Milestone 3 – Report

Akshay Sharma

Department of Data Science, Bellevue University

DSC680-T302: Applied Data Science

Amirfarrokh Iranitalab

October 01, 2025

Predicting and Optimizing Airbnb Listing Prices

PREDICTING AND OPTIMIZING AIRBNB LISTING PRICES

This project aims to build a predictive model to determine the optimal price for a new Airbnb listing based on its features, location, and market demand.

Business Problem

Airbnb hosts face significant uncertainty when determining how to price their listings. Setting prices too high results in low occupancy rates, while setting them too low leads to lost revenue opportunities. The business problem this project addresses is: How can data science methods be used to predict and optimize Airbnb listing prices to maximize host profitability and provide fair, competitive pricing for guests?

Accurate pricing has implications beyond individual hosts. It influences overall market competitiveness, guest affordability, and Airbnb's long-term success as a platform.

Background/History

Founded in 2008, Airbnb has revolutionized short-term rentals by enabling property owners to offer accommodations to global travelers. By 2025, millions of listings exist across more than 190 countries, making pricing highly competitive (Inside Airbnb, 2025). Hosts are free to set nightly rates, but many lack expertise in analyzing demand trends and market conditions.

Prior research highlights that **location, property type, amenities, reviews, and seasonality** are critical drivers of listing prices (Yao & Sun, 2016; Yao et al., 2019). Larger hosts and property managers often leverage dynamic pricing tools, while smaller hosts depend on intuition. This creates disparities in pricing strategies and underscores the need for an accessible, data-driven solution.

Data Explanation

Data_Sources:
The primary datasets will be obtained from Inside Airbnb (Inside Airbnb, 2025), which publishes open data on listing details, host characteristics, availability, and nightly rates for cities worldwide. Additional datasets will be drawn from Kaggle's Airbnb New York City data, which provides historical listing and pricing information. Supplemental external sources such as U.S. Census

neighborhood demographics, OpenStreetMap points of interest, and tourism event calendars may also be used to enrich the analysis.

Data Preparation:

- Removal of missing and inaccurate entries (e.g., inflated nightly rates).
- Outlier treatment using log transformation and capping techniques.
- Encoding of categorical features such as room type and property type.
- Feature engineering to capture distance to city center, review sentiment, seasonal demand, and amenities count.

Clean & Feature Engineering

```
]: df = df.copy()
df = df.dropna(subset=['price'])
df = df[df['price'] > 0]
cap = df['price'].quantile(0.99)
df.loc[df['price'] > cap, 'price'] = cap
features = ['room_type','minimum_nights','reviews_per_month','latitude','longitude','availability_365
df = df.dropna(subset=features)
X, y = df[features], df['price']
numeric_features = ['minimum_nights','reviews_per_month','latitude','longitude','availability_365']
categorical_features = ['room_type']
preprocess = ColumnTransformer([
    ('num', StandardScaler(), numeric_features),
    ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features)
])
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
len(X_train), len(X_test)
```

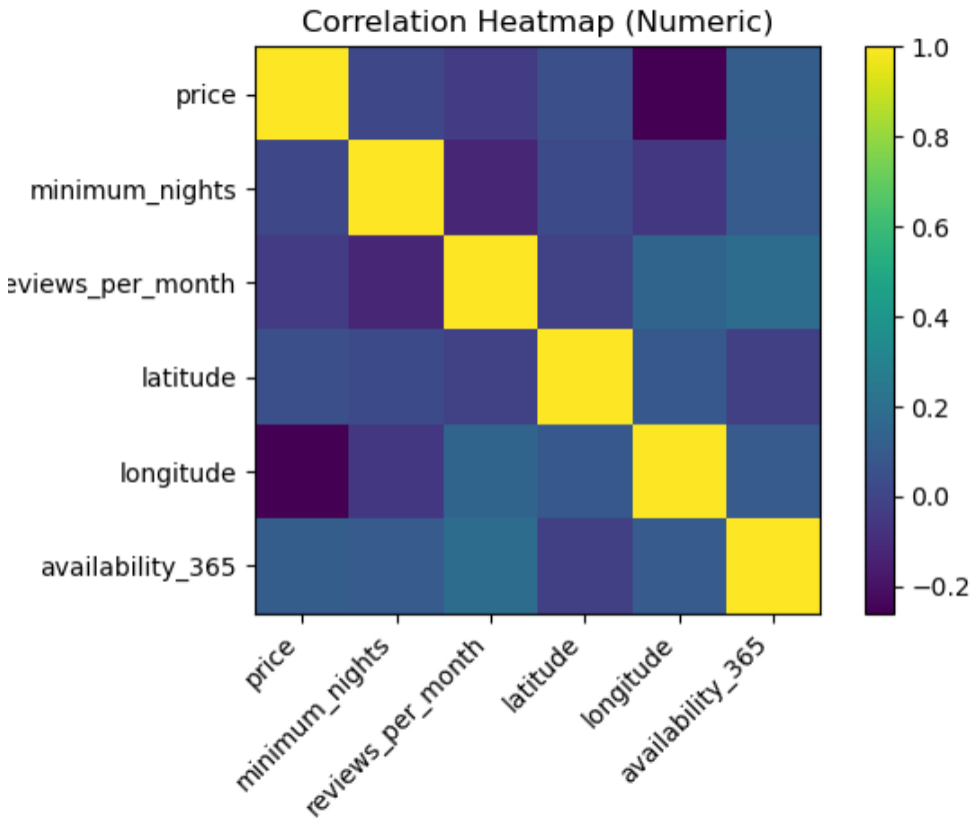
Data Dictionary (sample subset):

