PS Ifood - Case

Contexto: Construir modelo de classificação para suportar iniciativas de marketing

PS Ifood - Case

Mas antes de chegar lá, passamos pelo conhecimento de público e tratamento dos dados…

1. Data Transformation, Exploration, and Analysis

Person

Attribute	Description
ID	Customer's unique identifier
Year_Birth	Customer's birth year
Education	Customer's education level
Marital_Status	Customer's marital status
Income	Customer's yearly household income
Kidhome	Number of children in customer's household
Teenhome	Number of teenagers in customer's household
Dt_Customer	Date of customer's enrollment with the company
Recency	Number of days since customer's last purchase
Complain	1 if the customer complained in the last 2 years, 0 otherwise

Expenses

Attribute	Description
MntWines	Amount spent on wine in last 2 years
MntFruits	Amount spent on fruits in last 2 years
MntMeatProducts	Amount spent on meat in last 2 years
MntFishProducts	Amount spent on fish in last 2 years
MntSweetProducts	Amount spent on sweets in last 2 years
MntGoldProds	Amount spent on gold in last 2 years

Channels

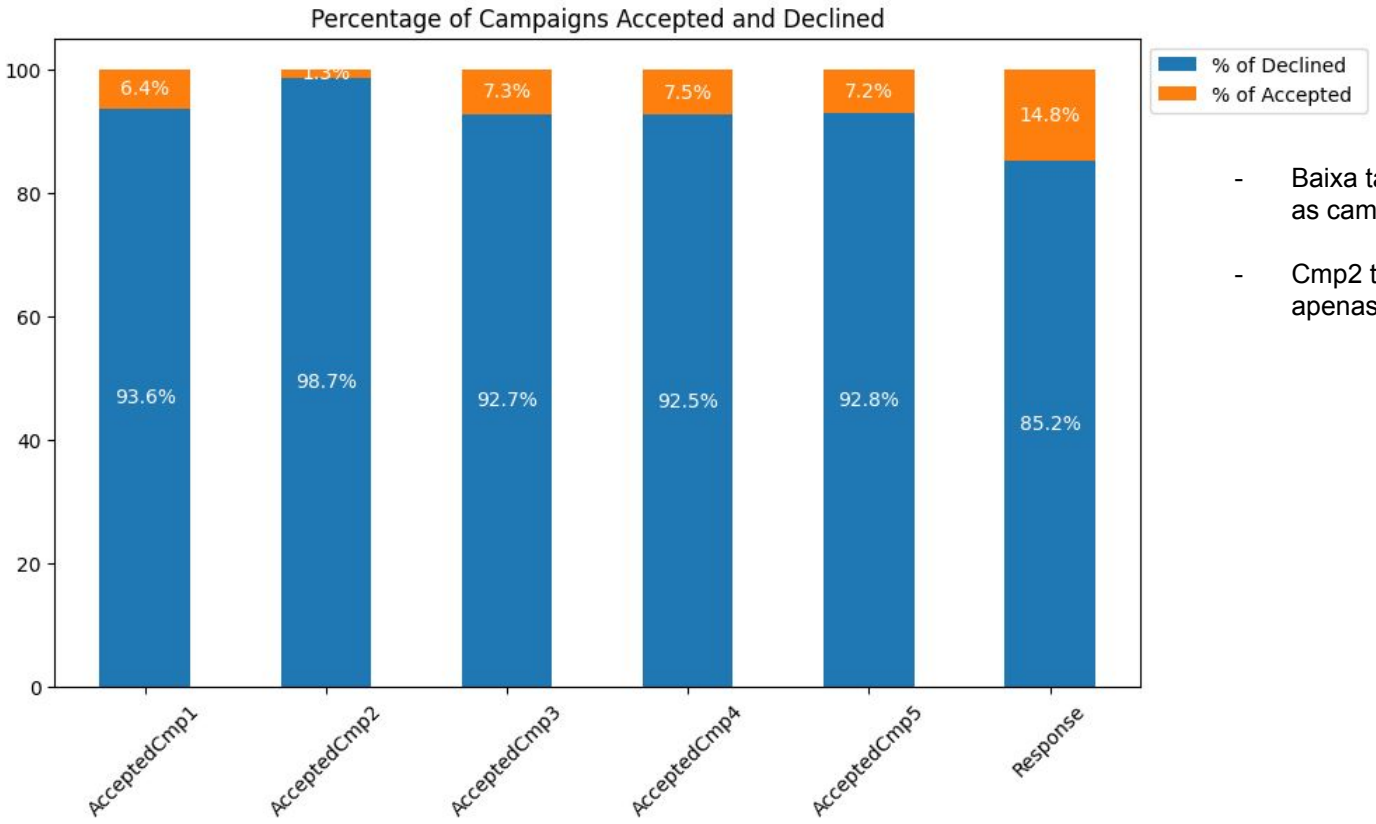
Attribute	Description
NumWebPurchases	Number of purchases made through the company's website
NumCatalogPurchases	Number of purchases made using a catalogue
NumStorePurchases	Number of purchases made directly in stores
NumWebVisitsMonth	Number of visits to company's website in the last month

Campaigns/Promotions

Attribute	Description
NumDealsPurchases	Number of purchases made with a discount
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

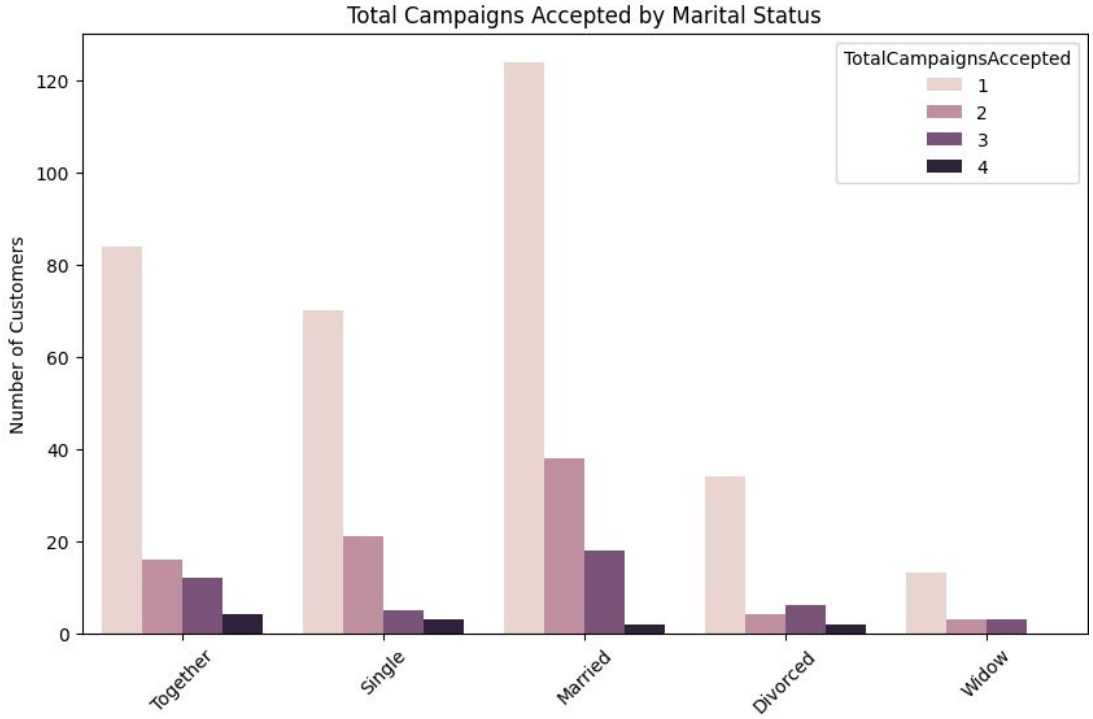
["Z_CostContact", "Z_Revenue"], duas outras variáveis presentes no dataset foram removidas pois era constantes e não trariam nenhum grau de explicabilidade para as análises/modelo.

Campaign Acceptance Rates



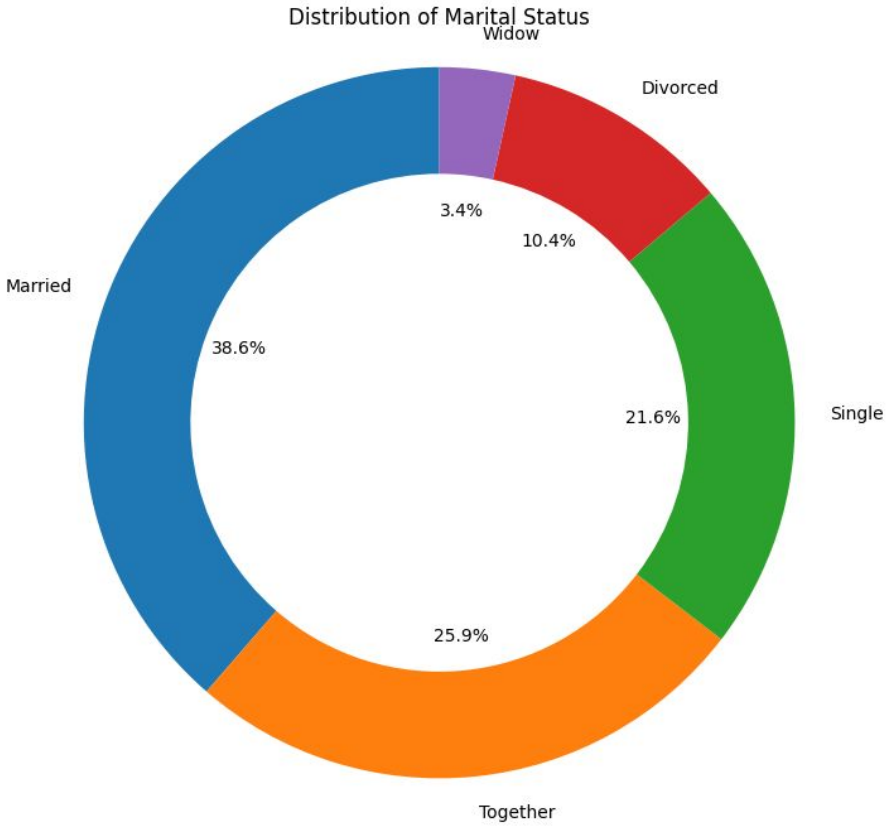
- Baixa taxa de aceitação para todas as campanhas.
- Cmp2 tem o pior desempenho, apenas 1.3% de aceitação.

