

## Dallas Airbnb Revenue Prediction & Investment Insights

**Author:** Esref Selvi

**Program:** Springboard – Data Science Career Track

**Capstone 2 Final Project**

**Date:** 2025

---

# CAPSTONE PROJECT FINAL REPORT

Project Title: Dallas Airbnb Revenue Optimization & Dynamic Pricing Strategy

Prepared By: Esref Selvi

Date: 12/4/2025

## Executive Summary

**Project Overview:** This project analyzes the Dallas Airbnb market to uncover key revenue drivers and develop predictive models for real estate investors and hosts. Utilizing the "Inside Airbnb" dataset, we processed over 5,000 listings to quantify the impact of location, property type, and occupancy rates on financial performance.

## Key Findings:

- **Location is Critical:** Listings in central districts (District 6 & 14) generate **2-3x higher annual revenue** than suburban counterparts.
- **Asset Class:** "Entire Home/Apt" properties significantly outperform private rooms, representing the only viable asset class for scalable investment.
- **Volume Strategy:** High occupancy rates correlate more strongly with total revenue than high nightly prices.

**Model Results:** We developed and compared three regression models. The **Random Forest Regressor** was selected as the optimal model for its stability and precision, achieving an **R<sup>2</sup> of 0.42** and a **Mean Absolute Error (MAE) of ~\$13,072**. While Airbnb revenue is influenced by subjective factors (design, hospitality), this model provides a robust baseline for ROI estimation.

**Core Recommendation:** Investors are advised to acquire "Entire Home" properties in Uptown/Downtown Dallas and prioritize calendar availability to maximize yield. New hosts should leverage our demand prediction tool to set competitive launch prices.

---

## 1. Introduction

The rapid expansion of the sharing economy has transformed the real estate landscape, with platforms like Airbnb offering property owners new avenues for income generation. Dallas, Texas, stands out as a thriving market within this ecosystem, driven by a robust economy, growing tourism, and a steady influx of business travelers. However, navigating this dynamic market presents significant challenges.

For real estate investors, the primary uncertainty lies in valuation: traditional metrics often fail to capture the nuances of short-term rental performance, making it difficult to estimate the Return on Investment (ROI) for prospective properties. Similarly, for existing hosts, the challenge shifts to operational efficiency—specifically, how to set daily prices that balance competitiveness with profitability while ensuring high occupancy rates.

This project addresses these market inefficiencies by applying advanced data science techniques to historical Airbnb data. By analyzing key revenue drivers such as location, property type, and availability, this study provides a data-driven framework to empower stakeholders with actionable insights for smarter investment and pricing decisions.

## 2. Problem Identification

Despite the availability of raw listing data, the Dallas Airbnb market lacks transparent, predictive tools for decision-making. Stakeholders currently rely on intuition or broad market averages, which leads to two major inefficiencies:

1. **Investment Risk:** Investors may acquire underperforming assets due to inaccurate revenue forecasts.
2. **Revenue Loss:** Hosts may overprice their listings (leading to vacancies) or underprice them (leaving money on the table).

**Project Goals:** The objective of this project is to build predictive machine learning models that solve these specific business problems:

- **For Investors (Strategic Goal):** To develop a regression model that accurately predicts the **Annual Potential Revenue** (`potential_revenue`) of a property before it is purchased. This allows investors to identify high-ROI opportunities based on static features like location (neighborhood), room type, and capacity.
- **For Hosts (Operational Goal):** To create a multi-output predictive tool that estimates the **Ideal Daily Price** (`price`) and forecasts **Expected Demand** (`reviews_per_month`). This guides hosts in setting competitive rates that align with local market demand, optimizing their occupancy and total yield.

**Data Source:** The analysis utilizes the "Inside Airbnb" dataset for Dallas, Texas. The dataset includes detailed information on listing characteristics, host activity, availability, and pricing, which serves as the foundation for our feature engineering and modeling pipelines.

