**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**João Rafael Quadros dos Santos**

**ANÁLISE DE REGRESSÃO E ASSOCIAÇÃO PARA GERAÇÃO DE INDICADORES NO CONTEXTO DE MARKET BASKET ANALYSIS.**

Belo Horizonte

2024

**João Rafael Quadros dos santos**

**ANÁLISE DE REGRESSÃO E ASSOCIAÇÃO PARA GERAÇÃO DE INDICADORES NO CONTEXTO DE MARKET BASKET ANALYSIS.**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte

2024 **SUMÁRIO**

[1. Introdução 4](#_Toc445198572)

[1.1. Contextualização 4](#_Toc445198573)

[1.1. O problema proposto 4](#_Toc445198574)

[2. Coleta de Dados 4](#_Toc445198576)

[3. Processamento/Tratamento de Dados 6](#_Toc445198582)

[4. Análise e Exploração dos Dados 11](#_Toc445198585)

[5. Criação de Modelos de Machine Learning 15](#_Toc445198586)

[6. Apresentação dos Resultados 24](#_Toc445198587)

[7. Links 26](#_Toc445198588)

# 1. Introdução

## 1.1. Contextualização

O trabalho apresentado visa demonstrar a aplicação prática de técnicas de Data Science e Big Data no contexto do varejo, especificamente na análise de cestas de compras (Market Basket Analysis). Em um mundo cada vez mais conectado, compreender o comportamento de compra dos consumidores é crucial para otimizar estratégias de marketing e melhorar a experiência do cliente. A análise de cestas de compras permite identificar padrões de consumo, possibilitando que redes de supermercados desenvolvam promoções mais eficazes e personalizadas. Utilizando um dataset público, a pesquisa visa não apenas a replicabilidade do estudo, mas também a validação dos resultados, contribuindo para o avanço das práticas de análise de dados na área do varejo.

## 1.2. O problema proposto

O objetivo central deste trabalho é investigar as relações entre produtos no setor de varejo global, utilizando dados de compras para desenvolver indicadores que sustentem decisões estratégicas de vendas e recomendações de produtos. A identificação de padrões de consumo é essencial, pois não apenas enriquece a experiência do cliente, mas também impulsiona as receitas por meio de promoções direcionadas. O dataset analisado é de domínio público e reflete transações de redes de supermercados em todo o mundo, abrangendo registros de compras dos anos de 2010 e 2011, com uma perspectiva geográfica abrangente.

# 2. Coleta de Dados

Os dados utilizados neste trabalho foram obtidos a partir de um dataset público disponível no Kaggle (<https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis/data>).O dataset contém informações sobre transações de compras em redes de supermercados ao redor do mundo, abrangendo os anos de 2010 e 2011. A estrutura do dataset é composta por várias colunas que descrevem os itens comprados, suas quantidades e preços. A seguir, apresentamos uma tabela com a descrição de cada campo/coluna do dataset:

|  |  |  |
| --- | --- | --- |
| Nome da Coluna | Descrição | Tipo |
| BillNo | Número de 6 dígitos atribuído a cada transação. | Nominal | |
| Itemname | Nome do produto. | Nominal | |
| Quantity | Quantidade de cada produto por transação. | Numérico | |
| Date | Data e hora em que cada transação foi gerada. | Numérico | |
| Price | Preço do produto. | Numérico | |
| CustomerID | Número de 5 dígitos atribuído a cada cliente. | Nominal | |
| Country | Nome do país onde reside cada cliente. | Nominal | |

**Figura 1 – Exemplo de registros do dataset.**

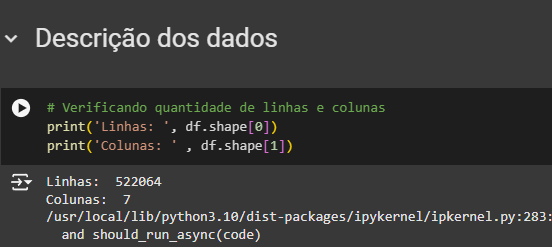


# 

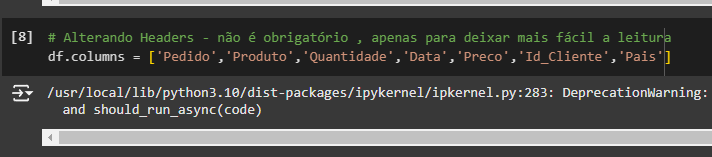
# 3. Processamento/Tratamento de Dados

O conjunto de dados analisado possui 522.065 registros e 7 colunas. Inicialmente, alteramos os cabeçalhos das colunas para facilitar a leitura, renomeando-os para 'Pedido', 'Produto', 'Quantidade', 'Data', 'Preco', 'Id\_Cliente' e 'Pais'.

**Figura 2 – Descrição dos dados.**



**Figura 3 – Alteração de cabeçalho.**



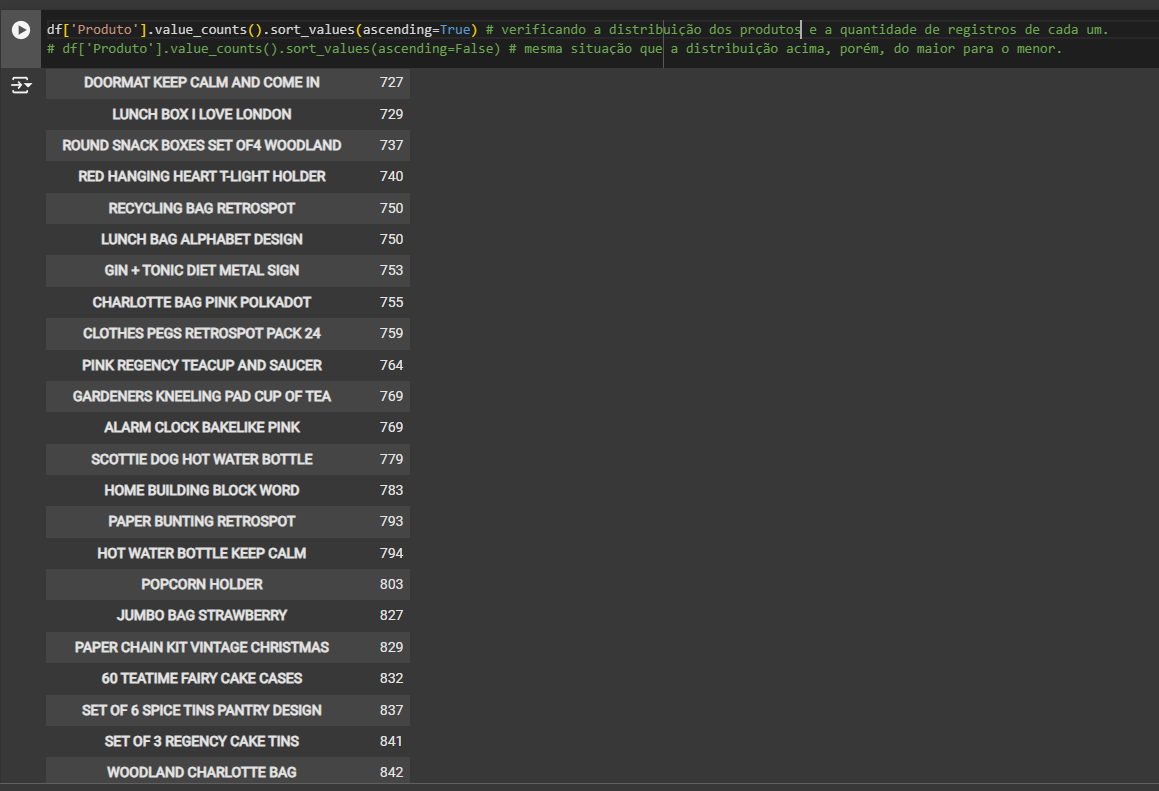
**Figura 4 – Análise Descritiva.**



Durante a análise descritiva, observou-se que os campos 'Quantidade' e 'Preco' contêm valores negativos, o que é anômalo e requer tratamento. Valores negativos em 'Quantidade' podem indicar erros de entrada de dados, enquanto os negativos em 'Preco' podem resultar de retornos ou ajustes incorretos. Esses dados inválidos precisam ser tratados para garantir a integridade das análises subsequentes.

