PROJETO DE CIÊNCIA DE DADOS: Perfil de Clientes (Clusterização)

Introdução

Esse projeto é uma iniciativa pessoal voltada para o aprendizado.

Durante o projeto um dataset sobre Perfil de Clientes é explorado, para passar por etapas de Análise, Transformação, Clusterização e entrega de análises.

Link para o dataset: Marketing Campaign (kaggle.com)

Tecnologias utilizadas:

Python 3 - Welcome to Python.org
Anaconda - Anaconda | The World's Most Popular Data Science Platform
Jupyter Notebook - Project Jupyter | Home
Power Bi - Power BI - Visualização de Dados | Microsoft Power Platform

Análise Exploratória e Limpeza e Transformação de Dados

Durante a fase de análise exploratória, eu busquei compreender o conjunto de dados e encontrar qualquer tipo de problema que possa causar complicações nas próximas fases do projeto.
O primeiro problema encontrado foi a presença de um outlier nos valores de renda, identificado ao utilizar o método "describe()" no dataframe. Para visualizar melhor o problema, utilizei 2 gráficos: Boxplot e Dispersão

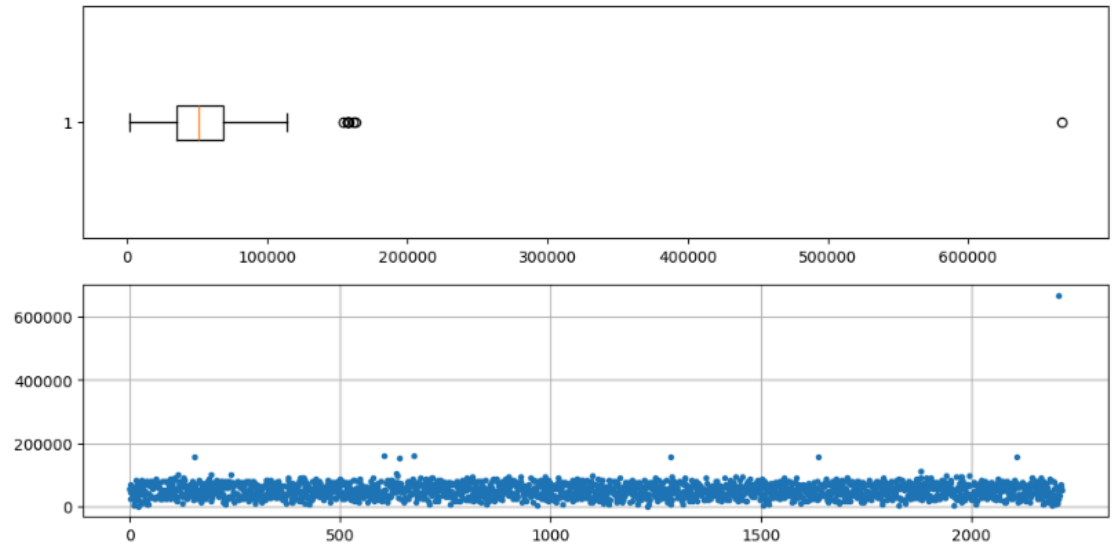


Figura 1: Detecção de Outlier

Após visualizar esse valor separadamente, cheguei a conclusão que esse valor pode ter sido um erro de digitação, uma vez que estando tudo dentro dos conformes, o valor da renda seja muito específico: “666666”.Como se trata de apenas um valor, mas que estava interferindo consideravelmente nas medidas devido ao seu alto valor, optei por remover essa observação do conjunto de dados.

Também é possível visualizar alguns outros outliers no boxplot, entretanto, eles parecem ser valores reais e não apresentam implicações significativas no conjunto de dados, portanto, não vejo necessidade de tratá-los.

O próximo problema identificado foi a presença de alguns valores nulos, detectados ao utilizar o método “isna().sum()”.Tratei esses valores substituindo-os pela média da respectiva coluna.

