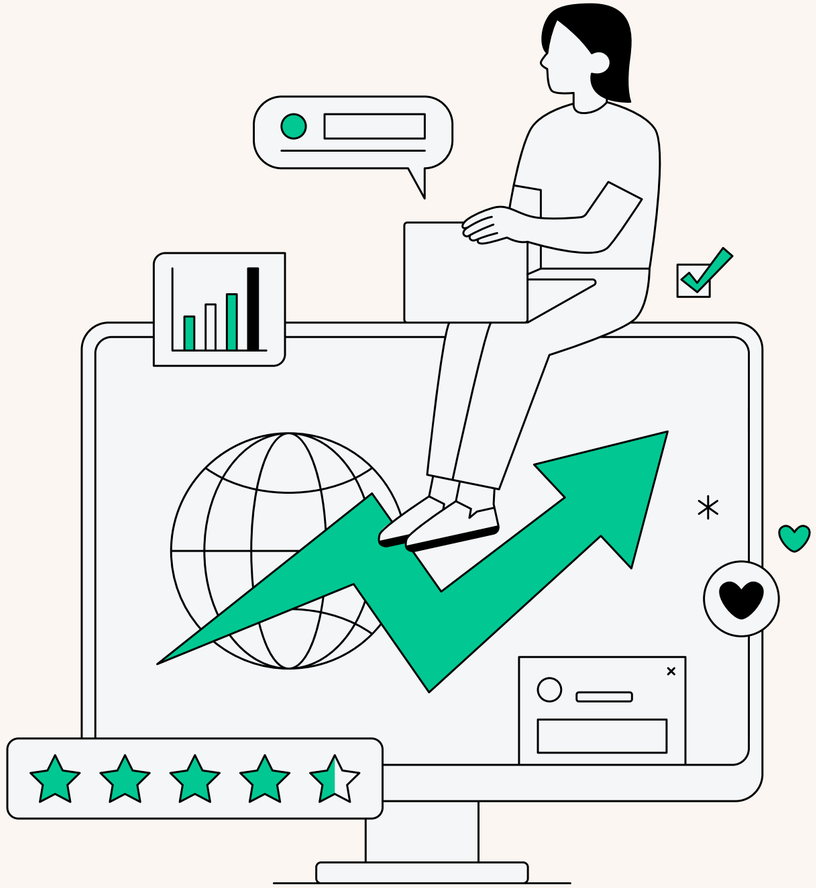
CUSTOMER PERSONALITY ANALYSIS

K MEANS--REGRESSION--CLV



Agenda & Objective

Introduction

- Leverage data to understand customer personalities through segmentation and behavior analysis
  - Analyze what drives success of campaigns via logistic regression
- Aim: Make informed decisions by analyzing customer needs, behaviors, and demographics

Objectives & Scope

- K-means Clustering: Identify customer segments based on similarities
  - Which is done through seeing their metrics on RFM : RFM variables are created from the raw data.
- CLV Analysis: Estimate long-term customer value (includes cohort and sensitivity analysis)
- Regression Analysis: Aimed to see what factors drive engagement in offered ad campaigns

Why This Dataset?

- Aligns with the 4 P’s of Marketing:
  - Product: Preferences
  - Price: Spending patterns
  - Place: Shopping channels
  - Promotion: Engagement data
- Enables informed strategies to enhance engagement and loyalty

What does our data say?

Dataset Overview

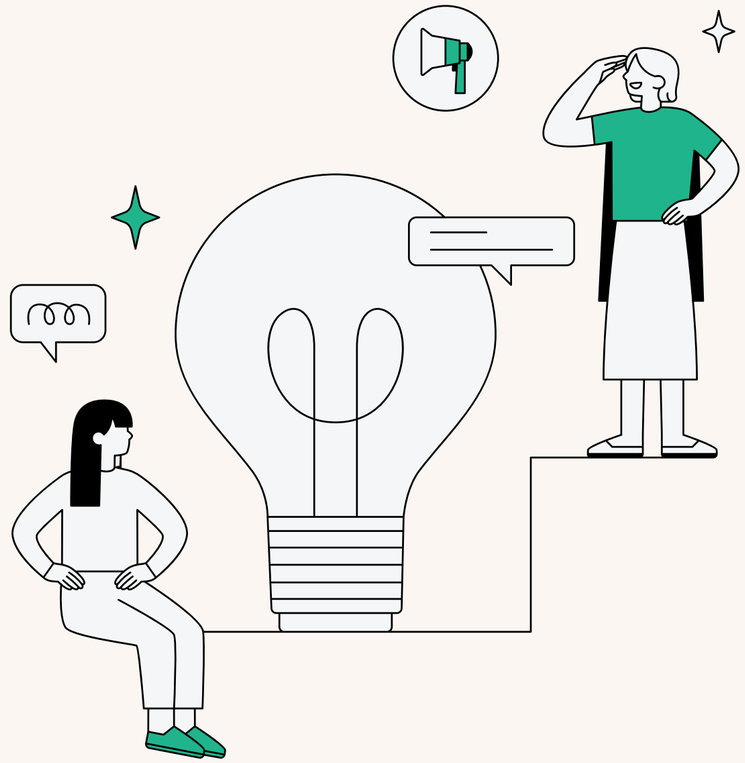
- Contains 2213 data points on demographics, purchasing patterns, engagement, and preferences

Key Attributes

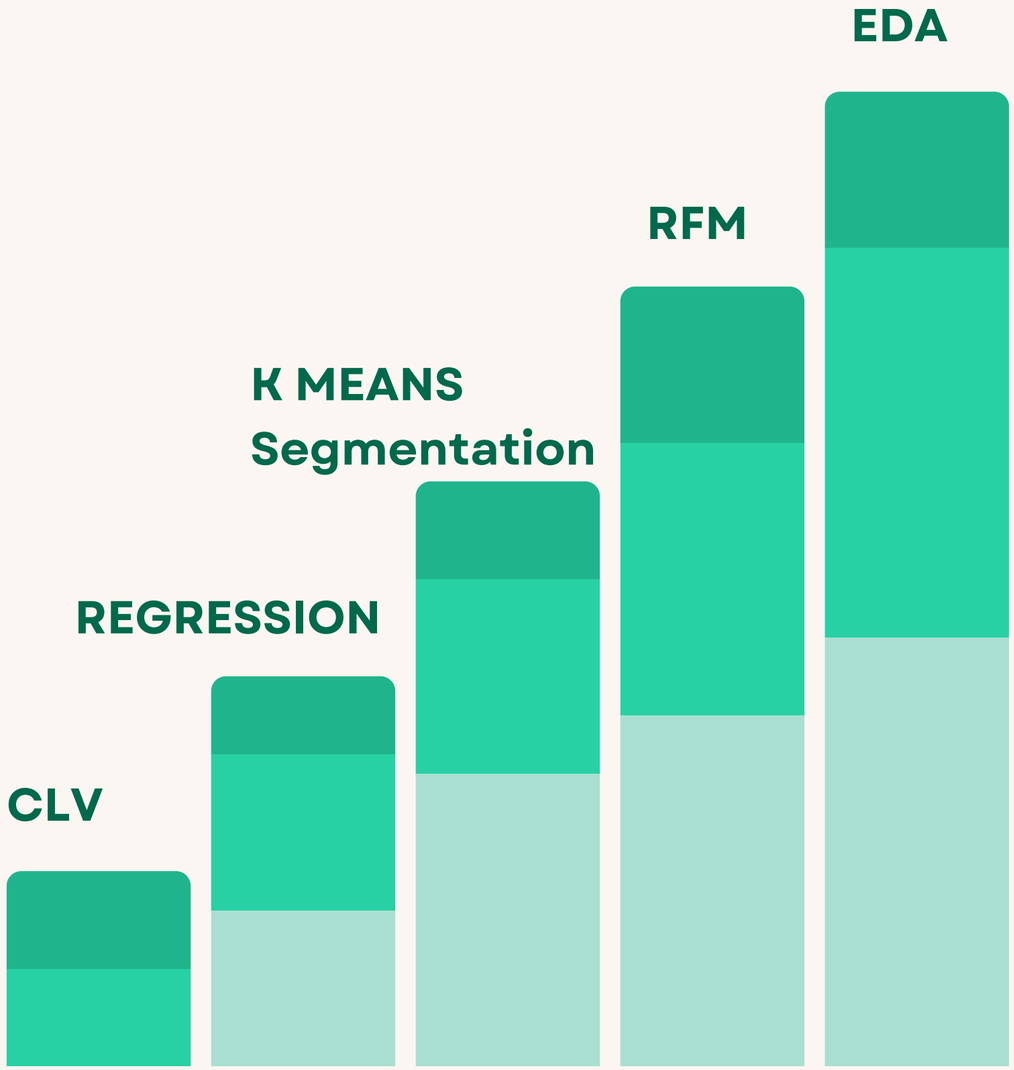
- Demographics: Age, income, education, marital status, household composition
- Product Spend: Preferences across categories like wines, meat, and sweets
- Engagement: Campaign interactions, discounts, and shopping channels
- Recency & Frequency: Tracks purchase timelines and behavior.

