

# Predict Customer Personality to Boost Marketing Campaign by using Machine Learning



**Created by:**

**Aziz Prabowo**

[azizprabowo128@gmail.com](mailto:azizprabowo128@gmail.com)

[linkedin.com/in/aziz-prabowo](https://www.linkedin.com/in/aziz-prabowo)

A fresh graduate in Data Science with high interest artificial intelligence, data science, and business analytics. Experienced in data cleaning, exploratory data analysis, visualization, machine learning, and basic deep learning through academic, bootcamps, courses, and personal projects.

Supported by:  
**Rakamin Academy**  
Career Acceleration School  
[www.rakamin.com](http://www.rakamin.com)

## Background

A company can grow rapidly when it understands its customers' behavior, enabling it to provide better services and benefits to customers who have the potential to become loyal customers. By processing historical marketing campaign data to improve performance and target the right customers for transactions on the company's platform, we focus on creating a cluster prediction model based on this insight, making it easier for the company to make decisions.

## Goal

Understand and analyze customer behavior using historical marketing data to enhance service offerings and deliver targeted benefits that increase customer loyalty and company growth.

## Objective

Develop a cluster prediction model that segments customers based on behavioral patterns, enabling the company to make data-driven marketing and service decisions.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   ID                     2240 non-null   int64
1   Year_Birth             2240 non-null   int64
2   Education              2240 non-null   object
3   Marital_Status         2240 non-null   object
4   Income                 2216 non-null   float64
5   Kidhome                2240 non-null   int64
6   Teenhome               2240 non-null   int64
7   Dt_Customer            2240 non-null   object
8   Recency                2240 non-null   int64
9   MntCoke                2240 non-null   int64
10  MntFruits              2240 non-null   int64
11  MntMeatProducts        2240 non-null   int64
12  MntFishProducts        2240 non-null   int64
13  MntSweetProducts       2240 non-null   int64
14  MntGoldProds           2240 non-null   int64
15  NumDealsPurchases      2240 non-null   int64
16  NumWebPurchases        2240 non-null   int64
17  NumCatalogPurchases    2240 non-null   int64
18  NumStorePurchases      2240 non-null   int64
19  NumWebVisitsMonth       2240 non-null   int64
20  AcceptedCmp3           2240 non-null   int64
21  AcceptedCmp4           2240 non-null   int64
22  AcceptedCmp5           2240 non-null   int64
23  AcceptedCmp1           2240 non-null   int64
24  AcceptedCmp2           2240 non-null   int64
25  Complain               2240 non-null   int64
26  Z_CostContact           2240 non-null   int64
27  Z_Revenue              2240 non-null   int64
28  Response               2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
```

This dataset contains **customer data** from a marketing campaign, with **2,240 total entries** and **29 columns** detailing various customer attributes.

- **Demographics & Household Information:** Includes **ID**, **Year\_Birth**, **Education**, **Marital\_Status**, and the number of children and teenagers in the household (**Kidhome**, **Teenhome**).
- **Financial & Purchasing Behavior:** Features like **Income** (with 24 missing values), **Mnt...** (monetary spending on different products), and **Num...** (number of purchases across various channels like web, catalog, and store).
- **Marketing Response & Engagement:** Data points such as **Recency** (days since last purchase), **Dt\_Customer** (enrollment date), **AcceptedCmp...** (campaign acceptance status), **Complain**, and **Response**.

[Link to Jupyter Notebook](#)

## Before Imputation

```
[89]: # Cek null values di dataset
      df.isnull().sum()[df.isnull().sum() > 0]

[89]: Income      24
      dtype: int64
```

Imputation

## After Imputation

```
[90]: # impute missing values in 'Income' column with median
      df['Income'].fillna(df['Income'].median(), inplace=True)

[91]: # Cek null values di dataset
      df.isnull().sum()[df.isnull().sum() > 0]

[91]: Series([], dtype: int64)
```

Missing values were imputed in the **Income** column using the **median value**. This was chosen because these three columns have a highly skewed data distribution, making the median a more appropriate imputation value than the mean, which is more sensitive to outliers.



## Duplicate Values

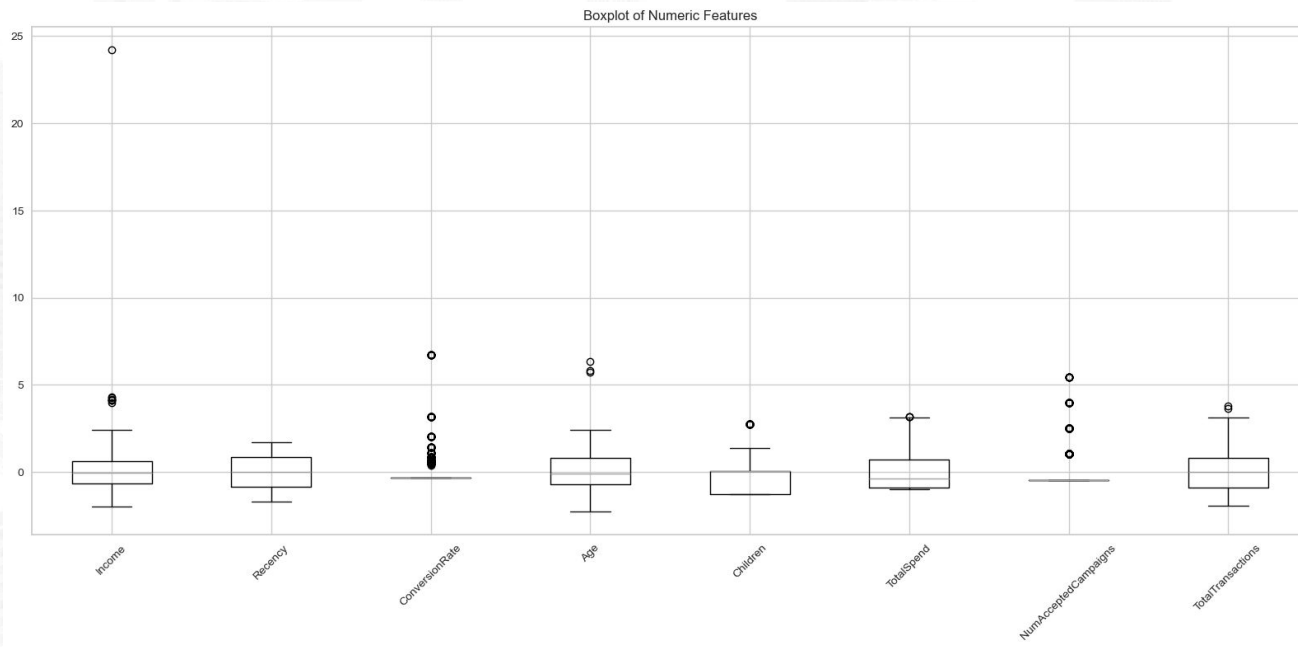
```
[212]: # check for duplicates  
df.duplicated().sum()
```