Customers by Campaigns Accepted



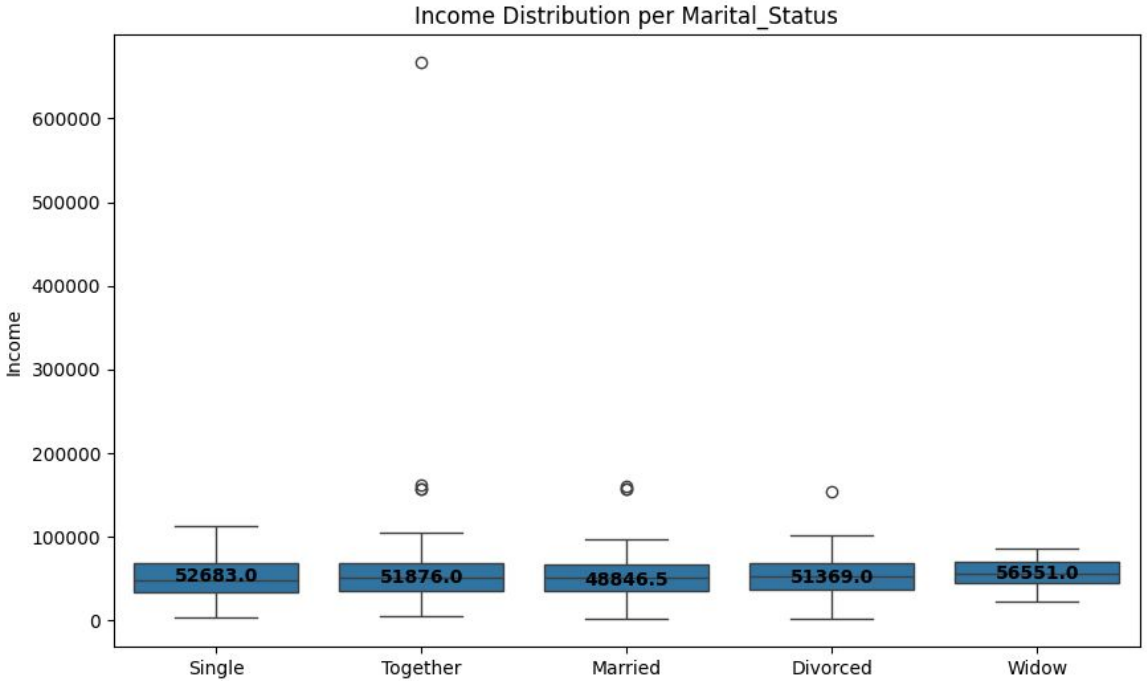
- Pessoas aceitam em média 1 campanha.
- Número de pessoas que aceita uma segunda campanha cai drasticamente.
- Dentro de Widow, não existe ninguém que aceitou 4 campanhas.
- Ninguém aceitou 5 campanhas.

Marital Status Share



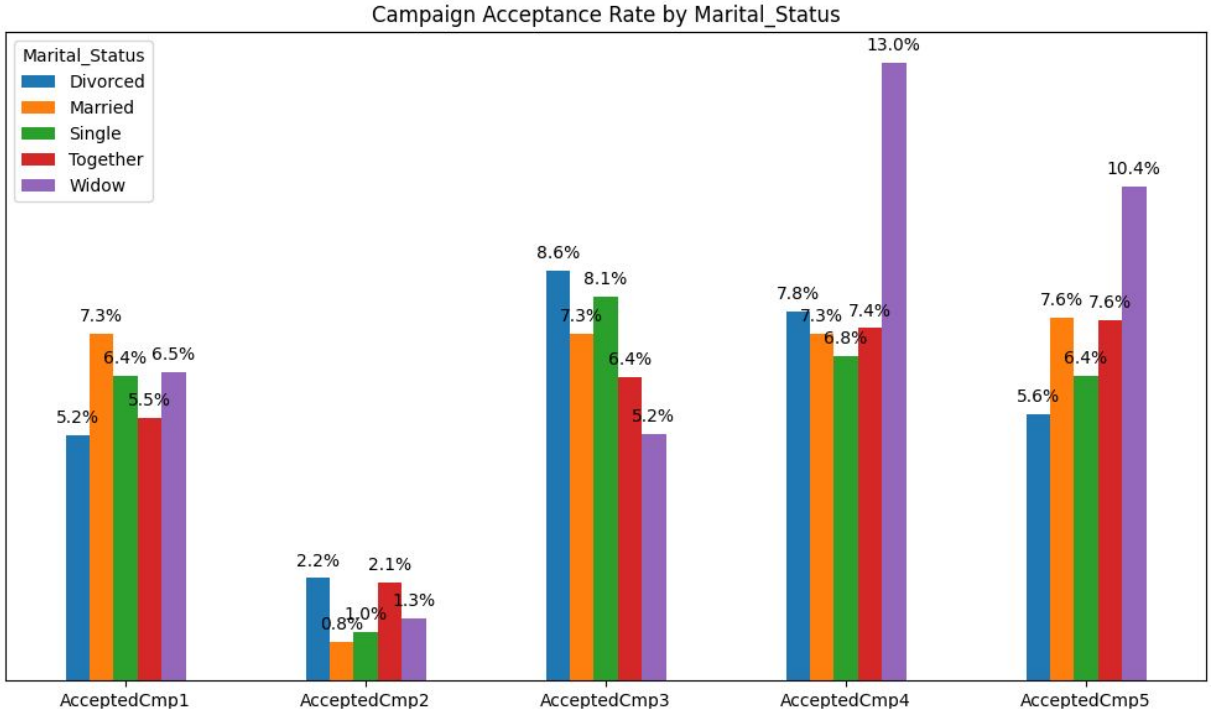
- Mais de 60% da base corresponde à pessoas morando com outras pessoas (Married + Together).

Income distribution



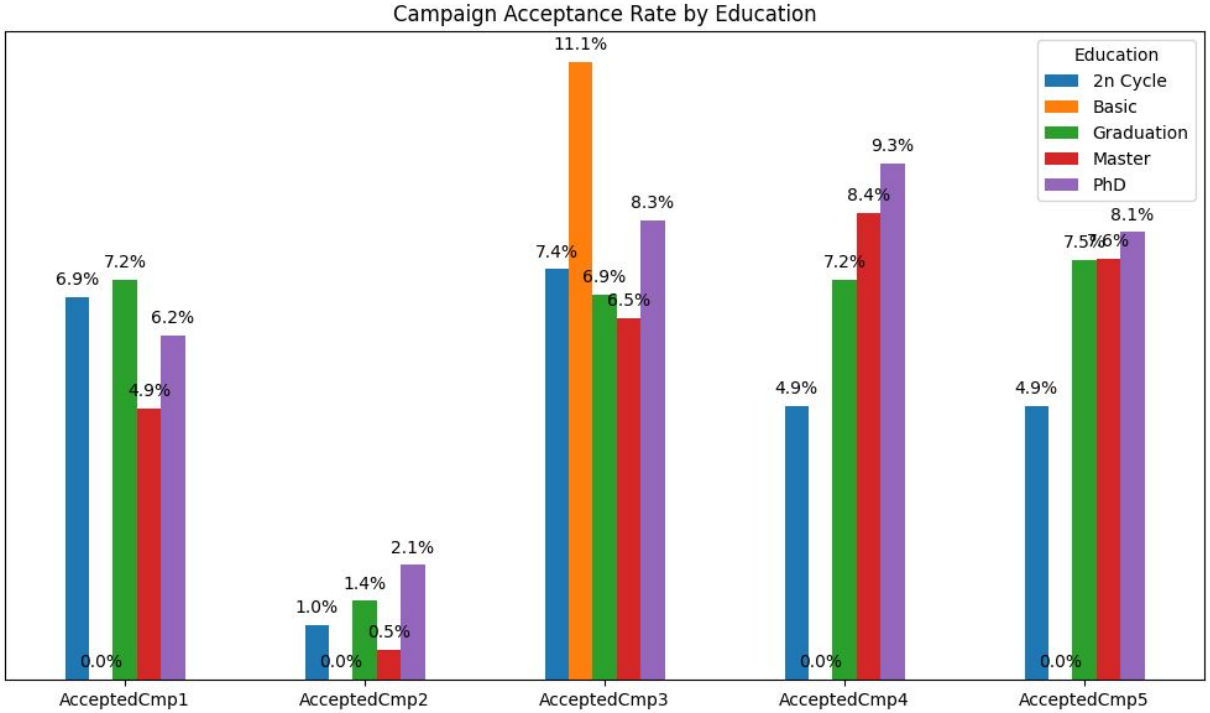
- Mediana Renda muito parecida entre os grupos.
- Já percebemos a presença de outliers.
- Pessoas viúvas são o grupo de maior renda.

Acceptance Rate by Marital Status for Each Campaign



- Pessoas viúvas corresponderam à maior parcela daqueles que aceitaram as campanhas 4 e 5.

Acceptance Rate by Education for Each Campaign



- Pessoas com nível de educação básica, foram as que mas aceitaram à campanha 3.
- Em contrapartida, esse público não aceitou nenhuma das outras campanhas.

