Task 1.3: Algebraic Feature Engineering

San Francisco & San Diego Airbnb Dataset

1. Overview

Dataset: San Francisco & San Diego Airbnb listings
Original Shape: 19,912 rows × 71 columns
Final Shape: 19,912 rows × 81 columns
New Features Added: 10 algebraic features
Data Quality: ✓No missing values, no infinite values

2. Algebraic Features Created

1. space_efficiency
Formula: beds / bedrooms
Description: Measures how efficiently space is utilized. Higher values indicate more beds per bedroom.

2. value_density
Formula: number_of_reviews / price
Description: Indicates popularity relative to price. Higher values suggest better value for money.

3. review_to_capacity_ratio
Formula: number_of_reviews / accommodates
Description: Shows review volume per guest capacity. Indicates listing popularity per person.

4. price_per_bedroom
Formula: price / bedrooms
Description: Cost per bedroom. Useful for comparing listings with different bedroom counts.

5. price_per_bathroom
Formula: price / bathrooms_numeric
Description: Cost per bathroom. Helps identify premium vs budget listings.

6. occupancy_rate
Formula: (365 - availability_365) / 365

Description: Estimated occupancy rate. Higher values indicate more frequently booked listings.

7. review_momentum

Formula: number_of_reviews_ltm / number_of_reviews

Description: Recent review activity. Values close to 1 indicate growing popularity.

8. host_portfolio_intensity

Formula: host_total_listings_count / accommodates

Description: Host business scale relative to property size. Identifies professional hosts.

9. booking_flexibility_score

Formula: 1 / (minimum_nights + 1)

Description: Booking flexibility. Higher values mean shorter minimum stays.

10. space_per_person

Formula: bedrooms / accommodates

Description: Privacy measure. Higher values indicate more private space per guest.

3. Statistical Summary

Feature	Mean	Median	Std Dev	Min	Max
space_efficiency	1.34	1.00	0.60	0.00	16.00
value_density	0.48	0.09	1.09	0.00	19.61
review_to_capacity_ratio	22.37	4.17	50.69	0.00	902.00
price_per_bedroom	128.28	113.00	68.25	6.50	680.00
price_per_bathroom	148.44	136.75	78.68	3.90	680.00
occupancy_rate	0.44	0.35	0.35	0.00	1.00
review_momentum	0.30	0.13	0.36	0.00	1.00
host_portfolio_intensity	74.94	2.00	374.59	0.06	4775.50
booking_flexibility_score	0.24	0.25	0.18	0.00	0.50
space_per_person	0.44	0.50	0.28	0.00	10.00

4. Key Insights

- Space Efficiency: Most listings have 1-2 beds per bedroom (median = 1.0)
- Value Density: Highly skewed distribution, indicating few listings have exceptional value
- Occupancy Rate: Average 44% occupancy, with high variance across listings
- Review Momentum: 30% of reviews are recent (last 12 months), showing active market
- Booking Flexibility: Most listings require 3-4 night minimum stays
- Host Portfolio: Median host has 2 listings, but some manage hundreds (professional hosts)
- Price per Bedroom: Average $128/bedroom, ranging from budget ($6.50) to luxury ($680)
- Space per Person: Average 0.44 bedrooms per guest, indicating shared accommodations

5. Output Files

- listings_with_algebraic_features.csv - Main dataset with all 81 features
- algebraic_features_distributions.png - Distribution plots for all 10 features
- algebraic_features_correlations.png - Correlation analysis with target variable
- algebraic_features_by_value_category.png - Feature comparison across value categories
- algebraic_features_statistics.csv - Detailed statistical summary