```
[212]: np.int64(183)
```

```
[213]: # Drop duplicates  
df.drop_duplicates(inplace=True, ignore_index=True)
```

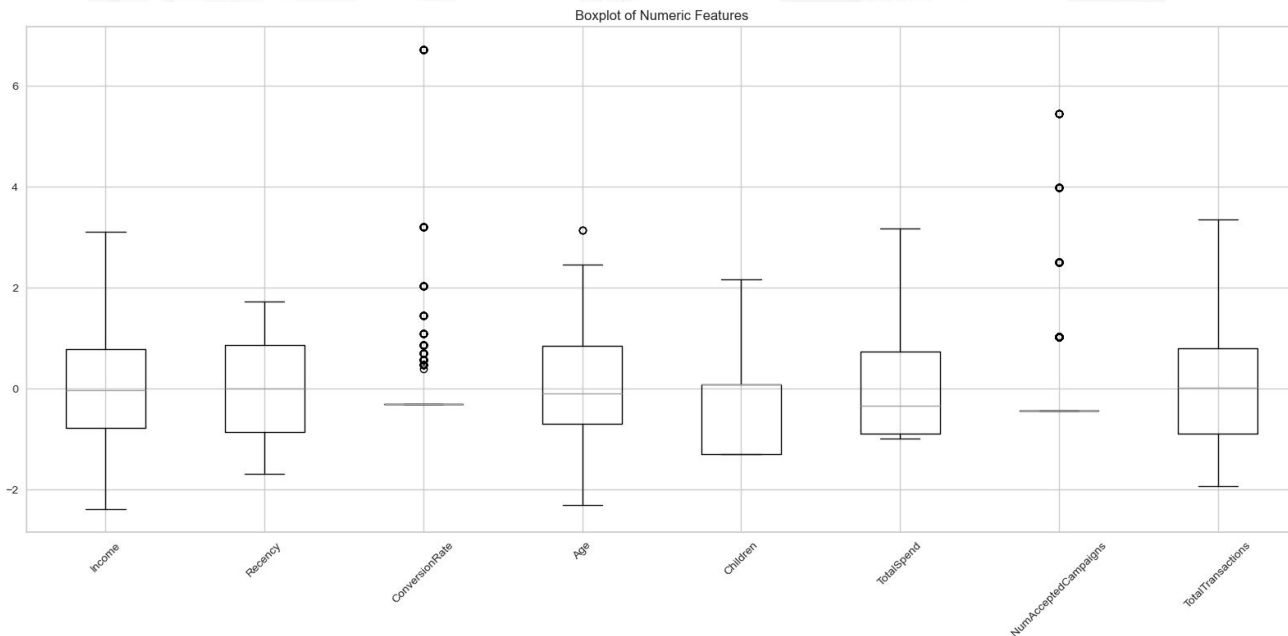
There are **183 duplicate rows** after the **ID** column was removed from the dataset. Therefore, these rows need to be removed from the dataset using the **drop\_duplicates** function.

## Before Outliers Handling



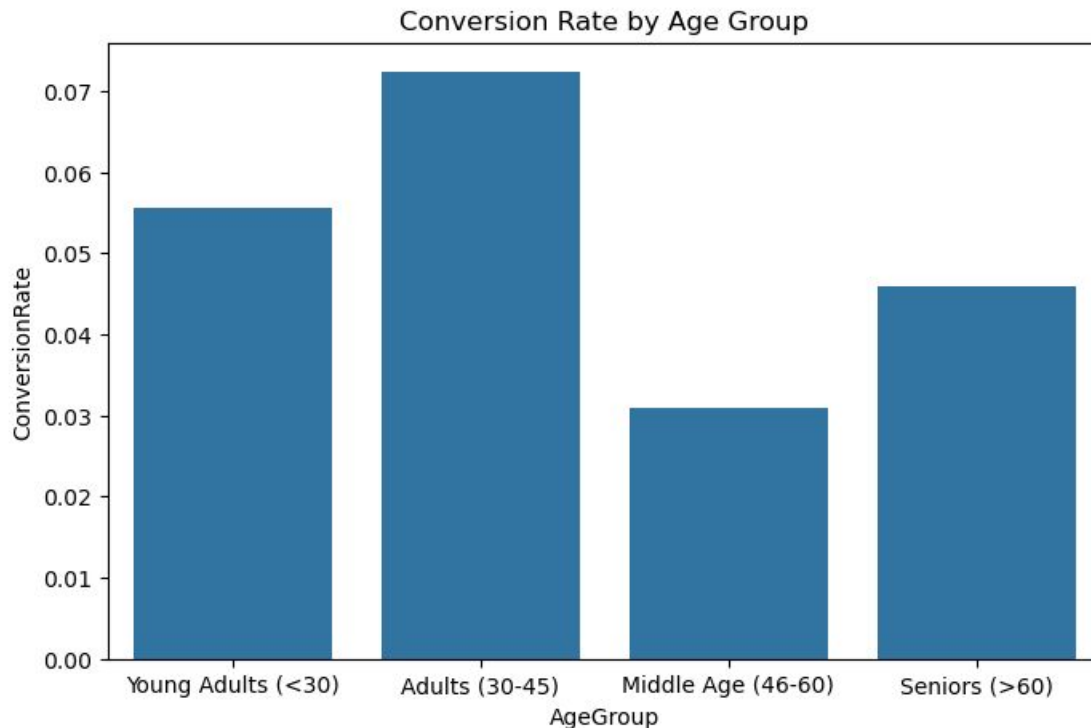
There are outlier values in almost all numeric columns in the dataset.

## After Outliers Handling



**Outliers are addressed by capping based on the IQR.** This is done to avoid missing the most valuable customer segments, which often fall outside the normal distribution.

## Younger Audiences Are More Likely to Convert

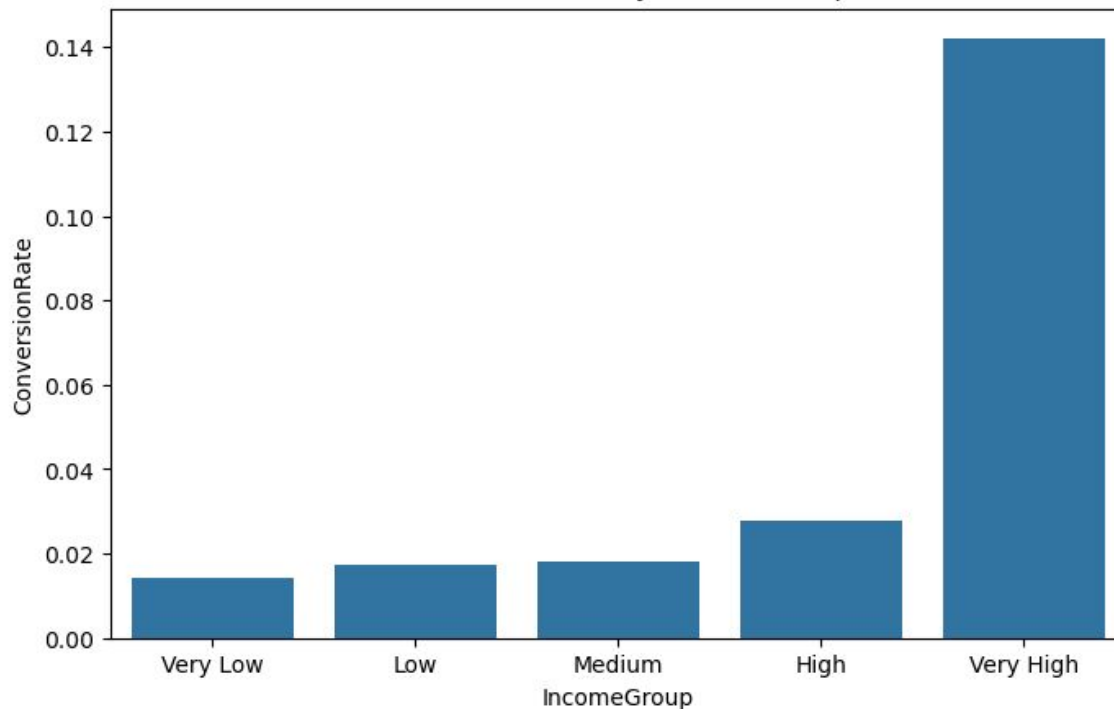


Customers in the **Young Adults (<30)** and **Adults (30–45)** age groups have a conversion rate of more than 5%, much higher than the Middle Age (46–60) and Seniors (>60) groups.



## High-Income Customers Drive Conversions

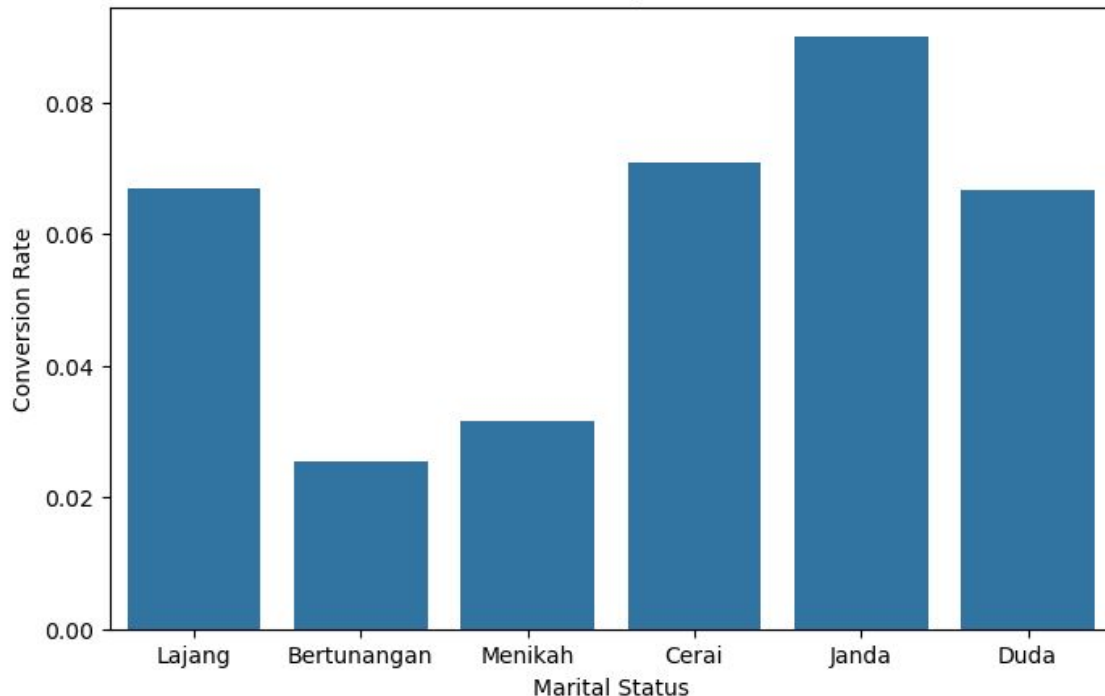
Conversion Rate by Income Group



Customers in the **Very High income group (>70 million)** showed the highest conversion rate, reaching **over 14%**. Conversely, customers in the High to Very Low income groups only showed a conversion rate of under 4%.

## Non-Marital Customers Convert More

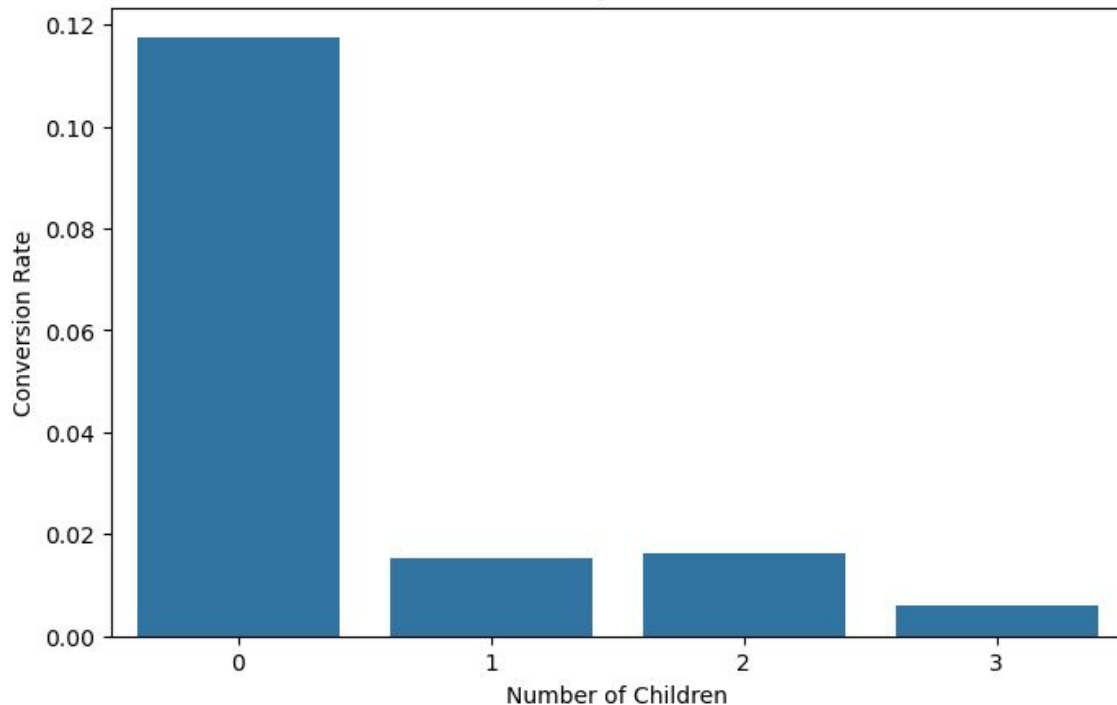
Conversion Rate by Marital Status



Customers who are not currently in a relationship—such as **single, widowed, widowed, or divorced**—show **higher conversion rates, above 6%**. Meanwhile, customers who are married or engaged have lower conversion rates, below 4%.

## Customers Without Children Convert Up to 5x More

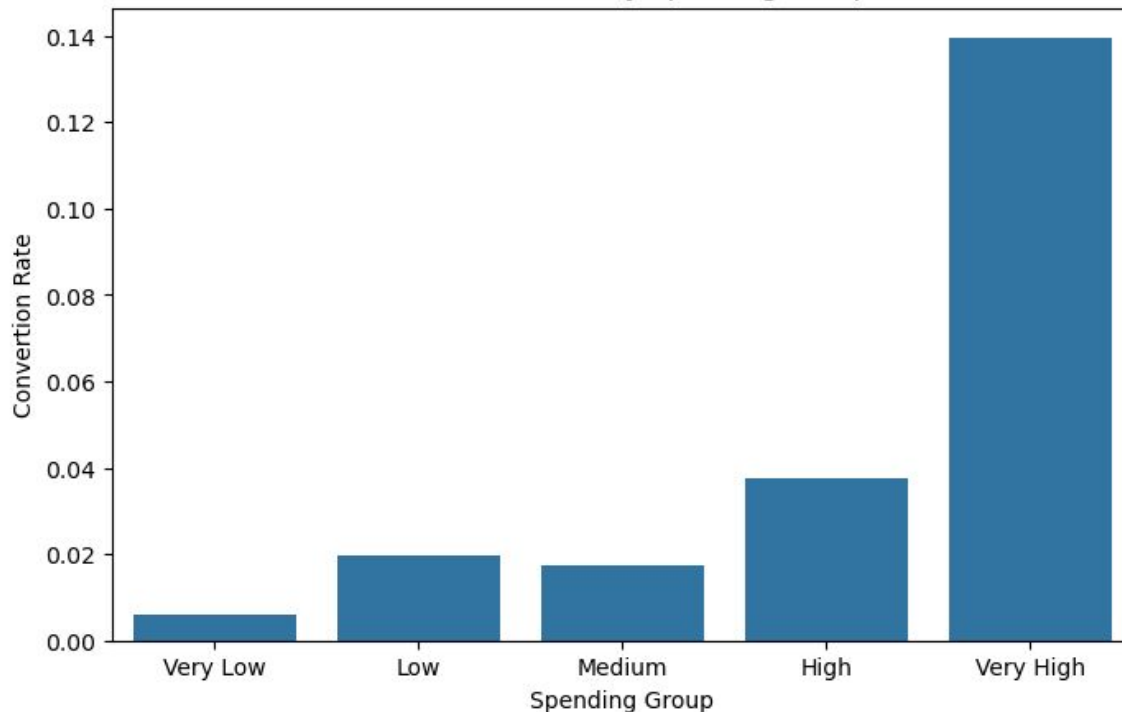
Conversion Rate by Number of Children



**Customers without children showed the highest conversion rates, above 10%.** Conversely, customers with one to three children only showed conversion rates below 2%.

## Top Spenders Are 3x More Likely to Convert

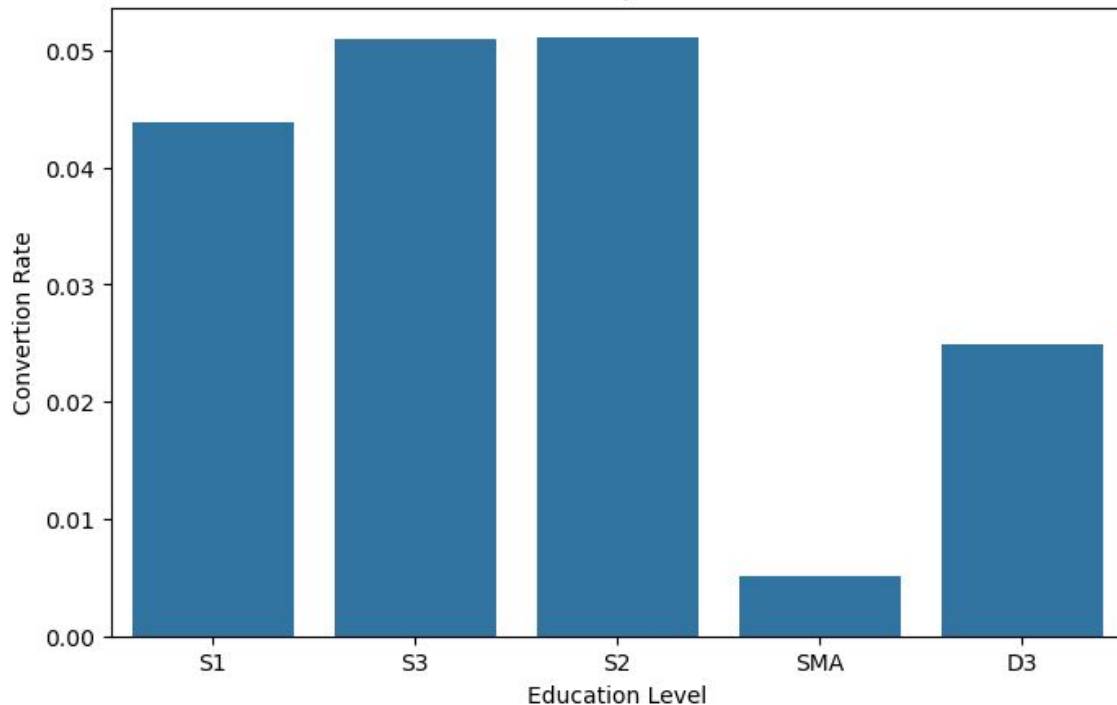
Conversion Rate by Spending Group



Customers **spending over 1 million rupiah** showed the **highest conversion rate, exceeding 12%**. Conversely, customers with lower spending—from Very Low to High—had a conversion rate of just under 4%.

## Higher Education, Higher Conversion

Conversion Rate by Education Level

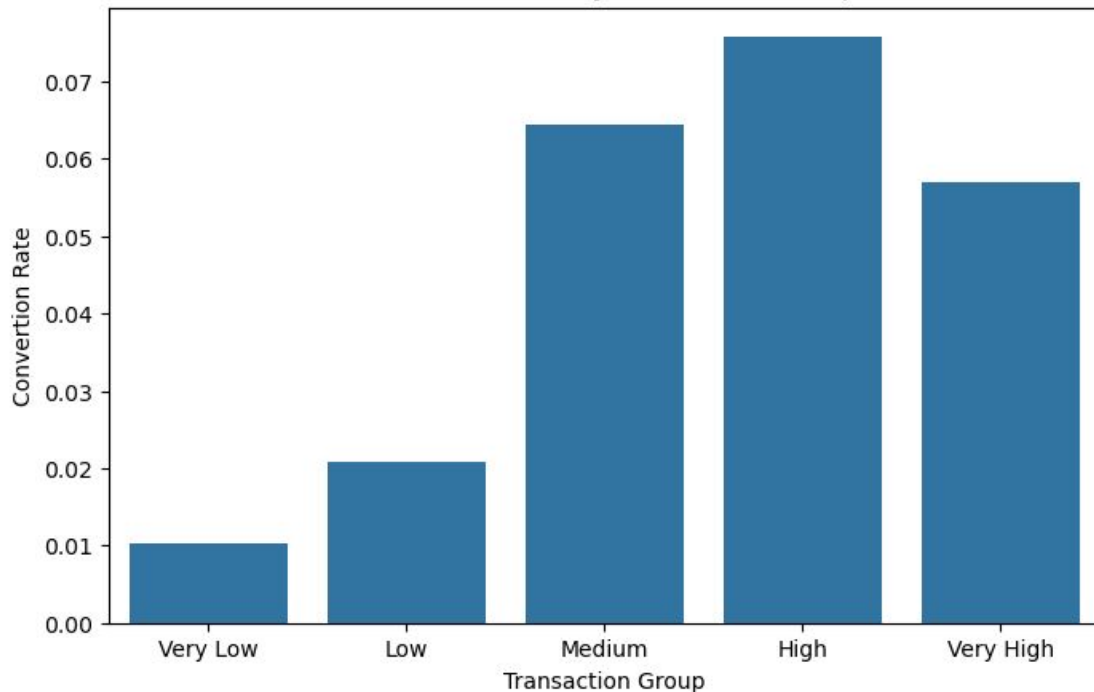


