AirBnB – Market Segmentation and Insights

MGMT 687- AI for Business Decisions

Aryan Saxena

Himanshu Sharma

Ayush Gupta

PURDUE UNIVERSITY

airbnb

Agenda

- Executive Summary
- Introduction to AirBnB
- Business Problem Overview
- Data Source Description
- Models Used and Results
- Insights and Implications
- Conclusion

PURDUE UNIVERSITY airbnb

Executive Summary

Objective

- Airbnb aims to enhance host network optimization by leveraging data-driven segmentation and predictive insights.
- The goal is to identify actionable strategies for improving host performance, revenue, and service quality while strengthening the Superhost program.

Modelling

- **Market Segmentation:** Hosts were grouped into six distinct clusters based on performance and characteristics using K-Means clustering.
- **Predictive Insights:** Statistical and machine learning models predicted revenue drivers, Superhost probabilities, and key factors influencing Superhost status.

Stakeholder Insights

- **High performers** (e.g., Clusters 2 & 4) excel in guest satisfaction and operational efficiency but **need retention strategies** to maintain leadership
- **Emerging potential hosts** (e.g., Cluster 3) show growth potential but **require support** to sustain their momentum
- **Underperformers** (e.g., Clusters 1, 5, & 6) face challenges with occupancy, pricing, and guest engagement but **present opportunities for targeted interventions**

PURDUE UNIVERSITY airbnb

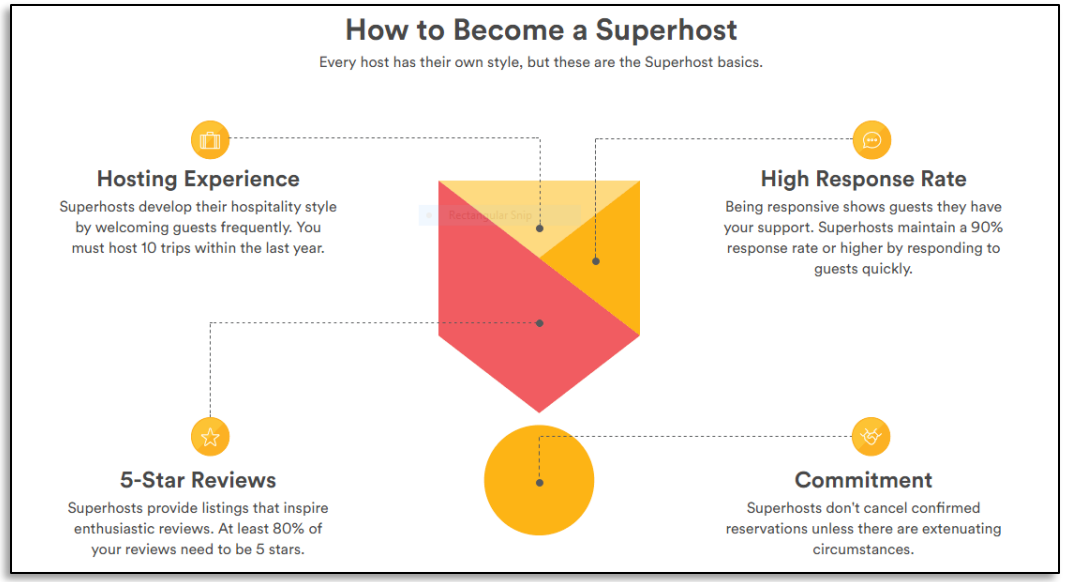
Introduction to AirBnB

- Airbnb is a global online marketplace that connects people looking for accommodations with hosts offering unique stays and experiences.
- **Founded in 2008**
- **Global Reach**: Airbnb operates in over **220 countries and regions** with millions of active listings.
- **Diverse Offerings**: From budget-friendly shared spaces to luxury villas, Airbnb caters to a wide range of travelers.
- **Community Focus**: Hosts and guests can interact directly, fostering a sense of community and trust through detailed profiles and reviews.
- **Beyond Stays**: In addition to accommodations, Airbnb offers **"Experiences"**, curated activities hosted by locals, enriching travel experiences.

PURDUE UNIVERSITY airbnb

What are Superhosts?

- **Definition**: Superhosts are Airbnb's top-performing hosts, recognized for exceptional service and reliability.
- **Evaluation**: Hosts are assessed quarterly, with Superhost status valid until the next evaluation.
- **Criteria**: Over 365 days, hosts must:
  - Host **10+ trips**.
  - Maintain **90% response rate**.
  - Have **0 cancellations**.
  - Earn **80% 5-star reviews**.
- **Benefits**: Superhosts gain higher visibility, attracting more bookings and ensuring a trusted experience for guests.
- **Significance**: The program drives service excellence, boosting Airbnb's reputation and customer loyalty.



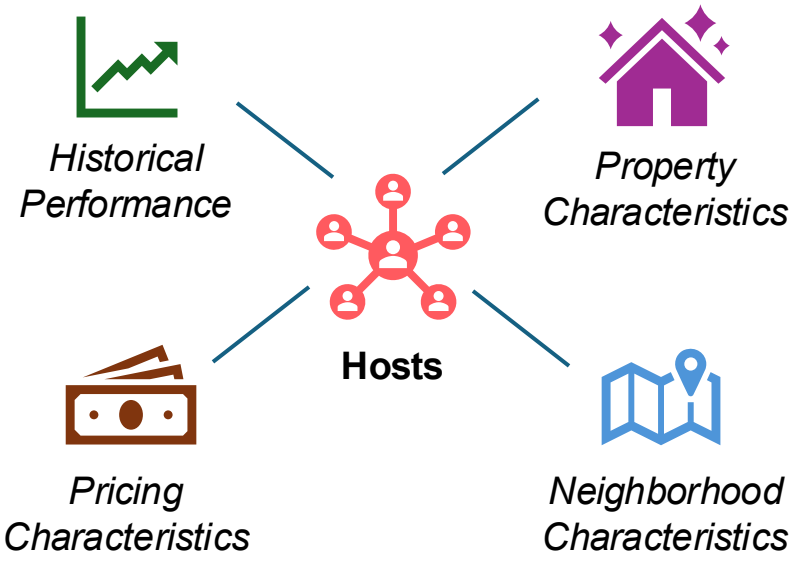
PURDUE UNIVERSITY airbnb

Business Problem Overview

We would like to help AirBnB create **Segmented Market Insights on their Host Network.** This can be accomplished as follows:

Step 1: Market Segmentation

Segment the hosts based on their performance and characteristics as shown below:



Step 2: Insight Generation

Once the hosts have been segmented, we will generate predictive insights on them:

1. Revenue Key Driver Prediction – What are the key drivers of revenue for each segment?

2. Superhost Probability Generation – In each segment, how many hosts have a high probability of being a Superhost in the next period?

