# Situação problema

A presente análise tem como objetivo auxiliar uma loja na elaboração de possíveis campanhas de marketings e na venda de serviços ou produtos novos.

Para tal é necessário à identificação de grupos de clientes-alvo e quais seriam suas características principais.

Portanto será analisado um dataset contendo algumas informações básicas dos consumidores do shopping onde essa loja está alocada, como a idade, renda, sexo, etc. A partir destas variáveis será elaborado um algoritmo de clusterização em Python para identificar quais seriam os grupos mais importantes para serem focados/trabalhados pela equipe de marketing.

# Análise Exploratória dos Dados

Examinando o glossário do dataset tem-se:

<desc\_v="/assets/img/arquivo.gif">

\*\*imagem desc\_variaveis\*\*

Observando a imagem nota-se a presença de três variáveis numéricas e cinco variáveis categóricas.

\*\*null\_val\*\*

Examinando o dataset carregado com as funções “.info()” e “.duplicated()” não é possível notar a repetição de algum tipo de consumidor em específico ou a presença de dados ausentes.

Nota-se também que não é possível identificar a qual tipo de ramo a loja analisada pertence por meio das features apresentadas.

Para um entendimento inicial dos dados será realizado uma visualização da distribuição de cada variável isoladamente.

# Variáveis numéricas

(AGE)

\*\*img\_2\*\*

Pelos gráficos e pelas tabelas têm-se:

- A média de idade dos compradores é de aproximadamente 36 anos.

- Existe uma grande concentração de pessoas com idades entre aproximadamente 25 a 28 anos.

- A menor idade é de 18 anos e a maior é 76 anos.

- 75% dos dados está concentrado na faixa abaixo dos 42 anos

(INCOME)

\*\*img\_3\*\*

Analisando o gráfico e a tabela têm-se:

- A média de renda anual é de aproximadamente 121,000 tendo um desvio de 40,000.

- É possível identificar visualmente outliers pelo gráfico, apartir da zona dos 200,000 de renda anual.

- O mínimo de renda anual é de aproximadamente 35,800 e a máxima é de 309,000.

- A maior concentração de valores de renda anual é na faixa de 106,000 à 113,000.

# Variáveis categóricas

Para uma melhor entendimento e vizualização dos dados as features categóricas serão decodificadas com base no 'describe segmentation' disponibilizado.

(Settlement size)

\*\*img\_4\*\*

Analisando o gráfico nota-se:

- A maior parte dos dados é composta por pessoas que moram em cidades "small city".

- O dataset possui menos menos pessoas vindas de 'big city'.

- 'mid-size city' e 'big city' possuem aproximadamente a mesma quantidade de dados.

(EDUCATION)

\*\*img\_5\*\*

Observando os gráficos nota-se que:

- Cerca de 50% do dataset é composto por 'high school' e por 'employee'.

- Menos de 2% das pessoas do dataset são 'graduate school'.

- Cerca de 30% à 32% do dataset é composto por 'unemployed Occupation'.

(SEX)

\*\*img\_6\*\*

Analisando os gráficos têm-se:

- A quantidade de pessoas solteiras e não solteiras é praticamente igual neste dataset.

- A quantidade de pessoas do sexo 'male' e 'female' é relativamente igual, tendo uma diferença de cerca de 5% entre eles.

(Income e age per sex)

\*\*img\_7\*\*

Observando os gráficos e as tabelas têm-se:

- A média de 'Income' do sexo 'male' é ligeiramente maior do que a do sexo 'female'.

- O valor mínimo de 'Income' e o desvio padrão do sexo 'female' são menores do que o sexo 'male', indicando uma distribuição mais concentrada.

- Os valores dos quartis (25%, 50% e 75%) para a classe 'female' são menores do que a classe 'male', resultando em valores esperados menores.

- A média de idades da classe 'male' é maior do que a da classe 'female'.

- A classe 'female' está mais concentrada na faixa dos 25 a 39 anos, enquanto a classe 'male' está concentrada entre os 29 e 44 anos.

(Income e age per Marital status)

\*\*img\_8\*\*

Observando os gráficos e as tabelas têm-se:

- A média de 'Income' para o 'single' é ligeiramente maior do que para o 'non-single'.

