IDS 400
Team 5 Group Project

Improving Marketing Campaign Efficiency

Members:
Christopher Hendro
Preet Godhani
Rama Hammad
Phu Huynh
Sheida Hassani

1) Introduction:

In this project, we aim to improve the marketing campaign efficiency of a business based on a Marketing Campaign dataset found on Kaggle. We will develop predictive models using Python and showcase various data visualizations to analyze the customer characteristics in the supplied dataset. Our end goal is to identify and precisely forecast the subscription behavior of the customers in this given business' marketing campaign. In other words, to predict who will respond to an offer for a product or service. By analyzing the characteristics of the customers in this dataset, we hope to identify the business's ideal customers. This will help the business better understand its customers to modify its product offerings according to the specific needs, behaviors, and concerns of its customers. Furthermore, our analysis can help this business decide if they could tailor specific campaigns and promotions to their ideal customers to increase this business' marketing campaign efficiency and boost the response rate of future marketing campaigns.

2) Research Questions/Problems:

- What are the key elements influencing customer spending and store visits?
- How can demographic characteristics such as age and family status influence customer behavior?
- Can promotional responses predict consumer spending and frequency of purchases?

3) Motivation:

In the analysis of the supplied dataset, we will present various data visualizations to provide a comprehensive understanding of the customer characteristics and behaviors within the dataset. For Exploratory Data Analysis (EDA), we will use visualizations such as histograms, scatter plots, etc. to explore the distribution of variables, identify outliers, and understand the relationship between different variables. A correlation heatmap will also be used to identify the correlation between the different variables in the dataset. On top of the data visualizations, we will use machine learning (ML) models such as *logistic regression, decision tree,* and *support vector machine* (SVM) to build the predictive models. These models will help us understand the information contained within the dataset to accurately predict the response rate of the customers in this business' marketing campaign. The combination of the machine learning models, as well as the data visualizations will aim to provide actionable insights that empower the business to make informed decisions regarding its marketing strategies, customer targeting, and overall marketing campaign efficiency.

4) Data Preparation and Cleaning:

The dataset we obtained for this analysis is obtained from Kaggle. It included **2240** records and **29** variables (see appendix for all variables). However, we used **11** variables to show the basic information for the data:

- Current_Age: Customer's age
- Family_Size: Family size of the customer
- Parent_Checking: Check if the customer is a parent

- Groups: Grouping customers by income and spending
- Income: Customer's yearly household income
- Spent: Total Spending on products
- AcceptedCmp1: if the customer accepted the offer in the 1st campaign
- AcceptedCmp2: if the customer accepted the offer in the 2nd campaign
- AcceptedCmp3: if the customer accepted the offer in the 3rd campaign
- AcceptedCmp4: if the customer accepted the offer in the 4th campaign
- AcceptedCmp5: if the customer accepted the offer in the 5th campaign

	Current_Age	Family_Size	Parent_Checking	Groups	Income	Spent	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5
count	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000
mean	55.179603	2.592509	0.714350	1.527527	52247.251354	607.075361	0.064079	0.013538	0.073556	0.074007	0.073105
std	11.985554	0.905722	0.451825	1.157237	25173.076661	602.900476	0.244950	0.115588	0.261106	0.261842	0.260367
min	28.000000	1.000000	0.000000	0.000000	1730.000000	5.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.000000	2.000000	0.000000	0.000000	35303.000000	69.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	54.000000	3.000000	1.000000	2.000000	51381.500000	396.500000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	65.000000	3.000000	1.000000	3.000000	68522.000000	1048.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	131.000000	5.000000	1.000000	3.000000	666666.000000	2525.000000	1.000000	1.000000	1.000000	1.000000	1.000000

5) Data Methods:

- **Data Overview:** We used a dataset with a variety of customer factors such as income, spending habits, demographic information, and promotional response rates.
- **Correlation Analysis:** We used a correlation matrix to determine the correlations between various consumer attributes. This study aids in determining which aspects are strongly linked to consumer expenditure and involvement.
- **Tools Used:** matplotlib, seaborn, numpy, and pandas
- **Visualization of Correlation Matrix:** By emphasizing noteworthy positive and negative correlations, the correlation heatmap offered a visual depiction of the connections between the variables.