Why It Matters

- Provides a full view of customer behavior to inform marketing decisions
- Enables data-driven strategies for improved satisfaction, loyalty, and growth



Methodology used in the analysis



K-MEANS CLUSTERING USING RFM AS VARIABLES

- Combining RFM variables into K-means clustering offers a powerful approach to customer segmentation. It enables a multi-dimensional understanding of the customer base, blending direct purchase behaviors with broader customer characteristics

- We did an in-depth analysis of RFM variables and integrated those into K-Means segmentation as it answers vital questions like who are the best customers, and who contribute to churn rate

- This holistic view supports more informed strategic decisions, helping to maximize customer value and enhance engagement strategies

```
df['Frequency'] = df['NumWebPurchases']+df['NumCatalogPurchases']+df['NumStorePurchases']
```

```
df['Monetary'] = df['MntWines']+df['MntFruits']+df['MntMeatProducts']+df['MntFishProducts']+df['MntSweetProducts']+df['MntGoldProds']
```

K Means Segmentation Using RFM as Variables

RFM Segmentation without normalization

	Recency	Frequency	Monetary
count	2212.000000	2212.000000	2212.000000
mean	49.050633	12.551537	606.711573
std	28.940794	7.208691	603.163013
min	0.000000	0.000000	5.000000
25%	24.000000	6.000000	68.750000
50%	49.000000	12.000000	396.000000
75%	74.000000	18.000000	1047.250000
max	99.000000	32.000000	2525.000000

Standardized RFM variables using z-score normalization.
Increases accuracy and usability of each variable

	count	mean	std	min	25%	50%	75%	max
Recency	2212.0	-1.365193e-17	1.000226	-1.695245	-0.865778	-0.001750	0.862278	1.726306
Frequency	2212.0	-8.030546e-18	1.000226	-1.741561	-0.909044	-0.076527	0.755989	2.698529
Monetary	2212.0	5.781993e-17	1.000226	-0.997819	-0.892102	-0.349423	0.730546	3.181101

- High differentiation in variable ranges which will lead to higher weight on some variables compared to others, making the data unreliable. Monetary variable would have a higher weight while recency and frequency variables would be negligible

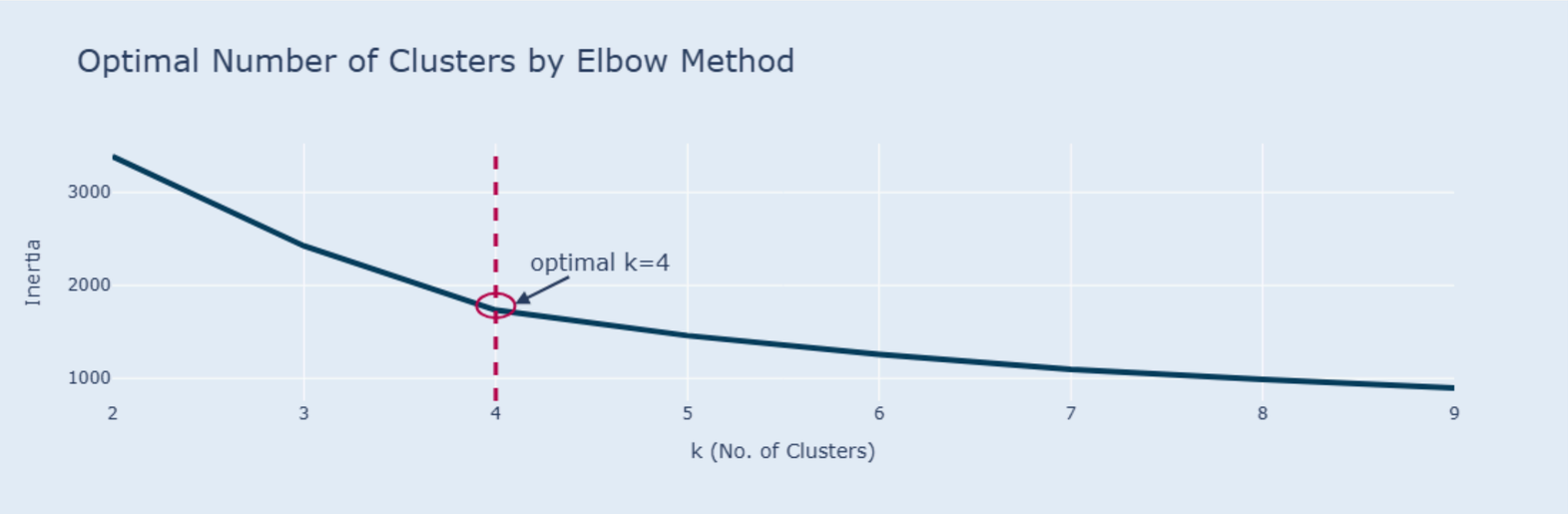
- Therefore, standardization re-scales the data to have a mean of 0 and a standard deviation of 1 so all variables effects can be seen and accounted for

Recency: The range from -1.693 to 1.727 suggests that while some customers purchased very recently (close to 1.727 standard deviations from the mean), others haven't engaged for a longer time (close to -1.693)

Frequency: The values range from -1.743 to 2.699, indicating that some customers purchase frequently (up to 2.699 standard deviations above the mean), while others may purchase infrequently (down to -1.743)

Monetary value: This suggests that some customers spend significantly more (up to 3.181 standard deviations above the average) compared to others who may not spend as much (down to -0.998)

CLUSTERING BY K-MEANS



First, determined the Number of Clusters using Elbow Method

Then, fitting the model using RFM variables with finalized 4 clusters

```
#k-means with k=4
kmeans = KMeans(n_clusters=4, random_state=2022)
kmeans.fit(df_rfm_scaled)
```

- The segmentation was based on categorizing the RFM variables from the raw data
- 4 optimal clusters were picked by scaling the RFM data
- Due to a limited number of data points, (2213) there could not be more clusters due to limited distribution
- Customers are divided into 4 clusters having different RFM characteristics

CUSTOMER SEGMENTATION ANALYSIS

In RFM analysis, each cluster's centroid represents the average (mean) values of Recency, Frequency, and Monetary metrics for the customers within that cluster

Clusters	Recency	Frequency	Monetary	Customer_Count
0	23.556465	7.034370	150.474632	611
1	73.491961	7.017685	146.393891	622
2	23.000000	19.760776	1186.963362	464
3	73.248544	19.285437	1181.161165	515

Cluster 0 = Low Recency, Low Frequency, Low Monetary
(Low-Spending Active Customers)

- Casual customers --> Transacted recently --> Low expenditures

Cluster 1 = High Recency, Low Frequency, Low Monetary
(Churned Low-Spending Customers)

- Low engagement --> Spent the least

Cluster 2 = Low Recency, High Frequency, High Monetary
(Best Active Customers, most valuable)

- Top customers -->High frequent expenditures --> Large contributers to revenue

Cluster 3 = High Recency, High Frequency, High Monetary
(Churned Best Customers)

- High expenditures -->High contributers to revenue-->Low frequency of purchase: Can indicate churning

Based on the boxplot and the centroid, we can see that the clusters separate the values of RFM into two groups, low and high

- **Recency**

🔽 Low : 22-23 days ; 🔼 High : 73 days

- **Frequency**

🔽 Low : 7 purchases ; 🔼 High : 19 purchases

- **Monetary Value**

🔽 Low : 146-151 USD ; 🔼 High : 1181-1184 USD

MARKETING STRATEGIES

CLUSTER 0

Low-Spending Active Customers

- Increase frequency and amount spent (monetary per purchase
- Create personalized offers to increase their purchase level in the form of vouchers or discounts
  - Conditional on if their transaction can reach the desired threshold

CLUSTER 1

Churned Low-Spending Customers

- Largest cluster out of the 4
- Promotions made to engage and resume spending
- Can be fulfilled by sending them emails, SMS, or notifications of personalized product recommendation with offers

CLUSTER 2

Best Active Customers

- Best customers deserve the best services
- Build connections and reward them with great offers and store points.
  - Done to reward loyalty
  - Make the customers feel part of a special club
  - Continue to drive majority of revenue

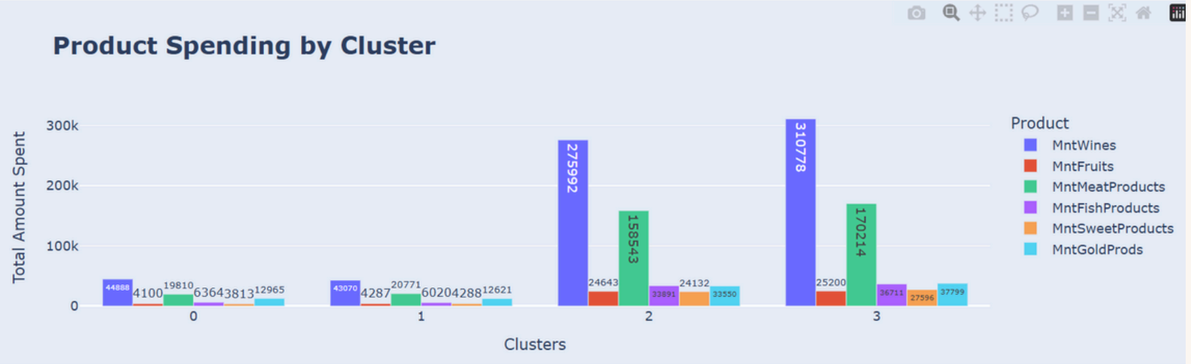
CLUSTER 3

Churned Best Customers

- This group was the revenue generators before. So we have an incentive to re-engage with them
- Conduct qualitative research of listening to their feedback and find out what will get them to come back
- Implement these changes and use that information for personalized product recommendations and price discounts

CUSTOMERS PROFILING

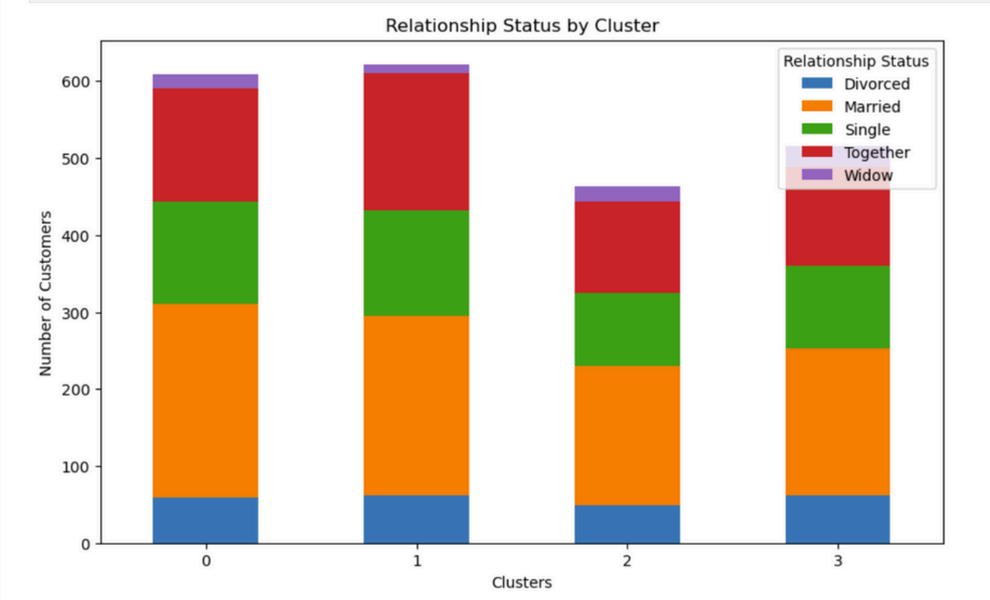
Analyze the clusters to the other features that reflect to the customer's profiles. Subsequently, the results can be used to enhance the personalized marketing strategy that has been created before



