# \*\*Introdução\*\*

O presente projeto tem como objetivo analisar os dados de um supermaket visando identificar os produtos mais comprados e os tipos de clientes mais recorrentes deste supermarket, para que com isto seja possível elaborar novas estratégias de vendas de produtos e o direcionamento de público-alvo para que o marketing possa focar em suas propagandas.

Para isto será utilizado dois datasets, o Market\_Basket\_Optimisation e o Supermarket\_CustomerMembers.

Portanto, neste projeto temos os seguintes objetivos:

- Identificar os tipos de consumidores;

- Identificar os produtos mais e menos vendidos;

- Identificar pares de produtos mais vendidos;

- Elaborar estratégias para aumentar a venda de produtos;

# \*\*Visualização dos dados\*\*

<img src="/US\_supermaket\_data/Project\_prints/img\_01.png ">

Nota-se que para o dataset dos produtos ocorre a presença de dados ausentes devido ao formato em que o dataset foi montado, onde a linha representa o conjunto de produtos comprados.

Neste formato quando uma pessoa compra menos produtos que a anterior ela gera colunas vazias, gerando os dados ausentes.

<img src="/US\_supermaket\_data/Project\_prints/img\_02.png ">

No customers dataset não existe a presença de dados ausentes, porém existe uma quantidade menor de dados para serem analisados.

# \*\*Análise Exploratória de Dados (EDA)\*\*

Inicialmente será explorado os dados do ‘customers dataset’, visando características importantes e insights relevantes para a formação de grupos, e após isto, será analisado o ‘products dataset’ tendo o objetivo de identificar os produtos mais vendidos e suas características relevantes.

## Dataset dos Clientes

<img src="/US\_supermaket\_data/Project\_prints/img\_03.png ">

Pelo gráfico vemos que a maior parcela dos consumidores deste supermercado é do sexo feminino.

<img src="/US\_supermaket\_data/Project\_prints/img\_04.png ">

Observando o gráfico nota-se que não existem valores muitos discrepantes de ‘annual income’ neste dataset. Observando os valores de annual income por gênero nota-se que a média do homem (62,600) é um pouco superior do que a média feminina (60,000).

O primeiro quartil para os homens (45,000) é maior que o das mulheres (39,500) indicando que a renda dos homens é um pouco mais concentrada e maior que a das mulheres.

<img src="/US\_supermaket\_data/Project\_prints/img\_05.png ">

Analisando o gráfico nota-se que a média do spending score é a mesma para os dois sexos, porém para o gênero feminino o range entre o primeiro e terceiro quartil é menor que o do gênero masculino, indicando que os dados estão mais concentrados.

Como o primeiro quartil feminino (35) é maior que o masculino (23) temos a ideia que as mulheres são mais propensas a comprar mais neste supermercado.

<img src="/US\_supermaket\_data/Project\_prints/img\_06.png ">

Analisando o gráfico percebe-se que a média das idades das mulheres (35) é menor do que a dos homens (37), além de ter um intervalo de idades mais concentrados (29-48) para as mulheres do que para os homens (27.5-51).

<img src="/US\_supermaket\_data/Project\_prints/img\_07.png ">

Com base nesses gráficos temos as seguintes considerações:

- No gráfico 'Age' x 'Spending Score' nota-se que idades maiores que 40 anos o spending score tende a estar entre o valor de 5 e 60, enquanto para idades menores que 40 anos o spending score oscila entre 20 a 100.

- No gráfico 'Age' x 'Annual Income' os maiores valores de income são encontrados na faixa de indade entre 28 à 54 anos. Para idades abaixo de 28 e acima de 54 o valor de income geralmente encontrado é abaixo de 88,000.

- No gráfico de 'Spending Score' x 'Annual Income' nota-se a formação de 5 grupos diferentes:

- Um grupo onde possui um valor alto de income (70,000 à 137,000) mas com spending score baixo, entre 1 à 40.

- Um grupo onde possui um valor alto de income (70,000 à 137,000) e com spending score alto, entre 63 à 97.

- Um grupo onde possui um valor baixo de income (16,000 à 38,000) e com spending score alto, entre 60 à 98.

- Um grupo onde possui um valor baixo de income (16,000 à 38,000) e com spending score baixo, entre 3 à 40.

- Um grupo com valores medianos de income (40,000 à 65,000) e de spending score (40 à 60).

