📊 Data Cleaning & Preprocessing Report

This document compares the original Mini Guide steps with the actual code implementation, **highlighting extra creative additions** made to improve the data cleaning workflow.

# Comparison Table

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.no** | **Mini Guide Description** | **Implemented Code** | **Extra Feature** |
| 1 | Import the dataset and explore basic info | Yes | Added full summary + original copy backup |
| 2 | Handle missing values using mean/median/imputation | Yes | Visualized missingness using `missingno` heatmap |
| 3 | Convert categorical features using encoding | Yes | Handled rare categories by grouping infrequent ones |
| 4 | Normalize/standardize numerical features | Yes | Log transformation for skewed features |
| 5 | Visualize outliers using boxplots and remove them | Yes | Used both IQR and Z-Score method |
| 6 | — | Yes | Dropped duplicate rows |
| 7 | — | Yes | Auto type correction for numeric/date columns |
| 8 | — | Yes | Final summary function for transparency |
| 9 | — | Yes | Cleaned dataset saved to CSV |

Summary:

The enhanced script goes beyond the basics by adding smart automation, handling data integrity (like duplicates, types), skewness correction, and a professional cleaning summary.