Por fim, eu removi clientes que tenham idade superior a 100 anos, pois eles representam um padrão de compra mais complicado para a análise, já que muitas vezes essas compras nem são realizadas por eles, mas sim por seu filhos e cuidadores, entretanto isso representa uma parcela muito pequena no conjunto, portanto não provoca mudanças significativas.

O que é Clusterização?

A clusterização é um técnica de aprendizado de máquina não supervisionada, que visa dividir um conjunto de objetos em grupos semelhantes. Essa técnica de agrupamento é utilizada para lidar com grandes volumes de dados, comumente usada para encontrar padrões e criar sistemas de recomendação.

Clusterização

Para realizar a clusterização, é necessário selecionar as colunas numéricas relevantes para o agrupamento. Neste projeto, escolhi 14 colunas que fornecem informações sobre o cliente e seus padrões de compra:

'Income', 'Age', 'Kidhome', 'Teenhome', 'MntWines', 'MntFruits', MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases'

Após a seleção das colunas, é necessário normalizar os dados, ou seja, ajustar todos os valores para uma mesma escala, o que é uma exigência para a entrada de dados no sklearn.

O próximo passo é determinar o número de clusters. Existem várias técnicas para definir o número ideal de clusters para um conjunto de dados. Neste projeto, optei por utilizar o método do cotovelo, que envolve a plotagem da curva de variância em função do número de clusters. Normalmente, o ponto onde a curva começa a se achatar indica o número ideal de clusters para o conjunto de dados.

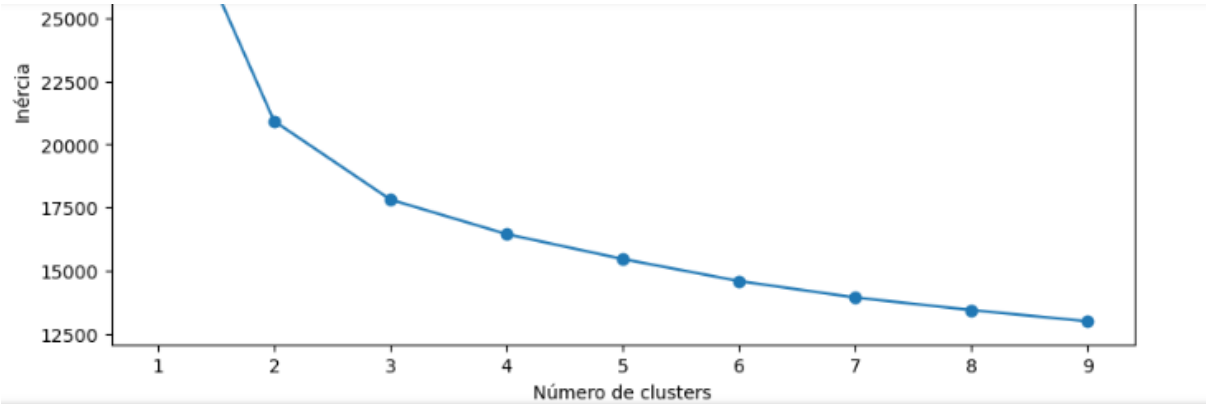


Figura 2 - Método do Cotovelo

Com base na forma da curva no gráfico, é concluído que 3 é um número interessante de clusters para este conjunto de dados.

Com tudo pronto, o próximo passo é treinar o algoritmo para concluir a etapa de clusterização.