A seguir, fizemos uma análise da distribuição dos países, produtos e pedidos, bem como verificações para identificar valores nulos e duplicados. Observamos que os campos 'Produto' e 'Id\_Cliente' apresentavam dados ausentes, e o conjunto continha 5.286 registros duplicados. Diante dos dados negativos e nulos, consideramos estratégias para o tratamento. Inicialmente, ponderamos substituir os valores negativos na 'Quantidade' e 'Preco' pela média ou mediana dessas colunas.

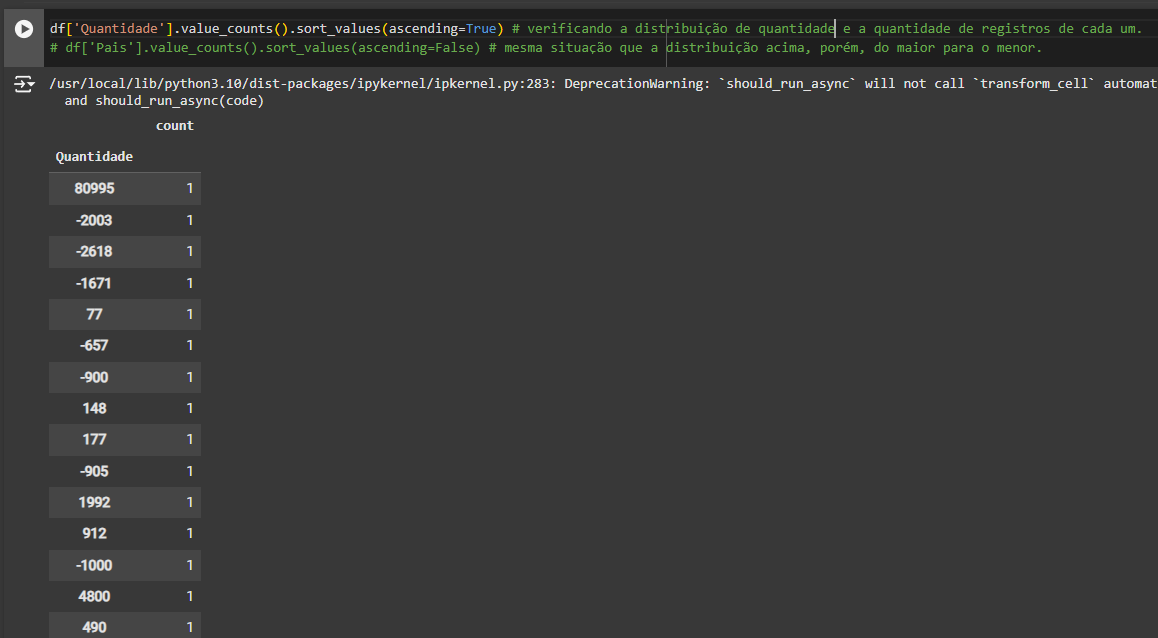
**Figura 5 – Valores para produto.**



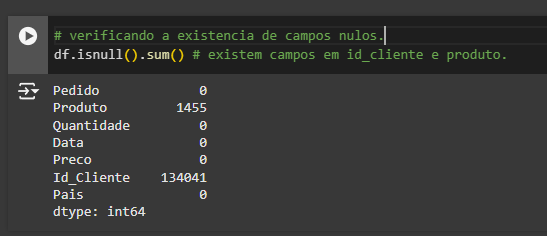
**Figura 6 – Valores para pedido.**



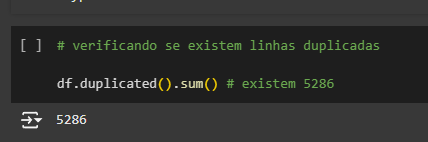
**Figura 7 – Valores para quantidades.**



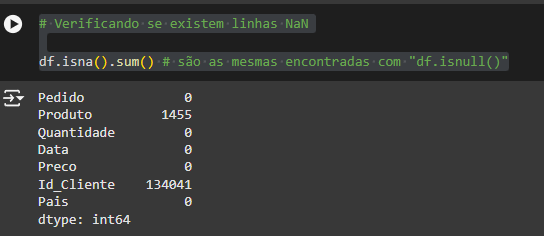
**Figura 8 – Valores Nulos.**



**Figura 9 – Valores Duplicados.**

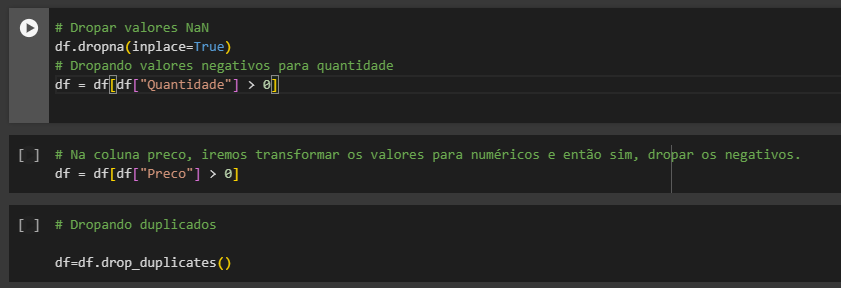


**Figura 10 – Valores “NaN”.**



No entanto, após uma análise mais aprofundada, decidimos pela exclusão de registros com dados inválidos ou ausentes, dado que a preservação da qualidade dos dados é fundamental para a validade da análise.

