Predict Customer Personality to boost marketing campaign by using Machine Learning

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Created by:
Bagus Ariobimo
bagusariobimo7@gmail.com
https://www.linkedin.com/in/bagus-ariobimo
https://github.com/riyouuyt

"A data enthusiast who has completed a course in this field and is ready to start his career. Have a excellent understanding in statistics, programming, and data processing. Proficient in using tools such as Python, R, and SQL. Able to collect, clean and analyze data with the necessary techniques. Have skills in data visualization and simple statistical modeling. Creative problem solver and always thrive for learning. Ready to contribute to data-driven projects and collaborate in teams. Committed to further developing data science skills and achieving significant results in data analysis."

Data Overview

Business Overview

Our company's rapid growth is intricately tied to our deep understanding of customer personalities. We use historical marketing campaign data to optimize performance and precisely target potential loyal customers, driving transactions on our platform. Our key strategy involves developing a predictive clustering model, enabling data-driven decisions. By clustering customers based on behavior and personality, we provide tailored services and personalized marketing, fostering customer loyalty. Our goal is to set industry standards in customer-centric operations and sustainable growth through continuous data-driven refinement.

Objective

Our primary objective is to optimize the marketing campaign by leveraging customer segmentation and data analytics. We aim to enhance customer engagement, increase conversion rates, and boost revenue while ensuring a seamless and personalized experience for our customers.

Goals

- **Segment-Specific Targeting**: Implement targeted marketing strategies for each customer segment, focusing on their specific characteristics and preferences.
- **Conversion Rate Optimization:** Continuously refine and improve our conversion funnels, using data-driven insights to enhance the customer journey and boost conversion rates.
- **Customer Engagement Enhancement:** Elevate the website experience and content to engage customers effectively. Develop loyalty programs and incentives to create lasting relationships.

Note: This is not a real company. The names "**ShopSavvy Emporium**" provided earlier are fictional names created for the purpose of this project. Please be aware that these names are entirely fictitious and not associated with any real businesses.

Data Preparation

Project Data Column Information:

- **Unnamed: 0**: An unnamed index or identifier column.
- **ID**: Customer identification number or code.
- **Year_Birth**: Year of birth of the customer.
- **Education**: The level of education attained by the customer.
- **Marital_Status**: Marital status of the customer.
- **Income**: Customer's income.
- **Kidhome**: Number of children in the household.
- **Teenhome**: Number of teenagers in the household.
- **Dt_Customer**: Date when the customer became a client.
- **Recency**: Number of days since the last purchase.
- **MntCoke**: Amount spent on Coke products.
- **MntFruits**: Amount spent on fruit products.
- **MntMeatProducts**: Amount spent on meat products.
- **MntFishProducts**: Amount spent on fish products.

For more information about the project, click here

Data Preparation

- **MntSweetProducts**: Amount spent on sweet products.
- **MntGoldProds**: Amount spent on gold products.
- **NumDealsPurchases**: Number of purchases made with deals or discounts.
- **NumWebPurchases**: Number of purchases made through the web.
- **NumCatalogPurchases**: Number of purchases made from catalogs.
- **NumStorePurchases**: Number of purchases made in physical stores.
- **NumWebVisitsMonth**: Number of web visits per month.
- **AcceptedCmp3**: Whether the customer accepted Campaign 3 (binary, likely a marketing campaign).
- **AcceptedCmp4**: Whether the customer accepted Campaign 4 (binary, likely a marketing campaign).
- **AcceptedCmp5**: Whether the customer accepted Campaign 5 (binary, likely a marketing campaign).
- **AcceptedCmp1**: Whether the customer accepted Campaign 1 (binary, likely a marketing campaign).
- **AcceptedCmp2**: Whether the customer accepted Campaign 2 (binary, likely a marketing campaign).
- **Complain**: Whether the customer has registered a complaint (binary).
- **Z_CostContact**: Cost of contacting the customer.
- **Z_Revenue**: Revenue generated from the customer.
- **Response**: Customer response to a marketing campaign (binary, likely indicating whether they responded positively to a campaign).

For more information about the project, click here

Feature Engineering

For more information about the project, click here

🚀 Feature Engineering 🚀

Rakamin Academy

📊 Introduction to Feature Engineering

In my quest to enhance the success of our marketing campaign, i embarked on a journey through data. Feature engineering was my guiding star, allowing me to uncover deeper insights into customer behavior and boost conversions. 💡

1. **Creating the Conversion Rate** 📈

I commenced by calculating the conversion rate, a pivotal metric that measures the percentage of website visitors who responded to our campaign. This served as the foundation for comprehending customer behavior. 🧮

The calculate conversion rate are from:

Total Responses / Total web visit

2. 🗓️ **Customer Age Insights**

I segmented our customers into five distinct age groups. This segmentation allowed us to gain insights into the preferences and behaviors of different age cohorts, spanning from children to senior adults. 👶👵

3. 💰 **Income Labeling**

We didn't stop at numerical data; we also created a meaningful "Income Level" feature. By categorizing income into four distinct labels—**Low Income, Moderate Income, High Income, and Very High Income**—we gained insights into spending behavior and purchasing power. 💼💎

🚀 Feature Engineering 🚀

Rakamin Academy

4. ⏰ Unlocking Recency Insights

Recency is a crucial factor in customer engagement. We segmented customers based on their recency of interaction with our brand. This allowed us to tailor our marketing efforts to customers who were recently active and those who might need a gentle nudge to re-engage. 🗓️

5. 🛒 Total Transactions Analysis

Total transactions give us a comprehensive view of customer engagement. We calculated the total number of transactions for each customer, shedding light on their loyalty and engagement with our products and services. 📊🛍️

6. 👪 Understanding Family Size

Family size can influence purchasing decisions. We engineered the "Family_Size" feature, providing insights into the composition of our customers' households. This information is invaluable for crafting family-centric marketing campaigns. 👪🏠

