Marketing Strategy Analysis According to Customer Behavior

How consumers decide when purchasing products or services, the personal, social and psychological parameters they are affected by, and their after-sales expectations from companies determine the behavior of consumers.

Businesses try to increase their market shares, brand values and profitability by using marketing strategies by dividing consumers into various groups and applying similar strategies to customers whose behaviors are similar.

Goals

- It is desired to examine the returns to the campaigns through customer behavior analysis and to determine the return effects on the campaign. -By analyzing customer behavior, it is desired to create a model to predict whether the customer will accept the offer in the last campaign.
- It is desired to analyze the customer base according to those who share similar characteristics and to obtain campaigns with better returns on the created clusters.

Task

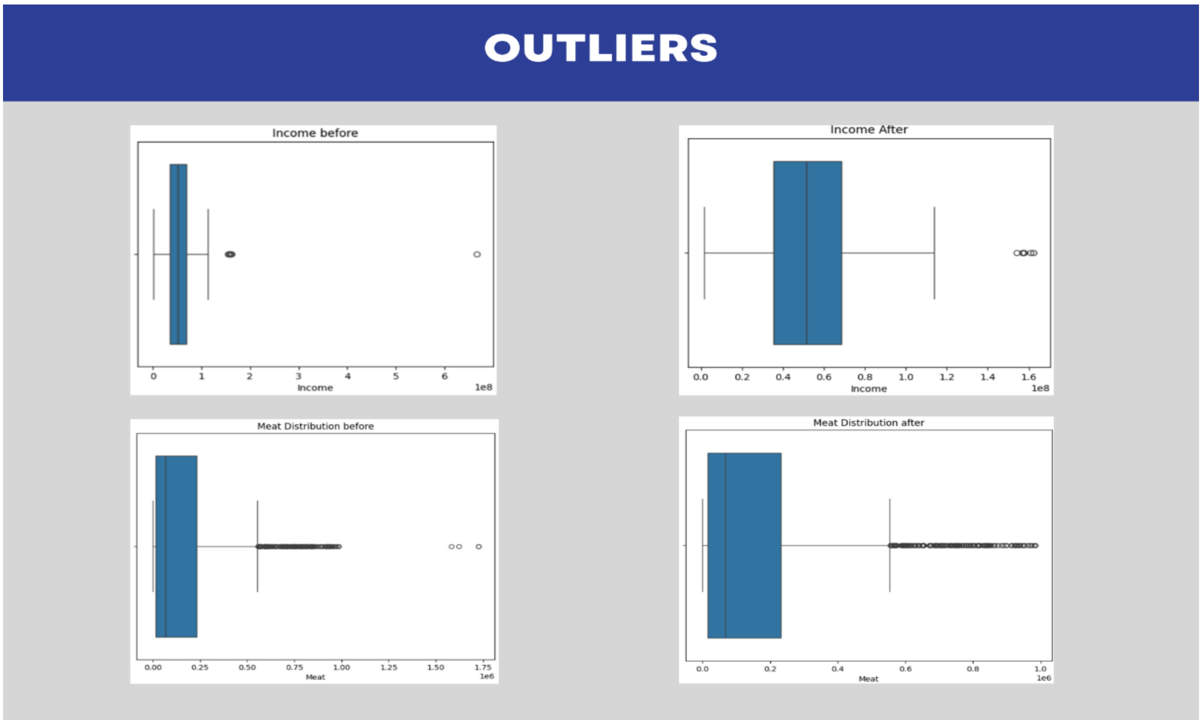
- Load Data
- Data Cleaning
- Feature Engineering
- Data Visualization
- Create Model
- Simple Logistic Regression: Using Metrics: Accuracy, Confusion Matrix, Recall, Precision, F1 Score
- Full Logistic Regression: Using Metrics: Accuracy, Confusion Matrix, Recall, Precision, F1 Score
- Clustering
- Scale Data
- KMeans: Silhouette score
- Analysis
- Visualization evaluation
- Insights

Data Cleaning

1) Handling Missing Value
2) Rename Variables : Education, Marial_Status, Products edited.
3) Remove Unused: ID, Z_CostContact, Z_Revenue removed.
4) Create New Columns: Age, Age Group, Number of Day, Total Spend, Family Size, Total Accepted Cmp, NumOfferAccepted1 (total campaign acceptance), NumOfferAccepted2 (first five campaign acceptances), IsParent created.
5) Categoric Variables
6) Outliers detected.