**Figura 11 – Dropando negativos e duplicados.**

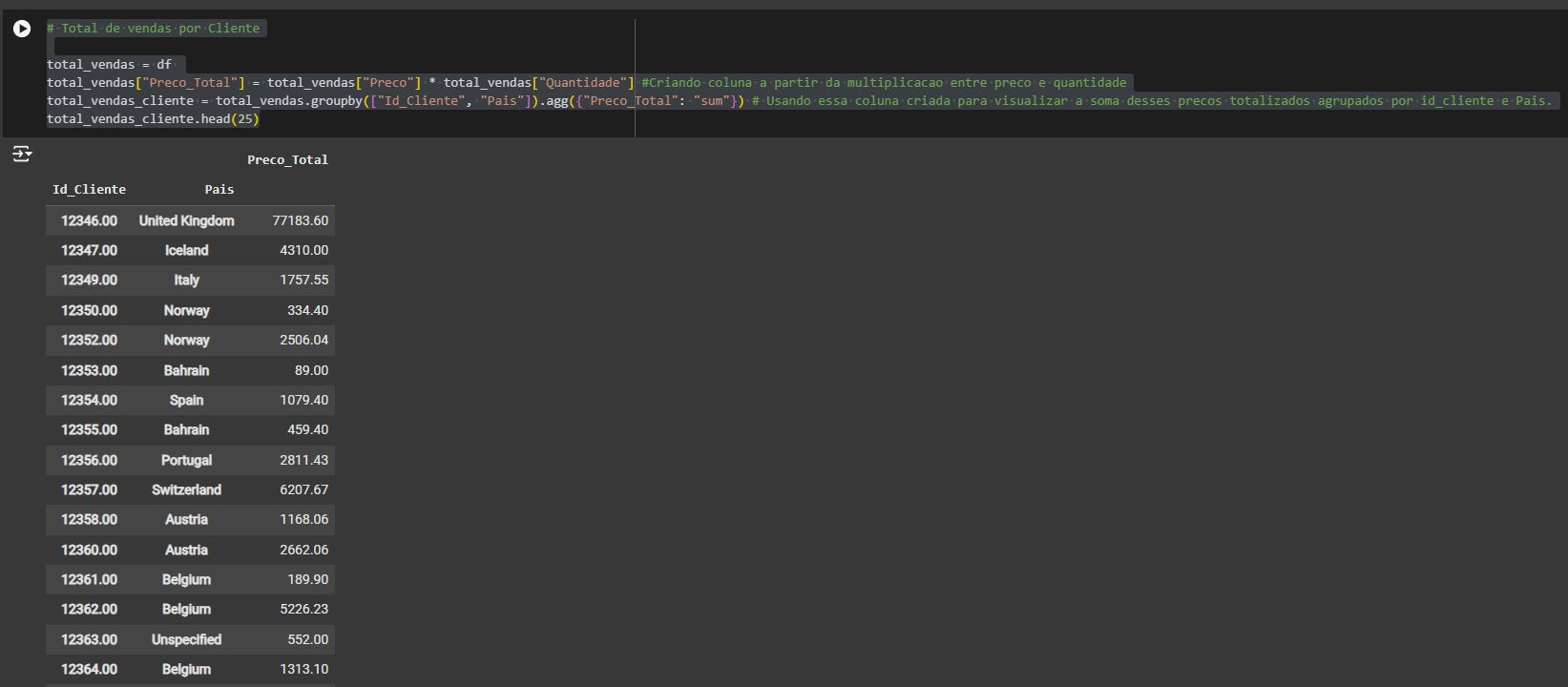


.

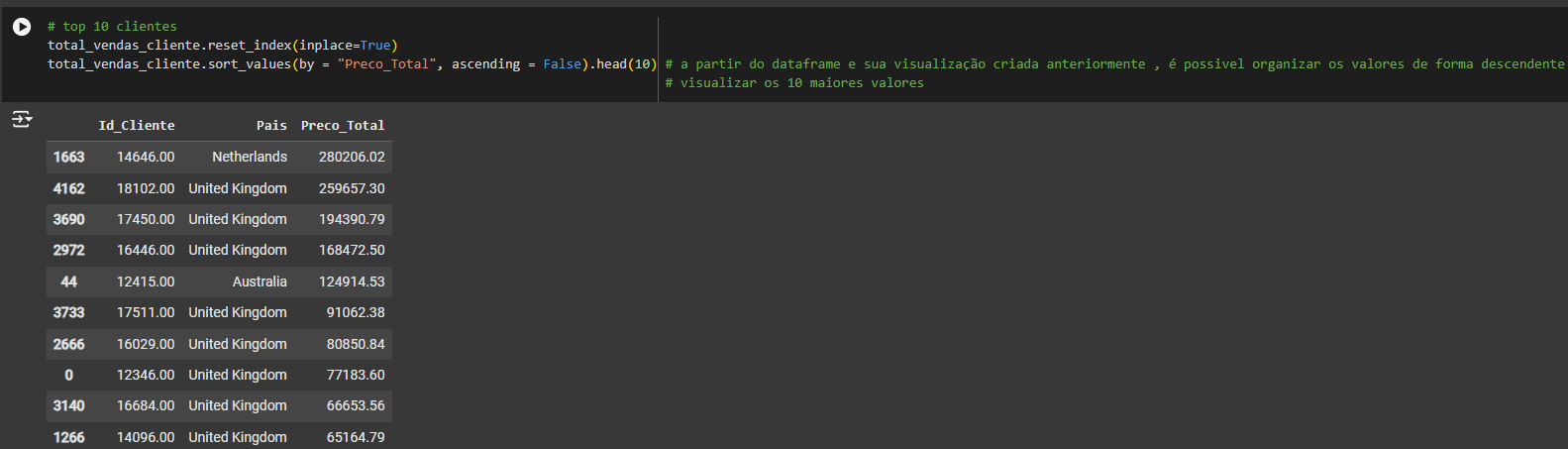
# 4. Análise e Exploração dos Dados

Após análise inicial do dataset, foi abordada a totalização das vendas por cliente, criando uma nova coluna chamada ‘Preco\_Total’, que resultou da multiplicação entre o ‘Preco’ e a ‘Quantidade’. Em seguida, os dados foram agrupados por ‘Id\_Cliente’ e Pais, permitindo identificar os clientes com as maiores vendas.