3. Superhost Status Key Drivers: What are the key drivers of hosts attaining superhost status?

PURDUE UNIVERSITY airbnb

Dataset Overview and Selection

We have selected Chicago as the city to run our analyses. A few summary data points:

120,217 rows

111 columns

13,689 Unique Hosts

25,845 Unique Properties

- **Key features:** Broad and diverse dataset, capturing extensive information about Airbnb listings, hosts, bookings, and neighborhood demographics.

- **Why Choose Chicago?**
  - **Similar Scale:** Other datasets are comparable in size, ensuring statistical robustness.
  - **Greater Variability:** Chicago's dataset reflects more diverse host and property profiles, enabling insights applicable to a wider range of markets.
  - **Fewer Missing Values:** Chicago has fewer missing values in critical metrics like bookings and revenue, ensuring more reliable and accurate analysis.

Chicago offers a similar scale compared to other datasets but provides more variability and lesser missing data, making it a better representation of a generic population.

PURDUE UNIVERSITY airbnb

Key Data Challenges

Missing Data

Booking-related, Revenue and Occupancy metrics metrics, had up to 31.7% missing values.

Moderate missingness (10–20%) was also observed in variables like available_days .

Skewness

Features such as revenue and tract_booking_share exhibited high skewness and variability.

The variability highlights diverse host strategies and guest behaviors.

Outliers

Columns like rating_ave_pastYear had 511 outliers detected using Z-score.

Revenue and booking metrics had thousands of outliers identified using IQR.

Correlation

Key features like tract_booking_share and tract_revenue_share showed high correlations (r > 0.8).

Dimensionality reduction via PCA was essential to handle these correlations.

The dataset presented significant challenges, including missing data, high variability, outliers, and multicollinearity, necessitating targeted preprocessing

Steps to Address Data Challenges

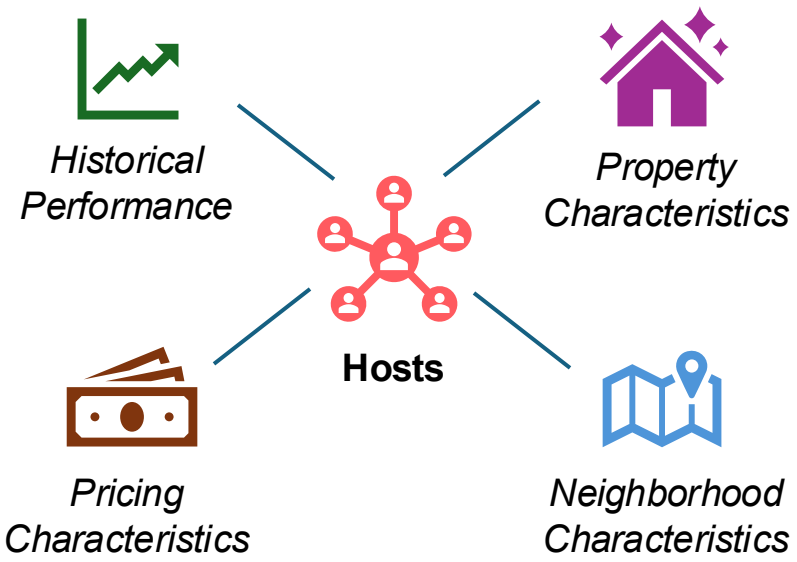
Missing Data	Skewness	Outliers	Correlation
Imputed missing values using the median for numerical features.	Applied normalization techniques (Yeo-Johnson transformation) to reduce skewness in key features.	Detected using Z-score and IQR, replaced with median values to reduce noise.	Dimensionality reduction using PCA to retain essential information while eliminating multicollinearity.

Preprocessing corrected ensures that the dataset is ready for robust analysis.

Market Segmentation – An Overview

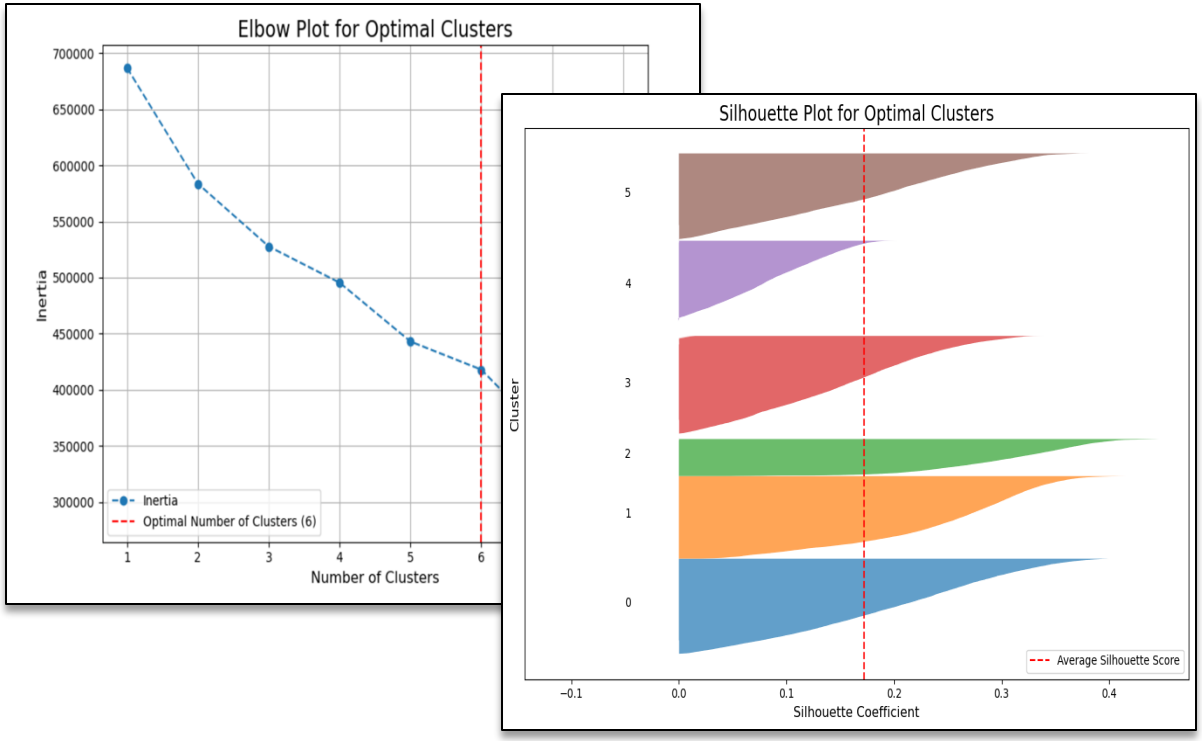
Segmentation Ideology

We've carried out the market segmentation based on the following characteristics:



