**Project Report**

**Project Title**

**Customer Intent Prediction & Pattern Discovery for e-Commerce businesses**

**Team Leader:**

Shivanoor Vignesh 21R11A0597

**Team Members:**

G.Sandhya Rani 21R11A0569

Dhathri 21R11A0545

Hasini 21R11A0518

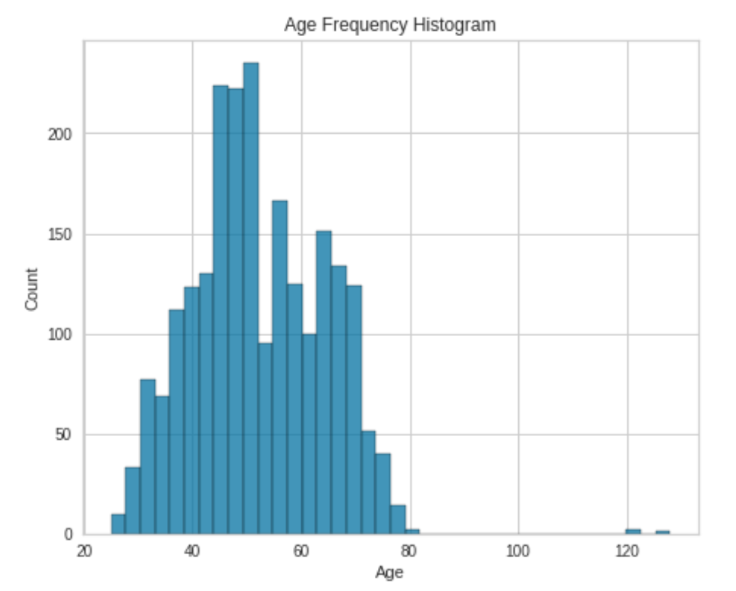
Pooja 21R11A0501

PSBN Sriya 21R11A0595

N Sai Charan 21R11A0588

MD.Akbar Ali 21R11A0586

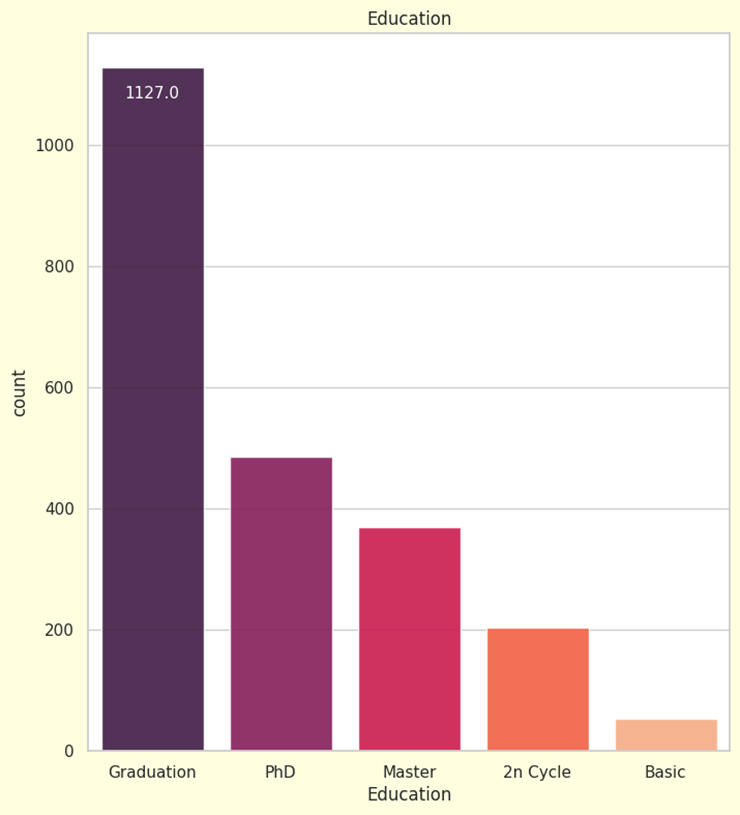
**EXPLORATORY DATA ANALYSIS**

* **Age distribution**

**Observation:**

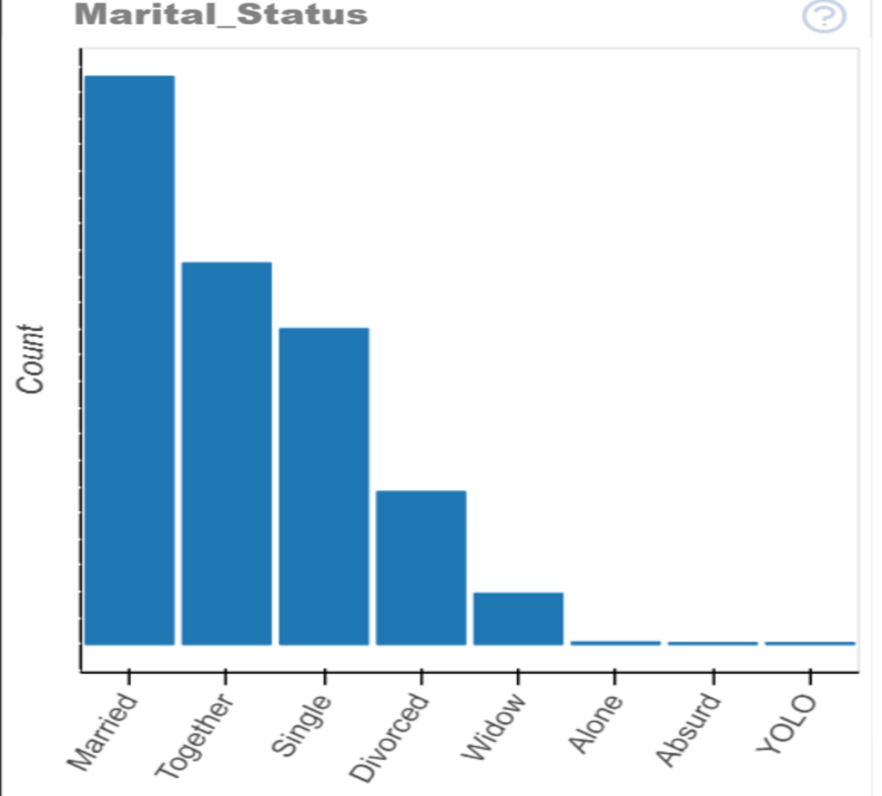
The histogram shows two prominent peaks:

* The first peak corresponds to customers aged around **40-50**, indicating a substantial user base within this age range.
* The second, smaller peak is around **60-70**, suggesting another segment of older customers.
* **Education of customers**



**Observation:**

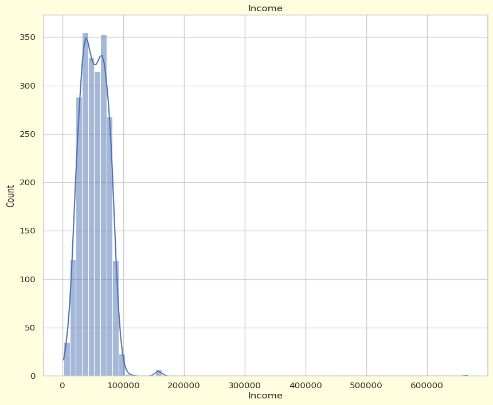
**Education Levels**:

* The graph illustrates the distribution of customers based on their educational levels.
