

A woman with long brown hair and glasses, wearing a denim jacket, is shown from the chest up. She is looking down at a laptop screen with a thoughtful expression, her hand near her chin. The background is blurred, showing an office environment.

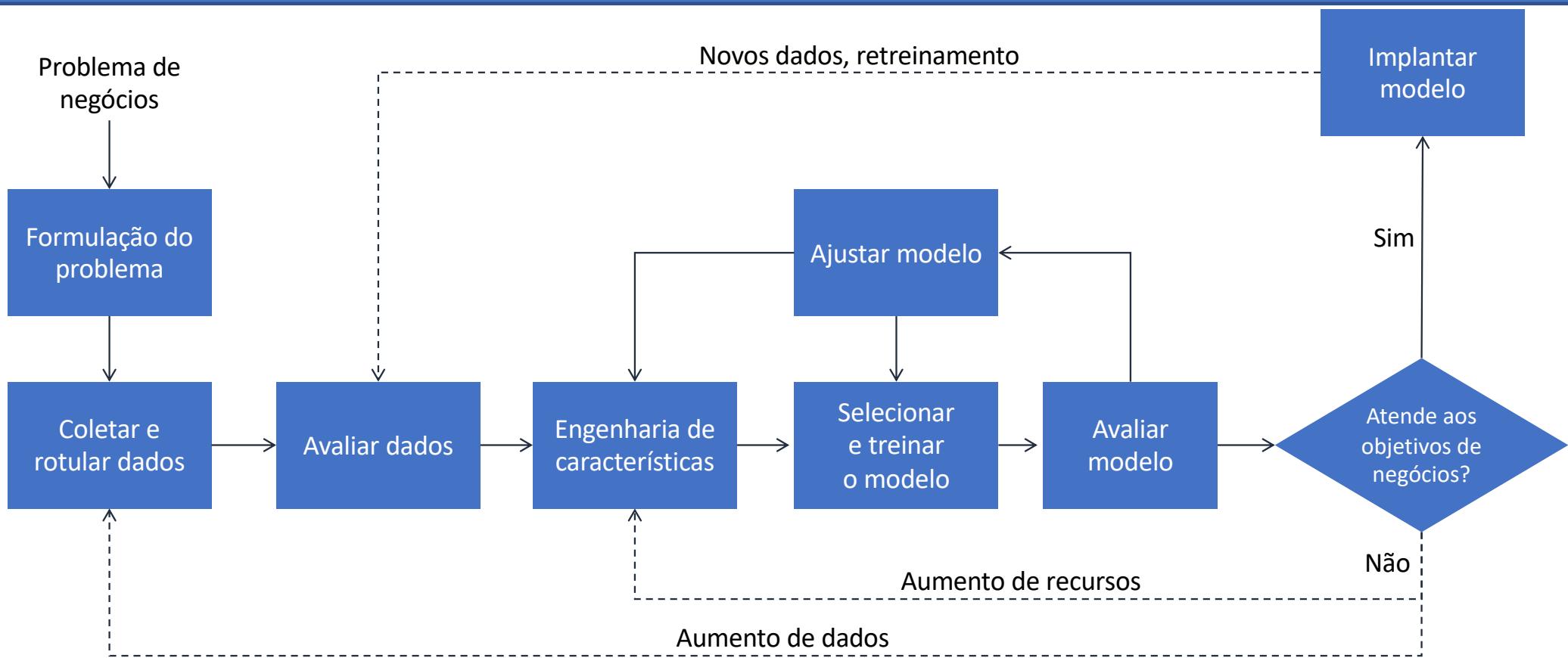
MBA IMPACTA

A black and white photograph of a classroom setting. Several students are seated at desks, facing forward. Some are looking at their laptops, while others have notebooks and pens open, appearing to take notes or work on assignments. The scene is lit from above, creating strong shadows and highlights on the students' faces and the desks.

LAB Coding AI

(2^a Aula)