Customers with **undergraduate and postgraduate education levels (S1, S2, S3)** show a **conversion rate above 4%**, higher than customers with lower education levels (D3 and SMA).



## Frequent Shoppers Convert More

Conversion Rate by Transaction Group



Customers with **more than 11 transactions—in the Medium to Very High categories—showed conversion rates above 5%.** Meanwhile, customers with fewer transactions (Very Low and Low) recorded significantly lower conversion rates.

## Before Encoding

	Education	Marital_Status
251	D3	Bertunangan
1813	S1	Lajang
1827	S1	Menikah
1574	S3	Menikah
1516	S1	Bertunangan

## After Encoding

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2057 entries, 0 to 2056
Data columns (total 25 columns):
 #   Column                                Non-Null Count  Dtype  
---  --
 0   Income                                2057 non-null   float64
 1   Dt_Customer                           2057 non-null   object  
 2   Recency                               2057 non-null   int64   
 3   NumWebVisitsMonth                     2057 non-null   int64   
 4   Complain                              2057 non-null   int64   
 5   Z_CostContact                         2057 non-null   int64   
 6   Z_Revenue                             2057 non-null   int64   
 7   Response                              2057 non-null   int64   
 8   ConversionRate                        2057 non-null   float64
 9   Age                                   2057 non-null   int64   
10   Children                             2057 non-null   float64
11   TotalSpend                           2057 non-null   int64   
12   SpendingGroup                        2057 non-null   category
13   NumAcceptedCampaigns                 2057 non-null   int64   
14   TotalTransactions                    2057 non-null   float64
15   TransactionGroup                     2057 non-null   category
16   AgeGroup                             2057 non-null   category
17   IncomeGroup                          2057 non-null   category
18   Marital_Status_Bertunangan            2057 non-null   float64
19   Marital_Status_Cerai                  2057 non-null   float64
20   Marital_Status_Duda                    2057 non-null   float64
21   Marital_Status_Janda                  2057 non-null   float64
22   Marital_Status_Lajang                 2057 non-null   float64
23   Marital_Status_Menikah                2057 non-null   float64
24   Education                             2057 non-null   float64
dtypes: category(4), float64(11), int64(9), object(1)
memory usage: 346.5+ KB
```