* The majority of customers fall into the **“Graduation”** category, which suggests that the company’s products or services resonate well with individuals who have completed their undergraduate education.
* However, there is a significant drop in customer counts for those with **PhD** and **Master’s** degrees. Additionally, very few customers fall into the **“2nd Cycle”** or **“Basic”** education categories.
* **Marital status of the customers**

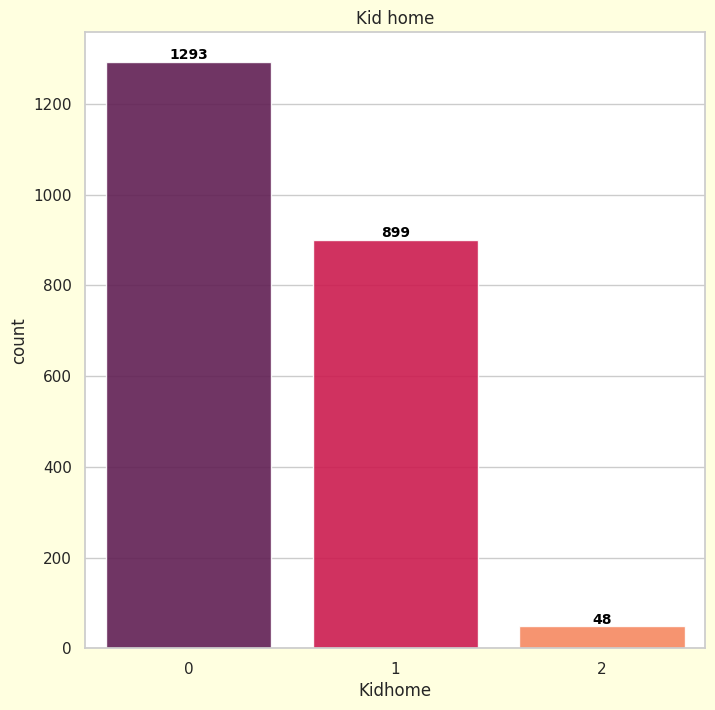
**Observation:**

* The graph represents the distribution of individuals based on their marital status.
* The highest count is in the **“Married”** category, followed by **“Together”** and **“Single”**.
* There is a significant drop in count as we move to categories like **“Divorced”**, **“Widow/Widower”**, **“Alone”**, **“Absurd”**, and **“YOLO”**.\

**Recommendation**: The company should focus its marketing strategies primarily on the **“Married”** and **“Together”** segments, as they constitute the largest customer base.