Com base nestas informações a expectativa é que ocorra a geração de aproximadamente de 4 a 6 grupos de consumidores baseado nestes dados tendo como features principais o spending score e o annual income.

## Dataset de Produtos

Como o dataset de produtos se trata da lista de compra de cada cliente, ele necessita ser manipulado e reestruturado para a remoção de dados ausentes.

Portanto a estrutura do dataset irá mudar de linhas de itens comprados para um formato onde cada coluna é um produto que já foi comprado e cada linha será uma compra, onde a junção de linha e coluna representará a quantidade de itens comprados para um determinado produto. Também será criado uma coluna para armazenar a quantidade total de itens comprados.

<img src="/US\_supermaket\_data/Project\_prints/img\_08.png ">

Analisando a tabela temos que:

- Existem 121 produtos únicos neste dataset.

- Não existe a repetição de um mesmo item em uma única compra.

- Em média um consumidor compra 4 produtos.

- O maior número de produtos comprados em uma única vez foi 20.

<img src="/US\_supermaket\_data/Project\_prints/img\_09.png ">

Com base no gráfico os top 10 itens mais comprados são:

- Mineral water, eggs, spaghetti, french fries, chocolate, green tea, milk, ground beef, frozen vegetagles, pancakes.

Destes itens, a sua maioria é composta por itens de consumo diário e por itens de breakfast.

Os top 10 itens menos comprados são:

- asparagus, water spray, napkins, cream, bramble, tea, mashed potato, chutney, chocolate bread,dessert wine.

Alguns destes itens são considerados itens de limpeza, de sobremesa e vegetais, porém não parece existir um padrão visível entre eles para que ocorra a sua baixa demanda.

Para identificar os pares de produtos mais comprados será utilizado as regras de associação.

# \*\*Regras de associação\*\*

## Conceitos principais

Para compreender as regras de associação, é necessário compreender quatro conceitos fundamentais:

- Suporte: O suporte é uma indicação da frequência com que o conjunto de itens aparece no conjunto de dados. Em outras palavras, isso é uma indicação da popularidade de um conjunto de itens em um conjunto de dados.

- Confiança: A confiança é uma indicação de quantas vezes a regra foi considerada verdadeira. Em outras palavras, a confiança indica a probabilidade de compra do item Y quando o item X é comprado.

- Lift: O Lift é uma métrica para medir a proporção entre X e Y que ocorrem juntos e a ocorrência de X e Y se eles fossem estatisticamente independentes. Em outras palavras, o lift ilustra a probabilidade de o item Y ser comprado quando o item X é comprado, ao mesmo tempo que controla a popularidade do item Y.

- Uma pontuação de Lift próxima de 1 indica que o antecedente e o consequente são independentes e a ocorrência do antecedente não tem impacto na ocorrência do consequente.

- Uma pontuação de Lift maior que 1 indica que o antecedente e o consequente são dependentes um do outro, e a ocorrência do antecedente tem um impacto positivo na ocorrência do consequente.

- Uma pontuação de Lift menor que 1 indica que o antecedente e o consequente se substituem, o que significa que a existência do antecedente tem um impacto negativo no consequente ou vice-versa.

- Convicção: A convicção mede a força de implicação da regra a partir da independência estatística. A pontuação de convicção é uma razão entre a probabilidade de X ocorrer sem Y enquanto eles eram dependentes e a probabilidade real de existência de X sem Y.

<img src="/US\_supermaket\_data/Project\_prints/img\_10.png ">

Com base no valor mínimo de suporte escolhido de 1% temos que no máximo ocorre a formação de um grupo composto por 3 itens.

<img src="/US\_supermaket\_data/Project\_prints/img\_11.png ">

Observando os pares de itens gerados com um valor de suporte maior que 4% tem-se que em todos os casos existe a presença do item 'mineral water'.

Também é perceptível que todos estes pares são gerados com os top 10 itens comprados neste supermarket.

<img src="/US\_supermaket\_data/Project\_prints/img\_12.png ">

Observando a tabela gerada composta pela associação de três itens, nota-se novamente a presença do item 'mineral water' em quase todos os grupos formados.

<img src="/US\_supermaket\_data/Project\_prints/img\_13.png ">

Observando a tabela é possível perceber existe uma grande tendência do item 'mineral water' ser o terceiro item a ser comprado.