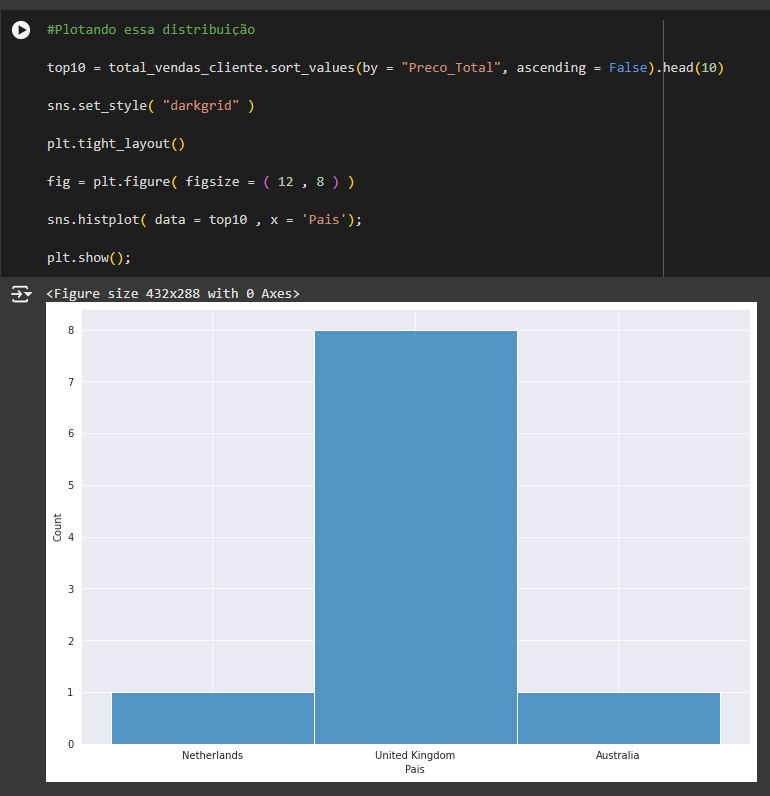
**Figura 12 – Criando Preco\_Total.**



**Figura 13 – Maiores Vendas.**

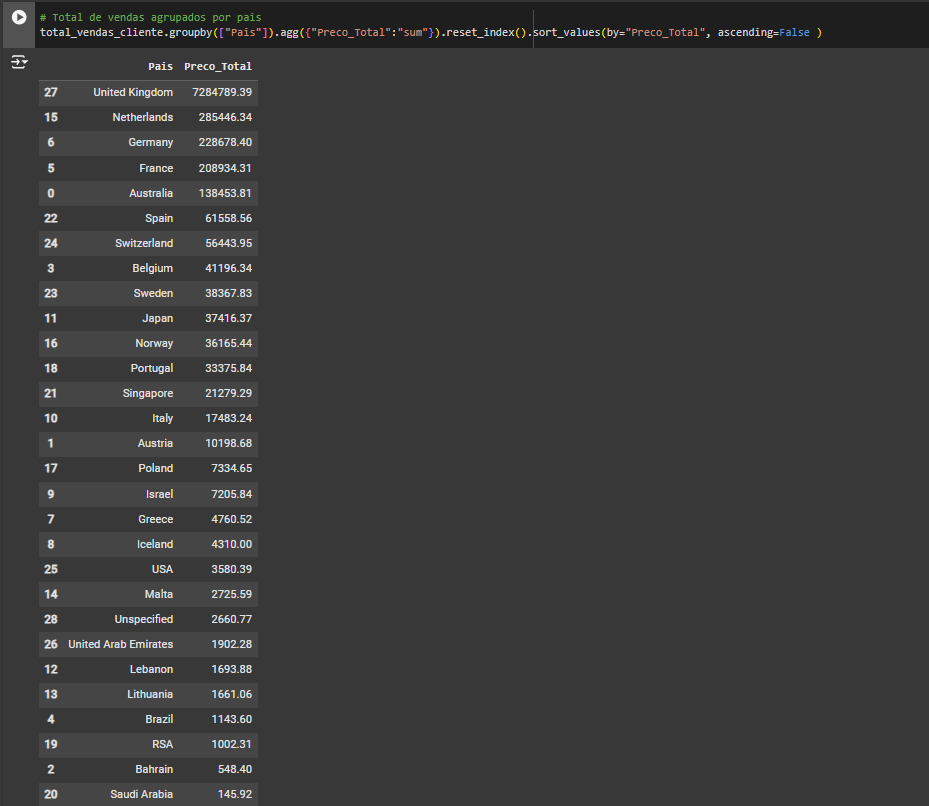


**Figura 14 – Distribuição dos 10 maiores clientes por país.**

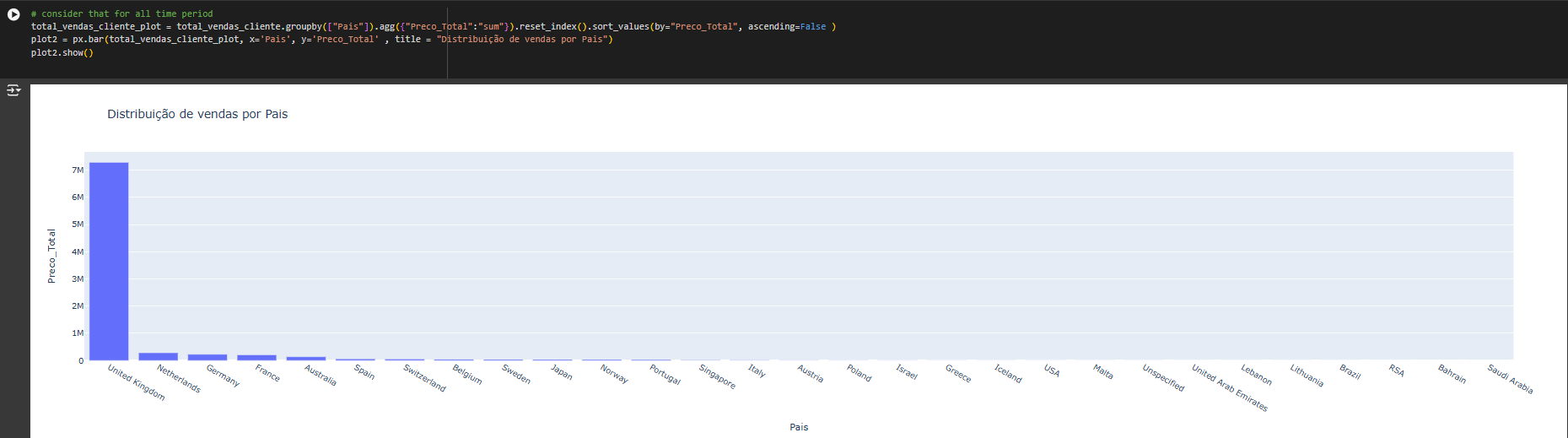


A análise da distribuição das vendas revela que 80% das dez maiores vendas estão concentradas no Reino Unido. Esse insight é significativo, pois indica um mercado forte e potencialmente leal, onde os clientes podem ter uma maior disposição para gastar. Essa concentração também sugere que estratégias de marketing e promoção podem ser especialmente eficazes nesse país, permitindo que as redes de supermercados aproveitem esse comportamento para otimizar suas vendas e atender melhor às necessidades dos consumidores locais.

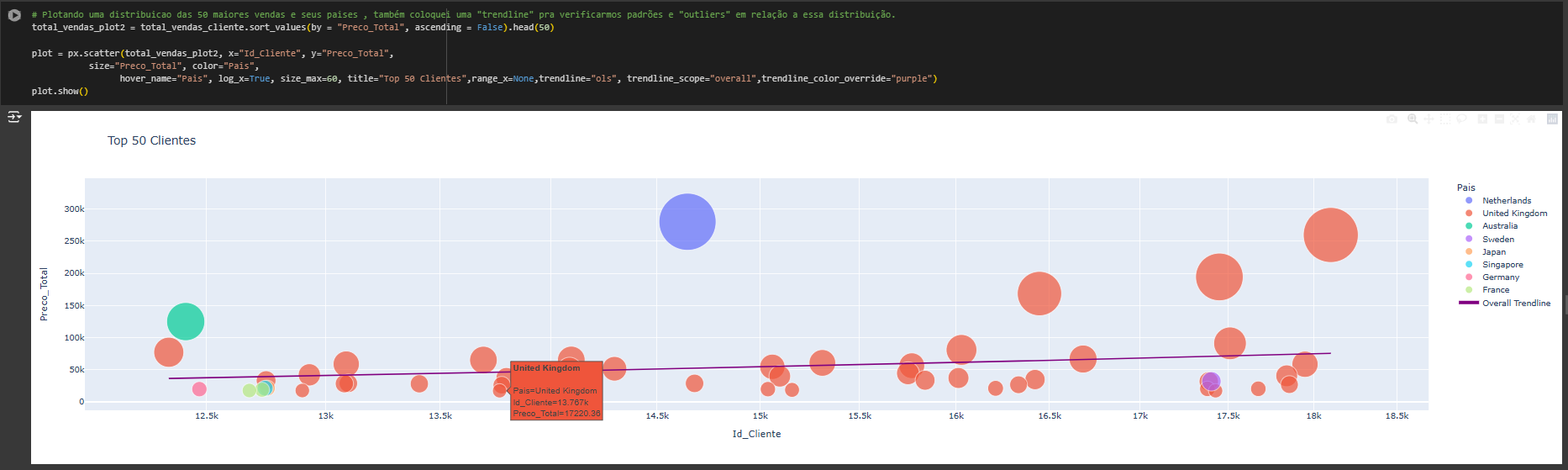
**Figura 15 – Total de vendas por país.**



**Figura 16 – Total de vendas por país 2.**



**Figura 17 – Distribuição das maiores vendas por país com trendline.**



A visualização das 50 maiores vendas mostra que, de maneira geral, as vendas seguem uma tendência próxima à linha de tendência (trendline), indicando um comportamento consistente entre os clientes. No entanto, o cliente da "Netherlands" se destaca como um outlier, apresentando vendas significativamente mais altas do que a média. Esse ponto fora da curva merece atenção, pois pode indicar um cliente especial ou uma situação particular que influenciou suas compras, sugerindo oportunidades para explorar essa dinâmica no mercado local.

O dataset não contém um valor total de vendas, o que limita as inferências sobre a relação entre variáveis. Portanto, foram levantadas duas hipóteses de análise futura:

1. **Regressão Supervisionada**: Criar um sub-dataset que inclua o total de vendas, permitindo prever vendas futuras com base em variáveis como país e cliente.
2. **Aprendizado Não Supervisionado**: Aplicar algoritmos de clusterização ou associação para explorar padrões de compra e identificar segmentos de clientes.

Essas abordagens visam enriquecer a análise e fornecer insights mais profundos sobre o comportamento de compra no contexto do varejo.

# .

# 5. Criação de Modelos de Machine Learning

Nesta seção, propomos duas abordagens distintas para a criação de modelos de machine learning, focando na previsão do valor total de vendas. A primeira hipótese envolve a criação de um sub-dataset que agrupa os dados por país e cliente.