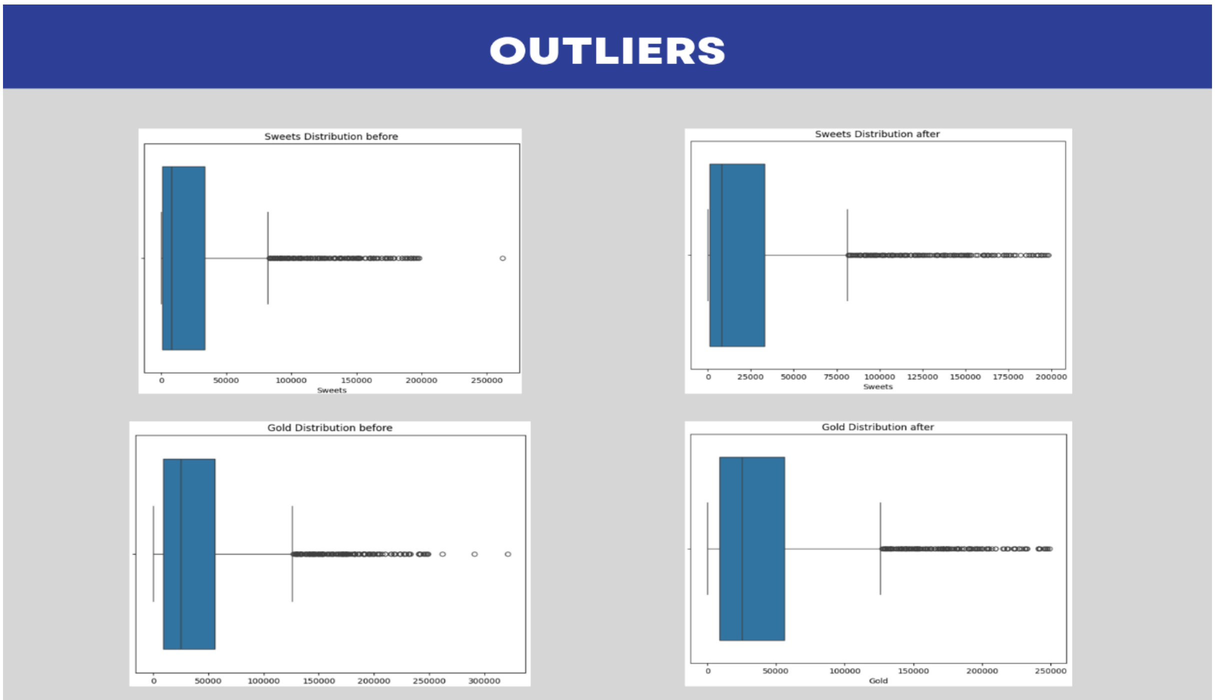
A box plot of the income variable is given before outliers and a box plot after clearing outliers.

Box plots before and after outliers in the Meat Distribution variable are given.



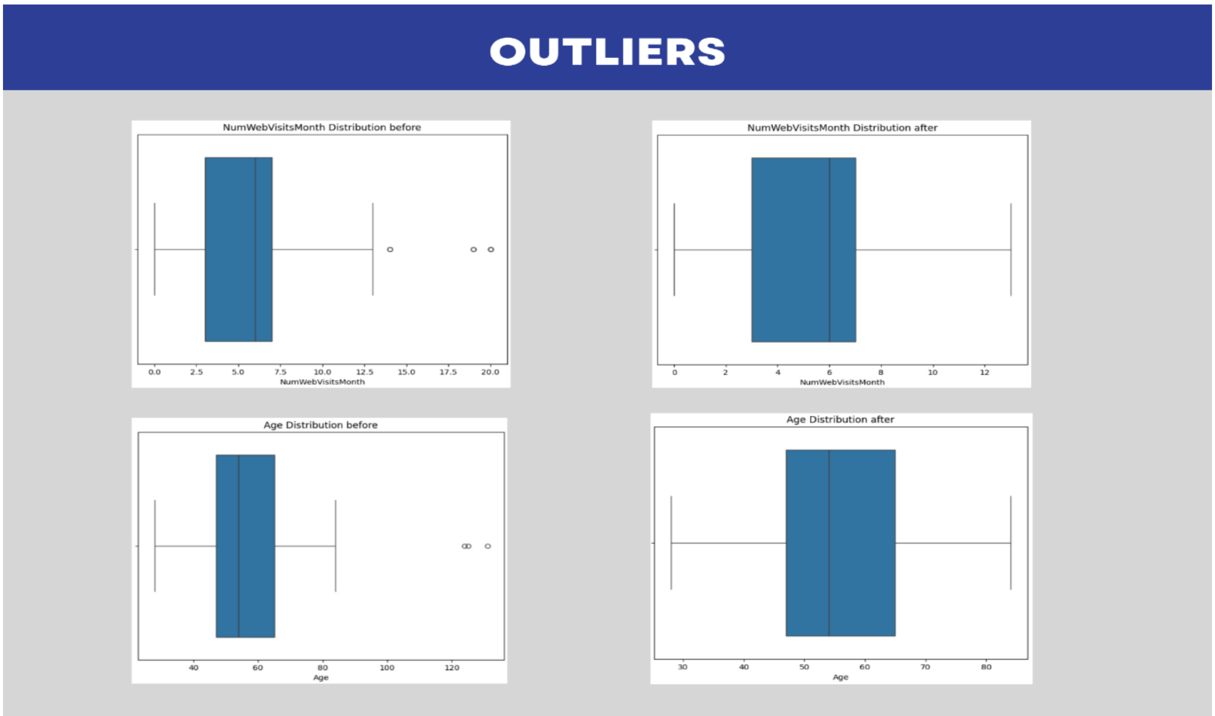
Box plots of the Sweets Disturbition variable before and after outliers are given.

Box plots of the Gold Disturbition variable before and after outliers are given.



Boxplots before and after NumWebVisitsMonth variable outliers are given.

Boxplots before and after Age variable outliers are given.