Model Finalization

We used an elbow plot and silhouette plot on the transformed data to identify the K=6 as the ideal number of clusters for our analysis:



PURDUE UNIVERSITY airbnb

Segment 1 – Hosts of Underperforming Properties

Historical Performance

❖ Almost negligible share of hosts are superhosts, with no gains in status

Property Characteristics

❖ Below-average size and capacity with minimal representation of entire apartments

Pricing Characteristics

❖ Significantly below-average pricing and low competitiveness in the local market

Neighborhood Characteristics

❖ Moderate superhost density but low overall population

PURDUE UNIVERSITY airbnb

Segment 2 – Hosts of Premium Properties

Historical Performance

- Nearly all hosts are superhosts, though no recent gains were observed
- Zero cancellations highlight strong reliability

Property Characteristics

- Significantly above-average size and capacity. No private rooms or entire apartments

Pricing Characteristics

- High pricing with highly competitive rates within the local neighborhood

Neighborhood Characteristics

- High superhost density but a relatively low population in the local area

PURDUE UNIVERSITY airbnb

Segment 3 – Rising Superhosts

Historical Performance

- ❖ All of the hosts are superhosts, with all gaining status recently
- ❖ Zero cancellations reflect strong reliability

Property Characteristics

- ❖ Average size and capacity. Some representation of entire apartments

Pricing Characteristics

- ❖ Near-average historical pricing with highly competitive rates in the local market

Neighborhood Characteristics

- ❖ Moderate superhost density and slightly above-average total population

PURDUE UNIVERSITY airbnb

Segment 4 – Stable Superhosts

Historical Performance

- Nearly all of the hosts are superhosts, with no recent gains
- Zero cancellations reflect strong reliability

Property Characteristics

- Below-average size and capacity with minimal representation of entire apartments

Pricing Characteristics

- Slightly below-average pricing with moderate competitiveness in the local market

Neighborhood Characteristics

- High superhost density and slightly above-average total population

PURDUE UNIVERSITY airbnb

Segment 5 – Potential Superhosts

Historical Performance

- Almost none of the hosts are superhosts, with no gains in status
- Zero cancellations suggest consistently good performance

Property Characteristics

- Significantly above-average size and capacity. Significant representation of entire apartments

Pricing Characteristics

- High pricing with highly competitive rates in the local market

Neighborhood Characteristics

- Moderate superhost density and below-average total population

PURDUE UNIVERSITY airbnb

Segment 6 – Budget-Friendly Hosts

Historical Performance

- A small share of the hosts are superhosts, with no gains in status
- Zero cancellations indicate consistent operations

Property Characteristics

- Below-average size and capacity with minimal representation of entire apartments and private rooms

Pricing Characteristics

- Slightly above-average pricing with the most competitive rates in the local market

Neighborhood Characteristics

- Low superhost density but significantly above-average total population

PURDUE UNIVERSITY airbnb

Revenue Prediction

What are each segment's key drivers of revenue?

Cluster	Features	Interpretation
High Performers Clusters 2 & 4	• Guest Satisfaction & Experience • Operational Efficiency	• High Performers generate revenue through exceptional guest satisfaction • Efficient host practices, like quick response times, help these clusters maintain a strong revenue position.
Emerging Potential Cluster 3	• Location Quality • Guest feedback	• This group benefits from historical pricing decisions but must continue refining strategies. • Hosts in this group can leverage the location's appeal in their listings.
Underperforming Segments Clusters 1,5,6	• Bookings • Price History	• Increasing occupancy is a critical area for improvement. • There is a need for effective pricing strategies to compete better in the market.

PURDUE UNIVERSITY airbnb

Revenue Prediction

What do these key drivers have in common?

- **Consistent Bookings:** Across all clusters, maintaining high occupancy is a key predictor of revenue. The ability to attract and retain guests consistently is fundamental for revenue generation.
- **Effective Pricing Strategies:** Historical pricing trends play a significant role in predicting revenue. Hosts who optimize their nightly rates are better positioned to maximize earnings.
- **Guest Satisfaction and Engagement:** Such metrics highlight the importance of positive guest feedback and high service quality. Satisfied guests not only leave positive reviews but are more likely to return.
- **Location Advantage :** Properties in desirable or highly rated neighborhoods tend to perform better, indicating that location is a critical factor in attracting bookings.
- **Property and Amenities Quality:** Such features influence revenue, reflecting the importance of offering competitive, well-equipped properties.
- **Operational Efficiency:** Attributes relating to this show that quick and professional responses improve guest experiences, contributing to revenue.

Revenue is primarily driven by occupancy, pricing, guest satisfaction, and location. Hosts who excel in these areas, regardless of cluster, are better positioned to maximize earnings.

Superhost Probability Generation

In each segment, how many hosts have a high probability of being a Superhost in the next period?

Cluster	Average Probability
1 - Underperformers	**8.0%**
2 – Premium Hosts	**92.5%**
3 – Rising Superhosts	**75.9%**
4 – Stable Superhosts	**92.5%**
5 – Potential Superhosts	**9.5%**
6 – Budget Friendly Hosts	**12.6%**

Key Takeaways

- **High Performers (2 and 4):** Stable, successful Superhosts who **need recognition and retention strategies**.
- **Emerging Potential (3):** Recent achievers with significant growth potential who require **support to sustain their momentum.**
- **Underperforming Segments (1, 5, and 6):** Struggle with low Superhost probabilities due to operational and service challenges but have **opportunities for improvement** through **targeted interventions, especially for "Potential Superhosts"**

The Average Probability for each cluster is calculated using the predicted probabilities of becoming a Superhost for all hosts within that cluster

PURDUE UNIVERSITY airbnb

Superhost Status Key Drivers

What are the key drivers of hosts attaining superhost status?

Cluster	Features	Interpretation
High Performers Clusters 2 & 4	• Guest Satisfaction & Engagement • Market Competitiveness	Successful Superhosts with high guest satisfaction and strong market positioning. Focus on retention and profitability maximization.
Emerging Potential Cluster 3	• Nightly Rate • Superhost History • Neighborhood influence • Property Characteristics	Recent Superhost achievers with growth potential. Provide support to sustain success through pricing tools and operational guidance.
Underperforming Segments Clusters 1,5,6	• Price and market Positioning • Neighborhood influence	Struggle with low Superhost probabilities due to weak pricing or service. Target improvements with training, better tools, and mentorship.

PURDUE UNIVERSITY airbnb

Key Insights and Implications

High Performers

Consist of stable Superhosts with high guest satisfaction and competitive positioning.