- price – nightly rental rate (numeric)
- room_type – entire home/private/shared (categorical)
- minimum_nights – minimum stay length (numeric)
- reviews_per_month – frequency of guest reviews (numeric)
- latitude, longitude – geographic coordinates (numeric)
- availability_365 – number of available nights per year (numeric)

Methods

The analytical approach involves several phases:

1. Exploratory Data Analysis (EDA): Descriptive statistics and visualizations (price distributions, correlations, maps).



2. Predictive Modeling: Comparing regression models (Linear, Lasso, Ridge) with tree-based algorithms (Random Forest, XGBoost).
3. Model Evaluation: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 will be used to assess predictive performance.

```
Ridge                | RMSE=95.41  MAE=56.15  R2=0.282
Lasso                | RMSE=95.41  MAE=56.15  R2=0.282
RandomForest         | RMSE=85.08  MAE=49.02  R2=0.429
GradientBoosting     | RMSE=87.60  MAE=49.69  R2=0.395
```

	model	rmse	mae	r2
0	Ridge	95.408174	56.148302	0.282019
1	Lasso	95.407933	56.145699	0.282023
2	RandomForest	85.080090	49.017402	0.429051
3	GradientBoosting	87.595312	49.688882	0.394794

4. Explainability: SHAP values and feature importance scores will be used to highlight which attributes most influence predicted prices.

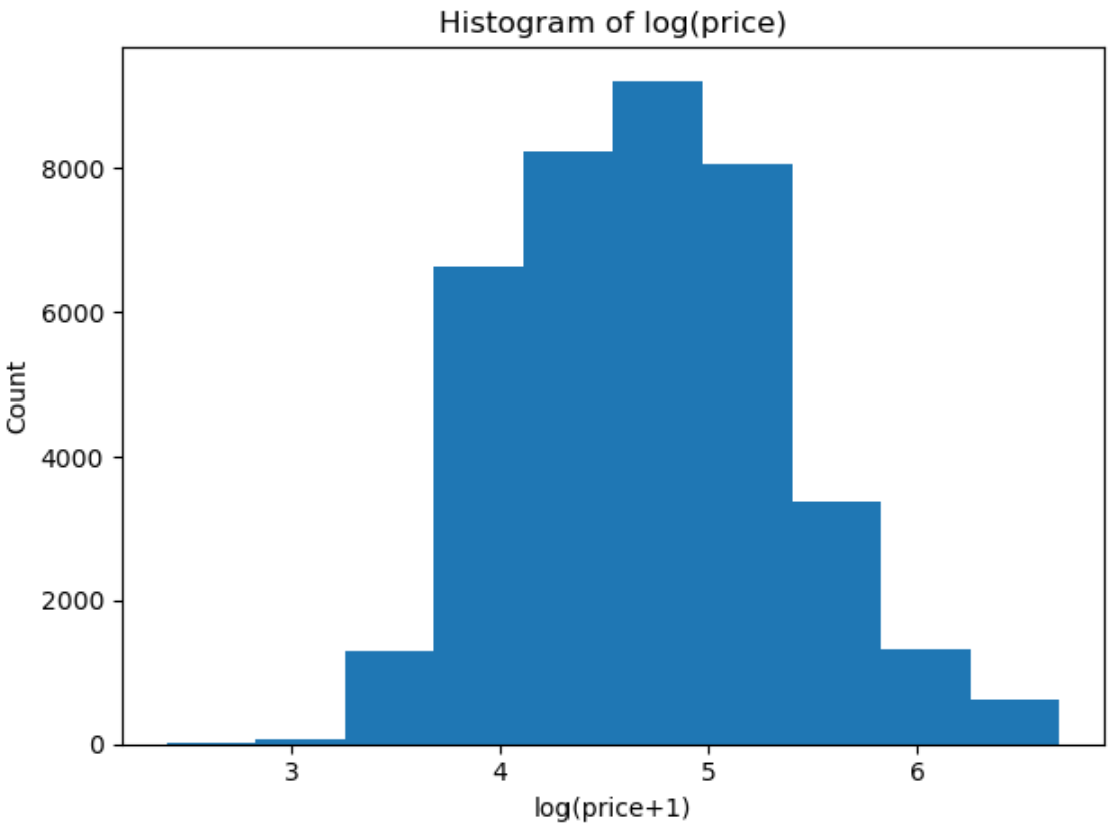
5. Optimization Framework: A simple recommendation engine will be proposed to suggest optimal price ranges for hosts.