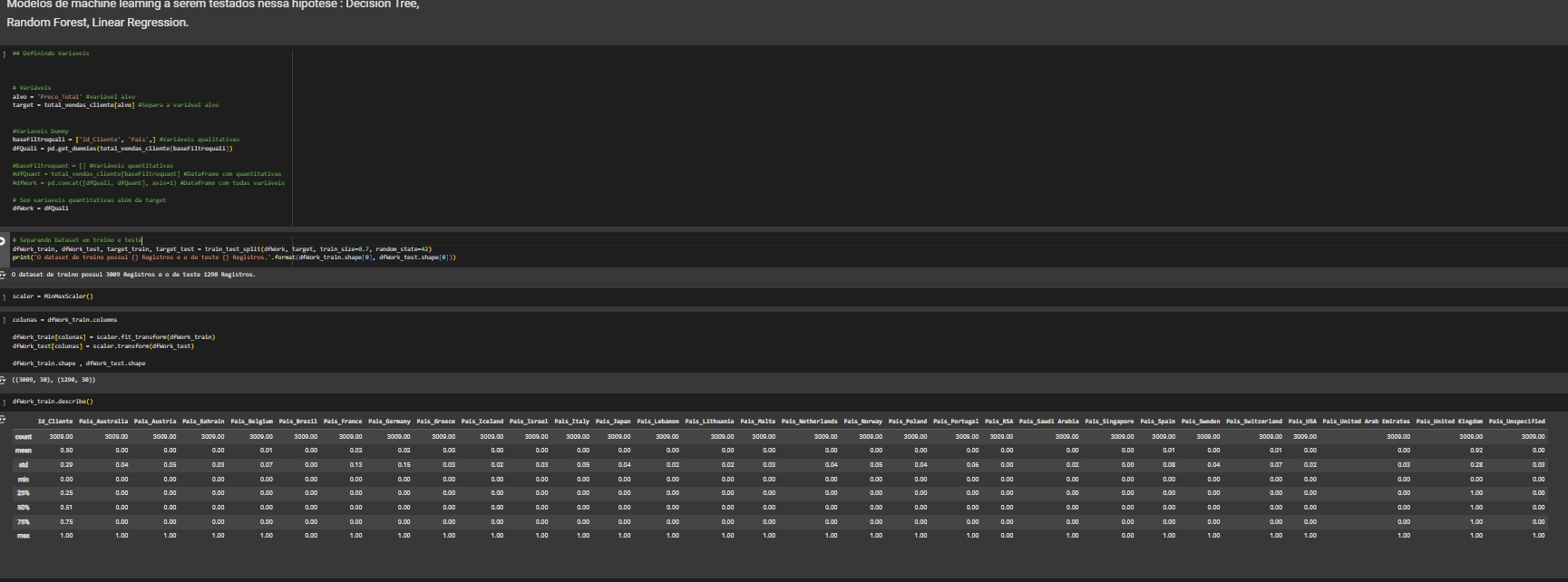
A segunda abordagem considera o uso de aprendizado não supervisionado, empregando algoritmos de clusterização ou associação.

**5.1 Hipótese 1 - Sub-dataset para Previsão de Preço Total**

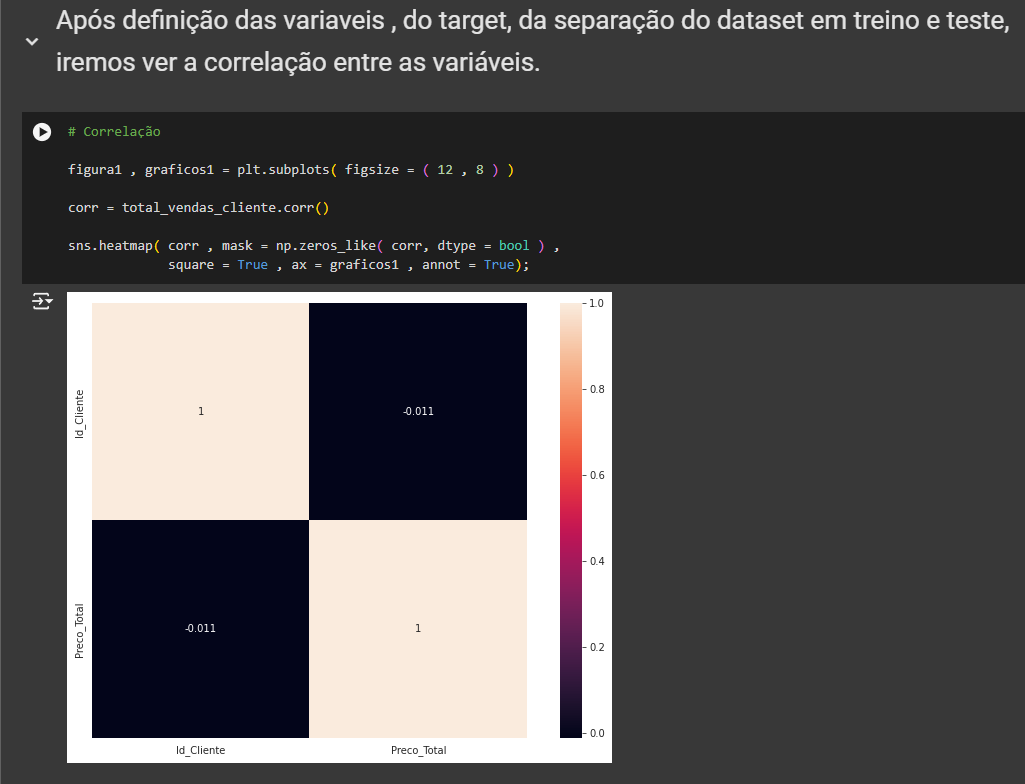
Nesta etapa, a criação de modelos de machine learning visa prever o ‘Preco\_Total’, utilizando um sub-dataset que agrupa dados por país e cliente. A análise exploratória inicial, incluindo histogramas e gráficos de dispersão, proporciona uma compreensão da distribuição e das relações entre as variáveis.

Os modelos a serem testados incluem Random Forest e Linear Regression, selecionados por sua capacidade de capturar diferentes padrões nos dados. Essa abordagem permitirá comparar o desempenho e a eficácia na previsão do valor total de vendas, fundamentando decisões estratégicas para o negócio.

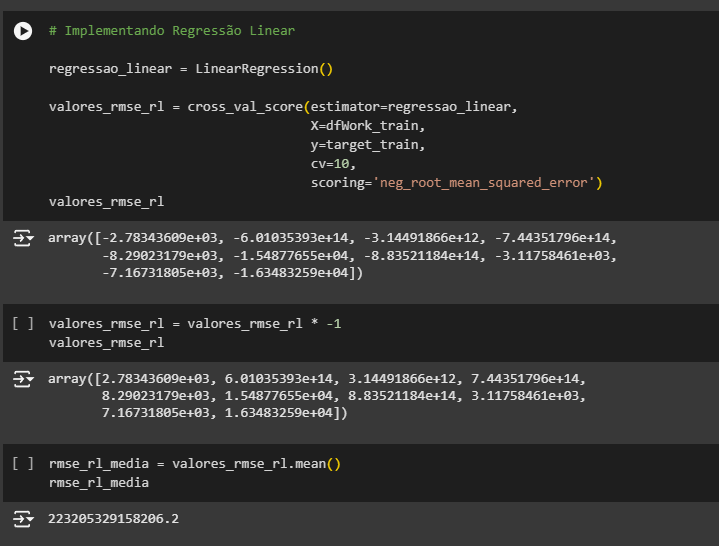
**Figura 18 – Definição das variáveis e divisão em treino e teste.**



