Explorando Dados e Modelos Não Supervisionados para Insights Empresariais

Ana Paula Vanderley

- Este trabalho buscou investigar e analisar os dados de uma concessionária de caminhões (DADOS REAIS algumas análises foram omitidas), em busca de fornecer insights para os gestores da empresa. Nossa abordagem apliquei técnicas estatísticas como análise exploratória de dados (EDA), com foco especial na análise temporal dos dados. Além disso, apliquei técnicas de modelagem não supervisionada, como Análise de Componentes Principais (PCA) e o algoritmo K-Means.

- Este processo incluiu uma EDA detalhada, onde foi verificado tendências, padrões e anomalias nos dados. Além disso, a análise temporal foi aplicada para compreender como o desempenho da empresa tem evoluído ao longo do tempo. Isso permitiu identificar sazonalidades, ciclos de negócios e outras informações temporais relevantes que possam influenciar as operações da concessionária.

- Após a conclusão da análise exploratória de dados, foi aplicado técnicas de modelagem não supervisionada para aprofundar a compreensão dos dados. Primeiramente, a Análise de Componentes Principais (PCA) para reduzir a dimensionalidade do conjunto de dados, simplificando-o enquanto mantemos a maior parte da variância original. Em seguida, o algoritmo K-Means para agrupar os dados em clusters significativos com base em suas características.

Objetivos:

- Identificar padrões e tendências nos dados.
- Compreender a evolução temporal do desempenho da empresa.
- Extrair insights acionáveis utilizando técnicas de modelagem não supervisionada.
- Fornecer recomendações e orientações estratégicas para os gestores da empresa com base nos resultados da análise.

VARIÁVEIS PRESENTE NO BANCO DE DADOS

Out[4]:

```
['DATA E HORA',
 'NaturezaOperacao',
 'DEPARTAMENTO QUE VENDEU (OFICINA OU BALCÃO)',
 'CONDIÇÃO DE PAGAMENTO',
 'NFItem_PercDesc',
 'QtdeEstoque',
 'VlMargemCont',
 'NOME DA PEÇA',
 'VENDEDOR',
 'VlDesc',
 'QUANTIDADE',
 'VALOR UNITÁRIO',
 'ProdCusto',
 'ProdPrecoValor',
```

```
 'ClasABC',
 'ValorICMS',
 'ValorICMSST',
 'ValorPisCofins',
 'ValorLucroBruto',
 'VALOR DA VENDA',
 'NFItem_VlBruto',
 'NFItem_PercMargemCont',
 'NFItem_PercMargemGer']
```

TAMANHO DA BASE

```
Total de registros: 9975
Total de variáveis: 23
```

NULOS POR VARIÁVEL

Out[ ]:

	Valores Nulos	Porcentagem de Nulos
DATA E HORA	0	0.00%
NaturezaOperacao	0	0.00%
DEPARTAMENTO QUE VENDEU (OFICINA OU BALCÃO)	0	0.00%
CONDIÇÃO DE PAGAMENTO	0	0.00%
NFItem_PercDesc	0	0.00%
QtdeEstoque	0	0.00%
VlMargemCont	0	0.00%
NOME DA PEÇA	0	0.00%
VENDEDOR	0	0.00%
VlDesc	0	0.00%
QUANTIDADE	0	0.00%
VALOR UNITÁRIO	0	0.00%
ProdCusto	0	0.00%
ProdPrecoValor	0	0.00%
ClasABC	0	0.00%
ValorICMS	0	0.00%
ValorICMSST	0	0.00%
ValorPisCofins	0	0.00%
ValorLucroBruto	0	0.00%
VALOR DA VENDA	0	0.00%
NFItem_VlBruto	0	0.00%
NFItem_PercMargemCont	0	0.00%
NFItem_PercMargemGer	0	0.00%

EDA (Análise Exploratória de Dados)

MEDIDAS DESCRITIVAS

	DATA E HORA	NaturezaOperacao	DEPARTAMENTO QUE VENDEU (OFICINA OU BALCÃO)	CONDIÇÃO DE PAGAMENTO	NFItem_PercDesc	QtdeEstoque	Vl
count	9975	9975	9975	9975	9975.000000	9975.000000	9975.000000
unique	3370	1	2	17	NaN	NaN	NaN
top	20/06/2019 15:35:00	VEN	OFI	BOL. 30 DIAS	NaN	NaN	NaN
freq	51	9975	5038	2286	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	8.612488	2.496040	141.114661
std	NaN	NaN	NaN	NaN	9.409881	6.318097	296.175585
min	NaN	NaN	NaN	NaN	0.000000	1.000000	-1438.500000
25%	NaN	NaN	NaN	NaN	0.000000	1.000000	25.175000
50%	NaN	NaN	NaN	NaN	9.990000	1.000000	59.820000
75%	NaN	NaN	NaN	NaN	14.985000	2.000000	137.190000
max	NaN	NaN	NaN	NaN	66.670000	260.000000	5899.800000