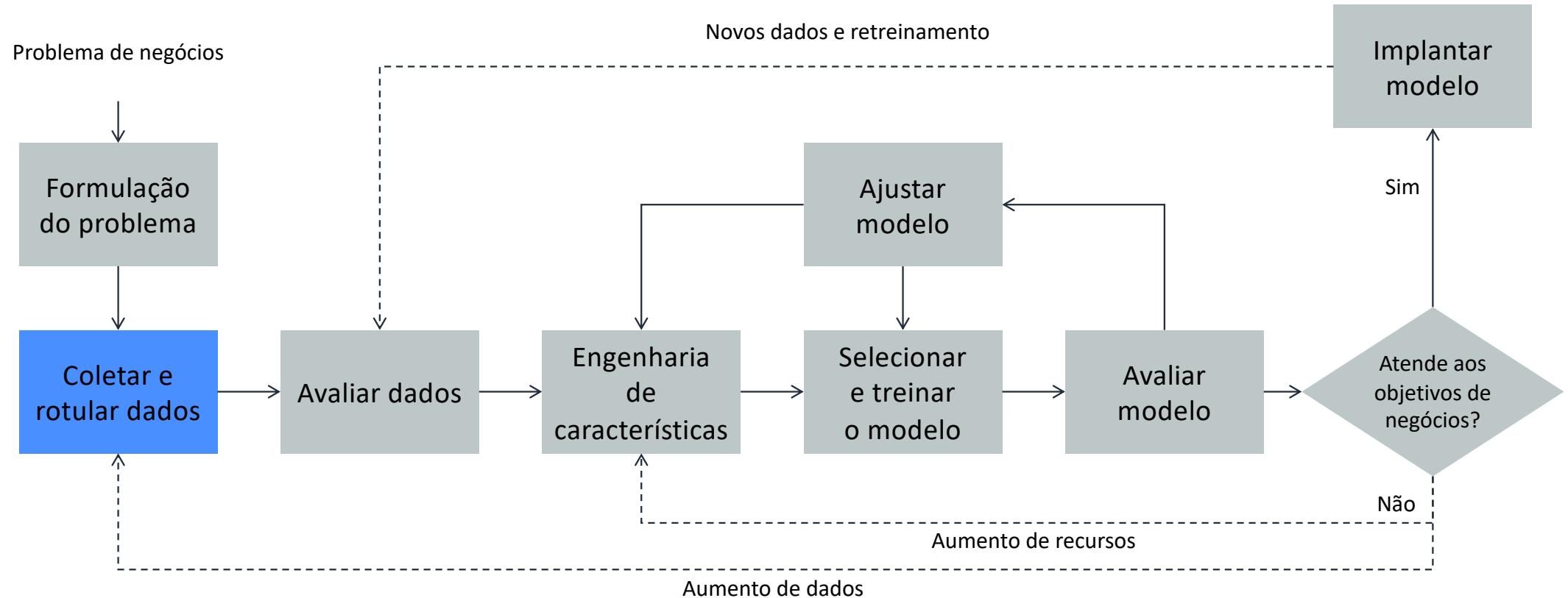
Pipeline de ML: implantação





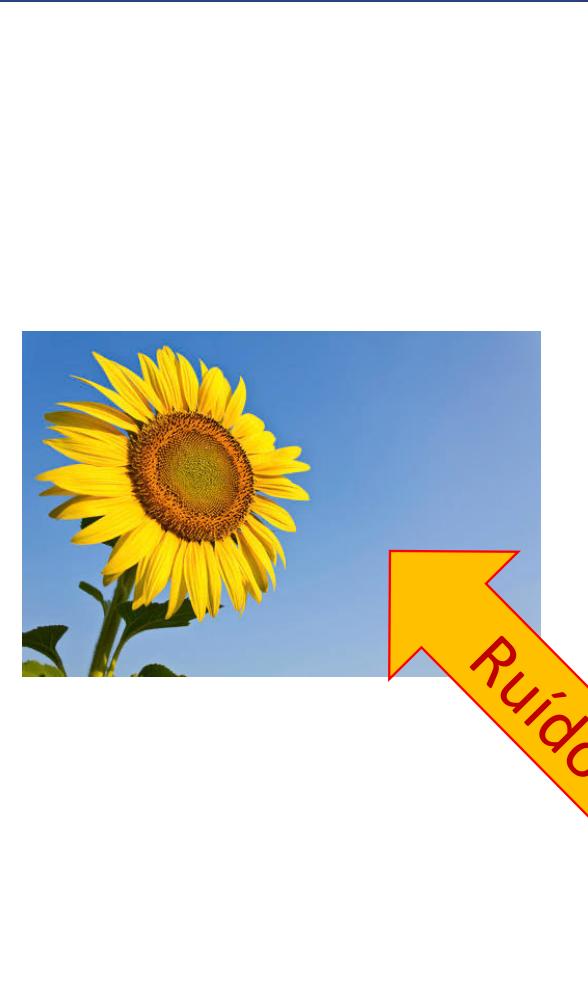
Coleta De Dados

Pipeline de machine learning



De quais dados você precisa?

- Quantos dados você tem e onde eles estão?
- Você tem acesso a esses dados?
- Qual solução você pode usar para trazer todos esses dados para um repositório centralizado?





girassol

with_leaves



girassol



girassol

Fontes de dados

- Dados privados: dados que os clientes criam
- Dados comerciais: Serasa Experian, SeReuters, Change Healthcare, Dun & Bradstreet etc.
- Dados de código aberto: dados que estão disponíveis publicamente (verifique se há limites de uso)
 - Kaggle (<https://www.kaggle.com/>)
 - Organização mundial da saúde
 - EUA Census Bureau
 - National Oceanic and Atmospheric Administration (EUA)
 - UC Irvine machine learning Repository
 - IBGE

Observações

Os problemas de ML precisam de uma grande quantidade de dados, também chamados de **observações** em que a resposta ou previsão de destino **já é conhecida**.

Cliente	Data da transação	Fornecedor	Valor da cobrança	Isso foi uma fraude?
ABC	10/5	Loja 1	10,99	Não
DEF	10/5	Loja 2	99,99	Sim
GHI	10/5	Loja 2	15,00	Não
JKL	10/6	Loja 2	99,99	?
MNO	10/6	Loja 1	99,99	Sim

