

Task 1.4: Categorical Feature Encoding

San Francisco & San Diego Airbnb Dataset

1. Overview

Dataset: San Francisco & San Diego Airbnb listings

Original Shape: 19,912 rows × 84 columns

Final Shape: 19,912 rows × 94 columns

New Features Added: 10 encoded features

Categorical Variables Encoded: room_type, property_type, neighbourhood_cleansed, value_category

2. Encoding Strategy

2.1 Room Type - One-Hot Encoding

Method: One-Hot Encoding (Binary)

Reason: Low cardinality (4 categories), no ordinal relationship

Columns Created: 4 (room_type_Entire home/apt, room_type_Hotel room, room_type_Private room, room_type_Shared room)

Distribution:

- Entire home/apt: 14,866 (74.7%)
- Private room: 4,557 (22.9%)
- Hotel room: 413 (2.1%)
- Shared room: 76 (0.4%)

2.2 Property Type - Label + Frequency Encoding

Method: Label Encoding + Frequency Encoding

Reason: High cardinality (66 categories), one-hot would create too many columns

Columns Created: 2 (property_type_label, property_type_frequency)

Label Range: 0 to 65

Frequency Range: 0.005% to 24.95%

Top 5 Property Types:

1. Entire home: 4,968 (24.95%)
2. Entire rental unit: 4,725 (23.73%)
3. Entire condo: 2,402 (12.06%)
4. Private room in home: 1,742 (8.75%)
5. Room in hotel: 1,410 (7.08%)

2.3 Neighbourhood - Target + Frequency Encoding

Method: Target Encoding + Frequency Encoding + Label Encoding

Reason: Very high cardinality (138 categories), geographic importance

Columns Created: 3 (neighbourhood_target_encoded, neighbourhood_frequency, neighbourhood_label)

Target Encoding: Mean of value_category per neighbourhood (0.0 to 2.0)

Frequency Range: 0.005% to 8.51%

Top 5 Neighbourhoods:

1. Mission Bay: 1,694 (8.51%)
2. Downtown/Civic Center: 1,204 (6.05%)
3. Pacific Beach: 991 (4.98%)
4. La Jolla: 816 (4.10%)
5. South of Market: 591 (2.97%)

2.4 Value Category - Label Encoding

Method: Label Encoding (Ordinal)

Reason: Target variable with ordinal relationship

Columns Created: 1 (value_encoded)

Mapping:

- Excellent_Value → 2 (6,568 listings)
- Fair_Value → 1 (6,773 listings)
- Poor_Value → 0 (6,571 listings)

3. Encoding Summary

Variable	Cardinality	Encoding Method	Columns Created	Range/Values
room_type	4	One-Hot	4	0 or 1
property_type	66	Label + Frequency	2	0-65, 0.005%-24.95%
neighbourhood	138	Target + Frequency + Label	3	0.0-2.0, 0.005%-8.51%, 0-137
value_category	3	Label (Ordinal)	1	0, 1, 2

4. Key Insights

- One-Hot Encoding: Best for low cardinality variables (room_type with 4 categories)
- Label Encoding: Memory-efficient for high cardinality (property_type with 66 categories)
- Frequency Encoding: Captures popularity information (most common property type: 24.95%)
- Target Encoding: Leverages predictive power (neighbourhood mean values range 0.0-2.0)
- Balanced Distribution: Value categories are well-balanced (33-34% each)
- Geographic Diversity: 138 unique neighbourhoods across SF and SD
- Property Diversity: 66 unique property types, dominated by "Entire home" (24.95%)
- Room Type Dominance: 74.7% are "Entire home/apt", only 0.4% are shared rooms

5. Data Quality

- ✓**No Missing Values:** All encoded columns are complete
- ✓**No Infinite Values:** All encodings are within valid ranges
- ✓**Consistent Encoding:** All 19,912 rows successfully encoded
- ✓**Reference Files:** Mapping files created for reproducibility

6. Output Files

- listings_with_categorical_encoding.csv - Main dataset with all 94 features
- property_type_encoding_map.csv - Property type encoding reference (66 types)
- neighbourhood_encoding_map.csv - Neighbourhood encoding reference (138 neighbourhoods)
- categorical_encoding_analysis.png - Distribution and encoding visualizations

- `encoding_methods_comparison.png` - Encoding methods comparison and analysis

7. Next Steps

- Dataset ready for model training with 94 features
- All categorical variables properly encoded
- Can proceed with feature selection and model building
- Encoding mappings saved for future data preprocessing