Bellevue University

Airbnb Price Predictor

Ashley Mayo

Milestone Three

April 21st, 2025

Professor Iranitalab

DSC680: Applied Data Science

**Introduction**

When two friend groups set out to plan the perfect bachelor and bachelorette weekends in New York City, their excitement quickly ran into the hard truth of budgeting for short term rentals. They both turned to Airbnb to find accommodations that could meet their needs: a central location, enough space for their guests, and the right vibe to celebrate before the big day. But with over 48,000 listings scattered across five cities and prices ranging from budget friendly to extravagant, how could they know what to expect?

This project took their party planning dilemma and transformed it into a data science problem. Could we predict the price of an Airbnb listing in New York City using historical data and machine learning? And if so, which factors such as room type, location, availability, or something else would influence those prices the most?

The "New York City Airbnb Open Data" (Kaggle, 2019) dataset provided the foundation for the analysis. It provided a rich snapshot of NYC's Airbnb landscape in 2019, including variables like neighborhood, room type, number of reviews, and availability. Cleaning the data involved filtering out extreme outliers such as listings over $1000 per night, removing long minimum night stays that would not appeal to weekend partygoers, and simplifying the features to focus on what mattered most. Latitude and longitude were removed to avoid overfitting and preserve interpretability. Missing values in review-related columns were filled with zeros, and categorical variables such as neighborhood group and room type were one hot encoded for modeling.

**Data Exploration**

A graph of a number of bars

AI-generated content may be incorrect.To predict Airbnb listing prices, three regression models were trained: Linear Regression, Random Forest, and Gradient Boosting. Each model was evaluated using performance metrics such as Mean Absolute Error, Root Mean Squared Error, and R squared. Among the three models tested, Gradient Boosting was identified as the most promising model, particularly after applying a log transformation to the target variable to reduce the skew caused by high-priced listings. This modeling approach aligns with the principle that effective data science involves both analytical rigor and strategic decision-making (Provost & Fawcett, 2013). As Zhang et al. (2017) noted, “machine learning models are especially effective in high-dimensional real estate data, where complex interactions often drive pricing outcomes.”

During data exploration, several patterns emerged. Manhattan stood out as the most expensive city. As shown in Graph One, average listing prices in Manhattan approach $175 per night, while Brooklyn, Queens, the Bronx, and Staten Island offer more affordable options. Groups seeking proximity to nightlife in the East Village or skyline views in Midtown could expect higher prices. In contrast, Brooklyn offered a more affordable option without sacrificing charm or access to popular areas such as Williamsburg or DUMBO. These patterns reflect broader urban economic trends, as “housing and rental markets are highly sensitive to neighborhood desirability and amenity access” (Glaeser et al., 2018).

A graph with blue squares

AI-generated content may be incorrect.Room type was among the most influential pricing factors. Entire home or apartment rentals commanded higher prices compared to private or shared rooms, but the added privacy made them a preferred option for group stays. Most listings supported short stays, typically with one-to-three-night minimums, aligning well with typical weekend travel schedules. Price fluctuations across the year appeared minimal, offering reassurance that booking in summer or fall would not drastically impact budgets.

A diagram of a graph

AI-generated content may be incorrect.The model predicted Airbnb prices with an average error of approximately $47, providing a reliable estimate for planning purposes. More importantly, the process demonstrated how data can transform complex tasks like trip budgeting into manageable and empowering experiences. It was trained on data from a single year (2019), which does not capture seasonal pricing, economic disruptions from the pandemic, or changes in Airbnb policies. Price stability was assumed within each city and room type category. Features such as amenities, precise addresses, or calendar-based booking patterns were excluded, limiting the model’s generalizability.

Price outliers presented an early challenge, skewing predictions. Class imbalance among pricing categories also required tuning to balance interpretability and accuracy. Removing features such as location coordinates helped reduce overfitting. This model has the potential to be extended into a real-time price estimator that incorporates calendar availability, host ratings, and user preferences to recommend listings. The same approach can also be adapted for other cities, offering support to both hosts and guests in global markets.