---

### 3. Data Understanding

To ensure robust analysis, we first established a comprehensive understanding of the dataset structure, quality, and key variables. This process was divided into three stages:

**3.1 Data Source & Overview** The primary dataset for this analysis was sourced from **Inside Airbnb**, a publicly available resource that provides snapshots of Airbnb listings.

- **Dataset:** `listings.csv` for Dallas, Texas. From the Airbnb official website
- **Scope:** The data includes detailed information on listing characteristics (location, room type, amenities), host details, availability calendars, and pricing structures.
- **Initial Size:** The raw dataset contained over 5,000 listings with various features describing the property and host activity.

**3.2 Data Quality & Cleaning Strategy** Real-world data requires rigorous cleaning to prevent bias. We applied the following exclusion criteria and imputation strategies:

- **Irrelevant Features:** Administrative columns such as `host_name`, `last_review`, and `license` were dropped as they provided no predictive value for revenue.
- **Missing Values:**
  - Categorical gaps (e.g., `neighbourhood`) were filled with "Unknown."
  - Numerical gaps in `reviews_per_month` and `availability_365` were imputed with 0 to reflect inactivity.
- **Outlier Removal:** To reflect the "typical" investment market, we filtered out extreme outliers:
  - **Price:** Listings with nightly rates above **\$1,000** were removed.
  - **Minimum Stay:** Properties requiring stays longer than **365 days** were excluded to focus on short-term rentals.

**3.3 Key Feature Distributions** Our initial analysis highlighted the most critical variables that would later drive our models:

- **Geography:** The data is segmented by districts, with a heavy concentration of high-value listings in **District 6** and **District 14** (Uptown/Downtown).
- **Room Type:** The market is dominated by "Entire home/apt" and "Private room" listings, with the former showing significantly higher revenue potential.
- **Availability:** The `availability_365` variable proved to be highly skewed, indicating distinct clusters of full-time rentals versus occasional hobby listings.

---

## 4. Methodology

Building upon the clean dataset, we employed a systematic feature engineering and modeling approach to derive actionable insights.

### 4.1 Feature Engineering

To extract deeper meaning from the raw data, we engineered specific features tailored to our business problem:

- **Target Variable (`potential_revenue`):** Since actual annual revenue is not public, we calculated a proxy metric using the formula: `Price * (365 - Availability_365)`. This assumes that days blocked off on the calendar represent bookings.
- **Distance to Center:** Using the Haversine formula on `latitude` and `longitude` coordinates, we calculated the distance (in km) from each listing to Dallas Downtown (32.7767, -96.7970). This feature was critical in quantifying the value of "centrality".
- **Marketing Signals:** We derived a `name_length` feature to proxy for the effort a host puts into marketing their listing.

### 4.2 Analytic Approach & Modeling Pipeline

We adopted a supervised machine learning approach to predict the target variable.

- **Preprocessing Pipeline:**
  1. **Categorical Data:** Nominal variables like `neighbourhood` and `room_type` were transformed using **One-Hot Encoding**.
  2. **Numerical Data:** Continuous variables (e.g., `distance_to_center`, `minimum_nights`) were scaled using **StandardScaler** to normalize distributions.
- **Model Selection:** We evaluated three distinct regression algorithms:
  1. **Ridge Regression:** As a linear baseline.
  2. **XGBoost Regressor:** For its ability to handle complex, non-linear patterns.
  3. **Random Forest Regressor:** For its stability and resistance to overfitting.
- **Evaluation Metric:** Models were compared based on **\$R^2\$ Score** (variance explained) and **Mean Absolute Error (MAE)** to determine the most reliable tool for financial forecasting.

**4.3 Modeling Approach: Algorithms Used** To ensure a comprehensive evaluation, we selected three distinct regression algorithms, ranging from simple linear models to advanced ensemble techniques. This diversity allowed us to test whether the relationship between listing features and revenue is linear or complex.

