**Marketing Mix Project Report – AI & ML Bootcamp**

**Objective**

The goal of this project was to perform data preprocessing, exploratory data analysis (EDA), and statistical hypothesis testing to understand factors that influence customer behavior in a marketing dataset. The analysis aligned with the four Ps of marketing—Product, Price, Place, Promotion—with an added focus on "People."

**Key Steps and Learnings**

**1. Data Cleaning and Preparation**

* Verified data importation of critical variables like Dt\_Customer and Income.
* Cleaned the Income column by removing $ and ,, and converted it to numeric.
* Converted Dt\_Customer to datetime format for proper temporal analysis.
* Imputed missing Income values using group-wise averages by Education and Marital\_Status.

**Learned:** How to handle missing values using domain-relevant groupings and how to clean and convert string-formatted numerical data.

**2. Feature Engineering**

* Created new variables:
  + Total\_Children = Kidhome + Teenhome
  + Age = Current Year - Year\_Birth
  + Total\_Spending = Sum of all product expenditure columns
  + Total\_Purchases = Sum of all purchase channels (web, catalog, store)

**Learned:** How to create derived features that can add analytical value for behavioral segmentation.

**3. Outlier Detection and Treatment**

* Used boxplots and histograms to inspect the distribution of key variables.
* Applied IQR (Interquartile Range) method to cap outliers.

**IQR Calculations and Formulas:**

To detect and treat outliers, we applied the following formulas:

* **Q1 (First Quartile):** 25th percentile of the data
* **Q3 (Third Quartile):** 75th percentile of the data
* **IQR (Interquartile Range):** IQR = Q3 - Q1
* **Lower Bound:** Lower = Q1 - 1.5 \* IQR
* **Upper Bound:** Upper = Q3 + 1.5 \* IQR
* Any value outside the lower or upper bound was capped using:

df[column] = df[column].clip(lower=Lower, upper=Upper)

**Learned:** Best practices for identifying and handling outliers using visual and statistical methods.

**4. Encoding Categorical Variables**

* Applied **Ordinal Encoding** to Education using a logical order.
* Applied **One-Hot Encoding** to nominal variables like Marital\_Status and Country.

**Learned:** How to prepare categorical variables for modeling while preserving semantic meaning.

**5. Correlation Analysis**

* Generated a correlation heatmap to observe relationships between numerical features.
* Applied masking and layout improvements for readability.

**Learned:** How to use visual correlation matrices to guide feature selection.

**6. Hypothesis Testing**

Performed t-tests and correlation analysis to validate the following:

* Older customers prefer in-store shopping.
* Parents are more likely to shop online.
* Sales in one channel may cannibalize others.
* U.S. customers have higher total purchase volumes than other countries.

**Learned:** How to apply statistical testing to support or reject marketing-related hypotheses.

**7. Visual Insight Generation**

Used Seaborn and Matplotlib to explore:

* Top/bottom revenue-generating products
* Age vs campaign response
* Country-level campaign performance
* Children vs total spending
* Complaints by education level

**Learned:** How to select the right chart types and interpret plots for actionable insights.

**Potential Interview Questions and Sample Responses**

**Conceptual:**

1. **How did you handle missing values in the dataset?**
   * I used group-based imputation. Specifically, I imputed missing Income values using the average income for each combination of Education and Marital\_Status, which helped maintain logical consistency across segments.
2. **Why did you choose IQR for outlier treatment?**
   * The IQR method is robust to extreme values and doesn’t assume normality. It’s particularly effective for identifying and treating outliers without distorting the distribution as much as z-score methods might.
3. **How did you determine which features to engineer?**
   * I created features that aligned with real-world customer behavior insights, such as total number of children, age, and total spending. These help represent consumer lifecycle, household size, and value.
4. **What insights did you gather from the correlation heatmap?**
   * It revealed strong positive correlations between product expenditures, and a moderate negative correlation between the number of children and total spending.
5. **Why did you use both ordinal and one-hot encoding?**
   * Education is inherently ordered, so ordinal encoding preserved that. In contrast, variables like Country and Marital\_Status are nominal, so one-hot encoding was more appropriate to avoid implying any order.

**Analytical Thinking:**

1. **How would you validate whether online sales are cannibalizing store sales?**
   * I calculated the Pearson correlation between store purchases and the sum of online channels (web + catalog). A strong negative correlation would suggest potential cannibalization.
2. **If given more data, how would you predict customer lifetime value?**
   * I’d build a regression model using features like age, income, total spending, frequency of purchases, and channel preference. Time-based RFM metrics could also be introduced.
3. **Which feature(s) would you select for a predictive model and why?**
   * I’d prioritize Age, Income, Total\_Spending, and Response, along with encoded categorical variables. These features are directly related to purchase power, behavioral trends, and campaign engagement.

**Visualization:**

1. **What plot types did you use to identify top-performing products?**
   * I used bar plots to rank product categories by total revenue. This provided a clear visual of which categories were driving sales.
2. **How did you ensure your visualizations were interpretable?**

* I used consistent color palettes, rotated x-axis labels where needed, and added titles, axis labels, and tight layouts to avoid crowding.

**Business Insight:**

1. **What marketing strategies would you recommend based on your findings?**

* Target younger customers through online campaigns, emphasize product categories with high margins, and personalize offers for parents who prefer online shopping for convenience.

1. **Were there any surprising patterns in customer behavior?**

* Yes, there was a negative correlation between the number of children and total spending, suggesting that households with more children may be more budget-conscious or constrained by other expenses.

**Final Thoughts**

This project strengthened my skills in data cleaning, feature engineering, statistical analysis, and visualization. It also helped me understand how to extract business insights from data, an essential competency for any data science or ML role.