11 rows × 24 columns

DESCRIÇÃO DAS VARIÁVEIS CATEGORICAS

```
Coluna: NaturezaOperacao
Valores únicos: ['VEN']
Categories (1, object): ['VEN']
Valor da moda: VEN
--------------------------------------------------------------------------
Coluna: DEPARTAMENTO QUE VENDEU (OFICINA OU BALCÃO)
Valores únicos: ['OFI', 'BLC']
Categories (2, object): ['BLC', 'OFI']
Valor da moda: OFI
--------------------------------------------------------------------------
Coluna: CONDIÇÃO DE PAGAMENTO
Valores únicos: ['BOL. 30 / 60 DIAS', 'BOL. 30 DIAS', 'BOL. 30 / 60 /
90 DIAS', 'RECEBIDO ANTECIPADAMENTE (PECAS / SERVICOS)', 'A VISTA',
```

```
..., 'BOL. 28 / 56 DIAS', 'EMPRESAS DO GRUPO', 'C / APRESENTACAO',
'BOL. 30 / 60 / 90 / 120 / 150 / 180 DIAS', 'BOL. 30 / 60 / 90 / 120 /
150 DIAS']
Length: 17
Categories (17, object): ['A VISTA', 'BOL. 28 / 56 DIAS', 'BOL.
28/56/84 DIAS',
                          'BOL. 28/56/84/112 DIAS', ..., 'GARANTIA',
'PARCELAS VARIADAS (R)',
                          'RECEBIDO ANTECIPADAMENTE (PECAS /
SERVICOS)', 'VENDA INTERNA']
Valor da moda: BOL. 30 DIAS
--------------------------------------------------------------------------
Coluna: VENDEDOR
Valores únicos: ['ANA', 'FERNANDO', 'ICARO', 'ISABELA', 'JOAO',
'LEANDRO', 'STELLA']
Categories (7, object): ['ANA', 'FERNANDO', 'ICARO', 'ISABELA',
'JOAO', 'LEANDRO', 'STELLA']
Valor da moda: ICARO
--------------------------------------------------------------------------
Coluna: ClasABC
Valores únicos: ['B3 ', 'C? ', 'A3 ', 'B2 ', 'A1 ', 'A2 ', 'B1 ', 'C4
', '   ']
Categories (9, object): ['   ', 'A1 ', 'A2 ', 'A3 ', ..., 'B2 ', 'B3
', 'C4 ', 'C? ']
Valor da moda: A3
--------------------------------------------------------------------------
```

VALORES ÚNICOS POR VARIÁVEL

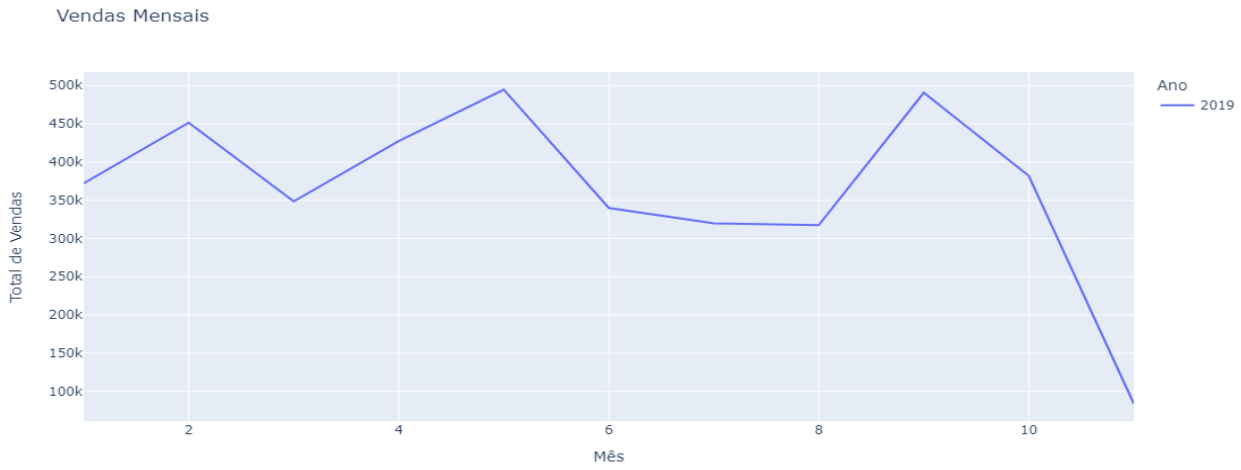
Out[ ]:

	index	Valores únicos
0	DATA E HORA	3370
1	NaturezaOperacao	1
2	DEPARTAMENTO QUE VENDEU (OFICINA OU BALCÃO)	2
3	CONDIÇÃO DE PAGAMENTO	17
4	NFItem_PercDesc	590
5	QtdeEstoque	45
6	VlMargemCont	6644
7	NOME DA PEÇA	1372
8	VENDEDOR	7
9	VlDesc	3492
10	QUANTIDADE	45
11	VALOR UNITÁRIO	3499
12	ProdCusto	4846
13	ProdPrecoValor	3194

	index	Valores únicos
14	ClasABC	9
15	ValorICMS	715
16	ValorICMSST	42
17	ValorPisCofins	2344
18	ValorLucroBruto	5874
19	VALOR DA VENDA	5772
20	NFItem_VlBruto	4265
21	NFItem_PercMargemCont	3760
22	NFItem_PercMargemGer	3760

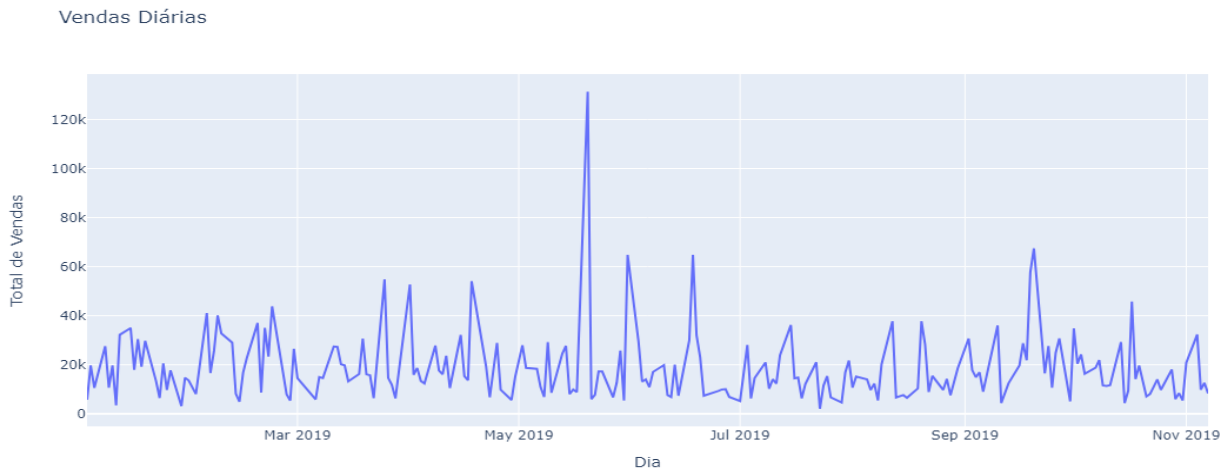
ANÁLISE TEMPORAL

VALOR DAS VENDAS MENSAIS



VALOR DAS VENDAS DIÁRIAS

```
           Dia  VALOR DA VENDA
0   2019-01-02         5753.46
1   2019-01-03        19738.17
2   2019-01-04        10624.62
3   2019-01-07        27587.74
4   2019-01-08        10703.95
..         ...             ...
212 2019-11-01        20884.54
213 2019-11-04        32298.31
214 2019-11-05         9820.70
215 2019-11-06        12603.99
216 2019-11-07         8342.53
```



LUCRO ANUAL DA EMPRESA

- O lucro total é um dos principais indicadores do desempenho financeiro de uma empresa. Ele reflete a capacidade da empresa de gerar lucro a partir de suas operações e demonstra sua eficiência na gestão de custos.
- Analisar o lucro total permite avaliar a viabilidade financeira do negócio a longo prazo. Um lucro consistente e crescente indica que a empresa está saudável financeiramente e tem potencial para crescimento sustentável.

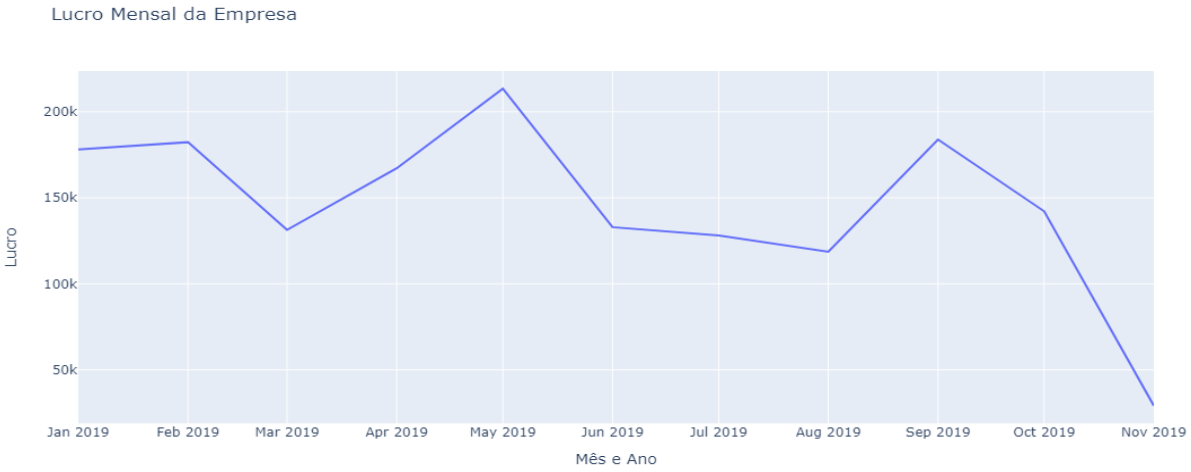
- O lucro total de uma empresa é calculado subtraindo-se os custos totais das receitas totais durante um determinado período de tempo. Em outras palavras, a fórmula básica para calcular o lucro é:

Lucro Total

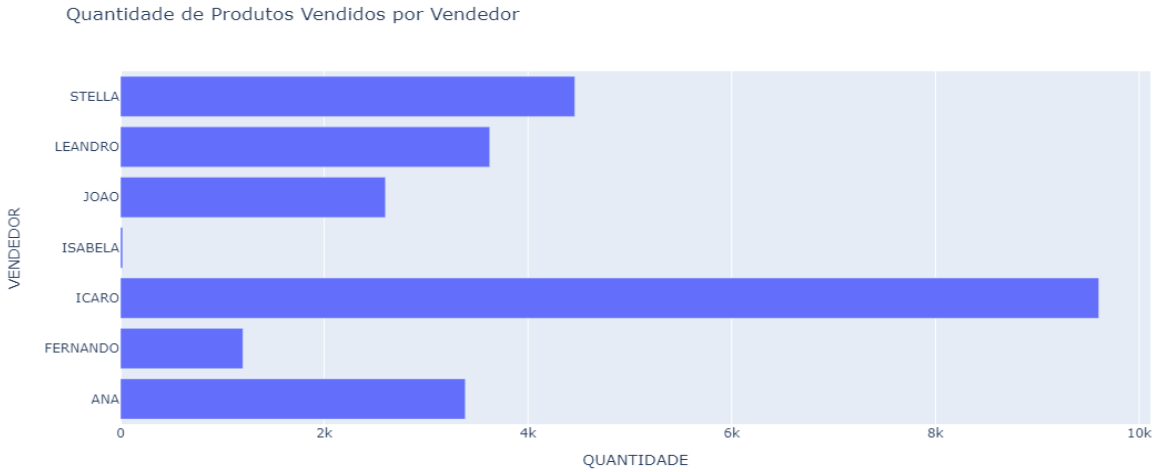
Receita Total − Custos Totais Lucro Total=Receita Total−Custos Totais

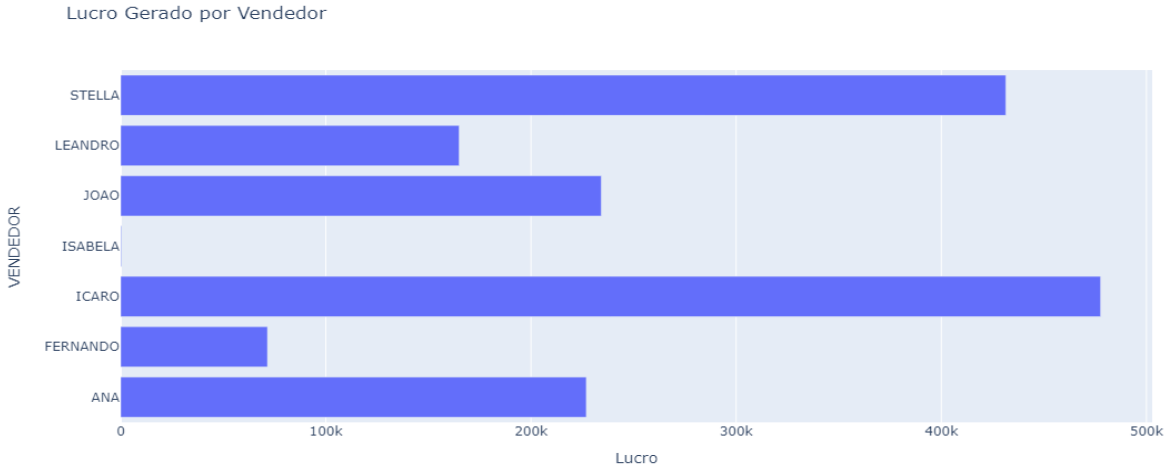
```
Lucro total da empresa: R$ 1607205.18
```