6) Machine Learning Model Development:

In this section, we develop machine learning models to provide predictions of the target variable. Our target variable is "Response", which is a binary variable:

· 1 if the customer accepted the offer in the last campaign,
· 0 otherwise

In order to develop the models, first we split the data into 80% training dataset and 20% testing dataset. Based on the preprocessing and visualization phase(such as correlation matrix) we include the following features for developing the model:

['Income', 'Recency', 'NumDealsPurchases', 'NumWebPurchases',' NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth', 'Complain', 'Spent', 'Status_Shorten', 'Current_Age', 'Parent_Checking', 'Family_Size', 'Education_Shorten', 'Promo_Accepted_Sum', 'Response']

Then by applying the x.unique(), we found the variable we should apply dummies on.

The packages we used here include:

- sklearn.model_selection => train_test_split
- sklearn.linear_model => LogisticRegression
- sklearn.tree => DecisionTreeClassifier
- sklearn.preprocessing => StandardScaler
- sklearn.metrics => accuracy_score, precision_score, recall_score, f1_score

We developed three models including Logistic Regression, Decision Tree, and SVC.

6.1) Logistic Regression

The logistic regression model was implemented to predict the binary response. The model achieved an accuracy of 86.71% on the testing set and an accuracy of 86.11% for the training dataset. Also, the confusion matrix was generated for testing the dataset to evaluate the model's performance in classifying the responses accurately.

	0	**1**
0	369	13
1	46	16

Here are the evaluation metrics in more detail:

- Recall: 0.25806451612903225
- Precision: 0.5517241379310345
- F Score: 0.3516483516483517
- Accuracy: 0.8671171171171171

6.2) Decision Tree

A decision tree classifier was used as a non-linear model alternative to logistic regression. After training on the same dataset, the decision tree model demonstrated an accuracy of 82.43% on the testing set and an accuracy of 99.20% on the training dataset which shows overfitting. The confusion matrix for the testing dataset is as follows.

	0	**1**
0	336	46
1	32	30

The evaluation metrics for our decision tree model are:

- Recall: 0.4838709677419355
- Precision: 0.39473684210526316
- F Score: 0.43478260869565216
- Accuracy: 0.8243243243243243

Decision tree with less overfitting:

To reduce the overfitting in our decision tree model, we applied some strategies which include:

- Limiting the depth of the tree: limiting the depth of the tree can decrease the complexity of the tree.
- Minimum sample split: increasing the sample split can prevent our model from generating a lot of splits.
- Minimum samples leaf: This strategy can make the leaves have more than one sample and as a result, the tree is less fitted to the training data.

We applied these strategies to our decision tree model and could increase the accuracy for the testing dataset (and also decrease the accuracy for the training dataset so that both accuracies become closer to each other). The accuracy for our training dataset for our new decision tree is 89.89%, and the accuracy for our testing dataset is 87.61%.

Here is the confusion matrix for our second decision tree model:

	0	**1**
0	363	19
1	36	26

the evaluation metrics in more detail are as follows:

- Recall: 0.41935483870967744
- Precision: 0.5777777777777777
- F Score: 0.4859813084112149
- Accuracy: 0.8761261261261262

6.3) Support Vector Classifier (SVC)

The Support Vector Classifier was employed to handle potentially complex relationships in the data. The model achieved an accuracy of 86.48% on the testing dataset and an accuracy of 85.10% on the training dataset. SVCs are known for their effectiveness in high-dimensional spaces. Here is the confusion matrix for the testing dataset.

	0	**1**
0	374	8
1	52	10

The evaluation metrics for svc are:

- Recall: 0.16129032258064516
- Precision: 0.5555555555555556
- F Score: 0.25
- Accuracy: 0.8648648648648649

Based on the accuracy of the models, we see that our new decision tree model, the one with less overfitting, has the highest performance in comparison with the other models. After that, the logistic regression model performed slightly better than the SVC and considerably better than the basic decision tree model(the one with overfitting). Our final ranking for the model is as follows.