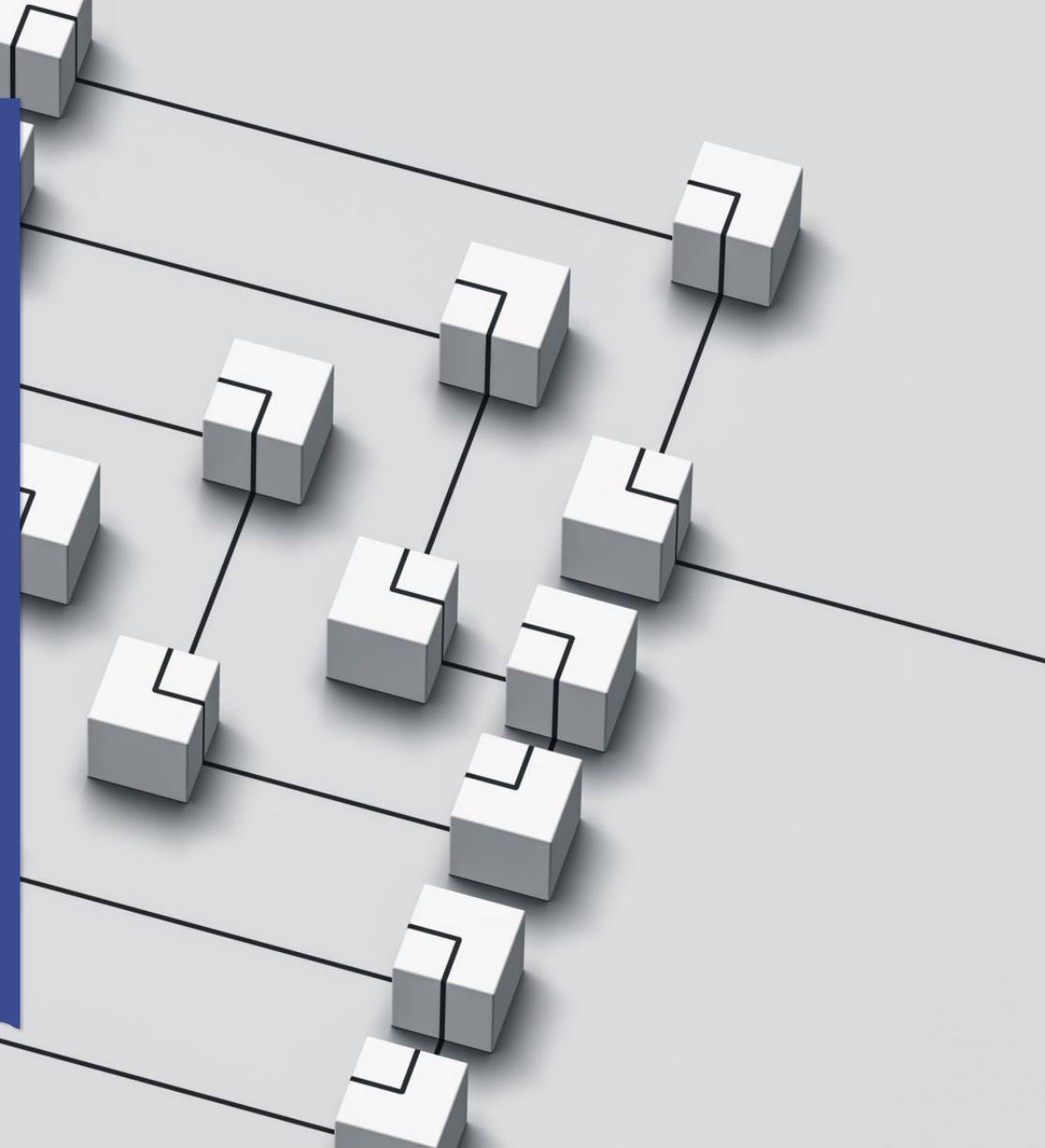
Observações

, que utilizam aprendizado supervisionado,

Os problemas de ML precisam de uma grande quantidade de dados, também chamados de **observações** em que a resposta ou previsão de destino **já é conhecida**.

Cliente	Data da transação	Fornecedor	Valor da cobrança	Isso foi uma fraude?	
ABC	10/5	Loja 1	10,99	Não	
DEF	10/5	Loja 2	99,99	Sim	
GHI	10/5	Loja 2	15,00	Não	
JKL	10/6	Loja 2	99,99	?	
MNO	10/6	Loja 1	99,99	Sim	

Normalmente, os dados
são distribuídos entre
vários sistemas e
provedores de dados
diferentes. O desafio é
reunir todas essas fontes
de dados em algo que um
modelo de machine
learning possa consumir.





Extract, Transform, Load (ETL-
Extrair, transformar e carregar)
é um termo comum para obter
dados para machine learning.

Extract, Transform, Load (ETL)

As etapas em um processo de extração, transformação e carregamento (ETL) são definidas da forma a seguir.

- **Extrair** - Extraia os dados das fontes para um único local.
- **Transformar** - Durante a extração, os dados podem precisar ser modificados, os registros correspondentes podem precisar ser combinados ou outras transformações podem ser necessárias.
- **Carregar** - Por fim, os dados são carregados em um repositório onde o processo criador dos modelos possam acessar.

ETL com Python

```
import boto3, requests, zipfile, os, io  
url = 'http://url.com/somezipfile.zip'  
folder='./extracts'
```

} Importações e variáveis

```
r = requests.get(url, stream=True)  
thezip = zipfile.ZipFile(io.BytesIO(r.content))  
thezip.extractall(folder)
```

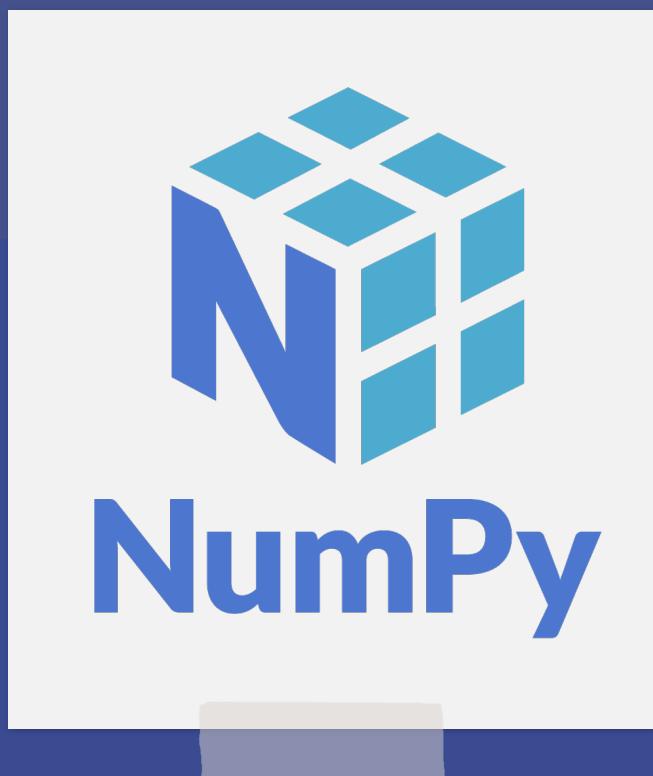
} Fazer download e extrair

```
s3 = boto3.client('s3')  
bucket = 'bucketname'  
with os.scandir(folder) as dir:  
    for f in dir:  
        if f.is_file():  
            s3.upload_file(  
                Filename=os.path.join(folder,f.name),  
                Key=f.name, Bucket=bucket)
```

