Data Preprocessing Pipelines

Nandan

September 25, 2025

Reference: `02_Stage1_staging.ipynb`

1 Systematic Missing Value Treatment

1.1 Objective

Address missing competitor pricing data using forward-fill imputation followed by group median fallback, following Section 2.3 methodology for handling dirty hotel pricing data.

1.2 Implementation

Applied two-stage missing value treatment:

$$\text{Stage 1: Forward Fill} : P_{i,t} = P_{i,t-1} \text{ if } P_{i,t} = \text{NA} \tag{1}$$

$$\text{Stage 2: Group Median} : P_{i,t} = \text{median}(P_i) \text{ if still NA} \tag{2}$$

1.3 Results

- Missing competitor prices: 30 observations (1.6% of total)
- Non-random pattern: Aqua Pacific (16 missing), Courtyard Marriott (11 missing)
- Treatment successful: 0 remaining missing values

2 Data-Driven Price Validation and Conservative Outlier Treatment

2.1 Methodology

Implemented data-driven price bounds using 99th percentile thresholds rather than arbitrary limits:

$$\text{Upper Bound} = 1.5 \times \max(\text{P99}_{focal}, \text{P99}_{competitors}) \tag{3}$$

$$\text{Calculated Threshold} = \$1498 \text{ (vs arbitrary } \$5000) \tag{4}$$

Applied conservative IQR-based outlier detection:

$$\text{Outlier Threshold} = Q_{0.75} \pm 3.0 \times (Q_{0.75} - Q_{0.25}) \tag{5}$$

2.2 Critical Signal Preservation

$$\text{Outliers Removed} : 169 \text{ focal} + 100 \text{ competitor observations} \tag{6}$$

$$\text{Competitive Correlation} : 0.440 \text{ (strengthened from initial)} \tag{7}$$

Conservative 3.0 multiplier prevented signal destruction that would occur with aggressive 1.5x multiplier.

3 Base Rate Extraction and Temporal Alignment

3.1 Critical Methodological Change

Following Professor Dan's feedback, implemented base rate focus since focal hotel has room-type pricing while competitors provide only base rates:

$$\text{Base Rate}_{focal,t} = \min_{r \in \text{room types}} P_{focal,r,t} \tag{8}$$

3.2 Base Rate Characteristics

$$\text{Base Rate Observations} : 365 \text{ daily observations} \tag{9}$$

$$\text{Base Rate Range} : \$209 - \$379 \tag{10}$$

$$\text{Temporal Overlap} : 364 \text{ days (Sep 2025 - Sep 2026)} \tag{11}$$

3.3 Alignment Results

- Overlap shapes: Focal (364 observations), Competitors (1720 observations)
- Daily density: Focal 1.0, Competitors 4.7 observations per day
- Perfect base-rate-to-base-rate competitive comparison achieved

4 Cross-Hotel Normalization Implementation

4.1 Location-Scale Robust Normalization

Following Section 3.1, applied robust normalization using median and median absolute deviation:

$$\text{Focal Base Rate Parameters} : \text{Median} = \$279.00, \text{MAD} = \$20.00 \tag{12}$$

$$\text{Competitor Parameters} : \text{Median} = \$298.00, \text{MAD} = \$36.57 \tag{13}$$

$$\text{Normalization Transform} : P_{i,t}^{(norm)} = \frac{P_{i,t} - M_i}{\text{MAD}_i} \times \text{MAD}_{ref} + M_{ref} \tag{14}$$

4.2 Normalization Impact

Base rate normalization creates proper scale alignment for competitive analysis while preserving relative price variation within each hotel's distribution.

5 Feature Engineering for Instrumental Variables

5.1 Temporal Instruments

Created cyclical temporal features for Stage 1 modeling:

$$\text{Day Cycle} : \sin(2\pi \cdot \text{day\_of\_week}/7), \cos(2\pi \cdot \text{day\_of\_week}/7) \tag{15}$$

$$\text{Monthly Cycle} : \sin(2\pi \cdot \text{month}/12), \cos(2\pi \cdot \text{month}/12) \tag{16}$$

$$\text{Weekend Indicator} : \mathbf{1}[\text{day\_of\_week} \geq 5] \tag{17}$$

$$\text{Week of Year} : \text{ISO calendar week number} \tag{18}$$

5.2 Price Dynamics Features

Constructed lagged price features for dynamic analysis:

$$\text{Base Rate Lags} : P_{focal,t-k} \text{ for } k \in \{1, 2, 3\} \tag{19}$$

$$\text{Price Changes} : \Delta P_{focal,t} = P_{focal,t} - P_{focal,t-1} \tag{20}$$

$$\text{Competitor Lags} : P_{comp_i,t-k} \text{ for each competitor } i \tag{21}$$

6 Data Structure Preparation for 2SRI

6.1 Base Rate Focus

Created simplified structure focusing on base rate competition:

- Focal base rate daily series: 364 observations
- Competitor price matrix: 364 dates × 5 hotels
- Eliminated room-type complexity for clean competitive modeling

6.2 Stage 1 Preparation

Structured data for flexible Stage 1 modeling:

$$\text{Focal Daily Structure} : \text{Single base rate per day} \tag{22}$$

$$\text{Competitor Matrix Structure} : \text{Hotel-date pivot with normalized prices} \tag{23}$$

$$\text{Temporal Features} : \text{Comprehensive instrument set} \tag{24}$$

7 Quality Validation Post-Processing

7.1 Competitive Relationship Validation

Final validation of base rate competitive relationships:

$$\text{Base Rate Correlation} = 0.440 \tag{25}$$

$$\text{Sample Sizes} : 364 \text{ focal}, 1720 \text{ competitor observations} \tag{26}$$

$$\text{Warning Threshold} : \text{No correlation weakening detected} \tag{27}$$

7.2 2SRI Readiness Assessment

Base rate approach successfully creates:

- Meaningful competitive correlation (0.440 ¿ 0.2 threshold)
- Clean temporal alignment without room-type complexity
- Proper base-rate-to-base-rate competitive comparison
- Foundation for Stage 1 flexible modeling implementation

8 Export Pipeline for Stage 1 Implementation

8.1 Output Structure

Created analysis-ready datasets:

- `focal_hotel_clean.csv`: Base rate time series with features
- `competitors_clean.csv`: Normalized competitor pricing data
- `focal_daily_aggregated.csv`: Daily focal base rates
- `competitor_price_matrix.csv`: Pivoted competitor matrix