LUCRO MENSAL



ANALISE DESEMPENHO DOS FUNCIONARIOS



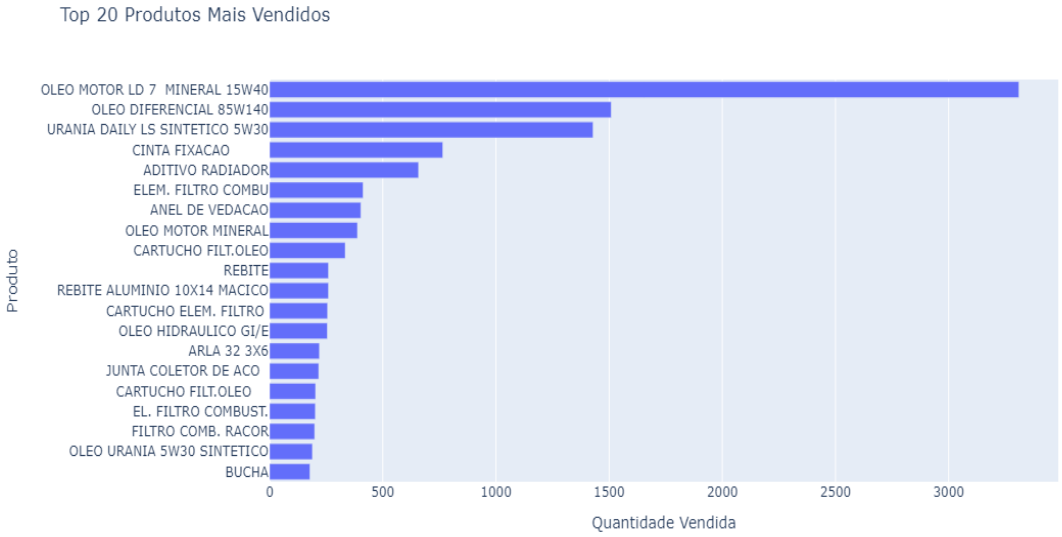


ANÁLISE DE PRODUTOS E ESTOQUE

A análise de produtos e estoque ajuda a empresa a entender quais produtos estão em alta demanda e quais estão com baixo desempenho. Isso permite uma gestão mais eficiente do inventário, garantindo que haja disponibilidade adequada dos produtos mais procurados e evitando o excesso de estoque de itens com baixa rotatividade.

PRODUTOS MAIS VENDIDOS

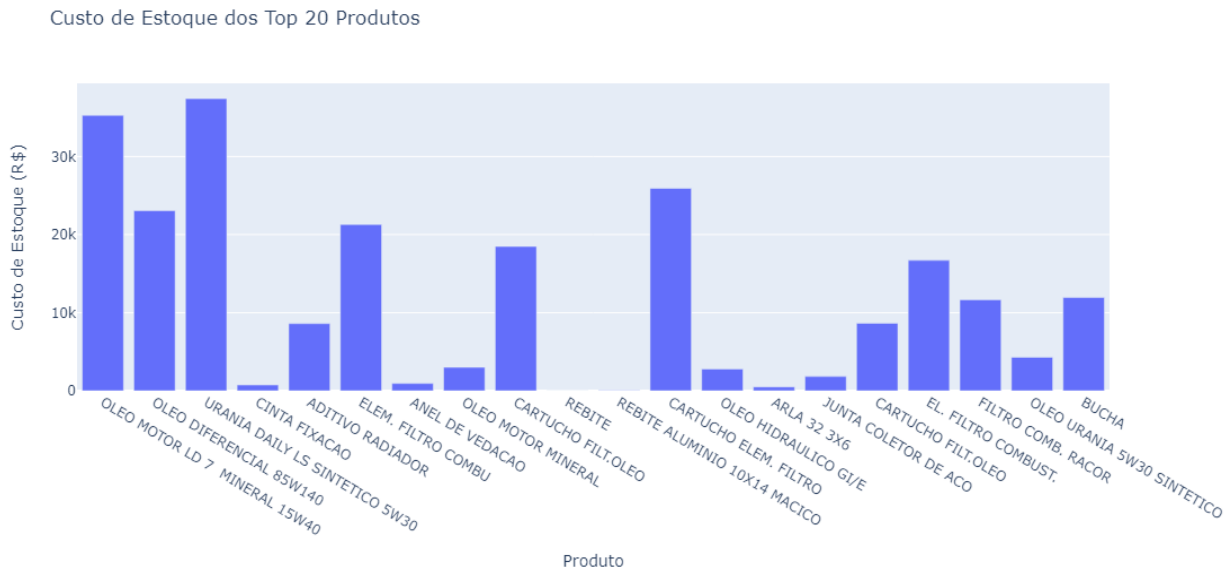
	NOME DA PEÇA	QUANTIDADE
0	OLEO MOTOR LD 7 MINERAL 15W40	3310
1	OLEO DIFERENCIAL 85W140	1509
2	URANIA DAILY LS SINTETICO 5W30	1429
3	CINTA FIXACAO	765
4	ADITIVO RADIADOR	658
5	ELEM. FILTRO COMBU	413
6	ANEL DE VEDACAO	403
7	OLEO MOTOR MINERAL	388
8	CARTUCHO FILT.OLEO	334
9	REBITE	260
10	REBITE ALUMINIO 10X14 MACICO	260
11	CARTUCHO ELEM. FILTRO	256
12	OLEO HIDRAULICO GI/E	255
13	ARLA 32 3X6	220
14	JUNTA COLETOR DE ACO	217
15	CARTUCHO FILT.OLEO	203
16	EL. FILTRO COMBUST.	202
17	FILTRO COMB. RACOR	199
18	OLEO URANIA 5W30 SINTETICO	189
19	BUCHA	178



