# Data Quality and Performance in Action

## 1. Data Quality Assessment and Improvement

**Dataset Selection and Initial Assessment**

For this project, I selected an Excel sample dataset which contains responses from a survey conducted in 2024 to understand gaming habits, preferences, and player behaviours. The dataset provides valuable insights into the gaming landscape, helping developers, marketers, and researchers identify trends and preferences in the gaming community. The data was assessed based on five quality dimensions: accuracy, completeness, consistency, timeliness, and uniqueness.

Assessment

- **Accuracy**: Data was cross-verified with known valid ranges.

- **Completeness**: Some columns have up to 40% missing values, especially those with duplicate headers (columns from N to V are duplications of columns E to M), as seen in Figure 1.

- **Consistency**: Formatting inconsistencies were noted, especially in the ‘*Location*’ and ‘*What is your favourite game?*’ fields.

- **Timeliness**: Timestamps are within a reasonable range (October 23–25, 2024), indicating data was collected over three days.

- **Uniqueness**: No duplicate rows were found in the dataset.

Data Quality Improvement Steps

I focused on improving completeness and consistency. The steps were as follows:

1. **Completeness**: Removing duplicated columns (from N to V).
2. **Consistency**: Selecting one format for ‘*Bangalore, Karnataka*’ (as seen in Figure 2) and ‘*Bhubaneswar, Odisha*’ (as seen in Figure 3) on ‘*Location*’, using Capital first and State second. And one format for ‘*Call of Duty*’ on ‘*What is your favorite game?*’ (as seen in figure 4).

\*Metrics\*: After cleaning, consistency was improved to 95%, and duplicate entries were entirely removed (see Figure 5).

Screenshots and Annotations

A white paper with black text

Description automatically generated

**Figure 1** shows the initial state of the data, with highlighted areas indicating missing values and duplicates.

A screenshot of a computer screen

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

**Figures 2 and 3** show the inconsistent data, on ‘*Location*’ field.

**Figure 4** shows the inconsistent data on ‘*What is your favorite game?*’ field.

A screen shot of a computer

Description automatically generated

**Figure 5** shows the final state of the data, after changes.

## 2. Database Schema Design with SQL as a DDL

**Schema Design**

A star schema was chosen to optimize data performance and quality. The central fact table holds transaction data, while dimension tables cover customers, products, and dates. This structure supports efficient querying, reducing redundancy (as seen in figure 10).

**Table Definitions**

- **Customers Table:** Contains unique customer IDs, names, and locations (as seen in figure 6).

- **Products Table**: Includes product IDs, names, and category information (as seen in figure 6).

- **Transactions Table:** Holds transaction ID, customer ID (foreign key), product ID (foreign key), transaction date, and amount (as seen in figure 6).

\*Assumptions\*: Data types include integers for IDs, decimals for amounts, and strings for descriptive fields. The Date from Transactions table is in DATE format to enable efficient querying.

**SQL DDL and DML Commands**

The sample SQL commands used to define these tables, to add some sample data can be seen in Figure 6:

A screenshot of a computer

Description automatically generated  
**Figure 6** shows the MySQL queries to create the tables: Customers, Products, Transactions, queries to add some sample data and queries to show the data from the tables

A screenshot of a computer

Description automatically generated

**Figure 7** shows the Customers table

A screenshot of a computer

Description automatically generated

**Figure 8** shows the Products table

A screenshot of a computer

Description automatically generated

**Figure 9** shows the Transactions table

A screenshot of a computer

Description automatically generated

**Figure 10** displays the schema creation in SQL, showing relationships among tables.

This document demonstrates a complete workflow, from data quality assessment to schema design, aimed at enhancing both data quality and performance.