University of Toronto

STA302 – Methods of Data Analysis I — Summer 2025

STA302
Final Project Report

Nicholas Koh Ben Sat Long Pingshan Yilin Liu Hen Lin

June 17, 2025

Contents

Research Question

Do hosts with larger portfolios charge different nightly prices after accounting for listing location, room type, availability, reviews, and other listing characteristics?

1 Introduction

Due to its fluctuating seasonal nature, the economic factors associated with travel are constantly changing, sometimes as quickly as overnight. Alongside ticket prices and other passenger fares, lodging costs are a major element when planning and budgeting for travel. Hotels have historically served as temporary residences for travellers worldwide, though as of the late 2000s, professional establishments have encountered competition from local property owners via organizations like Airbnb.

Community-driven offerings vary in price, often cheaper than professional alternatives. However, like the standard of hotel service, costs may fluctuate. Prices change according to quality, demand, services, etc., but an element less considered may be a host's portfolio. As hotel chains operate and offer service at multiple locations, so can Airbnb hosts.

Hosts must account for operational costs, and as the scale of an operation grows, so too do their costs. Demand fluctuates with the seasons, and pricing will naturally follow. Unlike these common variables, expenses accrued in the hospitality industry are a constant factor for determining revenue and profit.

While hotels and Airbnb hosts do not share the same expenses, not all hosts offer singular spaces. As a result, the number of rooms, locations, etc., can alter a service's valuation. While other factors help determine prices, the number of properties used for hospitality may influence the decision.

Whether they rise or fall, changes in pricing are our primary concern. The nature of linear regression will provide insight into the impact of our desired variable: host properties and listings. While accounting for other significant predictors, i.e., listing location, room type, nights stayed, and their reviews, linear regression will help analyze the effects of our response variable against others.

Should host listings influence costs, travellers looking to book an Airbnb may want to consider this variable for their travel plans.

2 Data Description

The dataset used in this study is the *New York City Airbnb Open Data* from Kaggle (Dgomonov, 2019). The original curator of the dataset compiled the data by collecting publicly available listings from Airbnb's website via Inside Airbnb (Inside Airbnb, 2019). Its primary purpose was to provide transparency about short-term rental activity in urban areas by including information such as listing prices, host activity, geographic location, and room features. Our current study extends the use of this dataset by analyzing pricing strategies relative to host portfolio size.

Our response variable is **price**, reflecting the nightly cost of each Airbnb listing, which is central to understanding host behavior. As a continuous variable, it is appropriate for linear regression and provides direct insight into how listing characteristics and host behavior influence rental pricing.

We selected seven predictors for the preliminary model:

1. **calculated_host_listings_count**: measures how many listings a host manages. Larger portfolios may signal professional operations, potentially affecting pricing strategies.
2. **neighbourhood_group** (categorical): indicates the borough where the listing is located. Prices vary by location due to differences in demand, tourism, and living costs.
3. **room_type** (categorical): specifies the level of privacy offered (e.g., entire home vs. shared room). Listings with more privacy usually command higher prices.
4. **minimum_nights**: sets the minimum stay requirement. Higher values may discourage short-term bookings, affecting pricing flexibility.
5. **availability_365**: captures how many days per year the listing is available. More availability may indicate a full-time rental, which could influence pricing patterns.
6. **number_of_reviews**: reflects total guest feedback. More reviews can build credibility and justify higher prices.
7. **reviews_per_month**: measures recent guest activity. Higher review rates suggest consistent demand, which may lead to higher pricing.

Each predictor is relevant to the research question as they represent either host behavior or guest engagement. The inclusion of both continuous and categorical variables enables subtle modeling of price variability across different market segments.