7. 📅 Recency Grouping

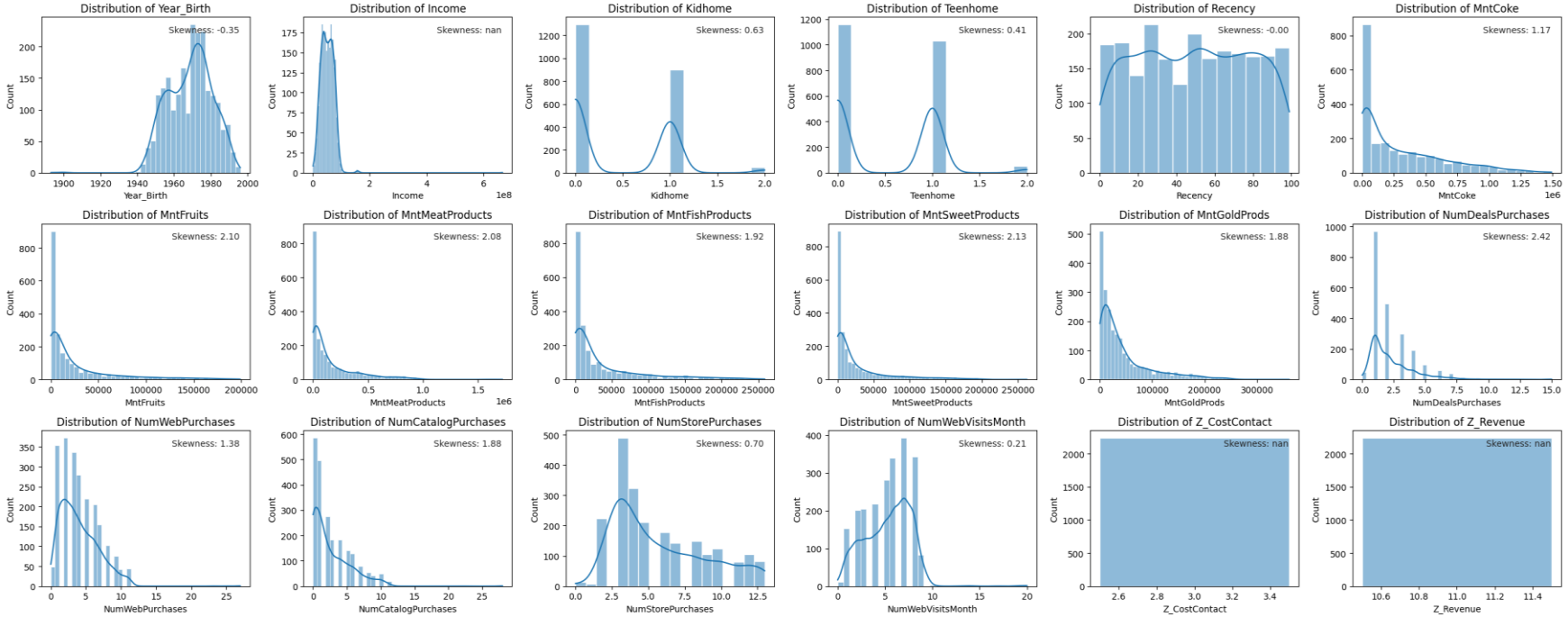
To further refine our strategies, we grouped customers by recency into distinct segments. This allowed us to tailor our communication and offers based on how recently customers interacted with our brand, ensuring relevance and engagement. 📊🔍

📈Exploratory Data Analysis📈

Exploratory Data Analysis

Rakamin Academy

1. Univariate EDA for numeric



Summary of Distribution Characteristics

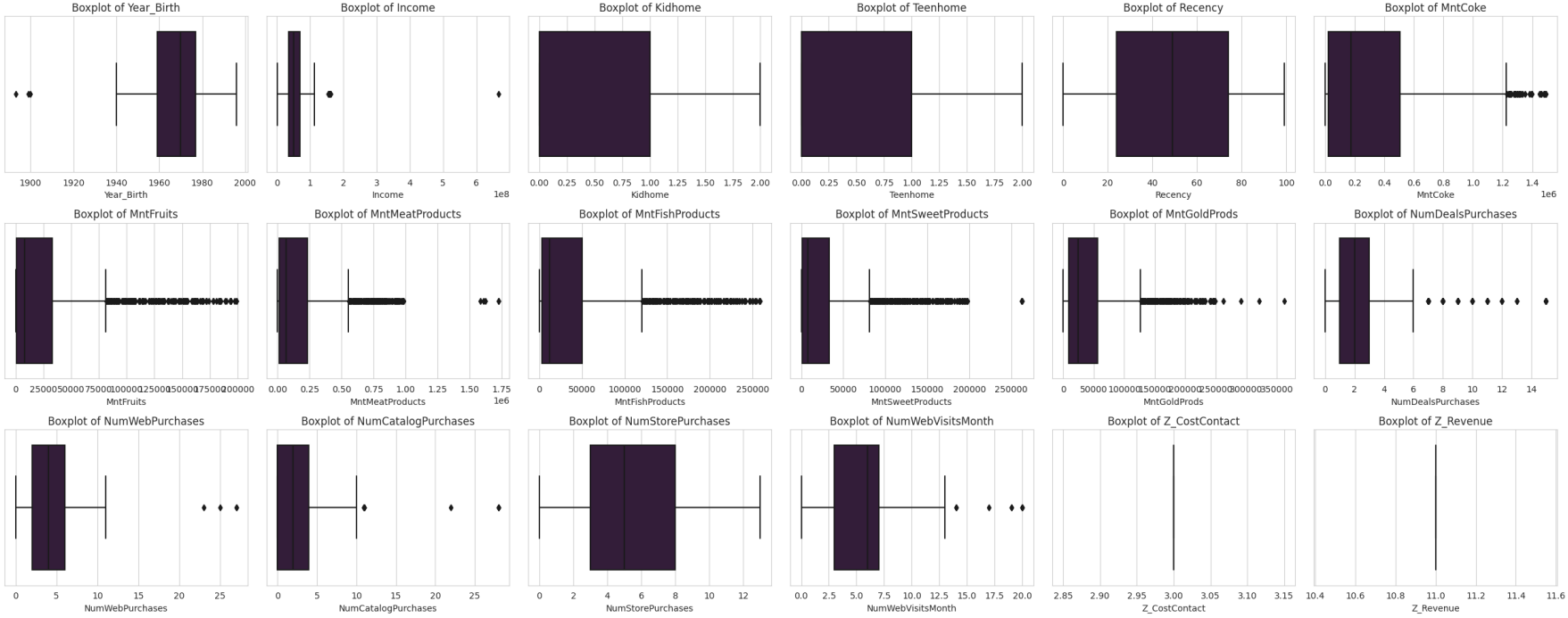
1. **Skewness**: Measures the asymmetry in data distribution.
   - Most columns are **highly positively skewed** (skewness > 1), indicating a longer tail on the right side.
   - Columns with skewness values between -0.5 and 0.5 are **approximately symmetrical**.
   - "Kidhome" and "Teenhome" have a **bimodal distribution**, suggesting two distinct modes.

