Name :- Pranjal Dixit

E-mail :- dixitpranjal77@gmail.com

Contact No. :- 7014600035

Data Preprocessing :-

Steps followed for pre-processing -

- Firstly, I removed all the blank cells for the columns - 'CustomerID', 'TransactionDate', 'PricePerUnit', 'TotalAmount' ; As we don't have any knowledge about these. So, there is no meaning of keeping them in the dataset.
- Then I changed the type of TransactionDate to 'DateFormat'.
- Now, there is a mismatch between productID and productCategory column as we have different productID for the same product Category, thus we need to update this 'ProductID' column.
- Handling the negative quantity values. I think this is a human error, where the quantities are mistakenly entered negative. Or negative quantity values can also refer to product being returned.
- Removing column where Quantity values are 0, as there is no transaction for that particular customer.
- Recalculating TotalAmount column as it might contain negative values because of negative quantities.
- Updating Empty cells of Discount columns to 0.
- Updating empty paymentType cells to 'Trust Points' if 'TrustPointsUsed' is not equal to 0, otherwise updating it to 'Credit Card/Cash'.
- Insert a new column to find out the total amount to be paid by the customer.

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

=>

- I first looked at the column names and their descriptions to understand the type of data and the expected values.
- Then, identified potential issues like missing values, negative values, or incorrect data types such as missing Customer IDs, missing and incorrectly formatted dates, negative values in Quantity and

TrustPointsUsed, and missing values in PricePerUnit and PaymentMethod..

- Then I planned to address missing values, correct data types, handle outliers, and ensure data consistency and started performing Data Pre-processing.

Data Aggregation and Grouping :-

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

=>

- Total number of Transactions - 1
- Number of distinct Customers - 2
- Total number of purchases by each Customer - 3
- Total number of unique products - 4
- Product sales by each category - 5
- Total quantities sold - 6
- Total quantities of each product category type - 7
- Total Trust points used by each customer - 8
- Amount paid by each customer - 9
- Max, Min and Average discount applied - 10
- Payment Method Transactions Grouping - 11
- Calculate the percentage for each payment type - 12
- Number of transactions for each day - 13
- **Create time frames groups and Calculate the total sales - 14**
- **Recency, Frequency and Monetary (RFM) Analysis - 15**

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

=>

- TransactionID - Count (provide insights into the volume of transactions.)
- CustomerID - Count, Unique Count (Counting unique CustomerID values can help determine the number of distinct customers.)
- TransactionDate - Count (number of transactions per day, month) (helps in analyzing time series data analysis.)

- ProductID - Count by product ID(Counting transactions per product can tell which products are more frequently purchased.)
- ProductCategory - Count by product Category(Counting transactions per product can tell which products are more frequently purchased.)
- Quantity - Sum, Average (Summing the Quantity gives the total quantity sold, while averaging can show what is the average purchase quantity.). We can even group the quantity by product type to check how many quantities of each product are being sold.
- PricePerUnit - Average, Minimum, Maximum (Averaging gives the typical price per unit, while minimum and maximum show price range.)
- TotalAmount - Sum, Average, Maximum, Minimum (Summing the TotalAmount shows total revenue, while averaging gives the average transaction amount.)
- TrustPointsUsed - Group (grouping by CustomerID can provide how many trust points were used by each Customer)
- PaymentMethod - Count, Percentage (Counting occurrences of each payment method helps understand payment preferences. Calculating percentages can show the proportion of each method.)
- DiscountApplied - Average, Maximum, Minimum (averaging shows the average discount per transaction.)

Data Validation :-

How do you know, your preprocessing was correct?

- By ensuring 'TotalAmount' column is correctly calculated as 'Quantity' * 'PricePerUnit'.
- By validating data type for each column like Quantity should be integers, TransactionDate should be in datetime format.
- By ensuring that missing or 'NaN' values have been handled appropriately like filling in with 0's or removal of those rows as needed.
- By checking discounts are applied correctly, and negative values are handled properly.

How will you validate your results?

- By Comparing summary statistics (like means, medians, ranges) before and after preprocessing.

- Verify that TotalAmount is non-negative and correctly reflects Quantity and PricePerUnit.
- Review a subset of records to ensure the preprocessing logic is good.
- Ensuring columns expected to have unique values do not have duplicates.
- Ensuring there are no negative quantities or prices.

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

- Define what the expected results should be for each preprocessing step.
- Validate data integrity, consistency, and accuracy.

What are the edge cases you can think of?