Deals Purchased by Marital Status

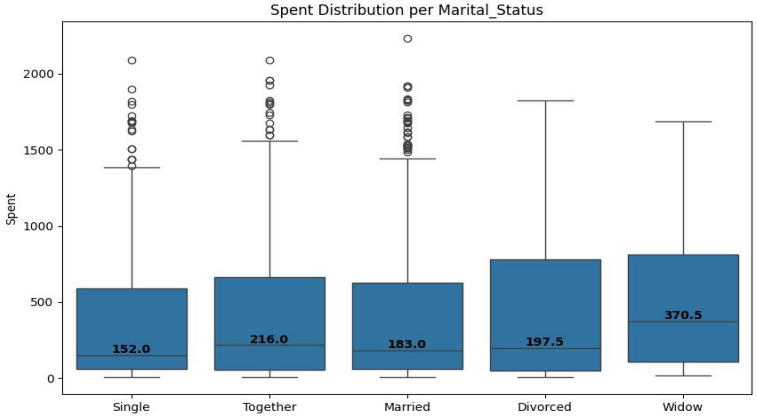
	Marital_Status	TotalPurchases	NumDealsPurchases	% DealsPurchases
0	Divorced	3535	565	16.0
1	Married	12922	2067	16.0
2	Single	6877	1034	15.0
3	Together	8594	1348	15.7
4	Widow	1286	180	14.0

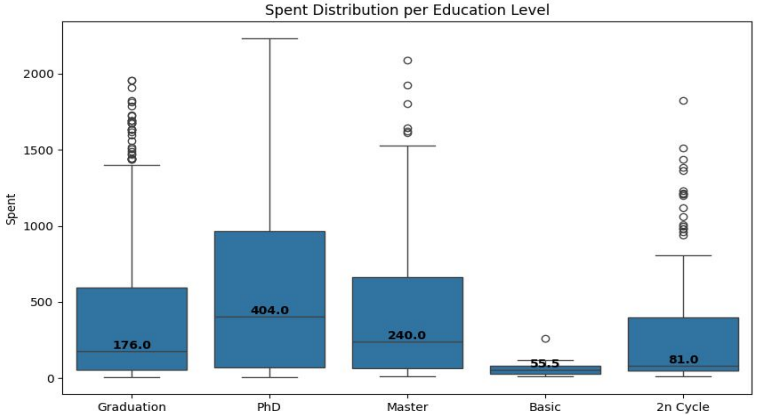
Deals Purchased by Education

	Education	TotalPurchases	NumDealsPurchases	% DealsPurchases
0	2n Cycle	2802	456	16.3
1	Basic	379	97	25.6
2	Graduation	16872	2602	15.4
3	Master	5506	895	16.3
4	PhD	7655	1144	14.9

- Percentual de compras com cupom, independe de Marital Status.
- Escolaridade baixa parece ter mais efeito na busca por compras com cupom. 25,6% das compras de quem tem no máximo Educação Básica, estão atreladas à cupom.

Spending by Marital Status and Education



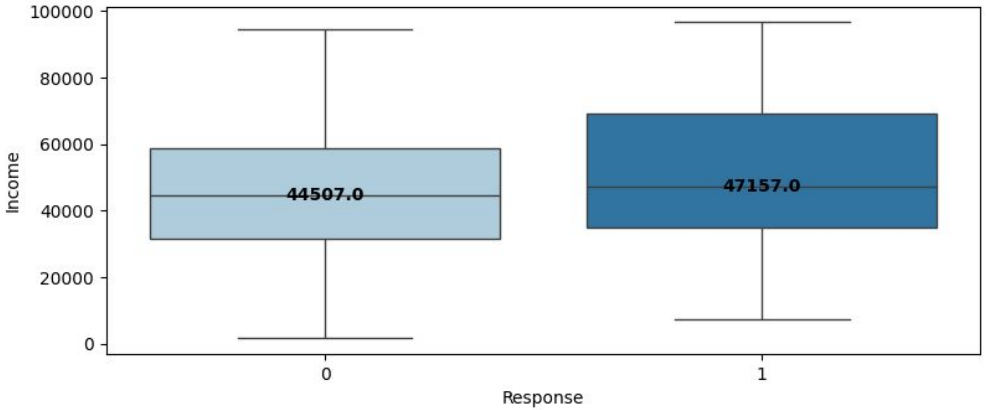


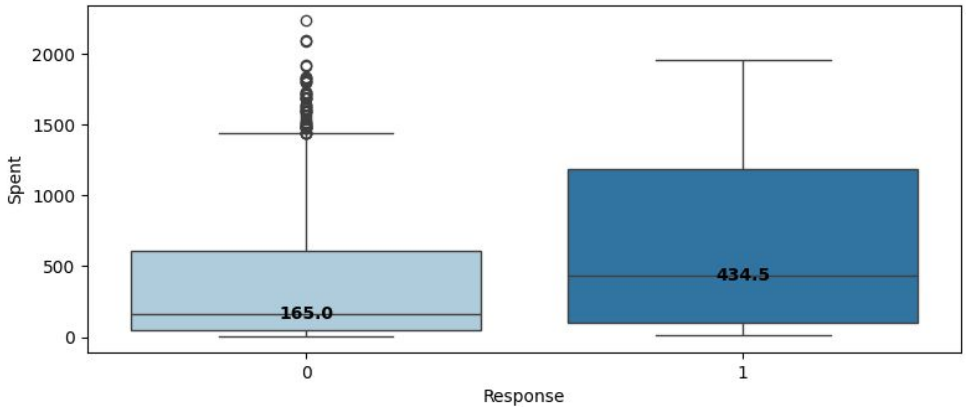
- Nível de gastos tem relação com Marital Status, mas varia mais pelo nível de escolaridade.
- Pessoas viúvas são quem mais gastam. Gastam quase o dobro da mediana dos outros grupos.
- Pessoas com PhD gastam 8x mais que o grupo que menos consome.

Response

Olhando para a variável Response (target)...

Income x Spent | Response



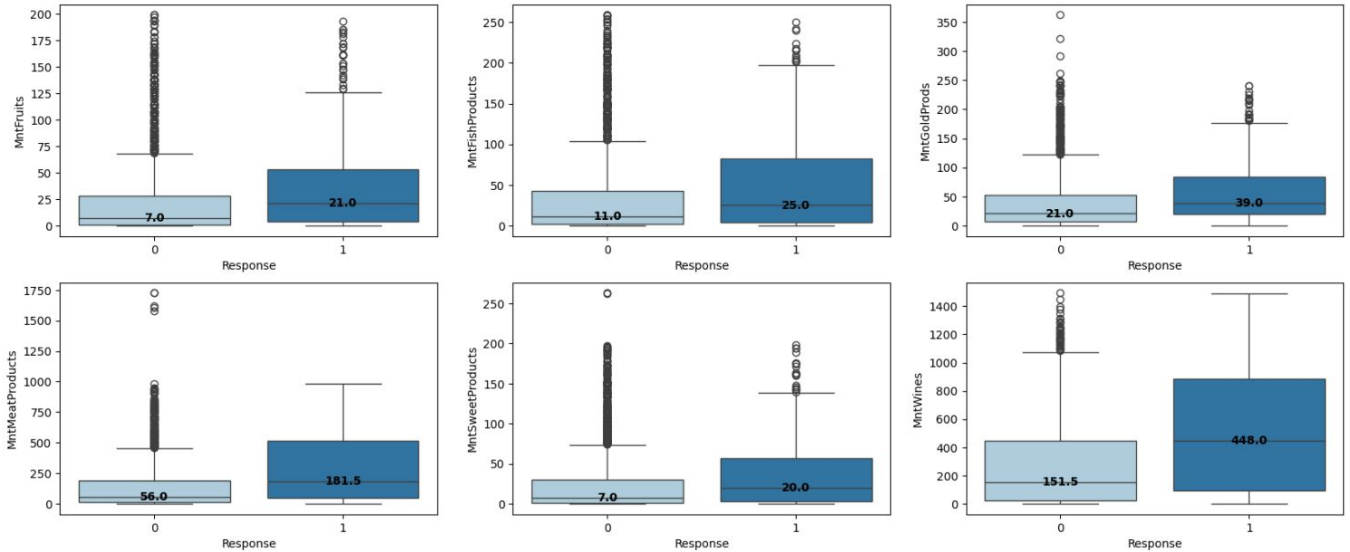


Para facilitar, chamaremos:
G1: Grupo que aderiu a Response
G0: Grupo que não aderiu à Response.

Embora à renda do grupo que aderiu (G1) e não aderiu à última campanha (G0) seja muito parecida, o grupo que adere gasta **2.6x** mais.

