Data Selection & Exploration

1. Dataset Description

1. **Data Sources & Structure**
    - **Sources:** Amazon Reviews datasets for **Books**, **Watches**, and **Electronics**.
    - **File Format:** Originally in TSV (tab-separated values) format; merged into a single dataset.
    - **Schema Overview (common columns):**
        - marketplace (VARCHAR)
        - review_id (BIGINT or VARCHAR)
        - customer_id (BIGINT) – *If available for cross-category analysis*
        - product_id (VARCHAR)
        - star_rating (FLOAT or INT)
        - helpful_votes (INT)
        - total_votes (INT)
        - verified_purchase (VARCHAR/BOOLEAN)
        - review_headline (VARCHAR)
        - review_body (TEXT)
        - review_date (DATE)
        - category (VARCHAR) – *A custom field labeling each record as "Books," "Watches," or "Electronics."*
2. **Dataset Size & Volume**
    - The merged dataset exceeds **100,000 total records** across all three categories.
    - Each category contributes thousands (or tens of thousands) of reviews, ensuring substantial coverage for analysis.
3. **Data Types & Missing Values**

- **Data Types:**
  - Numeric columns (e.g., star_rating, helpful_votes, total_votes).
  - Text columns (e.g., review_body, review_headline).
  - Date column (review_date).
  - Categorical columns (category, verified_purchase).
- **Missing Values:**
  - Some reviews may lack certain fields (e.g., missing review_headline or review_body).
  - Inconsistent or null entries in helpful_votes or total_votes if reviewers did not receive or cast votes.
- **Initial Handling:**
  - Identified these missing values during a preliminary scan in **SQL Server Management Studio (SSMS)**.
  - Will address these in the next phase (ETL) to ensure data consistency.

4. **Key Attributes**
   - **review_id** or **customer_id** can serve as a unique identifier (depending on availability).
   - **product_id** links to the item being reviewed, which is crucial for cross-category analysis.
   - **category** helps distinguish among books, watches, and electronics.

---

2. Business Problem

Focus: Cross-Selling & Customer Overlap

- **Primary Question:**
  Are there customers who purchase across multiple categories (books + watches + electronics)?

- **Supporting Questions:**
    1. *Which purchasing patterns exist across categories?*
        - Do customers who buy certain electronics also tend to buy certain books or watches?
    2. *Can we identify bundle opportunities or product recommendations?*
        - Is there a correlation between specific electronics and watch purchases that might indicate a potential bundle?
        - Are there popular book genres frequently bought alongside certain tech items?
- **Goal:**
  Leverage the reviews dataset to uncover insights into cross-category behavior, ultimately guiding **marketing strategies, personalized recommendations, and inventory decisions**. By pinpointing overlaps, the business can target customers with more relevant promotions and optimize stock for high-demand bundle opportunities.