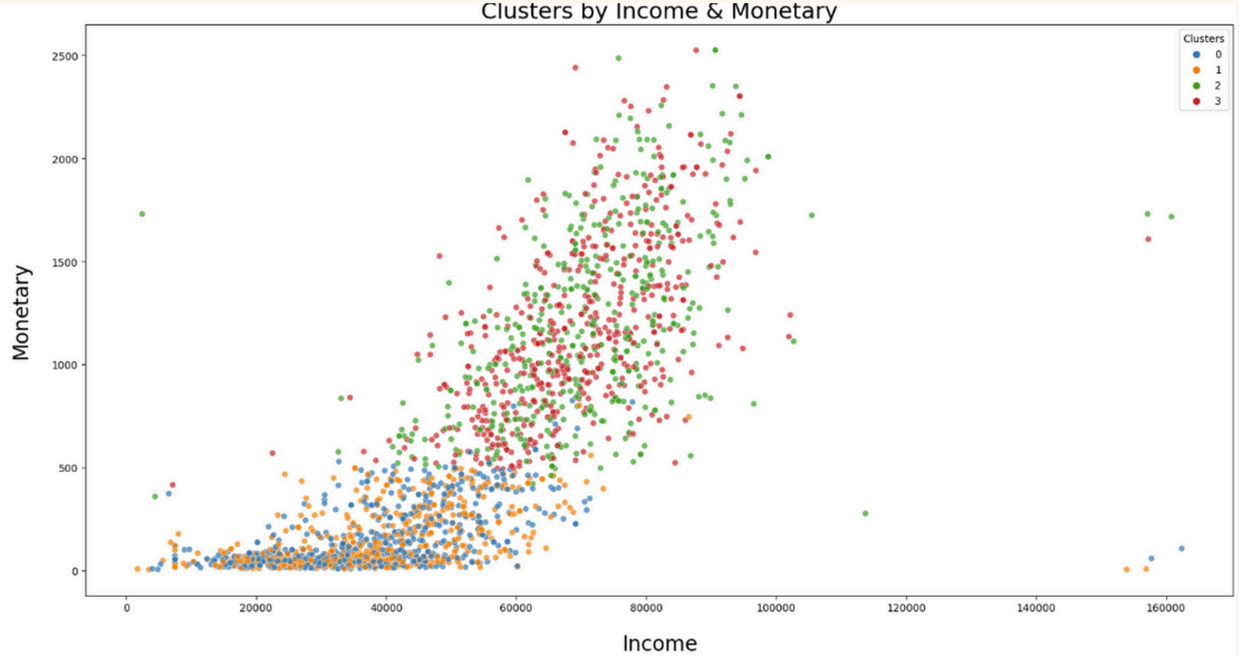
Customers from all segments have spent most of their money on Wine and Meat products

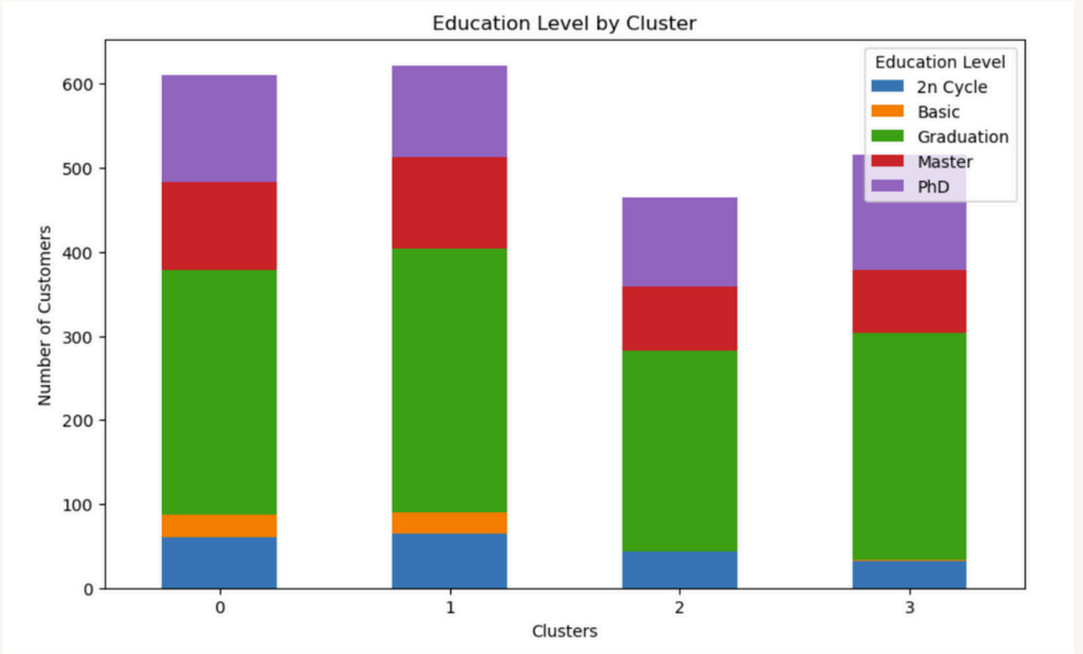
Cluster 1 and 0 have more kids than cluser 2 and 3

The average age per cluster is 50 years old. With a total of a 6 year age gap between cluster 0 and cluster 3
We also notice that as age increases so does expenditure on different products which is represented by clusters 2 and 3



Clusters 0, 1, and 3 have similar distributions, with "Married" being the dominant category
Cluster 2 shows slightly fewer customers, but the relationship distribution is consistent across clusters





- The chart shows customers with higher monetary tend to have higher income
  - Notice how the cluster 2 and 3 (Our Best Customers) are dominated by the high earners
- Nonetheless, there are some customers with high income that are low spenders. These are the customers that can be considered highly potential to increase their transaction

- We see that clusters 0 and 1 have more people with basic education and they spend the least compared to clusters 2 and 3

DEALS & SALES CHANNEL VS CLUSTERS

Promotion

NumDealsPurchases: Number of purchases made with a discount

Sales Channel

- NumWebPurchases: Number of purchases made through the company's web site
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's web site in the last month

Number of purchases made with a discount (If a customer has accepted any deals=1, else =0)

Clusters	Recency	Frequency	Monetary	Customer_Count
0	23.556465	7.034370	150.474632	611
1	73.491961	7.017685	146.393891	622
2	23.000000	19.760776	1186.963362	464
3	73.248544	19.285437	1181.161165	515

The deals received good responses from all the clusters, especially in the cluster 0 and 1 (low-spending customers). Despite their low frequency, these customers bought more discounted items than the best customers. Knowing the price sensitive behavior, offering them a deal will work best to increase their transactions

- Cluster 3: Has the highest total purchases across all purchase types
- Store purchases dominate, followed by Web purchases and Catalog purchases
- Cluster 2: Significant contribution from Store purchases, with Web purchases also being substantial
- Catalog purchases are noticeably higher in this cluster compared to others
- Cluster 1: Balanced distribution but with Store purchases leading, followed by Web purchases
- Cluster 0: Lowest total purchases among all clusters
- Store purchases again take the lead, but Catalog purchases are relatively low

CAMPAIGN ANALYSIS BASED ON CLUSTERS

AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise

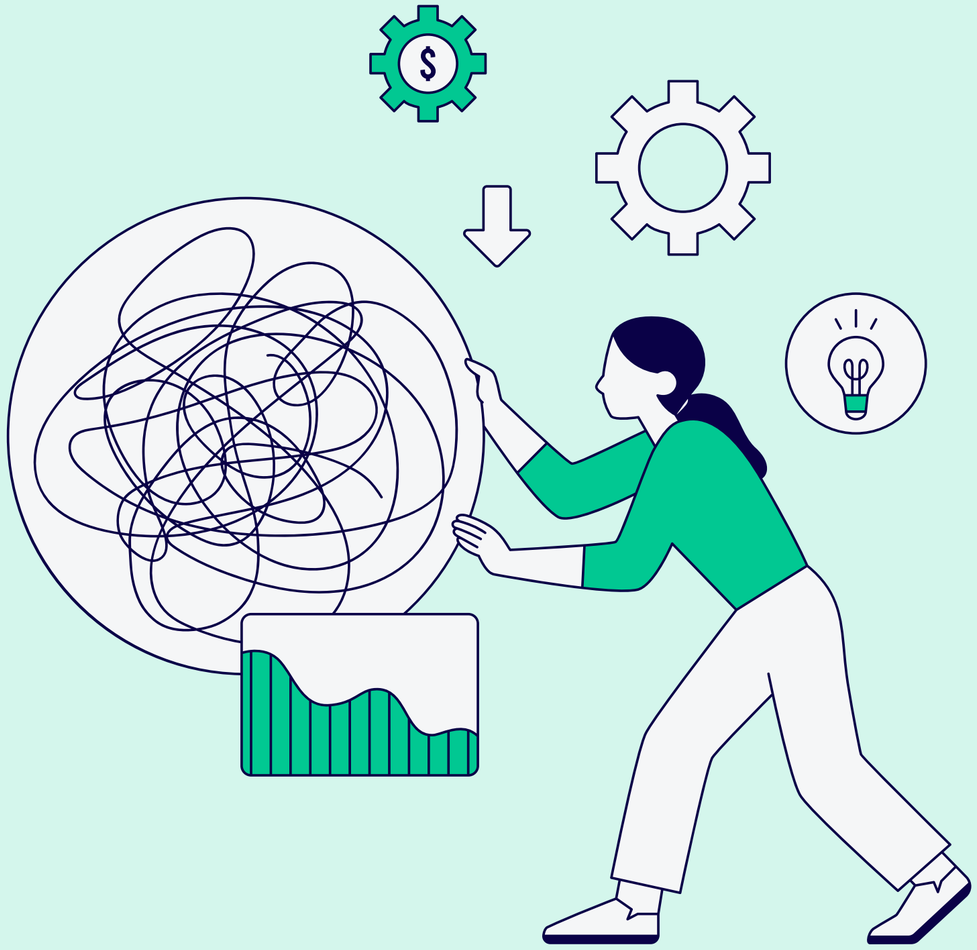
AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise

AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise

AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise

AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise

Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

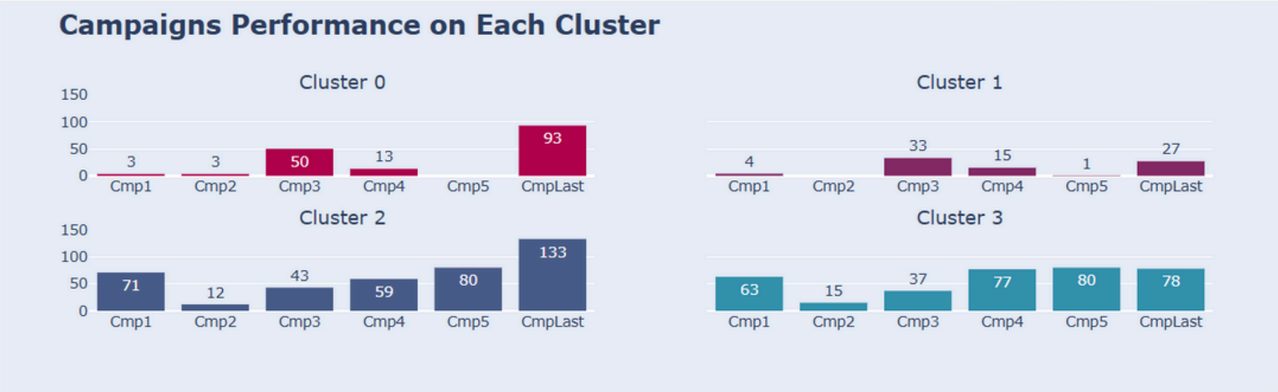


In contrary to the deals, the campaigns seems underperforming with far less acceptance than deals. Besides, cluster 2 and 3 (best customers) responded more than the low-spending customers. The marketing team might take this into consideration to craft more personalized campaigns for low-spending customers

Individual campaign performance

- Most campaigns failed to attract low-spending customers (cluster 0 and 1), where only the third and last campaign that could get higher acceptance . In addition, the second campaign received the least responses in all clusters while the last got the highest

- In the next campaign, the strategy can be formulated according to this evaluation. Which kind of offerings gave the best acceptance in each cluster and which one was not should be taken into account when designing the new campaign



REGRESSION ANALYSIS

Goal: To study the effect of demographic variables and Acceptance of Campaigns 1 to 5 on ‘Response’ variable i.e the 6th campaign (y- var.)

- The idea was to study the how the demographic variables (x-variables : Income, Age, Educational status, Marital Status, Kidhome, Teenhome etc.) affect the response to the 6th Campaign
- We also made an additional x variable Accp(1-5) which indicated '1' if a customer responded to an one of the 5 campaigns, 0 otherwise
- We ran 5 logistic regressions, 4 on each cluster & 1 across all clusters

REGRESSION ANALYSIS (cluster 0)

Model parameters (Variable Response):

Source	Value	Standard error	Wald Chi-Square	Pr > Chi²	Wald Lower	Wald Upper	Odds ratio	Odds ratio	Odds ratio
Intercept	-1.558	1.155	1.821	0.177	-3.821	0.705			
Age	0.020	0.013	2.326	0.127	-0.006	0.047	1.021	0.994	1.048
Edu_2n	-0.377	0.516	0.535	0.465	-1.388	0.634	0.686	0.250	1.885
Edu_Basic	-1.550	0.840	3.404	0.065	-3.197	0.097	0.212	0.041	1.101
Edu_Grad	-0.305	0.320	0.905	0.341	-0.933	0.323	0.737	0.393	1.381
Edu_Master	0.078	0.381	0.041	0.839	-0.670	0.825	1.081	0.512	2.281
Edu_PhD	0.000	0.000							
Mar_Single	-0.279	0.725	0.148	0.700	-1.700	1.142	0.757	0.183	3.133
Mar_Divorced	-0.822	0.791	1.080	0.299	-2.373	0.728	0.439	0.093	2.072
Mar_Married	-1.298	0.727	3.185	0.074	-2.723	0.127	0.273	0.066	1.136
Mar_Together	-1.208	0.744	2.639	0.104	-2.666	0.250	0.299	0.070	1.283
Mar_Widow	-0.979	0.000							
Income	0.000	0.000	2.924	0.087	0.000	0.000	1.000	1.000	1.000
Kidhome	0.681	0.264	6.678	0.010	0.165	1.198	1.976	1.179	3.313
Teenhome	-0.454	0.307	2.190	0.139	-1.054	0.147	0.635	0.348	1.159
Accp(1-5)	2.266	0.309	53.853	<0.0001	1.661	2.872	9.644	5.265	17.665

Significant

- Kidhome (C= 0.681, O.R = 1.976)
- Accp(1-5) (C=2.266, O.R = 9.644)

Insights

Family Dynamics: The positive association with Kidhome indicates that customers with children in this cluster are more likely to respond to campaigns. This may be due to family-focused priorities or the appeal of products suited to households with children

Past Engagement: **Accp(1-5)** is a critical predictor, implying that past responders are highly likely to respond again. This could suggest a loyal or highly engaged customer base in this cluster

REGRESSION ANALYSIS (cluster 1)

Model parameters (Variable Response):

Source	Value	Standard error	Wald Chi-Square	Pr > Chi²	Wald Lower	Wald Upper	Odds ratio	Odds ratio	Odds ratio
Intercept	0.001	1.907	0.000	1.000	-3.736	3.738			
Age	-0.028	0.024	1.327	0.249	-0.076	0.020	0.972	0.927	1.020
Edu_2n	-0.154	0.921	0.028	0.867	-1.959	1.651	0.857	0.141	5.213
Edu_Basic	-15.924	1866.672	0.000	0.993	########	3642.686			
Edu_Grad	0.063	0.617	0.011	0.918	-1.146	1.273	1.065	0.318	3.571
Edu_Master	0.494	0.703	0.493	0.483	-0.884	1.871	1.638	0.413	6.497
Edu_PhD	0.000	0.000							
Mar_Single	-1.239	1.177	1.108	0.293	-3.547	1.068	0.290	0.029	2.911
Mar_Divorced	-1.092	1.226	0.794	0.373	-3.495	1.310	0.335	0.030	3.706
Mar_Married	-1.992	1.182	2.841	0.092	-4.308	0.324	0.136	0.013	1.383
Mar_Together	-1.704	1.171	2.116	0.146	-4.000	0.592	0.182	0.018	1.807
Mar_Widow	0.000	0.000							
Income	0.000	0.000	0.537	0.464	0.000	0.000	1.000	1.000	1.000
Kidhome	-0.001	0.440	0.000	0.999	-0.863	0.862	0.999	0.422	2.367
Teenhome	-0.267	0.519	0.266	0.606	-1.285	0.750	0.765	0.277	2.116
Accp(1-5)	2.091	0.469	19.889	<0.0001	1.172	3.009	8.090	3.228	20.275

Significant

- Accp(1-5) (C= 2.091, O.R = 8.090)

Insights

Past campaign acceptance (Accp(1-5)) is by far the most influential factor for accepting Campaign 6, with a very high odds ratio and statistical significance. This suggests that targeting customers who accepted previous campaigns could be a highly effective strategy within this cluster

REGRESSION ANALYSIS (cluster 2)

Model parameters (Variable Response):

Source	Value	Standard error	Wald Chi-Square	Pr > Chi²	Wald Lower	Wald Upper	Odds ratio	Odds ratio	Odds ratio
Intercept	0.881	1.049	0.704	0.401	-1.176	2.937			
Age	0.001	0.009	0.019	0.889	-0.017	0.020	1.001	0.983	1.020
Edu_2n	-0.934	0.484	3.723	0.054	-1.883	0.015	0.393	0.152	1.015
Edu_Basic	0.000	0.000							
Edu_Grad	-0.215	0.307	0.488	0.485	-0.817	0.388	0.807	0.442	1.474
Edu_Master	-0.372	0.399	0.867	0.352	-1.154	0.411	0.690	0.315	1.508
Edu_PhD	0.000	0.000							
Mar_Single	-0.858	0.617	1.936	0.164	-2.066	0.350	0.424	0.127	1.420
Mar_Divorced	-0.762	0.657	1.345	0.246	-2.050	0.526	0.467	0.129	1.692
Mar_Married	-2.045	0.608	11.314	0.001	-3.236	-0.853	0.129	0.039	0.426
Mar_Together	-2.152	0.628	11.739	0.001	-3.383	-0.921	0.116	0.034	0.398
Mar_Widow	-0.871	0.000							
Income	0.000	0.000	0.145	0.703	0.000	0.000	1.000	1.000	1.000
Kidhome	-0.064	0.404	0.025	0.874	-0.856	0.728	0.938	0.425	2.071
Teenhome	-1.621	0.297	29.752	<0.0001	-2.204	-1.039	0.198	0.110	0.354
Accp(1-5)	1.809	0.254	50.597	<0.0001	1.311	2.308	6.106	3.709	10.052

Significant

- Edu_2n (C= -0.934, O.R = 0.393)
- Mar_Married (C=-2.045, O.R = 0.129)
- Mar_Together (C=-2.152, O.R = 0.116)
- Teenhome (C=-1.652, O.R= 0.198)
- Accp(1-5) (C=1.809, O.R=6.106)

Insights

- **Past campaign acceptance (Accp(1-5))** is the most influential factor, greatly increasing the likelihood of accepting Campaign 6, with both high statistical significance and a large effect size
- **Marital status** is also important, particularly for married and cohabiting individuals, who are significantly less likely to accept the campaign
- **Teenagers** in the household significantly decrease the likelihood of acceptance, suggesting that households with teenagers may have different priorities or constraints that affect campaign response
- **Education** doesn't seem to play a big role except that basic education is shown to be a detterant

REGRESSION ANALYSIS (cluster 3)

Model parameters (Variable Response):

Source	Value	Standard error	Wald Chi-Square	Pr > Chi^2	Wald Lower	Wald Upper	Odds ratio	Odds ratio Lower	Odds ratio Upper
Intercept	-3.779	1.319	8.213	0.004	-6.364	-1.195			
Age	-0.016	0.012	1.791	0.181	-0.040	0.008	0.984	0.961	1.008
Edu_2n	-0.533	0.741	0.518	0.472	-1.985	0.919	0.587	0.137	2.506
Edu_Basic	-10.090	1788.696	0.000	0.995	-3515.870	3495.691			
Edu_Grad	-0.186	0.335	0.307	0.579	-0.842	0.471	0.831	0.431	1.602
Edu_Master	0.163	0.459	0.126	0.722	-0.737	1.063	1.177	0.479	2.896
Edu_PhD	1.486	0.000							
Mar_Single	-0.457	0.614	0.554	0.457	-1.660	0.747	0.633	0.190	2.110
Mar_Divorced	0.182	0.623	0.085	0.771	-1.040	1.403	1.199	0.354	4.068
Mar_Married	-1.642	0.608	7.299	0.007	-2.833	-0.451	0.194	0.059	0.637
Mar_Together	-1.429	0.614	5.411	0.020	-2.633	-0.225	0.240	0.072	0.799
Mar_Widow	0.324	0.000							
Income	0.000	0.000	9.019	0.003	0.000	0.000	1.000	1.000	1.000
Kidhome	0.098	0.601	0.026	0.871	-1.081	1.277	1.103	0.339	3.585
Teenhome	-0.243	0.294	0.685	0.408	-0.820	0.333	0.784	0.441	1.395
Accp(1-5)	2.090	0.315	43.938	<0.0001	1.472	2.707	8.082	4.357	14.991