2. **Kurtosis**: Measures tailedness or peakedness compared to a normal distribution.
   - Several columns have **high positive kurtosis**, implying heavier tails and peaks.
   - Notably, "NumDealsPurchases," "NumWebPurchases," and "NumCatalogPurchases" exhibit this characteristic.

3. **Type of Distribution**: Describes the distribution based on skewness values.
   - Most columns are **highly positively skewed**.
   - "Year_Birth," "Recency," and "NumWebVisitsMonth" are **approximately symmetrical**.
   - "Kidhome" and "Teenhome" exhibit a **bimodal distribution**.
   - "Z_CostContact" and "Z_Revenue" have a **uniform distribution** with constant values.

Exploratory Data Analysis

Rakamin Academy



📈Exploratory Data Analysis📈

Rakamin Academy

🔍 Outlier Analysis Summary 🔍

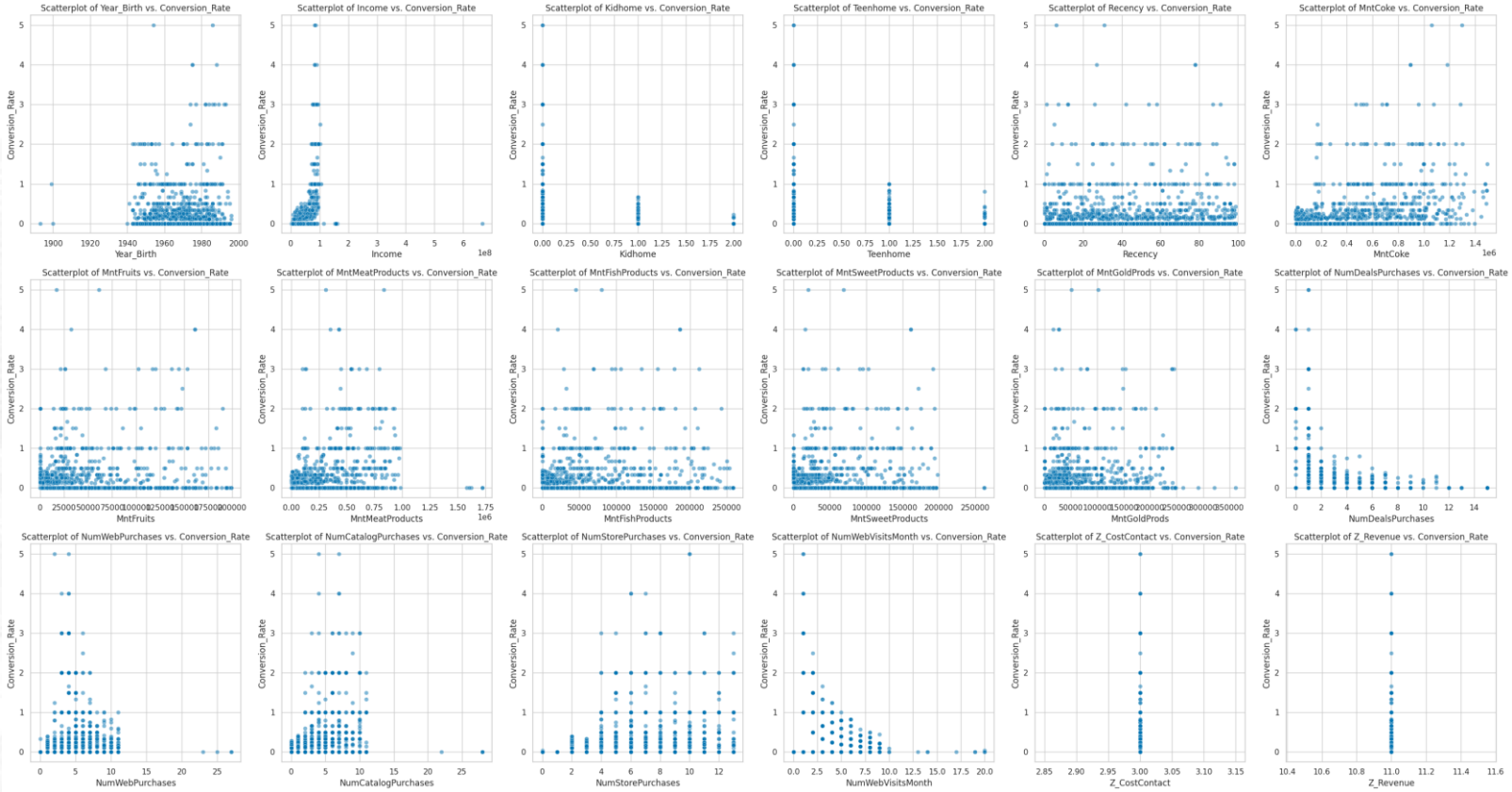
- **Year_Birth**: Detected outliers in birth years, but the impact is minimal (0.13% outliers).
- **Income**: Some extreme income values but relatively low impact (0.36% outliers).
- **Kidhome & Teenhome**: No outliers found in the number of children at home.
- **Recency**: No outliers detected in the recency of purchases.
- **MntCoke to MntGoldProds**: Significant outliers in spending on various product categories (ranging from 1.56% to 11.07% outliers).
- **NumDealsPurchases to NumCatalogPurchases**: Outliers found in the number of deals and catalog purchases (ranging from 1.03% to 3.84% outliers).
- **NumWebPurchases & NumStorePurchases**: Minimal outliers in web and store purchases (below 0.18% outliers).
- **NumWebVisitsMonth**: Detected outliers in web visits, but the impact is relatively low (0.36% outliers).
- **Z_CostContact & Z_Revenue**: No outliers in contact cost and revenue.

