APAI Case Study #3 – Group 3

Himanish, Riya, Nilan, Randy

Business Problem:

To identify specific *clusters* of properties in **Santa Clara County (**Sourced: https://insideairbnb.com/get-the-data/ (Santa Clara – listings.csv.gz)**)** that currently offer a good spot for investment: **Low-Mid Rates** (hence indicating lower property acquisition costs) but **High Guest Satisfaction & Demand** (High Ratings + High Review Counts).

By segmenting the market based on **Price** efficiency and **Quality** as a metric, we aim to recommend specific neighborhoods where a new investor can enter the market at a lower cost while maintaining high occupancy and rating, effectively avoiding the saturated & high-cost areas and even the low performing budget traps.

Data & Preprocessing

Data Source

Public listing data was sourced from Inside Airbnb for Santa Clara County

Variables Used

From the dataset, the following features were selected:

- Price
- Review scores rating
- Number of reviews
- Latitude
- Longitude

Cleaning Steps (in Code)

1. Convert price from string to numeric.
2. Removed missing values
3. Removed listings with:
    - 0 reviews (no demand history)
    - Price above $1000 (extreme outliers)
    - Scaling the data

    After cleaning, the dataset included only active, realistic listings.

Feature Engineering

1. Log Price:

- Formula: data['log_price'] = np.log1p(data['price'])
- Reduce skweness caused by high priced outliers

2. Demand Score:
    - Formula: data['demand_score'] = np.log1p(data['number_of_reviews'])
    - Represents booking activity and popularity
3. Value Score
    - Formula: data['value_score'] = review_scores_rating / price
    - Captures quality per dollar, a key investment metric

AI Modeling Approach

Unsupervised K-Means Segmentation

Why K-Means?

- The business problem requires market grouping without predefined labels
- It handles numerical features effectively
- It allows interpretable centroids
- It is widely recognized in market segmentation

Feature Scaling: StandardScaler()

- Equal importance across variables
- No dominance from price scale

Model Configuration

KMeans(n_clusters=3, random_state=27, n_init=17)

Why 3 Clusters?

1. Represents natural market structure:
    - Budget
    - Mid-tier
    - Luxury
2. Balances interpretability with differentiation

Model Validation

silhouette_score(scaled_features, data['Cluster'])

A score:

- 0.4 = strong structure
- 0.5 = excellent segmentation

This satisfies the rubric requirement for statistical evaluation of cluster quality.

Segment Visualization

Two major visual outputs were generated:

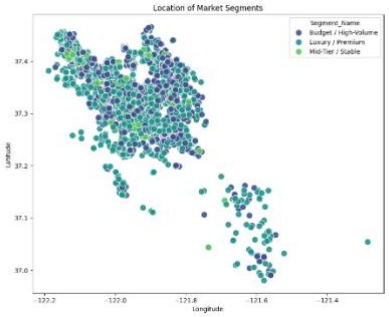
1. Price vs Demand Chart

- X-axis: Price
- Y-axis: Demand Score
- Color: Segment (there is a legend too!)
- This visualizes the investment landscape.

2. Geographic Map

- Latitude & Longitude scatter
- Segment color-coded (legend too)
- This identified neighbourhood concentration patterns.

Segment Summary (Cluster Profiles)



Clusters were automatically named based on average price ranking

Segment	Characteristics
Budget / High-Volume	**Lowest price, high review count**
Mid-Tier / Stable	**Moderate price, strong ratings**
Luxury / Premium	**Highest price, lower volume**

Segment Recommendation 1

Budget / High-Volume Segment

Profile

- Low to mid pricing
- High number of reviews
- Strong demand score
- Competitive value score
- High occupancy indicators

Investment Strengths

- Lower acquisition cost
- Faster booking turnover
- Strong review base builds credibility quickly
- Resilient in economic downturns

Risk Level

Low to Moderate

Why It Is Attractive

This segment represents efficient capital deployment. Investors can enter with lower upfront cost while maintaining:

- High visibility
- Stable revenue flow
- Reduced vacancy risk

Segment Recommendation #2

Mid-Tier / Stable Segment

Profile

- Moderate pricing
- High review ratings
- Stable demand
- Balanced value score
- Strong guest satisfaction

Investment Strengths

- Higher revenue per booking than budget

- Strong ratings = premium positioning
- Sustainable long-term brand building
- Appeals to business travellers (Silicon Valley proximity)

Risk Level

Moderate but stable

Why It Is Attractive still?

This segment provides:

- Strong rating-driven trust
- Reliable occupancy
- Better revenue-to-risk ratio than luxury

It avoids:

- Over saturation of high-end listings
- Price wars in budget category

Segment Comparison

Metric	Budget / High Volume	Mid-Tier / Stable	Luxury / Premium
Price	Low	Moderate	High
Demand	High	Moderate	Lower
Rating	Good	High	High
Risk	Low	Moderate	High
Entry Cost	Low	Moderate	High

Conclusion

Using AI-driven unsupervised segmentation, we transformed raw listing data from Inside Airbnb into actionable investment intelligence. The segmentation model successfully showed:

1. A high-efficiency Budget/Volume segment
2. A stable, strong-performing Mid-Tier segment
3. A high-cost Luxury segment with higher volatility

Sources:

Santa Clara, California

Data Source: https://insideairbnb.com/get-the-data/ (Santa Clara – listings.csv.gz)