Data Cleaning in

SQL

1.Import Data

First, import the Excel data into a SQL database table using a tool like SQL Server Management Studio.

2. Identify Missing Values

Use SQL queries to identify any missing or null values in the dataset. This helps in understanding the extent of missing data and planning for imputation or removal.

3. Remove Duplicates

Utilize SQL's 'DISTINCT' keyword or 'GROUP BY' clause to identify and remove duplicate rows from the dataset. This ensures that each observation is unique.

4. Standardize Data Formats

Use SQL functions like UPPER, LOWER, TRIM, etc., to standardize text formats and remove leading or trailing spaces. This ensures consistency in the data.

5. Correct Data Types

Convert data types of columns as needed using SQL's CAST or CONVERT functions. For example, convert string representations of numbers to actual numeric types.

6. Handle Outliers

Identify and handle outliers using SQL queries. This might involve filtering out extreme values or applying statistical techniques for outlier detection.

7. Normalize Data

Normalize the data if necessary to reduce redundancy and improve data integrity. This might involve splitting data into separate tables and establishing relationships between them.

8. Validate Constraints

Validate data against defined constraints such as foreign key constraints, unique constraints, etc., to ensure data integrity and consistency.

9. Impute Missing Values

If appropriate, impute missing values using techniques like mean imputation, median imputation, or predictive modeling.

10. Review and Validate

Finally, review the cleaned dataset to ensure that it meets the quality standards and is ready for analysis. Validate the results against the original Excel file to ensure accuracy.

Follow me

RESHARE/REPOST