**Scenarios**

**Data Preprocessing**

1. In the given CSV file do the preprocessing.

* Loaded the CSV file into a DataFrame in python.
* Handle missing values (e.g., fill NaNs, drop rows).
* Convert data types (convert decimal to integer) in python
* Removed blank entries in customerid and transactiondate in SQL and imported the CSV to the database.

1. Write the steps you followed.

* Data Cleaning
* Data Transformation
* Data Validation
* Data Visualization

1. What was your thought process when you first saw the data.

* Identify the structure of the data (columns, data types).
* Detect and decide how to manage missing values.
* Look for any inconsistencies or anomalies in the data.
* Determine if any columns need to be transformed or normalized.

**Data Aggregation and Grouping**

1. What all fields among them you think can be aggregated? Name them.

* TotalAmount
* DiscountApplied
* Quantity
* Trustpoints

1. What kind of aggregation (for every column) would make sense and why?

* TotalAmount: Sum, Average (to understand total and average sales).
* Quantity: Sum (to understand total quantity sold by products)
* DiscountApplied: Sum, Average (to understand total and average discounts given for products or customers).
* Trustpoints: Sum (to understand each customers earned turstpoints)

**Data Validation**

1. How do you know, your preprocessing was correct?

* Check for any remaining missing values.
* Ensure numerical values fall within expected ranges.
* Verify data types of each column.
* Ensure no duplicates are present.
* Verify the related columns align logically ( TotalAmount should equal PricePerUnit multiplied by Quantity if applicable)

1. How will you validate your results?

* Compared with a known subset of the data.
* Cross-validate with other data sources if available.
* Used statistical methods to check for anomalies (check for outliers in TotalAmount)

1. Do you follow any specific validation process for all questions? Explain.

* Conducted an initial inspection of the data.
* Perform preprocessing steps to clean the data.
* Checked the data after each preprocessing step.
* Conducted a final validation of the entire dataset.

1. What are the edge cases you can think of?

* Absence of values in essential columns.
* Presence of outliers in numerical fields.
* Incorrect or inconsistent data types.

1. What all data integrity points you want to mention for the given scenario?

* Ensured the relationships between tables are maintained (e.g., Customerid in transactions should exist in the customer table).
* Ensured all entries follow the proper format.
* Ensured all required fields are filled
* Accuracy: Validated the data accurately represents real-world values

**Data Visualisations**

1. What all projections are possible out of the data.

* Customer ID who ordered more in quantity
* Total Sales by Product Category
* Which product has made the maximum/least sales
* Customer ID with number of Transactions
* Total discounts used by each customers
* Number of transactions made in particular day
* Pre-ordered product list ignored pending payment
* Total Trust Points Used by Each Customer
* List the products that were returned
* Sales Contribution by Payment Method
* Average Discount by Product Category
* Most used payments
* Customerid with most pending payments
* List the pending payment process

1. How would be know if the data is linearly projected?

* Ensure the linearity assumption is valid for the data. This can be done by examining the relationship between the independent and dependent variables.

1. For all the different combinations of possible projections, what are the suitable graphical representation? (Eg: Line Chart or Bar Graph)

* Line chart to represent the average sales for payment method
* Bar graph represents total sales by product, total trust points earned by customer & average discount for each product.

Output everything in Python. Also, other than visualization, insert the csv in sql and do everything in sql. Attach the code.