- Os valores de 'Income' para a categoria 'non-single' estão ligeiramente mais concentrados.

- os valores dos quartis (25%, 50% e 75%) são bem próximos entre as duas classes.

- A média de idade da classe 'single' é maior do que a classe 'non-single'.

- A classe 'single' aparenta ter dados mais dispersos entre as faixas de 20 a 50 anos, enquanto a classe 'non-single' aparenta ter dados mais concentrados na faixa de 20 a 40 anos.

(Income e age per education)

\*\*img\_9\*\*

Observando os gráficos e as tabelas nota-se:

- A média de income das classes 'graduate school' e 'university' são próximas e maiores que as classes 'high school' e 'unknown'.

- A classe 'graduate school' possui uma kurtosis muito baixa, indicando que seus dados estão muito pulverizados. Isso indica que sua média e demais medidas não são representativas.

- Observando a tabela, nota-se que as métricas da classe 'high school' e 'unknown' são bem próximas e similares. Isto indica que a maioria dos dados da classe 'unknown' podem pertencer a classe 'high school'.

- As classes 'unknown', 'high school' e 'university' aparentam ter dois pontos de concentração em cada uma das distribuições.

- As classes 'unknown' e 'high school' possuem distribuições de idades mais jovens, possuindo média de idades menores, enquanto as classes 'university' e 'graduate school' possuem idades mais avançadas, possuindo médias de 57 e 54 anos.

- A menor idade da classe 'university' é de 38 anos.

- A maior idade da classe 'unknown' é de 33 anos.

(Income e age per occupation)

\*\*img\_10\*\*

Observando as tabelas e os gráficos nota-se:

- A média, os quartis e o desvio padrão da classe 'unemployed' são menores do que as demais classes.

- Pelo gráfico nota-se que a classe 'employee' possui valores bem concentrados em torno da mediana e da média.

- A classe 'highly qualified employee' possui uma média maior que as demais classes, porém possui um desvio padrão bem maior que as outras, indicando que os valores de 'Income' são bem dispersos.

- As distribuições de idade da classe 'employee' e 'unemployed' são similares.

- A classe 'highly qualified employee' possui uma média de idades maior que as demais classes.

- Na classe 'unemployed' observa-se duas concentrações de valores, uma próxima dos 25 anos e outra próxima dos 40 anos.

(Income e age per size city)

\*\*img\_11\*\*

Pelo gráficos e as tabelas é possivel observar:

- As distribuições de 'big city' e 'mid-size city' são muito similares.

- A média e desvio padrão da classe 'small city' é menor do que as demais classes.

- As três classes possuem distribuição de idades similares, tendo uma pequena diferença da média para a classe 'small city' , que possui um valor menor.

(age per income)

\*\*img\_12\*\*

Pelos gráficos é possível observar:

- Existe uma relação de linearidade entre 'Age' e 'Income' crescente.

- Existe uma concentração de dados entre a faixa de 'Age' entre 25 a 30 anos e de 'Income' entre 100,000 e 130,000.

- Existem outliers após a faixa de 200,000 de 'Income'.

# Correlação dos dados

Visando evitar possíveis problemas de multicolinearidade na resolução do problema será utilizado o método de correlação de Pearson, visando identificar a correlação linear entre as variáveis.

Como valor de corte será utilizado o valor de 0.75 e de -0.75, sendo esses valores definidos empiricamente em projetos anteriores.

\*\*img\_13\*\*

Como nenhuma variável ultrapassou os valores de corte nenhuma variável será removida e/ou modificada.

# Feature Engerring !!!!!!!!!!

\*\*Escalonando os dados\*\*

Como o objetivo do projeto é a elaboração de grupos de clientes-alvos será adotado o modelo de clusterização por meio do modelo K-means para a identificação destes grupos.

Por ser tratar de um modelo de clusterização baseado em distâncias euclidianas, os dados devem ser escalonados para que o modelo utilizado tenha maior precisão na geração dos grupos.

Como as variáveis 'Education', 'Occupation' e 'Settlement size' são variáveis categóricas e ordinais (possuem uma hierarquia definida) não será realizado nenhum procedimento de enconding nessas variáveis.