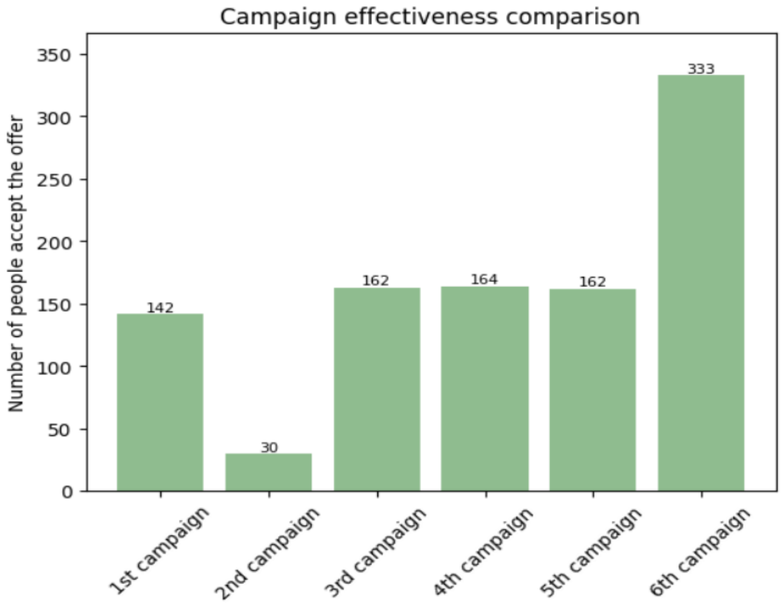


7) Reduce Categories: Marital Status edited.

PreProcessing

Before we start modeling, we organize the data set in the way we will use it.

1) Removed columns.
2) Observe Promotional campaigns.



Acceptance status according to the campaigns was observed more clearly by visualizing them.

According to the graph, the last campaign is more successful.

The second campaign is an unsuccessful campaign.

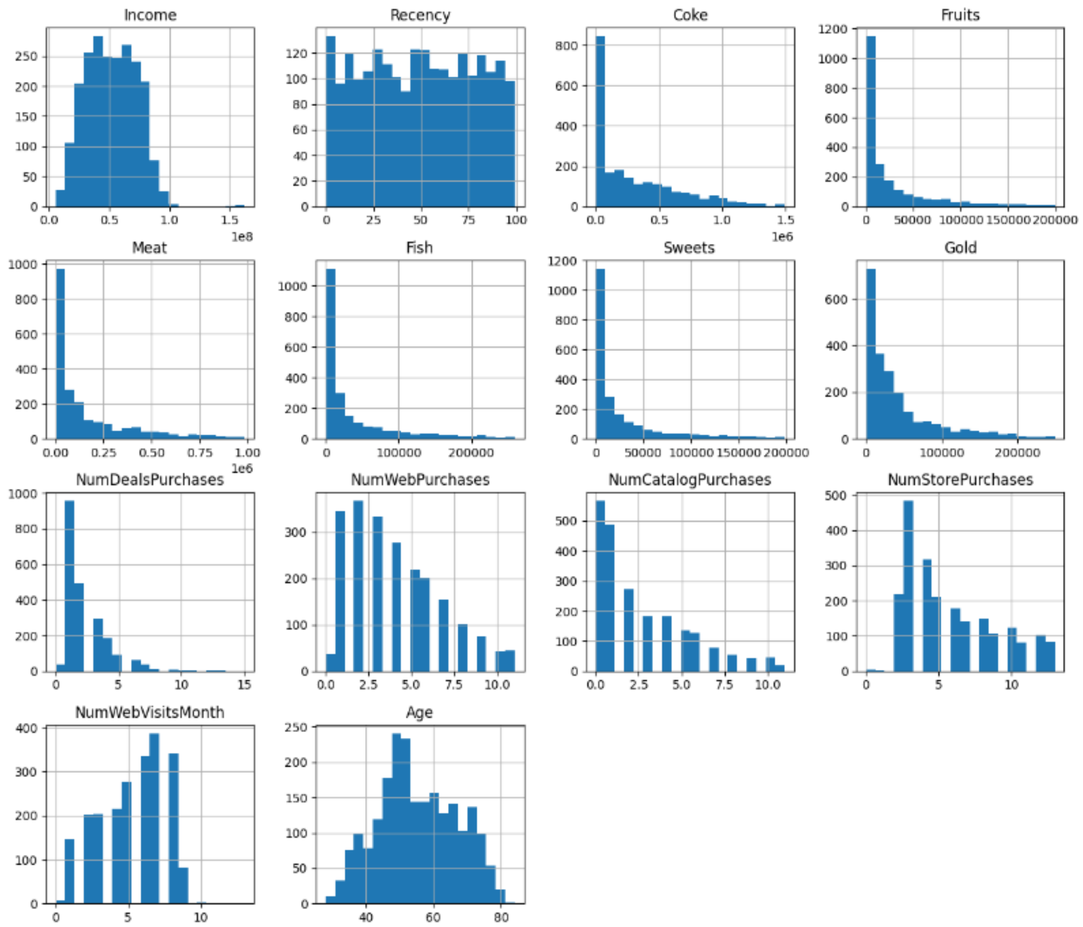
Other campaigns are similar in terms of acceptance status.