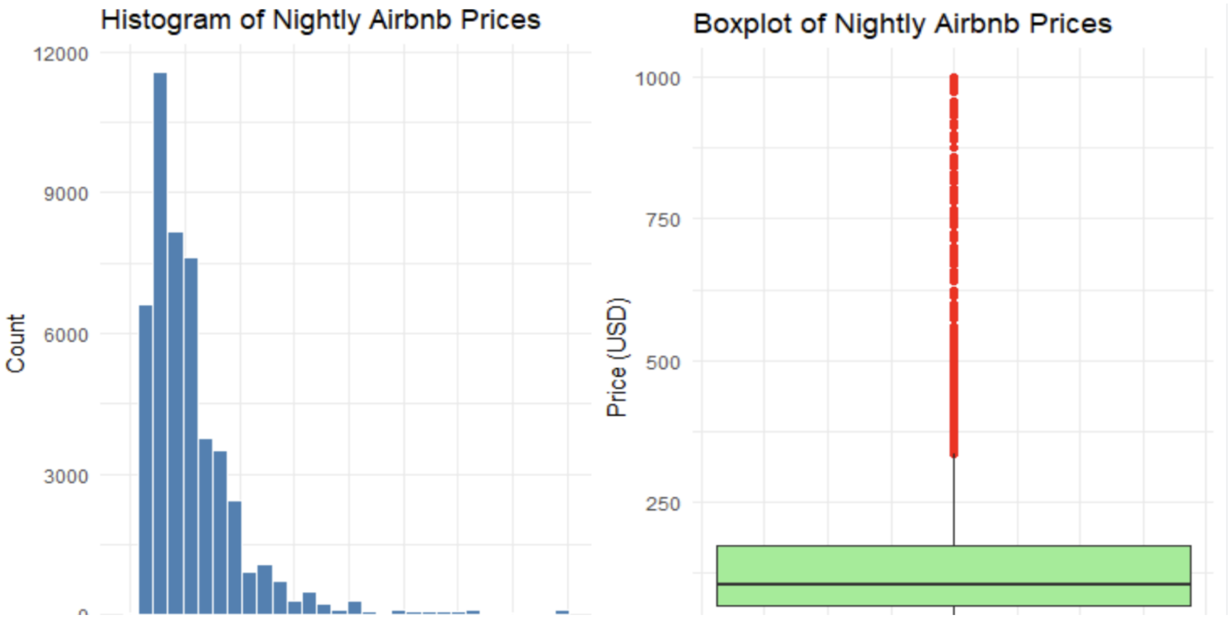

Figure 1: Distribution of Nightly Airbnb Prices: Histogram and Boxplot.
Left – Histogram: Nightly prices are heavily right-skewed; most stays cost under $250, with a long tail of luxury listings approaching $1,000.
Right – Boxplot: The median is just above $100, the inter-quartile range spans roughly $50–$170, and numerous upper-tail outliers underscore the premium segment.

To select our final seven predictors, we began by reviewing all available columns in the NYC Airbnb dataset and excluded those that were mostly incomplete (e.g., host response time), redundant with other variables, or conceptually overlapping availability features. From there, we kept only variables that were clearly linked to price in prior research or exploratory analysis.

3 Preliminary Results

We evaluated the four key regression assumptions for our preliminary model by using diagnostic plots to assess how well they are met. To address violations seen in an initial raw price model, we log-transformed the response variable (*price*), and all plots here use *log-price* on the y-axis.

3.1 Linearity

The Response vs. Fitted plot shows a reasonably linear upward trend, and the Residuals vs. Fitted plot displays a mostly random scatter of points centered around zero. There's no strong systematic curve, suggesting that the log transformation has helped stabilize the relationship between the predictors and price. The linearity assumption appears with no serious violation.

3.2 Homoscedasticity (Constant Variance)

The Standardized Residuals vs. Fitted plot indicates that the spread of residuals remains fairly consistent across the range of fitted values. While there's some slight widening at the higher end, the majority of residuals fall within a stable band. This suggests that the assumption of constant variance is largely met with only mild heteroscedasticity in the most expensive listings.