} Fazer upload para
o Amazon S3



Segurança dos Dados



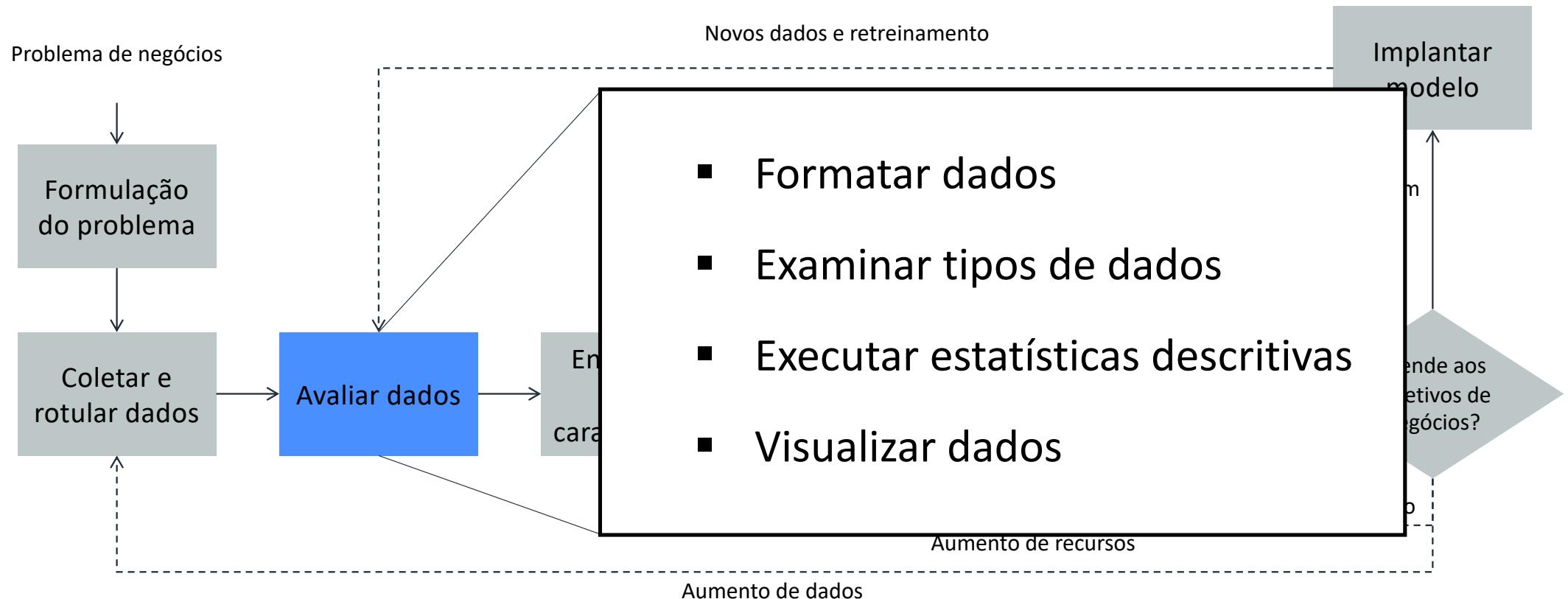
NumPy é o pacote fundamental para computação científica com Python.

Para computação intensiva de dados, o Numpy nos fornece uma ampla gama de métodos que tornam a manipulação de dados em Python muito rápida e fácil. Enquanto o Python é mais lento na execução em comparação com outras linguagens como Fortran durante o loop, o Numpy acelera as operações do Python convertendo o código repetitivo para a forma compilada.

Avaliação Dos Dados



Pipeline de machine learning





A primeira etapa na avaliação dos dados é garantir que eles estejam no formato correto para análise.

Você deve compreender seus dados



Cliente	Data da transação	Fornecedor	Valor da cobrança	Isso foi uma fraude?
ABC	10/5	Loja 1	10,99	Não
DEF	10/5	Loja 2	99,99	Sim
GHI	10/5	Loja 2	15,00	Não
JKL	10/6	Loja 2	99,99	?
MNO	10/6	Loja 1	99,99	Sim

Uma das bibliotecas Python de código aberto mais populares é a *pandas*. Ela pode reformatar dados de vários formatos em uma representação tabular de seus dados em linhas e colunas.



Carregamento de dados no pandas



- Reformata dados em representação tabular (DataFrame)
- Converte formatos comuns, como valores separados por vírgula (CSV), JavaScript Object Notation (JSON), Excel, Pickle e outros

```
import pandas as pd
url = "https://somewhere.com/winequality-red.csv"
df_vinho = pd.read_csv(url, ';')
```

Pandas DataFrame

Um DataFrame pode ser definido como uma
*"Estrutura tabular genérica rotulada em 2D
com coluna potencialmente heterogênea"*.