Pela tabela vemos uma grande presença dos itens 'ground beef', 'spaghetti' e 'milk' como antecedentes que possuem maior valor de confiança, porém em sua maior parte são comprados em pares para depois resultar na compra de 'mineral water'. Em geral todos esses casos estas hipóteses possuem valores de lift e convicção altos, reforçando a confiabilidade desta regra.

<img src="/US\_supermaket\_data/Project\_prints/img\_14.png ">

Observando o gráfico temos que a maioria das regras criadas possui lift maior que 1 e valor de confiança menor que 50%. Isto nos dá a ideia de que as regras criadas possuem bastante força, mas que algumas não tem uma ocorrência forte nas listas de compras examinadas.

<img src="/US\_supermaket\_data/Project\_prints/img\_15.png ">

Olhando a tabela que leva em consideração a compra de somente 1 item como antecessor, percebe-se a forte presença do item 'mineral water' novamente.

Um outro produto que foge um pouco desta regra é o item 'ground beef' com o 'spaghetti' possuindo um valor de confiança de cerca de 40% juntamente com um lift muito elevado de 2.3.

<img src="/US\_supermaket\_data/Project\_prints/img\_16.png ">

Observando as regras que possuem somente um item de antecessor e um item de consequência com o menor valor de confiança tem-se que quando a pessoa compra 'mineral water' ela tem baixa chance de comprar junto itens como 'cereals', 'red wine' e 'avocado'.

Nota-se também que essas regras apresentadas relacionam itens de maior demanda como itens antecessores e itens de média demanda como consequência.

# \*\*Clusterização dos Clientes\*\*

Para a elaboração do modelo de machine learning de clusterização dos grupos de clientes será utilizado o modelo K-means, por ser um modelo amplamente utilizado para esta aplicação. Com base nisto será realizado o escalonamento das variáveis, sendo o gênero classificado em 0 e 1 e as demais variáveis (‘Age’, ‘Income’, ‘Spending score’) com um Scaler \*\*‘MinMax’\*\*, para melhorar o desempenho do algoritmo de clusterização.

## Modelo K-Means

Para a implementação do modelo k-means inicialmente deve ser selecionado a quantidade de grupos a serem gerados. Para isto será utilizado quatro critérios diferentes para identificar a quantidade de clusters que devem ser gerados. Os critérios em questão são:

- The Elbow method, based in inertia.

- Silhouette criterion.

- Calinski and Harabasz criterion.

- Davies and Bouldin criterion.

Uma simples explicação dos ultimos três critérios pode ser encontrada no seguinte link: [How to measure clustering performances](<https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c>).

<img src="/US\_supermaket\_data/Project\_prints/img\_17.png ">

Observando os gráficos tem-se que:

- Para o critério de inércia ocorre um ponto suave de inflexão com 4 clusters.

- Para o critério de Silhouette o ponto que apresenta o maior valor de score é o que possui 2 clusters.

- Para o critério de Calinski and Harabasz o gráfico não apresenta um ponto de inflexão positivo, logo não existe um valor de clusters por meio deste critério.

- Pelo critério de Davis and Bouldin o ponto que apresenta menor valor de score é o que possui 2 clusters.

Com estes resultados temos que será gerado uma segmentação com 2 cluster ou com 4 clusters. Na minha opinião dois clusters não seriam significativos, visto que o dataset possui poucos dados e poucas features, assim provavelmente ocorreu uma segregação por meio de dois grupos de renda ou spending score.

Portanto será utilizado a clusterização com 4 clusters no decorrer 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 próprias 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 src="/US\_supermaket\_data/Project\_prints/img\_18.png ">

Observando o gráfico nota-se que a explicabilidade dos dados com três componentes é de 89,4% sendo um valor razoável para a visualização dos dados.

<img src="/US\_supermaket\_data/Project\_prints/img\_19.png ">

Analisando o gráfico nota-se que os quatro grupos gerados estão bem isolados, mostrando que a clusterização foi eficiente.

## Examinando os clusters

<img src="/US\_supermaket\_data/Project\_prints/img\_20.png ">

Pelo gráfico nota-se que a estão bens distribuídas, sendo a classe 2 com a maior quantidade de dados (57) e a classe 4 com a menor quantidade de dados (40).