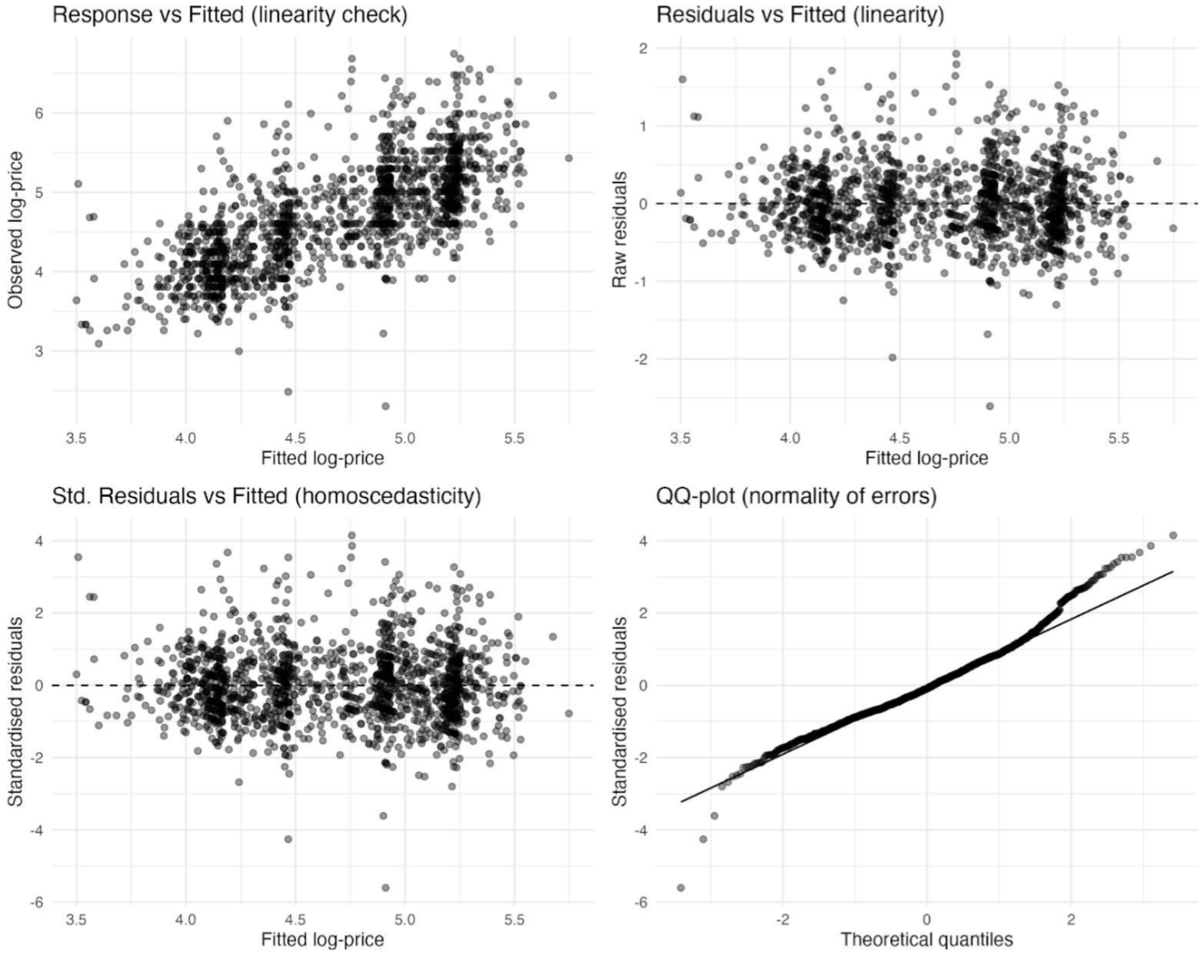

Figure 2: Diagnostic plots for assessing linear regression assumptions: linearity, homoscedasticity, and normality of residuals.

3.3 Normality of Errors

While the Q–Q plot generally supports the normality assumption, we observe mild deviations in both tails, particularly the upper tail, suggesting that a few high-end listings are not fully captured by the model even after log transformation. Given our relatively large sample size, we rely on the robustness of linear regression to slight non-normality, especially when it occurs only in the tails and does not heavily affect central residual behavior. As such, no further transformation was applied. Nonetheless, we acknowledge this limitation and interpret results for high-priced listings with caution.

The main predictor, **calculated_host_listings_count**, has a small, statistically insignificant effect, suggesting that hosts with more listings may charge slightly more, but the difference isn't conclusive at this stage.

