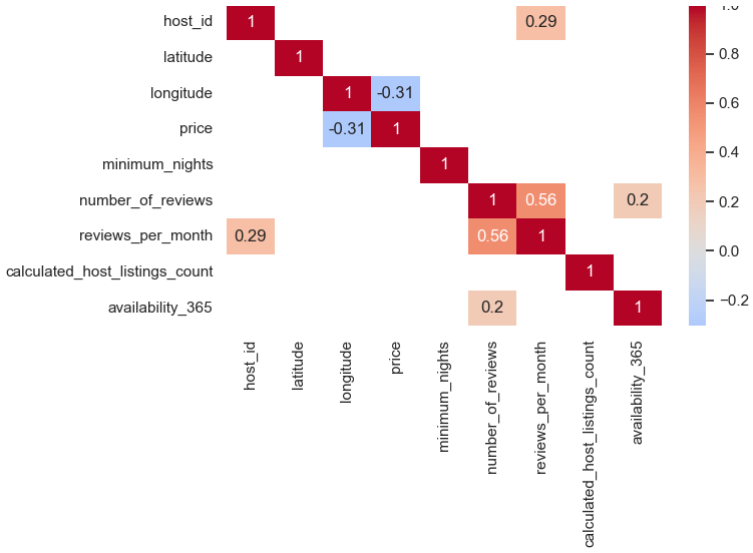

Figure 4 – Correlation matrix for multiple factors

The most expensive option was an entire home or apartment, followed by a private room and a shared room. The number of reviews appeared to have little impact on price. Listings that required 8 to 30 nights were usually more expensive. Stays longer than 30 nights often came with discounts (see Figure 5).