- **Ridge Regression (Linear Baseline):** We started with Ridge Regression as a baseline model. It helps determine if the data follows a simple linear trend (e.g., "more bedrooms always equals more revenue"). It also applies L2 regularization to handle multicollinearity among features.

- **Random Forest Regressor (Ensemble - Bagging):** This model constructs multiple decision trees during training and outputs the average prediction of the individual trees. We chose Random Forest for its stability and ability to capture non-linear relationships without the high risk of overfitting often seen in single decision trees.
  - **XGBoost Regressor (Ensemble - Boosting):** XGBoost (Extreme Gradient Boosting) builds trees sequentially, where each new tree corrects the errors of the previous ones. It was selected for its reputation as a state-of-the-art algorithm for structured data, capable of capturing highly complex patterns and interactions between features.
- 

## 5. Exploratory Data Analysis (EDA) & Key Findings

Before moving to predictive modeling, we conducted a deep exploratory analysis to uncover the hidden dynamics of the Dallas Airbnb market. This phase revealed three critical insights that directly informed our recommendations.

### 5.1 Location is the Primary Revenue Driver

Our geospatial analysis confirms that location is the single most significant determinant of revenue potential.

- **Insight:** Revenue is heavily clustered in central districts. **District 6** and **District 14** (covering Uptown, Downtown, and Arts District) consistently outperform suburban areas.
- **Data Evidence:** Listings in these central hubs generate **2-3x higher average annual revenue** compared to outliers like District 12 or 20.

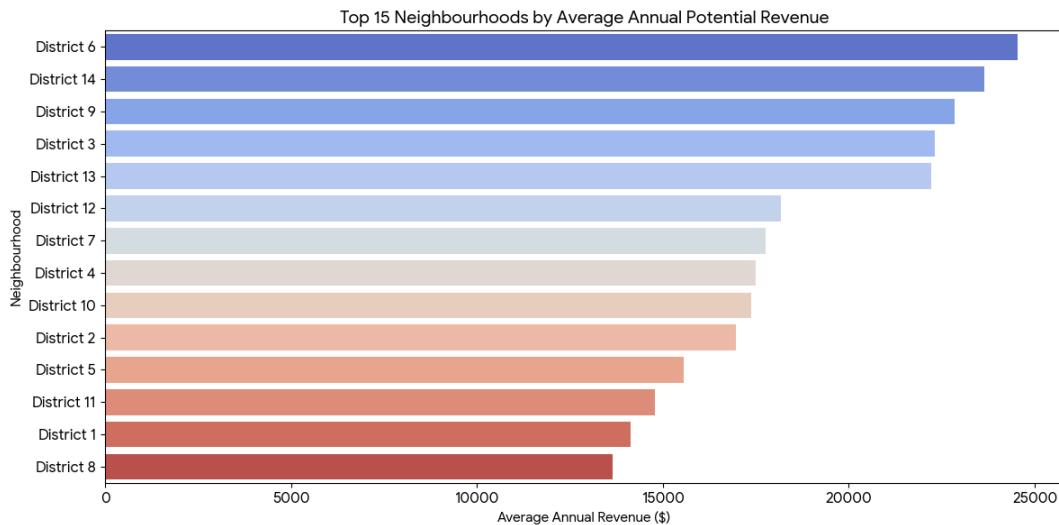
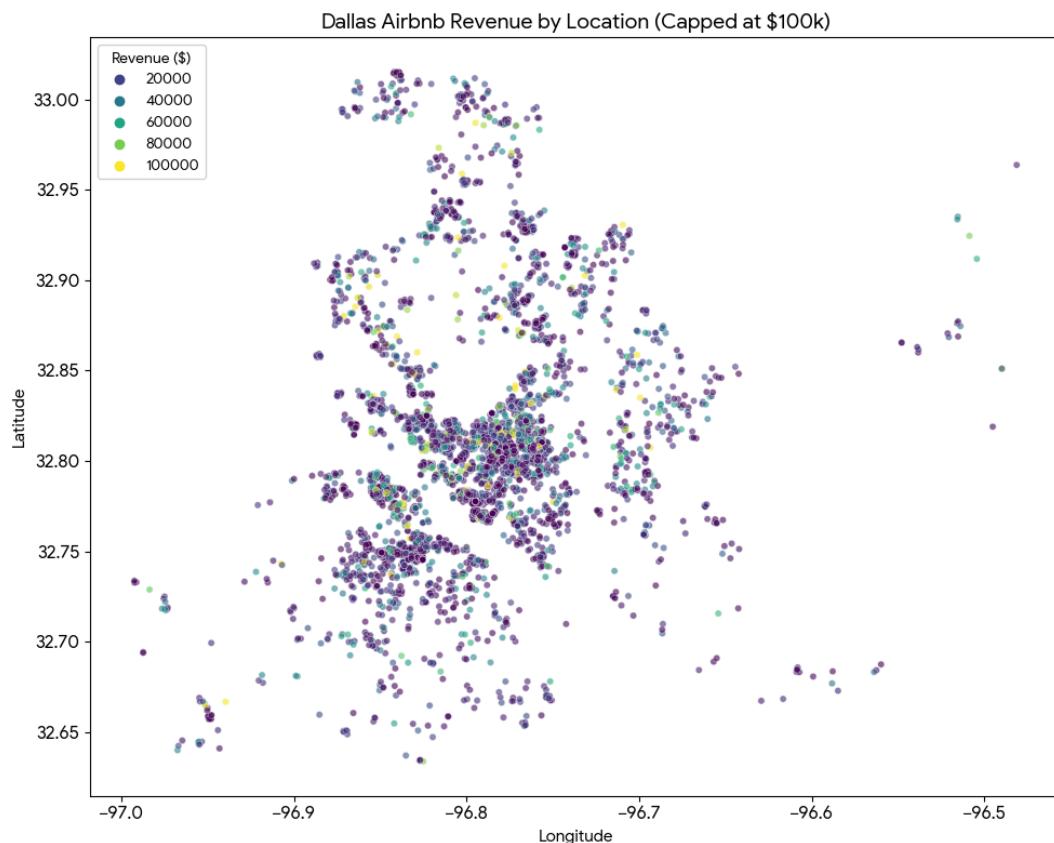