Room type shows strong effects: compared to entire homes (the baseline), private rooms and shared rooms are associated with substantially lower prices (about 54% and 70% lower, respectively). Among neighborhood groups, Manhattan and Brooklyn listings are significantly more expensive than the reference group, consistent with real-world pricing trends.

Overall, the transformed model appears to satisfy the core regression assumptions well enough to support early conclusions.

4 Model Selection

To build a reliable and meaningful model of nightly Airbnb prices, we compared two versions of a linear regression model, one using the original variables and one using transformed variables. We then checked for any problematic observations that might have had too much influence on the model and finally tested for multicollinearity among the predictors. Each decision we made was backed by data, and we made sure to explain why every choice was appropriate.

4.1 Transformations: What We Did and Why

We started by comparing two models:

- `model_raw` used the original versions of the variables.
- `model_xfrm` used transformations for some variables that had skewed distributions.

The variables we transformed in the second model included:

- *price* → log(price)
- *minimum_nights* → $\sqrt{\text{minimum\_nights}}$
- *availability_365* → $\sqrt{\text{availability\_365}}$
- *number_of_reviews* → $\log(1 + \text{number\_of\_reviews})$
- *reviews_per_month* → $\log(1 + \text{reviews\_per\_month})$

These transformations helped reduce skew and stabilize variance. For example, the square root transformation helped compress long tails from very high minimum night values. The log transformations captured the idea that extra reviews become less informative after a certain point (e.g., going from 0 to 10 reviews matters more than going from 100 to 110).

The price variable itself was also log-transformed to deal with skew and uneven variance. This helped make the relationship between the response and predictors more linear and easier to interpret. Since our response was ‘log(price)‘, exponentiating the coefficients let us talk about percent changes in price.

What was interesting was how even small transformations like $\sqrt{\text{minimum\_nights}}$ noticeably improved residual plots and made the relationship between predictors and price much clearer. While transformations can sometimes complicate interpretation, in our case they actually made the model more intuitive.

4.2 Model Comparison Metrics

We used AIC, BIC, and adjusted R**²** to compare the two models

Table 1: Model Comparison: Raw vs. Transformed

Model	AIC	BIC	Adjusted R^2
Raw	2065.441	2135.064	0.4961
Transformed	2048.874	2118.497	0.5014

The transformed model had a noticeably lower AIC (better) and a slightly higher adjusted R^2. These results told us that the transformations improved the model, so we chose the transformed model (`model_xfrm`) to continue our analysis.

4.3 Influential Observations and Outliers

After we chose our working model, we looked for listings that might have had too much influence on the results. These are usually outliers, either with unusual predictor values or with predictions that are way off.

Table 2: Criteria for Flagging Influential Points

Tool Used	**Purpose**	**Threshold**
Standardized Residuals	Detect extreme prediction errors	> 3
Leverage (hat values)	Detect unusual predictor combinations	$> 2\times$ average leverage
Cook's Distance	Assess model sensitivity to individual cases	$> 4/n$ (with $n = 1565$)

Applying this tri-criterion approach to the full dataset ($n = 1565$) yielded 122 flagged observations, which is about 7.8% of the sample. Table 3 summarizes a subset of the most influential cases, sorted by Cook's Distance.

These listings were not random. Many were either extremely expensive entire homes in high-demand neighbourhoods like Manhattan or unusually low-priced listings with unique availability patterns. Their prices deviated from predicted values more than others, and their leverage scores suggested unusual combinations of predictor values.

To assess whether these points materially distorted the model, we refitted the regression without them (`model_drop`) and compared the coefficients against the full model (`model_work`). Table 4 shows the relative differences:

Table 3: Subset of Most Influential Listings (sorted by Cook's Distance)

Listing ID	Cook's D	Leverage	Std. Residual
25285364	0.074	0.021	-3.9
14097205	0.068	0.018	$+3.7$
37985340	0.065	0.019	-3.5

Table 4: Coefficient Stability After Dropping Influential Points

Term	**Full Model**	**Dropped**	**Relative Diff**
calculated_host_listings_count	0.00075	–0.00877	19.28%
ln_reviews	–0.00078	0.00821	12.59%
ln_reviews_month	–0.09290	–0.08013	13.74%
Intercept	4.799	5.014	4.45%

While some differences were noticeable (12–19%), most core predictors stayed fairly stable. This suggested that even though these listings had some influence, they didn't completely change the story. Since they represent real market cases (e.g. luxury listings in Manhattan), we decided to keep them in the final model but make note of their influence.

