Customer Analysis - Project Deliverable 1

Group Members: 21L-5306, 21L-5482, 21L-____

Part 1: Statistical Analysis

- Used .info() and .describe() to summarize the dataset.
- Computed descriptive statistics including mean, median, standard deviation, quartiles, and interquartile range (IQR) for all numerical features.
- Counted unique values and frequency distributions for all categorical variables.

Part 2: Data Wrangling, Transformation, and Visualization

Data Cleaning

- Missing values were found only in the 'Income' column. As the missing data represented less than 2% of the dataset and showed no patterns, those rows were dropped.
- No duplicated rows were found in the dataset.

Data Transformation

- Converted 'Dt_Customer' to datetime format and derived 'Customer_Tenure' and 'Customer_Year'.
- Converted 'Education' and 'Marital_Status' columns to categorical types for efficient analysis.

Feature Engineering

- Created additional columns for analytical purposes:
  - Customer_Tenure
  - Total_Spending
  - Spending_Ratio
  - Engagement_Count
  - Discount_Sensitivity
  - Family_Size

Univariate Visual Analysis and Findings

- Income: Right-skewed distribution showing that most customers earn below $100,000.

- Year_Birth: Majority born between 1950 and 1980, indicating an older customer base.
- Product Spending: Spending is highest on wines and lowest on fish/sweet products.
- Campaign Responses: Very low acceptance rates across all campaigns.
- Recency: Spread evenly, implying varied engagement recency among customers.
- Web Visits: Most customers visit 0-2 times per month.
- Complaints: Almost all customers had no complaints, showing possible satisfaction or disinterest in reporting.

Categorical Visual Analysis and Findings

- Education: Majority of customers are graduates or hold a PhD/Master's degree.
- Marital Status: Most customers are married or living together. Some categories like 'YOLO' and 'Absurd' are minimal and may need grouping.

Bivariate Analysis

- Correlation Heatmap: Showed moderate positive relationships among purchase-related features.
- Income vs Education: Boxplot showed that higher education levels correlate with higher income.
- Marital Status vs Education: Count plot revealed diverse education distribution across marital statuses.

Normalization and Standardization

- Normalized spending-related features using MinMaxScaler to bring them within [0, 1] range.
- Standardized visit and engagement-related features using StandardScaler for normality.
- Applied log1p transformation on 'Income' and 'Total_Spending' to handle skewness and zero values.

Dimensionality Reduction

- Removed features with near-zero variance: 'Z_CostContact' and 'Z_Revenue'.
- Label encoded all categorical features before applying PCA.
- PCA was used to reduce dimensionality while retaining 95% of explained variance, resulting in 22 principal components.
- Exported the cleaned dataset before PCA and the final reduced dataset for modeling.

Data Validation

- Confirmed completeness by handling missing data appropriately.
- Ensured consistency by converting data types and handling invalid category entries.
- Verified accuracy through manual checks on statistical summaries and visual validation.

Deliverables Summary

- Final Cleaned Dataset (before PCA): cleaned_before_pca.csv
- Final PCA-Reduced Dataset: cleaned_processed_data.csv
- Python Script: project_dev_1.py
- Analysis Report: (this document)