Strategic Steps:

- Recognize and reward top-performing hosts to maintain loyalty.
- Offer advanced pricing tools to further optimize revenue.
- Promote their listings in high-demand markets to maximize profitability.

Emerging Potential

Hosts recently achieving Superhost status show significant growth potential with moderate guest satisfaction and pricing strategies.

Strategic Steps:

- Provide targeted support, such as mentorship and operational tools, to sustain momentum.
- Encourage investment in quality enhancements.
- Highlight their listings through promotions.

Underperforming Segments

These clusters struggle with low guest satisfaction, weak pricing strategies, and inconsistent service levels.

Strategic Steps:

Launch training programs to improve service quality and operational efficiency.

Provide dynamic pricing tools to align rates with market demand.

Incentivize participation in host improvement programs.

Conclusion

Market Segmentation and Insights:

- ❑ Through advanced data analysis, Airbnb's host network has been segmented into six distinct groups, providing a clear understanding of host performance dynamics.
- ❑ This segmentation serves as a foundation for identifying growth opportunities and addressing key challenges.

Strategic Takeaways:

- ❑ **High Performers (Clusters 2 & 4):** Focus on retention and profitability by leveraging their strengths in guest satisfaction and operational excellence.
- ❑ **Emerging Potential (Cluster 3):** Support and sustain the growth trajectory of these hosts with tailored tools and mentorship.
- ❑ **Underperformers (Clusters 1, 5, & 6):** Address gaps in service, pricing, and engagement through structured interventions.

Path Forward:

- ❑ Implement targeted strategies to enhance host performance across segments, with a focus on maximizing revenue and improving guest satisfaction.
- ❑ Strengthen the Superhost program by empowering hosts with data-driven insights, pricing tools, and operational resources.

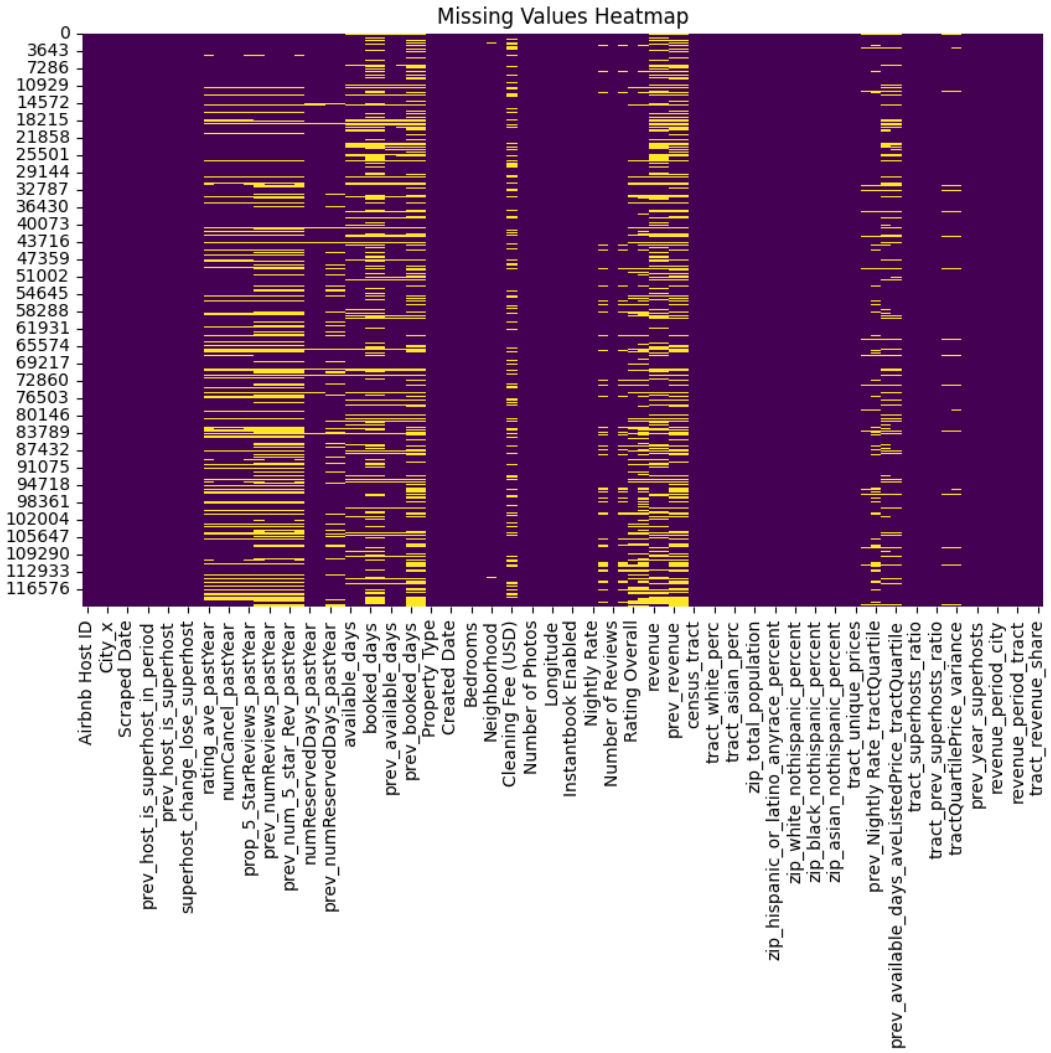
Broader Implications:

- ❑ By optimizing host performance and reinforcing the Superhost program, Airbnb can bolster its reputation, improve customer loyalty, and drive sustainable growth in the highly competitive marketplace.