1. Accuracy of our new Decision Tree model(with less overfitting): 87.61%
2. Accuracy of logistic regression: 86.71%
3. Accuracy of SVC: 86.48%
4. Accuracy of basic Decision tree(the one with overfitting): 82.43%

7) Data Result:

7.1) Correlation between each variable

a) Graph Description: A correlation heatmap showing the correlation between the different variables across within the dataset

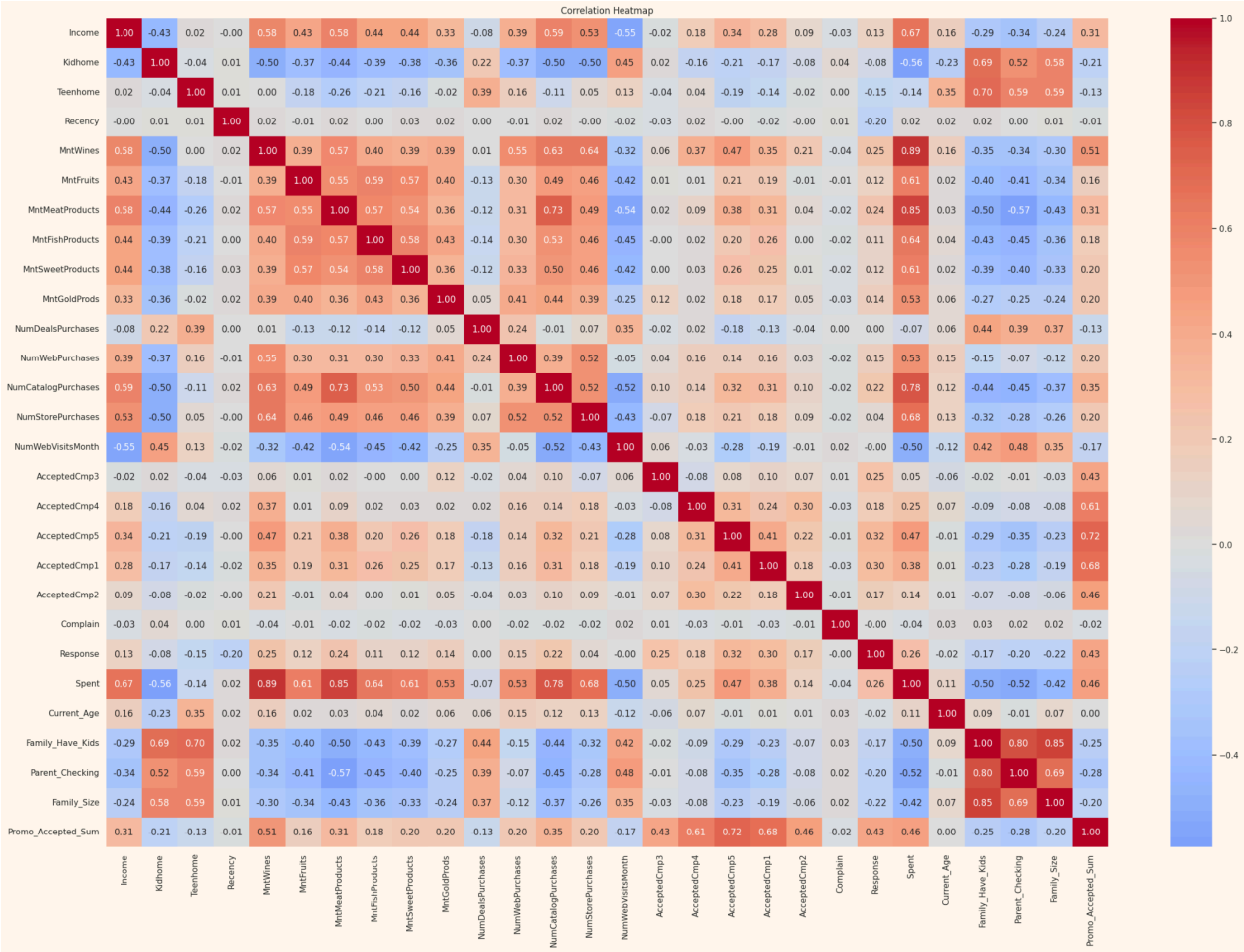
b) Analysis and Insight:

- Spending and child in household relationship: Initial analysis of the correlation between spending variable and KidHome/TeenHome variables indicate a negative correlation, suggesting customers with children in the household spend less relative to customers without children.

- Spending and products sold relationship: Initial analysis shows a high correlation between the spending variable and the MntWines/MeatProducts variable. Indicating a high amount of customer spending on Wine and Meat products.

- Responses and child in household relationship: Initial analysis shows a slight negative correlation between the response variable and the KidHome/TeenHome variable. Suggesting that customers with children responded less to the most recent marketing campaign relative to customers without children.

c) Conclusion of the result:
The visual representation of the correlation between the different variables contained within the dataset provides us with some basic initial insights into the relationship between the different variables. Gives us a basic understanding of the customer behavior towards the marketing campaign based on certain customer characteristics. This will be useful for businesses to understand how responses to previous marketing campaigns have been influenced by customer behavior, allowing businesses to shape future campaigns to better align with the preferences and behaviors of different customer segments.



