DATA QUALITY REPORT

Dataset: Customer360Insights.csv

Number of Records: 2000 rows × 23 columns

1. Introduction

This report outlines the data cleaning methodology applied to the Customer360Insights dataset to ensure data quality, reliability, and usability for further analysis.

2. Data Cleaning Methodology

2.1 Handling Missing Values

Missing values were identified in several key columns, specifically:

- **'OrderConfirmationTime'**
- **'PaymentMethod'**
- **'OrderReturn'**
- **'ReturnReason'**

Actions Taken:

- **'OrderConfirmationTime'**, **'PaymentMethod'**, and **'OrderReturn'** were filled with "Unknown" to retain completeness without misrepresenting the data.
- **'ReturnReason'** was replaced with an empty string (') to distinguish missing reasons from valid data.

Result: All columns now have consistent, non-null values, ensuring uninterrupted analysis.

2.2 Handling Outliers

Outliers were detected in the following numeric columns:

- **'CreditScore'**
- **'MonthlyIncome'**
- **'Cost'**
- **'Price'**
- **'Quantity'**

Actions Taken:

- Applied a percentile capping technique, setting values below the 5th percentile and above the 95th percentile to the nearest boundary. This preserved the central distribution while mitigating the impact of extreme values.

Result: Outliers capped, preventing skewed analysis.

2.3 Feature Engineering & Data Transformation

Several transformations were applied to enhance data utility:

- **Datetime Conversion:** Columns such as SessionStart, CartAdditionTime, OrderConfirmationTime, and SessionEnd were converted to datetime format.
- **Session Duration:** A new column SessionDuration (in minutes) was derived from SessionEnd - SessionStart.

- **Date Extraction:** New columns SessionYear, SessionMonth, and SessionDay were created from SessionStart for improved time-based analysis.
- **Categorical Encoding:**
    - ❖ Gender was binary encoded (Male = 1, Female = 0).
    - ❖ Other categorical columns (Country, State, Category, and PaymentMethod) were label-encoded to convert textual data into numeric format for modeling.

Result: Enhanced dataset structure for better analysis and machine learning compatibility.

3. Final Dataset Overview

- **Final Data Shape:** 2000 rows × 27 columns (including engineered features)
- **Completeness:** All missing values addressed.
- **Consistency:** Outliers capped, datetime columns standardized, and categorical variables encoded.
- **Readiness:** Dataset is now clean, structured, and ready for advanced analytics or machine learning workflows.

4. Conclusion

The data cleaning methodology successfully addressed missing values, outliers, and data transformation. The resulting dataset ensures improved data quality, making it suitable for reliable insights and predictive modeling.

Would you like to extend this report to include data visualization summaries or data profiling metrics?