Data Visualization

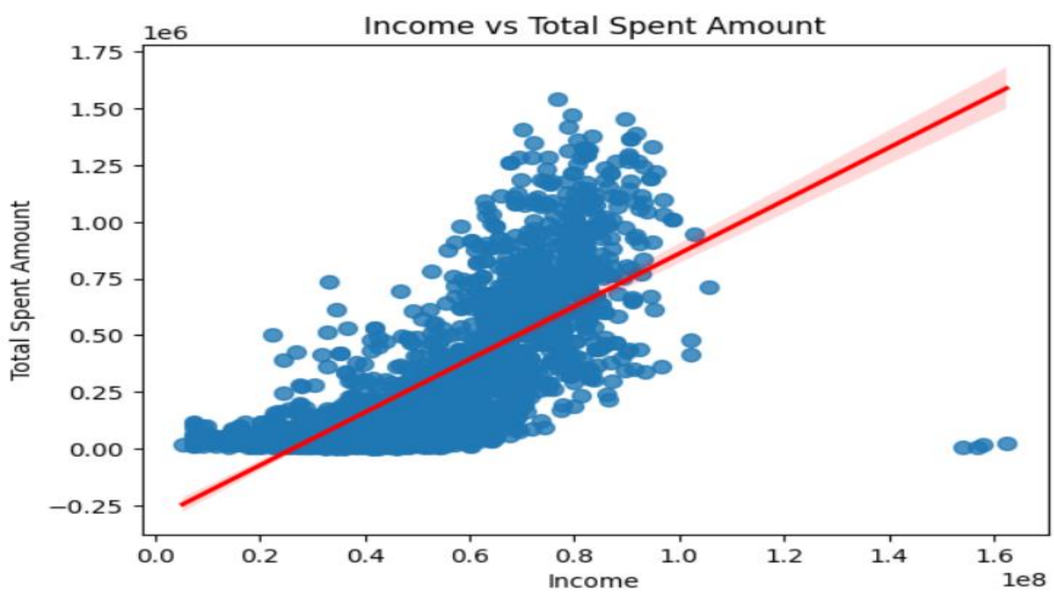
Below is the correlation matrix. Variables with high correlation can be observed in this matrix. According to matrix:

- There is a strong positive correlation of 0.73 between the total amount of expenditure and income.
- It seems that the most spent product type is Meat with 0.95.
- Catalog Buying has better correlation than in-store, web and deal purchases.
- Customer interest in the last campaign is better than other campaigns. 0.73 strong positive correlation.

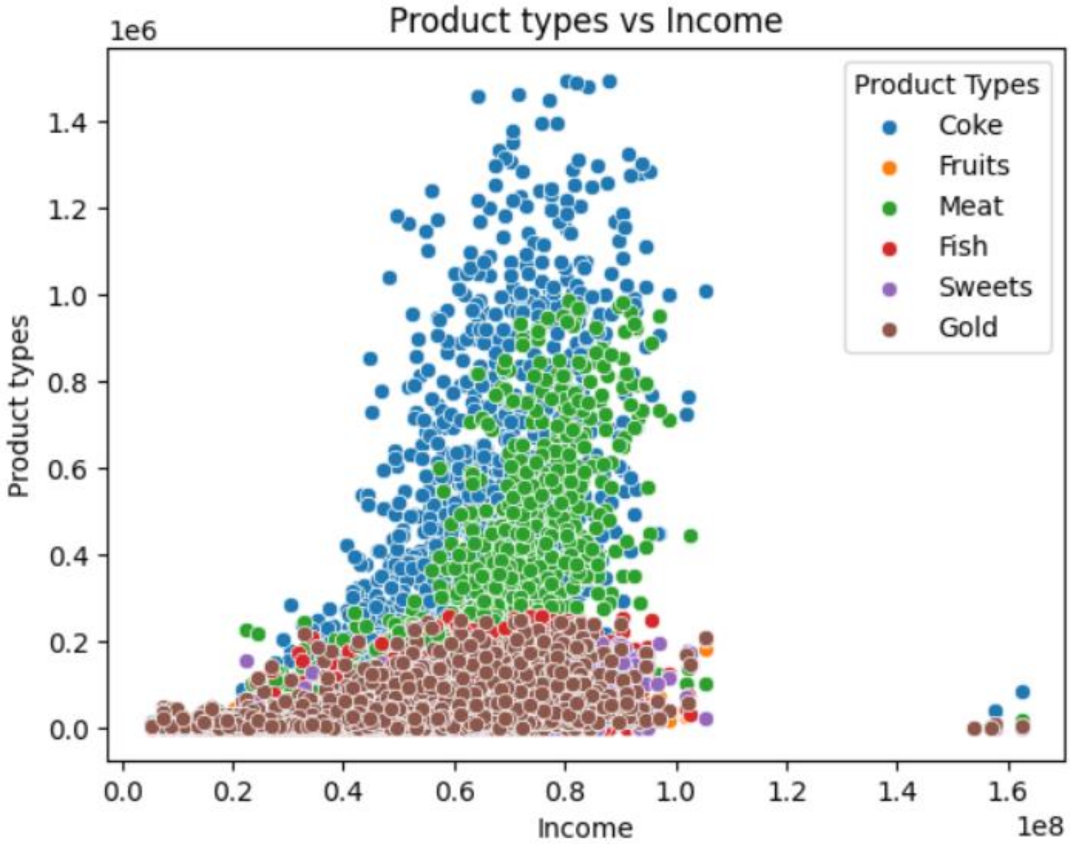
It was created to see the distribution of relevant variables.