ESTOQUE

```
Quantidade total em estoque: 24898
Valor total em estoque: R$ 2420673.48
```

Atualmente a empresa tem um total de 24.898 itens em estoque. Essa é uma métrica fundamental que nos ajuda a entender a escala das operações e a capacidade de atender à demanda dos clientes. Além disso, o valor total de estoque é de R$ 2.420.673,48. Essa cifra reflete o investimento significativo que empresa fez em produtos e nos fornece uma visão clara do valor desses ativos físicos.



CIÊNCIA DE DADOS

A estatística multivariada compreende um conjunto de técnicas que analisam simultaneamente um conjunto de variáveis que caracterizam os objetos ou indivíduos de uma amostra. Usualmente essas técnicas são classificadas em técnicas de dependência ou interdependência.

- Nas técnicas de dependência uma variável (variável dependente) é explicada por outras variáveis (variáveis independentes). Temos como exemplo modelos de regressão múltipla e a análise discriminante.
- Nas técnicas de interdependência nenhuma variável é considerada dependente ou independente, mas todas as variáveis são analisadas simultaneamente com a finalidade de encontrar uma estrutura para todo conjunto de variáveis. São exemplos a análise fatorial, a análise de cluster, análise de componentes principais e escalonamento multidimensional.

NESSE TRABALHO IREMOS UTILIZAR O PCA E O K-MEANS

PCA

- Em cenários empresariais, geralmente lidamos com conjuntos de dados complexos que possuem várias variáveis, e algumas dessas variáveis podem estar correlacionadas ou possuir multicolinearidade. Isso pode dificultar a interpretação dos dados e afetar a eficácia de algoritmos de aprendizado de máquina, como o K-Means, que exigem que as variáveis sejam independentes.

- Nesse caso usaremos o PCA para Redução de dimensionalidade do Conjuntos de dados. O PCA ajuda a reduzir a dimensionalidade dos dados, mantendo a maior parte da variância original. Isso simplifica o conjunto de dados, tornando-o mais gerenciável e fácil de interpretar.
- Ao reduzir a dimensionalidade dos dados com o PCA, os clusters identificados pelo algoritmo K-Means podem se tornar mais distintos e interpretáveis. Isso facilita a compreensão das características dos grupos e a formulação de insights acionáveis para tomada de decisões empresariais.

VARIÁVEIS DE INTERESSE

Out[ ]:

```
['QtdeEstoque',
 'VlMargemCont',
 'VlDesc',
 'QUANTIDADE',
 'VALOR UNITÁRIO',
 'ProdCusto',
 'ProdPrecoValor',
 'ValorLucroBruto',
 'VALOR DA VENDA']
```

ESCALONANDO OS DADOS

TREINANDO O MODELO

Out[ ]:

```
PCA
PCA()
```

VARIÂNCIA EXPLICADA DE CADA COMPONENTE

```
Variância explicada por cada componente principal:
[6.60255346e-01 2.23661860e-01 6.54274423e-02 2.65962960e-02
 2.02526886e-02 3.06189603e-03 6.94341232e-04 5.01297156e-05
 3.45485062e-31]
```

NÚMERO DE COMPONENTES PRINCIPAIS QUE EXPLICAM 95% DA VARIABILIDADE

Out[ ]:

```
3
```

	NOME DA PEÇA	QtdeEstoque	VlDesc	ProdPrecoValor	ValorLucroBruto	Componente_1	Componente_2	Componente_3
0	ABRAC.TUBOS BORRA	2	0.0	19.82	54.17	-1.030573	-0.099680	0.062690
1	ANEL	1	0.0	12.00	24.18	-1.093517	-0.324928	0.090907
2	ANEL DE VEDAÇÃO TANQUE	1	0.0	21.06	54.30	-1.022503	-0.325241	0.060396
3	EL. FILTRO COMBUST.	1	0.0	113.87	209.33	-0.630754	-0.327128	-0.075613
4	FILTRO	1	0.0	97.31	164.93	-0.696554	-0.326478	-0.058841

