Section 3: Demographics Inference Methodology

Problem Statement
The objective is to infer user demographics (age group and gender) using only transactional signals, since explicit demographic details are unavailable. This framework outlines assumptions, scoring methodology, and validation plans.

Assumptions
1. Product categories and brands exhibit strong demographic signals.
2. Spending patterns and transaction frequency correlate with purchasing power and age.
3. Time-of-purchase provides behavioral cues (e.g., late-night purchases skew younger).

Step 1: Feature Definition
- Category Weights: Assign demographic likelihoods (e.g., Kids products → 25–40, Electronics high-value → male <40, Beauty/Fashion → female).
- Spending Power: High spend → 40+, mid spend → 25–40, low spend → <25.
- Transaction Frequency: Frequent buyers likely younger/tech-savvy.
- Time-of-Day: Late-night → <25, daytime → 25–40 or 40+.

Step 2: Scoring Model
- Rule-based weighted scoring: Each user receives probability scores across buckets (<25, 25–40, 40+ for age; male/female for gender).
- Example: If 60%+ spend is in Beauty/Apparel/Footwear, user is weighted more toward female. If most spend is in Electronics and Audio, skew toward male.
- Behavioral indicators (spending frequency, transaction time) adjust these probabilities.

Step 3: Synergy Across Features
The framework combines signals from categories, brands, spending, and timing. The synergy between multiple weak signals helps increase confidence even in the absence of explicit demographic labels.

Step 4: Validation Plan
Once true demographics are available:
- Evaluate precision, recall, and F1-score for each demographic bucket.
- Construct confusion matrices to understand misclassifications.
- Benchmark against random assignment to measure lift.

Operationalization
1. Use the rule-based system for immediate deployment.
2. As labeled data accumulates, refine into a supervised ML model (logistic regression, XGBoost).
3. Implement threshold-based classification (accept predictions only if probability ≥ 70%).

Ethical Considerations
- Predictions should never be used for discriminatory targeting.
- Outputs must be communicated as probabilistic, not deterministic.

Conclusion
This methodology leverages transactional data to construct a demographic inference framework. By combining category signals, spending power, frequency, and behavioral cues, we can generate probabilistic demographic predictions. Future validation and ML refinement will strengthen accuracy while maintaining transparency and fairness.