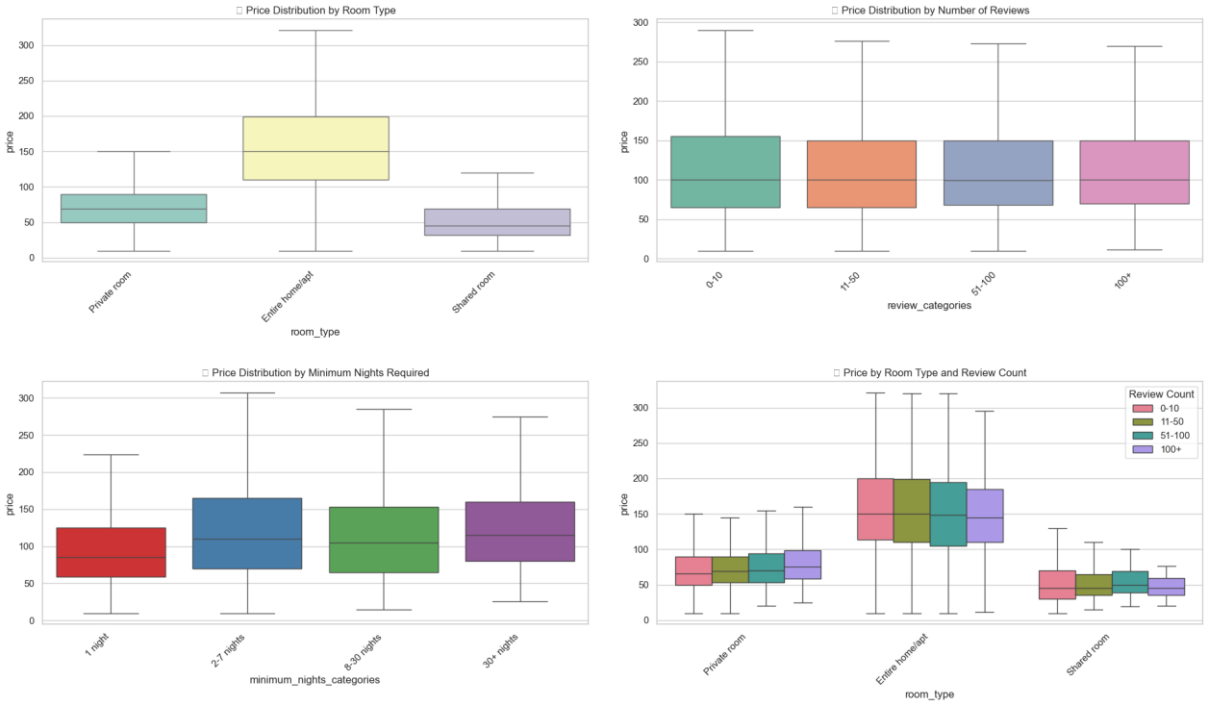

Figure 5 - Analyses on Price and Contributing Factors

Most listings were between $60 and $80. The left-skewed distribution shows that there are fewer premium properties ($200–$300), though they were usually nicer units (see Figure 6).

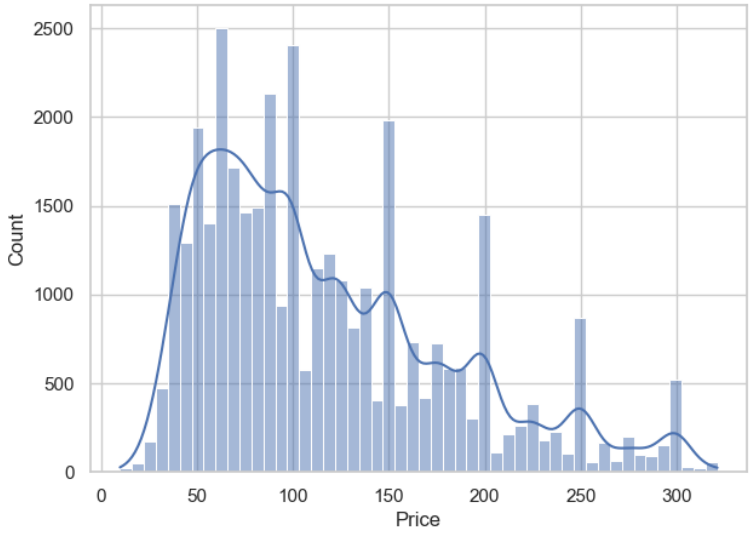

Figure 6 - Count of Units by Price

There was a weak negative correlation ($r \approx -0.0188$, $p \approx 0.0003$) between the price and number of reviews. The price and occupancy relationship can be seen in Figure 7 below.

Figure 7 - Distribution of Reviews by Unit Price

Data Analysis - Analysis of Machine Learning

K-Means clustering was used on price and number of reviews to look for patterns at the neighbourhood level. Linear regression was used to find out how several factors, like minimum nights, review activity, host qualities, room type, and neighbourhood group, affected listing pricing (Schroeder et al., 2017).

Data Analysis - Results of Linear Regression

Table 1 shows that listing features and the location influence prices. Minimum night requirements (-0.22, $p < 0.001$) and quantity of reviews (-0.03, $p < 0.001$) have a small negative effect on pricing. This means that tougher stay criteria and more reviews are linked to somewhat cheaper prices. On the other hand, the number of hosts listed (+0.11, $p < 0.001$) and the availability of the property for the whole year (+0.05, $p < 0.001$) have positive effect. This means experienced hosts and properties that are always accessible can charge somewhat higher fees.

Room type is another important factor. Private rooms are, on average, 77 units cheaper than whole homes, and shared rooms are 103 units less. Location is also very important. Listings in Manhattan (+52 units) and Brooklyn (+23 units) cost significantly more than the baseline, but listings in Queens only slightly more (+11 units).

Overall, the regression shows room type and location are the major factors that affect price, whereas host activity and booking criteria have a smaller effect. This shows how important property features and location are in determining market value.