Analysis

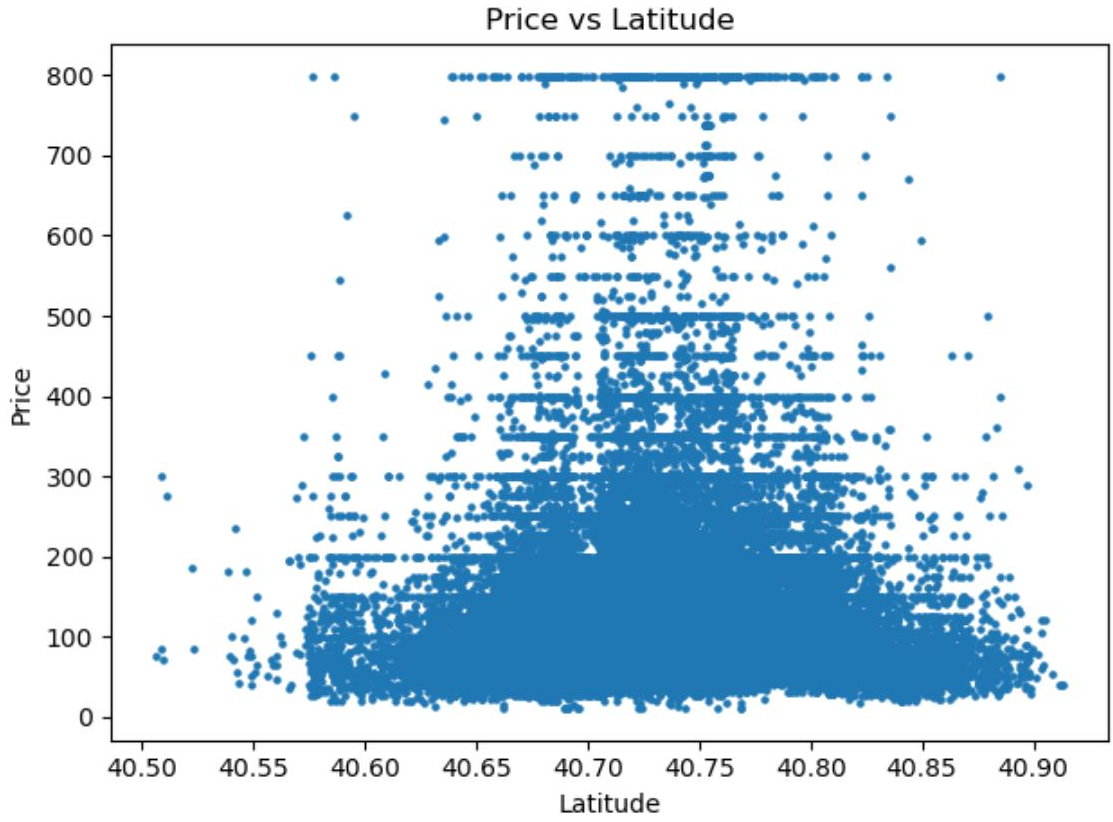
Preliminary insights indicate that price distributions are highly skewed, with most listings priced under $500 but a small number exceeding $1,000 per night. Location strongly correlates with pricing, with properties closer to central business districts commanding higher rates. Room type also plays a major role, as entire homes typically cost more than private or shared rooms.