The **Marital\_Status** column is encoded using the **One-Hot Encoding** method, while the **Education** column uses the **Ordinal Encoding** method. This selection is based on the characteristics of each column: **Education** has an intrinsic value order (such as SMA < D3 < S1 < S2 < S3), so it is more appropriate to use Ordinal Encoding. Meanwhile, **Marital\_Status** does not have a logical order between its categories, so One-Hot Encoding is more appropriate.

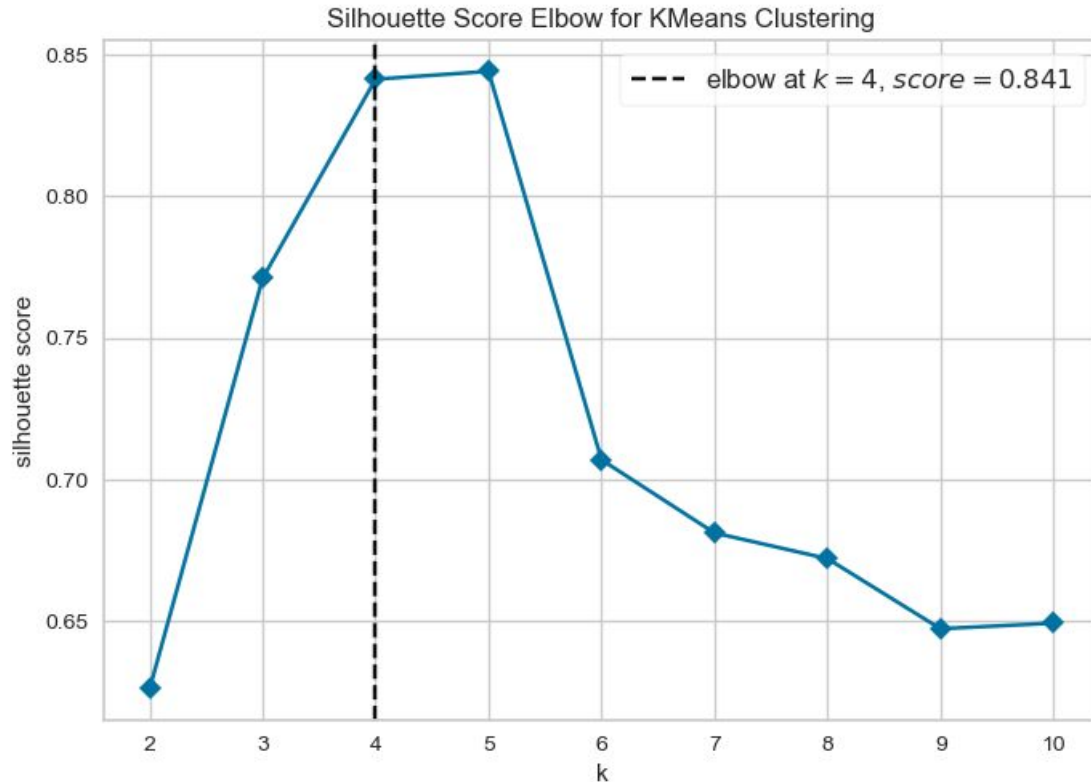
## Before Standardization

	Age	Income	Children	Recency	TotalSpend	NumWebVisitsMonth	TotalTransactions	NumAcceptedCampaigns
count	2057.000	2.057000e+03	2057.000	2057.000	2057.000	2057.000	2057.000	2057.000
mean	56.164	5.194200e+07	0.943	48.974	606313.563	5.318	14.864	0.300
std	11.745	2.095013e+07	0.720	28.989	602922.034	2.440	7.654	0.678
min	29.000	1.730000e+06	0.000	0.000	5000.000	0.000	0.000	0.000
25%	48.000	3.570100e+07	0.000	24.000	69000.000	3.000	8.000	0.000
50%	55.000	5.138150e+07	1.000	49.000	396000.000	6.000	15.000	0.000
75%	66.000	6.827400e+07	1.000	74.000	1047000.000	7.000	21.000	0.000
max	93.000	1.171335e+08	2.500	99.000	2514000.000	20.000	40.500	4.000