Predictor	Coefficient	P-value	Interpretation
const	119.55	0.000	When all predictors are 0, the average price is ~119 units.
minimum_nights	-0.22	0.000	More minimum nights required → lower price.
number_of_reviews	-0.03	0.000	More reviews → slightly lower price (popular listings tend to be cheaper).
reviews_per_month	+0.03	0.859	Not statistically significant → no meaningful effect on price.
calculated_host_listings_count	+0.11	0.000	Hosts with more listings tend to charge slightly higher prices.
availability_365	+0.05	0.000	Listings available year-round are priced slightly higher.
room_type_Private room	-76.89	0.000	Private rooms are ~77 units cheaper than entire homes.
room_type_Shared room	-102.92	0.000	Shared rooms are ~103 units cheaper than entire homes.
neighbourhood_group_Brooklyn	+23.37	0.000	Listings in Brooklyn are ~23 units more expensive than the baseline area.
neighbourhood_group_Manhattan	+51.89	0.000	Listings in Manhattan are ~52 units more expensive than the baseline area.
neighbourhood_group_Queens	11.29	0	Listings in Queens are ~11 units more expensive than the baseline area.

Table 1- Predictors and Their Statistics

Data Analysis - K-Means Clustering

The clustering results revealed three distinct market segments:

1. Budget-Friendly (Cluster 0): Listings priced between 50–100 units with moderate reviews (20–50). These properties often serve cost-sensitive travelers and are in less popular areas, indicating lower market maturity.

2. Standard (Cluster 1): Listings priced between 100–150 units with relatively high reviews (40–70). They represent well-established properties in attractive locations, balancing affordability and reliability.
3. Luxury (Cluster 2): Listings priced between 150–250 units with consistently high reviews (60+). Concentrated in high-demand areas, these properties reflect strong reputations, premium services, and target affluent customers.

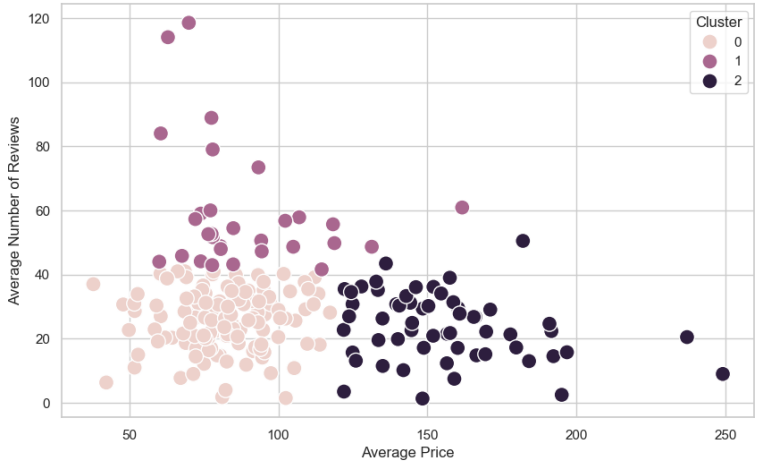

Figure 8 - K-Means Clustering of Neighbourhoods

Standard hosts benefit from the balance between competitive price and service quality, luxury hosts benefit from the exclusivity and first-rate amenities, and budget hosts benefit from the boosted occupancy and reviews. These insights assist hosts in differentiating marketing approaches for different client segments.

Findings and Recommendations

The results suggest that there is a gap, with strong demand near major attraction points and a lot of opportunity for growth with Airbnb. Competitiveness can be strengthened through value-based pricing by selling low-cost properties in the outskirts, along with selling premium properties in high-demand areas. Modification of the pricing architecture comes into play as well. Also, targeted Airbnb marketing based on the companion special offers further improves customer relations, builds company image, and maximizes profitability.

Conclusion

This report identifies key business opportunities for Airbnb in New York City, emphasizing neighbourhood-focused strategies to enhance customer satisfaction and revenue. Leveraging data analysis and machine learning supports evidence-based decision-making, strengthening Airbnb's competitive advantage and long-term growth.

