:

📘 Important Topics – Theory Explanations for Data Wrangling Project

---

1. Data Wrangling

Data wrangling refers to the process of converting raw data into a clean and structured format for analysis. This includes detecting and correcting errors, handling missing values, type conversions, outlier treatment, and transforming data into formats suitable for modeling. It is a crucial first step in the data science pipeline as it directly impacts the accuracy and effectiveness of subsequent analyses.

---

2. Missing Values Handling

Missing values occur when no data value is stored for a variable in an observation. In datasets, they can arise from human error, sensor faults, or incomplete records. These values can be handled using various strategies such as deletion, mean/median/mode imputation, or advanced techniques like interpolation or predictive modeling. In this project, we used mean imputation to fill missing scores and dropped rows when values couldn’t be resolved.

---

3. Type Conversion

Type conversion ensures that each column in a dataset has the appropriate data type. For instance, numeric columns should not contain strings or mixed types. Improper data types can lead to errors during computation or statistical analysis. Python's `pandas` library provides methods like `astype()` and `to_numeric()` to explicitly convert column types. In our code, this was used to clean up non-numeric values from the 'English' column.

---

4. Outlier Detection

Outliers are values that are significantly different from other observations in the dataset. They can distort statistical analyses and affect machine learning models. Common methods of outlier detection include the Z-score, IQR method, and visual techniques like boxplots. In this project, we used Z-score to mathematically identify outliers and corrected them by marking extreme values as missing.

---

5. Z-Score Method

The Z-score standardizes values by measuring how many standard deviations a data point is from the mean. It is useful in identifying outliers, especially in normally distributed data. If the Z-score is greater than a threshold (typically 3 or -3), the point is considered an outlier. We applied this technique to filter out statistically abnormal academic scores.

---

6. Log Transformation

Log transformation is used to reduce skewness in data and stabilize variance. It is particularly useful when dealing with positively skewed distributions. The function `np.log1p()` is often used to handle zero or low values safely. In our case, we applied this transformation to the GPA column to prepare it for potential statistical modeling.

---

7. Boxplot Visualization

Boxplots are graphical representations of data distribution. They show the median, quartiles, and outliers in a concise format. They are particularly helpful for comparing multiple columns or identifying anomalies in numeric data. We used boxplots to visualize subject-wise marks and highlight any unusually high or low values.

---

8. Data Cleaning

Data cleaning is the process of correcting or removing incorrect, corrupted, duplicated, or incomplete data within a dataset. It ensures consistency and integrity, which is essential for accurate data analysis. Cleaning can involve operations like trimming whitespace, removing duplicates, correcting formats, or standardizing units. Our project addressed several cleaning tasks, including fixing incorrect entries and removing invalid rows.

---

9. Synthetic Data Generation

Synthetic data is artificially generated data that simulates real-world data. It is useful for testing models or demonstrating data workflows. In our project, we used random functions like `np.random.randint()` and `np.random.choice()` to create a mock dataset representing student academic performance.

---

10. Dataframe Operations with Pandas

The pandas library in Python provides powerful data structures like DataFrames for working with structured data. Operations such as slicing, filtering, aggregation, and transformation are made easy with its API. In this project, pandas was the core tool used to manage, clean, and transform the dataset efficiently.