Amount Spent per Categories | Response

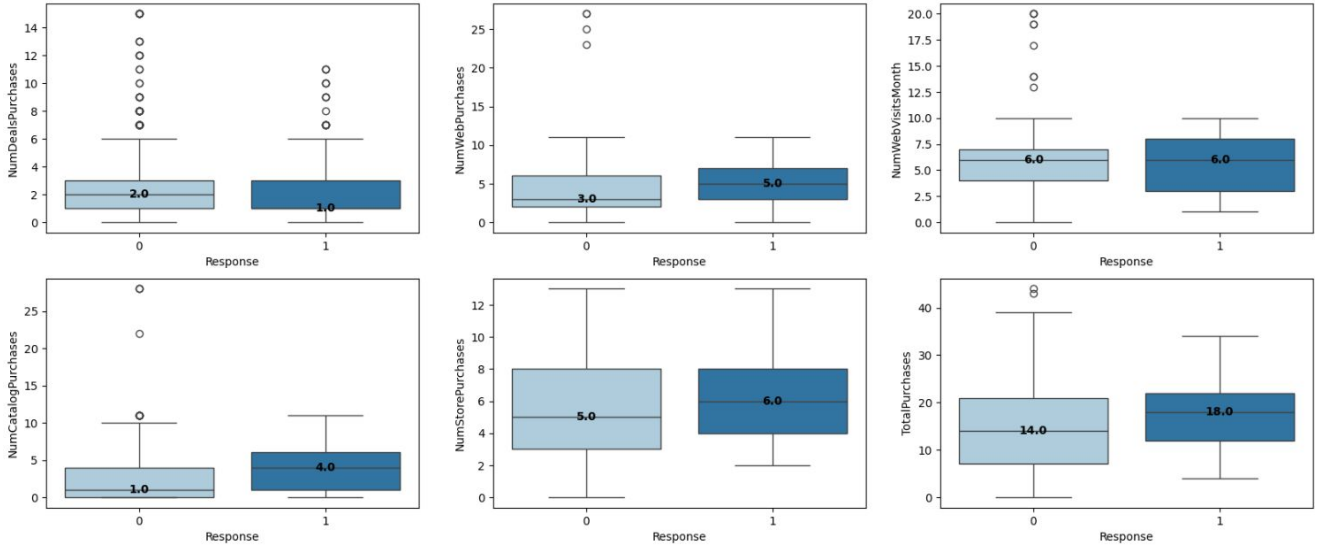
Quem respondeu tem maior padrão de consumo em todas as categorias.



O gasto de G1, é puxado principalmente por **Wine e Meat Products**.

Para as outras categorias G1 gasta entre **2-3x** mais que G0.

Purchases and Web visits | Response



O **G1** faz **menos compras com utilização de cupom**, embora tenha uma frequência alta de visitas web.

O **G0** tem a mesma mediana de visita ao site, que o **G1** É o grupo que mais faz compras com cupom.

2. Data Preparation and Cleaning for Customer Segmentation

Agora vamos remover os outliers...

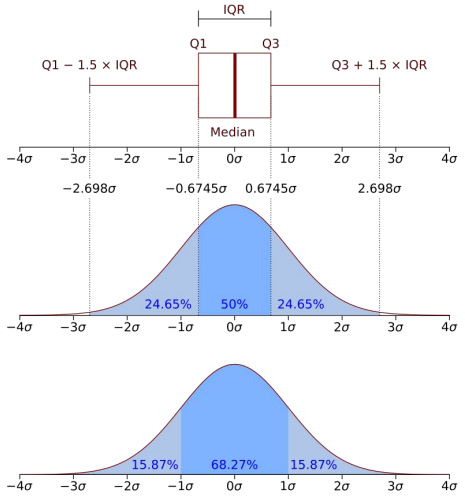
InterQuartile Range (IQR)

Metodologia utilizada para remoção de outliers

Outliers are typically defined as values below

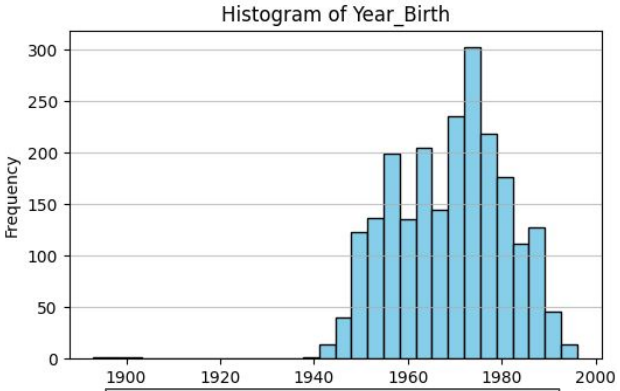
- Q1 - 1.5 x IQR
  or above
- Q3 + 1.5 x IQR

where
Q1 = First Quartil (25%),
Q3 = Third Quartil (75%) and
IQR = Q3 - Q1

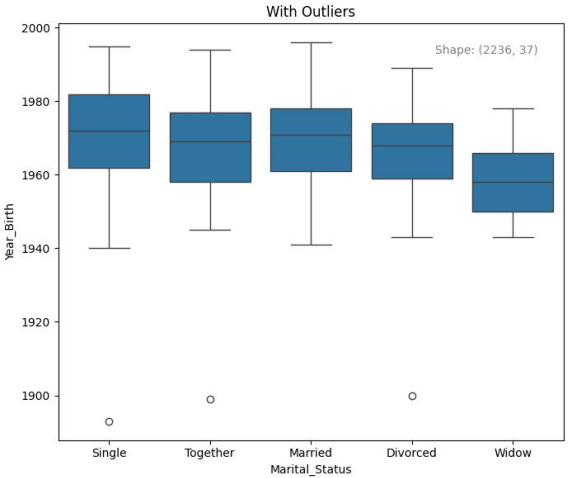


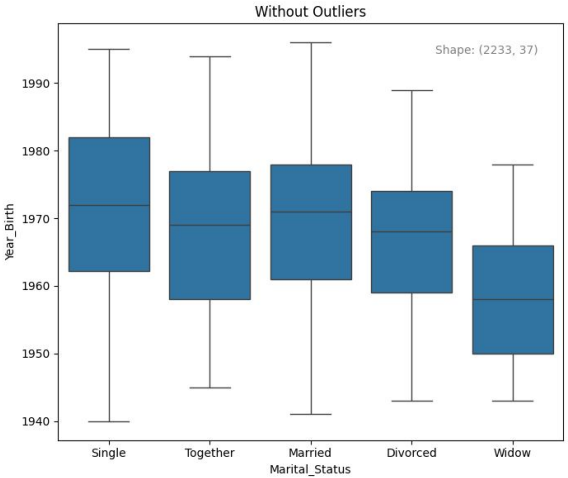
Nós iremos variar o multiplicador de IQR à depender da distribuição e da cauda da variável observada no histograma. Para variáveis com caudas longas, principalmente aquelas referentes à Montantes de Gastos, consideramos um multiplicador menos conservador (3) de forma a manter mais dados para criação dos modelos.

Year_Birth



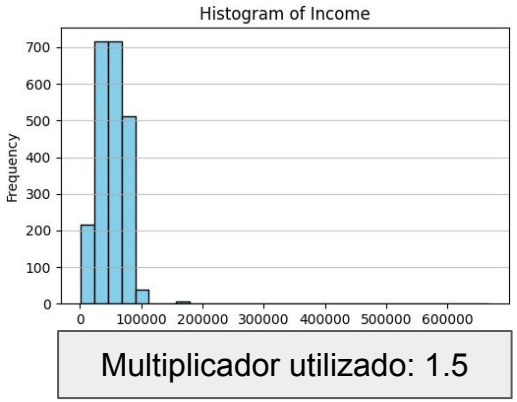
Multiplicador utilizado: 1.5



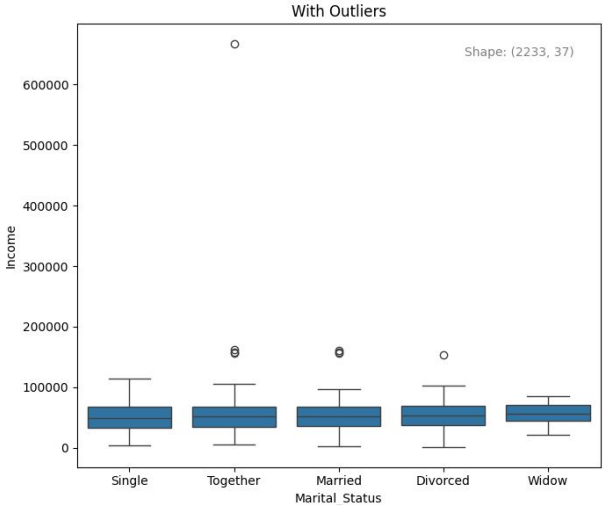


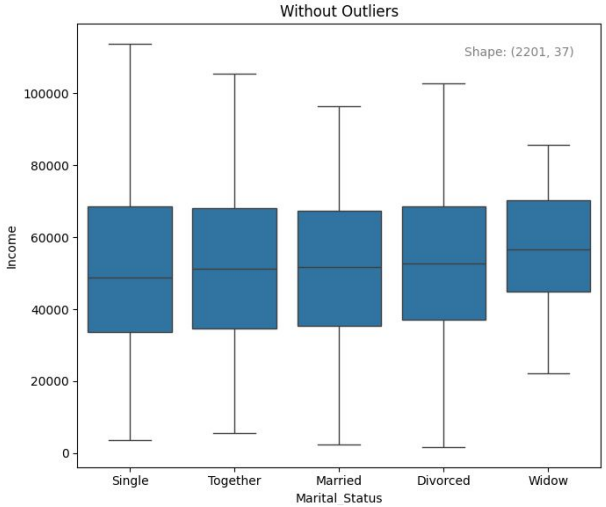
- Utilização de MaritageStatus apenas para facilitar visualização, a remoção de outliers considera unicamente Year_Birth.
- Após à remoção de outliers, nós perdemos -0.1% do nosso dataset.

