Contents

[**Customer Segmentation** 2](#_Toc160992822)

[**Context:** 2](#_Toc160992823)

[**About the Dataset:** 2](#_Toc160992824)

[**Summary Statistics:** 2](#_Toc160992825)

[**Handling Missing Values:** 2](#_Toc160992826)

[**Distribution of Numerical Features:** 2](#_Toc160992827)

[**Analysis of Categorical Variables:** 2](#_Toc160992828)

[**Relationships between Variables:** 2](#_Toc160992829)

[**Correlation Analysis:** 3](#_Toc160992830)

[**Feature Engineering:** 3](#_Toc160992831)

[**Machine Learning Models:** 3](#_Toc160992832)

[**Data Visualization:** 3](#_Toc160992833)

[**Segmentation Analysis:** 3](#_Toc160992834)

[**Conclusion:** 3](#_Toc160992835)

**Customer Segmentation**

**Context:**

Customer Personality Analysis involves a thorough analysis of our company's ideal customers. It aids us in better understanding our customers, allowing us to adjust our products according to their specific needs, behaviors, and concerns. By segmenting customers into groups based on similarities, we can tailor our products to meet the specific needs and behaviors of each group.

**About the Dataset:**

Our dataset consists of 29 columns and 2240 entries. It includes various features such as ID, year of birth, education, marital status, income, number of children at home, customer recency, and spending amounts on different products. There are also columns indicating the number of deals, web, catalog, and store purchases, as well as responses to marketing campaigns.

**Summary Statistics:**

The dataset contains 2240 observations. Most variables have no missing values, except for the 'Income' column, which has been imputed. The mean income of customers is approximately $52,247, with a standard deviation of $25,037, indicating some variability in income levels. The majority of customers have no children at home. Customers have varying levels of spending on different product categories. Other variables, such as accepted campaigns, complaints, and contact costs, have binary or constant values across observations.

**Handling Missing Values:**

Missing values in the 'Income' column have been filled with the mean income value.

**Distribution of Numerical Features:**

We visualized the distribution of numerical features through histograms, revealing insights into the distribution of data for each feature.

**Analysis of Categorical Variables:**

We examined the distribution of categorical variables using count plots, providing insights into the distribution of different categories within each variable.

**Relationships between Variables:**

Scatter plots were used to visualize relationships between variables, such as income vs. wine spending, and income vs. wine spending with age. Outliers were detected and removed before plotting to ensure accurate visualization of relationships.

**Correlation Analysis:**

A heatmap was used to visualize the correlation matrix of numerical features, providing insights into the relationships between different variables.

**Feature Engineering:**

We calculated the age of customers from their year of birth and total spending by summing up spending columns, introducing new features for analysis.

**Machine Learning Models:**

We trained a linear regression model to predict income based on customer features, achieving a mean squared error of approximately 97681117.38 on the test set.

**Data Visualization:**

Pair plots were created to visualize relationships between key variables, providing insights into the distribution and relationships between variables.

**Segmentation Analysis:**

We performed K-means clustering to segment customers based on their features, visualizing the segments using scatter plots. Outliers were detected and removed before segmentation analysis. Three distinct segments were identified based on income and total spending.

**Conclusion:**

* The dataset provides valuable insights into customer demographics, behaviors, and spending patterns.
* Relationships between variables, such as income and spending, are visualized, helping us understand how different factors are related.
* Segmentation analysis reveals distinct customer segments based on income and spending behavior, allowing us to tailor our marketing strategies and product offerings to better meet the needs of each segment.
* Machine learning models provide predictive capabilities for customer income, enabling us to make informed decisions based on customer data.