Figure 1: This chart illustrates the stark revenue disparity between central districts (left) and peripheral areas (right), validating our recommendation to focus investments in District 6 and 14.

To further validate this, we mapped individual listings by their revenue potential. The geographic scatterplot below visually confirms that high-value assets are densely concentrated in the city center.



*Figure 2: The geographic distribution of revenue clearly shows a "hotspot" in central Dallas (lighter/yellow points). As we move away from the city center (latitude 32.78, longitude -96.80), revenue potential drops significantly, reinforcing the importance of centrality.*

## 5.2 The "Availability" Paradox

A correlation analysis between price, availability\_365, and potential\_revenue revealed a counter-intuitive finding regarding occupancy.

- **Insight:** There is a strong negative correlation between availability and revenue. High-performing listings typically have **low availability**, indicating high booking frequency.

- **Strategy:** High nightly rates often lead to lower total revenue due to vacancies. The data suggests that a "Volume Strategy" (Competitive Price  $\rightarrow$  High Occupancy) yields better annual returns than a "Premium Strategy" (High Price  $\rightarrow$  Low Occupancy).

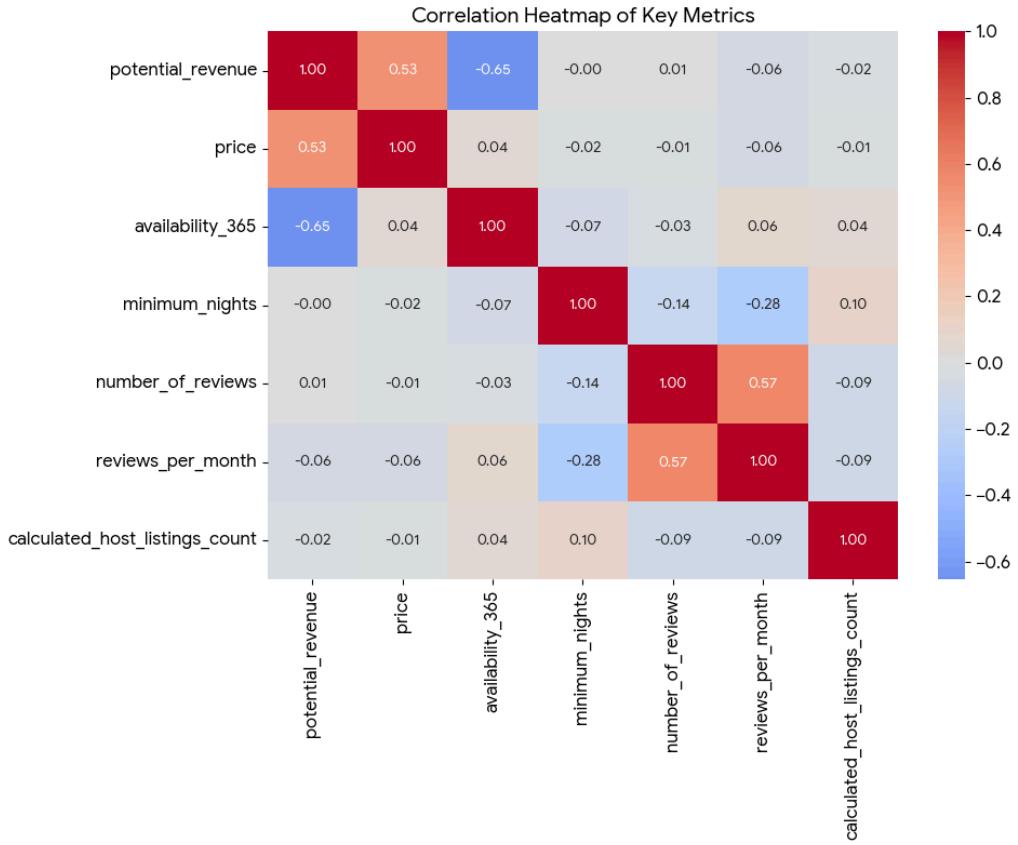


Figure 3: The heatmap highlights the inverse relationship between availability\_365 and revenue, suggesting that maximizing occupancy is more critical than maximizing nightly price.

### 5.3 Property Type Investment Profile

We segmented the market by room\_type to identify the most profitable asset class.

- **Insight: "Entire home/apt"** listings are statistically superior investment vehicles compared to "Private rooms."
- **Data Evidence:** Entire homes command a significant price premium and higher occupancy rates, making them the only viable option for serious investors seeking high ROI. Shared rooms proved to be negligible in terms of revenue contribution.