Overall, the dataset contains outliers in income, spending on various product categories, and some purchase-related variables. These outliers should be further investigated for data quality and potential impact on the analysis.

Exploratory Data Analysis

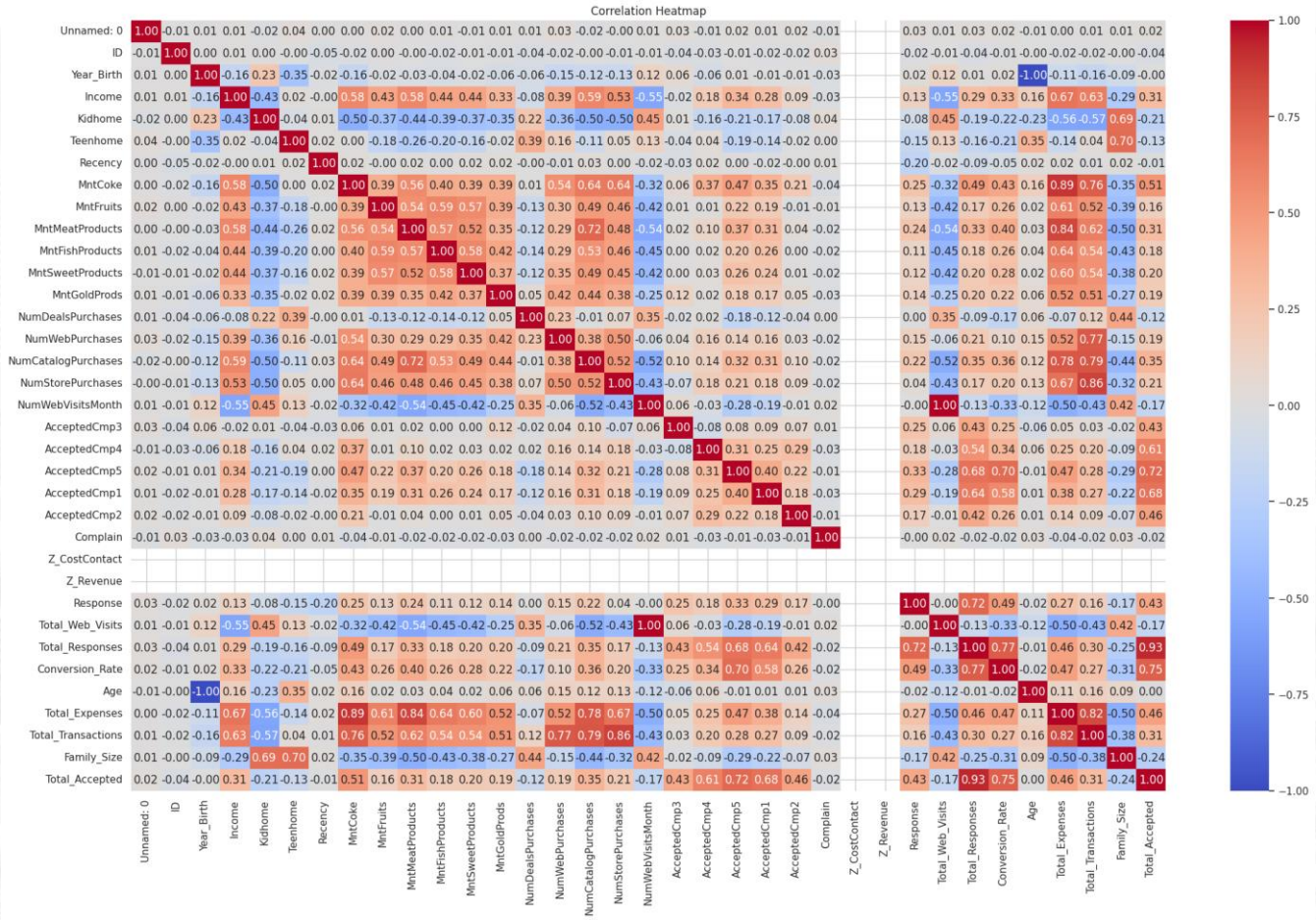
Rakamin Academy

2. Correlation Exploratory Data analysis



Exploratory Data Analysis

Rakamin Academy



Exploratory Data Analysis

Rakamin Academy

Key Correlations for Conversion Rate Analysis

In our exploration of the data, we've uncovered intriguing correlations between various factors and our Conversion Rate. These insights are crucial for tailoring our marketing strategies.

Top Positive Correlations:

- **Total Responses:** A strong positive correlation of 0.766 suggests that higher total responses lead to a better conversion rate.
- **Total Accepted:** With a correlation of 0.750, a higher number of total accepted offers positively impacts conversion.
- **AcceptedCmp5:** This campaign, with a correlation of 0.700, plays a significant role in boosting conversion.
- **AcceptedCmp1:** Positive correlation of 0.577 highlights its effectiveness in driving conversions.
- **Response:** Respondents exhibit a positive correlation of 0.486 with conversion.