For weekend travel groups, Brooklyn or Queens offer affordable options, while entire home rentals provide added privacy. Filtering listings by availability and minimum night requirements ensures a better match with travel plans. Hosts can apply the model’s insights to set competitive, yet fair, prices based on listing characteristics.

To bring this model to life as a practical tool for travelers or Airbnb hosts, a web-based application could be developed using Python-based frameworks such as Streamlit or Flask. These platforms would allow users to input key details, such as neighborhood, room type, number of nights, and availability, and receive an instant price prediction generated by the trained Gradient Boosting model.

For deployment, cloud-based services like Heroku or AWS (Amazon Web Services) could host the application and make it publicly accessible. The user interface would prioritize simplicity, guiding users through a few dropdown menus and sliders, while the back end handles the prediction logic in real time. With additional development, the tool could incorporate interactive maps, calendar-based pricing trends, and personalized recommendations based on user preferences. Integrating this solution into travel planning platforms or Airbnb analytics dashboards could further enhance its usability and real-world impact.

Careful oversight is essential when applying pricing algorithms, especially in housing markets where affordability and accessibility are already under strain. Without proper safeguards, automated recommendations could inadvertently reinforce existing inequalities, contribute to gentrification, or displace long-term residents in high-demand neighborhoods. These unintended consequences highlight the importance of building models that are not only accurate but also responsible. Transparency in model logic, regular bias audits, and input from community stakeholders should be prioritized. Airbnb, for example, has emphasized its commitment to fair housing and equitable access (Airbnb, n.d.), and data scientists working with such platforms must align with these goals. Ultimately, ethical data science practices should strive to reduce disparities, not deepen them, by ensuring fairness, inclusivity, and accountability in both design and deployment.

**Conclusion**

Planning a weekend getaway for a bachelor or bachelorette party in New York City involves countless decisions. From choosing the right neighborhood to staying within a set budget. Through the lens of data science, these decisions become more manageable and grounded in insight. By analyzing Airbnb listings and exploring key factors such as location, room type, and availability, this project transformed a complex pricing puzzle into a practical guide for group travelers.

For party planners hoping to celebrate near the energy of Manhattan or the charm of Brooklyn, the model provides a reliable estimate of what to expect. Whether it's a skyline view in Midtown or a cozy spot near Williamsburg, the data helps narrow down options and set realistic expectations. More broadly, this process demonstrates how predictive modeling can bridge the gap between personal needs and market realities.

The findings extend beyond a single celebration. This work highlights how machine learning can bring clarity to everyday choices, making travel less stressful and more strategic. While the model is not without limitations, it serves as a foundation for smarter tools that can be adapted to new cities, new features, and evolving traveler needs.

Ultimately, this project shows that data isn't just about numbers, it's about storytelling, decision-making, and empowerment. When used thoughtfully, it has the power to turn the chaos of planning into something purposeful, whether for a milestone weekend or any journey ahead.

10 Questions the Audience would ask:

1. Why did you choose 2019 data, and how might results change if newer or post-pandemic data were used?

I chose 2019 data because it reflects stable, pre-pandemic market conditions. Using post-pandemic data might show different trends due to shifts in travel behavior and demand. Including newer data in future models could help account for ongoing market changes.

1. How would you handle changes in demand patterns due to events like the COVID-19 pandemic or shifts in tourism?

I’d retrain the model regularly with updated data and include features like public health alerts or seasonal tourism indicators to capture demand shifts. Adapting the model over time ensures it remains relevant and accurate in changing conditions.

1. What were some of the key features removed during data cleaning, and how did that affect the model's performance or generalizability?

I removed columns with high missing values or irrelevant identifiers. This helped streamline the model and reduce noise, improving generalizability. While some detail was lost, it improved efficiency and model focus.

1. Why was Gradient Boosting the best-performing model over Random Forest or Linear Regression? What specific metrics led you to that decision?

Gradient Boosting outperformed others based on MAE, RMSE, and R² scores. It captured non-linear patterns better than Linear Regression and was more precise than Random Forest in this case. Its ability to fine-tune predictions gave it the edge in accuracy.