Income



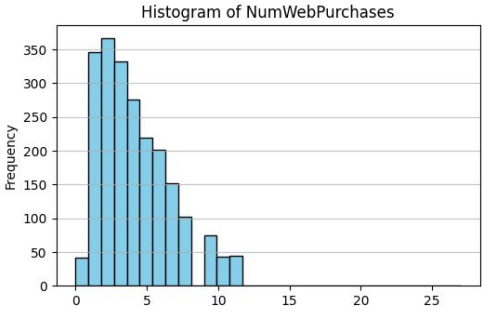
Multiplicador utilizado: 1.5



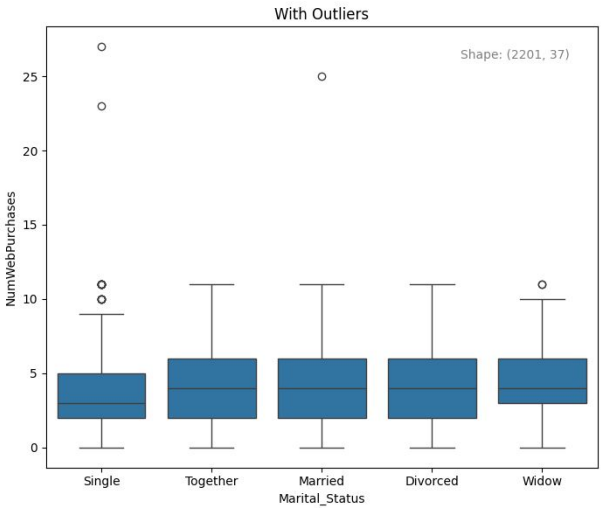


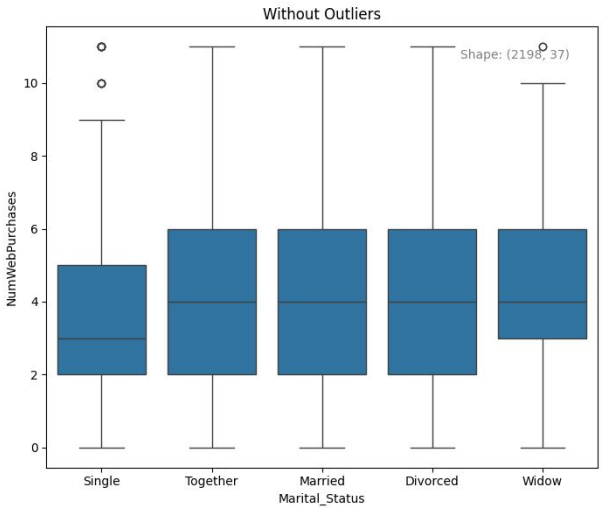
- Utilização de MaritageStatus apenas para facilitar visualização, a remoção de outliers considera unicamente Income.
- Após à remoção de outliers, nós perdemos -1.4% do nosso dataset.

NumWebPurchases



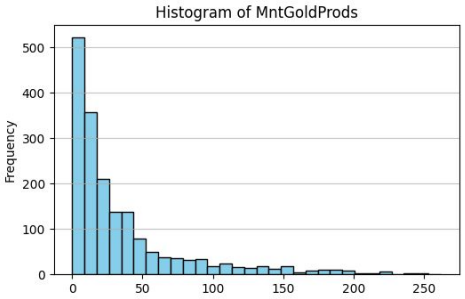
Multiplicador utilizado: 1.5



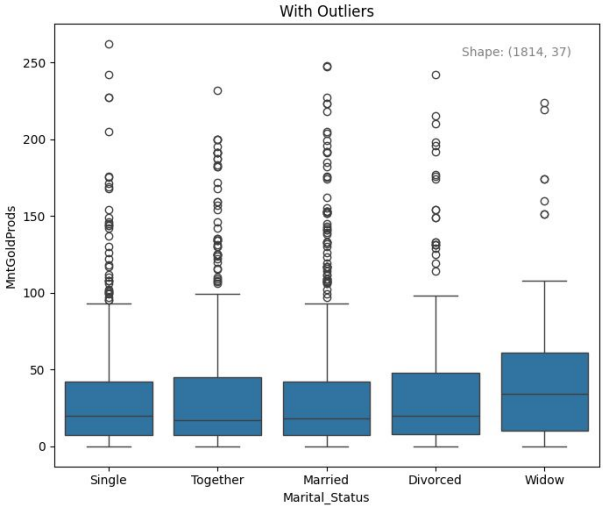


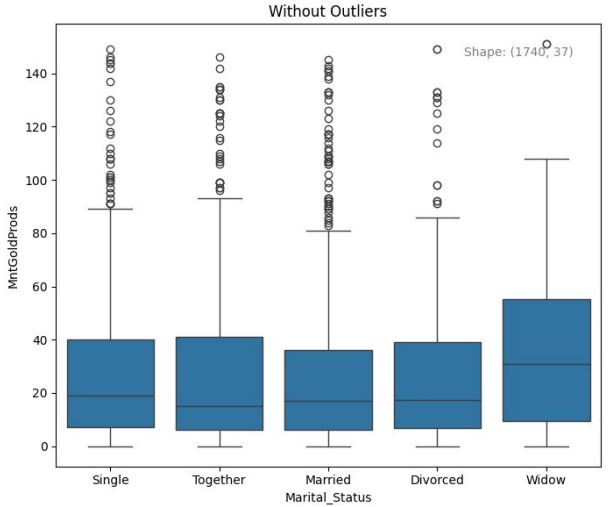
- Utilização de MaritageStatus apenas para facilitar visualização, a remoção de outliers considera unicamente NumWebPurchases.
- Após à remoção de outliers, nós perdemos -0.1% do nosso dataset.

MntGoldProds



Multiplicador utilizado: 3.0





- Utilização de MaritageStatus apenas para facilitar visualização, a remoção de outliers considera unicamente MntGoldProds.
- Após à remoção de outliers, nós perdemos -5.4% do nosso dataset.

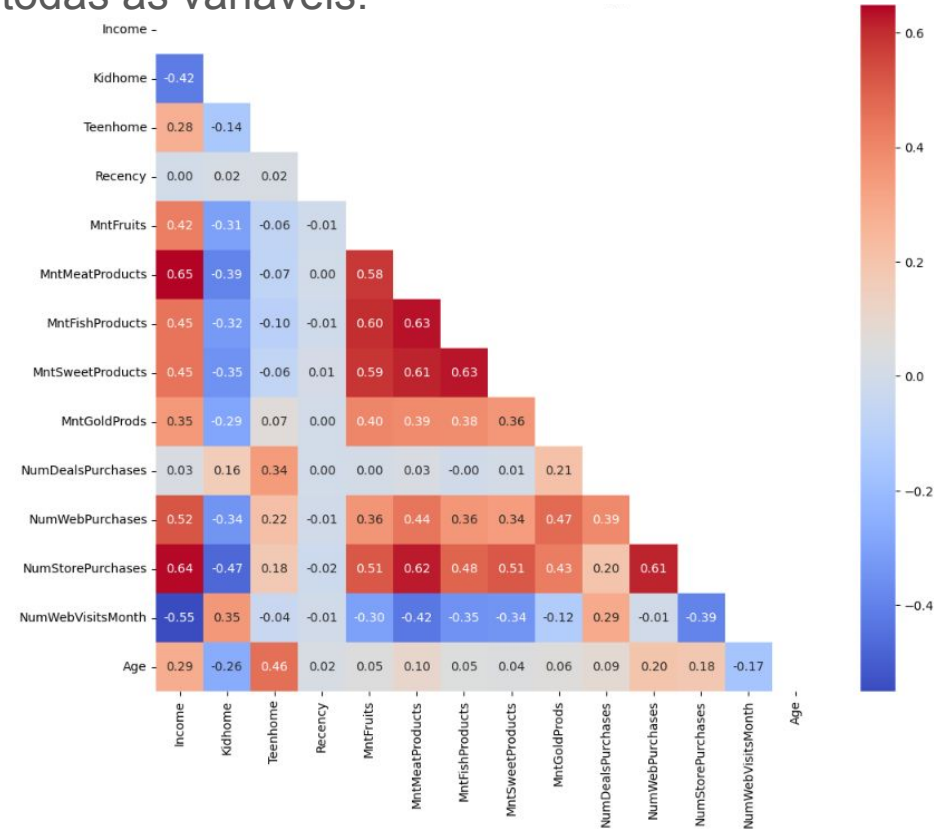
Após performar à limpeza, ficamos com 1.740 observações, 78% do dataset original.

3. Clusterização

Depois de feito os tratamentos, removido os outliers, podemos começar nosso processo de segmentação…

Correlation Matrix

KidHome: Ter crianças em casa, apresenta uma correlação negativa com quase todas as variáveis.

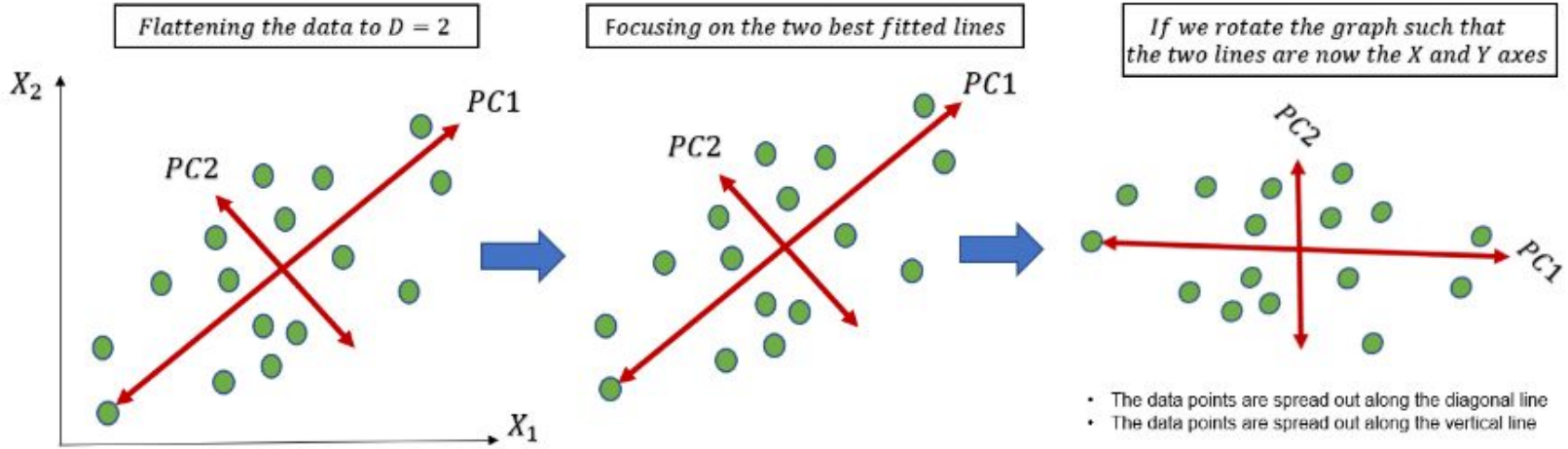


- KidHome has a significant negative correlation with almost all other features unlike TeenHome.