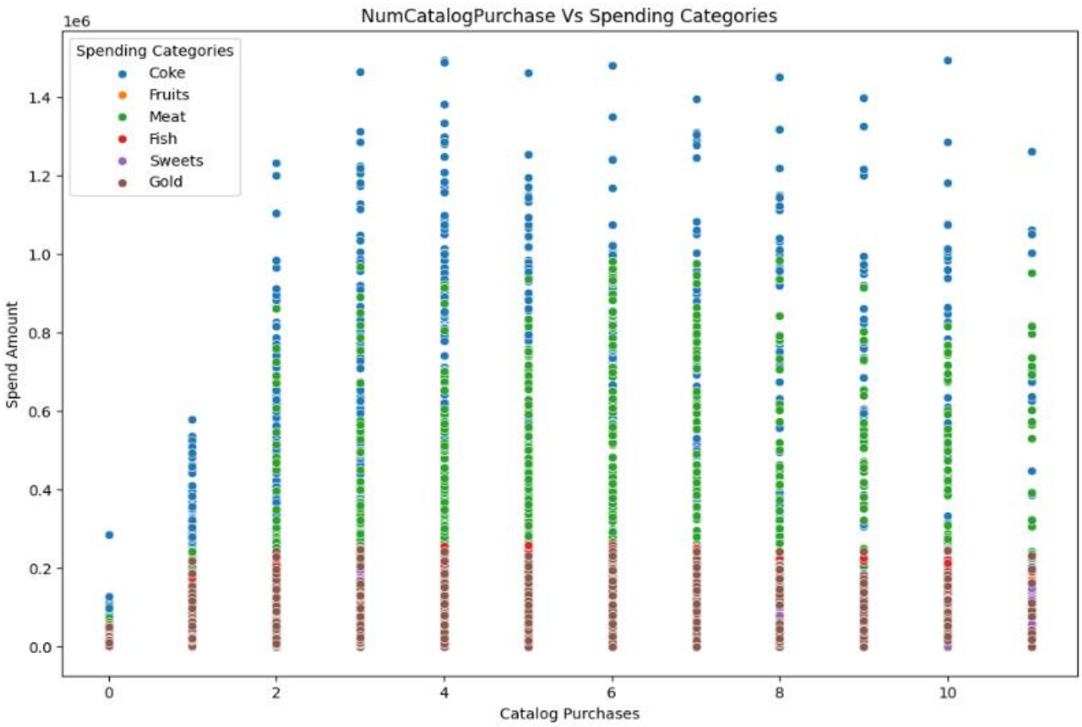
The relationships of some variables with each other were graphed with a scatter plot. There appears to be a positive relationship between income and total spending amount. It seems that the total spending amount of high-income customers is also higher.



Sales amounts were visualized according to revenue. Those with higher incomes bought the most cola. The flesh follows him. At least gold was exchanged.



The types of products sold were visualized according to catalog sales. It is seen that Gold sales are low here too.

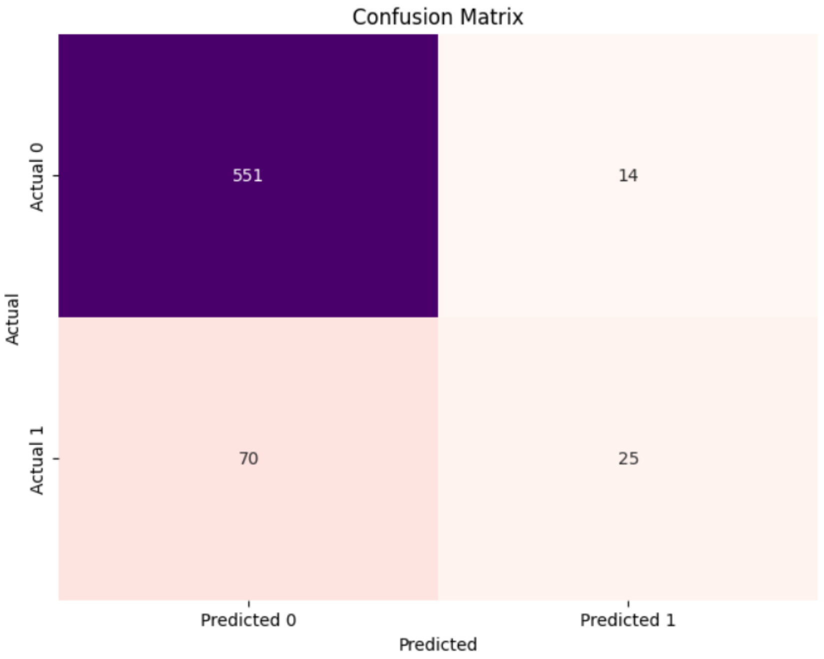


Modeling

1) Simple Logistic

The data set was divided into training and validation data sets and then the logistic regression model was applied. The statistical summary of the model was displayed using the *statsmodels* library.

The aim is to examine how a particular dependent variable (AcceptedCmp6) can be predicted by a logistic regression model.



Confusion matrix and other metrics were calculated.

Accuracy: 0.8727

Precision: 0.6410

Recall: 0.2632

F1 Score: 0.3731

- Precision: 0.6410 64.1% of the customers that the model predicted as "1" actually accepted the campaign.
- Recall: 0.2632 26.3% of the customers who accepted the campaign were correctly detected by the model.
- Accuracy: 0.8727 86.6% of the model's total predictions are correct
- F1 Score: 0.3731 It is used to evaluate whether the model performs well in terms of both precision and recall. In this case, a 38.7% F1 Score was obtained.