- By handling of NaN, blank cells, or unexpected values in columns like CustomerID, TransactionDate, PricePerUnit
- Handling negative values in various columns to ensure they are correctly handled or flagged.
- Errors in date formats in TransactionDate that may affect parsing as in the dataset, it is a text object.

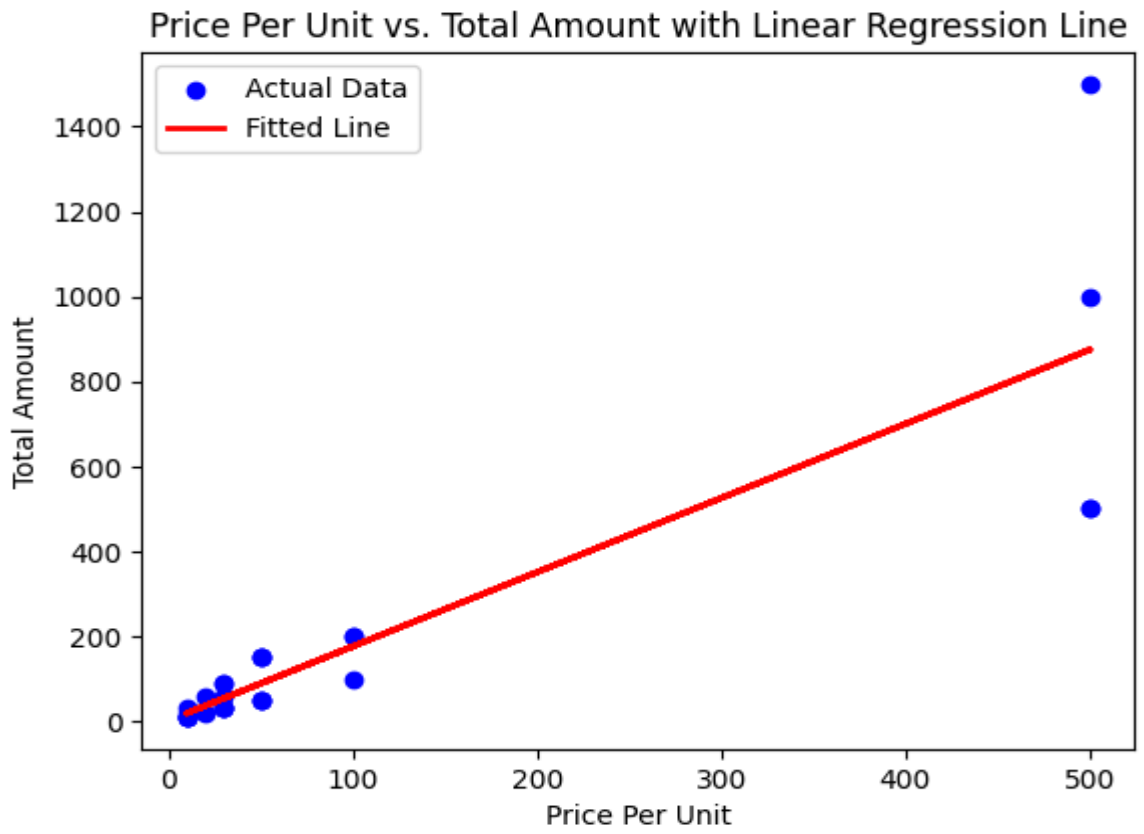
What all data integrity points you want to mention for the given scenario?

- All calculated fields are accurate based on the raw data.
- Data is consistently formatted and follows the rules
- Verify that all fields are populated and missing values are handled appropriately.
- Unique identifiers do not have duplicates.

Data Visualization :-

How would be know if the data is linearly projected?

To determine if the data is linearly projected, we will perform regression analysis. We can use PricePerUnit as the independent variable (X) and TotalAmount as the dependent variable (Y) to check for a linear relationship between these two variables. This is the result which we got after performing regression analysis.



What all projections are possible out of the data?

- Transactions Over Time (Line Chart)
- Sales over TimeFrames (Line Chart)
- Sales by Product Category (Bar plot)
- Sales by Payment Method (Bar plot)
- Number of Transactions per Hour (Bar plot using seaborn library)
- Relationship between Discount Applied and Amount to be paid (Scatter plot)
- Relationship between Trust Points Used and Amount to be paid (Scatter plot)
- Distribution of Payment Method (Pie chart)
- Market Share of Product Categories (Pie chart)
- Distribution of Amount Paid (Histogram)
- Frequency of Discount Applied (Histogram)
- Correlation between different Numeric Variables (Heatmap)

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

=>

- **Line Chart**
- **Bar Plot**

- **Scatter Plot**
- **Histogram**
- **Pie Chart**
- **HeatMap**