## After Standardization

	Age	Income	Children	Recency	TotalSpend	NumWebVisitsMonth	TotalTransactions	NumAcceptedCampaigns
count	2057.000	2057.000	2057.000	2057.000	2057.000	2057.000	2057.000	2057.000
mean	-0.000	-0.000	0.000	-0.000	0.000	-0.000	0.000	-0.000
std	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
min	-2.313	-2.397	-1.310	-1.690	-0.998	-2.251	-1.942	-0.443
25%	-0.695	-0.775	-1.310	-0.862	-0.891	-0.977	-0.897	-0.443
50%	-0.099	-0.027	0.079	0.001	-0.349	0.298	0.018	-0.443
75%	0.838	0.780	0.079	0.864	0.731	0.722	0.802	-0.443
max	3.137	3.113	2.162	1.726	3.165	3.271	3.350	5.454

## Elbow Method & Silhouette Score

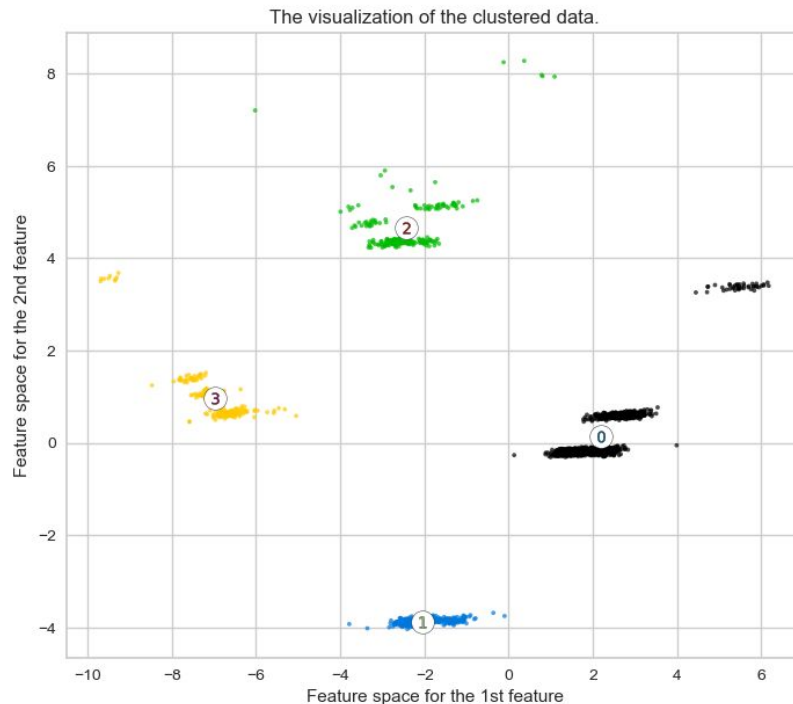
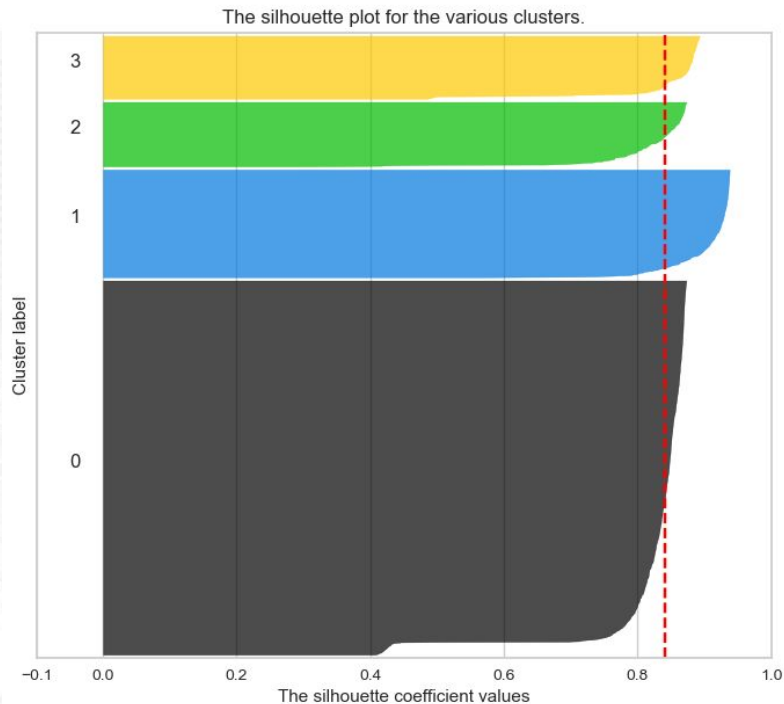


The results of the clustering analysis show that the elbow method with evaluation using the silhouette score produces an **optimal number of clusters of four clusters**, with a fairly high silhouette score value of 0.84.