Other Positive Correlations:

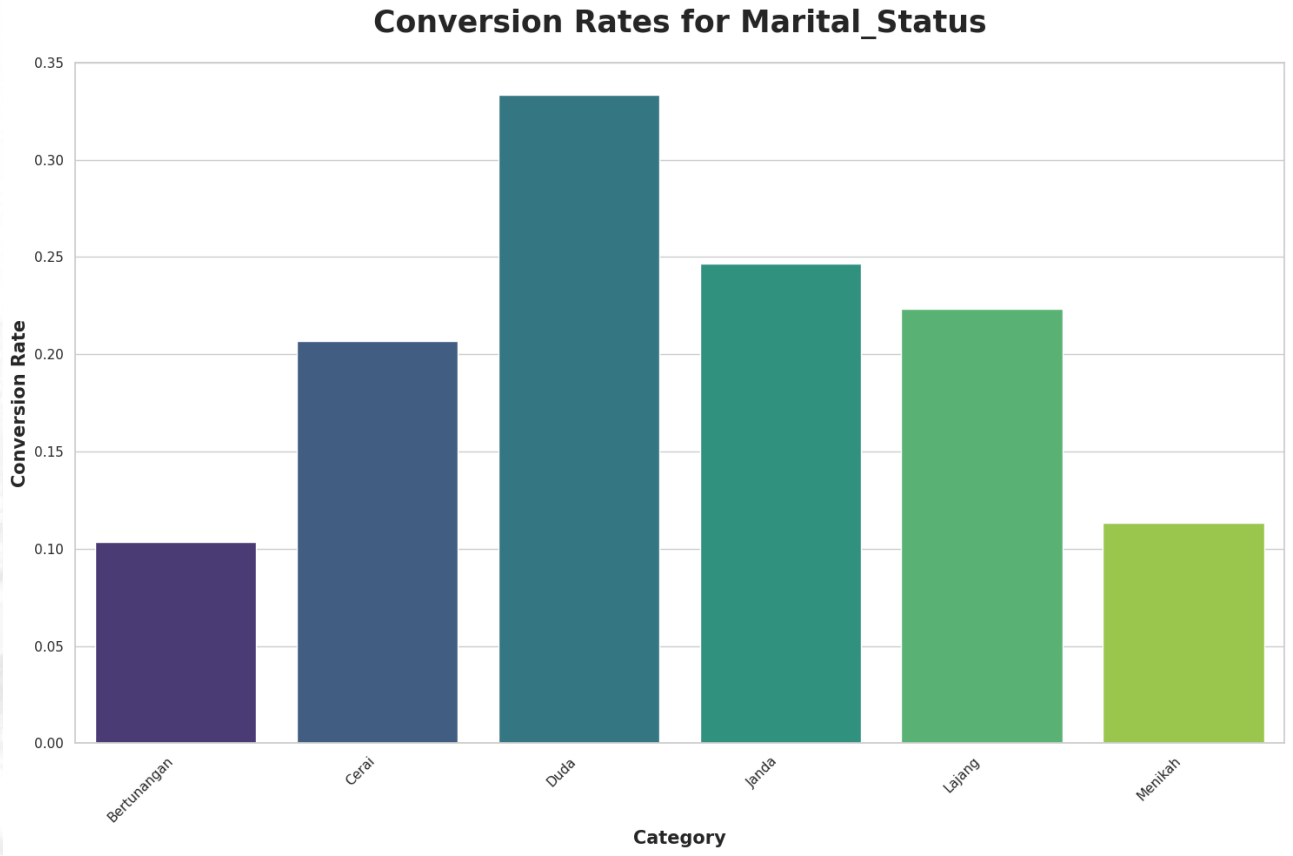
Total Expenses, MntCoke, MntMeatProducts, NumCatalogPurchases, and more contribute positively to conversion.

Negative Correlations:

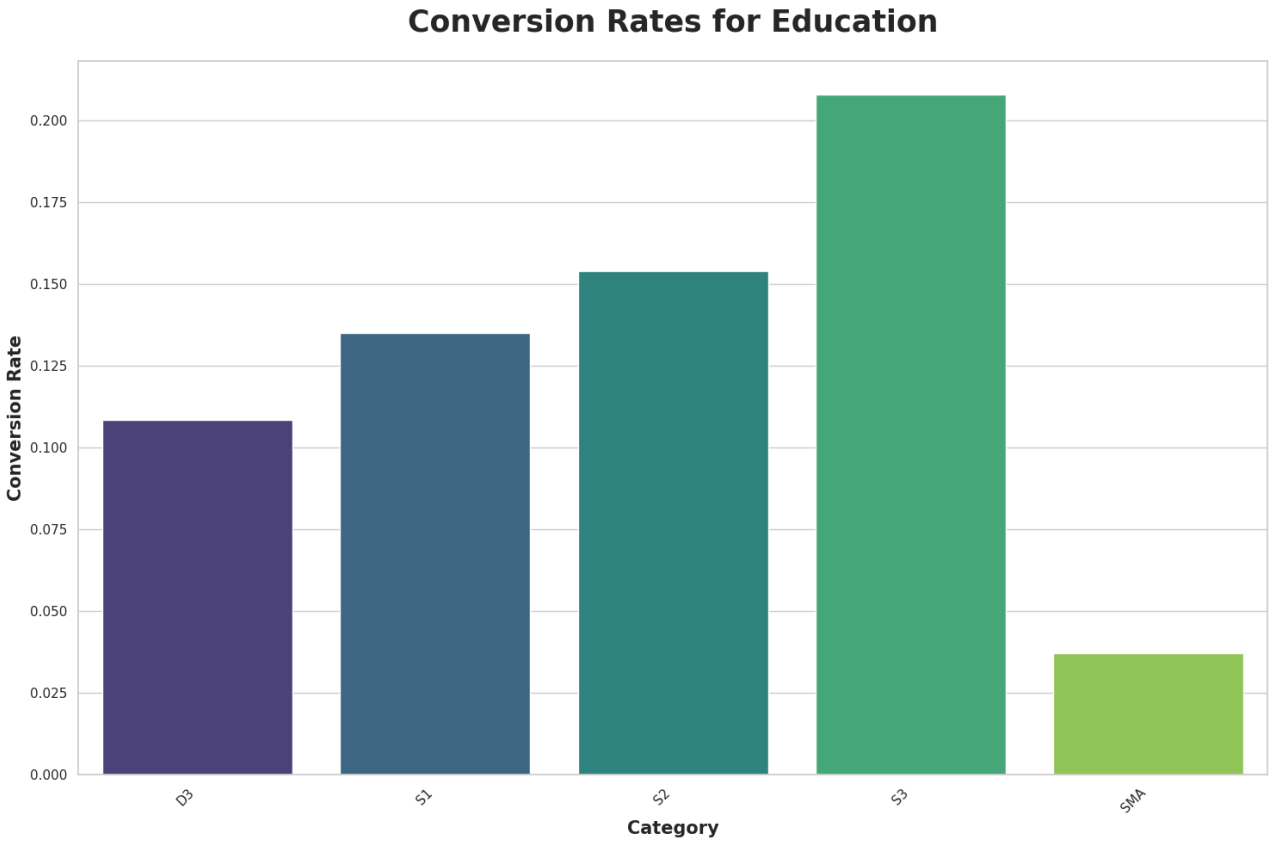
Z_CostContact and Z_Revenue exhibit no correlation with Conversion Rate, indicating they don't significantly impact our campaigns.