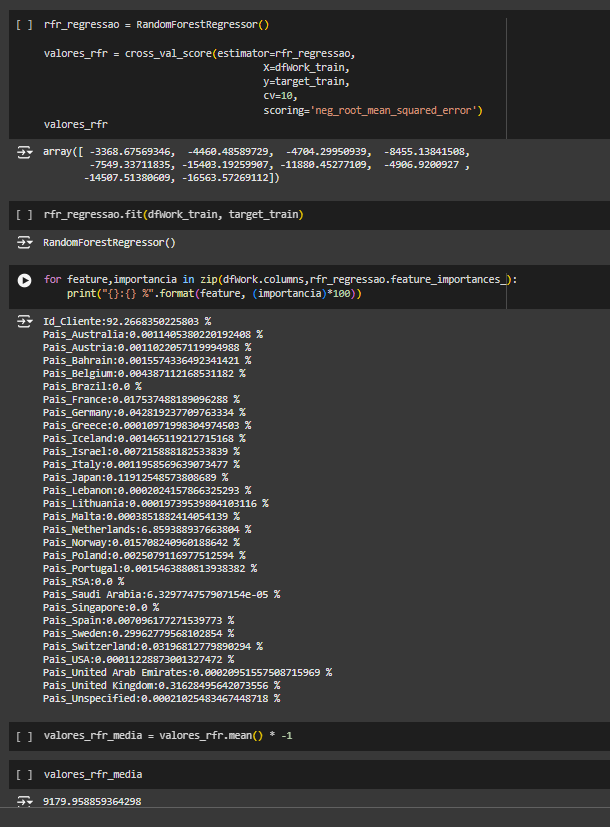
**Figura 19 – Correlações.**



**Figura 20 – Regressão Linear.**



**Figura 21 – Random Forest.**

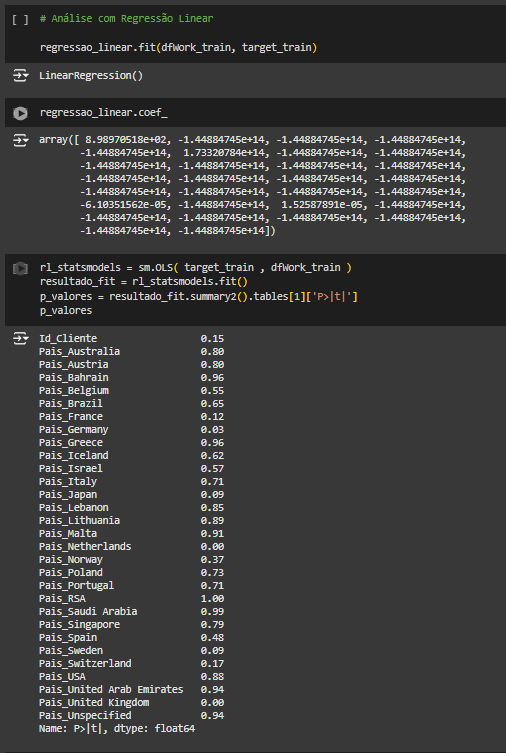


**Figura 22 – Comparando modelos.**

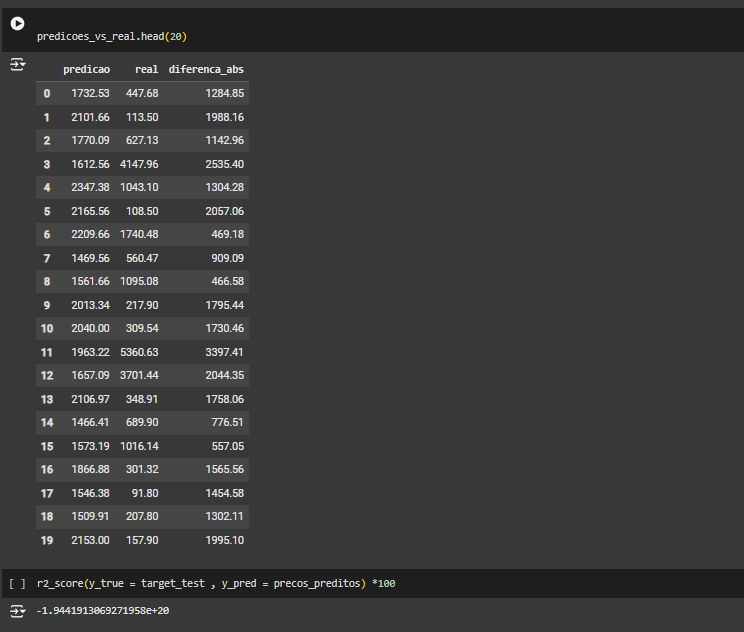


A partir da comparação dos resultados dos modelos, optou-se por utilizar a regressão linear.

**Figura 23 – Predição com regressão linear.**



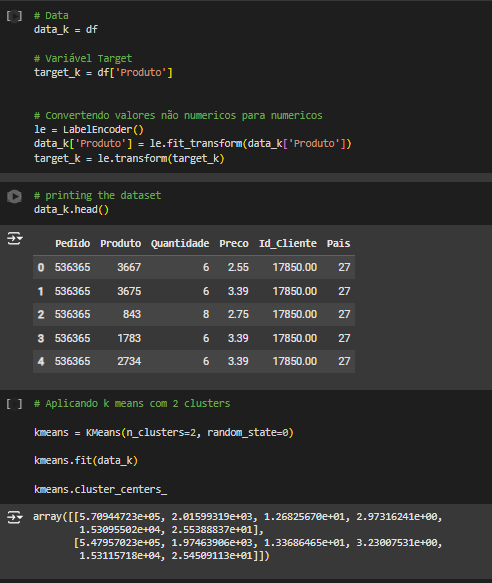
**Figura 24 – Resultados Regressão Linear.**



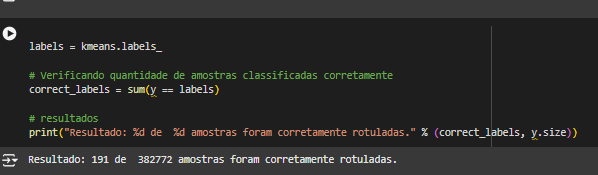
**5.2 Hipótese 2: Aprendizado Não Supervisionado**

Nesta hipótese, propomos a utilização de técnicas de aprendizado não supervisionado, como algoritmos de clusterização e associação, para explorar o dataset. O objetivo é identificar grupos de produtos ou clientes com características similares, permitindo a descoberta de padrões ocultos que podem não ser evidentes em análises tradicionais. A abordagem de associação pode, por exemplo, revelar quais produtos são frequentemente comprados juntos, oferecendo insights valiosos para a criação de promoções e estratégias de marketing direcionadas. Esta análise contribuirá para um melhor entendimento do comportamento do consumidor e para a otimização das operações de vendas. Nesse estudo aplicamos duas abordagens, k-means e apriori.

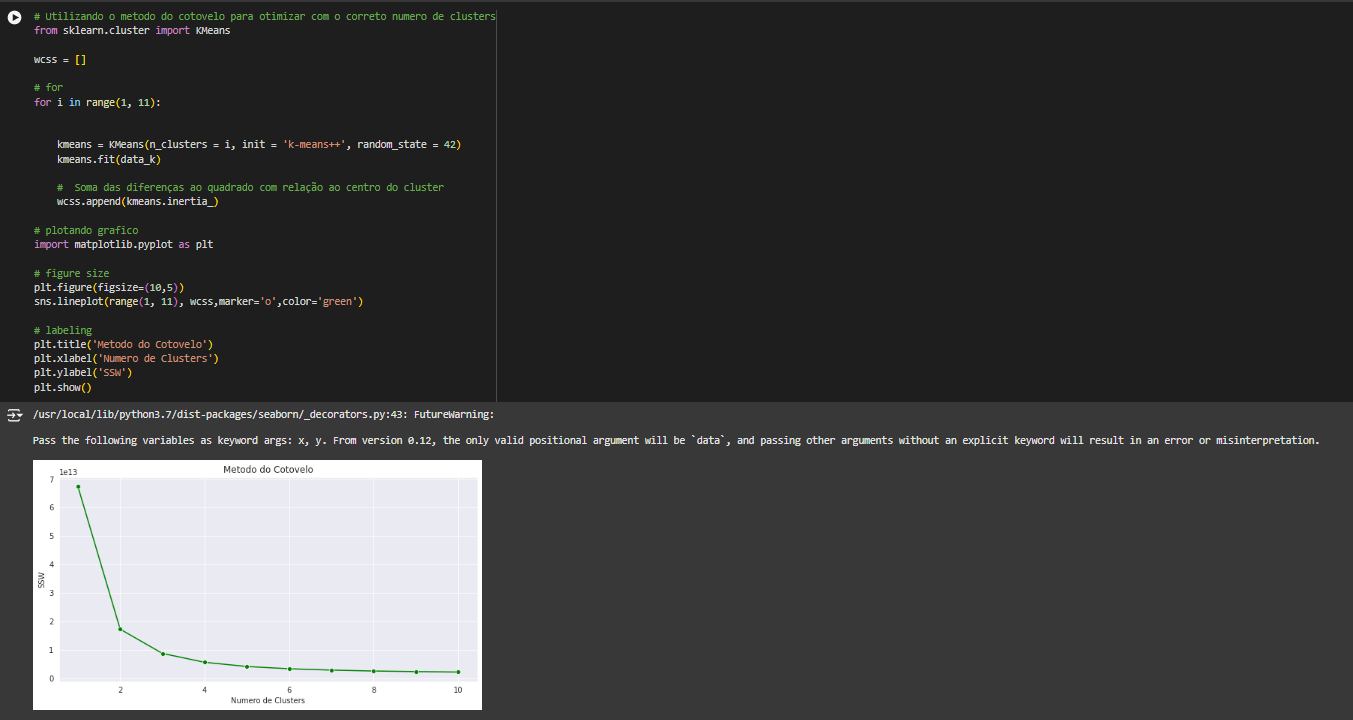
**Figura 25 – Ajustando dataset para kmeans**