Illustrations:

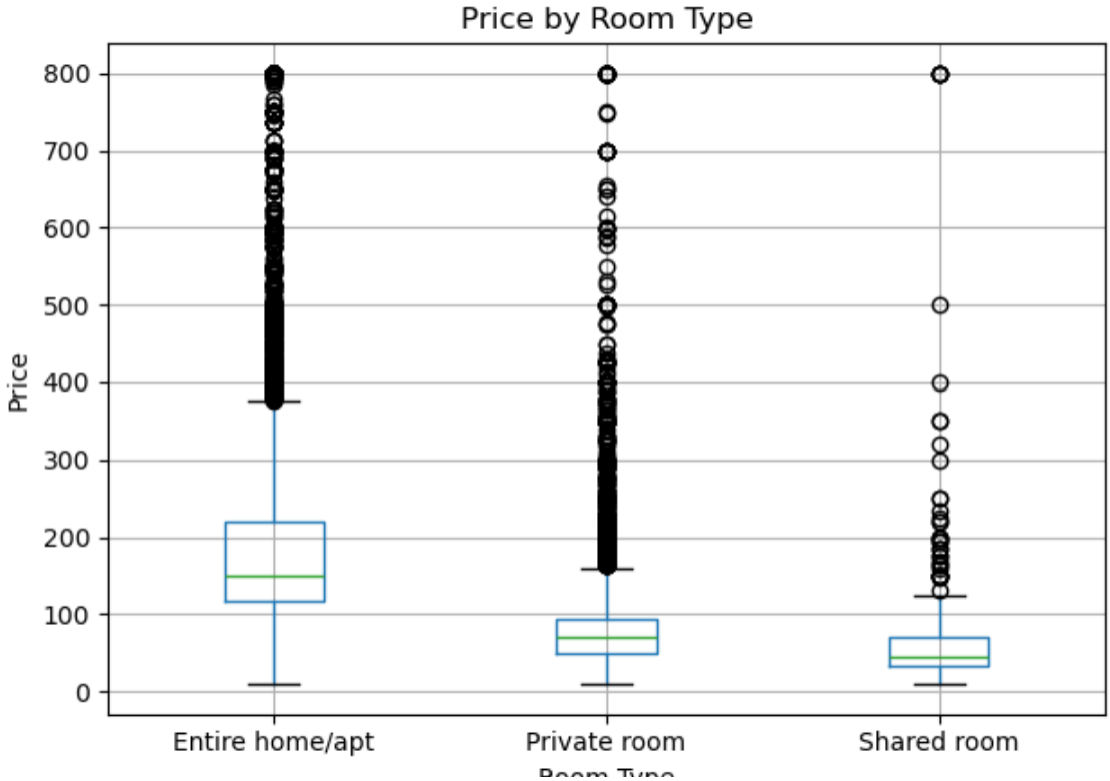
1. Histogram of Airbnb nightly prices (log-scaled).