These insights guide us in crafting data-driven marketing strategies to optimize our Conversion Rate. Let's harness these correlations for success!

3. EDA for Categorical Columns



3. EDA for Categorical Columns



Predict Customer Personality to Boost Marketing Campaign

I have executed a robust data preprocessing phase, aiming to prepare the dataset for comprehensive analysis. Let's explore the remarkable steps i've accomplished through a series of powerful actions:

- **Data Cleaning** 🧹: The primary mission was to cleanse the data from imperfections, including missing values, duplicates, and the relentless outliers. This step was pivotal in ensuring data quality.
- **Outlier Handling** 🚫: We addressed outliers fearlessly, making informed decisions on whether to remove, transform, or retain them based on rigorous analysis.
- **Standardization** 📏: The focus was on standardizing the data, ensuring uniformity across all variables, a critical step in robust analysis
- **Feature Encoding** 🛠️**:** Before we began into diving onto the machine learning, the categorical data has to encoded so that the data can be able fit in to the machine

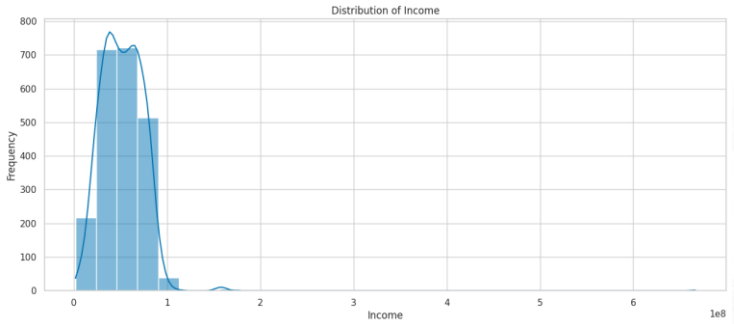
These systematic actions have laid the groundwork for our upcoming analyses. Our objective is to gain profound insights into our data, enabling us to formulate intelligent solutions based on our findings. Together, we embark on a journey towards deeper knowledge! 🌟"

For more information about the project, click here

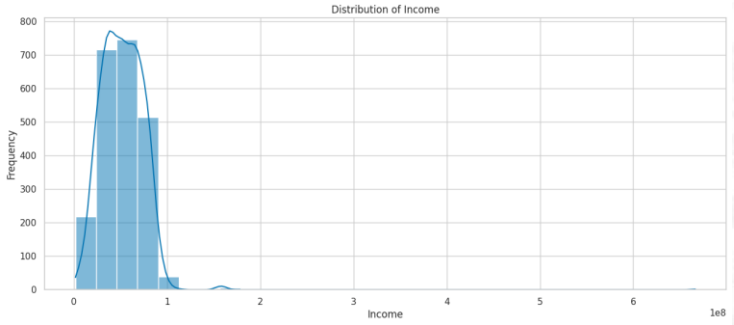
Data Preprocessing

In the dataset, we have a total of 2240 rows and 29 columns. Notably, the 'Income' column contains 2216 non-null values, while 24 values are missing. Rather than imputing these missing 'Income' values, we have chosen to remove the rows with missing data. This approach simplifies the dataset and ensures that only complete and available information is used for analysis. Removing the null data helps maintain the integrity of the dataset, avoiding any potential bias introduced by imputing missing values with a specific measure.

Before the preprocessing



After the preprocessing



For more information about the project, click here

Data Preprocessing

The other data cleansing process, we executed two key steps to enhance the quality and relevance of our dataset.

1. Removing Duplicate Rows
We initially conducted a check for duplicate rows within the dataset. Fortunately, no duplicated rows were identified. Duplicate rows, if present, could lead to inaccuracies in our analysis, and their removal ensures data integrity.

2. Dropping Unnecessary Columns
To streamline our dataset and focus on the most relevant features, we dropped several columns deemed unnecessary for our analysis. These columns, namely **'Unnamed: 0,' 'ID,' 'Dt_Customer,' 'Z_CostContact,' and 'Z_Revenue,'** were excluded from the cleaned dataset.

By eliminating these columns, we simplify the data structure, making it more manageable and conducive for subsequent analysis and modeling. This step also aids in improving computational efficiency and data interpretability.

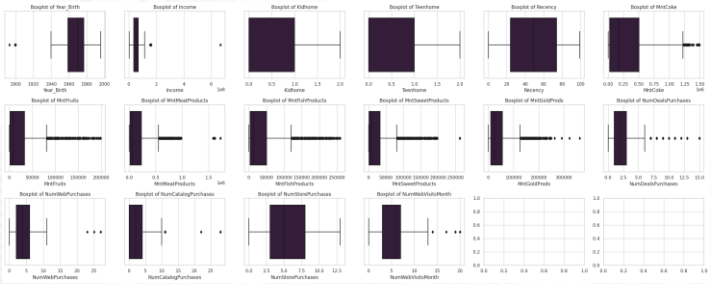
Benefits of Data Cleansing
- Enhances data quality and accuracy.
- Streamlines the dataset, improving its relevance.
- Reduces the risk of errors in analysis and modeling.
- Enhances computational efficiency.

These data cleansing actions align with best practices for ensuring that our dataset is well-prepared for further exploration and modeling.