# Implementação do modelo

\*Modelo K-Means\*

Para a implementação do modelo k-means inicialmente deve ser estipulado a quantidade de grupos a serem gerados pelo algoritmo. Para isto será utilizado dois critérios diferentes para identificar esta quantidade de clusters. Os critérios utilizados serão:

- The Elbow method (baseado na inércia).

- Critério de Calinski and Harabasz.

O 'Elbow method' ou 'método do cotovelo' é um método para definir a quantidade de clusters baseado no erro gerado por cada grupo. neste método o objetivo não é selecionar o menor erro, mas sim a quantidade de clusters que a partir dele não ocorre variação significativa do erro. Com base nisso, a avaliação deste método é feita graficamente, relacionando o erro gerado e a quantidade de clusters, visando identificar esse ponto de inflexão onde o erro não diminui significativamente.

O 'Calinski and Harabasz criterion' é um método que leva em consideração do quão semelhante um objeto é ao seu próprio cluster (coesão) em comparação com outros clusters (separação). Aqui, a coesão é estimada com base nas distâncias dos pontos de dados em um cluster ao centróide do cluster e a separação é baseada na distância dos centróides do cluster ao centróide global. Ele também é avaliado graficamente e valor mais alto desse método significa que os clusters são densos e bem separados, embora não exista um valor de corte “aceitável”. Precisamos escolher aquela solução que fornece um pico ou pelo menos uma curva abrupta no gráfico linear.

\*\*img\_crit1\*\*

\*\*img\_crit2\*\*

Observando os gráficos tem-se que:

- Para o método do cotovelo não ocorre um ponto de inflexão muito percepitível, tendo uma ligeira alteração nas curvas com 4 e 6 clusters.

- Para o critério de Calinski and Harabasz o primeiro ponto de inflexão positivo ocorre com número de clusters igual à 6.

Como o critério de Calinski and Harabasz apresentou um valor plausível, será utilizado 6 clusters para o resto das análises.

# Visualização dos clusters

Após o processo de clusterização é necessário a realização de uma verificação visual e analítica sobre os grupos gerados, visto que essa é uma técnica de aprendizado não supervisionada. Para isso, pode-se utilizar as proprias variáveis do dataset para tentar realizar a visualização e entendimento desses clusters, porém é complexo identificar quais seriam as principais variáveis a serem utilizadas neste processo.

Com o intuito de ter uma melhor explicabilidade dos clusters gerados, será utilizado a técnica do 'PCA' (Análise dos Componentes Principais). Esta técnica tem como objetivo a redução de dimensionalidade (quantidade de variáveis) de um dataset por meio da criação de novas variáveis que expliquem de maneira satisfatória os dados originais.

Portanto, a utilização desta técnica tem como objetivo criar novas dimensões para uma melhor visualização dos clusters gerados

\*\*img\_PCA\*\*

Observando o gráfico nota-se que a explicabilidade do dados com três componentes é de 84,3% sendo um valor que eu considero razoável para a vizualização dos dados.

\*\*img\_final\_cluster\*\*

Analisando o gráfico gerado tem-se que as classes 1 e 4 estão bem isoladas podendo representar grupos distintos de clientes-alvo.

Já as classes 6 e 3 aparentam estar muito próximas, necessitando verificar se estes grupos realmente possuem características diferentes entre eles. O mesmo problema acontece com as classes 2 e 5, necessitando também serem analisados.

# Análise dos clusters

\*Clusters 3 e 6\*

\*\*img\_c3\_c6\*\*

Nota-se pelos gráficos que:

- Ambos os clusters (3 e 6) são compostos unicamente pela classe 'male' na variável 'Sex' e pela classe 'single' pela variável 'Marital status'. Isto pode indicar que essas duas variáveis tiveram grande impacto na elaboração do modelo de clusterização.

- O cluster 3 é composto majoritariamente pelas classes 'big city' e 'mid-size city' na variável 'Settlement size', já o cluster 6 é composto majoritariamente pela classe 'small city' com alguns poucos dados de 'mid-size city'.

- Nota-se que o cluster 6 quase não possui dados da classe 'highly qualified employee' para a variável 'Occupation' e é composto em sua maior parte por 'unemployed'. Em comparação o cluster 3 é o oposto, havendo uma parcela maior de dados de 'highly qualified employee' e poucos de 'unemployed'.