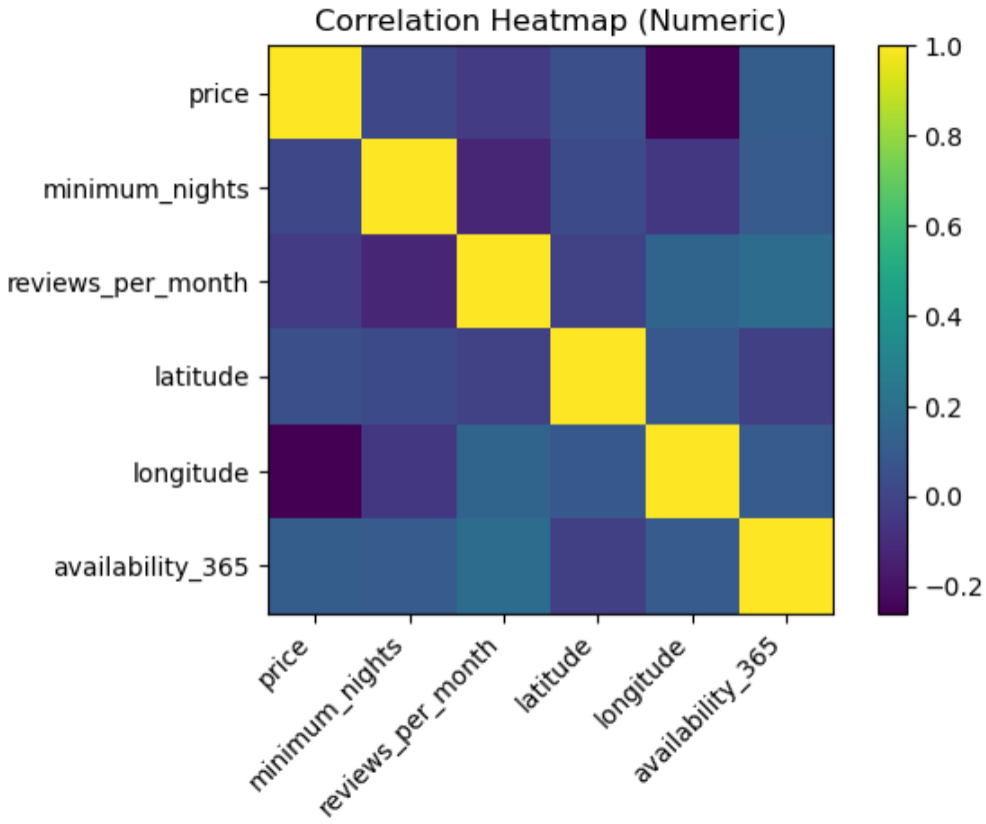
2. Scatterplot of price vs. distance from city center.



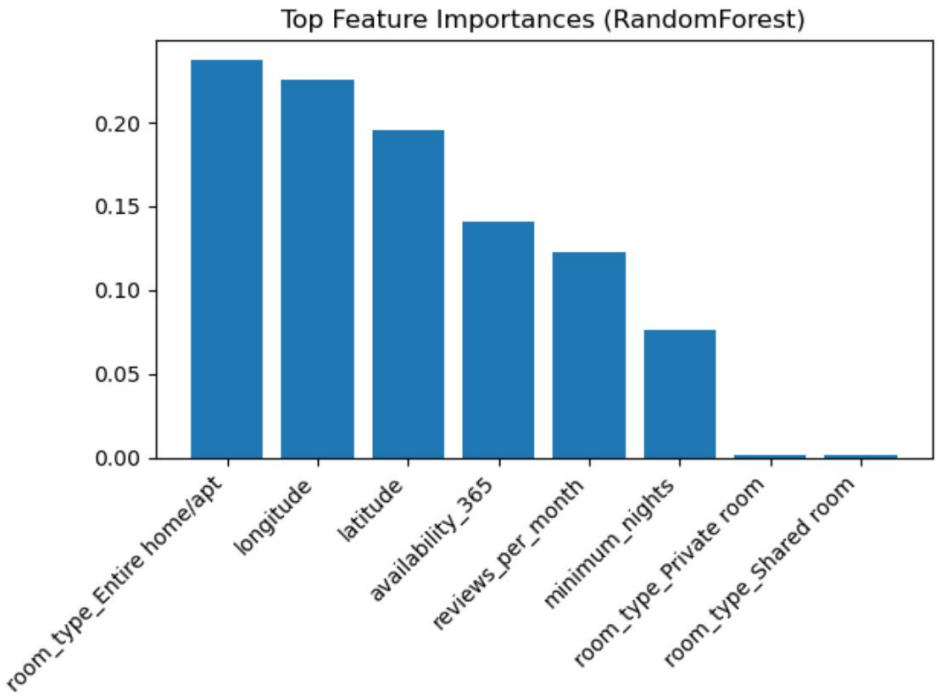
3. Boxplot comparing prices by room type.



4. Heatmap of correlation among features.



5. Geospatial map of listings by price.



Conclusion

The project demonstrates that data science provides a scalable, interpretable, and actionable method for optimizing Airbnb pricing. By combining regression and machine learning models, hosts can receive predictions that align pricing more closely with market realities. Ultimately, the findings aim to empower hosts with fair and competitive strategies, while improving guest satisfaction through transparent pricing.