- To capture more variability from both variables, we created a new variable called ChildrenAtHome, which combines the two.

Principal Component Analysis (PCA)

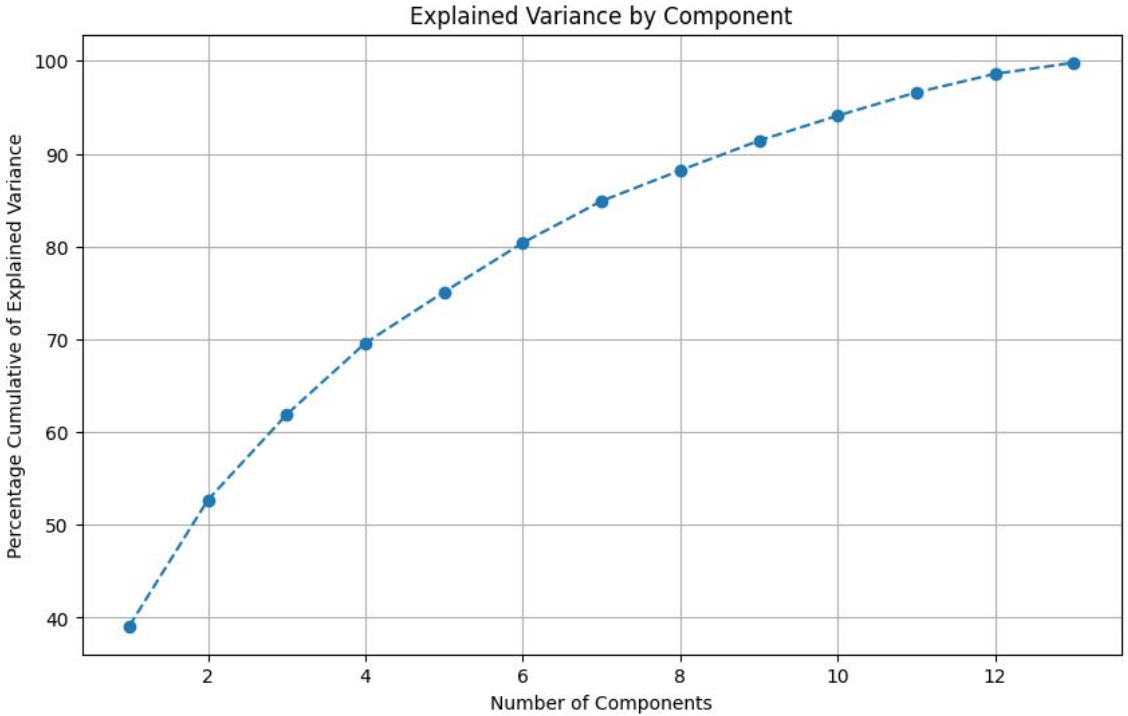
Performs linear transformation of the original data, to reduce dimensionality.



- first we plot X1 versus X2
- the two best fitting of these variables become our principal components.
- rotate to transform pc1 and pc2 in our X and Y axes
- now we have an adjusted combination of the two variables.

Principal Component Analysis (PCA)

Reduce the dimensionality of the original dataset, selecting a smaller number of components



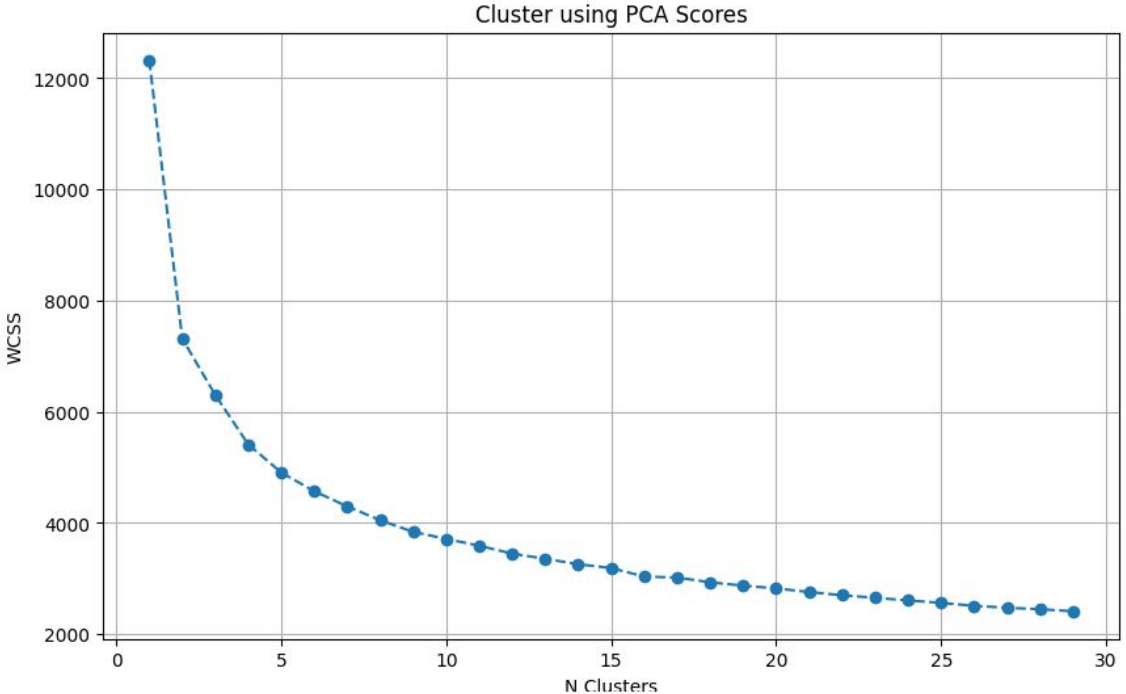
- From graph we can see that it's possible to retain almost 85% of the variation by using only 7 out of the 13 features

Number of Clusters

Vamos usar os 7 principais componentes obtidos a partir da abordagem PCA para definir o número de clusters…

Elbow Method

Observe how the sum of the clusters' variance decreases as the number of clusters increases



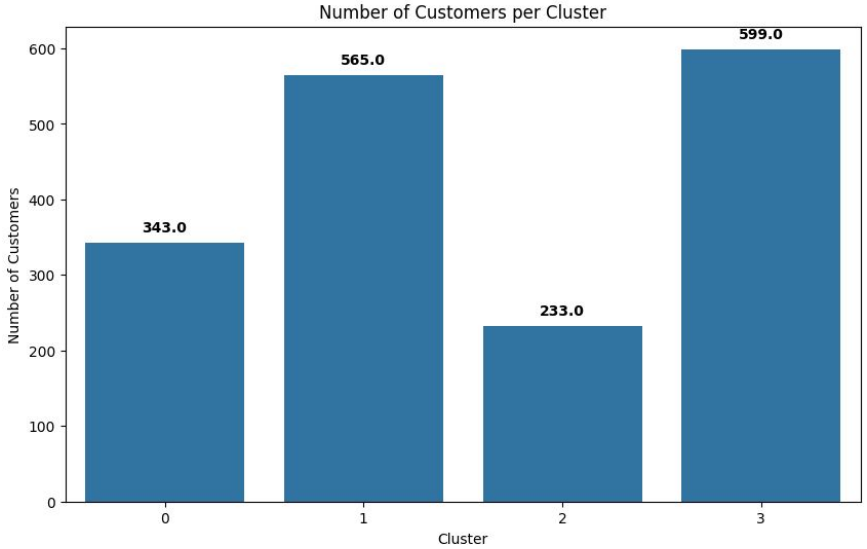
- Kmeans will perform n combinations of clusters, using the PCA scores.
- After that we take the sum of the clusters' variance and plot it.
- Visually (Elbow Method), we observe that using four clusters achieves a degree of stability in the process.