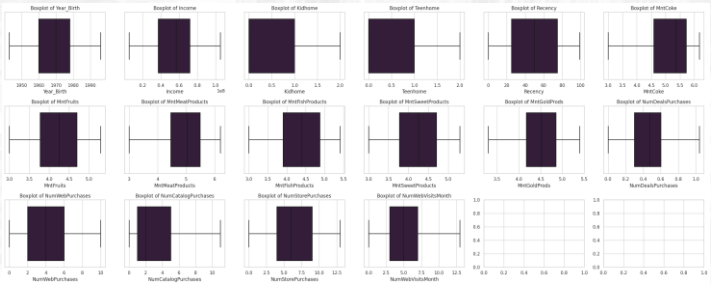
Handling Outlier

Rakamin Academy

Before the preprocessing



After the preprocessing



The next step data pre-processing journey has uncovered some critical observations regarding outliers in specific columns. Let's break down the key insights:

Outlier Identification: 🚫🔍

- We've identified outliers in multiple columns, including **MntGoldProds**
- , **MntGoldProds**, and various spending categories (e.g.**., MntCoke, MntFruits**).
- Some of these outliers are exceptionally extreme, such as an Income above 600M and MntCoke spending above 1.2M.

Outlier Handling: 🔄📈

- To address these outliers, we applied a log transformation to the data.
- As a result, we were able to mitigate the impact of extreme values, which is particularly crucial due to our relatively small dataset (2240 rows).
- This step aligns with best practices for handling outliers, as it avoids the need for data deletion, preserving valuable information.

Further Steps: 📜🖌️

- For columns with a large number of remaining outliers, we may consider additional methods or transformations to better understand and manage these data points.
- Removing outliers based on Interquartile Range (IQR) or Z-score could be a next step to enhance data quality.

Standardization & Encoding

Rakamin Academy

Feature Standardization & Encoding 📏🔤

Now that i've successfully managed outliers in our data, let's delve into the next crucial steps in our data preprocessing journey: feature standardization and encoding. 🧱🔄

1. Feature Standardization: 📏📐

Standardization is the process of transforming our numeric features to have a mean of 0 and a standard deviation of 1. This step is essential for ensuring that all numeric features have the same scale, preventing any single feature from dominating the analysis or modeling process.

2. Feature Encoding: 🔤🔢

Feature encoding allows us to work with categorical data effectively. Here's what i've done:

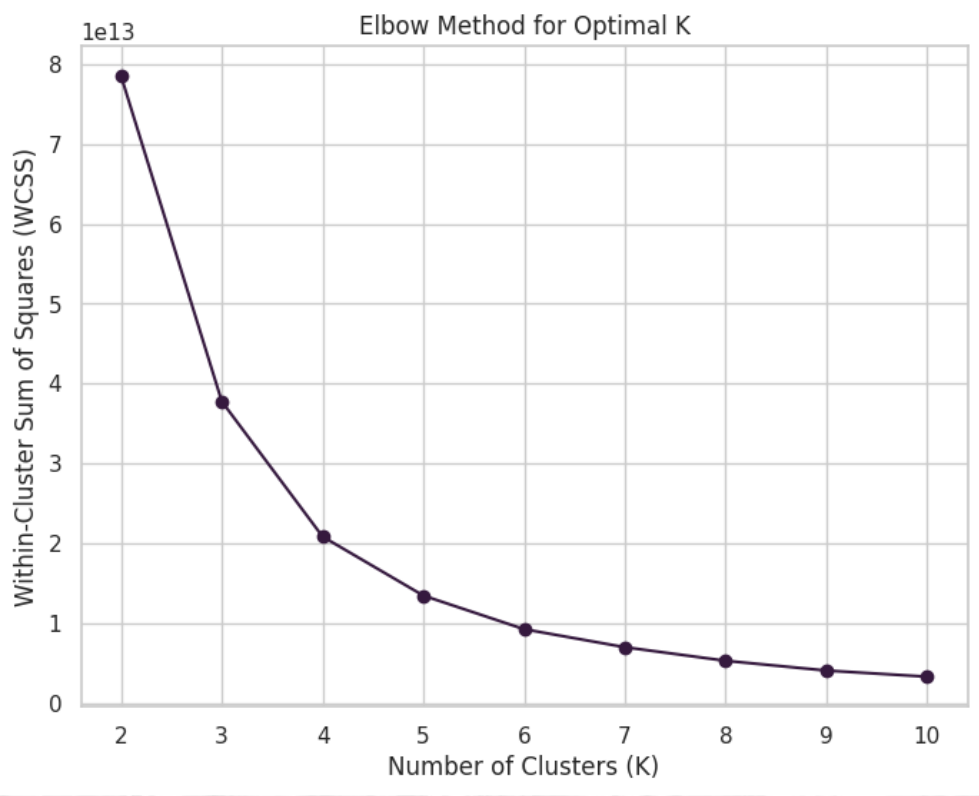
- I applied One-Hot Encoding to the **“Marital_Status"** columns. This technique converts categorical variables into binary vectors, making them suitable for machine learning models.
- For the **"Education"** column, we performed Label Encoding. This method assigns unique numerical values to each category, preserving the ordinal relationships in the data.

For more information about the project, click here

Data Modeling

Elbow Plot

- In our quest for the ideal number of clusters, we conducted an elbow plot analysis.
- The plot revealed a distinct **"elbow point"** at **K=4**, suggesting that four clusters were a suitable choice for our customer segmentation.
- This decision allows us to strike the right balance between granularity in understanding customer behavior and practicality in tailoring marketing campaigns.



For more information about the project, click here

Data Modeling

Rakamin Academy

Silhoutte Score

- As we examined the Elbow Method plot, we observed a significant "elbow" forming at 4 clusters. This suggests that 4 clusters is an appropriate choice, balancing complexity and meaningful segmentation.
- Further validating our choice, the Silhouette Score for our clustering was an impressive **0.62**, reinforcing the quality and separation of the clusters we've created.
- Let's move forward and explore how these clusters can power our targeted marketing strategies and elevate customer engagement.