Com o algoritmo treinado, é possível ver os resultados por meio de um gráfico de distribuição de valores pela idade e renda:

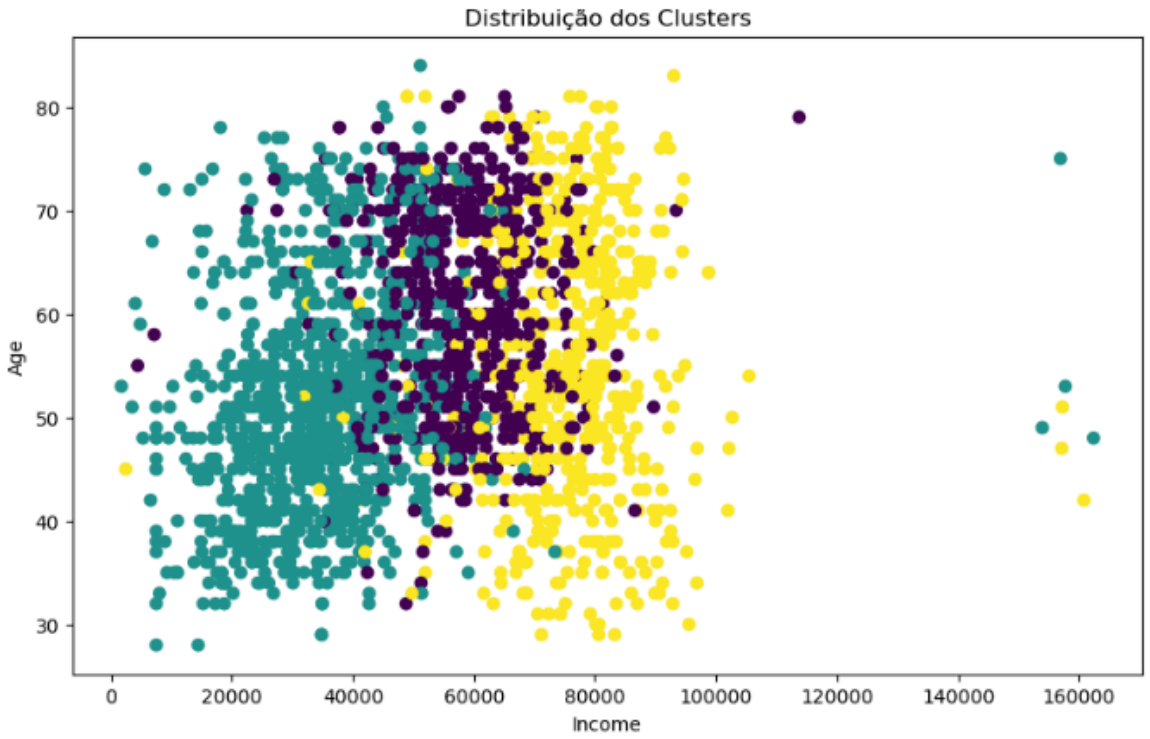


Figura 3: Distribuição dos Clusters

Análises e Insights (Power BI)

Como último passo do projeto, levar os dados preparados para o Power BI é uma excelente forma de montar análises e extrair valor dos dados. O conjunto de dados é focado no perfil dos clientes, permitindo a compreensão de padrões entre os grupos.

Para aproveitar ao máximo os dados disponíveis, dividi o dashboard em quatro partes:

Relação de Idade - Esta seção tem como objetivo apresentar todas as informações relevantes relacionadas à idade do cliente, como a distribuição por faixas etárias, nível de educação, renda, principais produtos adquiridos e canais de compra utilizados.

Relação Familiar - Esta seção fornece informações sobre a situação familiar do cliente, como estado civil e número de filhos, e como esses fatores se relacionam com outras características de cada cliente.

Aceitação das campanhas - A empresa realizou diversas campanhas. Esta seção indica os resultados dessas campanhas, o número de clientes que responderam a elas e os perfis desses clientes.

Distribuição de Compras de Produtos - Mostra a distribuição de compra dos produtos, baseada na renda dos clientes e do grupo, com a possibilidade de filtrar por idade.

