Data Cleaning with SQL

Project Overview

This project focuses on performing data cleaning on employee-related datasets using SQL. The goal is to identify and address issues such as missing values, duplicates, outliers, and integrity problems.

Steps Performed

1. Viewing Raw Data

Employee table

PerformanceRating table

2. Searching for Missing Values

Counting missing EmployeeID values in the Employee table.

Counting missing EmployeeID values in the PerformanceRating table.

3. Identifying Duplicate Entries

Checking for duplicate EmployeeID values in the Employee and PerformanceRating tables using partitioning.

4. Searching for Outliers & Invalid Data

Detecting negative or unrealistically high salaries.

Ensuring valid age range (18-100 years).

Checking if YearsAtCompany exceeds Age - 18.

5. Checking for Incorrect Rating Values

Ensuring EnvironmentSatisfaction, JobSatisfaction, and WorkLifeBalance values are between 0 and 5.

6. Data Integrity Check

Identifying EmployeeIDs present in PerformanceRating but missing in Employee.

Yahya Ali