7.2) General Income and Spending Analysis:
a) Graph Description: A scatter plot showing the link between income and spending across the dataset, with a trend line to indicate the overall trend.

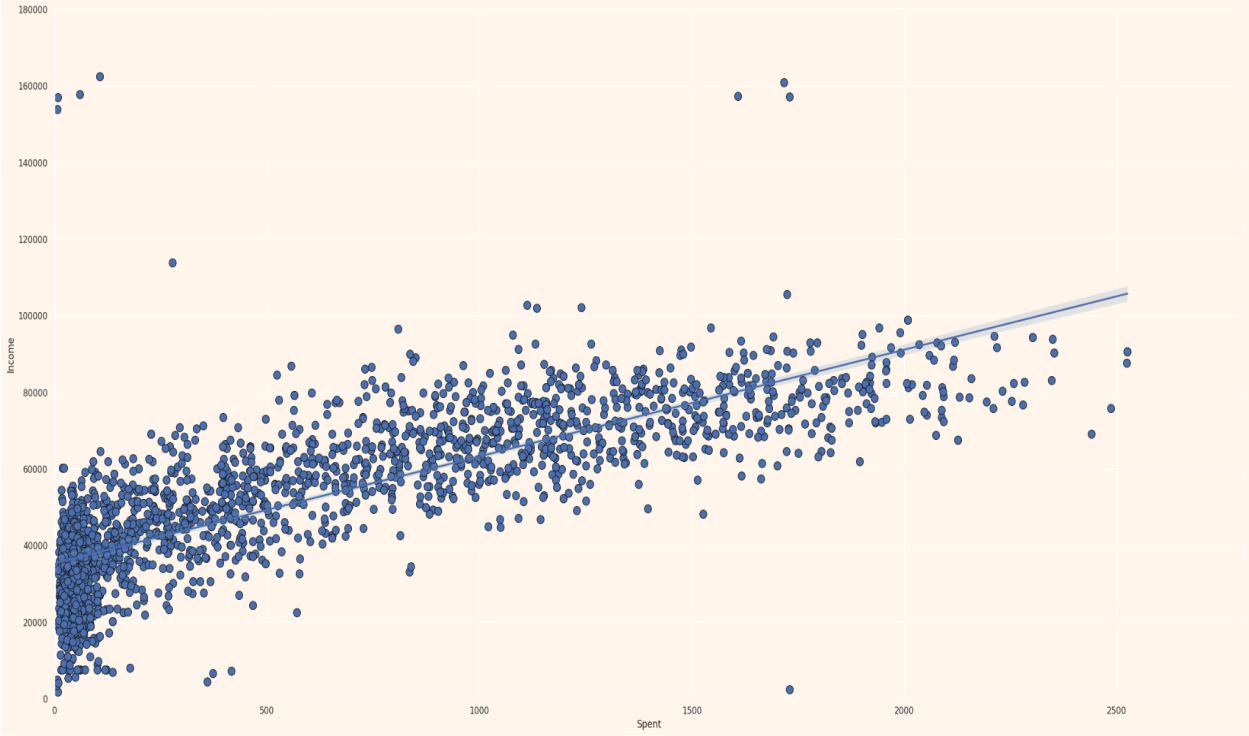
b) Analysis and Insight:

- Income and Spending Relationship: The trend line indicates a favorable association between income and spending. However, the data points indicate that there is tremendous variation in how much people spend relative to their income.

- Marketing Implications: This information is critical for understanding the elasticity of spending in relation to income fluctuations, which can assist build price-sensitive marketing strategies.

c) Conclusion of the result:

These visual analyses indicate diverse consumer habits within each group and provide a broader understanding of spending patterns based on income. This dual perspective is useful for designing both wide and focused marketing strategies, ensuring that promotional efforts are tailored to the unique needs of various consumer categories.



