Title: Data Wrangling on Academic Performance Dataset

Theory Concepts

1. Data Wrangling:
Data wrangling (or data cleaning) is the process of transforming and mapping raw data into a more usable format. In real-world datasets, data often contains issues like missing values, inconsistencies, incorrect types, and outliers. These must be handled properly to ensure accurate analysis.

2. Missing Values:
Missing data can distort analysis and lead to incorrect conclusions. Common techniques include:

- Mean/Median imputation
- Dropping rows/columns
- Predictive imputation

3. Inconsistencies:
Non-numeric or incorrectly typed entries (e.g., text in numeric columns) are identified and either corrected or removed.

4. Outliers:
Outliers are extreme values that deviate significantly from other observations. These can affect models and statistics.

- Detection via boxplots or Z-score method
- Handling by capping, removal, or transformation

5. Data Transformation:
Used to modify the distribution of data or change scale. For example:

- `np.log1p()` is used to reduce skewness and normalize the data.

Algorithm (in short)

1. **Create dataset** with columns: `Student_ID`, `Math`, `Physics`, `English`, `Attendance`, `GPA`.
2. **Introduce errors**:
    - Missing value in `Math`
    - Out-of-range value in `Physics`
    - Non-numeric entry in `English`
3. **Scan for missing values** using `df.isnull().sum()` and for type inconsistencies using `pd.to_numeric(errors='coerce')`.
4. **Visualize numeric data** with `boxplot()` to identify outliers.
5. **Fix issues**:
    - Replace missing `Math` value with column mean.
    - Convert `Physics` scores >100 to NaN.
    - Convert `English` to numeric, invalid entries become NaN.
    - Drop rows with any remaining NaN values.
6. **Re-check outliers** using Z-score method.
    - Identify rows with Z-score > 3.
    - Remove these outliers.
7. **Apply transformation** on `GPA` using `np.log1p()` to normalize the skewed distribution.
8. **Visualize** the result using histograms before and after transformation.

Conclusion

Through this exercise, we successfully simulated real-world data issues and performed comprehensive wrangling steps including:

- Handling missing values with imputation
- Fixing data type inconsistencies
- Detecting and removing outliers
- Applying transformation to reduce skewness

These operations improve data quality and prepare it for accurate analysis or machine learning modeling. The final dataset is cleaner, consistent, and better suited for further processing or visualization tasks.