## Case Study: Smart Retail Analytics

**Scenario.** You are part of the data science team of a large retail chain. The company wants to **predict customer churn** and **optimize marketing campaigns**. You have access to several datasets:

1. **Online Sales Logs**: Contains timestamps, items purchased, quantities, and prices. Some entries have missing timestamps.
2. **Customer Profiles**: Includes age, gender, email, loyalty status, and signup date. About 20% of emails are missing, and some loyalty status labels are inconsistent.
3. **In-store Purchases**: Weekly transactions collected via point-of-sale systems. Some store IDs are duplicated or missing.
4. **Customer Surveys**: Voluntary feedback on satisfaction, collected quarterly. Many surveys are incomplete, with some customers responding multiple times.

Your task is to **prepare a data handling plan** that will enable accurate modeling and insightful analysis.

### Part 1: Understanding and Evaluating Data

**Questions:**

1. Which datasets would you prioritize for your analysis, and why?
2. How would you assess the reliability and completeness of each dataset?
3. Identify potential biases that could arise from using these datasets.

### Part 2: Handling Missing Data

**Questions:**

1. For each dataset, identify the most critical missing data and its potential impact.
2. Propose strategies to handle missing values (e.g., imputation, removal) and justify your choices.
3. Discuss the trade-offs between data completeness and data quality in this context.

### Part 3: Cleaning and Correcting Data

**Questions:**

1. How would you detect and handle inconsistent entries (e.g., loyalty status, duplicate store IDs)?
2. How would you handle outliers in purchase amounts or transaction counts?
3. How would you document your cleaning decisions to maintain transparency and reproducibility?

### Part 4: Data Integration

**Questions:**

1. How would you merge the online and in-store purchase datasets?
2. What issues might arise from mismatched or missing customer IDs, and how would you address them?
3. After merging, how would you validate that the integrated dataset accurately represents customers' behavior?

# Retail Analytics

## Part 1: Understanding and Evaluating Data

1. **Prioritization:**
   - **Online Sales Logs**: High priority — contains behavioral data critical for predicting churn.
   - **Customer Profiles**: Essential — demographic info improves predictive accuracy.
   - **In-store Purchases**: Medium priority — useful if modeling total customer behavior.
   - **Customer Surveys**: Low priority — valuable for qualitative insights but sparse and biased.
2. **Reliability Assessment:**
   - Check for missing values, duplicates, inconsistent formats.
   - Evaluate the source: internal systems (more reliable) vs. voluntary surveys (less reliable).
3. **Potential Biases:**
   - Survey data may be biased toward highly engaged customers.
   - Missing emails or IDs could disproportionately affect certain demographics.

## Part 2: Handling Missing Data

1. **Critical Missing Data:**
   - Missing timestamps in online sales → affects chronological analysis.
   - Missing loyalty status → affects segmentation.
   - Missing emails → affects merging datasets.
2. **Strategies:**
   - **Impute missing timestamps** using order sequence or remove if impossible.
   - **Impute loyalty status** using mode within demographic group or remove rows if necessary.
   - **Drop customers with missing IDs/emails** only if unavoidable for integration.
3. **Trade-offs:**
   - Imputation preserves sample size but may introduce errors.
   - Removing rows ensures data quality but reduces dataset size and may introduce bias.

## Part 3: Cleaning and Correcting Data

1. **Inconsistencies:**
   - Standardize loyalty status labels (e.g., “Gold,” “gold,” “GOLD” → “Gold”).
   - Remove duplicate store IDs or verify via external reference.
2. **Outliers:**
   - Detect extreme purchase amounts using statistical methods (IQR, Z-score).
   - Investigate outliers: errors vs. genuine large purchases. Remove or cap extreme errors.
3. **Documentation:**
   - Keep a **data cleaning log** with all decisions, rationale, and transformations applied.

## Part 4: Data Integration

1. **Merging Datasets:**
    - Use **customer ID** as primary key for joining online and in-store purchases.
2. **Handling Mismatched/ Missing IDs:**
    - Attempt fuzzy matching on names/emails if IDs are missing.
    - Flag and exclude entries that cannot be confidently matched.
3. **Validation:**
    - Compare aggregated totals (e.g., total purchases per customer) before and after merge.
    - Check for unexpected duplication or missing records.