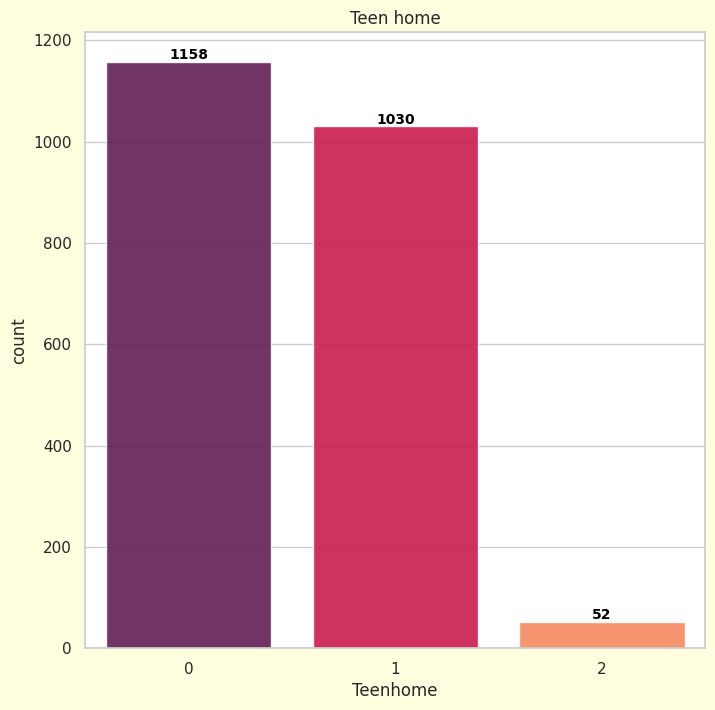
* **Income of customers**

**Observation:**

* The graph depicts income distribution within a specific population or dataset.
* Most data points cluster on the left side, indicating a significant portion of the population with incomes below approximately 100,000.
* Few data points exist at higher incomes, suggesting an exponential decline in count as income increases.
* The company’s ideal customer base primarily falls within the lower to middle-income bracket.
* Key Observations:
  + **Target Audience**: Focus efforts on understanding and meeting the needs of customers in the lower to middle-income range.
  + **Affordability**: Develop affordable products that address specific concerns of this income segment.
  + **Market Segmentation**: Optimize marketing resources by analyzing which customer segment is likely to buy the product.
  + **Behavior Over Words**: Observe actual behavior (purchase patterns, browsing history, engagement metrics).
  + **Customer Attitude**: Gather direct feedback from customers through surveys, reviews, and social media interactions.
* **No. of Kids and children at home**

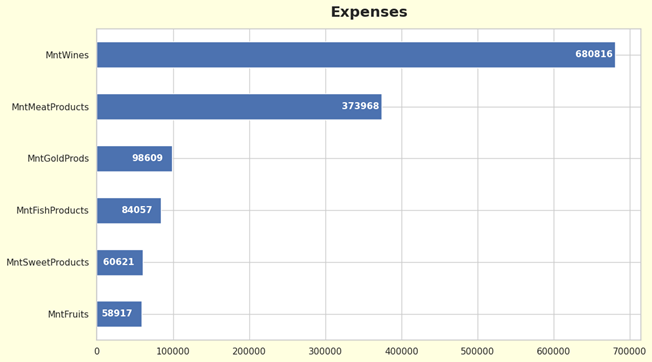
**Observation:**

* **No Kids at Home (0)**: Largest segment (1293 customers) without children. Their priorities and spending habits may differ.
  + **One Kid at Home (1)**: Second-largest segment (899 customers) representing a significant portion of the customer base.
  + **Two Kids at Home (2)**: Smallest segment (48 customers) but still relevant.
* **Number of teenagers in customer’s household.**



**Observation:**

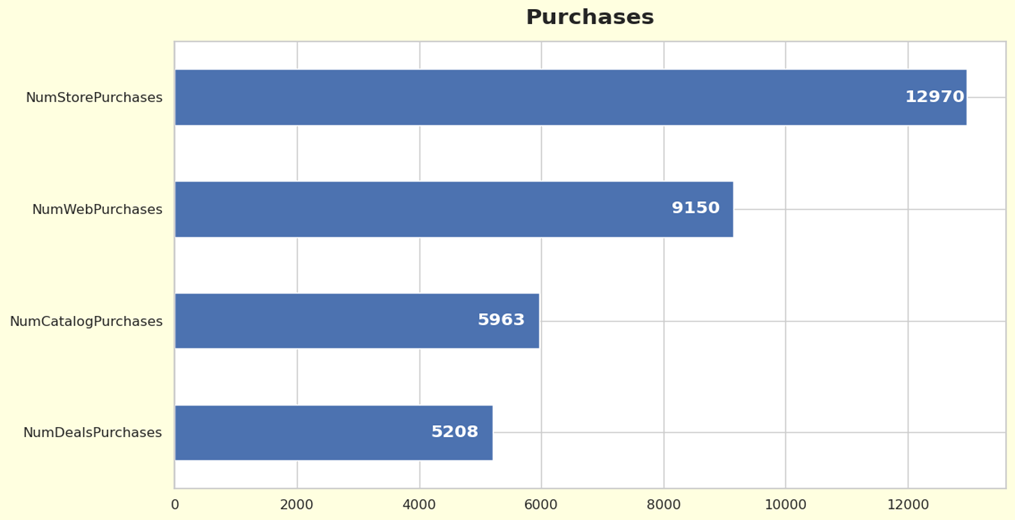
* + The graph depicts the distribution of the number of teenagers in customers’ households.
  + Three categories exist: ‘0’, ‘1’, and ‘2’ for “Teenhome.”
  + The majority of customers have no teenagers (1158 count for category ‘0’).
  + The second-largest segment comprises households with one teenager (1030 count for category ‘1’).
  + A small number of customers have two teenagers (52 count for category ‘2’).
* **EXPENSES AND PURCHASES MADE BY THE CUSTOMERS:**

**Expenses:-**

**Observation:**

* + The graph represents expenses across different product categories.
  + **Wines** have the highest expense, followed by **Meat Products**. These two categories dominate the expenditure.
  + **Gold Products**, **Fish Products**, **Sweet Products**, and **Fruits** have relatively lower expenses.