While keeping influential points is often risky, in this case, it reflects the actual diversity in the marketplace. Still, we interpret coefficients for these predictors with caution, especially when discussing edge cases or high-end properties.

4.4 Predictor Selection and Multicollinearity

Next, we checked whether any predictors were too similar to each other (which could confuse the model). We did this using Variance Inflation Factors (VIFs). A VIF above 5 is usually a cause of concern.

In our model, all VIFs were well below that threshold. The highest VIF was around 2.0, which is very safe. For example:

- `room_typePrivate room`: VIF = 1.0000
- $\sqrt{\text{avail\_365}}$: VIF = 2.02

We briefly considered whether `ln_reviews` and `ln_reviews_month` might overlap too much since both relate to guest feedback. However, they represent distinct concepts: one reflects the cumulative volume of feedback, while the other captures recent listing activity. Their low VIFs confirmed there was no strong collinearity. Notably, despite their apparent similarity, both variables *contributed* meaningfully to the model, reinforcing the interpretation that engagement quantity and recency capture different behavioural dimensions among guests.

We kept *all* predictors in the model, and each one has a clear reason for being included:

- `calculated_host_listings_count`: main variable of interest, measures host scale

- `neighbourhood_group`: location-based price differences
- `room_type`: affects price due to level of privacy
- $\sqrt{\text{min\_nights}}$ and $\sqrt{\text{avail\_365}}$: capture booking rules and availability
- `ln_reviews` and `ln_reviews_month`: show social proof and recent activity

After transforming the right variables, checking for overly influential listings, and confirming that no predictors caused redundancy, we finalized a clean and reliable regression model. This model explains about 50.1% of the variation in nightly Airbnb prices. Its coefficients are interpretable, stable, and based on real-world logic.

While our model performs well, we always need to interpret with caution, especially around cases that fall outside the average, such as luxury listings or inactive hosts. Nonetheless, we are confident in its use for answering our research question about how host scale affects pricing.

5 Final Model Inference and Results

Table 5: Final Model Inference and Results

Term	**Exp(Est.)**	**Std. Err.**	**2.5% CI**	**97.5% CI**	***p*-value**
(Intercept)	121.437	0.107	98.491	149.729	2.5×10^{-283}
calculated_host_listings_count	1.000	0.001	0.999	1.002	4.9×10^{-1}
neighbourhood_groupBrooklyn	1.275	0.102	1.043	1.559	1.8×10^{-2}
neighbourhood_groupManhattan	1.728	0.102	1.414	2.113	1.1×10^{-7}
neighbourhood_groupQueens	1.094	0.106	0.888	1.346	4.0×10^{-1}
neighbourhood_groupStaten Island	0.946	0.160	0.692	1.294	7.3×10^{-1}
room_typePrivate room	0.456	0.024	0.435	0.478	1.1×10^{-175}
room_typeShared room	0.302	0.080	0.258	0.353	6.7×10^{-48}
sqrt_min_nights	0.926	0.012	0.905	0.947	4.8×10^{-11}
sqrt_avail_365	1.018	0.002	1.014	1.021	1.3×10^{-20}
ln_reviews	0.999	0.012	0.976	1.023	9.5×10^{-1}
ln_reviews_month	0.911	0.030	0.858	0.967	2.3×10^{-3}

5.1 Narrative Interpretation and Model-Fit Metrics

Table 5 (above) summarises the final model's coefficients, standard errors, confidence intervals, and p-values.

Our model uses log(price) as the response, so exponentiated coefficients indicate the multiplicative effect on price for a one-unit change in the predictor.