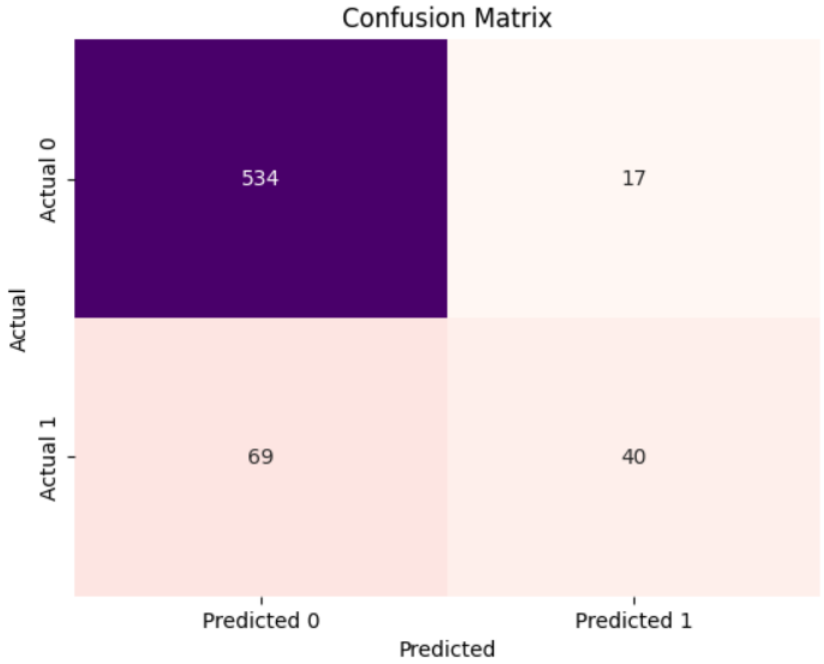
Looking at these values, we can observe that the model performance is low.

2) Full Logistic

Goal: AcceptedCmp6, is estimated based on other columns

Dependent Variable: AcceptedCmp6

Argument: All columns



Accuracy: 0.87

Precision: 0.70

Recall: 0.37

F1 Score: 0.48

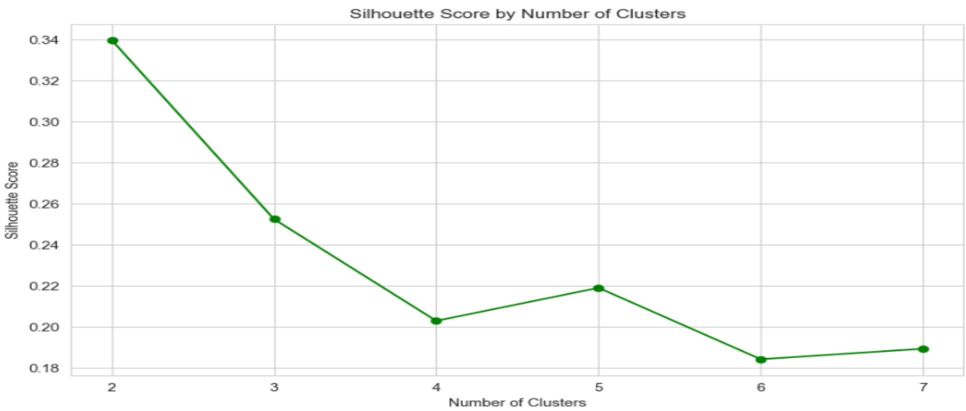
- Precision: 0.70 percent of samples predicted as positive are actually positive. Good value but could be improved depending on the problem.
- Recall: 0.37, percent of truly positive samples were correctly predicted as positive. It was more successful than the previous model, but it could have been higher.
- Accuracy: 0.87 The rate at which the model predicts correctly is quite high.
- F1 Score: 0.48, balanced as a combination of precision and recall.

Clustering

1) **KMeans n_clusters=3**

K-Means is frequently applied in customer segmentation and market segmentation processes. It works by dividing large clusters to obtain clusters that are as discrete as possible to minimize the value of the error parameter.

Silhouette_score is used in such clustering algorithms. Varies between -1 and 1. A high value indicates that the object matches well with its cluster.

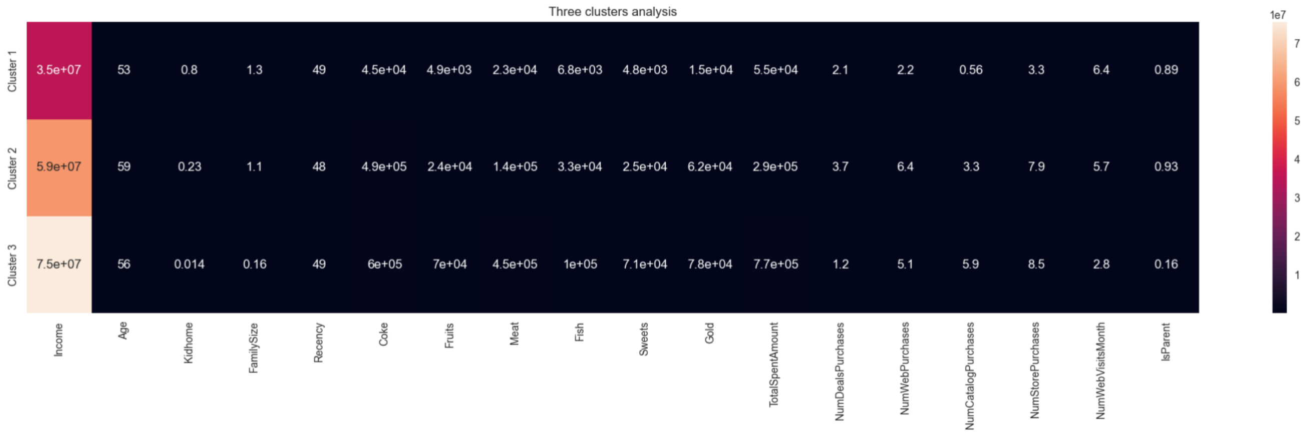


SilhouetteVisualizer class in the Yellowbrick library was used for the visualization process.

