ANKIT GUPTA

Data Analyst | Prompt Engineer

DATA CLEANING & TRANSFORMATION

Data cleaning and transformation are crucial steps in the data analytics process to ensure the data is accurate, consistent, and ready for analysis. Here are some key things that a data analyst should check during data cleaning and transformation

DATA CLEANING CHECKLIST	DATA TRANSFORMATION CHECKLIST
□ Missing Values Handling	□ Data Normalization/Scaling
□ Outliers Detection & Handling	□ Feature Engineering
□ Duplicate Records Removal	□ Data Type Conversion
□ Data Formatting Consistency	□ Data Integration
□ Handling Typos & Inconsistencies	□ Domain Knowledge Utilization
□ Data Quality Assessment	□ Documentation

DATA CLEANING

DATA CLEANING CHECKLIST:

□ **Missing Values:**

Identify and handle missing values appropriately. This might involve imputation techniques like mean, median, or mode substitution, or more advanced methods like predictive modeling to estimate missing values.

Example:

Identify missing values in the "Age" column of a dataset containing customer information.

Action:

Use mean imputation to fill missing age values based on the average age of customers.

DATA CLEANING

DATA CLEANING CHECKLIST:

□ **Outliers:**

Detect and handle outliers that could skew analysis results. Techniques like z-score, IQR (Interquartile Range), or clustering can be used to identify outliers and decide whether to remove, transform, or treat them separately.

Example:

Detect outliers in the "Income" column of a dataset containing salary information.

Action:

Remove outliers that are beyond three standard deviations from the mean salary.

DATA CLEANING

DATA CLEANING CHECKLIST:

□ **Duplicate Records:**

Check for and remove duplicate records to avoid redundancy in the dataset, which could bias analysis results. This involves identifying identical rows or records based on key attributes.

Example:

Check for duplicate entries in the "Customer ID" column of a customer database.

Action:

Remove duplicate customer records based on unique customer IDs.

DATA CLEANING

DATA CLEANING CHECKLIST:

□ **Data Formatting:**

Ensure consistency in data formatting across different fields, such as date formats, numeric formats, and categorical variables. Standardizing formats improves data quality and facilitates analysis.

Example:

Ensure consistency in date formats across different date columns.

Action:

Convert all date formats to YYYY-MM-DD format for uniformity.

DATA CLEANING

DATA CLEANING CHECKLIST:

□ **Handling Typos and Inconsistencies:**

Identify and correct typos or inconsistencies in the data, such as variations in spelling, capitalization, or naming conventions. This improves the accuracy and reliability of analysis results.

Example:

Identify inconsistent spellings of product names in a sales dataset.

Action:

Standardize product names by correcting typos and ensuring consistent spelling.

DATA CLEANING

DATA CLEANING CHECKLIST:

❑ **Data Quality Assessment:**

Perform checks to assess overall data quality, including assessing data completeness, accuracy, and consistency. Visualization tools and statistical metrics can help in identifying potential data quality issues.

Example:

Assess data completeness in a sales dataset.

Action:

Check for missing values in key columns like "Order ID" and "Customer ID."

DATA TRANSFORMATION

DATA TRANSFORMATION CHECKLIST:

□ Data Normalization/Scaling:

Normalize or scale numeric features to bring them to a similar scale, especially when using algorithms sensitive to feature scales like K-means clustering or gradient descent-based methods.

Example:

Normalize numeric features like "Height" and "Weight" in a dataset containing biometric information.

Action:

Use min-max scaling to scale all numeric features between 0 and 1.

DATA TRANSFORMATION

DATA TRANSFORMATION CHECKLIST:

□ **Feature Engineering:**

Create new features or transform existing ones to enhance the predictive power of the dataset. This could involve techniques like binning, one-hot encoding categorical variables, or creating interaction terms.

Example:

Create a new feature "Total Revenue" by combining "Quantity" and "Unit Price" columns in a sales dataset.

Action:

Multiply the "Quantity" column by the "Unit Price" column to calculate total revenue for each transaction.

DATA TRANSFORMATION

DATA TRANSFORMATION CHECKLIST:

❑ Data Type Conversion:

Convert data types appropriately, ensuring compatibility with analysis tools and algorithms. For example, converting string variables to numeric or categorical variables to factors

Example:

Convert categorical variables like "Gender" into numeric format for analysis.

Action:

Use one-hot encoding to convert categorical variables into binary format (e.g., Male = 1, Female = 0).

DATA TRANSFORMATION

DATA TRANSFORMATION CHECKLIST:

❑ **Data Integration:**

Merge or join multiple datasets if needed, ensuring consistency and coherence across different sources. This involves identifying common key variables and combining datasets accordingly.

Example:

Merge customer demographic data with transaction data for analysis.

Action:

Use common identifiers like "Customer ID" to merge the two datasets into a single dataset.

DATA TRANSFORMATION

DATA TRANSFORMATION CHECKLIST:

□ **Domain Knowledge:**

Utilize domain knowledge to validate data and make informed decisions during the cleaning and transformation process. Understanding the context of the data helps in identifying anomalies and making appropriate transformations.

Example:

Understand the business context of a dataset containing website traffic data.

Action:

Identify relevant metrics for analysis based on the business goals, such as conversion rate, bounce rate, etc.

DATA TRANSFORMATION

DATA TRANSFORMATION CHECKLIST:

□ **Documentation:**

Document all data cleaning and transformation steps undertaken, including reasons for decisions made and any assumptions or transformations applied. This ensures transparency and reproducibility of the analysis process.

Example:

Document all transformations performed on a dataset containing stock market data.

Action:

Maintain a log detailing each transformation step, including the rationale behind the transformation and any assumptions made.