Significant

- Mar_Married (C=-1.642, O.R = 0.194)
- Mar_Together (C=-1.429, O.R = 0.240)
- Accp(1-5) (C=1.809, O.R=3.709)

Insights

- **Past campaign acceptance (Accp(1-5))** is the most influential factor in predicting Campaign 6 acceptance, with a large effect size and high statistical significance
- **Marital status**, specifically being married or living together, has a strong negative association with the likelihood of campaign acceptance, similar to the other clusters

REGRESSION ANALYSIS (All clusters)

Model parameters (Variable Response):

Source	Value	Standard error	Wald Chi-Square	Pr > Chi²	Wald Lower	Wald Upper	Odds ratio	Odds ratio	Odds ratio
Intercept	-3.415	0.518	43.523	<0.0001	-4.430	-2.401			
Age	-0.003	0.006	0.252	0.616	-0.015	0.009	0.997	0.986	1.009
Edu_2n	-0.046	0.282	0.027	0.870	-0.598	0.506	0.955	0.550	1.659
Edu_Basic	-1.128	0.752	2.248	0.134	-2.602	0.347	0.324	0.074	1.414
Edu_Grad	0.223	0.161	1.906	0.167	-0.094	0.539	1.250	0.911	1.715
Edu_Master	0.432	0.207	4.338	0.037	0.025	0.838	1.540	1.026	2.311
Edu_PhD	0.863	0.000							
Mar_Single	1.440	0.339	18.071	<0.0001	0.776	2.105	4.223	2.173	8.203
Mar_Divorced	1.468	0.358	16.790	<0.0001	0.766	2.170	4.340	2.151	8.759
Mar_Married	0.501	0.335	2.241	0.134	-0.155	1.157	1.650	0.857	3.179
Mar_Together	0.471	0.345	1.869	0.172	-0.204	1.146	1.602	0.815	3.147
Mar_Widow	1.594	0.000							
Income	0.000	0.000	2.050	0.152	0.000	0.000	1.000	1.000	1.000
Kidhome	0.110	0.146	0.565	0.452	-0.177	0.396	1.116	0.838	1.486
Teenhome	-0.836	0.141	34.931	<0.0001	-1.113	-0.559	0.434	0.329	0.572
Accp(1-5)	2.011	0.143	197.284	<0.0001	1.730	2.291	7.470	5.642	9.890

Significant

- Edu_Master (C=0.432, O.R = 1.540)
- Mar_Single (C=1.440, O.R = 4.223)
- Mar_Divorced (C=1.468, O.R = 4.340)
- Teenhome (C=-0.836, O.R = 0.434)
- Accp(1-5) (C=2.011, O.R=7.470)

Insights

- **Marital Status**: Being single or divorced significantly raises the probability of accepting Campaign 6, possibly reflecting different life priorities or lifestyle flexibility
- **Education Level**: Higher education, particularly a master's degree, is positively associated with campaign acceptance
- **Teenhome**: Households with teenagers are less likely to accept the campaign, which might imply greater budget constraints or differing priorities in such households
- **Previous Campaign Engagement (Accp(1-5))**: This is the most impactful predictor. If customers have previously accepted other campaigns, they are substantially more likely to respond positively to Campaign 6

RECOMMENDATIONS (BASIS REGRESSION)

Summary

- Past acceptance of campaigns is consistently the strongest positive predictor of future campaign acceptance
- Marital status, particularly being married or cohabiting, shows a negative impact on acceptance likelihood across clusters
- Age, education, income and household composition (children/teenagers) generally do not have significant or positive effects across clusters

Future Steps

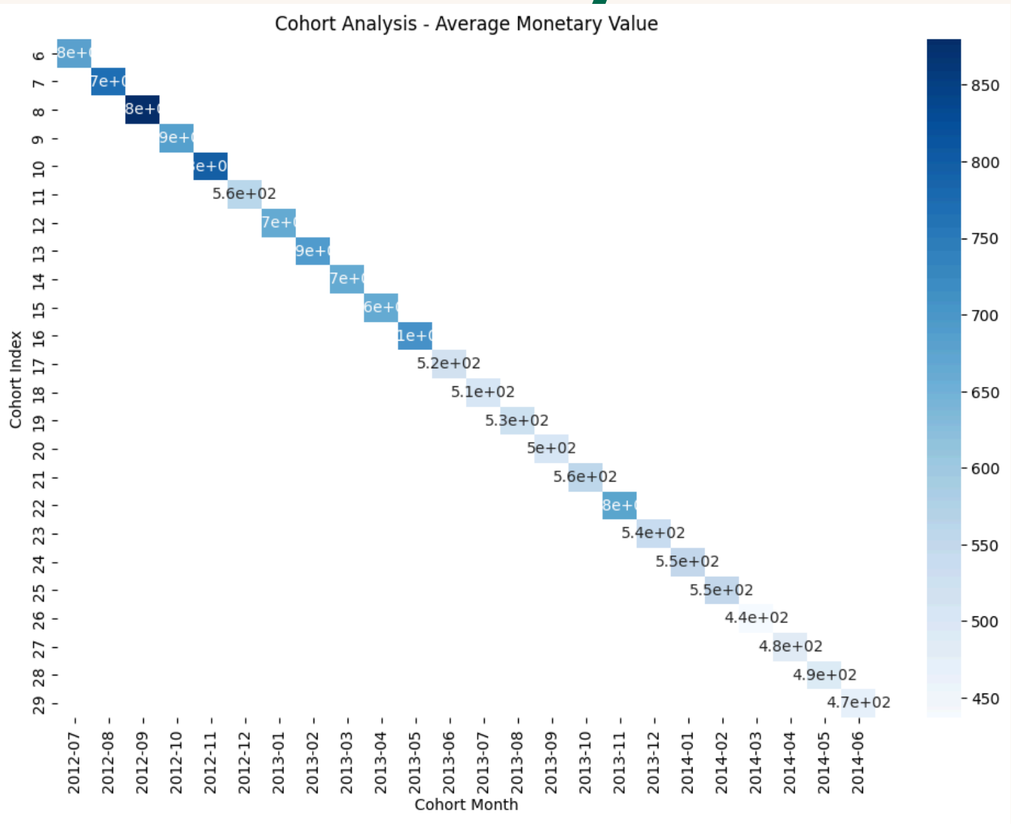
- Target Loyal Customers Who Accepted Previous Campaigns :
  - Personalized re-targeting
  - loyalty incentives or exclusive early access
- Refine Messaging for Married and Cohabiting Individuals
  - Family-focused messaging
  - Family-oriented discounts
- Develop Family-Friendly Campaign Variants to Appeal to Households with Kids & Teenagers

CLV ANALYSIS

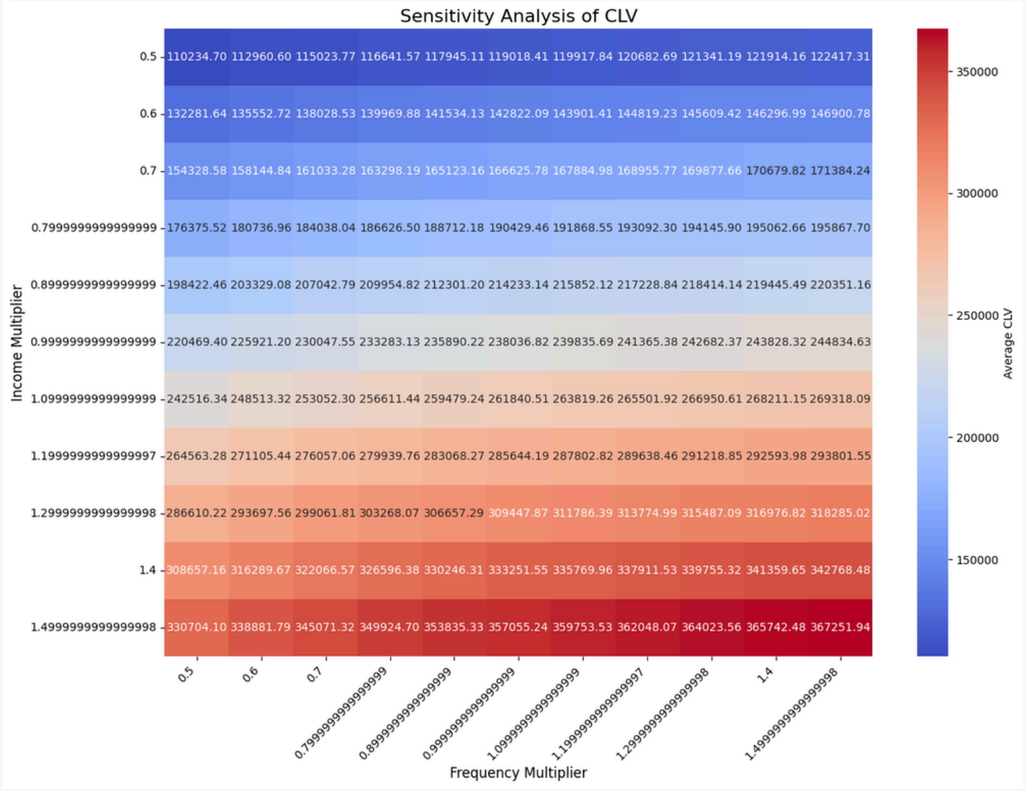
Customer Lifetime Value

Customer Lifetime Value = Average Order Value × Purchase Frequency Rate × Average Customer Lifetime

CLV Analysis



- Cohort Grouping: Rows represent customer cohorts based on their first purchase month
- Cohort Index: Indicates months since first purchase (y-axis).
- Average Spending: Darker cells = higher spending; lighter cells = lower spending
- *Insight: Spending declines over time within each cohort, highest when customers are newly acquired*



- Frequency Multiplier (x-axis): Reflects changes in purchase frequency
- Income Multiplier (y-axis): Reflects changes in average customer spending
- Average CLV: Darker red = higher CLV; blue = lower CLV
- *Insight: Increasing income or frequency significantly boosts CLV*

The retention rate assumption of 75% is based on general industry standards for companies in sectors like retail and consumer goods- HBR/Customer Retention Reports