Interpreting the SilhouetteVisualizer graphic:

- A score approaching 1 indicates that a data point fits well within its cluster and is well separated from other cluster members.
- A score approaching 0 indicates that a data point does not fit properly into its cluster and cannot be distinguished from other clusters.
- A score approaching -1 indicates that a data point is incorrectly clustered and is more similar than other cluster members.

K Means model was created using n_clusters=3.



Analysis was made by examining it with a detailed heatmap.

Cluster 3 Insights:

- The focus should be on this cluster. Because almost every type of product has high sales.
- Those with high income are in this cluster.
- Focus should be on catalog and store sales. Because the most spending is here.
- Online purchasing is high but website visits are low.

Cluster 2 Insights:

- This customer should not have any room points.
- Purchases are low. They visit the website but do not buy the product.

Cluster 1 Insights:

- A cluster with a high concentration of customers who have more children at home compared to other clusters.
- Too many products are purchased. The website is highly visited.
- The second focus should be this cluster. They are customers with average income. They are more likely to purchase medium or low cost products.

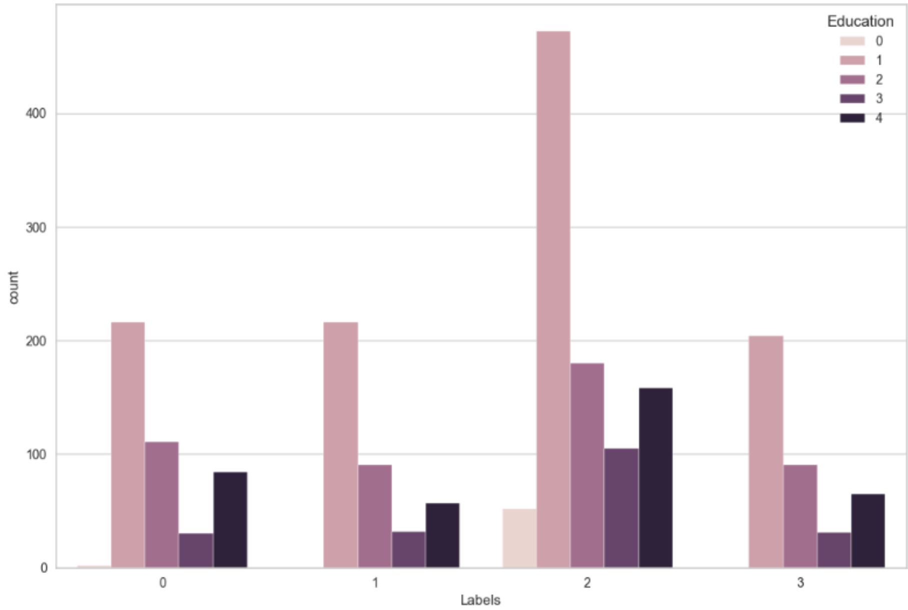
2) **KMeans n_clusters=4**

A new model was created for n_clusters=4 by making changes to the data set. Necessary variables were edited using LabelEncoder. Scale operation was performed using StandardScaler() and the data was ready.

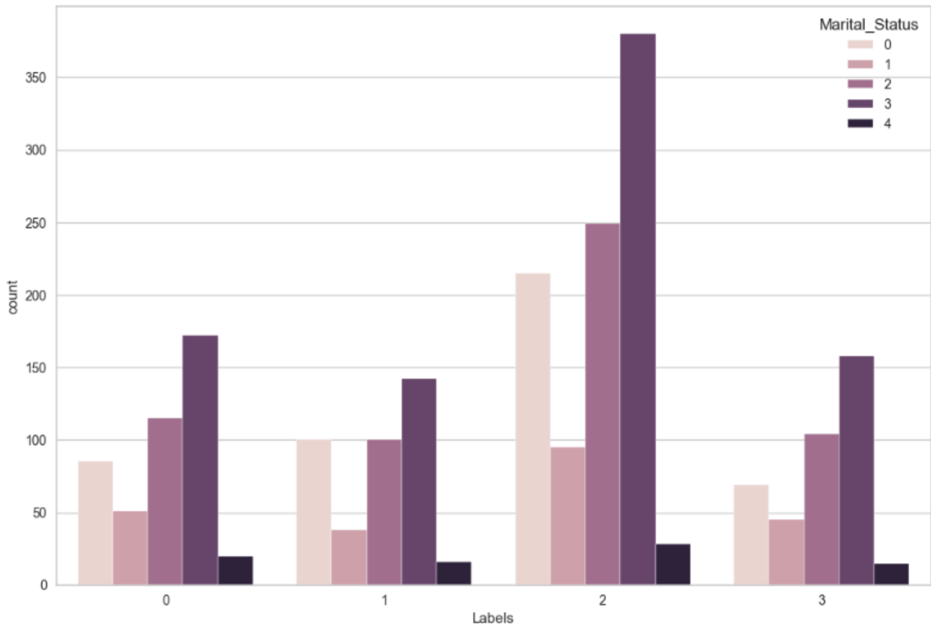
Silhouette Score calculation was also made here. KMeans model was created for n_clusters=4.