Pandas DataFrame

`df_wine.shape`

Número de instâncias

`df_wine.head(5)`

Número de atributos

Colunas/atributos

	acidez fixa	acidez volátil	ácido cítrico	açúcar residual	cloreto	dióxido de enxofre livre	dióxido de enxofre total	densidade	pH	sulfatos	álcool	qualidade
0	7,4	0,70	0,00	1,9	0,076	11,0	34,0	0,9978	3,51	0,56	9,4	5
1	7,8	0,88	0,00	2,6	0,098	25,0	67,0	0,9968	3,20	0,68	9,8	5
2	7,8	0,76	0,04	2,3	0,092	15,0	54,0	0,9970	3,26	0,65	9,8	5
3	11,2	0,28	0,56	1,9	0,075	17,0	60,0	0,9980	3,16	0,58	9,8	6
4	7,4	0,70	0,00	1,9	0,076	11,0	34,0	0,9978	3,51	0,56	9,4	5

Linhas/instâncias

Pandas DataFrame

Cada coluna em um DataFrame é uma série. Uma série é uma matriz rotulada unidimensional. Uma série pode armazenar dados de qualquer tipo.

varatos	álcool	qualidade
0,56	9,4	5
0,68	9,8	5
0,65	9,8	5
0,58	9,8	6
0,56	9,4	5

Nomes de índice e coluna

```
df_vinho.columns
```

```
Index(['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

```
df_vinho.index
```

```
RangeIndex(start=0, stop=1599, step=1)
```

Nomes de índice e coluna

Junto com os dados, você pode carregar um DataFrame com um *índice* (rótulos de linha) e *colunas* (rótulos de coluna). Se você carregar os dados de um CSV com uma linha de cabeçalho, as colunas serão criadas a partir da primeira linha do arquivo. No entanto, você pode alterar esse comportamento.

Esquema do DataFrame

Quando você executa a análise de dados, é importante ter certeza de que usa os tipos de dados corretos. Em muitos casos, a pandas infere corretamente os tipos de dados corretos ao carregar dados, e você pode continuar.

Se você tiver conhecimento ou acesso a um especialista no domínio de negócio, eles geralmente poderão detectar problemas de tipo de dados.

Esquema do DataFrame

`df_wine.dtypes`

```
qualidade      int64
acidez fixa    float64
acidez volátil  float64
ácido cítrico   float64
açúcar residual float64
cloreto         float64
dióxido de enxofre livre float64
dióxido de enxofre total float64
densidade       float64
pH              float64
sulfatos         float64
álcool           float64
dtype: object
```

`df_wine.info()`

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1597 entradas,
0 a 1598
Colunas de dados (total de 12 colunas):
qualidade          1597 int64 não nulo
acidez fixa        1597 float64 não nulo
acidez volátil     1597 float64 não nulo
ácido cítrico       1597 float64 não nulo
açúcar residual    1597 float64 não nulo
cloreto            1597 float64 não nulo
dióxido de enxofre livre 1597 float64 não nulo
dióxido de enxofre total 1597 float64 não nulo
densidade          1597 float64 não nulo
pH                 1597 float64 não nulo
sulfatos            1597 float64 não nulo
álcool              1597 float64 não nulo
dtypes: float64(11), int64(1)
uso de memória: 162,2 KB
```

Esquema do DataFrame

Se você não tiver os tipos de dados corretos, deverá descobrir o motivo. Muitas vezes, pode ser uma coluna numérica com dados ausentes ou um único valor de texto. Depois de analisar os dados, você pode convertê-los usando astype:

```
df_data['col'] = df_data['col'].astype('int')
```

No exemplo, você converte a coluna chamada **col** no tipo **int**, e substitui a coluna dentro do DataFrame **df_data**.

Use estatísticas descritivas para saber mais sobre o conjunto de dados

Estatísticas descritivas

- Use estatísticas descritivas para **obter insights sobre seus dados** antes de limpá-los.



Estatísticas gerais



Estatísticas
de atributo



Estatísticas
multivariadas

Estatísticas descritivas

As *estatísticas gerais* incluem o número de linhas (instâncias) e o número de colunas (recursos ou atributos) no seu conjunto de dados. Essas informações, relacionadas às *dimensões* de seus dados, são importantes. Por exemplo, ele pode indicar que você tem muitos recursos, o que pode levar à alta dimensionalidade e performance insatisfatória do modelo.

Estatísticas descritivas

As *estatísticas de atributos* são outro tipo de estatística descritiva, especificamente para *atributos numéricos*. Elas dão uma melhor ideia da forma de seus atributos, incluindo propriedades como média, desvio padrão, variação, valor mínimo e valor máximo.

Estatísticas descritivas

Estatísticas multivariadas analisam
relacionamentos entre mais de uma variável,
como correlações e relacionamentos entre seus
atributos.

Características estatísticas

`df_wine.describe()`

	acidez fixa	acidez volátil	ácido cítrico	açúcar residual	cloreto	dióxido de enxofre livre	dióxido de enxofre total	pH	sulfatos	álcool	qualidade
contagem	1599,00	1599,00	1599,00	1599,00	1599,00	1599,00	1599,00	1599,00	1599,00	1599,00	1599,00
média	8,32	0,53	0,27	2,54	0,09	15,87	46,47	3,31	0,66	10,42	5,64
desvio padrão	1,74	0,18	0,19	1,41	0,05	10,46	32,90	0,15	0,17	1,07	0,81
mín.	4,60	0,12	0,00	0,90	0,01	1,00	6,00	2,74	0,33	8,40	3,00
25%	7,10	0,39	0,09	1,90	0,07	7,00	22,00	3,21	0,55	9,50	5,00
50%	7,90	0,52	0,26	2,20	0,08	14,00	38,00	3,31	0,62	10,20	6,00
75%	9,20	0,64	0,42	2,60	0,09	21,00	62,00	3,40	0,73	11,10	6,00
máx.	15,90	1,58	1,00	15,50	0,61	72,00	289,00	4,01	2,00	14,90	8,00