<img src="/US\_supermaket\_data/Project\_prints/img\_21.png ">

Analisando o gráfico temos:

- As classes 2 e 4 pertencem à um grupo de pessoas mais jovens, girando em torno de 18 à 40 anos.

- A classe 2 possui uma grande concentração de pessoas na faixa dos 30 anos.

- A classe 3 pertence à um grupo de pessoas adultas e idosos, indo de 34 anos até 68 anos, sendo mais concentrada na faixa de 41 à 54 anos.

- a classe 1 engloba uma faixa alta de idades, indo de 19 anos até 70 anos, onde em média a pessoa possui 49 anos.

<img src="/US\_supermaket\_data/Project\_prints/img\_22.png ">

Com base nesse gráfico é possível concluir:

- A classe 4 representa um grupo que tem a tendência a gastar mais em suas compras, possuindo em sua maioria um spending score entre 58 e 89.

- A classe 3 representa um grupo de pessoas que não tende a gastar muito, variando o spending score de 5 à 60, porém sendo mais concentrado na faixa de 39 a 50.

- A classe 2 também representa um grupo de pessoas que gastam mais, porém neste caso o spending score é mais concentrado entre o valor de 62 e 93.

- A classe 1 representa um grupo de pessoas que em geral não gastam muito, tendo dois intervalos de concentração, um com spending score variando entre 1 e 25 e outro intervalo indo de 32 até 60.

<img src="/US\_supermaket\_data/Project\_prints/img\_23.png ">

Observando o gráfico não é possível perceber uma diferença entre o annual income de cada classe.

<img src="/US\_supermaket\_data/Project\_prints/img\_24.png ">

Neste gráfico nota-se que as classes 4 e 1 são compostos por pessoas do sexo masculino enquanto as classes 3 e 2 são compostos pelas pessoas do sexo feminino.

# \*\*Conclusão\*\*

Na segmentação dos clientes foram identificados quatro grupos distintos de compradores, nos quais os principais a serem focados são os grupos 1 e 3, visto que os grupos em si possuem annual income relativamente similar, mas estes dois grupos possuem valores de spending score menores que 60.

Como esses dois grupos são formados por homens com idades na faixa de 40 à 60 anos, o ideal é identificar quais são os produtos mais comprados e mais desejados desses tipos de consumidores para elaborar estratégias de marketing focando neste público e em alguns casos concedendo descontos em compras com programas de fidelidade para tentar impulsionar o número de vendas.

Em relação aos produtos mais vendidos foram identificados os seguintes itens:

- Mineral water, eggs, spaghetti, french fries, chocolate, green tea, milk, ground beef, frozen vegetagles, pancakes.

Destes itens, a sua maioria é composta por itens de consumo diário e por itens de breakfast.

Os top 10 itens menos comprados são:

- asparagus, water spray, napkins, cream, bramble, tea, mashed potato, chutney, chocolate bread,dessert wine.

Alguns destes itens são considerados itens de limpeza, de sobremesa e vegetais, porém não parece existir um padrão visível entre eles para que ocorra a sua baixa demanda. Como o dataset de produtos comprados é considerado relativamente pequeno não existe uma ocorrência significativa dos produtos menos comprados, como por exemplo 'asparagus' foi comprado somente 1 vez em 7500 compras, com isso não é possível realizar uma análise confiável visando aumentar o número de vendas desses produtos.

visando aumentar o número de vendas nota-se que das regras de associação criadas o produto 'mineral water' tem um grande impacto como item consequente tanto após compras com um ou dois itens, além de ser o produto mais vendido da loja. Analisando os itens com menor confiança entre as compras com dois itens o 'mineral water' é geralmente o primeiro item a ser comprado, indicando que quando a pessoa compra este item ela tende a não comprar outros itens como 'cereals', 'red wine' e 'avocado', que são itens de média demanda.

Com base nisto uma estratégia que pode ser implementada é alocar o produto 'mineral water' no fim dos corredores desta loja, onde nestes corredore podem se encontrar produtos como itens 'ground beef', 'spaghetti' e 'milk', que são itens que possuem uma confiança e suporte altos de antecessores do item de 'mineral water'. Isto também pode ocasionar no aumento de venda de outros itens, visto que 'mineral water' é o produto de maior demanda os consumidores podem acabar comprando outros produtos no meio do trajeto até este item.