## **Data Quality Report**

### **Overview**

This report summarizes the initial data quality assessment and cleaning steps performed on the Amazon Clothing Sales dataset.

### **Missing Values**

* Several columns contained missing values, notably:
  + **Categorical fields:** region, device\_type, payment\_method, customer\_age\_group, and brand.
  + **Numerical fields:** delivery\_days, and price.

### **Missing Value Handling**

* For categorical fields (region, device\_type, payment\_method, customer\_age\_group), missing values were imputed using the **mode** (most frequent category), ensuring minimal impact on data distribution.
* For brand, due to a high variety of unique values, missing entries were filled with the placeholder "Unknown" to avoid bias.
* For the numeric field delivery\_days, missing values were imputed using the **median**, providing robustness against outliers.
* For missing price values, unit price was calculated based on related fields (final\_price, quantity, and discount\_percent) using the formula:  
   price=final\_pricequantity×(1−discount\_percent100)\text{price} = \frac{\text{final\\_price}}{\text{quantity} \times (1 - \frac{\text{discount\\_percent}}{100})}price=quantity×(1−100discount\_percent​)final\_price​

### **Data Consistency Checks**

* Verified consistency between price, discount\_percent, and final\_price using the above formula; discrepancies were minimal.
* Standardized categorical values with inconsistent capitalization (e.g., "baby" vs "Baby").
* Corrected typographical errors in sub\_category (e.g., "Jack3ts" corrected to "Jackets").

### **Outliers and Anomalies**

* No extreme outliers detected in numeric fields that required removal.
* Values in review\_rating and is\_returned columns considered valid despite appearing as extremes (e.g., rating = 1, returned = 0).
* Identified some suspicious orders via anomaly detection (e.g., high discount + high quantity + returned), which were analyzed separately.

### **Duplicates and Integrity**

* No duplicate records found based on order\_id.
* Date fields converted and validated for correct formats.