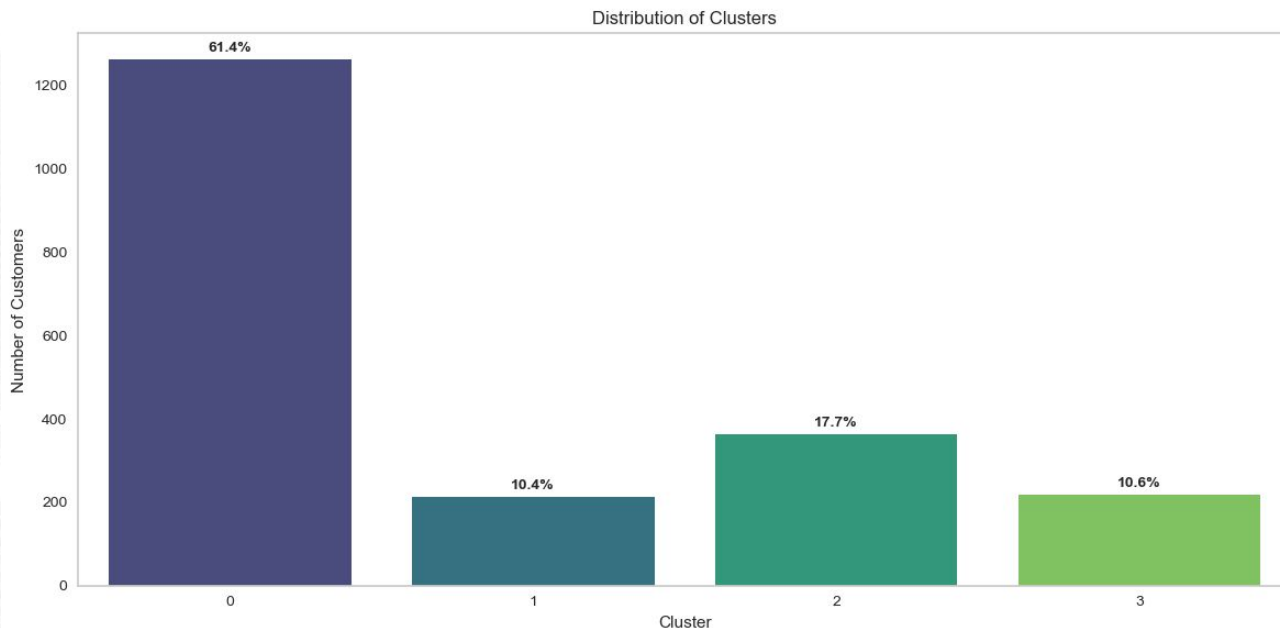
## Silhouette Analysis & Cluster Graph

Silhouette analysis for KMeans clustering on sample data with  $n\_clusters = 4$





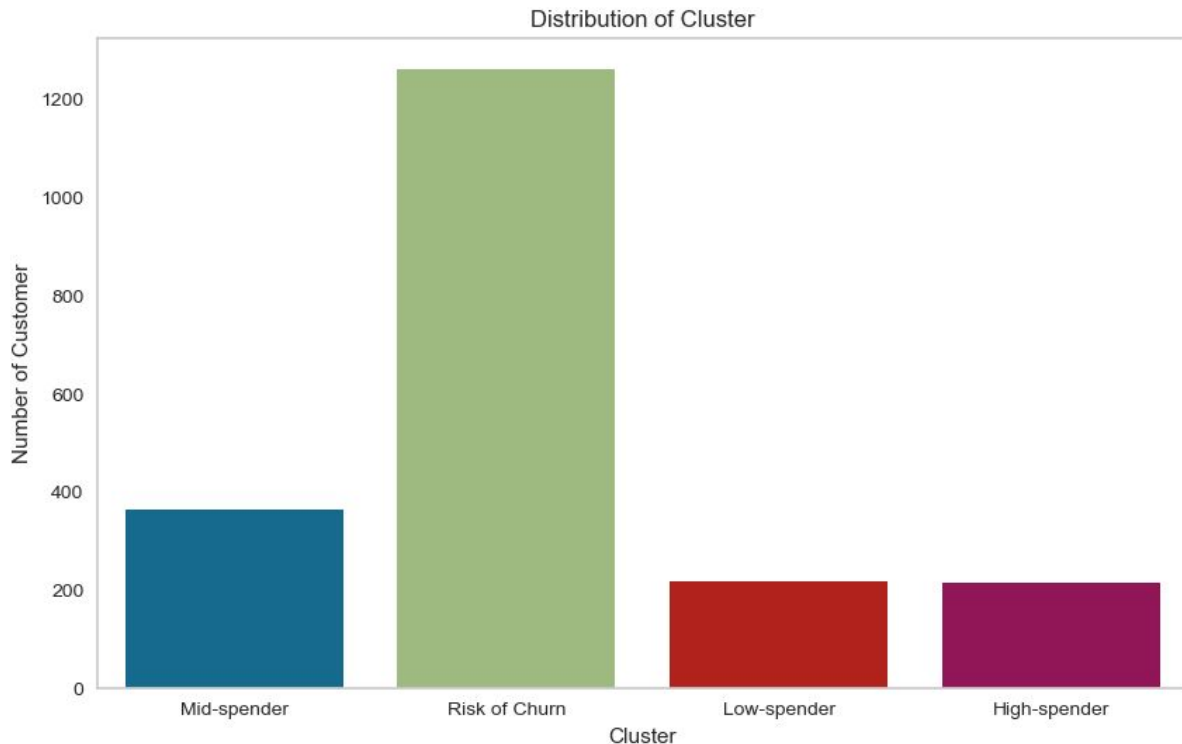
## Distribution of Clusters



Of these four clusters, **Cluster 0 dominates with the largest proportion, accounting for 61.4% of total customers.** This means that more than half of these customers share similar characteristics and can be considered the company's primary segment.

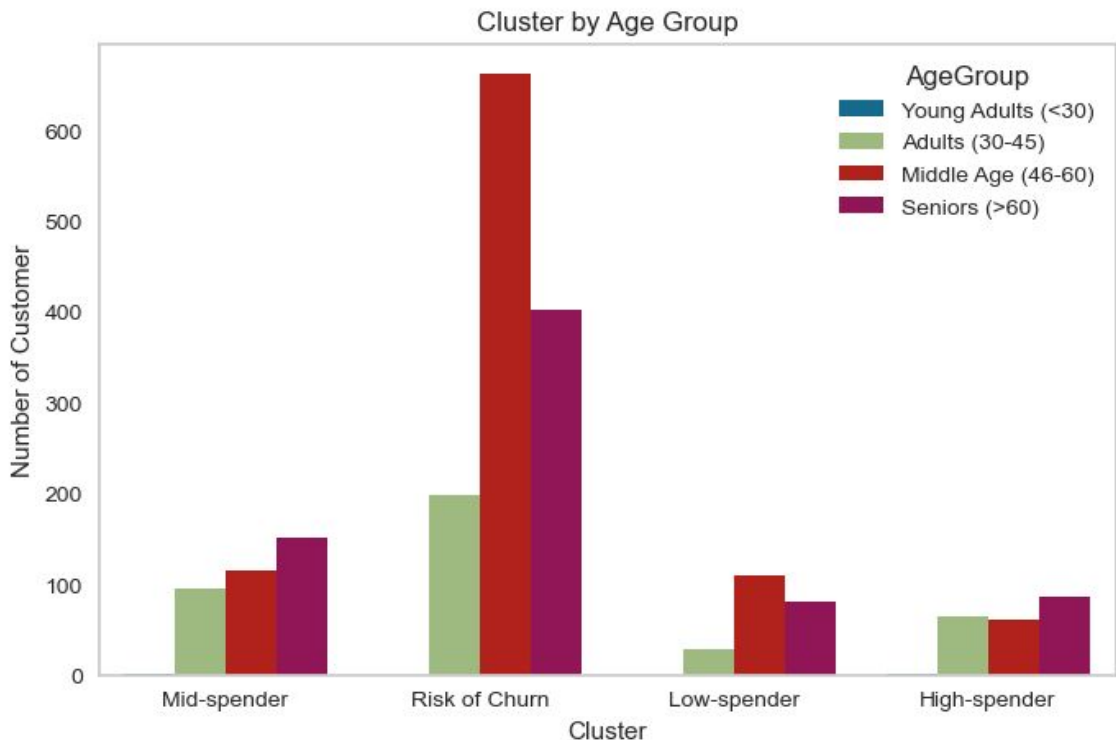
Meanwhile, **Cluster 1 and Cluster 3 represent groups with similar and smaller proportions of customers,** and therefore may represent specialized or niche segments with more specific needs.

## Most Customers Are at Risk of Churning



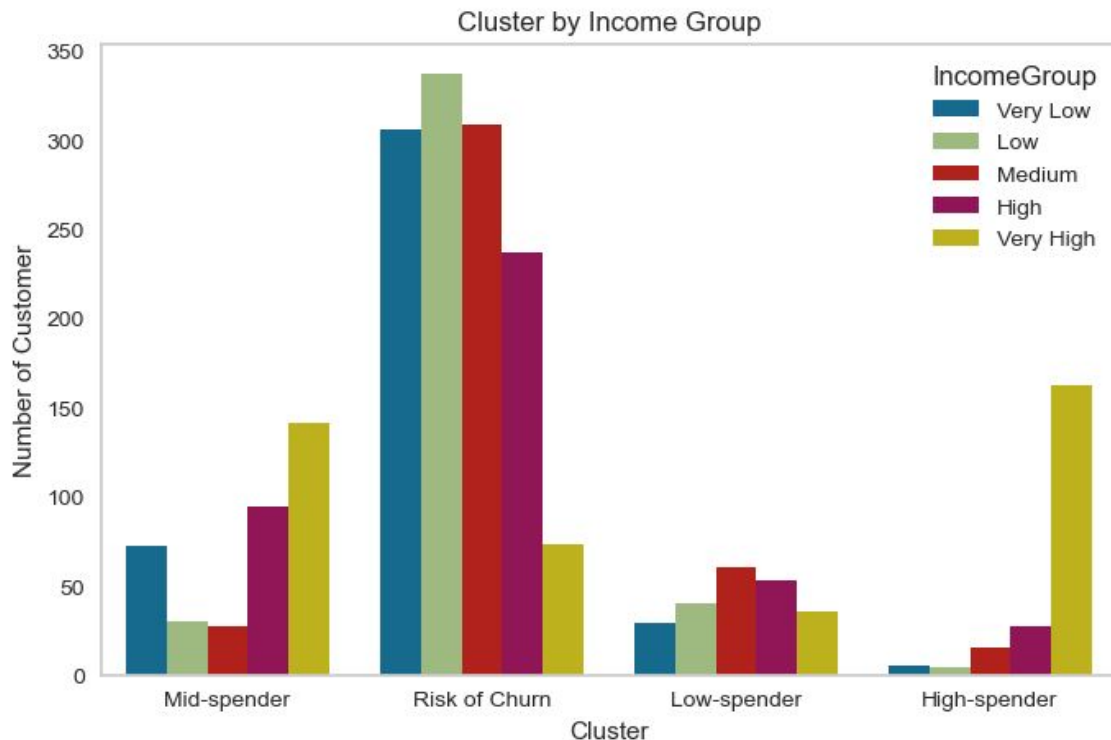
**More than 1,200 customers fall into the Risk of Churn cluster, far more than Mid-Spender ( $\pm 390$ ), Low-Spender, and High-Spender ( $\pm 200$ ).**