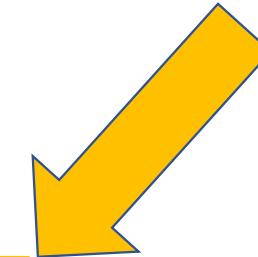
Estatísticas categóricas identificam a frequência de valores e desbalanceamentos de classe

`df_carros.head(5)`

	preço_compra	preço_manutenção	portas	pessoas	porta-malas	segurança	classe
0	m_alta	m_alta	2	2	pequeno	baixa	n_aceito
1	m_alta	m_alta	2	2	pequeno	med	n_aceito
2	m_alta	m_alta	2	2	pequeno	alta	n_aceito
3	m_alta	m_alta	2	2	med	baixa	n_aceito
4	m_alta	m_alta	2	2	med	med	n_aceito

`df_carros.describe()`

	preço_compra	preço_manutenção	portas	pessoas	porta-malas	segurança	classe
count	1728	1728	1728	1728	1728	1728	1728
unique	4	4	4	3	3	3	4
top	baixa	baixa	5more	more	grande	baixa	n_aceito
freq	432	432	432	576	576	576	1210

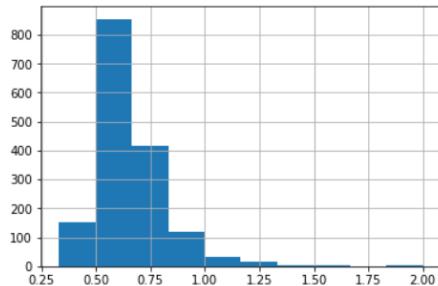


Crie visualizações
com Pandas para
examinar o conjunto
de dados em mais
detalhes

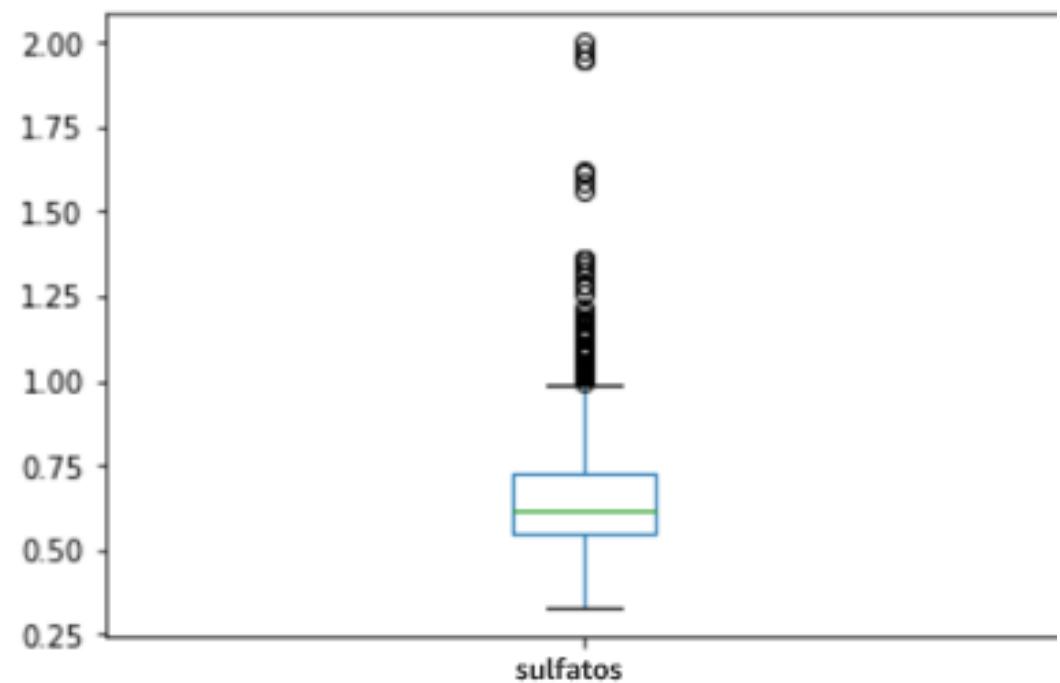


Plotagem de estatísticas de atributo

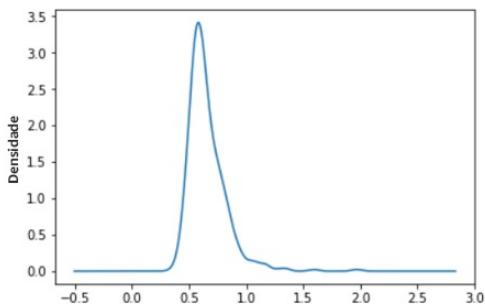
```
df_wine['sulphates'].hist(bins=10)
```



```
df_wine['sulphates'].plot.box()
```