Relação de Idade

Relação Familiar

Aceitação das Campanhas

Distribuição de Compras de Produtos

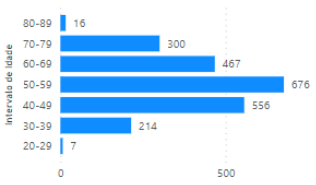
Média de Idade

55,10

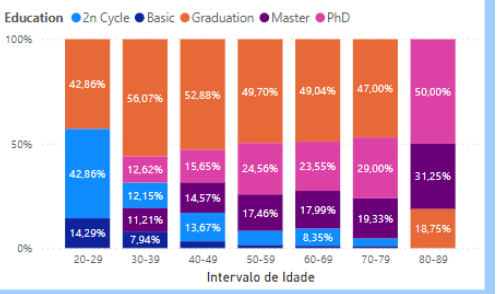
Média de Renda

52 Mil

Contagem de Clientes por Idade



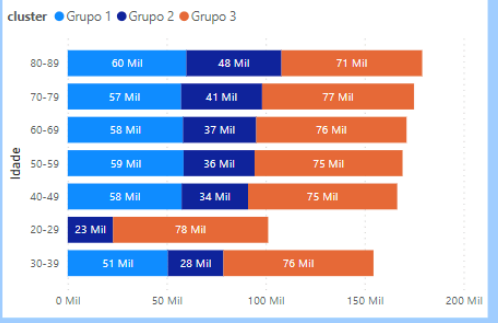
Contagem de Cliente por Idade e Educação



Total de Forma de Compra por Idade

Intervalo de Idade	Catalog	Deals	Store	Web
20-29	25	9	41	18
30-39	532	356	1178	723
40-49	1276	1236	2955	2080
50-59	1575	1727	3772	2786
60-69	1473	1173	2955	2060
70-79	990	681	1929	1380
80-89	84	19	129	93
Total	**5955**	**5201**	**12959**	**9140**

Distruibuição de Clientes por Renda e Idade por Grupo



Total Gasto em Produto por Idade

Intervalo de Idade	Fish	Fruits	Gold Prods	Meat	Sweet	Wines
20-29	270	117	209	1631	293	1424
30-39	8594	6664	9485	43807	6648	55480
40-49	20409	13878	22862	82593	14337	135275
50-59	21434	15971	27110	101012	16568	192663
60-69	19088	13372	22297	81798	12986	169924
70-79	12744	8012	15920	57753	9161	115426
80-89	1392	739	463	4781	559	9837
Total	**83931**	**58753**	**98346**	**373375**	**60552**	**680029**

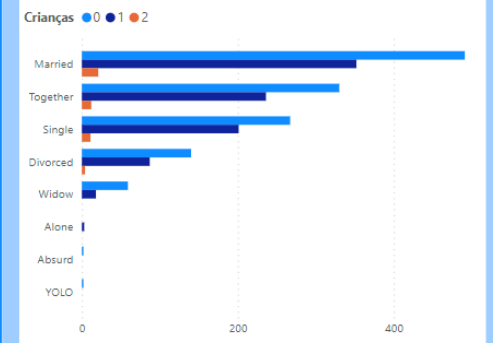
Relação de Idade

Relação Familiar

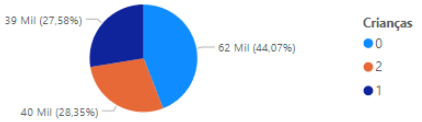
Aceitação das Campanhas

Distribuição de Compras de Produtos

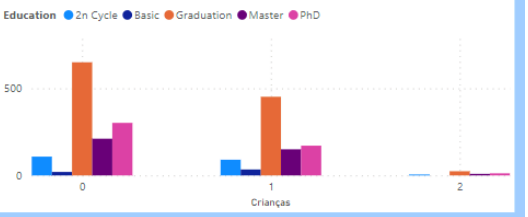
Contagem de Clientes por Estado Civil por Crianças em Casa