Appendix

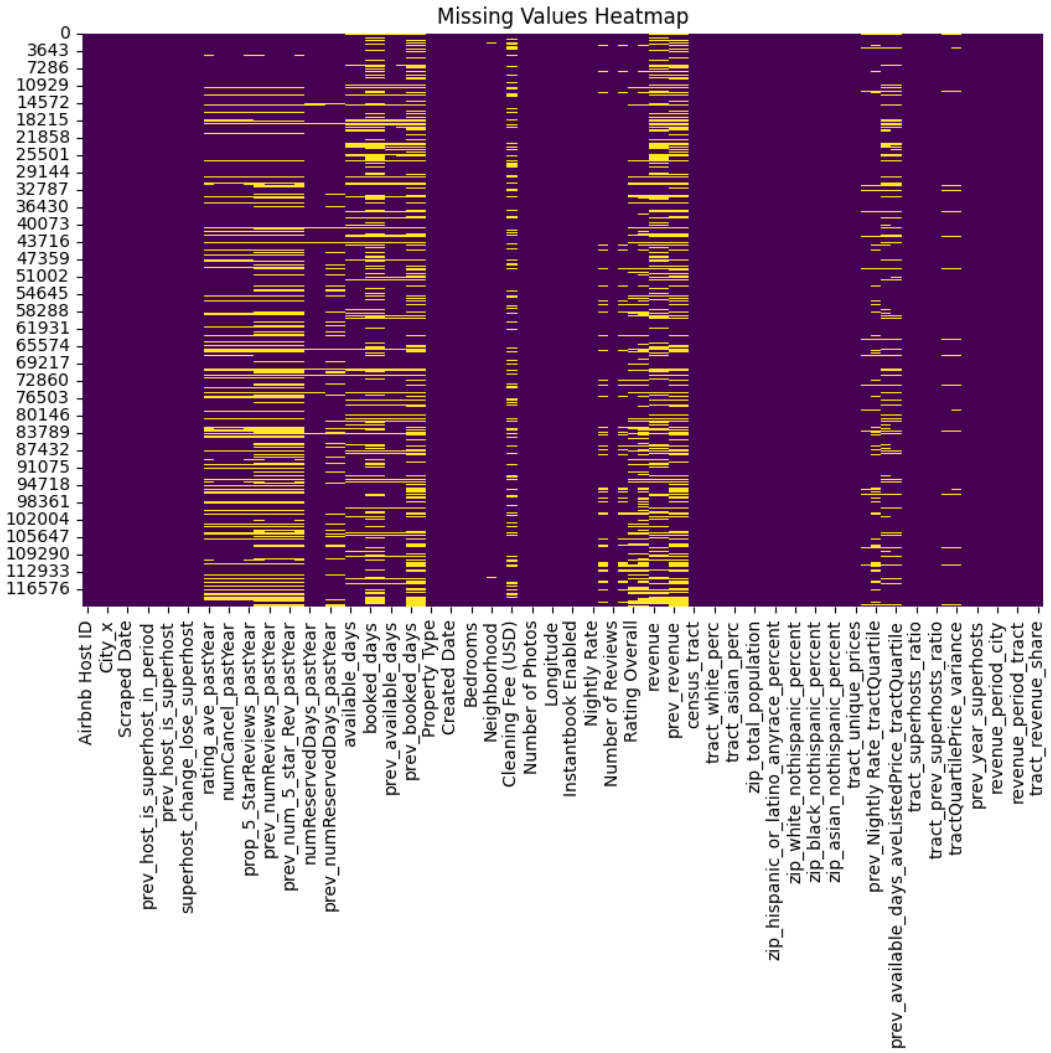
PURDUE UNIVERSITY airbnb

Data Preprocessing – Missing Values



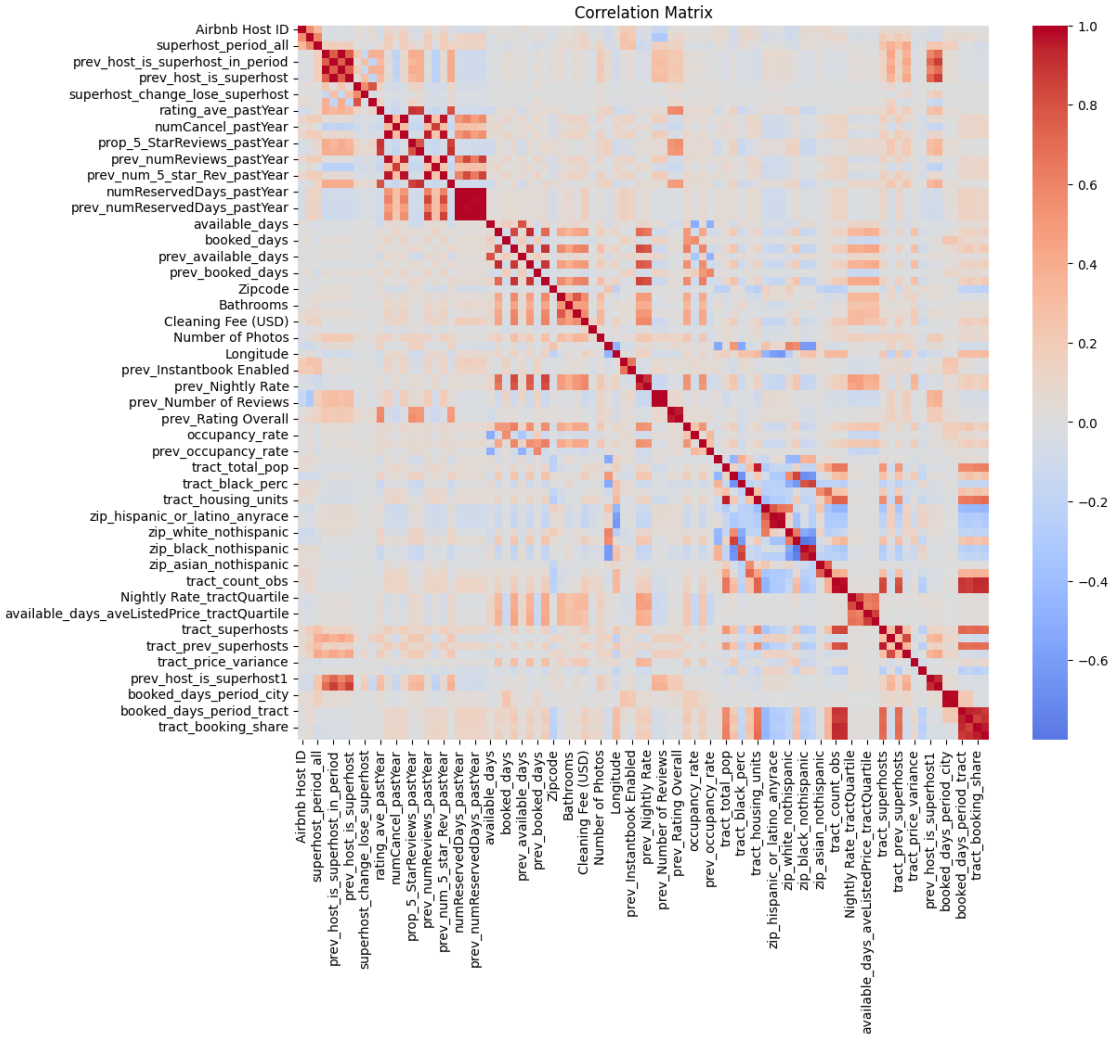
PURDUE UNIVERSITY airbnb

Data Preprocessing – Missing Values



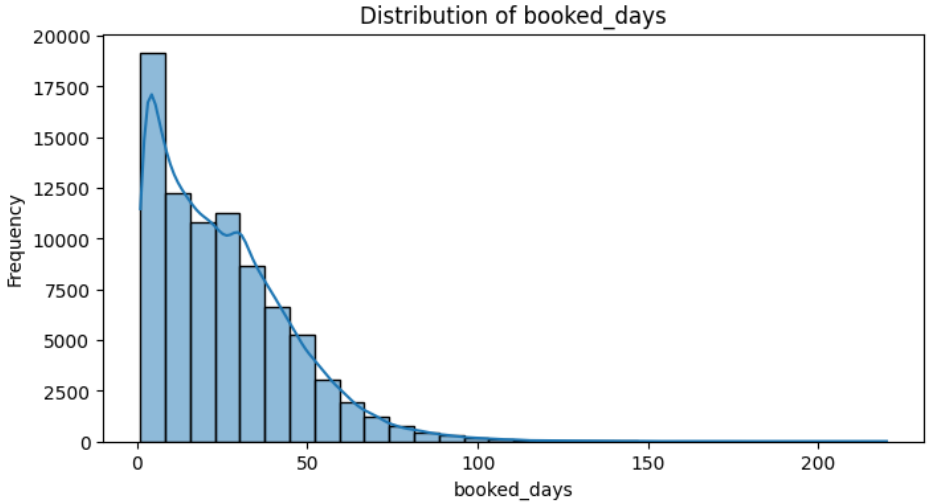
PURDUE UNIVERSITY airbnb

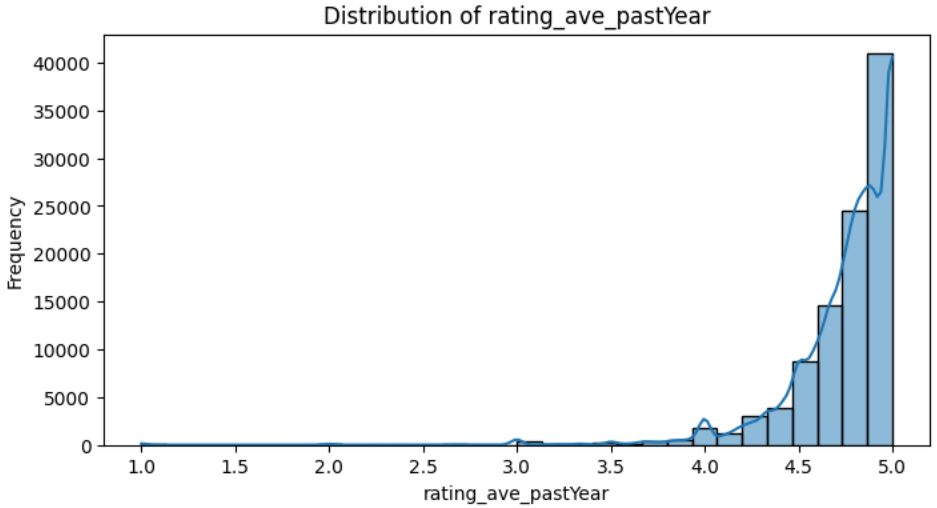
Data Preprocessing – Correlation

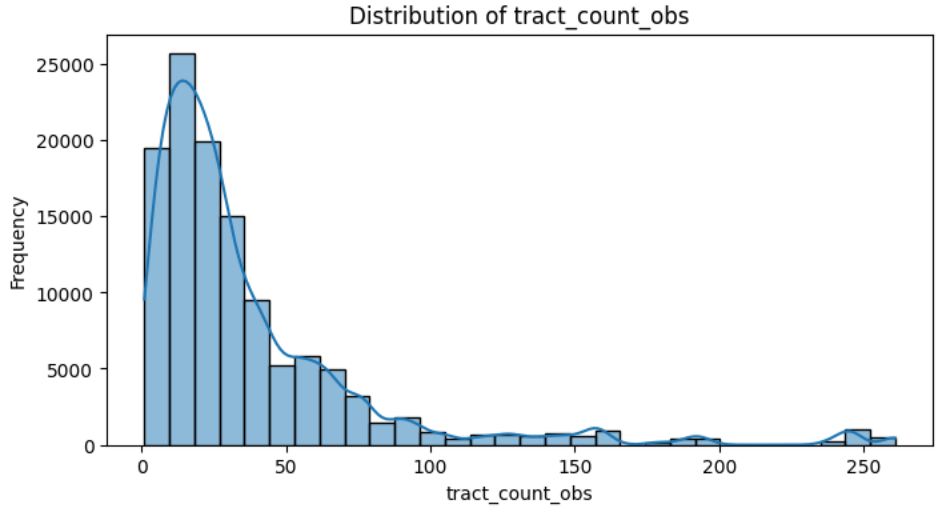


PURDUE UNIVERSITY airbnb

Data Preprocessing – Skewness

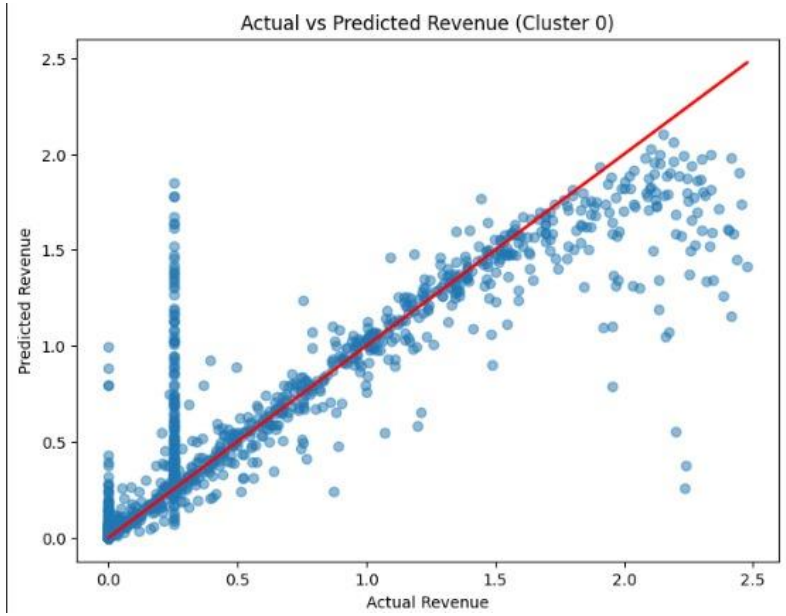


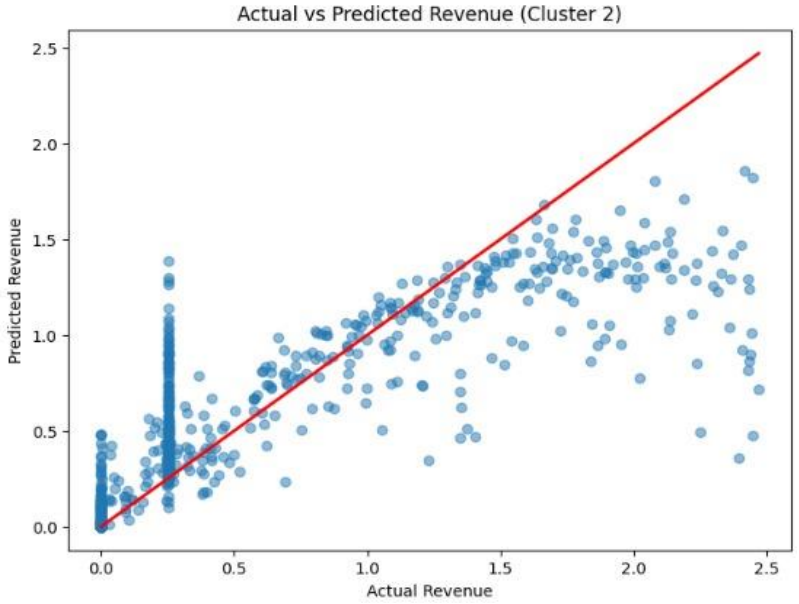


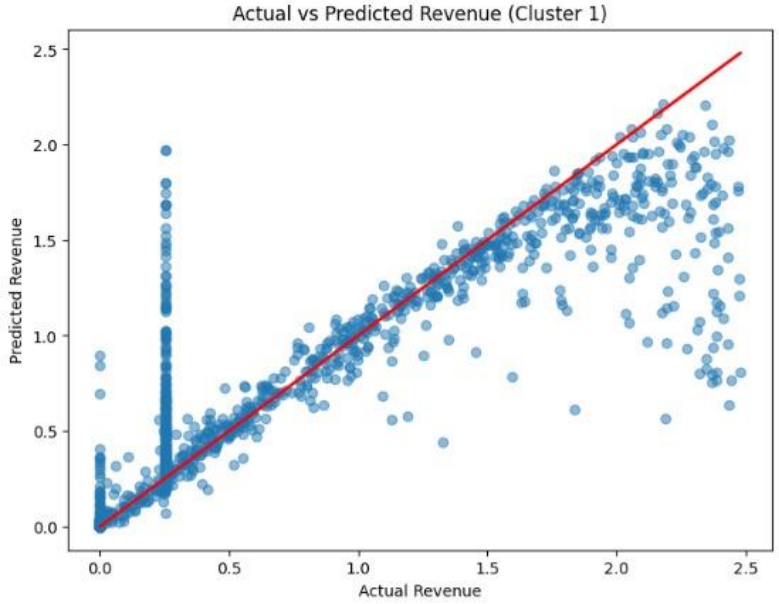


PURDUE UNIVERSITY airbnb

Revenue Prediction Model Results – Clusters 1 - 3

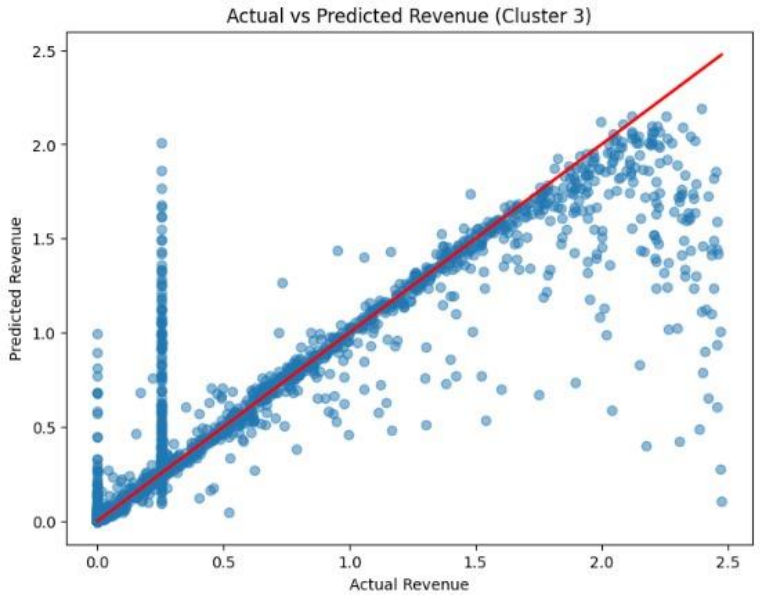


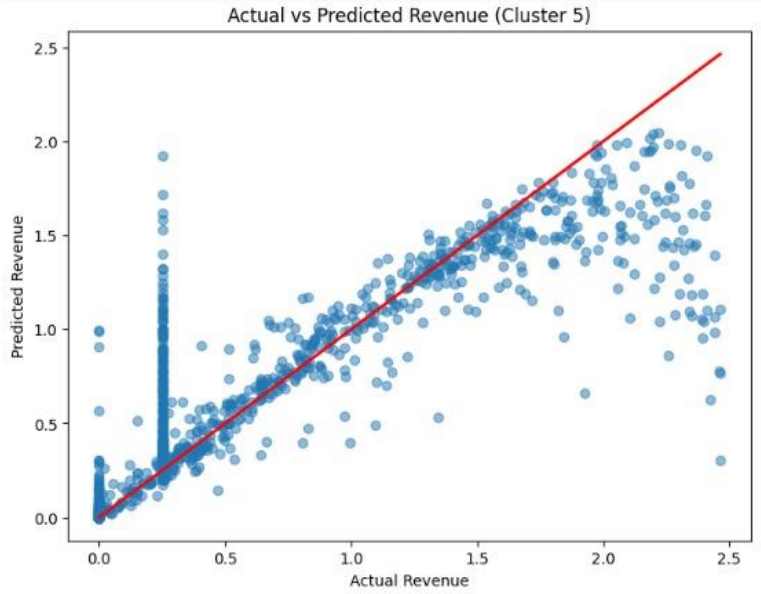


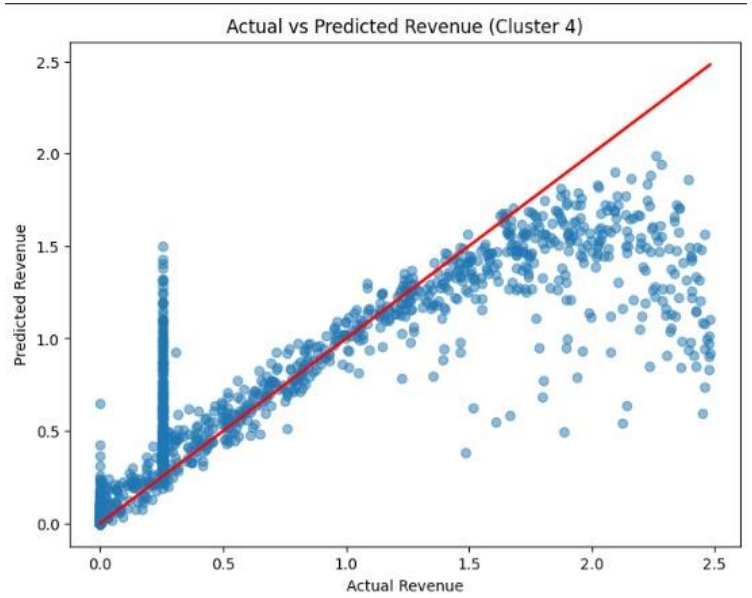


PURDUE UNIVERSITY airbnb