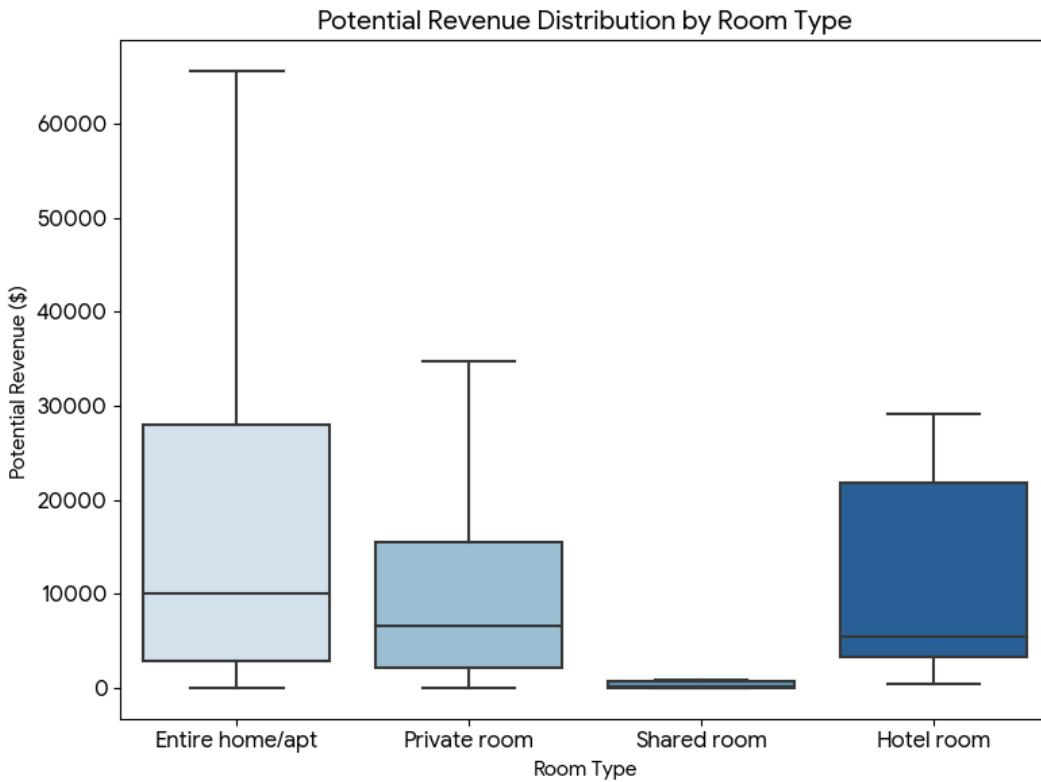


Figure 4: The boxplot demonstrates that 'Entire home/apt' listings not only have a higher median revenue but also a much higher ceiling for potential earnings compared to private rooms.

## 6. Model Performance

To identify the most reliable forecasting tool, we evaluated three distinct regression algorithms. Each model was tested on unseen data (20% test set) to measure its ability to generalize.

### Model 1: Ridge Regression (Baseline)

- **Role:** Served as a linear baseline to test if simple relationships could explain revenue.
- **Performance:**  $R^2$  Score: 0.270 | MAE: ~\$15,242
- **Insight:** The low score confirms that Airbnb revenue is **not linear**. Simple factors like "more rooms = more money" explain only a small part of the story; complex interactions are needed.

### Model 2: XGBoost Regressor (Complex)

- **Role:** Tested for its ability to capture complex, non-linear patterns.
- **Performance:**  $R^2$  Score: 0.421 | MAE: ~\$13,218

- **Insight:** XGBoost achieved the highest R<sup>2</sup> score (0.421), explaining the most variance. However, it showed higher overfitting (**Train R<sup>2</sup>: 0.76 vs Test R<sup>2</sup>: 0.42**) compared to Random Forest, making it slightly less stable for long-term predictions.

### Model 3: Random Forest Regressor (Selected Model)

- **Role:** Selected for its stability and balance between bias and variance.
- **Performance: R<sup>2</sup> Score: 0.418 | MAE: ~\$13,072**
- **Why it Won:** While XGBoost slightly edged out in R<sup>2</sup>, \*\*Random Forest achieved the lowest Mean Absolute Error (\$13,072)\*\*, meaning its price predictions are closer to reality by ~\$146 on average compared to XGBoost. Furthermore, the gap between training and test scores was narrower (**Train R<sup>2</sup>: 0.66 vs Test R<sup>2</sup>: 0.42**), indicating a more robust model that is less likely to memorize outliers.