**Purchases:-**



**Observation:**

1. **In-Store Purchases:**

* **Dominant Channel**: In-store purchases are the most preferred method. It’s crucial to maintain the quality of in-store experiences.
* **Customer Engagement**: Leverage in-store interactions for upselling, cross-selling, and personalized recommendations.

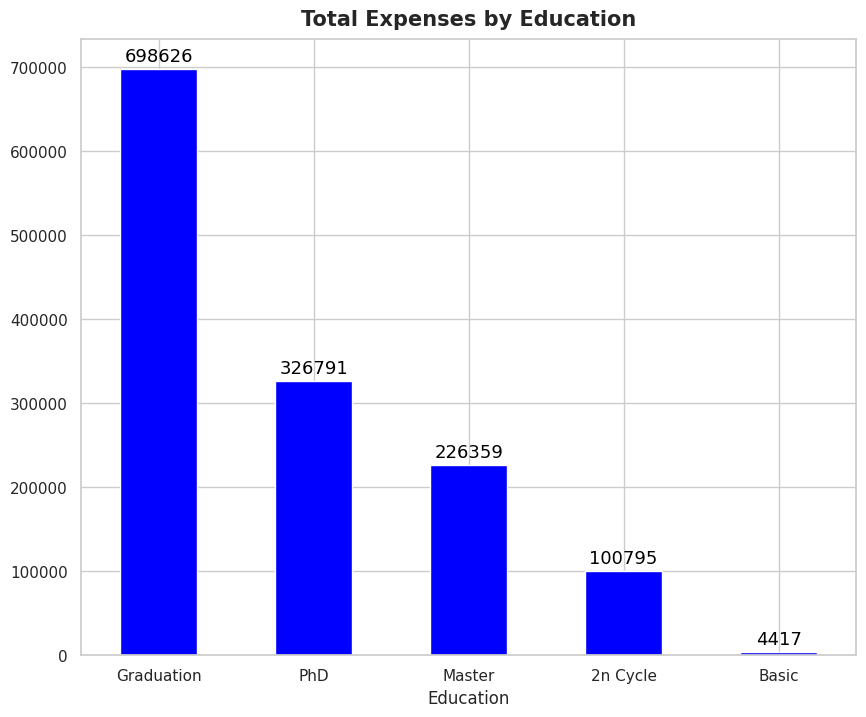
1. **Web Purchases:**

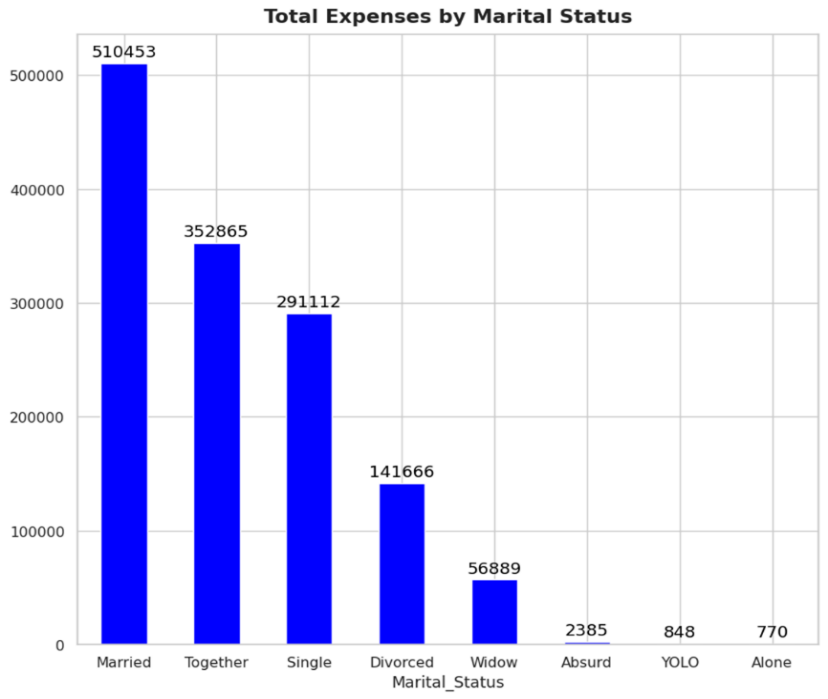
* **Growing Potential**: Web purchases are the second most popular channel. Enhance the online shopping experience to boost web sales.
* **Digital Marketing**: Invest in targeted digital marketing campaigns to attract and retain web customers.

