

## Q&A

### 1. What are missing values and how do you handle them?

Missing values represent absent data points in a dataset. In this analysis, the 'Income' column had missing values, which were handled using the `.drop_duplicates()` function.

### 2. How do you treat duplicate records?

Duplicate records were identified and removed using the `.drop_duplicates()` function (though no duplicates were found in this dataset).

### 3. Difference between `dropna()` and `fillna()` in Pandas?

`dropna()` removes rows or columns with missing values, while `fillna()` replaces missing values with a specified value (e.g., the median, mean, or a constant). In this analysis, `fillna()` with the median was used for the 'Income' column.

### 4. What is outlier treatment and why is it important?

Outlier treatment involves identifying and handling extreme values in a dataset. Although the analysis visualized potential outliers using boxplots, no specific outlier treatment was performed in this task. Outliers can skew statistical measures and affect model performance.

### 5. Explain the process of standardizing data.

Data standardization involves transforming data to have a consistent format and scale. In this analysis, text values in the 'Marital\_Status' columns were standardized using `replace` function.

### 6. How do you handle inconsistent data formats (e.g., date/time)?

Inconsistent date/time formats were handled by converting the 'Dt\_Customer' column to a consistent datetime format (`%d-%m-%Y`).

### 7. What are common data cleaning challenges?

Common challenges include missing values, inconsistent data formats, duplicate records, outliers, and incorrect data types.

### 8. How can you check data quality?

Data quality can be assessed through various methods, including checking for missing values, duplicates, inconsistent formats, and data types. Visualizations (histograms, boxplots) and summary statistics can also reveal potential data quality issues.