STRATEGIC RECOMMENDATIONS

RECOMMENDATIONS

1

- Conduct both a mix of qualitative and quantitative research for each cluster based on the original data found in the K means cluster analysis
- Ensures each promotion/discount offers per segment has the most accurate information tailored per segment for driving re-engagement and increasing revenues

2

- Increase efficiency when it comes to the campaigns so it can generate more revenue
- Either increase the appeal of the campaigns to clusters 0 and 1 so they can engage with them which is friendly to low spenders
- Or make the campaigns to clusters 2 and 3 more lucrative so they spend more per purchase when they apply a deal on a shopping trip

3

- Drive more engagement towards campaign 6 as it has been the most successful (331 customers accepting it) compared to the average of campaigns 1-5 (157.25 customers accepting it)
- Conduct advertising that focus more on consumers that have a higher education and single. Remove any elements that focuses on children or having a family lifestyle

THANK YOU

APPENDIX

Why:

We aimed to understand the long-term value of our customers using CLV, an essential metric that helps in strategic marketing and customer retention

How:

- AOV - Average revenue per purchase
- Purchase Frequency - Customer buying patterns
- Customer Lifespan - Estimated based on churn rate

What:

CLV identifies high-value customers, guiding targeted marketing and retention

Application:

- Sensitivity Analysis: Impact of AOV and frequency changes on CLV
- Cohort Analysis: Track customer value over time
- Segmentation Analysis: Differentiate high, medium, and low-value customers

```
# Compute revenue per customer
revenue_per_customer = data.groupby('ID')['Total_Spend'].sum().mean()

# Compute average purchases per customer
purchases_per_customer = data['Total_Purchases'].mean()

# Calculate Average Order Value (AOV)
AOV = revenue_per_customer / purchases_per_customer
print("Average Order Value (AOV):", AOV)

# Calculate the total number of unique customers
total_customers = data['ID'].nunique()

# Calculate Purchase Frequency
frequency = purchases_per_customer / total_customers
print("Purchase Frequency:", frequency)
```

```
Average Order Value (AOV): 48.33763146520674
Purchase Frequency: 0.0056742934315209825
```

```
[ ] # Assuming a retention rate of 75%(Industry Standards)
    retention_rate = 0.75
    churn_rate = 1 - retention_rate
    print("Churn Rate:", churn_rate)

    # Calculate Expected Customer Lifespan
    expected_customer_lifespan = 1 / churn_rate
    print("Expected Customer Lifespan:", expected_customer_lifespan)
```

```
Churn Rate: 0.25
Expected Customer Lifespan: 4.0
```

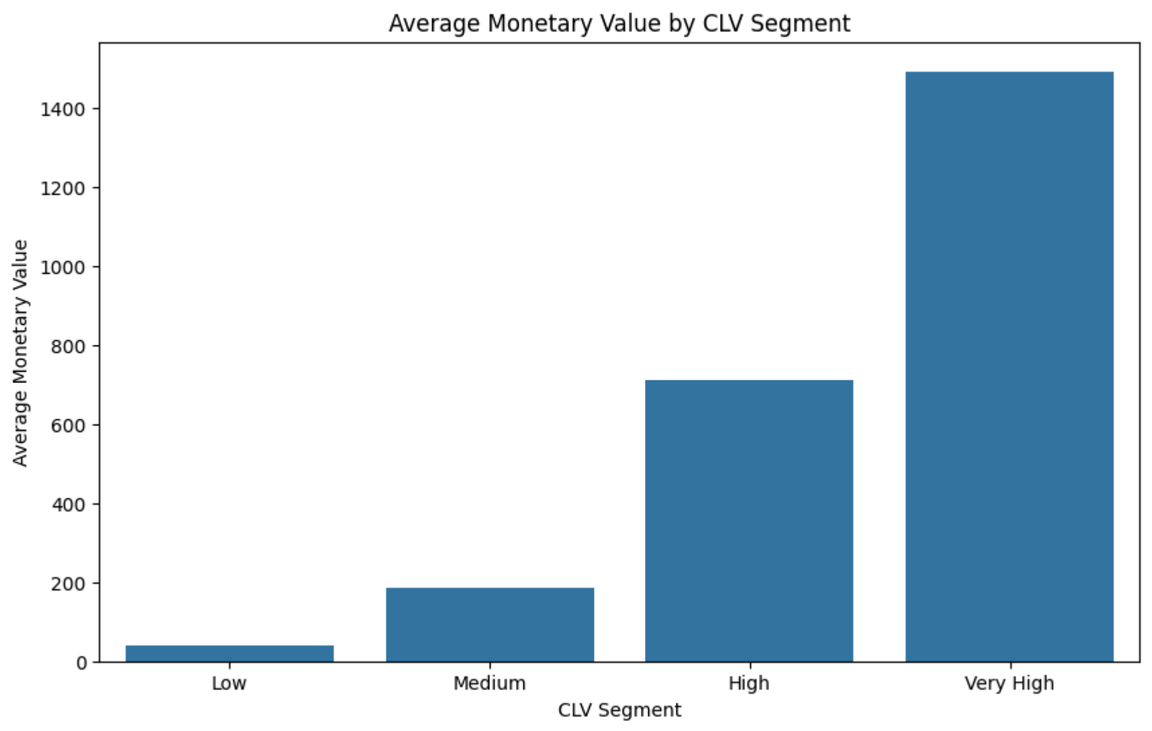
Calculate Customer Lifetime Value

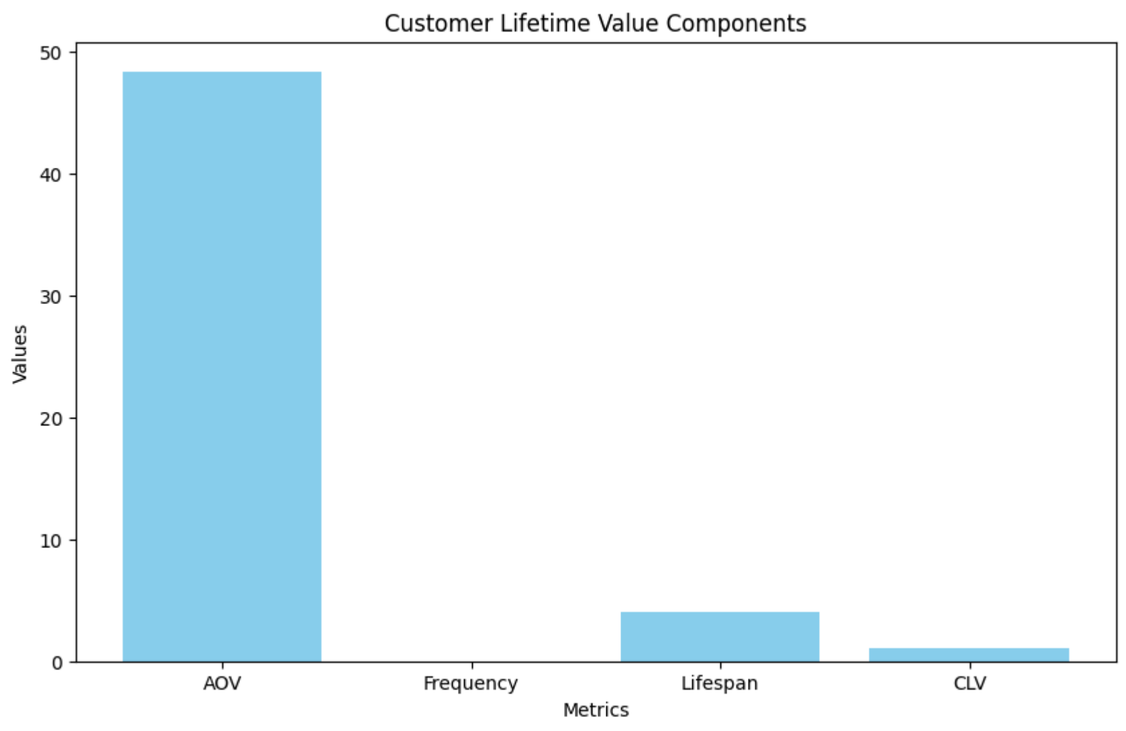
```
[ ] # Calculate Customer Lifetime Value (CLV)
    CLV = AOV * frequency * expected_customer_lifespan
    print("Customer Lifetime Value (CLV):", CLV)
```

```
Customer Lifetime Value (CLV): 1.0971276188732182
```

The retention rate assumption of 75% is based on general industry standards for companies in sectors like retail and consumer goods- HBR/Customer Retention Reports

CLV Analysis





- Low: 39.3 | Medium: 184.7 | High: 712.3 | Very High: 1492.5
- Clear increase in revenue as we move to "High" and "Very High" segments
- Strategic Focus: Retain and target "High" and "Very High" customers

- Frequency: Improving purchase frequency can boost CLV significantly
- Retention: Higher retention extends customer lifespan, increasing CLV
- Investment: Address AOV disparity to optimize lifetime value

CLV CALCULATION

Link to the working document:
https://colab.research.google.com/drive/1_tNvMzGSQA5IeIxbMUp_KDzdhVWQ83ZP?usp=sharing

Data Aggregration

```
# Create a new column for total spend by summing all product expenditure columns
data['Total_Spend'] = (data['MntWines'] + data['MntFruits'] + data['MntMeatProducts'] +
                       data['MntFishProducts'] + data['MntSweetProducts'] + data['MntGoldProds'])



# Calculate total purchases per customer
data['Total_Purchases'] = (data['NumWebPurchases'] + data['NumCatalogPurchases'] +
                           data['NumStorePurchases'])
```

Your paragraph text

```
# Display a sample of the modified dataset
data[['ID', 'Total_Spend', 'Total_Purchases']].head()
```

	ID	Total_Spend	Total_Purchases
0	5524	1617	22
1	2174	27	4
2	4141	776	20
3	6182	53	6
4	5324	422	14

The purpose of this code is to preview and check the content of the modified dataset, specifically focusing on the ID, Total_Spend, and Total_Purchases columns. This is useful for understanding how the data is organized and ensuring that the calculations or data extractions were done correctly.

We create new columns for **Total_Spend and Total_Purchases** to aggregate the monetary value and purchase frequency for each customer. This data will be crucial for calculating metrics like Average Order Value (AOV) and Purchase Frequency.

