ZEOTAP ASSIGNMENT

TASK-1

Data pre-processing for the merged dataset (Customer,Product,Transaction)

1. Handling Missing Values: The merged dataset may contain missing values in any of the columns (e.g., `TotalValue`, `Price`, or customer details). Identifying and appropriately handling these missing values is crucial to ensure accurate analysis.

2. Data Type Conversion: The `TransactionDate` and `SignupDate` columns should be converted to datetime objects for time-series analysis. Ensuring that numerical columns (like `Price` and `TotalValue`) are in the correct format is also essential.

3. Outlier Detection: Identifying and handling outliers in numerical fields can prevent skewed results during analysis. For instance, unusually high prices or transaction values could distort insights.

4. Normalization/Standardization: Depending on the analysis, normalizing or standardizing certain features may be necessary, especially if machine learning models are to be applied later.

5. Categorical Encoding: Categorical variables such as `Category` or `Region` may need to be encoded into numerical formats for certain analyses or modeling techniques.

6. Date Feature Engineering: Extracting additional features from dates (like year, month, day of the week) can provide more granular insights into trends and patterns.

Analysis Interpretation

- Contingency Table: This shows how many customers from each region purchased products from each category.
- Chi-Squared Test: A low p-value (typically < 0.05) indicates a significant association between product categories and customer regions.
- Heatmap: This visual representation helps identify patterns in how different regions engage with various product categories.
- Scatter Plot: This will help you visually assess whether there is a linear relationship between product prices and customer purchases. A positive correlation would show points trending upwards.
- Correlation Matrix: The values in the matrix will indicate how strongly related the two variables are (values closer to 1 or -1 indicate strong correlations)

Data Overview

- Products: Info on product IDs, names, categories, and prices.
- Customers: Info on customer IDs, names, regions, and signup dates.
- Transactions: Info on transaction IDs, customer IDs, product IDs, dates, quantities, total values, and prices.

Key Analyses

1. Price vs. Total Value:
   - Correlate product prices with transaction totals (Price × Quantity).
   - Higher prices typically lead to higher transaction values, especially in bulk purchases.
2. Region vs. Product Categories:
   - Analyze which product categories are most popular in different regions.
   - Helps identify regional preferences (e.g., Electronics popular in South America).

Distribution Analysis

1. Product Price Distribution:
   - Use histograms/box plots to understand price ranges and outliers.
2. Customer Signup Dates:
   - Analyze signup trends over time to spot peaks, possibly linked to marketing campaigns.

Five key business insights:

1. Product Category Popularity:
The demand for products is evenly distributed across categories, with Books, Electronics, and Clothing showing slightly higher sales than Home Decor. This suggests a balanced interest in these categories, with room to promote Home Decor products.

2. Regional Customer Distribution:
South America leads in customer representation, followed by Europe, North America, and Asia. Marketing strategies should prioritize South America while exploring growth opportunities in underrepresented regions like Asia.

3. Seasonal Sales Trends:
Monthly sales data indicates significant fluctuations, with peaks around mid-year (June-August) and a decline towards the end of the year (October-November). Seasonal campaigns can be planned to capitalize on high-sales months.

4. High-Value Products Driving Revenue:
Products like smartwatches and high-priced books contribute significantly to revenue. Focusing on these high-value items through targeted promotions could further boost profitability.

5. Customer Retention Opportunities:
A large portion of customers joined recently (2023-2024), indicating potential for retention-focused strategies such as loyalty programs or personalized offers to maintain engagement and drive repeat purchases.