Data validation component :

Purpose of Data Validation Component :

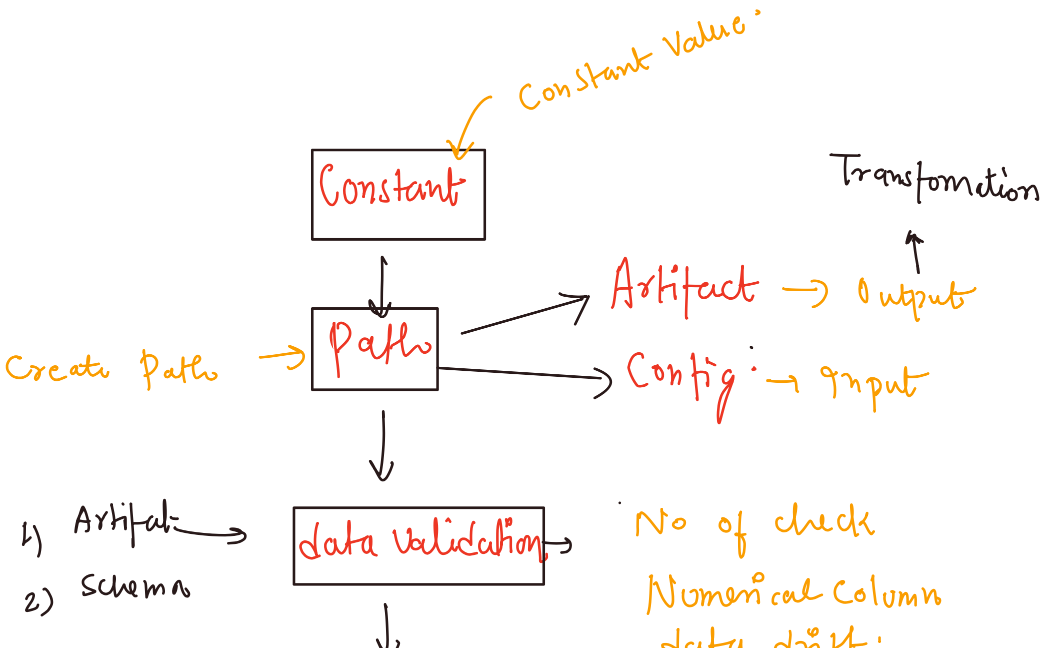
1. **Data Reading:** Reads the trained and test datasets.
2. **Column Validation**: Validates the presence of required columns in both datasets.
3. **Numerical Column Validation:** Validates the presence of required numerical columns in both datasets.
4. **Data Drift Detection:** Detects data drift between the trained and test datasets.
5. **Artifacts Creation:** Constructs and returns a DataValidation Artifacts object containing validation results and file paths

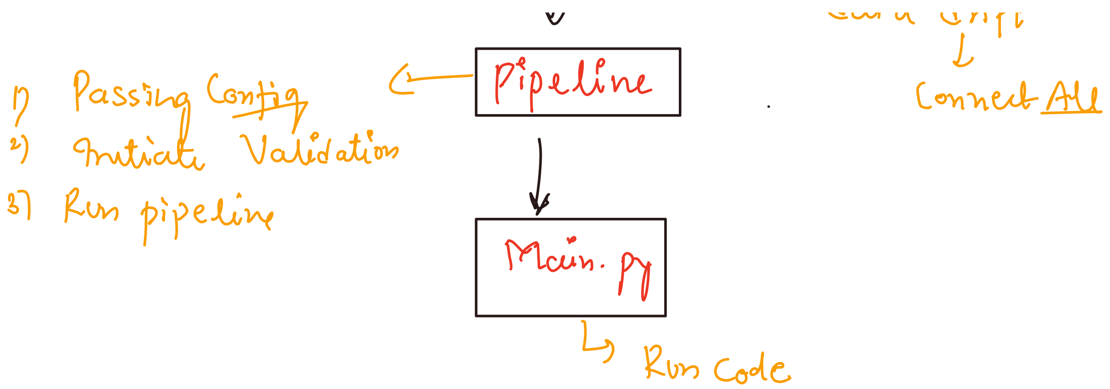
Data validation means :

Data Quality: Ensuring the quality and integrity of data by checking for missing values, outliers, inconsistencies, and errors.
Data Completeness: Verifying that all required data fields and features are present and complete.
Data Consistency: Checking the consistency and coherence of data across different sources, datasets, or time periods.

Purpose of Column Validation

The purpose of the validate_number_of_columns method is to validate whether a given pandas DataFrame contains the correct number of columns as specified in a schema configuration. It ensures that the DataFrame aligns with the expected structure defined by the schema configuration, which is crucial for subsequent data processing and analysis task.





1) Passing Config
2) Initiate Validation
3) Run pipeline

New Changes → 1) Utils Folder & Main_utils files

2) Config folder → schema file.