- Summary:
- The Elbow Method plot indicated a clear "elbow" at 4 clusters, implying this is an optimal choice for segmentation.
- The Silhouette Score of **0.62** confirmed the quality of our clustering and the meaningfulness of the resulting clusters.

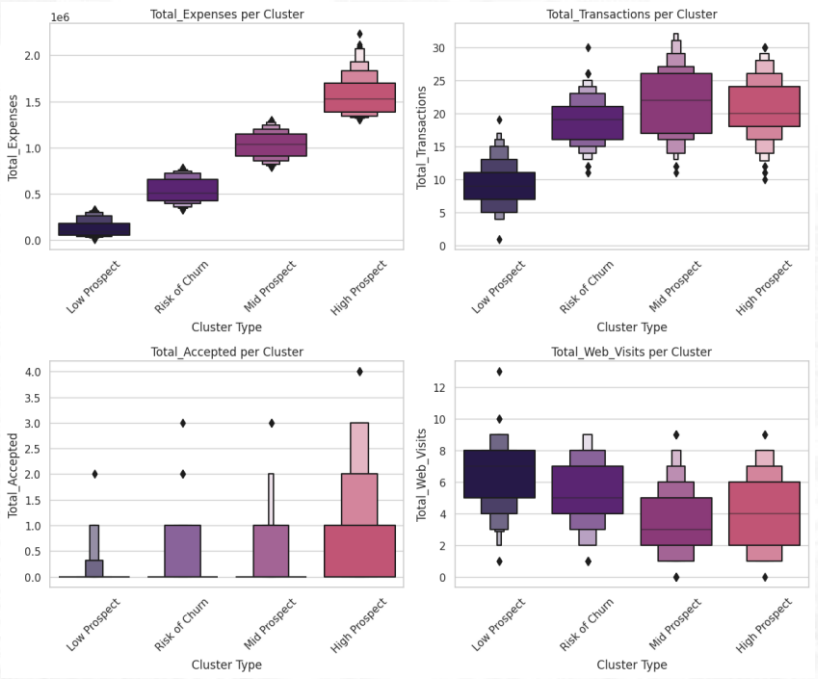
For more information about the project, click here

Customer Personality Analysis for Marketing Retargeting

Rakamin Academy

Cluster Summary Based on The Plot

- **low Prospect** : they were very low at spending and buying something on their money to actually buy one of the product, but yet the had the most total clicked on the website.
- **Risk of Churn** : this cluster is a mid spender and buyer on the event and has second of the most clicked total on the website.
- **Mid Prospect** : this is the trusty prospect to actually interested on spending their money on the campaign even though they had very low clicked total on the website.
- **High Prospect** : this is the most guarantee prospect we had, they had very high expenses and transactions of all of this cluster, even though just like the Mid Prospect they had very low clicked total, but when they do clicked it's a guarantee that they will accepted



For more information about the project, click here

Customer Personality Analysis for Marketing Retargeting

Rakamin Academy

Business Recommendation 🚀
Based on the clustering analysis of customer segments, we propose the following recommendations for optimizing the marketing campaign:

Segment-Specific Targeting 🎯:

- Tailor marketing efforts to each customer segment's unique characteristics.
- For the "Low Prospect" segment, focus on converting high website engagement into actual purchases by providing compelling incentives. 💰
- The "Risk of Churn" segment should be targeted with personalized offers and retention strategies to prevent them from exploring alternatives. 🌓
- "Mid Prospect" customers represent a potential growth opportunity. Enhance their website experience and encourage purchases through user-friendly interfaces. 📈
- Prioritize the "High Prospect" group by offering exclusive and high-value products or services, as they have shown strong purchase potential. 🌟

Conversion Rate Optimization 📊:

- Implement data-driven strategies to improve conversion rates for each segment.
- A/B testing and targeted promotions can be used to refine conversion funnels for different clusters.
- For the "Low Prospect" and "Risk of Churn" groups, consider implementing retargeting campaigns to re-engage potential customers who have shown interest. 🔄

Customer Engagement Enhancement 💬:

- Develop loyalty programs and incentives tailored to the "High Prospect" group to increase their loyalty and lifetime value. 🏆
- Use data analytics to identify high-value products or services for this segment.
- By implementing these recommendations, we can build stronger customer relationships, improve conversion rates, and increase revenue. Our success will depend on the ability to adapt strategies in real-time and create exceptional experiences for all customers. Let's embark on this journey to drive growth and satisfaction across all clusters! 🌟🚀💰

Customer Personality Analysis for Marketing Retargeting

Rakamin Academy

Calculating The Impact

The analysis of the financial impact of total expenses by cluster unveils critical insights that can guide our marketing strategy. Here's what we've uncovered:

Mid Prospect Dominance

- The "Mid Prospect" cluster stands out with a total expense impact of **$285,418,000.00**. Their willingness to spend is a valuable asset, even though they have low website engagement.
- This segment represents a significant revenue opportunity. By enhancing their website experience and offering personalized incentives, we can tap into their potential.

High Prospect Potential

- The "High Prospect" cluster showcases immense promise with an impact of **$135,285,000.00**. These customers have high expenses and transactions.
- Focusing on retaining their loyalty and providing exclusive, high-value products can further boost revenue.

Risk of Churn Challenge

- The "Risk of Churn" cluster, while not the highest spender, still contributes significantly with an impact of **$243,629,000.00**.
- Their moderate spending and high website engagement suggest that retaining their interest is crucial. Implementing personalized retention strategies can be a game-changer.

Low Prospect Engagement Opportunity

- "Low Prospect" customers, with a total expense impact of **$66,891,000.00**, may not be big spenders, but their extensive website engagement is noteworthy.
- With targeted efforts and enticing incentives, we can turn this high engagement into increased sales.
- In summary, this analysis provides a clear roadmap for optimizing our marketing strategy. By segmenting our approach, we can unlock the potential of each customer group, boosting revenue and ensuring long-term success. The financial impact of each segment emphasizes the value they bring to our business and underscores the importance of tailored engagement strategies.

Thanks for participating