CLV CALCULATION

Calculate Key Metrics for CLV

```
# Compute revenue per customer
revenue_per_customer = data.groupby('ID')['Total_Spend'].sum().mean()
```

https://colab.research.google.com/drive/1_tNvMzGSQA5IeIxbMUp_KDzdhVWQ83ZP#printMode=true

---


```
# Compute average purchases per customer
purchases_per_customer = data['Total_Purchases'].mean()

# Calculate Average Order Value (AOV)
AOV = revenue_per_customer / purchases_per_customer
print("Average Order Value (AOV):", AOV)

# Calculate the total number of unique customers
total_customers = data['ID'].nunique()

# Calculate Purchase Frequency
frequency = purchases_per_customer / total_customers
print("Purchase Frequency:", frequency)
```

```
Average Order Value (AOV): 48.33763146520674
Purchase Frequency: 0.0056742934315209825
```

```
# Assuming a retention rate of 75%(Industry Standards)
retention_rate = 0.75
churn_rate = 1 - retention_rate
print("Churn Rate:", churn_rate)

# Calculate Expected Customer Lifespan
expected_customer_lifespan = 1 / churn_rate
print("Expected Customer Lifespan:", expected_customer_lifespan)
```

```
Churn Rate: 0.25
Expected Customer Lifespan: 4.0
```

We **assume** a retention rate of 75% and calculate the churn rate.

The expected customer lifespan is the reciprocal of the churn rate, giving us the average duration a customer stays active.

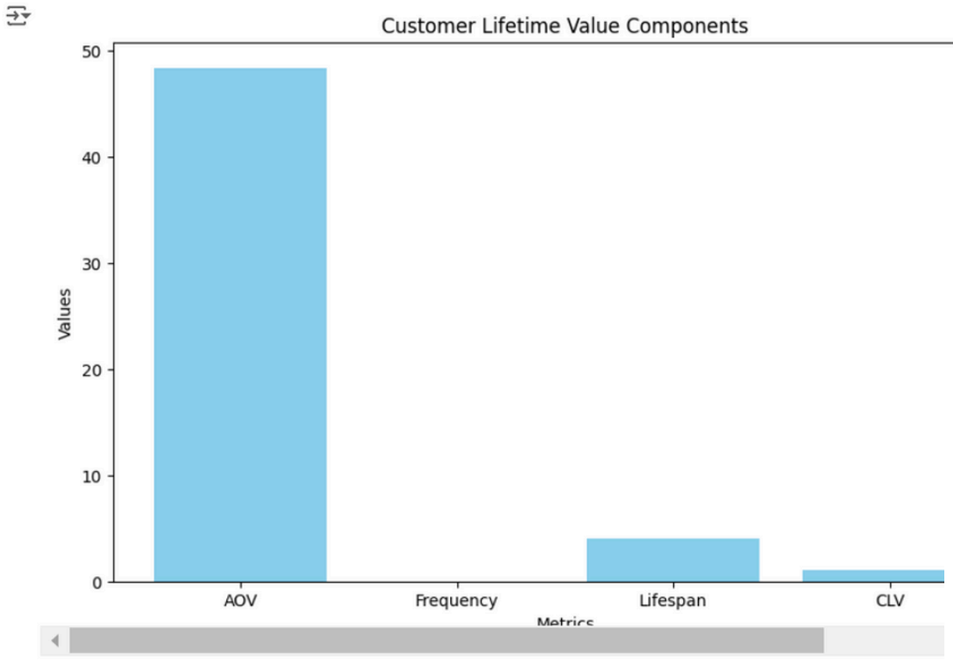
Calculate Customer Lifetime Value

```
# Calculate Customer Lifetime Value (CLV)
CLV = AOV * frequency * expected_customer_lifespan
print("Customer Lifetime Value (CLV):", CLV)
```

```
Customer Lifetime Value (CLV): 1.0971276188732182
```

Codes for Plotting and Analysis

```
# Visualize components of CLV
labels = ['AOV', 'Frequency', 'Lifespan', 'CLV']
values = [AOV, frequency, expected_customer_lifespan, CLV]

plt.figure(figsize=(10, 6))
plt.bar(labels, values, color='skyblue')
plt.title('Customer Lifetime Value Components')
plt.xlabel('Metrics')
plt.ylabel('Values')
plt.show()
```



1. Cohort Analysis

Understanding the cohorts with the help of CLV

```
# Extract Cohort Month
data['CohortMonth'] = data['Dt_Customer'].dt.to_period('M')

# Calculate Cohort Index
data['CohortIndex'] = (data['Dt_Customer'].dt.year - data['Dt_Customer'].dt.year.min()) * 12 + (data['Dt_Customer'].dt.month - data['Dt_Cust

# Cohort Analysis for CLV
cohort_data = data.groupby(['CohortMonth', 'CohortIndex'])['Monetary'].mean().unstack(0)

# Visualize Cohort Analysis
plt.figure(figsize=(12, 8))
sns.heatmap(cohort_data, annot=True, cmap='Blues')
plt.title('Cohort Analysis - Average Monetary Value')
plt.xlabel('Cohort Month')
plt.ylabel('Cohort Index')
plt.show()
```

Codes for Plotting and Analysis

2. Segmentation Analysis with CLV

```
# Create a new column for CLV segmentation
data['CLV_Segment'] = pd.qcut(data['Monetary'], 4, labels=['Low', 'Medium', 'High', 'Very High'])

# Analyze CLV across segments
segment_analysis = data.groupby('CLV_Segment')['Monetary'].mean()

print(segment_analysis)

# Visualize CLV Segments
plt.figure(figsize=(10, 6))
sns.barplot(x=segment_analysis.index, y=segment_analysis.values)
plt.title('Average Monetary Value by CLV Segment')
plt.xlabel('CLV Segment')
plt.ylabel('Average Monetary Value')
plt.show()
```

```
# Convert results to a DataFrame
sensitivity_df = pd.DataFrame(sensitivity_results, columns=['Income Multiplier', 'Frequency Multiplier', 'Average CLV'])

# Use pivot_table for better formatting
sensitivity_pivot = sensitivity_df.pivot_table(index='Income Multiplier', columns='Frequency Multiplier', values='Average CLV')

# Adjust the plot size and format tick labels
plt.figure(figsize=(14, 10))
sns.heatmap(sensitivity_pivot, annot=True, fmt=".2f", cmap='coolwarm', cbar_kws={'label': 'Average CLV'})
plt.title('Sensitivity Analysis of CLV', fontsize=16)
plt.xlabel('Frequency Multiplier', fontsize=12)
plt.ylabel('Income Multiplier', fontsize=12)

# Rotate the tick labels for readability
plt.xticks(rotation=45, ha='right', fontsize=10)
plt.yticks(fontsize=10)
plt.show()
```

Codes for Plotting and Analysis

Demographic Visualizations

Age Distribution: A histogram showing the distribution of customer ages.

Income Distribution: A histogram showing the distribution of customer income.

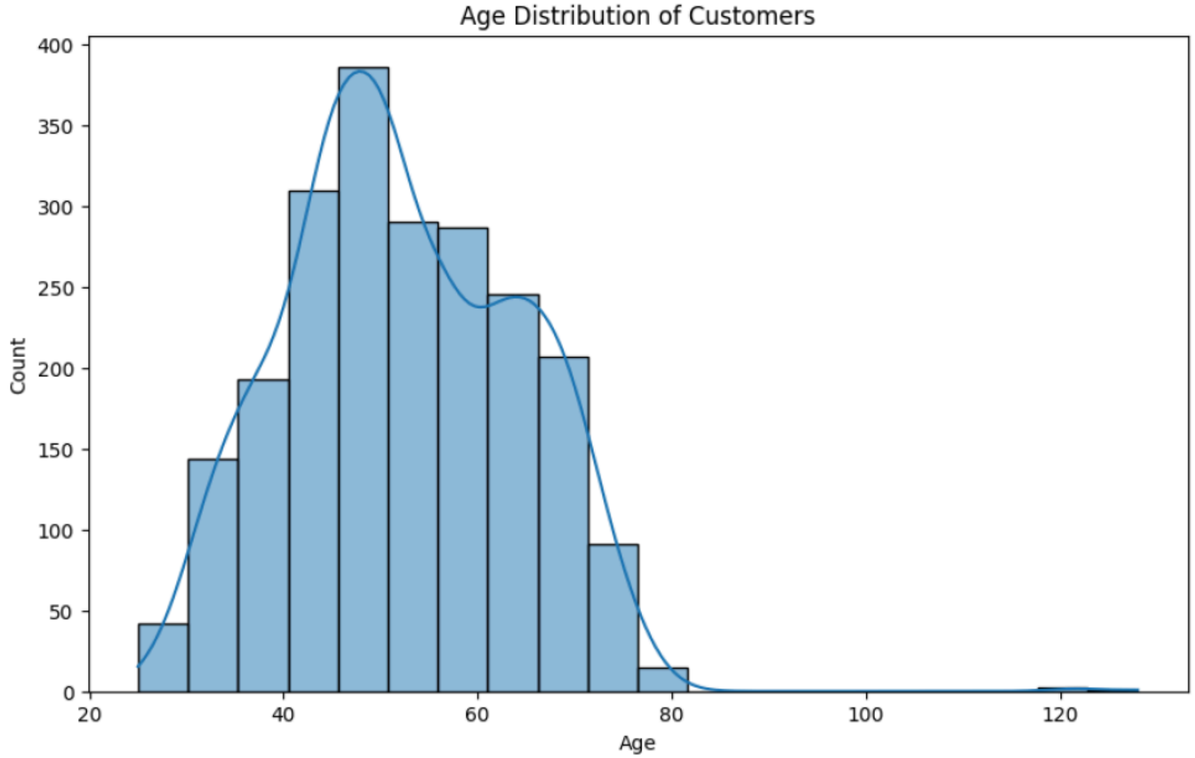
Marital Status: A bar chart showing the count of customers by marital status.