**Figura 26 – Resultados com 2 clusters**

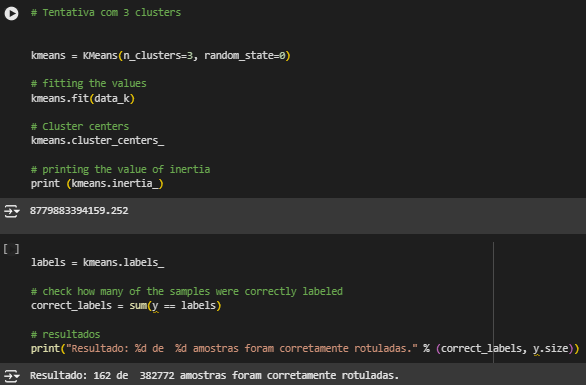


**Figura 27 – Aplicando método do cotovelo para otimização.**



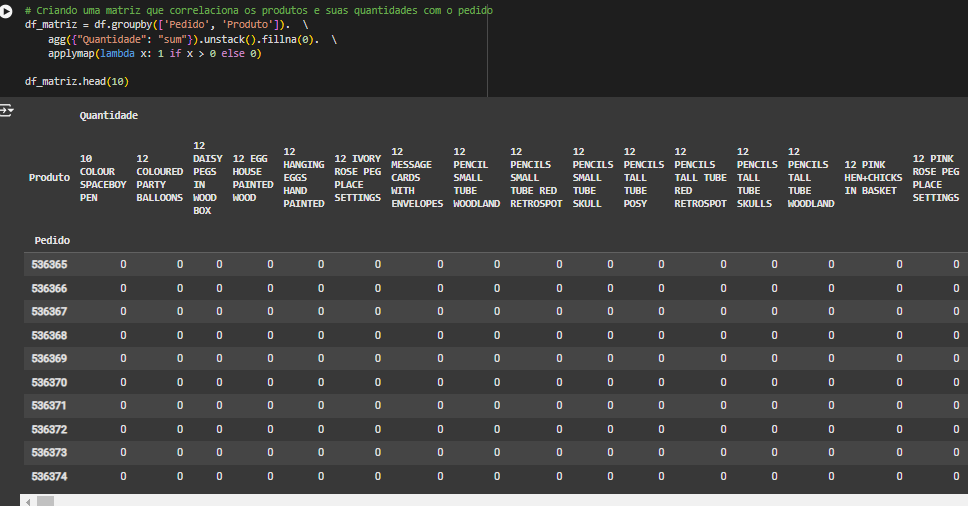
# O método do cotovelo é uma maneira prática de escolher o número certo de clusters ao usar o K-Means. A ideia é simples: você plota a soma dos erros quadráticos (SSE) para diferentes números de clusters e observa o gráfico. Conforme você adiciona mais clusters, a SSE diminui, mas chega um ponto onde essa diminuição desacelera—é o "cotovelo". Esse ponto sugere um bom equilíbrio, evitando a complexidade desnecessária e garantindo que os clusters sejam significativos e úteis para a análise.

**Figura 28 – Teste com 3 clusters.**

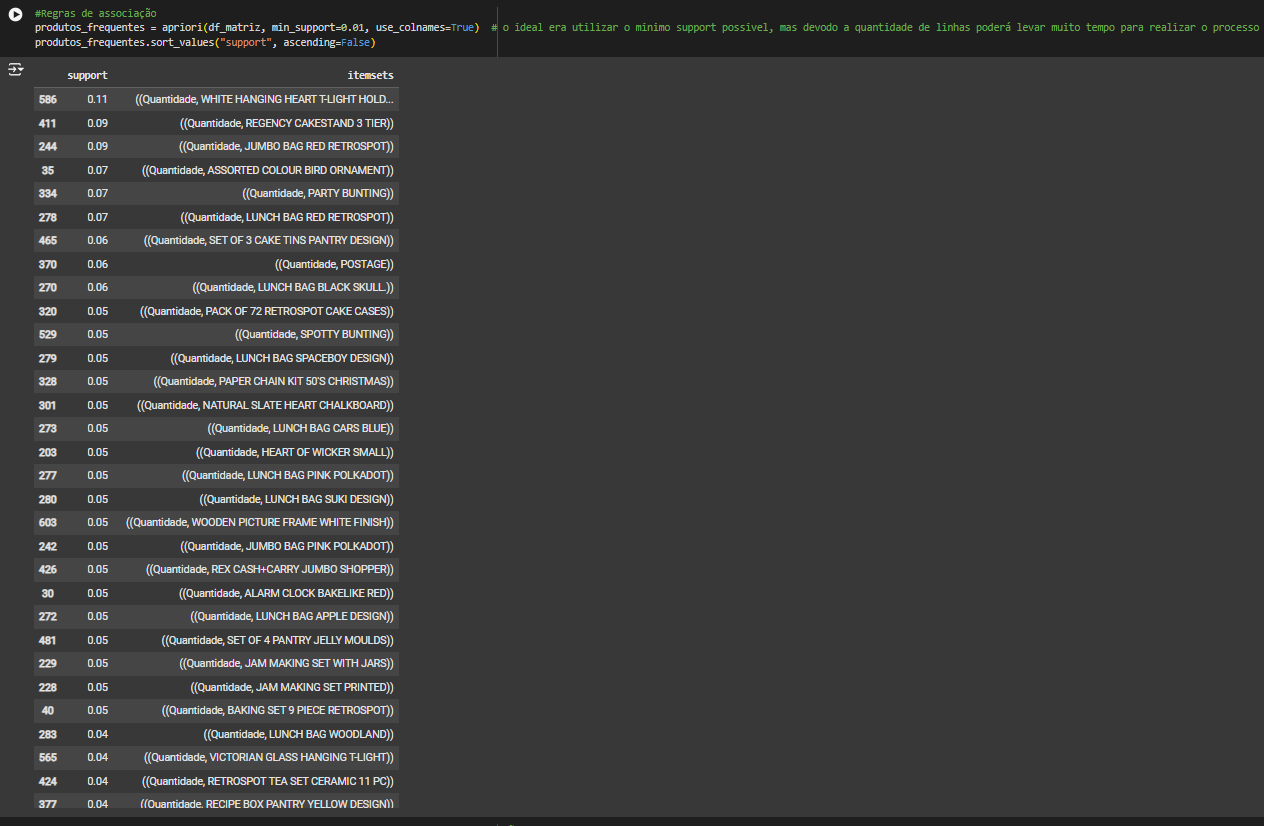


# Em seguida, foi utilizado a abordagem Apriori, a análise se concentra nas associações entre produtos nas compras dos clientes. A ideia é identificar quais produtos são frequentemente adquiridos juntos. Isso nos permite descobrir padrões interessantes no comportamento do consumidor, como, por exemplo, se a compra de um determinado item geralmente leva à compra de outro. Com essas informações, é possível otimizar estratégias de marketing, criar promoções mais eficazes e oferecer recomendações personalizadas, melhorando a experiência de compra do cliente.