- **calculated_host_listings_count** has a coefficient of 1.0007, meaning the nightly price rises by $\approx 0.07\%$ for each additional listing a host manages (statistically insignificant). Host scale alone therefore does not meaningfully influence pricing, contrary to what one might expect from a more professionalised host.
- **neighbourhood_group:** compared with Queens, Brooklyn listings cost about 20–30% more, Bronx and Staten Island are slightly cheaper, and Manhattan listings are typically more than double the nightly price of Queens. Location is thus a major driver of pricing; centrality and prestige carry significant premiums.
- **room_type:** private rooms ($\exp\beta = 0.46$) are priced $\approx 54\%$ lower than entire homes; shared rooms ($\exp\beta = 0.30$) are $\approx 70\%$ lower. This aligns with expectations: privacy is highly valued.
- $\sqrt{\text{minimum\_nights}}$: each unit increase multiplies price by 0.93; longer minimum stays slightly reduce pricing, perhaps to attract longer commitments.
- $\sqrt{\text{availability\_365}}$: coefficient 1.018 suggests that greater availability modestly increases price; full-time rentals may signal reliability.
- log(reviews) shows a tiny effect, implying that once a listing gains credibility, additional reviews scarcely move the price.
- log(reviews_month) ($\exp\beta = 0.911$) indicates that higher recent review rates correlate with lower prices—possibly because budget-friendly listings receive reviews more frequently.

The coefficients listed above reveal valuable insights into the research question. Host portfolio size has little to no effect on price after controlling for room type, location, and listing features. THis implies that professionalization via scale does not necessarily translate into premium pricing, possibly due to price competition or standardized service expectations on Airbnb. Location and privacy level dominate price determination, showing clear and intuitive effects. Guest feedback adds marginal value and recent review activity could even suggest budget listings. Booking constraints have small but interpretable roles in shaping pricing dynamics.Our model proves a clear answer to the research question: while larger portfolios may hint at professionalism, they do not significantly affect Airbnmb nightly pricing after accounting for the other key factors. Instead, location, privacy level, and availability are the most influential determinants of price.

As have listed above, we chose the second transformed model rather than the first raw model which used the original versions of the variables, since some variables that had skewed distributions need a transformation. Here we assess the model performance by using different metrics(Table 6).

Table 6: Model-Fit Metrics for Competing Specifications

Model	**AIC**	**BIC**	**Adjusted** R^2
Raw (untransformed)	2065.441	2135.064	0.4961
Transformed (final)	2048.874	2118.497	0.5014

The normal R2 provides a measure of how well observed outcomes are replicated by the model. For the adjusted R2 specifically, the model complexity (i.e. number of parameters) affects the R2 and thereby captures their attributes in the overall performance of the model. The closer R2 is to 1, the greater is said to be the degree of linear association between the covariates X and the response variable Y. Therefore, adjusted R2 can be interpreted as a less biased estimator of the population R2. Here the difference of the R-squared value between two models is small, making it an indecisive factor.

Overall, host portfolio size has little impact on price after controlling for other variables. Location and privacy level dominate price determination.

6 Discussion & Conclusion

6.1 Purpose of Final Model

Our goal is to produce a model that determines how Airbnb listing counts influence prices. By parsing our dataset and tailoring it to our needs, analyzing the portfolios of Airbnb hosts has yielded several valuable insights for our purposes. With data from thousands of listings across New York City per Kaggle's New York City Airbnb Open Data[1], analysis has shown a distinct relationship between prices and chosen predictors.

6.2 Findings and Interpretations

Despite the purpose of our research and choice of main predictor, calculated_host_listings_count, patterns suggest a stronger influence from other predictors within the majority of sampled listings. Prices are predominantly determined by our alternative predictors, i.e. neighbourhood_group (region), room_type (entire home vs. shared room), etc.. While a higher volume of host listings is more akin to the professional hospitality industry, these values are not a strong or consistent factor in the price of Airbnb listings.