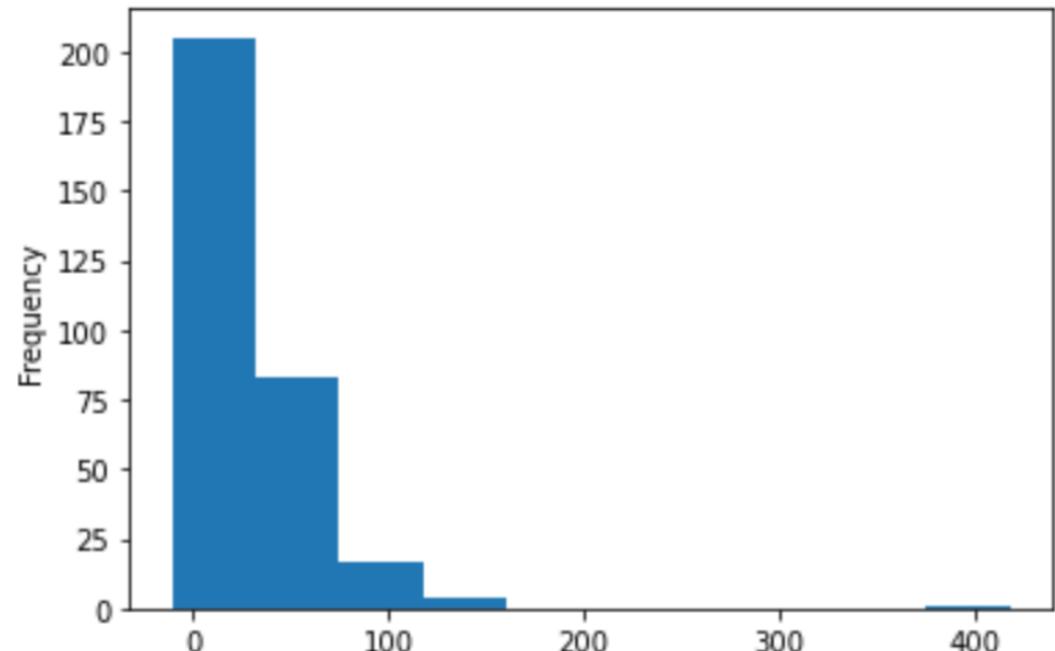
```
df_wine['sulphates'].plot.kde()
```



Histogramas

Um *histograma* geralmente é uma boa técnica de visualização para ver o comportamento geral de um recurso específico. Com um histograma, você pode responder perguntas, como: *os dados do recurso são normalmente distribuídos? Quantos picos existem nos dados? Esse recurso específico tem alguma distorção?*

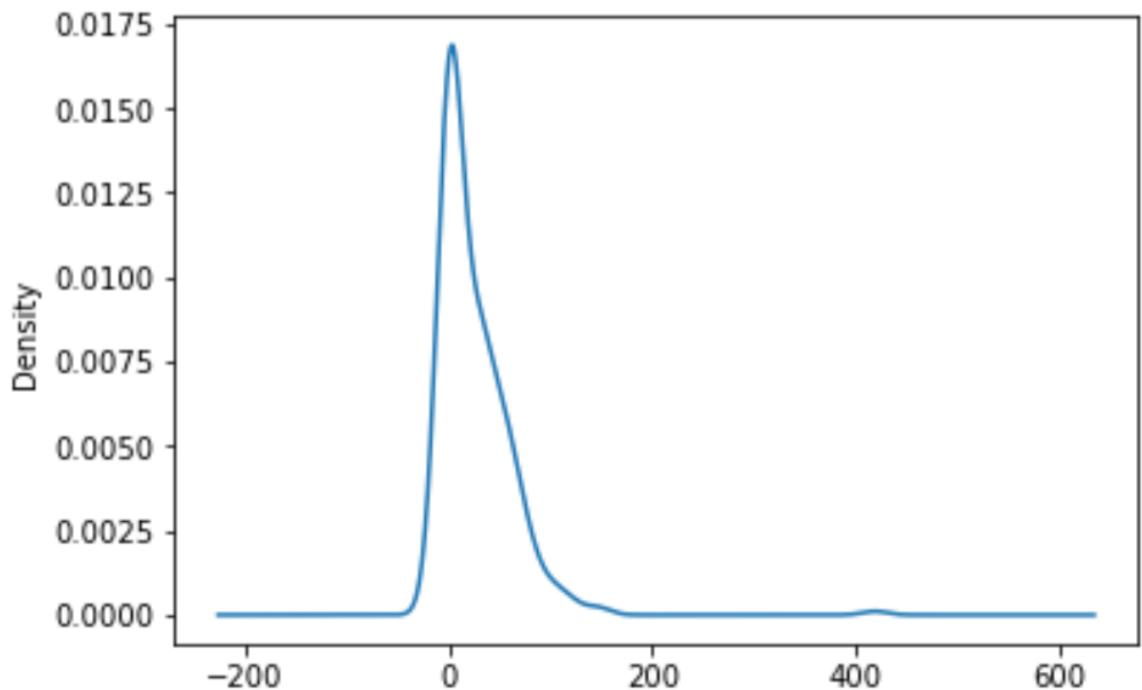
```
df['degree_spondylolisthesis'].plot.hist()
```



Gráficos de Densidade

Além dos histogramas, você pode usar *gráficos de densidade* e *gráficos de caixa* para recursos numéricos para ter uma ideia do que está dentro desse recurso específico.