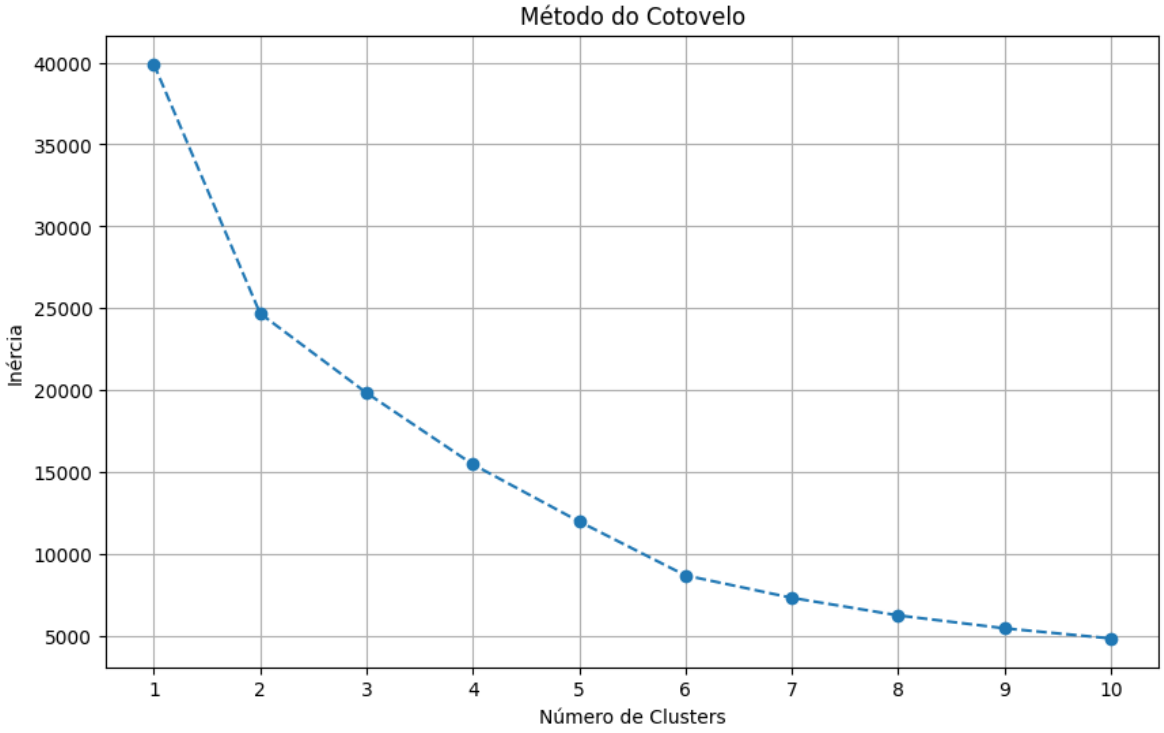
DADOS FINAIS

Out[ ]:

```
['NOME DA PEÇA',
 'QtdeEstoque',
 'VlDesc',
 'ProdPrecoValor',
 'ValorLucroBruto',
 'Componente_1',
 'Componente_2',
 'Componente_3']
```

GRÁFICO DE COTOVELO QUE NOS INDICA O NÚMERO DE CLUSTERS IDEAL PARA TRABALHAR

Nesse exemplo foram 2



K-MEANS

- O k-means é usado para agrupar dados em conjuntos distintos com base em suas características. O objetivo do algoritmo K-means é dividir um conjunto de pontos de dados em "K" grupos (clusters) diferentes, onde cada ponto de dados pertence a um cluster com base em sua proximidade com os outros pontos de dados no mesmo cluster.

- A utilização do modelo K-Means nos proporcionará uma visão abrangente da estrutura de preços e lucro bruto da empresa e nos ajudará a identificar oportunidades para uma gestão mais eficiente.

- Ao agrupar produtos com características semelhantes, será possivel implementar estratégias de gestão de marketing mais direcionadas e eficazes, reduzindo custos, melhorando a disponibilidade de produtos e aumentando a satisfação do cliente.

- O modelo K-Means agrupará os dados em clusters com base na similaridade entre os Produtos vendidos e suas preços. Ele calculará os centroides de cada cluster, que representam os "centros" dos grupos de itens semelhantes. O algoritmo K-Means tentará minimizar a variação intra-cluster e maximizar a variação inter-cluster, agrupando os itens com base nos seus preços e de lucro gerado por cada produto para a empresa de forma coesa e distintiva.