1. You mentioned one-hot encoding for categorical features, did that impact model complexity or training time significantly?

One-hot encoding did increase model complexity slightly but didn’t significantly affect training time due to the limited number of categorical variables. It allowed the model to process categorical features effectively without introducing bias.

1. The model had an average error of $47, how would that impact a real user’s planning or budget decisions? Is that margin of error acceptable in this context?

A $47 error could influence budget planning, especially for longer stays, but it’s generally acceptable for early-stage price estimation. Users can use this range as a guideline while considering other listing-specific factors.

1. Were there any surprising findings during the analysis? For example, did any City or room type behave unexpectedly in terms of pricing?

Yes, some neighborhoods in Queens had unexpectedly high prices, and private rooms in certain areas were more expensive than anticipated. These insights highlight the importance of location and property type in pricing variability.

1. How would you extend this model to other cities or markets with different characteristics, like rural or seasonal tourist destinations?

I’d collect localized data and retrain the model, incorporating features like seasonality, local events, and transportation access for rural or niche markets. Adapting feature selection based on regional characteristics would enhance performance.

1. Could this model be used by Airbnb hosts to price competitively? What additional data would you need to make that practical?

Yes, with added data like competitor pricing, booking frequency, and seasonal trends, this model could help hosts price more competitively. Making the model dynamic would allow it to react to market shifts in real time.

1. How would you deploy this model as a user-facing tool? What platforms or frameworks would you consider?

I’d deploy it as a web app using Flask or Streamlit, possibly hosted on Heroku or AWS for accessibility and scalability. The interface would allow users to input listing details and receive price estimates instantly.

**References**

Airbnb. (n.d.). *Airbnb's commitment to fair housing*. <https://www.airbnb.com/help/article/2701>

Glaeser, E. L., Gyourko, J., & Saks, R. E. (2018). Why have housing prices gone up? *American Economic Review: Papers & Proceedings, 108*(5), 1–6.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O’Reilly Media.

Zhang, C., Zhao, Y., Li, X., & Cao, J. (2017). Analyzing short-term rental price determinants with machine learning algorithms. *Procedia Computer Science, 122*, 801–807. <https://doi.org/10.1016/j.procs.2017.11.439>

**Appendix**

To support the analysis presented in the main body of this project, three key visualizations have been included in this appendix. These charts provide deeper insight into pricing trends across New York City Airbnb listings and highlight the model's inner workings.

Graph One, titled “Average Airbnb Price by City in NYC”, displays the average listing price across the five boroughs. As expected, Manhattan stands out significantly with the highest average nightly price, approaching $175. Brooklyn follows, still relatively expensive, while the Bronx, Queens, and Staten Island offer more affordable options for travelers. This graph visually supports the narrative that location plays a substantial role in Airbnb pricing, especially for bachelor or bachelorette groups aiming for both proximity and budget.

Graph Two, “Top 10 Feature Importances - Gradient Boosting”, showcases the most influential variables in the Gradient Boosting model—the best-performing algorithm used for predicting listing prices. Among the top predictors are room type (particularly the presence of private or shared rooms), Manhattan listings, and host availability (measured by availability\_365). This graph demonstrates the strength of categorical variables like room type and neighborhood in driving price predictions. The model’s reliance on these features validates the initial data cleaning and feature engineering strategy used in the project.

Graph Three, “Airbnb Price Distribution by Room Type (Log Scale)”, presents a boxplot comparing nightly prices across private rooms, entire homes/apartments, and shared rooms. Displayed on a logarithmic scale to better handle extreme values, this chart reveals that entire homes/apartments command the highest median prices—an expected outcome for group-oriented travel scenarios. Private rooms generally fall in the middle range, while shared rooms represent the most economical option. This visualization aligns directly with the project's central narrative: that room type strongly influences price, and travelers can make informed choices based on their privacy and space preferences.

Together, these three figures provide a solid visual foundation for understanding Airbnb pricing dynamics in NYC. They support the model findings, reinforce the story-based problem framing, and enhance the project’s credibility as both a data science application and a real-world decision-making tool.