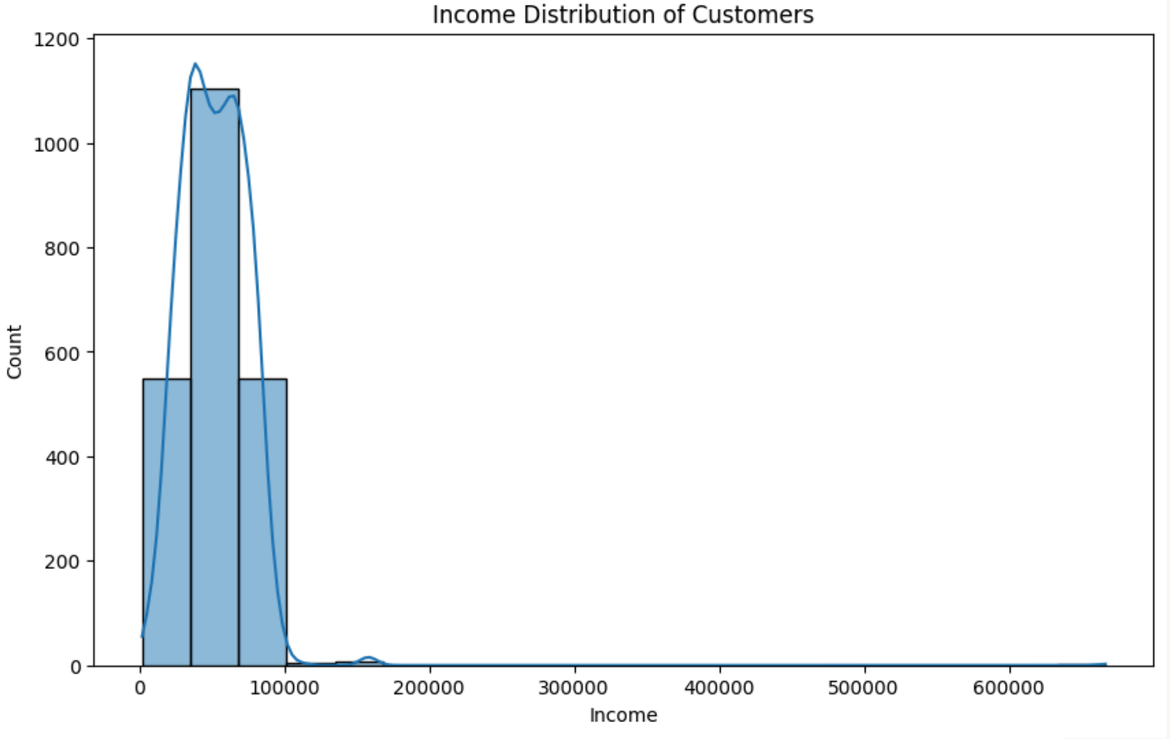
Education Level: A bar chart showing the distribution of customers by education level.

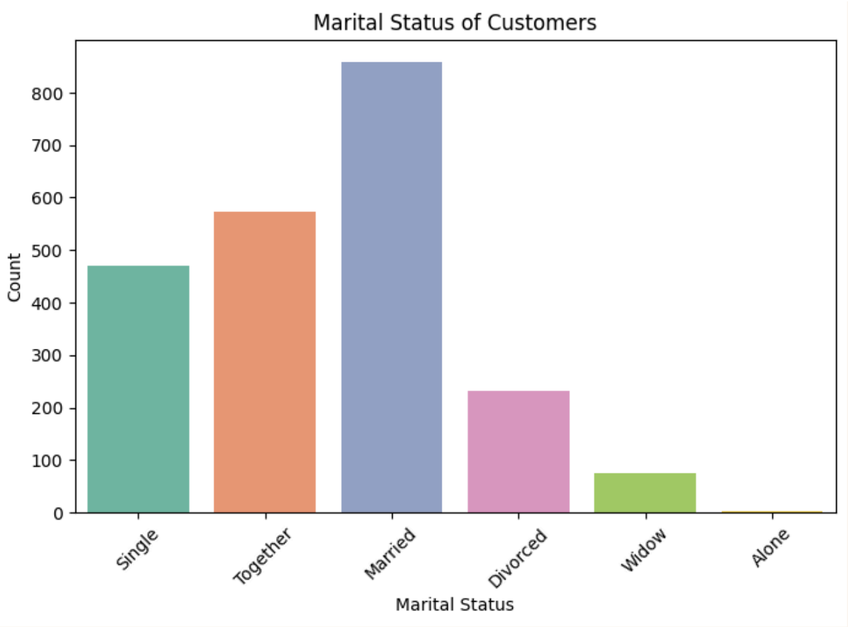
Income vs. Age: A scatter plot showing the relationship between age and income, colored by marital status.

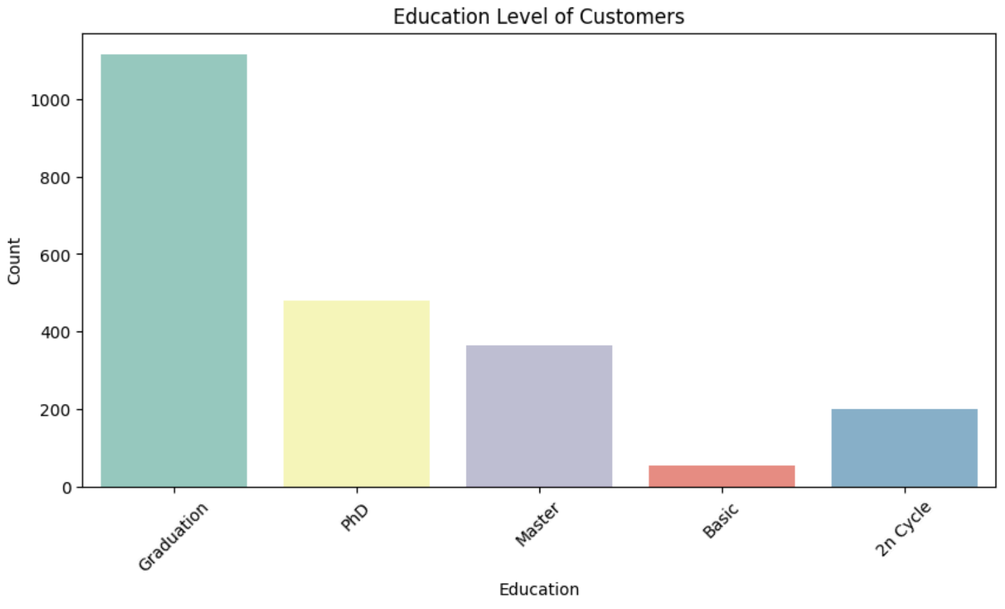
```
# Import necessary libraries for visualization
import seaborn as sns

# 1. Age Distribution
plt.figure(figsize=(10, 6))
sns.histplot(data['Age'], bins=20, kde=True)
plt.title('Age Distribution of Customers')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()

# 2. Income Distribution
plt.figure(figsize=(10, 6))
sns.histplot(data['Income'], bins=20, kde=True)
plt.title('Income Distribution of Customers')
plt.xlabel('Income')
plt.ylabel('Count')
plt.show()
```

```
# 3. Marital Status Distribution
plt.figure(figsize=(8, 5))
sns.countplot(x='Marital_Status', data=data, palette='Set2')
plt.title('Marital Status of Customers')
plt.xlabel('Marital Status')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()

# 4. Education Level Distribution
plt.figure(figsize=(10, 5))
sns.countplot(x='Education', data=data, palette='Set3')
plt.title('Education Level of Customers')
plt.xlabel('Education')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()

# 5. Income vs. Age Scatter Plot
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Age', y='Income', data=data, hue='Marital_Status', palette='viridis')
plt.title('Income vs. Age Scatter Plot')
plt.xlabel('Age')
plt.ylabel('Income')
plt.show()
```


Age Distribution of Customers
Count
Age
0
50
100
150
200
250
300
350
400
20
40
60
80
100
120


Income Distribution of Customers
Count
Income
0
200
400
600
800
1000
1200
0
100000
200000
300000
400000
500000
600000


Marital Status of Customers
Count
Marital Status
0
100
200
300
400
500
600
700
800
Single
Together
Married
Divorced
Widow
Alone

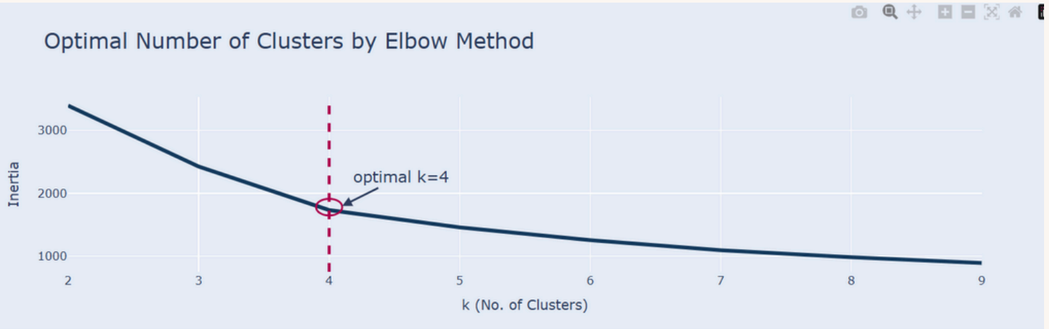

Education Level of Customers
Count
Education
0
200
400
600
800
1000
Graduation
PhD
Master
Basic
2n Cycle

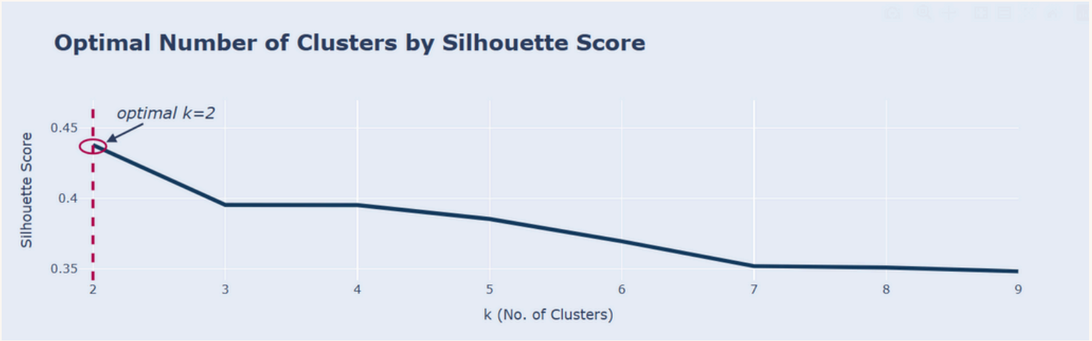
Determine The Number of Clusters

There are two methods which frequently used to find the optimal number of clusters.

First, the Elbow Method where each inertia (mean squared distance) will be plotted against the number of cluster in ascending order. Generally, the optimal k is the last number of cluster that reduce the inertia significantly until adding more cluster will not help much.

Second, calculating the Silhouette Score, which is the mean silhouette coefficient over all the instances. Results that vary between -1 and +1. A value close to +1 means the instance is well assigned inside its cluster and far from the other cluster, while -1 indicates it may have been incorretly assigned.





The optimal cluster based on the elbow method is 4, while according to silhouette score is 2. Since silhouette score method considers intra and inter clusters, it might produce more separated clusters.
However, two clusters might be too general for customer segmentation which require more specific cluster to build personalized offers. Thus, I will use k=4 to perform clustering.