Out[ ]:

```
KMeans
KMeans(n_clusters=2, random_state=42)
```

DADOS SEGMENTADOS

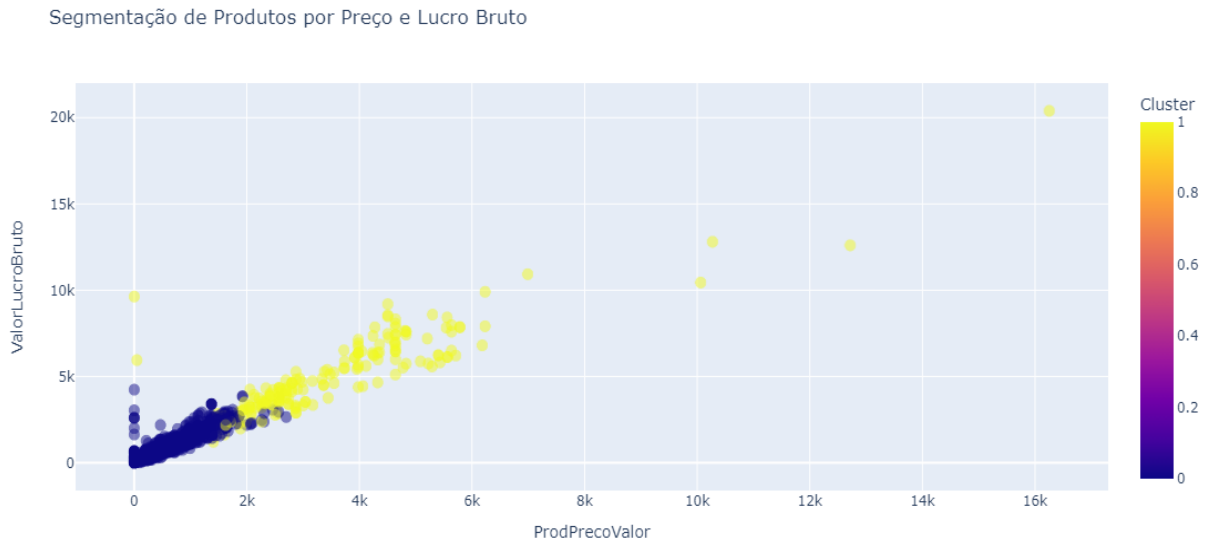
Temos aqui as variáveis indicadas pelo PCA como as influentes no noso conjunto de dados e usaremos elas para representar os clusters no qual definimos 2 clusters.

Out[ ]:

	NOME DA PEÇA	QtdeEstoque	VlDesc	ProdPrecoValor	ValorLucroBruto	Cluster
0	ABRAC.TUBOS BORRA	2	0.00	19.82	54.17	0
1	ANEL	1	0.00	12.00	24.18	0
2	ANEL DE VEDAÇÃO TANQUE	1	0.00	21.06	54.30	0
3	EL. FILTRO COMBUST.	1	0.00	113.87	209.33	0
4	FILTRO	1	0.00	97.31	164.93	0
...	...	...	...	...	...	...
9970	VENTILADOR	1	24.47	131.14	220.35	0
9971	SUPORTE DIR.CENT.COMPL	1	0.00	323.15	608.08	0
9972	SUPORTE INFERIOR DIREI	1	0.00	240.57	452.05	0
9973	CARTUCHO DO FILTRO	1	37.77	59.34	67.39	0
9974	OLEO MOTOR LD 7 MINERAL 15W40	18	113.69	196.74	305.21	0

9975 rows × 6 columns

O Objetivo aqui é mostrar como o preço do produto se relaciona com o lucro bruto e como isso difere entre os clusters.



MÉDIAS POR CLUSTERS

```
Cluster
0     197.042039
1    3526.291832
```

Os produtos pertencentes ao cluster 0 tem média de valor de aproximadamente 197,04 , já os produtos pertencentes ao cluster 1 tem média de valor de 3.526,29.

- O cluster **0** nos mostra produtos com um preço médio relativamente baixo e um lucro bruto correspondente. O que nos indica que esses produtos podem estar posicionados no mercado como opções mais acessíveis ou de entrada. Indicando uma estratégia de penetração de mercado, onde a empresa busca atrair clientes oferecendo produtos a preços competitivos, com margens de lucro mais modestas.

  Uma estratégia para esse grupo de produtos seria focar em volumes de vendas para compensar margens de lucro menores.

- Já o cluster **1** são produtos com um preço médio muito mais alto e, presumivelmente, margens de lucro bruto também mais elevadas. São produtos premium ou de luxo, destinados a um segmento de mercado mais exclusivo.

  Nesse grupo de produtos podemos ver uma estratégia de diferenciação , onde a empresa busca se destacar da concorrência oferecendo produtos de alta qualidade com margens de lucro bem maiores.

O foco nesse caso está mais na qualidade do que na quantidade, podemos nesse caso intensificar as estratégias de marketing e vendas para atrair consumidores dispostos a pagar um preço mais alto por valor percebido superior.