Data Collection and Preprocessing Phase

Date: 31 July 2025
SkillWallet ID: SWUID20240141492
Project Title: Employee Performance Prediction using Machine Learning

---

Maximum Marks: 2 Marks

Data Quality Report:

The Data Quality Report highlights issues in the dataset sourced from Kaggle (Garment Worker Productivity Dataset). It includes identified issues, their severity, and proposed resolution plans to ensure data consistency and quality for model development.

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle Dataset	Missing values in `idle_time`, `idle_men`, and `incentive` columns	Moderate	Use **mean/median imputation** for numerical fields.
Kaggle Dataset	Presence of categorical variables such as `department`, `day`, and `quarter`	Moderate	Apply **Label Encoding or One-Hot Encoding**.
Kaggle Dataset	Outliers in `over_time` and `no_of_style_change` columns	High	Apply **IQR (Interquartile Range) method** to handle outliers.
Kaggle Dataset	Variations in `targeted_productivity` values due to scaling issues	Low	**Normalize or standardize** the feature.
Kaggle Dataset	Mixed data formats in `month` column (text and numbers)	Low	Convert all month values to a **uniform numeric format**.