## Older Customers Dominate Risk and Low-Spending Clusters



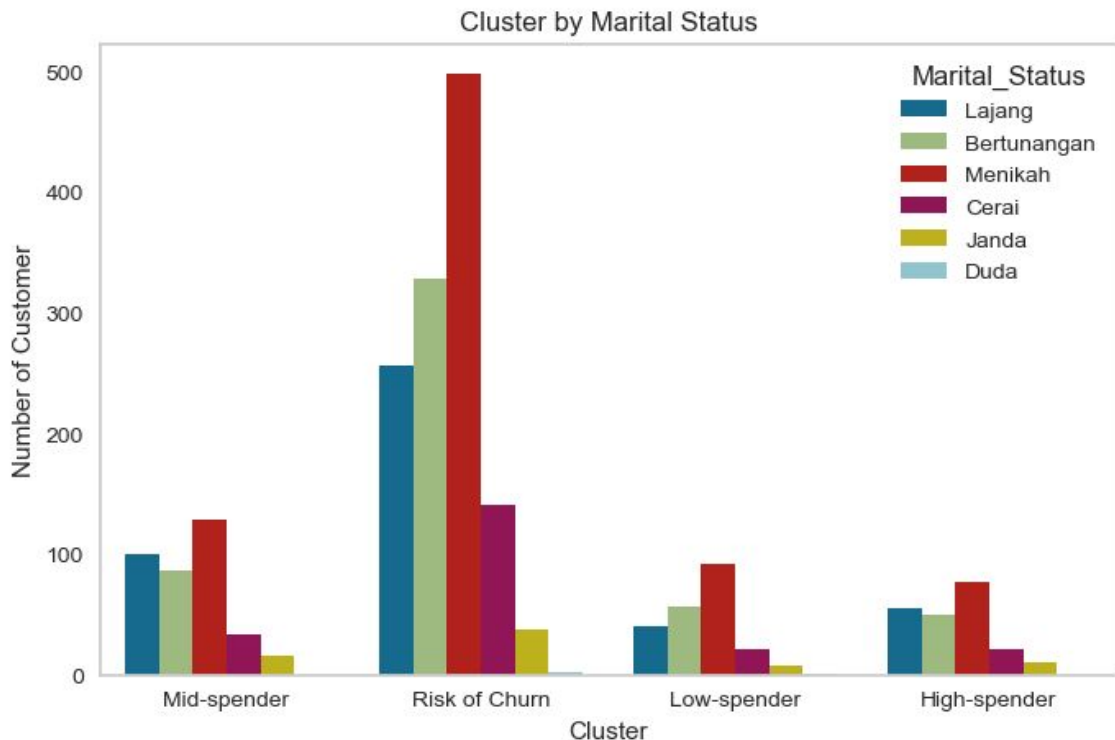
The risk of churn and low spenders is dominated by middle-aged customers (46–60), followed by seniors (>60) and adults (30–45). Seniors make up the majority of mid-spenders and high-spenders, followed by middle-aged and adult customers.

## Low-Income Customers Dominate Risk of Churn Cluster



The Risk of Churn cluster is dominated by customers with **Very Low** (1.73–32.23 million), **Low** (>32.23–44.94 million), and **Medium** (>44.94–58.17 million) income levels. In contrast, customers with **High** and **Very High** income are more commonly found in the **High-Spender** and **Mid-Spender** clusters.

## Married Customers Dominate All Clusters

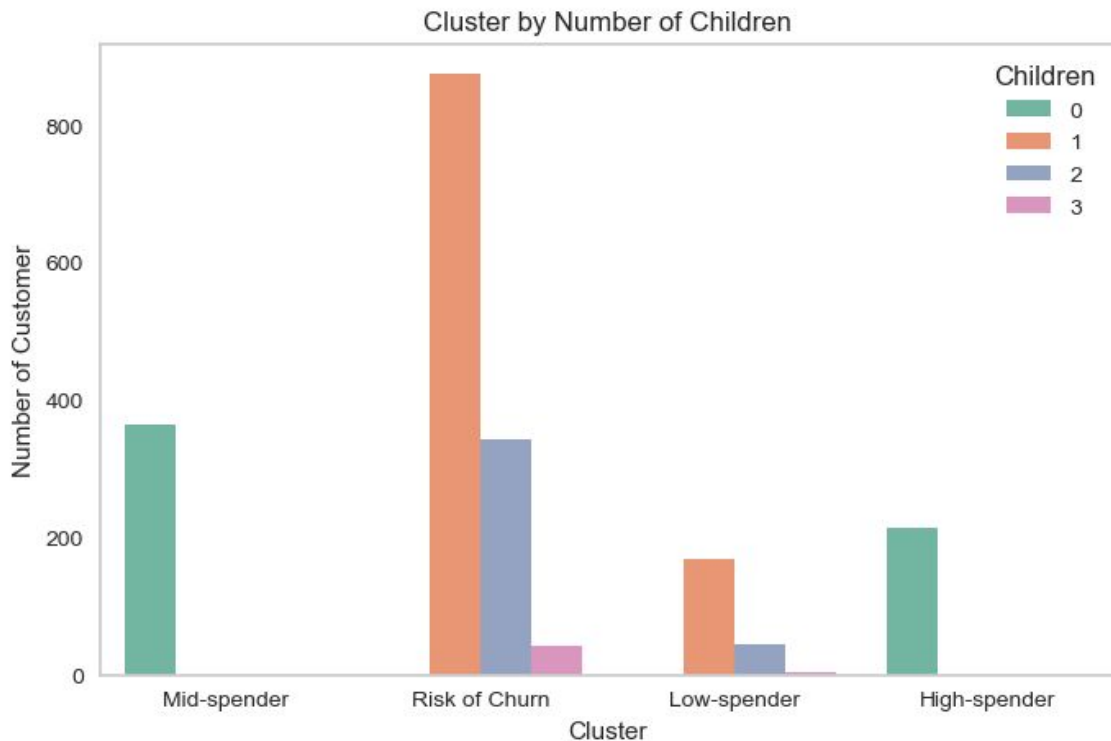


**Married** customers dominate across all clusters, followed by those who are **Engaged**, **Single**, **Divorced**, **Widowed**, and **Separated**.

**Marital status** appears to be associated with **stability** and **customer engagement** in transactions.



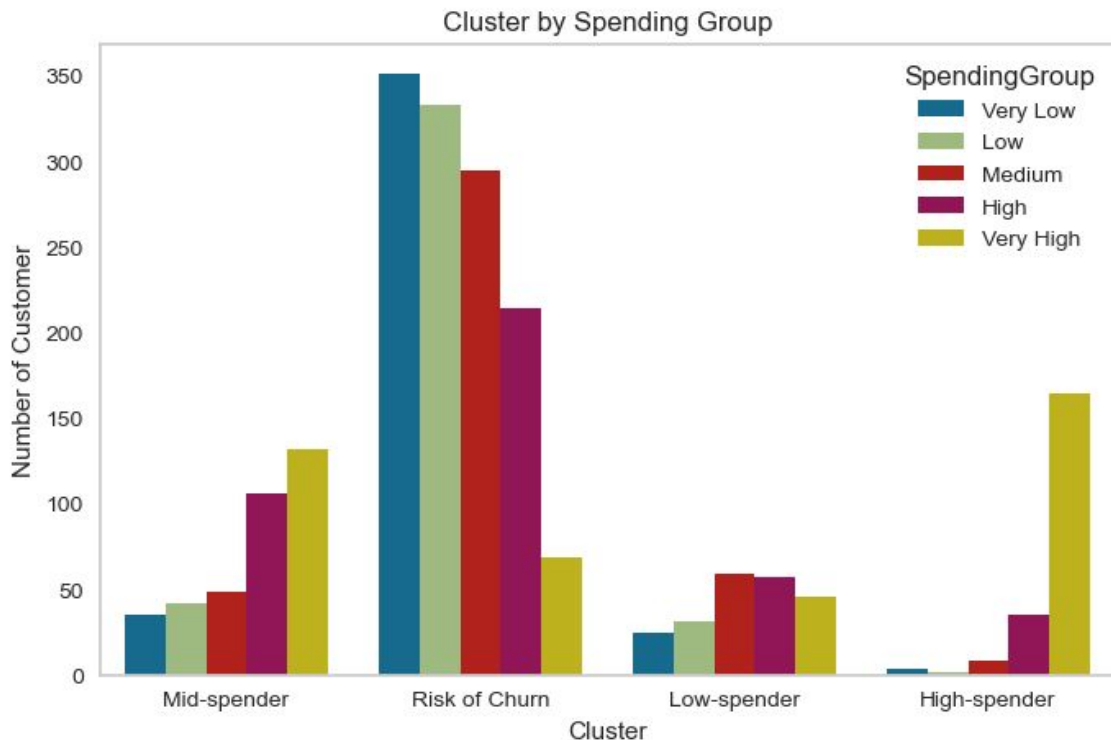
## Customers With Children Tend to Spend Less and Churn More



The **Risk of Churn** and **Low-Spender** clusters are dominated by customers **with children**.