- Existe uma diferença de renda entre os dois clusters, sendo o cluster 3 mais concentrado em torno de 130,000 a 140,000 enquanto o cluster 6 está na faixa dos 70,000 e 120,000.

- não existe uma diferença de distribuição de idades entre os dois clusters.

Por fim, os dois clusters algumas similaridades mas têm diferenças bem significativas, sendo o cluster 6 um agrupamento de homens solteiros com menor renda e de empregos de menor grau hierárquico e que moram majoritariamente em cidades pequenas, enquanto o cluster 3 um agrupamento de homens solteiros com uma renda maior e de empregos de maior grau hierárquico e que moram em cidades de médio/grande porte.

\*Clusters 2 e 5\*

\*\*img\_c2\_c5\*\*

Nota-se pelos gráficos que:

- Ambos os clusters (2 e 5) são compostos unicamente pela classe 'female' na variável 'Sex' e pela classe 'non-single' pela variável 'Marital status'.

- O cluster 5 é composto majoritariamente pelas classes 'big city' e 'mid-size city' na variável 'Settlement size', já o cluster 2 é composto majoritariamente pela classe 'small city'.

- Nota-se que o cluster 2 quase não possui dados da classe 'highly qualified employee' para a variável 'Occupation' e é composto em sua maior parte por 'unemployed'. Em comparação o cluster 5 é o oposto, havendo uma parcela maior de dados de 'highly qualified employee' e quase nenhum da classe 'unemployed'.

- Não existe uma diferença de renda visível mas o cluster 5 possui uma maior variabilidade de renda, enquanto o cluster 2 paresenta dados concentrados em torno do valor de 100,000.

- não existe uma diferença de distribuição de idades entre os dois clusters.

Da mesma forma que ocorreu entre os clusters 6 e 3 ocorreu entre os clusters 5 e 2, sendo o cluster 2 um agrupamento de mulheres 'non-single' de empregos de menor grau hierárquico e que moram majoritariamente em cidades pequenas, enquanto o cluster 5 é um agrupamento de mulheres 'non-single' de empregos de maior grau hierárquico e que moram em cidades de médio/grande porte.

\*Clusters 1 e 4\*

\*\*img\_c1\_c4\*\*

Examinando os clusters 1 e 4 temos:

- O cluster 1 é composto pela classe 'male' da variável 'Sex' e pela classe 'non-single' da variável 'Marital status'.

- O cluster 4 é composto pela classe 'female' da variável 'Sex' e pela classe 'single' da variável 'Marital status'.

- O cluster 4 não apresenta dados de pessoas da classe 'highly qualified employee' da variável 'Occupation', pessoas da classe 'big city' da variável 'Settlement size'. Este cluster tem 'Income' menores que 150,000, sendo a maioria das rendas concentradas em torno de 100,000.

- O cluster 1 é bem diversificado, tendo a mesma proporção de dados para todas as classes da variável 'Settlement size' e possui uma parcela dados da classe 'highly qualified employee' da variável 'Occupation'. A maior parte dos dados está concentrado em torno das idades de 25 a 30 anos.

# Conclusão

Com isso, foram descobertos 6 grupos distintos que podem ser trabalhados em conjunto com a equipe de marketing para que seja traçado novas estratégias para aumentar a quantidade de vendas ou novos produtos.

Os principais direcionamentos obtidos na análise são:

- Desenvolvimento de produtos/propagandas de itens de consumo com base nas caracteristicas de cada clusters.

- A renda das mulheres em média é menor, então é necessário levar isto em consideração na elaboração de produtos/promoções.

- Pessoas que moram em cidades pequenas tem uma renda menor do que pessoas que moram e cidades de médio/grande porte, então é necessário levar isto em consideração na elaboração de produtos/promoções.

- É possível observar uma certa linearidade crescente entre 'Age' e 'Income', então elaboração de planos de fidelidade podem ocasionar em lucros maiores ao longo do tempo.

- A maior parte do dataset analisado é composto por pessoas de 'Education' 'high school', então elaborar planos de desconto ou de captação de clientes que tenham pelo menos este nível de escolaridade pode ter resultados positivos.