📄 Viva Q&A: Data Wrangling on Academic Performance Dataset

---

1. What is data wrangling?

Ans: Data wrangling is the process of cleaning, transforming, and organizing raw data into a usable format for analysis.

---

2. Why is data wrangling important?

Ans: It ensures data quality, handles inconsistencies, and prepares data for accurate analysis or modeling.

---

3. Which Python libraries are used in this project?

Ans: `pandas`, `numpy`, `matplotlib.pyplot`, and `scipy.stats`.

---

4. How are missing values identified in a DataFrame?

Ans: Using `df.isnull().sum()` to count null entries column-wise.

---

5. How did you handle the missing value in the ‘Math’ column?

Ans: By replacing the missing value with the mean of the column using `df['Math'].fillna(df['Math'].mean(), inplace=True)`.

---

6. What is an outlier?

Ans: An outlier is a data point that significantly deviates from other observations in the dataset.

---

7. How did you detect outliers in the dataset?

Ans: Using a boxplot for visualization and Z-score method for numerical detection.

---

8. What is the Z-score?

Ans: It measures how many standard deviations a data point is from the mean. A Z-score > 3 is usually considered an outlier.

---

9. How were non-numeric values handled in the ‘English’ column?

Ans: By converting the column using `pd.to_numeric(errors='coerce')` which turns invalid entries into NaN.

---

10. Why were rows with NaN dropped at the end?

Ans: After attempting to fix issues, remaining NaN rows are dropped to avoid corrupt data during analysis.

---

11. What transformation was applied to the ‘GPA’ column?

Ans: A log transformation using `np.log1p()` to reduce skewness in the distribution.

---

12. What does `np.log1p()` do?

Ans: It applies log(1 + x) transformation, helping normalize positively skewed data.

---

13. What kind of error was introduced in the 'Physics' column?

**Ans:* An out-of-range score (120) was added, which exceeds the typical max of 100.

---

14. How was the Physics score above 100 handled?

Ans: Values above 100 were set to NaN using a conditional check.

---

15. What does the `dropna()` function do?

Ans: It removes rows or columns containing missing (NaN) values.

---

16. Why is type conversion important in data wrangling?

Ans: To ensure the dataset contains valid data types for computation and analysis.

---

17. How can we visualize numeric data distributions?

Ans: Using histograms or boxplots from `matplotlib.pyplot`.

---

18. What does the `boxplot()` help us identify?

Ans: It highlights the spread and outliers in numerical data.

---

19. Why were synthetic errors introduced into the dataset?

Ans: To simulate real-world data issues and demonstrate how to clean them.

---

20. What is the shape of the dataset before and after cleaning?

Ans: Initially 10 rows, some removed after handling missing values and outliers.

---

21. What is the role of `np.random.randint()` in this code?

Ans: It generates random integers for simulating student marks and attendance.

---

22. How does Z-score-based outlier detection work?

Ans: By calculating the standard deviation distance of each point from the mean.

---

23. Can you name other outlier detection methods?

Ans: IQR method, DBSCAN clustering, isolation forest, and visual inspection.

---

24. Why did we use `inplace=True` while filling missing values?

Ans: To modify the original DataFrame directly without reassignment.

---

25. What is the final outcome of the wrangling process?

Ans: A clean, consistent, and transformed dataset ready for further analysis or modeling.