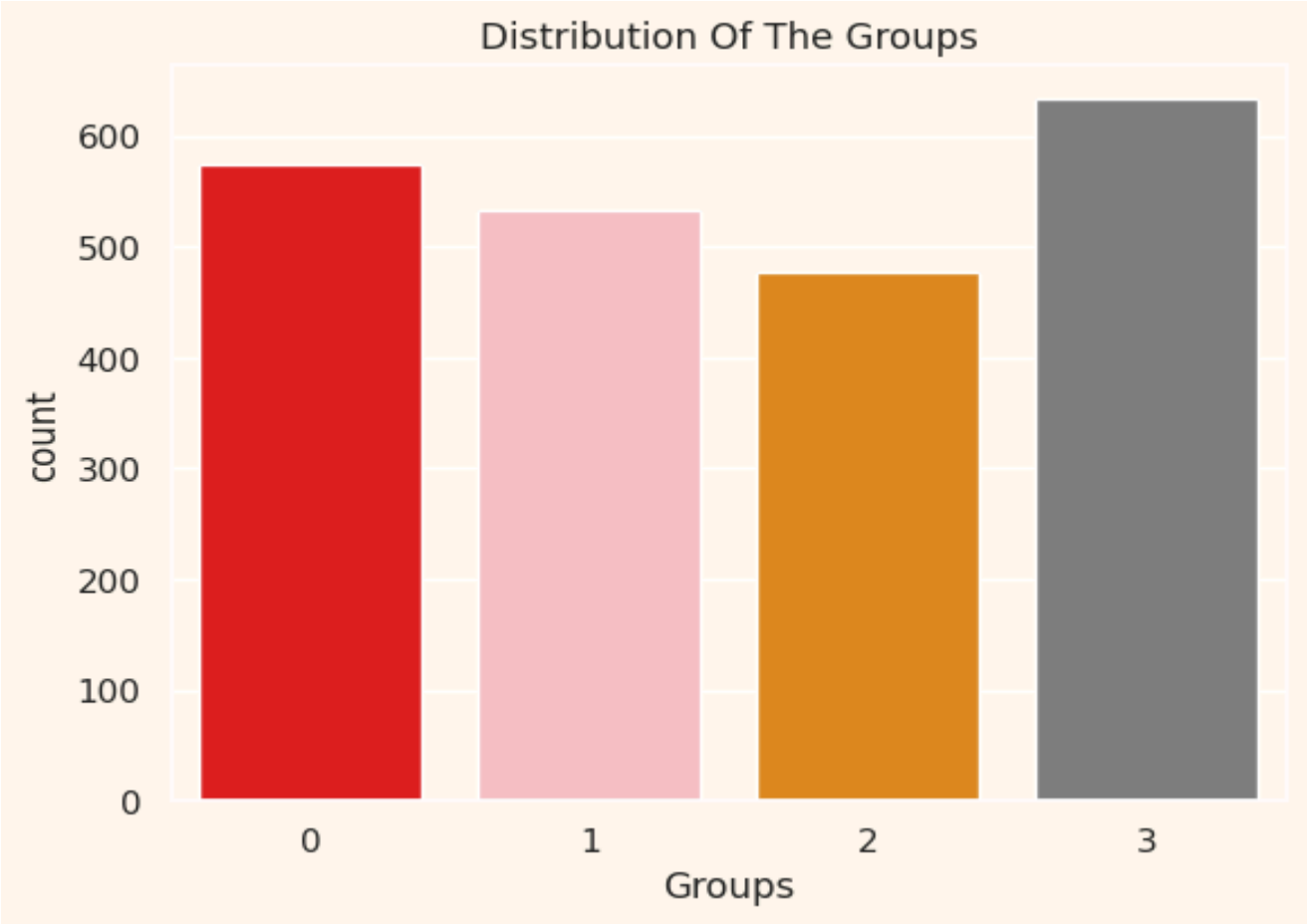
7.3) Distribution of 4 groups

a) Graph Description: This bar chart shows distribution counts for four consumer categories from the dataset:

- Group 0: Low Spending & Low Income (red)
- Group 1: High Spending & Average Income (pink)
- Group 2: High Spending & High Income (orange)
- Group 3: High Spending & Low Income (gray)

b) Analysis and insights:

- Group Distribution: The graphic depicts a fairly balanced distribution among the four groups, with Group 0 slightly bigger than the others. This implies a solid mix of consumers from various segments.
- Implications: The balanced distribution is suitable for conducting comparison analysis across groups to better understand different behaviors and preferences, which is important for targeted marketing tactics.