Cluster Analytics

Depois de definidos os clusters, adicionamos as informações de cada cluster ao dataset original e podemos analisar suas características…

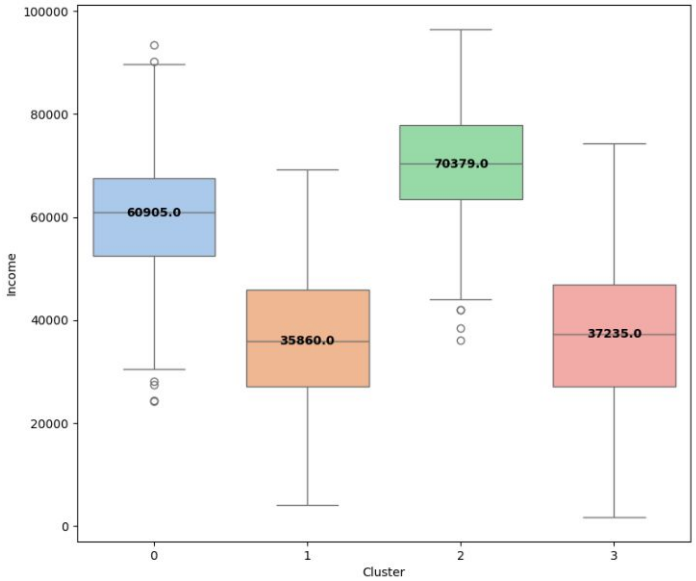
Customer per Cluster

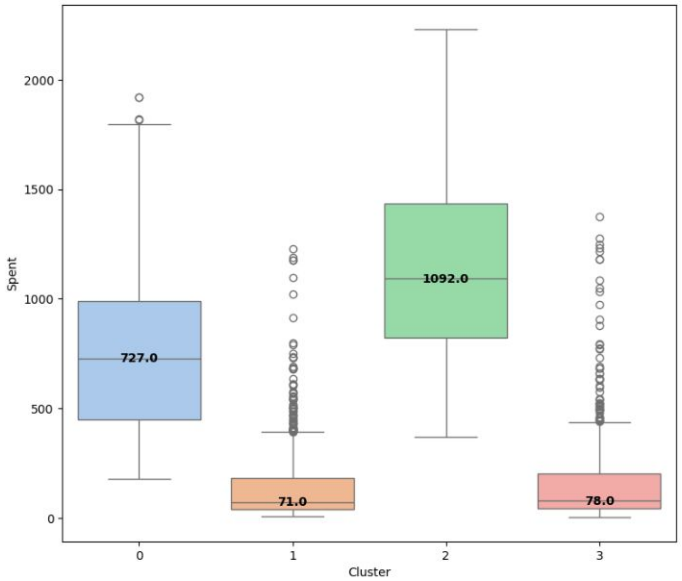
Clusters 1 e 3 tem a maior volumetria de clientes.



Income | Spent

No geral, renda está muito correlacionada com o padrão de consumo.

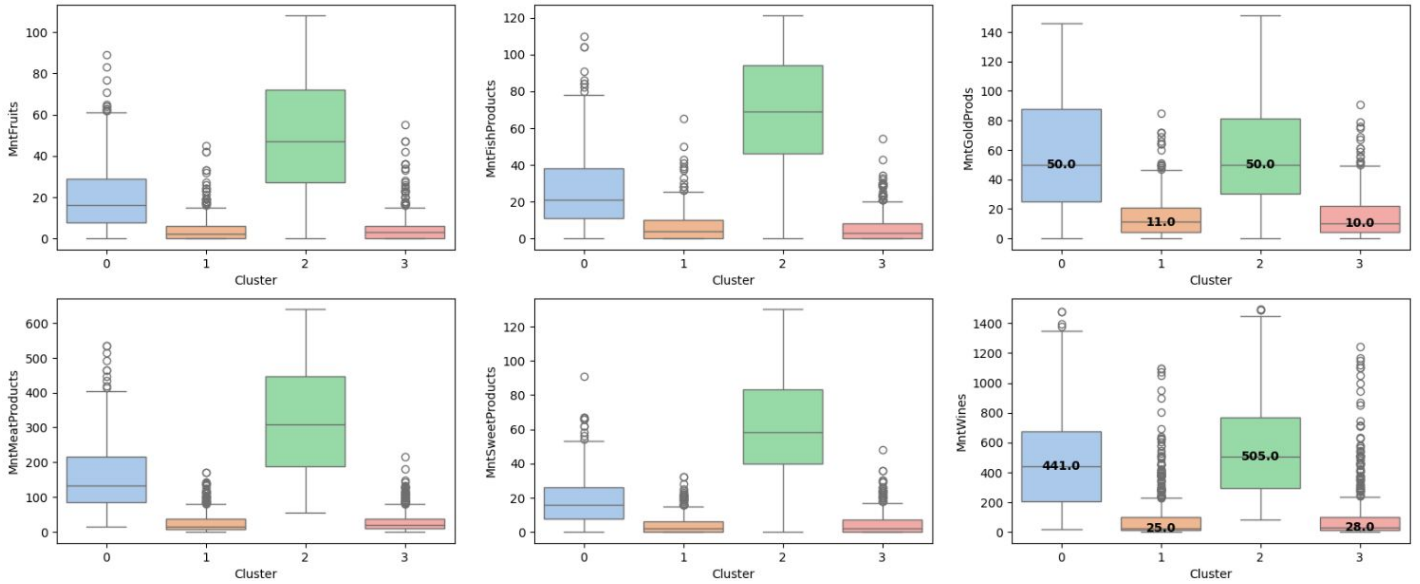




Cluster 2 se destaca com maior gasto e maior renda.

Cluster 1 e **Cluster 3** tem padrão de renda e de consumo parecidos.

Amounts Spent per Category

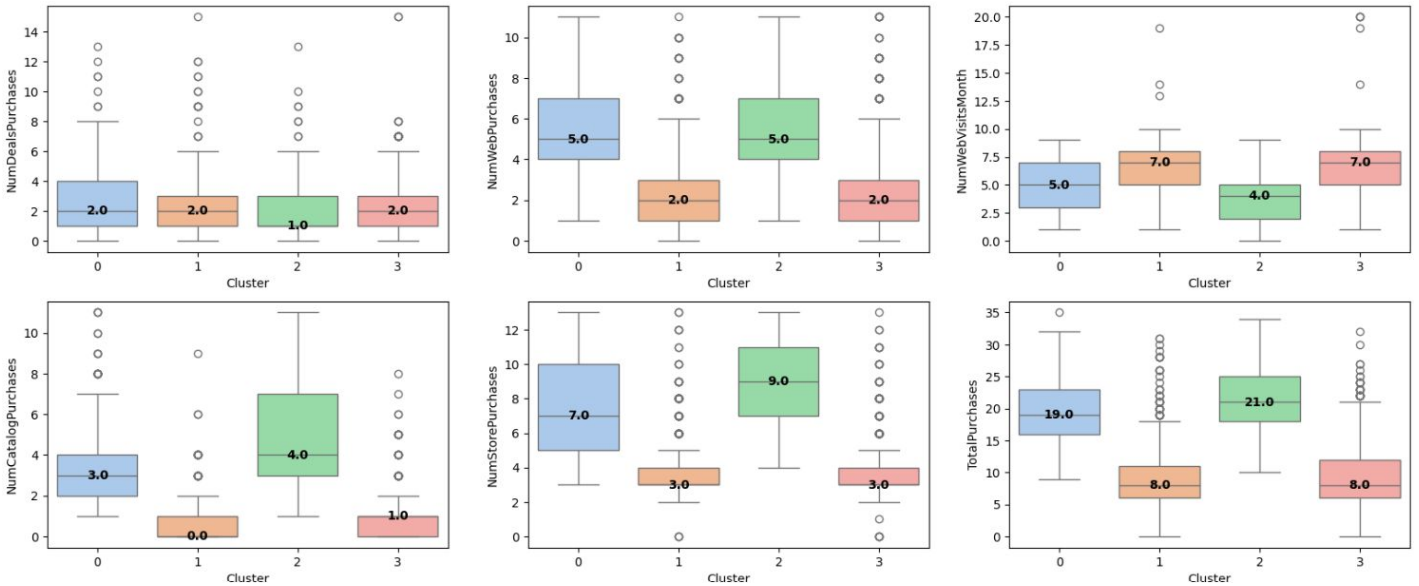


Cluster 2 é o que mais se destaca em todas as categorias de consumo.

Cluster 0 (2º maior renda) prioriza bens de alto valor agregado (Ouro e Vinhos).

Cluster 1 e **Cluster 3**, tem padrões de consumo baixos e parecidos.

Purchases and Web Visits

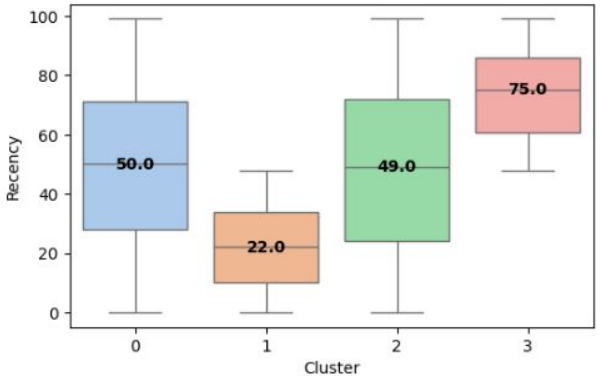


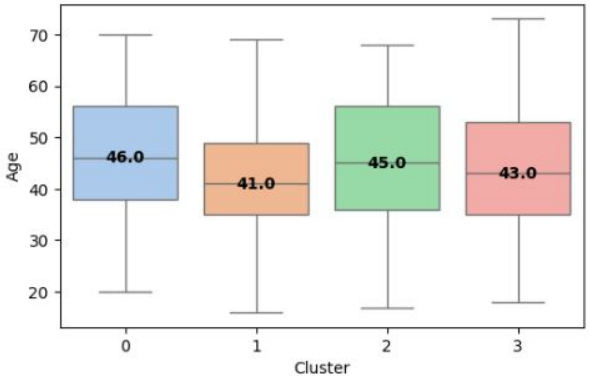
Cluster 2 é o que menos usa cupom, e o que menos visita o site.

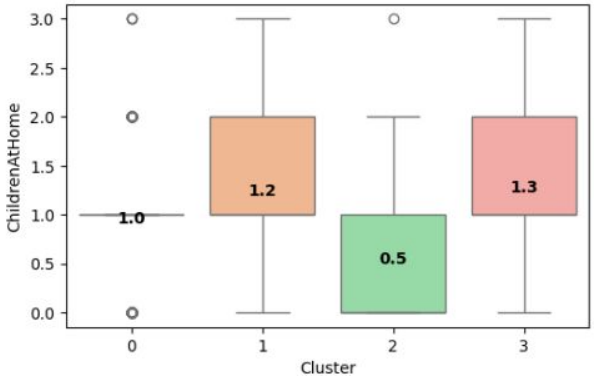
Cluster 0 visita bem o site, compra bastante na Web e tem à maior amplitude em cupom.

