Data Collection and Preprocessing Phase

Date	6 July 2025
Team ID	SWTID1749620997
Project Title	Online Payments Fraud Detection
Maximum Marks	2 Marks

Data Quality Report

The Data Quality Report will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source	Data Quality Issue	Severity	Resolution Plan
Dataset	Presence of skewed numerical variables like amount, oldbalanceOrg, etc.	Moderate	Applied log transformation to reduce skewness and normalize distributions.
Dataset	Outliers in `amount`, oldbalanceOrg, newbalanceDest columns	High	Detected outliers using IQR method and capped them to reduce model distortion.
Dataset	High cardinality in columns like nameOrig, nameDest	Moderate	Dropped these columns as they act like IDs and add no predictive value.
Dataset	isFlaggedFraud column with low variance (mostly 0s)	Low	Dropped due to low informational value and to prevent bias.
Dataset	Skewed class imbalance in isFraud target variable	High	Addressed later in modeling phase using techniques like oversampling/undersampling.
Dataset	Some columns lacked intuitive meaning for downstream models (e.g., `step` as timestamp)	Low	Retained `step` as-is but considered converting it to time-based bins if needed.

Dataset	No missing values found	Low	Verified using df.isnull().sum(); no imputation required.