7.4) Specific Income and Spending Analysis (4 groups of family income & spending)

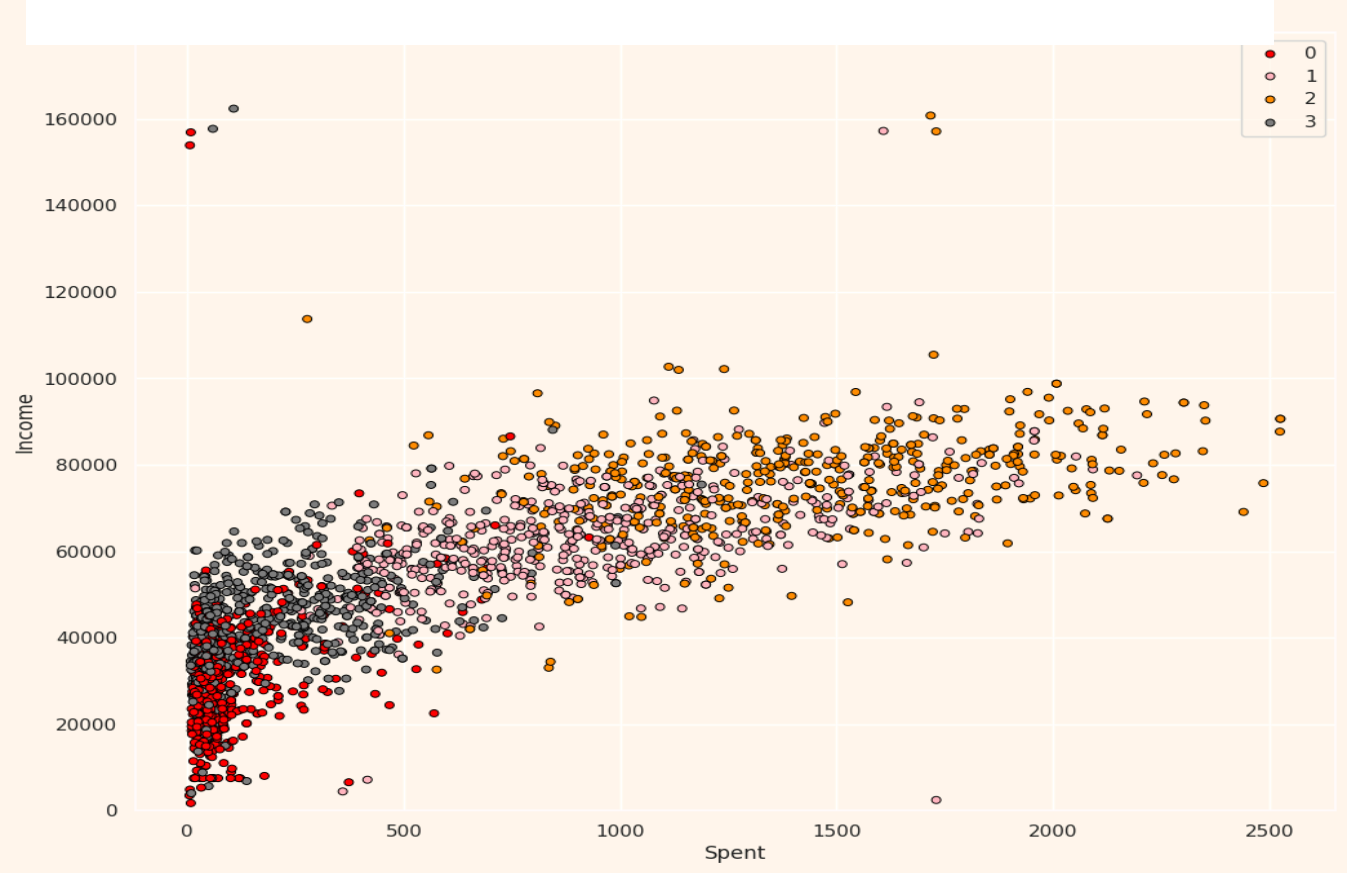
a) Graph Description: this scatterplot clusters consumers depending on their income and spending levels, color-coded by the group.

b) Analysis and insights:

- Behavioral Patterns: The plot demonstrates that increasing income does not always translate to higher expenditure, as seen in Groups 1 and 2.

- Targeted Marketing: Group 0, which often has a lower income and spends less, may require different marketing methods than Group 3, which comprises higher earners with a wider range of expenditure.

- Strategic Insights: Understanding these spending habits allows you to design financial solutions or marketing campaigns to each group's economic profile.