Clusters were observed according to educational status, marital status and age groups.

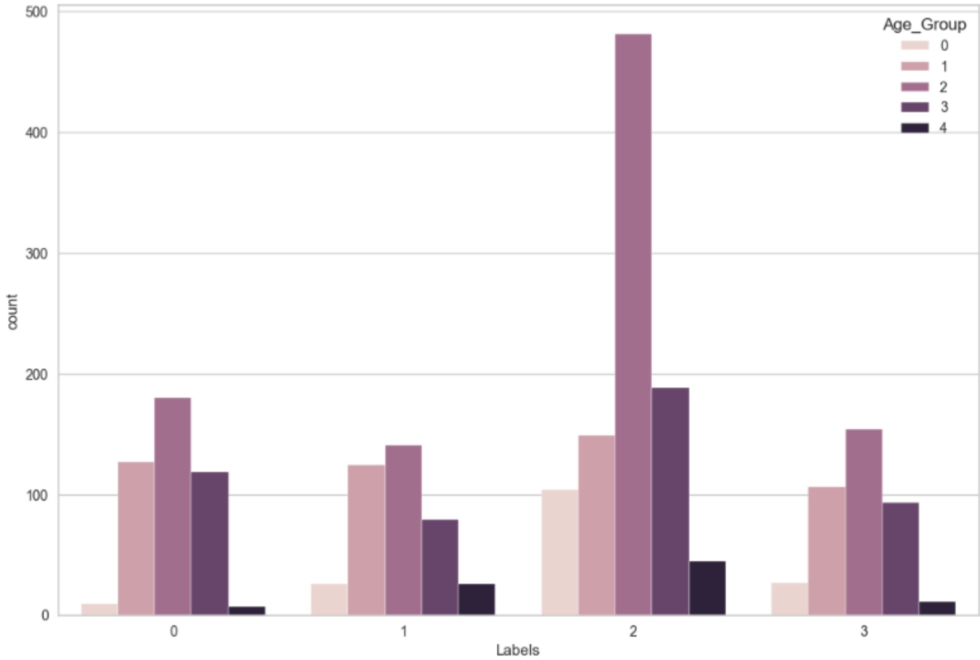
For Education:



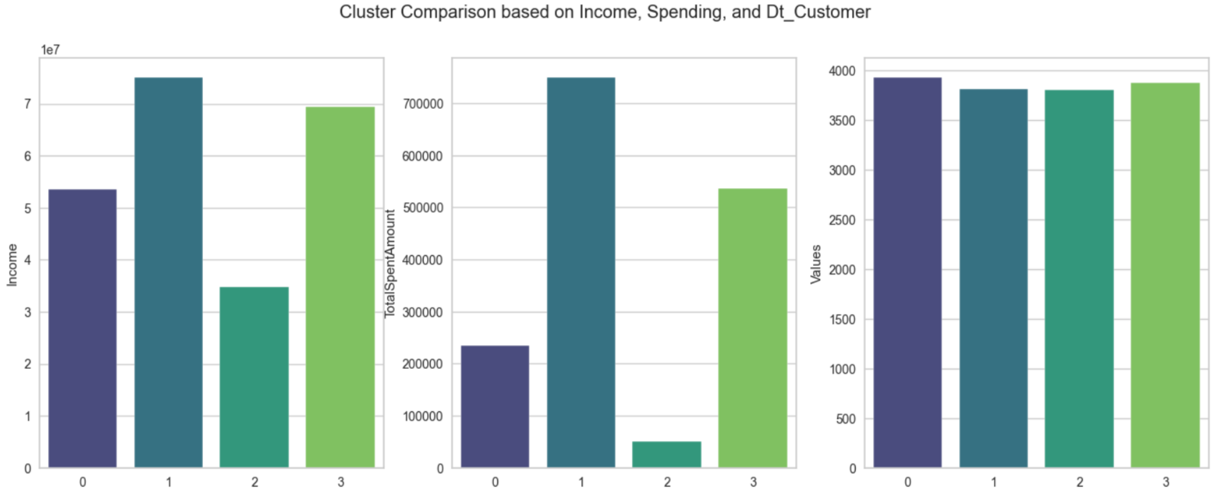
For Marital_Status:



For Age Group:



Customer clusters are analyzed according to income levels, total expenses and registration day variables.



Cluster 0 Insights:

- Low-income, old customer group with low shopping potential. Not a suitable target.

Cluster 1 Insights:

- High potential customer group. Campaign acceptance rates, income levels and total expenditures are also better. A very suitable target.

Cluster 2 Insights:

- Low income, low spending, slightly newer customers. The majority of married customers.

Cluster 3 Insights:

- Customer group with high income levels and spending above average.

Conculusion

- When we evaluate the models in general, the last campaign offer acceptance was estimated separately in logistic models according to previous campaign acceptances and other customer behaviors.
- In clustering models, in order to better evaluate customer profiles, analysis was made by dividing them into 3 clusters and then 4 clusters.
- Models have enabled us to analyze customer behavior indicating that customers with high incomes, total shopping volumes are high, and customers with fewer children have more income.

REFERENCES

Kaggle

Model Evaluation Metrics

Medium

Logistic Regression

Customer Behavior

KMeans Algorithm

KMeans Metrics

Silhouette Score

Silhouette Score Visualizer