1. **Catalog Purchases:**
   * **Underutilized**: Catalog purchases are relatively low. Investigate reasons behind this and explore ways to improve.
   * **Segment-Specific**: Analyze which customer segments respond well to catalogs and tailor accordingly.
2. **Deals Purchases:**
   * **Optimization Opportunity**: Deals purchases can be optimized. Evaluate deal effectiveness and adjust strategies.

* **Total Expenses by Education:-**

**Observation:**

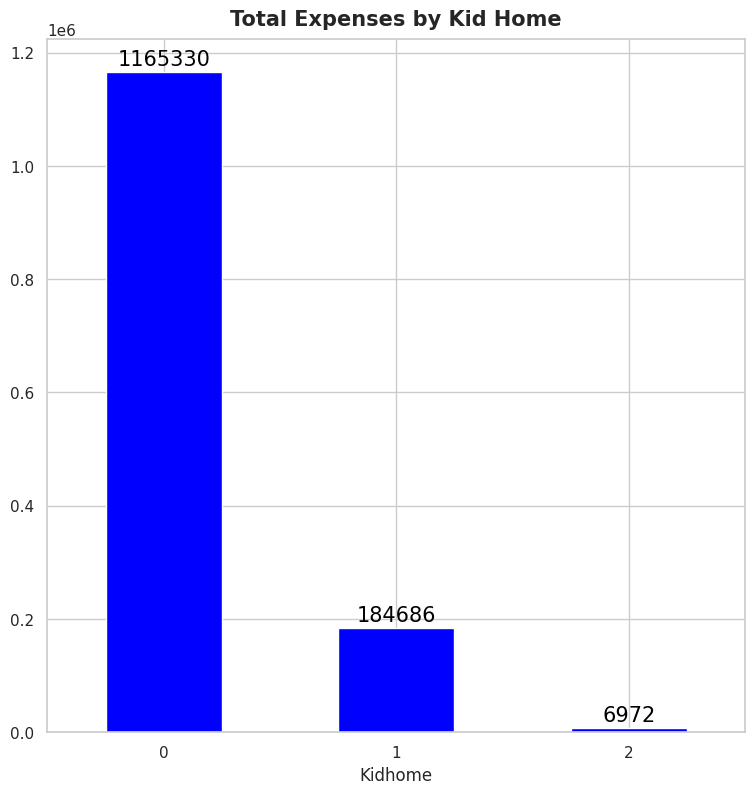
* The graph titled “Total Expenses by Education” reveals a clear correlation between **education level** and **total expenses**.
* Customers with a **graduation degree** exhibit the highest total expenses, followed by those with **PhDs**, **Master’s education**, **2nd cycle**, and **basic education**.
* Specifically:
  + **Graduation**: Highest total expenses at **698,626**.
  + **PhD**: Total expenses at **326,791**.
  + **Master’s education**: Total expenses at **226,359**.
  + **2nd cycle**: Total expenses at **100,795**.
  + **Basic education**: Lowest expenses at **4,417**.
* This suggests that people with **higher educational qualifications** tend to spend more. It could be due to a higher income level or different spending habits.

**Total expenses by Marital Status:**

**Observation:**

1. **Marital Status and Expenses**:

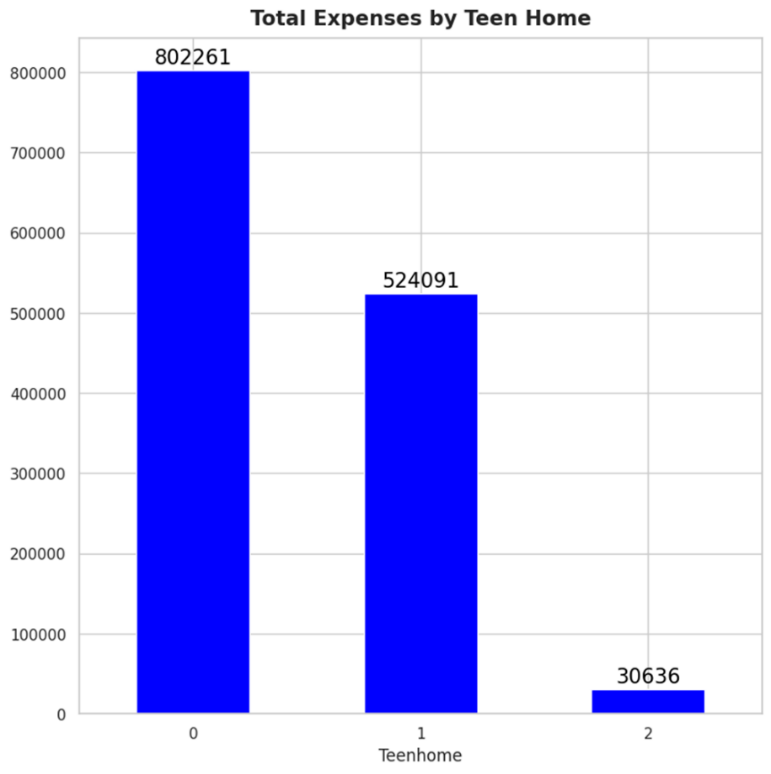
* The graph illustrates total expenses categorized by marital status.
* **Married individuals** have the highest expenses, followed by those who are **together**, **single**, **divorced**, and **widows**.
* Notably, there is a **drastic drop** in expenses for categories like **“Absurd,” “YOLO,” and “Alone.”** These categories might represent non-traditional or humorous self-descriptions of marital status.