Média de Renda por Crianças em Casa



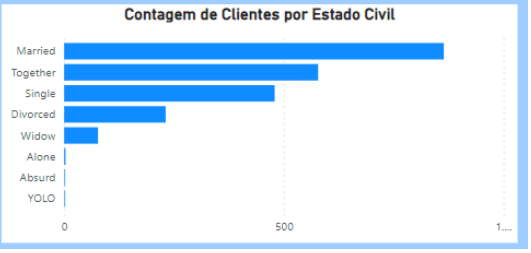
Contagem de Clientes por Educação e Crianças em Casa



Total Gasto em Produtos por Crianças em Casa

Crianças	Fruits	Gold Prods	Meat	Sweet	Wines	Fish
0	50791	77402	327850	52301	582604	72507
1	7637	20108	43901	8025	93835	11053
2	325	836	1624	226	3590	371
Total	**58753**	**98346**	**373375**	**60552**	**680029**	**83931**

Contagem de Clientes por Estado Civil



Total de Forma de Compra por Crianças em Casa

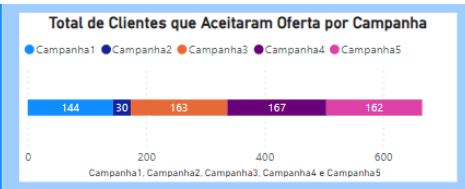
Crianças	Catalog	Deals	Store	Web
0	5112	2512	9326	6426
1	809	2553	3468	2580
2	34	136	165	134
Total	**5955**	**5201**	**12959**	**9140**

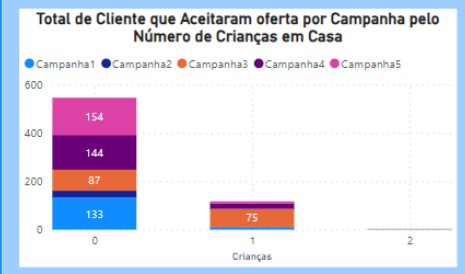
Relação de Idade

Relação Familiar

Aceitação das Campanhas

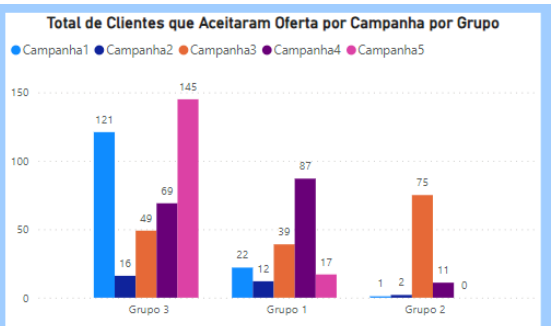
Distribuição de Compras de Produtos

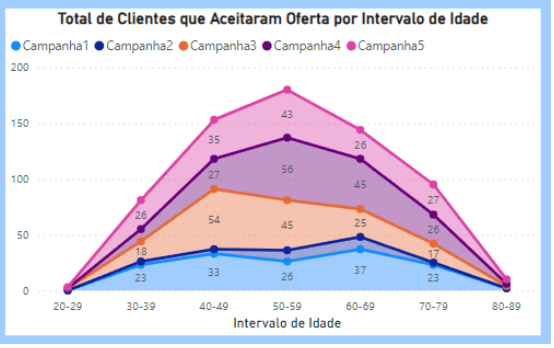




Total de Clientes que Aceitaram Oferta por Campanha por Educação

Education	Campanha1	Campanha2	Campanha3	Campanha4	Campar
PhD	30	10	40	45	
Master	18	2	24	31	
Graduation	82	16	78	81	
Basic	0	0	6	0	
2n Cycle	14	2	15	10	
Total	**144**	**30**	**163**	**167**	





Relação de Idade

Relação Familiar

Aceitação das Campanhas

Distribuição de Compras de Produtos

Intervalo de Ida...

- [ ] 20-29
- [ ] 30-39
- [ ] 40-49
- [ ] 50-59
- [ ] 60-69
- [ ] 70-79
- [ ] 80-89