Data Analysis Report

1. Introduction

This report presents an in-depth analysis of the Body Shop dataset. The objective was to perform data cleaning, exploratory data analysis (EDA), and identify relationships between variables. The findings aim to enhance understanding of the dataset and reveal patterns that could inform business strategies.

2. Data Cleaning

Initial Inspection:

- The dataset structure was examined using df.info() and df.describe().
- Date columns were converted to datetime format for proper handling.

Handling Missing Values:

- Missing values in numerical columns were imputed using the median.
- Categorical columns with missing values were filled with "Unknown."

Duplicate Records:

- Duplicate records were identified and removed, ensuring data integrity.

Outliers Treatment:

- Outliers in numerical columns were detected using the Z-score method (threshold of 3) and removed to prevent skewed analyses.

Standardization:

- Column names and categorical values were standardized to lowercase and underscores replaced spaces to ensure consistency.

3. Exploratory Data Analysis (EDA)

Univariate Analysis:

- Histograms and KDE plots were generated for numerical variables to visualize their distributions.

Bivariate Analysis:

- Pair plots were used to explore relationships between numerical variables.
- A correlation matrix highlighted the strength of relationships between these variables.

Multivariate Analysis:

- Grouped comparisons were conducted, excluding transaction_id and product_name to avoid irrelevant visualizations.
- Date variables were grouped into 8 bins to enhance the clarity of visualizations.

- Boxplots revealed differences in numerical variables across categories, providing insights into patterns and variations.

4. Key Findings

- **Data Integrity:** The dataset required extensive cleaning, with missing values and duplicates handled appropriately.
- **Outliers:** Removal of extreme values improved the overall robustness of the analysis.
- **Relationships:** The correlation matrix indicated notable relationships between some numerical features, which could be further explored for predictive modeling.
- **Temporal Trends:** Binning the dates provided a clearer view of trends over time, suggesting periodic variations.

5. Conclusion and Recommendations

- The dataset was successfully cleaned and prepared for deeper analysis.
- Insights from the visualizations can inform product performance analysis, sales trends, and other strategic decisions.
- Future work could involve building predictive models using these insights.