**Discussion Questions**

1 .What are the key advantages of using Pandas for data cleaning compared to other methods?

* **Ease of use**: Pandas has intuitive functions like .dropna(), .fillna(), .duplicated(), .replace()It makes sense quickly, without complex reasoning its natural and self-explanatory.
* We can inspect and clean data in the same framework (e.g., .isnull().sum(), .describe()).
* Handling missing values depends on whether they’re random or carry meaning.
* Documentation ensures reproducibility and transparency.
* Sometimes, dropping rows is safer than imputing to avoid bias.

1. How would you approach to handling missing values differ if the missing data was not random but had a pattern or meaning?

* If data is Missing Completely at Random (MCAR):that means missing completely by chance simple imputation (mean, median, mode) is fine. Dropping can also be fine.
* If data is Missing at Random (MAR): statistical/machine learning models is used to predict missing values based on other features. Example: filling missing income using education + occupation.
* If data is Not Missing at Random (NMAR): the missingness itself carries meaning, like income data not disclosed by participants so strong assumptions need to be made about the missingness mechanism or specialized statistical models are used.

1. What types of data quality issues might not be immediately visible through simple DataFrame inspection methods?

* **[Inconsistent Data Formats/Units:](https://www.google.com/search?sca_esv=89e7bcb64ef3cfc7&rlz=1C1JZAP_enUS933US933&cs=0&sxsrf=AE3TifNd3KsBP49492CL4r0XH3MXFO8t-w%3A1759423742003&q=Inconsistent+Data+Formats%2FUnits&sa=X&ved=2ahUKEwiCoZ6Y_IWQAxUHGFkFHS9NDV8QxccNegQIFRAD&mstk=AUtExfBti3vgdQmfxVnXRVL8fhWis8Y4cp36zwYduwsDoNOeUXVXciu6rQ19ME4PddIT4cn3TSVcMe174_A6XHBRMcHAwUqp32HM2Bdm1V5gmt5lcn_-ZmR17qp3XzsHvwhMD3g56bucKpPYiaJ_hgIGWS8KxaJ-io59fH4FlVIW22mF5e2lPSV2ThiLaE4oxqzZeczyhvY6JEAhRUl-_SUBaiiGyfVSLla3POhB8lJqrEeJtTw7VQUny0sFzmpBeIrJm--6DbycvgaxbuT_EPl6QhoX&csui=3" \t "_blank)**

(e.g., "USA", "U.S.A.", "United States" for a country; "kg", "kilograms" for units)

* Logical inconsistancies eg an age of 130,a negative price or start date after end date.

1. How would you document your data cleaning process to ensure reproducibility?

* **Use Markdown Cells for Explanations:** Clearly describing the purpose of each cleaning step, the rationale behind decisions
* **Provide Inline Comments:** Adding comments within code cells to explain specific lines or blocks of code, especially for complex or less intuitive operations.
* Display head(), info(), describe(), or specific value counts after significant cleaning steps to demonstrate the effect of the transformations.
* Use visualizations (e.g., histograms, box plots) to illustrate data distributions before and after cleaning, particularly for outlier

1. In what scenarios might it be better to remove rows with missing values rather than imputing them?

* When a very small percentage of rows contain missing values or extremely high number of missing values that it is considered incomplete row.
* When the missing value is absolutely crucial for analysis.
* When its MNAR situation where imputing with mean or median values misrepresent the data.