**Sales Data Analysis**

**Data Preprocessing**

1. In the given CSV file do the preprocessing.

The preprocssing of the data is given in sales\_analysis.ipynb file which is run on jupyter notebook.

1. Write the steps you followed.
2. Importing the necessary libraries and data set.
3. Analysing the dataset using data.info() and data.describe()
4. Converting TransactionDate to datetime format
5. Checking for null values using isna() function
6. Filling NaN values of CustomerID by Mode
7. Filling NaN values of TransactionDate by Forward fill method
8. Filling the NaN values of PricePerUnit by Mean of the ProductCategory
9. Filling the NaN values of TotalAmount by multiplying Qunatity and PricePerUnit
10. Filling NaN values of PayementMethod by Unknown
11. Filling the NaN values of DiscountApplied by Mean of the ProductCategory
12. Checking for duplicates
13. Checking for outliers using the boxplot for numerical columns and retaining the data which is greater than 0
14. What was your thought process when you first saw the data?

The data consisted of many missing values. Before preprocessing the data using python I analysed the data in excel using filter feature in detail. Primary task was to understand what this sales data consisted of and how the data is distributed.

**Data Aggregation and Grouping**

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

Fields like Quantity, TotalAmount, PricePerUnit, DiscountApplied, TrustPointsUsed, TransactionDate were aggregated.

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

 Quantity (Sum): To know the total amount of items sold.

 TotalAmount (Sum): To calculate the total sales revenue.

 PricePerUnit (Mean): To find the average price of products sold.

 DiscountApplied (Sum): To quantify total discounts given.

 TrustPointsUsed (Sum): To measure the total points used in transactions.

 TransactionDate (Unique): To see the number of unique dates.

**Data Validation**

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

 Ensuring no missing values remain, and all data types are as expected.

 Verifying that outliers have been handled appropriately, such as negative values being removed.

 Comparing preprocessed data with the original for consistency.

1. How will you validate your results?

 Checking statistical summaries like mean, median, mode before and after preprocessing.

 Plotting different graphs to visually inspect the distribution of data.

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

For every step, I check of the output. And also I rely on descriptive statistics, visual inspection, and comparisons with the original data to ensure correctness.

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

 There were Missing values in critical fields.

 Incorrect data types or formats in fields such as TransactionDate.

 Negative values in non-negative fields which was totally unaccepted.

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

* Ensuring that no duplicates or incorrect entries are present.
* Verifying that all transactions have valid customer and product IDs.
* Ensuring logical consistency (e.g., TotalAmount should be non-negative).

**Sales analysis using sql**

Uploaded preprocessed csv file and completed Sales analysis using sql queries in MySQL workbench. The file name is sales.sql .