Regarding computed values, evidence of host listing counts proving notably ineffective is similarly found in its coefficient. Examining their values, the coefficient for calculated_host_listings_count has the smallest impact at 1.0007, suggesting a ∼0.07% difference in price for each host's listing. (Sat, Section 5)

The geographical coefficients are strongest for Manhattan at 1.728, followed by Brooklyn at 1.275. Listings within these two areas can be interpreted as notably influenced by regional factors, such as local demand and perceived quality, in comparison to those of other sampled spaces. Our final model shows that this holds compared to many of our remaining predictors, including our main choice. Similarly, room_typeShared_room uses a small but strong coefficient of 0.302, suggesting ∼70% lower costs compared to average listings for whole homes. By comparison, room_typePrivate_room demonstrates a ∼54% reduction from the standard.

As noted in Section 4: Model Selection, review quantity and frequency are unique in reflecting a listing's popularity and activity, but carry less influence by comparison; ln_reviews and ln_reviews_month show low-impact coefficients of 0.999 and 0.911, suggesting little strength to these elements.

Due to its smallest coefficient compared to others, calculated_host_listings_count can be inferred to be a part in deciding the price of an Airbnb listing, however weak it may be. In the same manner that reviews do little, in both our model and reality, to alter a host's valuation of their services, so too does their sum of offerings. Instead, factors normally explored while browsing listings, i.e. region and room type, have a greater effect on price.

6.3 Recommendations

Our analysis shows that Airbnb pricing is mostly shaped by listing features like location and room type—not by how many listings a host manages. Based on these findings, we suggest the following:

- **For guests:** Focus on where the listing is and what kind of room it offers. These factors have the biggest influence on price. For example, private rooms tend to cost about 54% less than entire homes, and shared rooms about 70% less. Listings in central areas like Manhattan are also significantly more expensive than those in outer boroughs like Queens.

- **For hosts:** Simply having more listings doesn't lead to higher prices. The effect of host portfolio size was minimal and not statistically significant in our model. Hosts looking to increase earnings should instead focus on features that travelers value—like better locations, more privacy, and reliable availability.
- **For Airbnb:** When ranking listings or promoting hosts, the number of listings managed shouldn't be a major factor. Our results suggest that being a larger host doesn't necessarily mean offering better value or commanding higher prices.
- **For future studies:** While host portfolio size doesn't matter much now, that could change as the platform evolves and professional hosts become more common. Future research using newer datasets could reveal whether this trend shifts over time.

In short, it's the quality and characteristics of the listing and not the size of the host's portfolio that drive price. Location, room type, and availability continue to be the most important factors when it comes to Airbnb pricing.

6.3 Insight and Analysis

Although this data was collected in 2019, and this research thus pertains to the period, this information may still hold weight in the current year; hotels and Airbnb services have persisted through global disruptions of the early 2020s. The industry is not the same as it was following Kaggle's publishing, and for comparisons with present data, the values of this study remain relevant well beyond its time. This assumes significant changes in hotel businesses, regardless of the differences in scope between professional and individual services.

This also aids comparisons with future datasets, assuming pricing strategies shift after the completion of this study. Adjusting for inflation, models created using newer datasets may produce different results from 2019 or 2025 Airbnb data, and may be a point of reference for studies under a changed hospitality industry.

An ideal future analysis would have more detailed data for a deeper study. For a research question like this and a dataset of this size, linear regression still requires careful cleaning and thoughtful preparation – much like what we've done in this study. Different participants may utilize different transformations and produce different results. This variation is natural;

modelling involves experimentation, and the path to a final model commonly depends on how researchers interpret the data and define what makes a model "optimal." Whether the priority is interpretability or theoretical consistency, the choice of the final model will ultimately reflect a researcher's analytical judgment.

7 Contributions

- Hen Lin: Introduction and Conclusion.
- Ben Sat & Yilin Liu: Data exploration and variable summaries.
- Long Pingshan & Nicholas Koh: Latex formatting, Research question, dataset selection, and preliminary analysis.

Bibliography

Dgomonov. (2019, August 12). *New York City Airbnb open data.* Kaggle. https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Inside Airbnb. (2019). *Get the data.* Inside Airbnb. http://insideairbnb.com/get-the-data.html