Data Cleaning & Preprocessing Report

1. Data Examination

All three datasets (Customer Sales, Healthcare, and Social Media) were analyzed for missing values, inconsistencies, and outliers.

2. Findings & Issues Identified

(A) Customer Sales Data

- **Missing Values:** None detected.
- **Purchase Amount Format Issue:** Initially stored as a string with "$" symbols. **Fixed:** Converted to numerical format.
- **Outliers Analysis:**
    - The minimum purchase amount is **$100**, and the maximum is **$1500**.
    - There are no extreme anomalies in this range.

(B) Healthcare Patient Records

- **Missing Values:** None detected.
- **Date Format:** Admission and discharge dates are correctly formatted.
- **Outliers in Age:**
    - The minimum age is **28 years**, and the maximum is **60 years**.
    - No extreme outliers.

(C) Social Media Posts Data

- **Missing Values:** None detected.
- **Outliers in Likes & Comments:**
    - Likes range from **20 to 50** with no extreme anomalies.
    - Comments range from **5 to 20**, which is within normal engagement levels.

---

3. Proposed Data Cleaning Strategies

✅ For Customer Sales Data:

- Ensure all monetary values are stored as numbers for calculations.
- Further analyze purchase amounts if the dataset expands to detect fraudulent activities.

✅ For Healthcare Data:

- Validate admission and discharge dates to avoid incorrect or inconsistent entries.
- If the dataset grows, check for impossible ages (e.g., < 0 or > 120).

✅ **For Social Media Data:**

- Convert text-based date columns into a standardized datetime format.
- If there are bot-generated posts, remove those based on suspicious engagement patterns