7.5) Campaign Promotion Acceptance:

a) Analysis and insights:

- Group 0 has the highest involvement across all promotion categories, particularly where no promotions were accepted, but it still represents a significant proportion of people accepting one or more incentives. This group could be described as "promotion-sensitive," with varying levels of participation depending on the offer.

- Groups 1 and 2 show moderate levels of promotion acceptance. Notably, Group 1 is slightly more engaged in accepting a single promotion than Group 2. Both groups show little enthusiasm for taking three or four promotions, which could reflect a moderate

interest in promotions or selective preferences.

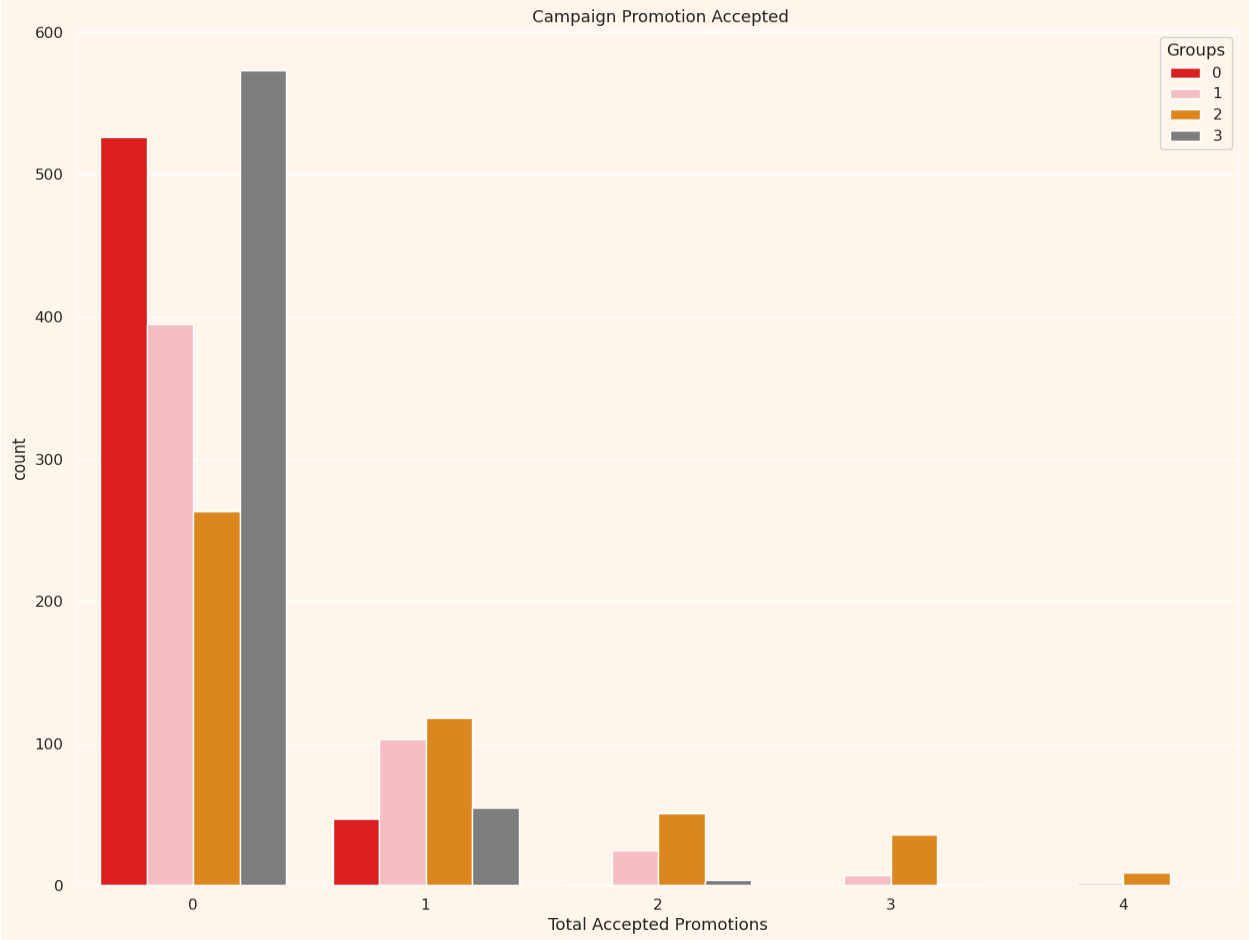
- Group 3, while having fewer overall numbers in the sample, has extremely low levels of promotion acceptance, notably in terms of accepted promotions. This group might be described as "promotion-resistant," choosing not to participate in promotional offers.

b) Strategic Implications:

- Marketing methods could be modified to boost interaction in Groups 0 and 1, which already show an interest in promotions. Special care should be used to create offers that are compelling enough to convert single promotion acceptors into multiple promotion acceptors.

