Data Preprocessing

Steps taken:

- Imported the libraries and loaded the data.
- Changed the data type of the date & time column.
- Dealt with the duplicate value in the TransactionID column.
- Handled missing values with mean and median.
- Validated data with descriptive statistics.
- Analysed the data.
- Visualised the data.

What was your thought process when you first saw the data?

I tried understanding the structure of the data, noting down the issues or inconsistencies that were clearly visible to my eye. I followed up with changing the data types of the required columns and handling missing values, to ensure my data is ready for analysis.

Data Aggregation and Grouping

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

   Fields for Aggregation:

   - Quantity (Agg function)
   - PricePerUnit (Agg function)
   - TotalAmount (Agg function)
   - DiscountApplied (Agg function)
   - TransactionDate (Agg method)
   - CustomerID (Agg method)
   - ProductID (Agg method)

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

   Quantity:

- **Aggregation:** Sum
- **Reason:** It can provide insights into total quantity sold per customer or per product category.

   PricePerUnit:

- **Aggregation:** Average, Sum

- **Reason:** It can provide insight into average price per unit per product, while total revenue can be derived from the sum.

TotalAmount:

- **Aggregation:** Average, Sum
- **Reason:** Total revenue from sales and average transaction amount.

TransactionDate:

- **Aggregation:** Group by week/month/year for time series analysis.
- **Reason:** It can reveal trends in the data.

DiscountApplied:

- **Aggregation:** Sum
- **Reason:** It can provide details of total discount availed/received per customer/ per category.

CustomerID & ProductID:

- **Aggregation:** Group by aggregation method
- **Reason**: It can be used to find data specific to each customer and product.

Data Validation

1. **How do you know, your preprocessing was correct?**
    - I review to ensure that the data before and after preprocessing to identify any inconsistencies, missing values, outliers, and incorrect data types.
    - Use descriptive statistics and compare these values before and after preprocessing to ensure that the data's integrity has been preserved.

2. **How will you validate your results?**
    - Cross-Verification with Original Data
    - Handling edge cases
    - Run a consistency check

3. **Do you follow any specific validation process for all questions? Explain.**
    - Understanding the dataset- I try to understand the structure of the data, what each variable means, what do the values imply(for example, in

this dataset minus before an integer means returned products) and what changes do I need to make.

- Data quality checks - I check and change the data types if required(for example, in this dataset, CustomerID and TransactionDate, both their data type has been changed), I try to fill in the missing values with mean, median or mode, replace or drop the duplicates, and others.
- Document the process- I keep documenting what changes I am doing and my findings, so that if there’s anything I want to revert or try to find any error on my part, I can quickly glance at the steps taken by me.

4. **What are the edge cases you can think of?**

- Missing Values
- Incorrect Data
- Outliers
- Data Imbalances

5. **What data integrity points do you want to mention for the given scenario?**

- Each TransactionID is unique.
- Consistency in calculations.
- Fill in the missing values as required.
- The TransactionDate column is of the correct data type.

Data Visualisations

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

- Sales trend Analysis
- Total Amount spent on each product category
- Total Amount spent by each customer
- Total Quantity sold in each product category
- Total Quantity bought by each customer
- Payment method distribution

2. **How would it be known if the data is linearly projected?**

Two of the methods are:

- Statistical test: correlation coefficient to measure the linearity between variables.
- Line Plot that draws a linear regression model along with a scatterplot.

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

    - Bar charts
    - Pie Charts
    - Scatter plots
    - Line charts
    - Box plots
    - Histogram
    - Heatmap