Revenue Prediction Model Results – Clusters 4 - 6

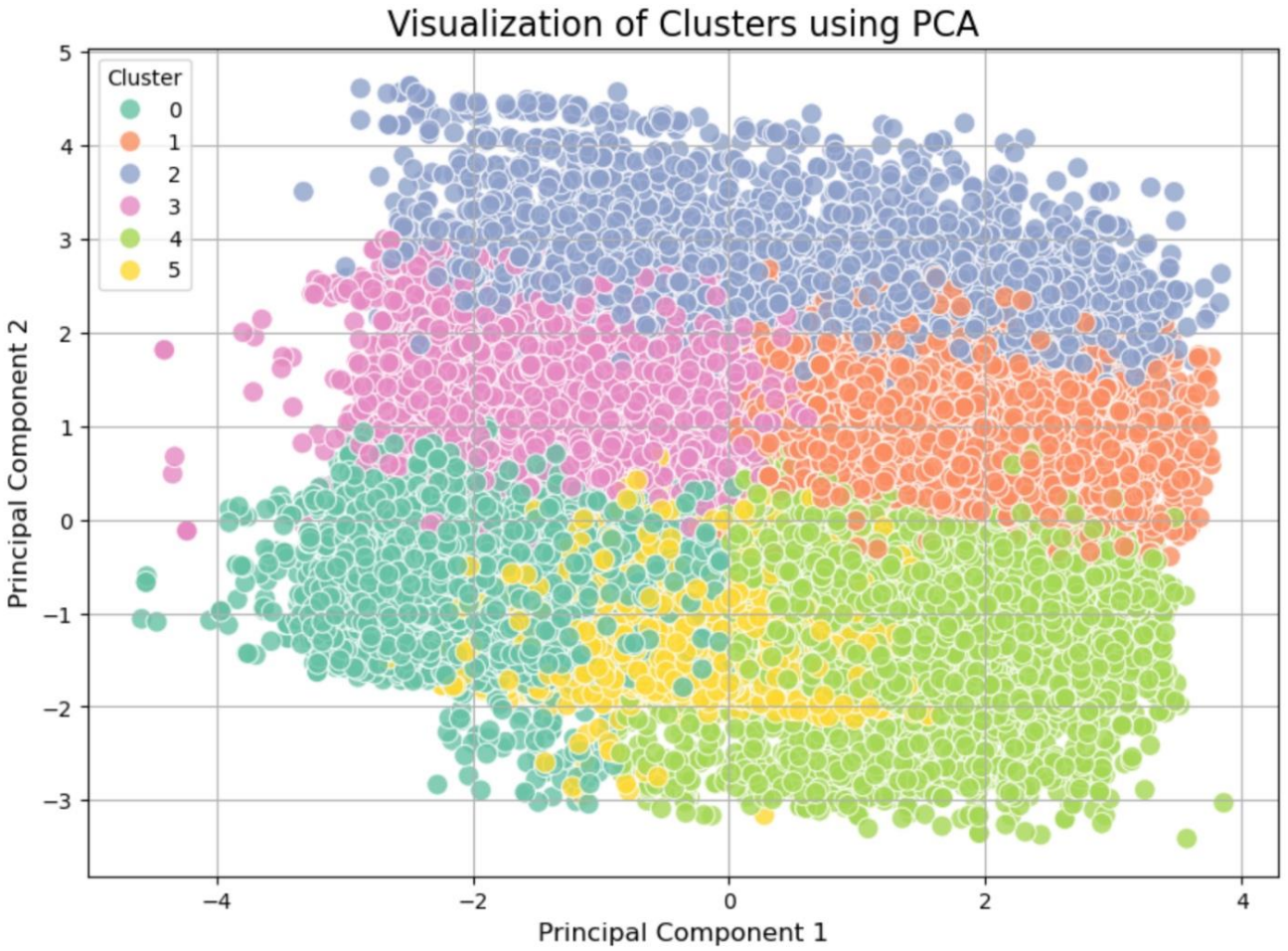






PURDUE UNIVERSITY airbnb

Clustering Results



PURDUE UNIVERSITY airbnb

References

Clustering Techniques:
Jain, A. K. (2010). *Data clustering: 50 years beyond K-means.* Pattern Recognition Letters, 31(8), 651-666.
Explores the foundational concepts of K-means clustering, its limitations, and extensions.
PCA and Dimensionality Reduction:
Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.
Covers the theoretical and practical aspects of PCA for reducing dimensionality and interpreting data.
Data Preprocessing and Feature Scaling:
Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.
Includes best practices for data cleaning, scaling, and feature engineering.
Power Transformations:
Box, G. E. P., & Cox, D. R. (1964). *An Analysis of Transformations.* Journal of the Royal Statistical Society: Series B, 26(2), 211-252.
Discusses methods like the Yeo-Johnson transformation for stabilizing variance and normalizing data distributions.
Machine Learning with Python:
Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
A practical guide to implementing machine learning techniques, including clustering, preprocessing, and regression.
Evaluation of Clustering Results:
Rousseeuw, P. J. (1987). *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 20, 53-65.
Provides methods to evaluate clustering performance and interpret results.
Superhost Prediction (Airbnb Context):
Zervas, G., Proserpio, D., & Byers, J. W. (2017). *The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry.* Journal of Marketing Research, 54(5), 687–705.
Contextual insights into Airbnb metrics and their impact on host and property success.
Data Aggregation Techniques:
Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media.
While focused on R, this book provides principles of data aggregation and grouping, which are transferable to Python.

PURDUE UNIVERSITY airbnb

<insert data comparison visualizations>

Chicago offers a similar scale compared to other datasets but provides more variability and lesser missing data, making it a better representation of a generic population.

<insert data preprocessing diagrams>

Market Segmentation

Models built

I. Super host
   A. Logistic regression
   B. Random Forest and XGBOOST
II. Market Segmentation
   A. K-means Clustering
III. Cancellation Prediction
   A. Random Forest
   B. XGBOOST
   C. LightGBM
   D. LSTM
IV. Revenue Prediction
   A. Gradient Boosting

Results

Business findings

References