- Group 3 may require a different approach, such as increasing the relevancy of the offerings or employing alternative marketing tactics in addition to standard promotions.

- Understanding the attributes that identify each group (demographics, buying history, preferences) may improve targeted marketing efforts by ensuring that promotions are not only visible but also relevant and enticing to each segment.



c) Conclusion of the result:

Analysis of promotion acceptance across groups offers insights into customer behavior during promotional efforts. It aids in determining which groups are more likely to engage in marketing initiatives, allowing for more efficient deployment of marketing resources and individualized campaign methods.

8) Business Insights:

The analysis conducted on the marketing campaign dataset from Kaggle has provided actionable business insights that can significantly enhance marketing strategies. The correlation analysis between various consumer attributes has revealed which factors are strongly linked to consumer spending and engagement. This understanding is crucial for tailoring marketing approaches that resonate with specific customer demographics. For example, recognizing that income levels have a strong association with spending behaviors allows for the development of price-sensitive marketing strategies. Furthermore, the segmentation of customers into four distinct groups based on income and spending habits enables targeted marketing tactics. Each group exhibits unique characteristics and preferences, which suggests that a one-size-fits-all approach may not be as effective. By understanding these differences, the business can tailor its promotional efforts to meet the specific needs of each segment, ensuring more personalized and effective marketing outreach.

9) Conclusions:

The extensive analysis of the customer behavior dataset from Kaggle has profoundly demonstrated the efficacy of utilizing data-driven approaches to optimize marketing strategies. This project has been foundational in illustrating how in-depth data analysis, coupled with advanced predictive modeling, can significantly enhance the precision and effectiveness of marketing campaigns. Through the rigorous application of exploratory data analysis (EDA), we have uncovered intricate patterns and relationships within the data that directly inform strategic marketing decisions.

Correlation analysis has been particularly instrumental, revealing pivotal insights into how variables such as income levels and spending habits are intertwined. These insights are critical, as they allow for the crafting of marketing strategies that are not only targeted but also economically savvy. For instance, the positive correlation between income and spending underscores a clear pathway for targeting higher-income segments with premium offerings, thus driving up potential revenue.

Further, the segmentation of the customer base into clearly defined groups based on spending and income has allowed for a more nuanced approach to campaign management. Tailoring marketing efforts to meet the specific characteristics and preferences of each segment ensures that campaigns are not only relevant but also more likely to resonate with the intended audience. This strategic segmentation facilitates the efficient allocation of marketing resources, maximizing impact while minimizing wasteful expenditure.

Predictive models, including logistic regression, decision trees, and support vector machines, have provided a forward-looking lens through which we can anticipate customer responses. The ability to predict which customer segments are most likely to engage with specific marketing

initiatives allows for proactive strategy adjustments, ensuring optimal outcomes from each campaign.

Moreover, the project has highlighted the flexibility required in marketing strategies as they adapt to economic variations. Understanding the elasticity of consumer spending relative to income fluctuations is crucial in developing resilient marketing strategies that can adapt to and thrive in varying economic conditions.

In conclusion, the project has not only yielded insightful revelations about customer behaviors and preferences but has also reinforced the indispensable role of data-driven methodologies in crafting effective marketing strategies. The findings advocate for the continuous integration of comprehensive data analysis into marketing planning and execution. Looking ahead, it is recommended that further research include more diverse variables, enhancing the granularity and accuracy of our predictive models. Additionally, adopting real-time analytics will empower marketers to dynamically refine strategies in response to evolving market conditions and consumer trends, ensuring sustained relevance and effectiveness of marketing efforts.

10) Data Source:
https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

Appendix

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Martial_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years
- NumDealsPurchases: Number of purchases made with discount
- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalog
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to the company's website in the last month
- AcceptedCmp1: 1 if the customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if the customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if the customer accepted the offer in the 3rd campaign, 0 otherwise

- AcceptedCmp4: 1 if the customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if the customer accepted the offer in the 5th campaign, 0 otherwise
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise
- Response (target): 1 if the customer accepted the offer in the last campaign, 0 otherwise