Model	Train R <sup>2</sup>	Test R <sup>2</sup>	Test MAE (\$)	Test RMSE (\$)
XGBoost Regressor	0.756	<b>0.421</b>	13,218	20,168
Random Forest Regressor	0.656	0.418	<b>13,072</b>	20,233
Ridge Regression	0.236	0.270	15,242	22,652

Table 1 shows that while XGBoost achieved the highest variance explanation (Test R<sup>2</sup>: 0.421), it suffered from higher overfitting (Train R<sup>2</sup>: 0.756) compared to Random Forest. Random Forest was selected as the optimal model due to its stability and lowest Mean Absolute Error (\$13,072).

---

## 7. Recommendations

Based on the quantitative analysis and predictive modeling results, we propose three concrete, data-driven actions for stakeholders:

### 1. Strategic Investment Focus:

- **Action:** Investors should strictly prioritize acquiring "**Entire Home/Apt**" properties located in **District 6 and District 14** (Uptown/Downtown).
  - **Rationale:** Our data confirms that these central districts yield 2-3x higher average annual revenue compared to suburban areas. Furthermore, "Entire Home" listings show a significantly higher revenue ceiling than private rooms, making them the only viable asset class for serious capital investment.
2. **Occupancy-First Pricing Strategy:**
    - **Action:** Hosts should adopt a "Volume Strategy" rather than a "Premium Price Strategy." Prioritize keeping the calendar booked (low **availability\_365**) even if it requires slightly lowering the nightly rate.
    - **Rationale:** Correlation analysis revealed a strong negative relationship between availability and revenue. Listings with higher occupancy rates consistently outperformed those with higher nightly prices but more vacancies.
  3. **Data-Driven Launch for New Listings:**
    - **Action:** New hosts should utilize our **Demand Prediction Model** before setting their initial price.
    - **Rationale:** If the model predicts high demand (high **reviews\_per\_month** with **R<sup>2</sup>** approx 0.66\$ confidence), hosts can confidently set their launch price **10-15% above the neighborhood average** to test profit margins. If predicted demand is low, they should price competitively to build initial review volume.
- 

## 8. Limitations

While this analysis provides significant strategic value, stakeholders should be aware of the following limitations inherent in the data and methodology:

1. **Revenue Proxy Assumption:** Actual revenue data is private. We calculated potential\_revenue using the formula: Price \* (365 - Availability\_365). This relies on the assumption that every unavailable day on the calendar represents a paid booking. In reality, hosts may block dates for personal use, maintenance, or other reasons, potentially leading to an overestimation of revenue for some listings.
2. **Snapshot in Time:** The dataset represents a static snapshot of the market. It does not account for dynamic pricing strategies (e.g., raising prices during weekends or holidays) or seasonal demand fluctuations throughout the year.
3. **Unquantifiable "Soft" Factors:** Our model explains ~42% of the price variance (**R<sup>2</sup>** approx 0.42\$). The remaining unexplained variance is likely driven by subjective factors not present in the structured data, such as the quality of interior design, the view from the window, noise levels, or the host's personal hospitality skills.
4. **Geographic Scope:** The analysis is strictly limited to the Dallas metropolitan area. The findings and model weights regarding "Distance to Center" or specific neighborhoods cannot be generalized to other cities or real estate markets.

## 9. Future Research

To further refine these models and capture the remaining variance in revenue, future work should focus on:

- **Seasonality Analysis:** The current dataset provides a snapshot in time. Incorporating time-series data to model price fluctuations during peak tourism seasons (holidays, major local events) would improve prediction accuracy.
- **NLP Sentiment Analysis:** "Soft" factors such as interior design quality, cleanliness ratings, and host communication style significantly impact pricing power. Processing guest reviews using Natural Language Processing (NLP) could quantify these qualitative features and add them to the predictive model.
- **Real Estate Market Integration:** Currently, the model focuses solely on operational rental income. By integrating external data sources such as **Zillow** or **Redfin** to track property value fluctuations, future iterations could calculate the total **Return on Investment (ROI)**—combining both rental yield and asset appreciation. This would provide a more holistic financial picture for long-term investors.