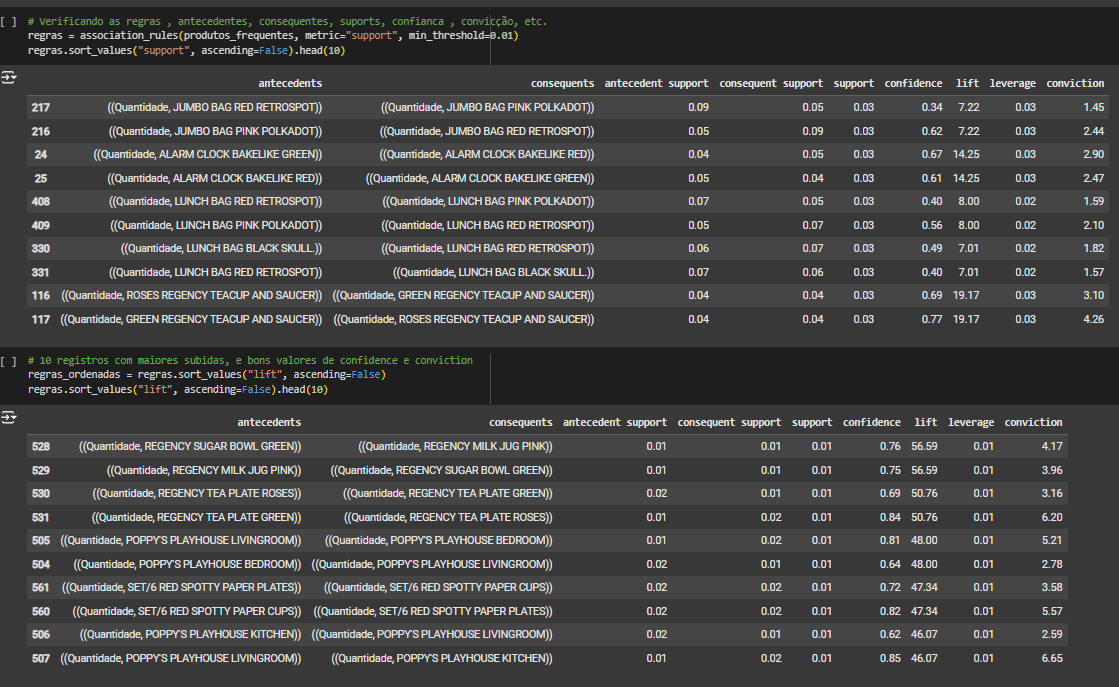
**Figura 29 – Matriz para correlação apriori.**



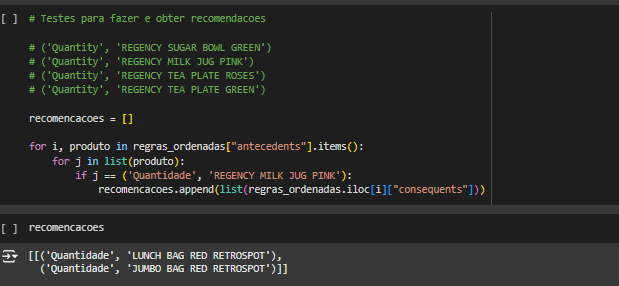
**Figura 30 – Regras de associação.**



**Figura 31 – Métricas e produtos com maior confiança.**



**Figura 32 – Testes de recomendação.**



# 6. Apresentação dos Resultados

Nesta seção, apresentamos os resultados da análise e dos modelos aplicdos, refletindo sobre as descobertas significativas e sua relevância.

**6.1 Avaliação da Hipótese 1**

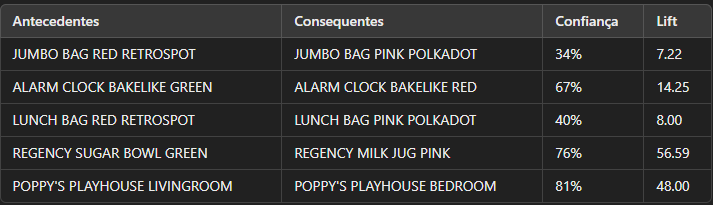
Ao examinar a primeira hipótese, a previsão do valor total de vendas com um modelo de regressão, observamos que o modelo apresentou um R² score de -1.94e+20, o que indica um grave problema de underfitting. Isso sugere que as variáveis selecionadas não possuem correlação suficiente para explicar a variabilidade dos dados, levando à conclusão de que o sub-dataset gerado não foi eficaz. Essa falha nos levou a descartar a abordagem de regressão e a buscar novas alternativas.

**6.2 Análise por Associação com Apriori**

Dando continuidade à nossa investigação, aplicamos o algoritmo Apriori para realizar uma análise por associação, aproveitando os registros de compra que envolvem múltiplos produtos por cliente. O objetivo foi identificar quais itens tendem a ser comprados juntos.

Os resultados das regras de associação revelaram as seguintes informações valiosas:

**Figura 33 – Melhores resultados dos produtos.**

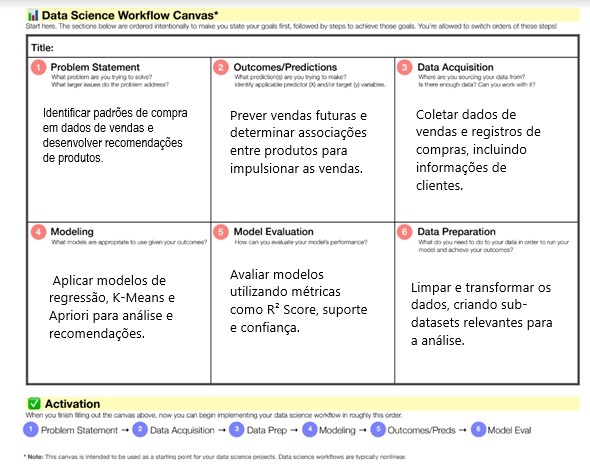


Essas regras demonstram que a maioria dos itens com altos valores de confiança (> 70%) e lift são produtos que podem ser estrategicamente agrupados em promoções. Por exemplo, a associação entre **"**JUMBO BAG RED RETROSPOT" e "JUMBO BAG PINK POLKADOT**"** poderia ser utilizada para oferecer um desconto na compra conjunta, incentivando o cliente a adicionar ambos os itens ao carrinho.

**6.3 Implicações para o Negócio**

Os insights obtidos a partir da análise por associação oferecem uma oportunidade significativa para melhorar as estratégias de marketing e vendas. A identificação de itens que costumam ser comprados juntos permite a implementação de recomendações personalizadas e promoções direcionadas. Isso pode não apenas aumentar a taxa de conversão, mas também melhorar a experiência do cliente ao facilitar escolhas relevantes durante o processo de compra.

Essas descobertas contribuem para a nossa missão inicial de entender melhor o comportamento do consumidor e maximizar as vendas através de insights orientados por dados.



# 7. Links

[*https://github.com/JRafaQuadros91/TCC-2024*](https://github.com/JRafaQuadros91/TCC-2024)

[*https://www.youtube.com/watch?v=\_ZrEfSkO3IQ*](https://www.youtube.com/watch?v=_ZrEfSkO3IQ)