In contrast, the **Mid-Spender** and **High-Spender** clusters consist exclusively of customers **without children**.

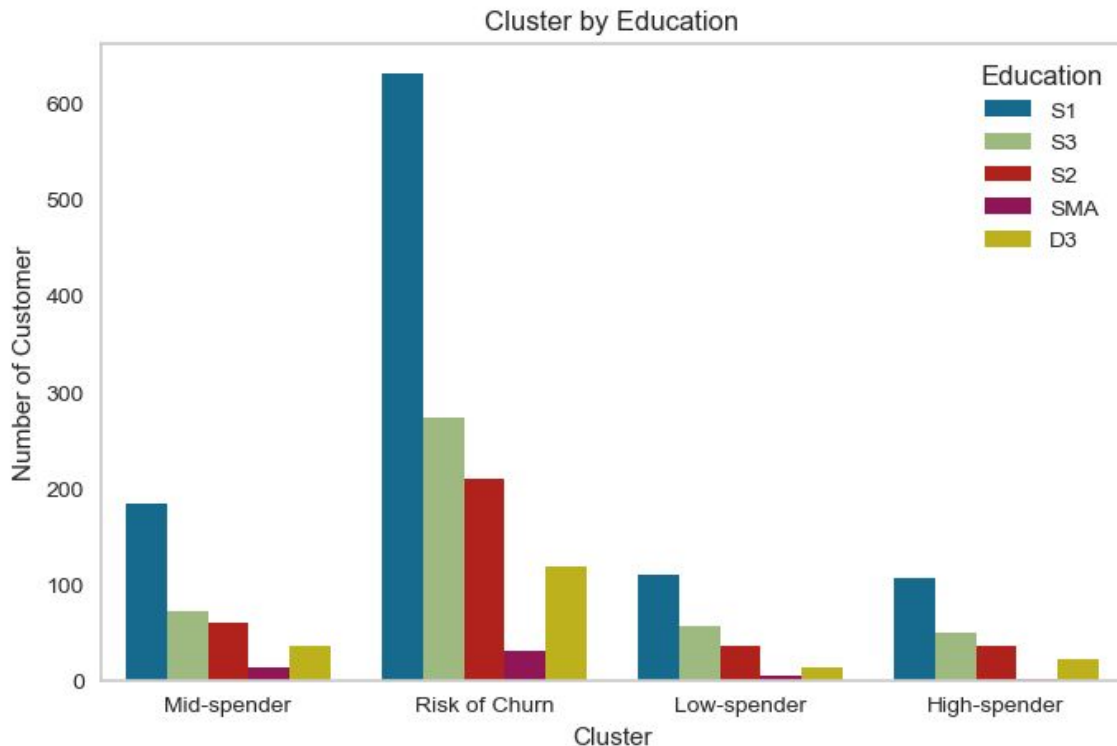
## High-Spenders Mostly Come From Very High-Income Group



The **Risk of Churn** cluster is dominated by customers with **Very Low to Medium** income levels. **Low-Spender** and **Mid-Spender** clusters have a more **diverse income distribution**.

In contrast, the **High-Spender** cluster is **mostly composed of** customers from the **Very High Income** group.

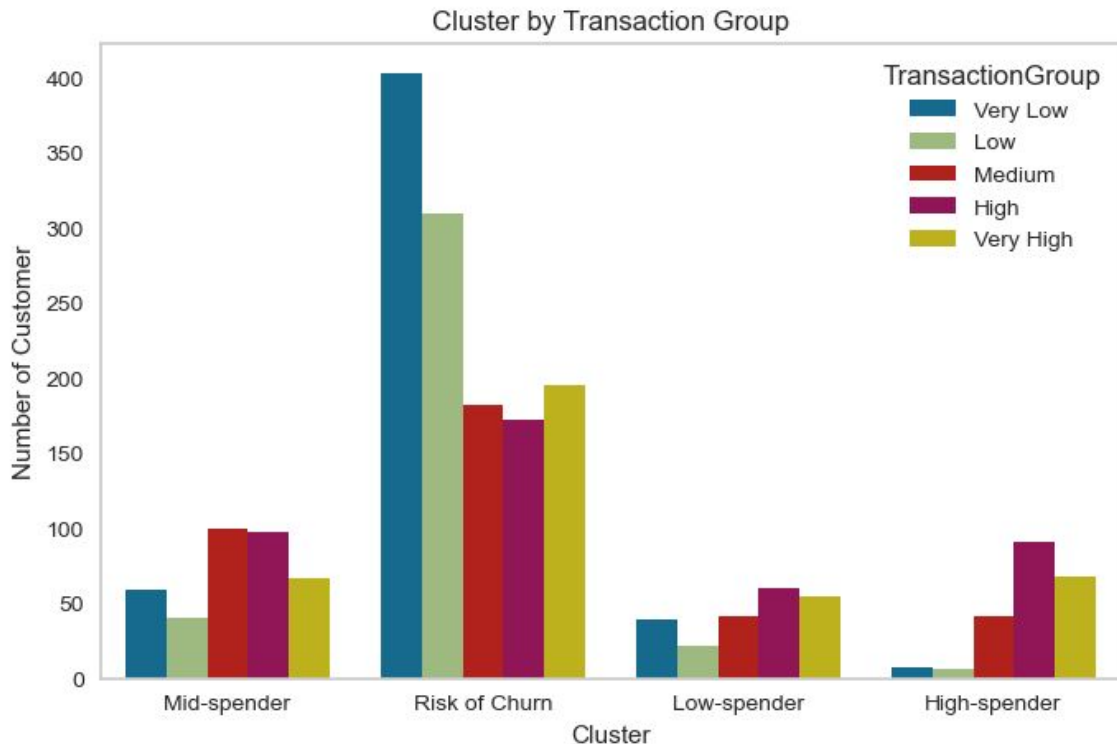
## Highly Educated Customers Dominate All Clusters



The majority of customers with **Bachelor's (S1), Master's (S2), and Doctoral (S3)** degrees dominate across all clusters.

This indicates that a **higher level of education** is associated with greater **customer engagement, purchasing power, and loyalty potential**.

## Very Low Transactions Dominate Risk of Churn Cluster



All clusters include customers with **varying numbers of transactions** (from **Very Low** to **Very High**).

However, the **majority of customers** with **Very Low transaction frequency** are concentrated in the **Risk of Churn** cluster.

## Risk of Churn Generates the Highest GMV Despite Inactivity



The **Risk of Churn** cluster generated the **highest GMV** (Rp 450 million), primarily due to its **very large customer base**.

Although customers in this cluster are currently **inactive**, they were **key contributors** to **historical revenue**.



## Recommendations

- **Re-engage the Risk of Churn Segment:** Target inactive customers with exclusive offers and personalized messages to encourage them to return and re-engage with the platform.
- **Retain and Upsell Mid & High Spenders:** Provide loyalty programs or exclusive promotions to drive repeat purchases and increase transaction value among high-potential, active customers.
- **Segment Based on Potential Value:** Combine historical data and customer lifetime value (CLV) to identify dormant but high-value customers for targeted reactivation efforts.
- **Optimize Targeting Using Stable Demographics:** Leverage insights from age, marital status, and education level to create more relevant and effective marketing campaigns tailored to each customer segment.

## Potential Impacts

- **Increase Customer Retention:** Reactivating just 20% of Risk of Churn customers (~240 customers) has the potential to generate an additional Rp92 million+ in GMV, based on average historical contribution.
- **Drive GMV Growth from Loyal Segments:** Upselling 10% of Mid and High Spender customers could result in an estimated Rp64 million in additional GMV.
- **Reduce Churn Among High-Value Customers:** Early identification of high-value customers at risk could help preserve up to Rp 450 million in GMV currently at risk of being lost.
- **Unlock Dormant GMV Potential:** If all Risk of Churn customers were successfully reactivated, the GMV potential could reach Rp450 million — equivalent to 31% of current total GMV.



**THANK YOU**