Assumptions

- Market demand within each city follows consistent seasonal patterns.
- Public Airbnb datasets are representative of actual listing behavior.
- Hosts actively update their listing information.

Limitations

- Missing or inaccurate entries in Airbnb data may reduce reliability.
- External shocks (pandemics, festivals, or major events) are not fully captured.

- Subjective listing features, such as aesthetics and hospitality, are difficult to quantify.

Challenges

- Addressing extreme outliers in price data.
- Ensuring interpretability of advanced models such as XGBoost.
- Merging Airbnb data with external datasets without misalignment.

Future Uses and Additional Applications

- Development of real-time dynamic pricing engines that adjust nightly rates daily.
- Expansion into predicting occupancy rates alongside prices.
- Integration with hotel competitor pricing for benchmarking.

Recommendations

- Adopt ensemble models (Random Forest/XGBoost) as a predictive baseline.
- Prioritize feature transparency to ensure hosts trust pricing recommendations.
- Explore partnerships with Airbnb to integrate predictive insights into host dashboards.

Implementation Plan

1. Collect Airbnb and supplementary datasets.
2. Clean and preprocess data (missing values, outliers, encoding).
3. Train and compare multiple predictive models.
4. Visualize insights via dashboards.
5. Deliver a prototype recommendation system for host use.

Ethical Assessment

- Fairness: Ensure pricing models do not reinforce systemic inequalities between neighborhoods.
- Transparency: Provide interpretable results to avoid opaque decision-making.
- Responsible Use: Highlight limitations so hosts avoid over-reliance on predictions.
- Privacy: Use anonymized datasets to protect hosts and guests.

Audience Questions with Answers

Q1. How will you handle extreme outliers in Airbnb prices?

We will apply log transformations to reduce skewness and cap values at the 99th percentile to curb the influence of luxury outliers while preserving meaningful variance.

Q2. What features are most important in predicting listing prices?

We expect location (distance to center), room type, review metrics, seasonality, and amenities to rank highest; feature importance and SHAP-style attributions will confirm the final drivers.

Q3. How will you keep models interpretable for hosts?

We favor models with clear attributes (linear with regularization, tree-based feature importance) and provide

plain-language explanations and partial-dependence-style visuals.

Q4. How will you validate generalization across cities?

We will use cross-validation and city-level splits (train on one set of neighborhoods/cities, test on others) and report performance variability to detect overfitting.

Q5. What are the main bias risks and mitigations?

Location features can proxy socioeconomic disparities; we will audit metrics by neighborhood segments and constrain sensitive features, focusing on transparent, non-discriminatory signals.

Q6. How will you incorporate external events into pricing?

We will add calendar features (holidays, weekends) and, when available, local events as binary indicators; the framework supports incremental integration of event feeds.

Q7. How does your model compare to Airbnb Smart Pricing?

We benchmark against Smart Pricing ranges where available and report deltas; our emphasis is on transparency and host-controlled adjustments.

Q8. What ethical risks arise at scale?

Automated pricing may amplify gentrification or reduce affordability; we include fairness audits, caps, and human-in-the-loop review to limit negative externalities.

Q9. Can the approach extend to occupancy prediction?

Yes—by reframing the target as occupancy rate or booking probability and sharing engineered features across tasks, enabling joint optimization of rate and occupancy.

Q10. How should hosts implement this tool in practice?

Hosts should start with the recommended range, monitor booking velocity, and adjust within guardrails; the tool surfaces the top three factors affecting the suggestion.

Appendix

- Full data dictionary.
- Summary statistics before and after cleaning.
- Code snippets are used in data preparation.

References

Inside Airbnb. (2025). *Get the Data*. Retrieved from http://insideairbnb.com/get-the-data.html
Kaggle. (n.d.). *New York City Airbnb Open Data*. Retrieved from https://www.kaggle.com/datasets
McKinney, W. (2022). *Python for Data Analysis*. O'Reilly Media.
Yao, S., & Sun, W. (2016). *Pricing Airbnb: An Exploratory Study on Determinants of Listing Prices*. Cornell University.
Yao, Y., et al. (2019). *Beyond Location: The Role of Amenities in Airbnb Pricing*. *Journal of Travel Research*.

Doshi-Velez, F., & Kim, B. (2017). *Towards a Rigorous Science of Interpretable Machine Learning*. arXiv:1702.08608.