Cluster 1 e **Cluster 3** são os que mais visitam Web, mas compram pouco por lá.

Recency | Age | ChildrenAtHome







Cluster 2 o de maior renda, é o que têm menos filhos, e o 2º de maior idade.

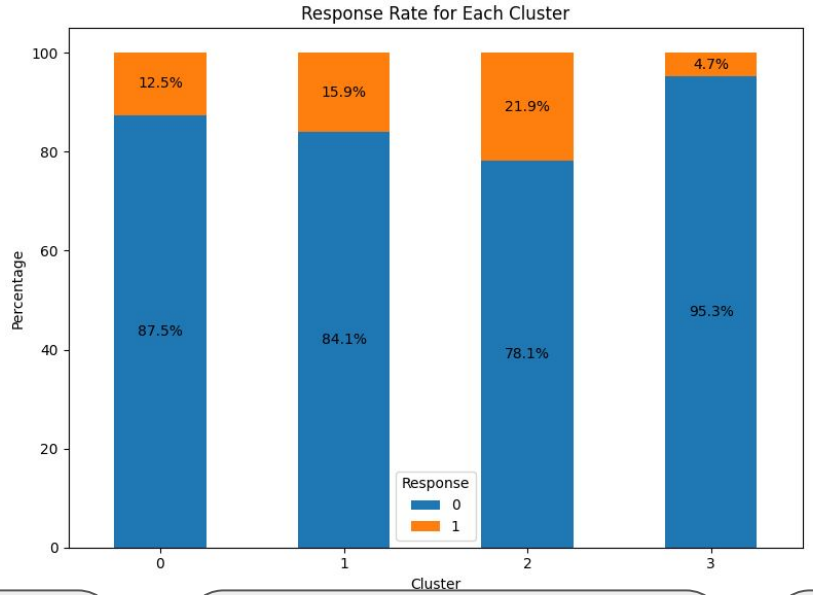
Cluster 0 é o de maior idade, segundo mais rico e com a segunda menor taxa de filhos.

Cluster 1 e **Cluster 3** tem à maior taxa de filhos e são os mais jovens.

Cluster 3 tem à maior taxa de filhos e faz tempo que não compra.

Response

Cluster 2 é o que tem à maior taxa de Response.



Cluster 2 além de ser o de maior renda, é o que tem à maior taxa de Response. Corrobora com as análises iniciais de Response x Income.

Cluster 3 tem a menor taxa de Response, mas tem uma alta taxa de visitas ao site. Além disso é o maior Cluster.

Cluster 1 tem a segunda maior taxa de Response, e é o 2º maior em volumetria. Aprendizados podem ajudar a melhorar a performance do **Cluster 3.**

Clusters

Cluster 0

Tem Cupom?

- 1.0 filhos
- 2º maior renda
- Visita bem o site
- Privilegia produtos de alto valor agregado (Ouro e Vinhos)
- 50 dias desde a última compra
- Maior amplitude no uso de cupom. (Metade desse público usa entre **2 e 4 cupons**)

Cluster 1

Vou dar uma voltinha depois eu passo aí…

- 1.2 filhos
- Menor Renda
- O mais recente (Apenas **22 dias** desde a última compra)
- **Visita muito o site**
- Mesmo volume de compra do Cluster 3 (mais antigo)

Cluster 2

Tô Podendo…

- 0.5 filhos
- Maior renda
- É o que menos visita o site.
- Tem um padrão de consumo alto, bem distribuído entre todas as categorias.
- O que menos usa cupom.

Cluster 3

Na volta à gente compra…

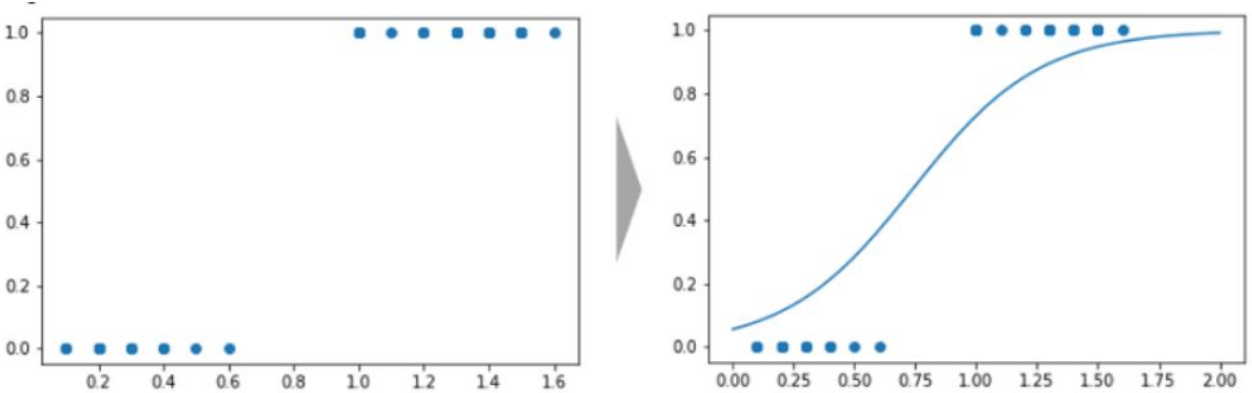
- 1.3 filhos
- 2º menor renda
- O mais antigo (**75 dias** desde a última compra)
- Visita muito o site
- mesmo volume de compra do Cluster 1

Predictive Model

Agora nosso modelo de previsão...

Predictive Model - Logistic Regression

Com as informações obtidas até aqui, podemos agora construir o modelo de previsão para a variável Response.



- Em vez de uma classificação determinística entre 0 e 1, a regressão logística nos permite obter uma relação contínua que estima a probabilidade de ocorrência de um evento. Isso nos fornece insights valiosos sobre a chance de um evento acontecer com base nas características dos usuários.

Predictive Model - Logistic Regression

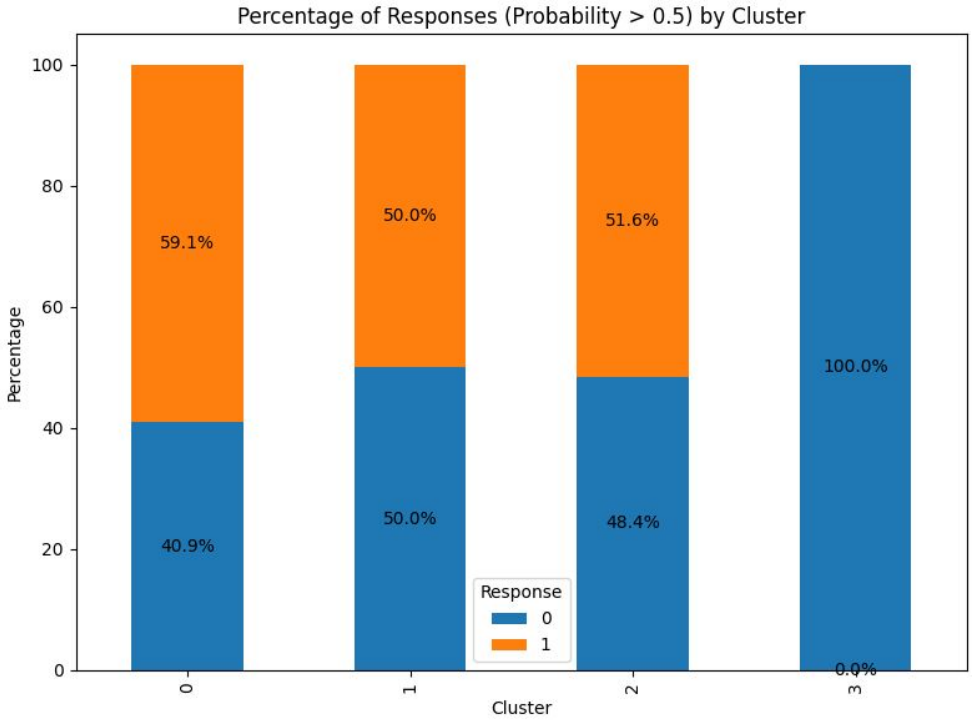
Mesmo preservando o 78% da base original, o modelo não performou muito bem.

	precision	recall	f1-score	support
0	0.89	0.97	0.93	457
1	0.45	0.15	0.23	65
accuracy			0.87	522
macro avg	0.67	0.56	0.58	522
weighted avg	0.84	0.87	0.84	522

- O modelo de previsão logística acerta 45% dos dados previstos, indicando uma alta taxa de falsos positivos.
- Aqui no entanto, nós temos a opção de abordar aqueles clientes com probabilidade de Response acima de acima de um certo valor, digamos 50%.

Percentage of Response with probability > 0.5

No geral, o modelo acerta 50% dos casos para indivíduos com probabilidade de Response maior que 0.5



- É muito difícil acertar para o cluster 4, uma vez que ele tem a menor taxa de Response, o que dificulta a generalização do modelo para estes casos.

- Além disso, tem a pior taxa de recência (75 dias desde a última compra). Seus indicadores de produtividade estão defasados.

Aprendizados

- O maior gasto do grupo que aderiu à Response, é puxado principalmente por Wines e Meat Products. Então deveríamos focar em Clientes classificados dentro dos **Clusters 2 e 0** com o objetivo de aumentar Response.

- O **Cluster 0** é o que o nosso modelo melhor consegue prever a probabilidade de Response 1.

- **Cluster 3** tem a menor taxa de Response, mas uma alta taxa de visita ao site. Talvez existam oportunidades de prospecção nesse canal.

For a better detailment of each step, consult:
https://colab.research.google.com/drive/1hHscJaSTtYEy8A3hC_JcHHeSwbHSzSzP#scrollTo=pwfiWliuLsXK