**Total expenses by kids:**

**Observation:**

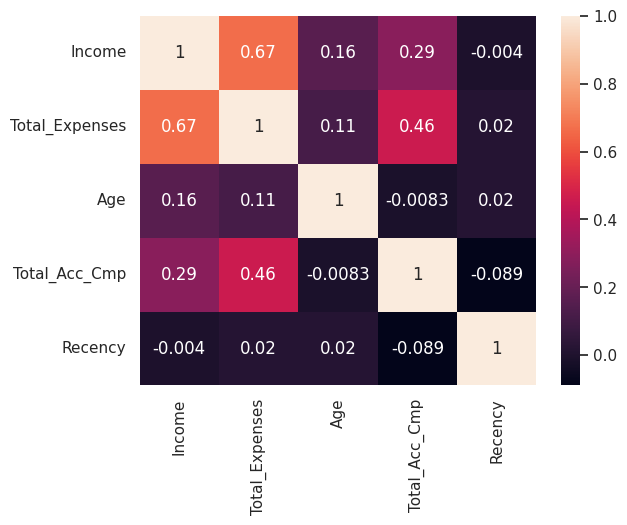
* Customers with **no kids at home** have the highest expenses, totaling **1,165,330**.
* As the number of kids at home increases:

1. For **1 kid**, the expense is **184,686**.
2. For **2 kids**, the expense drops significantly to **6,972**.

**Total expenses by teen:**

**Observation:**

Households without teenagers, denoted as **Teen home 0**, incur the highest expenses, suggesting they might have greater disposable income or distinct spending habits. On the other hand, households with one or two teenagers, labeled as **Teen home 1** and **Teen home 2**, show reduced expenses, indicating that teenagers’ presence likely influences budgeting and spending priorities within these families.

* **Heatmap**

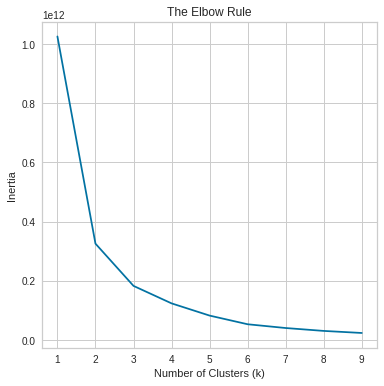
**Observation:**

In the dataset, **Income** and **Total Expenses** share a strong positive correlation of **0.67**, indicating that customers with higher incomes also have higher expenses on items like wine and meats, making them valuable for revenue. **Age** correlates weakly with **Income** (**0.16**) and **Total Expenses** (**0.11**), implying it’s not a strong predictor of economic behavior. **Total\_Acc\_Cmp** shows moderate positive correlations with **Income** (**0.29**) and **Total Expenses** (**0.46**), suggesting those who participate more in campaigns also spend more. Lastly, **Recency** has a slight negative correlation with **Total\_Acc\_Cmp** (**-0.089**), hinting that recent purchasers may be less campaign-responsive.

* **K-means clustering**

Clustering algorithm is used to determine the 'k' that will yield the most accurate results.

This is done using two popular methods:

1. The Elbow Method
2. The Silhouette Method

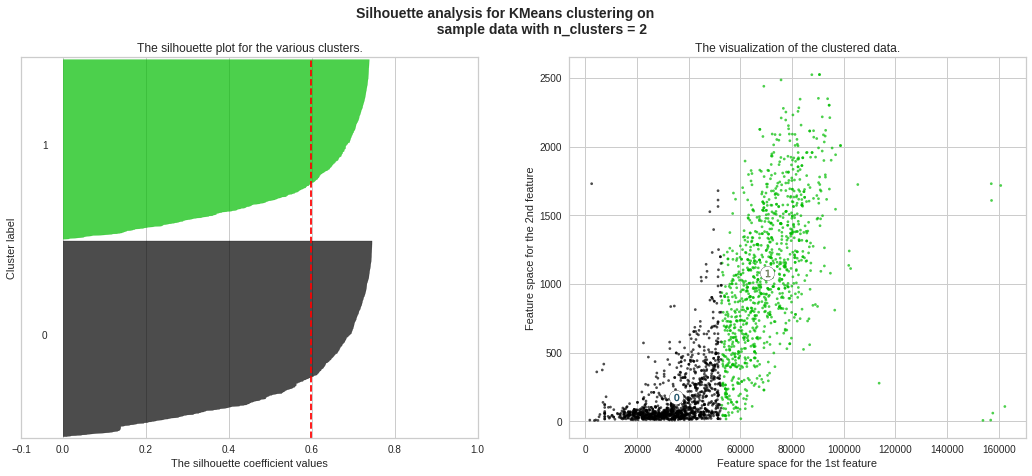
**Observation:**

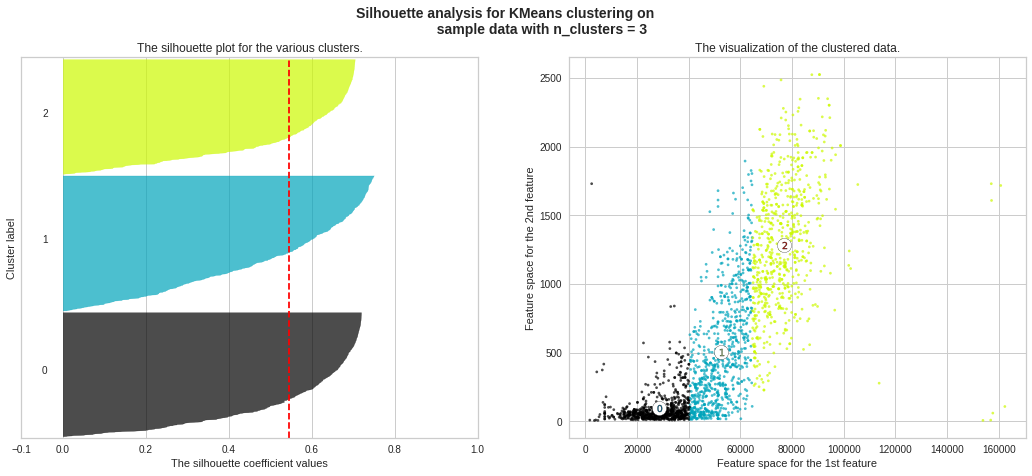
**In Elbow method:**

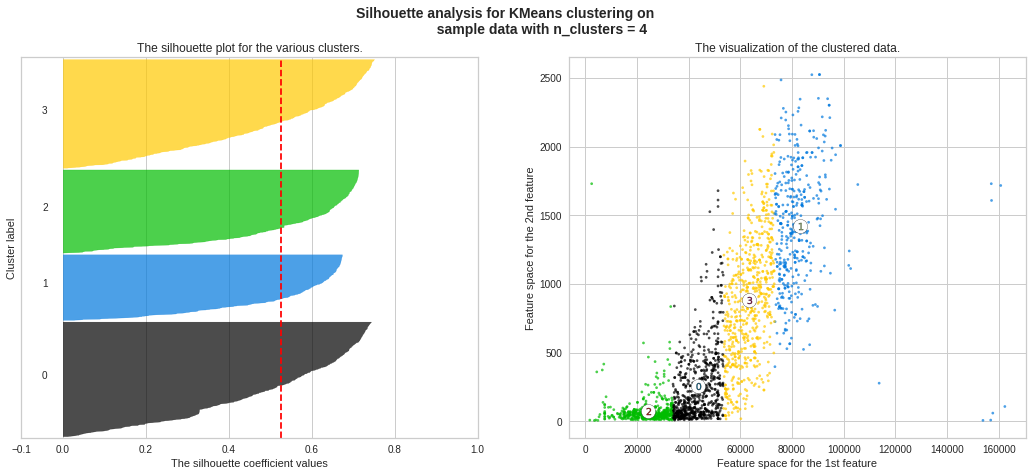
We do not obtain a sharp elbow point, as observed in the graph above. A good k value can be either 4 or 5. Further clarification can be obtained by performing the Silhouette Method.

* **Silhouette Method:**

Silhouette analysis for KMeans clustering on sample data with 2,3 and 4 clusters.





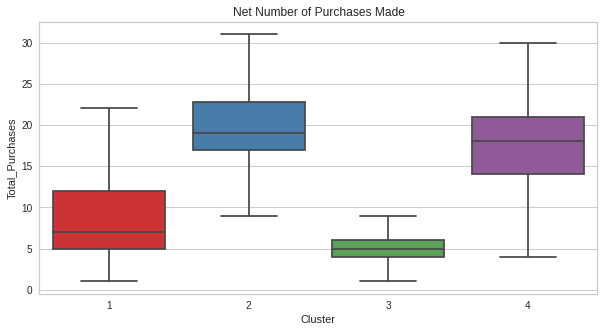


* **Analysis of Clustering Results**

# 

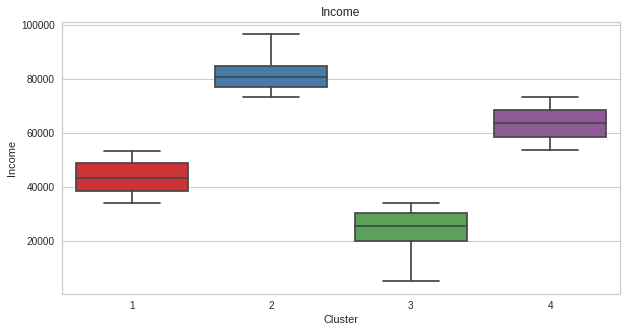
**Observation:**

Clusters 1 and 2 have a similar number of customers. However, Cluster 1 beats Cluster 2 marginally.



