EDA

1. df.info() and df.describe() insight

- All columns have 1000 non-null values, suggesting no missing data.

- Quantity - Most transactions involved purchasing between 2 and 4 items, with 3 being the median quantity.

- TotalValue - The TotalValue varies widely, as shown by the large standard deviation (~493.14).

- Price x and price y - The mean price is ~272.55,and the median is 299.93, indicating a slightly right-skewed price distribution where most products are priced around $300 but there are some lower-priced products

- df.isna() there were no missing value

2.Checked Outliers -

There were no outliers either in the price and Total value, checked with the help of box plot and Interquartile Range (IQR) method.

3.Perform feature engineering create TotalSpend and Purchase Frequency by grouping CustomerID and aggregating TotalValue and TransactionID Respectively.

- Also add TotalQuantity and MostPurchasedCategory, To know the quantity and most purchase item by the customer to better understand the customer behaviour and most selling item, so recommendation system can be improve.

- Books are the most purchased item, with 86.3% of customers buying books more than any other category.

- **Category**

  Books 270

Electronics 254
Home Decor 248
Clothing 228

- Create one more feature SignupDate, logic - If the signup date is within the last year (current_date - 365 days), the customer is classified as New otherwise as Returning.

- Analyze TransactionDate and Recency (Days) to understand the majority of customer behavior..
-

4.Data Visualisation

- Scatter plot(Recency vs Total Spend) - Returning customers seem to be spread across the full range of "Recency" and "Total Spend".

- Bar plot(Average Total Spend by Region) - There were no major difference

- Bar plot Total Spend by Category- No major differences, but Books are most purchased category

- Pie chart(Customer Type Distribution) - Returning customer = 64.2
  New customer = 35.8. *Main aim is to convert new customers into returning customers.

- Scatter plot(Bubble) Recency vs Purchase Frequency- how recently a purchase was made) and Purchase Frequency (how often purchases occur)

- Horizontal Bar - between customer name and TotalSpend. (Paul Parsons over 100000.

5.Splitted data into categorical and numeric and applied One Hot Encoding for categorical features and Standard Scaling for numeric features.