GROUP AIRBNB

1. Executive Summary

This report presents the culmination of our multi-phase analysis focused on identifying investment-worthy Airbnb listings in San Francisco and San Diego. By integrating machine learning models, geospatial analysis, and engineered features, we created a robust investment scoring framework. Our findings, delivered via an interactive Power BI dashboard, support strategic decision-making by investors and hosts, with actionable insights drawn from detailed listing analysis across both cities.

2. Project Objectives

- Predict Airbnb listing prices using advanced machine learning models
- Engineer and select impactful features (e.g., price per accommodate, proximity to downtown)
- Develop a composite investment score (out of 100) to evaluate listing potential
- Integrate additional features like simulated safety score and sentiment polarity
- Visualize insights using Power BI to assist hosts, investors, and Airbnb stakeholders
- Ensure scoring framework scalability for broader market applicability beyond initial city scope

3. Dataset Summary

- Source: InsideAirbnb.com – listings from both San Diego and San Francisco
- Structure: Combined datasets included listing, calendar, review, and neighborhood data
- Key Fields:
    - price, availability_365, review_scores_rating, latitude, longitude
- Engineered Features:
    - price_per_accommodate, bed_bath_ratio, distance_to_downtown_km, investment_score
- Data Cleaning:
    - Removed nulls in pricing/location
    - Imputed missing review scores

- Normalized numerical values and encoded categorical fields

4. **Modeling & Evaluation**

We trained and compared the following regression models:

- Linear Regression
- Random Forest
- Neural Network
- XGBoost (final model)

Model Results:

- XGBoost was selected due to:
  - R^2: 0.81
  - Lowest RMSE and MAE among all models
  - Stable performance across a wide range of prices
- Model performance was consistent across both cities, with accurate predictions maintained throughout San Diego and San Francisco datasets.

Evaluation Additions:

- Feature Importance Plot (XGBoost): highlighted top predictors like availability and rating
- Residual Histogram: confirmed normal error distribution

5. **Investment Scoring Framework**

A composite investment score (0–100) was created based on:

- Predicted price
- Review score rating
- Sentiment polarity
- Safety score (simulated)
- Availability
- Distance to downtown

Listings scoring 80+ were considered high-value. These listings were predominantly found in downtown San Diego, San Francisco, Pacific Beach, and coastal neighborhoods. The framework is structured for extensibility and may be adapted for use in other metropolitan markets with minimal recalibration.

6. **Power BI Dashboard Insights**

A fully interactive dashboard was developed with three primary layers:

Investor Dashboard:

- Metrics: ADR, RevPAR, Occupancy, ROI
- Filters: Room Type, Neighborhood, Price Range
- Top ROI zones: Downtown, Little Italy, Pacific Beach

Host Quality Dashboard:

- Metrics: Host response rate, average review score, review volume
- Listings with high ratings and fast responses had 20–30% higher ROI

Map Dashboard:

- Heatmaps of investment score and price
- Clustered markers show high-concentration of premium listings
- The dashboard also enables comparative filtering between San Francisco and San Diego, revealing cross-market trends in listing performance.

7. **Key Findings & Recommendations**

Findings:

- XGBoost was highly accurate in predicting prices
- Availability, review scores, and distance to downtown are major drivers
- Listings near tourist-friendly, central zones have the highest profitability

Recommendations:

- Investors: Prioritize listings with scores >80 near downtown/coastal areas
- Hosts: Focus on amenities, cleanliness, and guest response times
- Airbnb: Consider integrating sentiment/safety insights into their ranking algorithms

- Developers: Expand this into a pricing recommendation API or investment decision tool

These findings reflect patterns consistent across both cities, indicating that the investment scoring model is a scalable tool for broader urban applications.

8. **Limitations**

- Safety data was simulated; real crime datasets were unavailable
- Guest reviews were not included in text form due to dataset constraints
- The model is tailored to San Diego and San Francisco only; generalization would require re-training
- Seasonality trends (calendar-based revenue forecasting) were not modeled

Mitigating these limitations in future work—such as sourcing regional crime statistics and including seasonal forecasting—will improve scoring accuracy.

9. **Future Work**

- Extend the project to include comparative multi-city analysis (e.g., New York, Miami, Austin)
- Add deep sentiment analysis using guest reviews with NLP tools (BERT, VADER)
- Integrate real-world crime statistics for accurate safety modeling
- Apply Prophet or LSTM for time series-based occupancy and revenue prediction
- Build a Flask/Streamlit-based app to let investors input criteria and rank listings

These enhancements will support a dynamic investment tool adaptable across diverse geographies.

Conclusion

This final milestone successfully demonstrates the power of machine learning and geospatial data analysis to drive intelligent investment decision-making in the Airbnb market. Our XGBoost model achieved an R^2 of 0.81, confirming high accuracy in price prediction. Listings scoring above 80 were strongly concentrated in downtown and coastal areas, driven by key features such as availability, review scores, and proximity to attractions.

The interactive dashboards revealed that top-performing listings correlated with faster host response times and higher review volumes, underscoring the strategic value of our scoring framework for real estate investors, property managers, and short-term rental platforms.

Our research question was addressed by successfully identifying investment-worthy listings through a composite scoring framework. The analysis confirmed our hypothesis: that engineered features—particularly review scores, availability, and proximity to downtown—could reliably predict Airbnb profitability.

The integration of Power BI enabled clear strategic translation of model insights, guiding hosts and investors toward high-performing opportunities.

Based on these insights, we recommend that investors prioritize listings with composite scores above 80, particularly those located in downtown and coastal neighborhoods where demand and ROI consistently trend higher.

The scoring framework—highlighting availability, review quality, and proximity—offers clear guidance for identifying profitable opportunities. Meanwhile, hosts should focus on improving review volume, response rates, and guest satisfaction metrics, which were shown to significantly influence profitability and investment appeal.

While the model proved highly effective within its defined scope, limitations such as simulated safety data, absence of guest review text, and the lack of seasonality modeling may influence investment interpretations. These constraints should be considered when applying the framework to broader markets or long-term projections.

Looking ahead, this scoring framework could be adapted to enable comparative analysis across multiple U.S. cities. Expanding the model to include real crime data, NLP-based sentiment extraction, and seasonal forecasting would further refine its accuracy. These additions would support investors in identifying high-performing listings across diverse urban markets and ensure regional applicability at scale.