**Observation:**

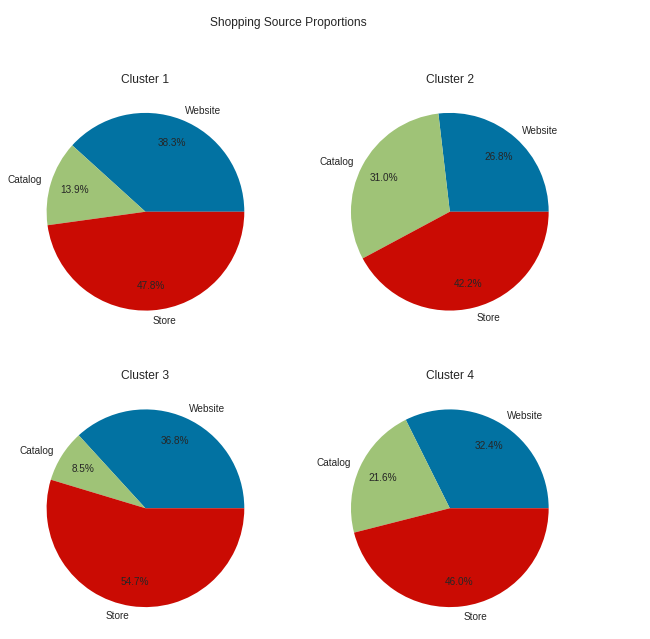
From the above box-whisker plots, we make the following inferences:

1. **Cluster 1**: Low amount of purchases
2. **Cluster 2:** Low amount of purchases
3. **Cluster 3:** Large amount of purchases
4. **Cluster 4:** Considerable amount of purchases.

**Observation:**

From the above box-whisker plots, we make the following inferences:

1. **Cluster 1:** Medium income
2. **Cluster 2:** Low Income
3. **Cluster 3:** High income
4. **Cluster 4:** Very high income



**Observation:**

All clusters make the maximum purchases at the store. Cluster 2 makes the maximum number of catalog purchases, possibly because catalogs are often released during the sale period.

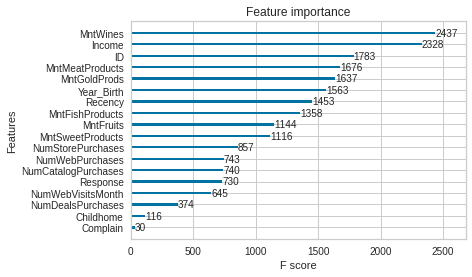
# ****KMeans Customer Analysis****

We can divide the customers into four categories:

1. **Platinum Customers: Cluster 4** (Highly educated, very high income, mostly buy from the store)
2. **Diamond Customers: Cluster 3** (Well educated, high income, mostly buy from the store)
3. **Gold Customers: Cluster 1** (Educated, medium income, highest website purchases)
4. **Silver Customers: Cluster 2** (Comaparatively lesser education, low income, highest catalog purchases)

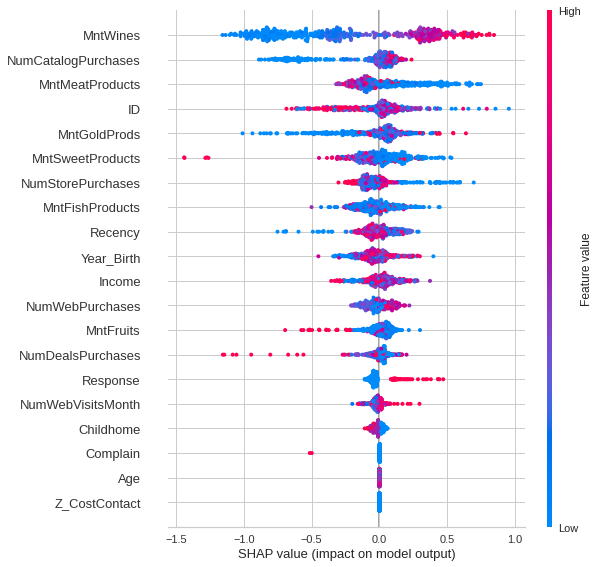
While online shopping gains traction, in-store purchases remain prevalent. Operating physical stores incurs significant expenses, including rent, electricity, staffing, and maintenance. In contrast, e-commerce operates at lower costs. To capitalize on this, the company should enhance its online presence and entice customers with attractive online shopping incentives.

# ****Feature Importance****

Using the XGBoost Classifier and the plot\_importance function Model was trained The feature-wise impact on the accuracy score has been detailed below and a decent accuracy score of 83.93% was achieved.

**Observation:**

From this visual, we are able to observe that the amount of money a customer spends on wine is a factor that hugely impacts the total number of campaigns the customer accepts. A close second factor is the income of the customer.

* **SHAP Analysis**

# 

# Analysis of SHAP Results

We can see that the amount of money a customer spends on wine is a huge determining factor of the campaigns they accept. It is possible that a lot of campaigns have been run for the alcoholic beverage department. It has been found earlier that families with children at home spend lesser money on wine - however, food items are a necessity. Running more campaigns on fruits and fish products will certainly be advantageous to more customers, thereby increasing profits.

Another noticeable observation is that the number of purchases made during deals has a very small impact on the total number of campaigns accepted. This may be an indication of lesser promotion of deals. By advertising deals, more customers will be interested in making purchases.

Finally, it would be helpful to increase the internet presence of the company, as the number of purchases made from the website impacts the accepted campaigns very less.