References

Bishop, C.M. (2016) *Pattern recognition and machine learning*. Springer.

Carvache-Franco, M. *et al.* (2023) 'Market segmentation in urban tourism: A study in Latin America,' *PLoS ONE*, 18(5), p. e0285138. https://doi.org/10.1371/journal.pone.0285138.

Chatterjee, S. and Hadi, A.S. (2006) *Regression analysis by example*. John Wiley & Sons.

Chawla, N.V. (2006) 'Data mining for Imbalanced Datasets: An Overview,' in *Springer eBooks*, pp. 853–867. https://doi.org/10.1007/0-387-25465-x_40.

Coles, P.A. *et al.* (2017) 'Airbnb usage across New York City neighborhoods: geographic patterns and regulatory implications,' *SSRN Electronic Journal* [Preprint]. https://doi.org/10.2139/ssrn.3048397.

De Jaureguizar Cervera, D., Yábar, D.C.P.-B. and De Esteban Curiel, J. (2022a) 'Factors affecting short-term rental first price: A revenue management model,' *Frontiers in Psychology*, 13. https://doi.org/10.3389/fpsyg.2022.994910.

De Jaureguizar Cervera, D., Yábar, D.C.P.-B. and De Esteban Curiel, J. (2022b) 'Factors affecting short-term rental first price: A revenue management model,' *Frontiers in Psychology*, 13. https://doi.org/10.3389/fpsyg.2022.994910.

Field, A. (2024) *Discovering statistics using IBM SPSS Statistics*. Sage Publications Limited.

Gelman, A. *et al.* (2013) *Bayesian data analysis*, *Chapman and Hall/CRC eBooks*. https://doi.org/10.1201/b16018.

Hastie, T., Tibshirani, R. and Friedman, J.H. (2001) *The elements of statistical learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Hinton, G. E. (2012). A practical guide to neural networks. Cambridge, MA: MIT Press.

Hocking, R.R. (2013) *Methods and applications of linear models: Regression and the Analysis of Variance*. John Wiley & Sons.

Jiao, J. and Bai, S. (2019) 'Cities reshaped by Airbnb: A case study in New York City, Chicago, and Los Angeles,' *Environment and Planning a Economy and Space*, 52(1), pp. 10–13. https://doi.org/10.1177/0308518x19853275.

Martinez, R.D. *et al.* (2017) *The impact of an AirBnb host's listing description 'Sentiment' and length on occupancy rates*. https://arxiv.org/abs/1711.09196.

McKinney, W. (2017) *Python for Data Analysis, 2nd Edition*.

Nivón, T. and Nivón, T. (2022) 'Data analysis on Airbnb in NYC | Data Science blog,' *Data Science Blog*, 9 July. https://nycdatascience.com/blog/student-works/data-analysis-on-airbnb-in-nyc/.

Perez-Sanchez, V.R. *et al.* (2018) 'The what, where, and why of Airbnb price determinants,' *Sustainability*, 10(12), p. 4596. https://doi.org/10.3390/su10124596.

Provost, F. and Fawcett, T. (2013) *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. 'O'Reilly Media, Inc.'

Schroeder, L.D., Sjoquist, D.L. and Stephan, P.E. (2016) *Understanding regression analysis: An Introductory Guide*. SAGE Publications, Incorporated.

Sharma, U. and Gupta, D. (2021) 'Analyzing the applications of internet of things in hotel industry,' *Journal of Physics Conference Series*, 1969(1), p. 012041. https://doi.org/10.1088/1742-6596/1969/1/012041.

Tufte, E. R. (2001) *The visual display of quantitative information*. 2nd ed. Graphics Press LLC.

Wedel, M. and Kamakura, W.A. (2000) *Market segmentation*, *International series in quantitative marketing*. https://doi.org/10.1007/978-1-4615-4651-1.