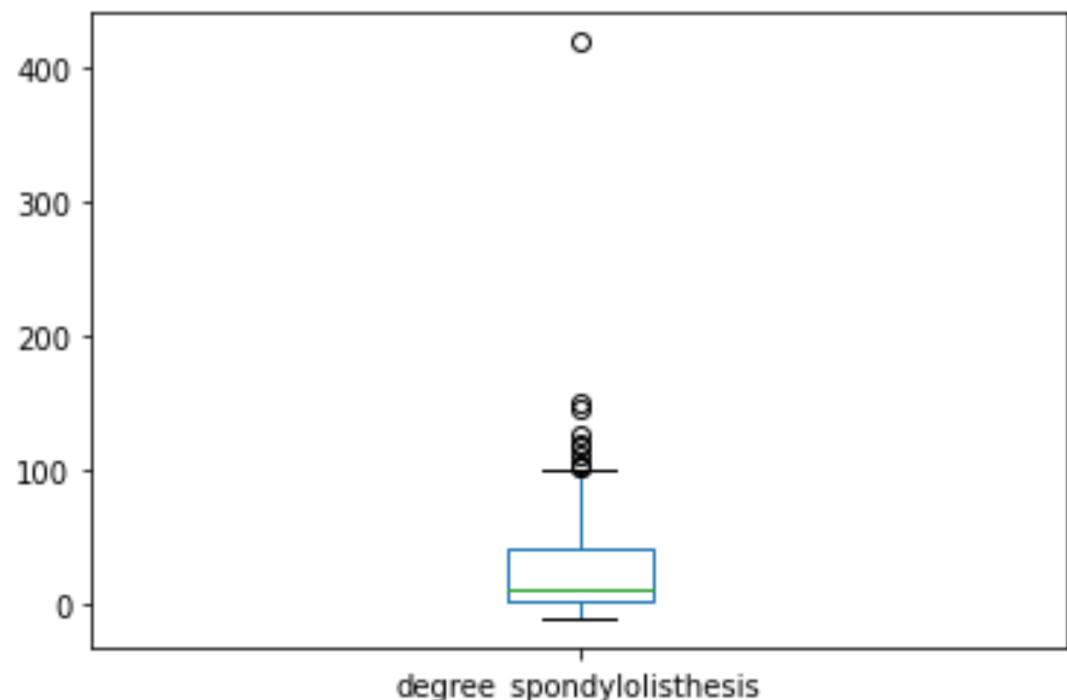
```
df['degree_spondylolisthesis'].plot.kde()
```



Gráficos de Caixa

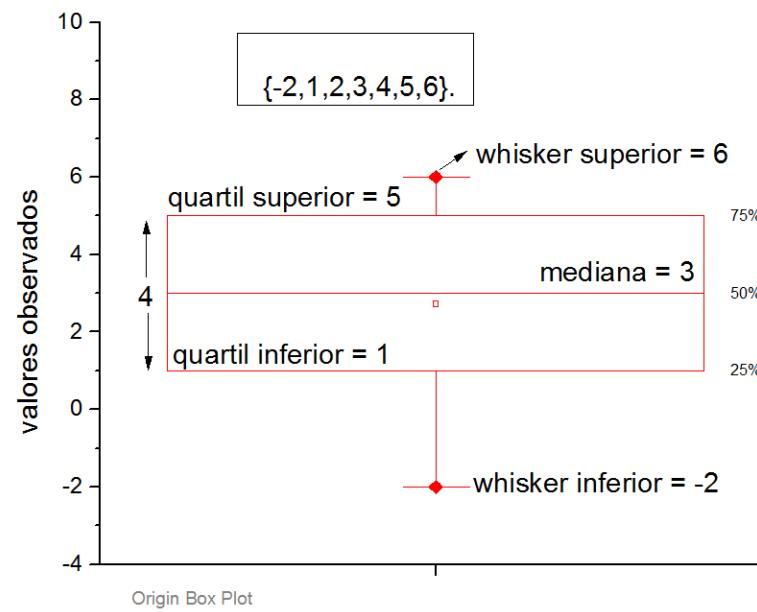
Um gráfico de caixa é um método para descrever graficamente grupos de dados numéricos por meio de seus quartis. A caixa se estende dos valores de quartil Q1 a Q3 dos dados, com uma linha na mediana (Q2). Os valores fora desse intervalo são representados pelas linhas que se estendem a partir da caixa. Essas linhas às vezes são chamadas de *whiskers*

```
df['degree_spondylolisthesis'].plot.box()
```



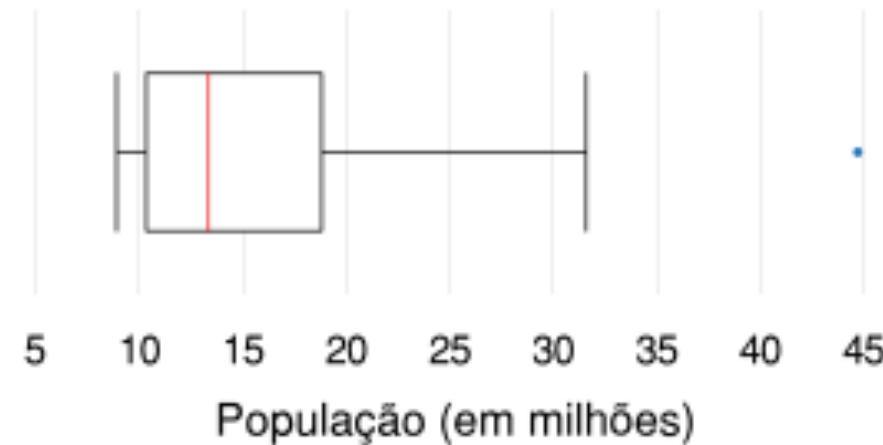
Gráficos de Caixa

Os whiskers se estendem a partir das bordas da caixa para mostrar o intervalo dos dados. A posição dos whiskers é definida por padrão como $1,5 * \text{IQR}$ ($\text{IQR} = \text{Q3} - \text{Q1}$) nas bordas da caixa. Pontos de discrepância são aqueles pontos que estão além do final dos whiskers.



Gráficos de Caixa

O box plot acima mostra a importância do cuidado com os *outliers* em análise de dados. A população de São Paulo é maior que a população dos demais estados brasileiros e isso não é um erro. Isto significa que nem sempre o *outlier* corresponde a um erro de arredondamento ou a um erro de observação.

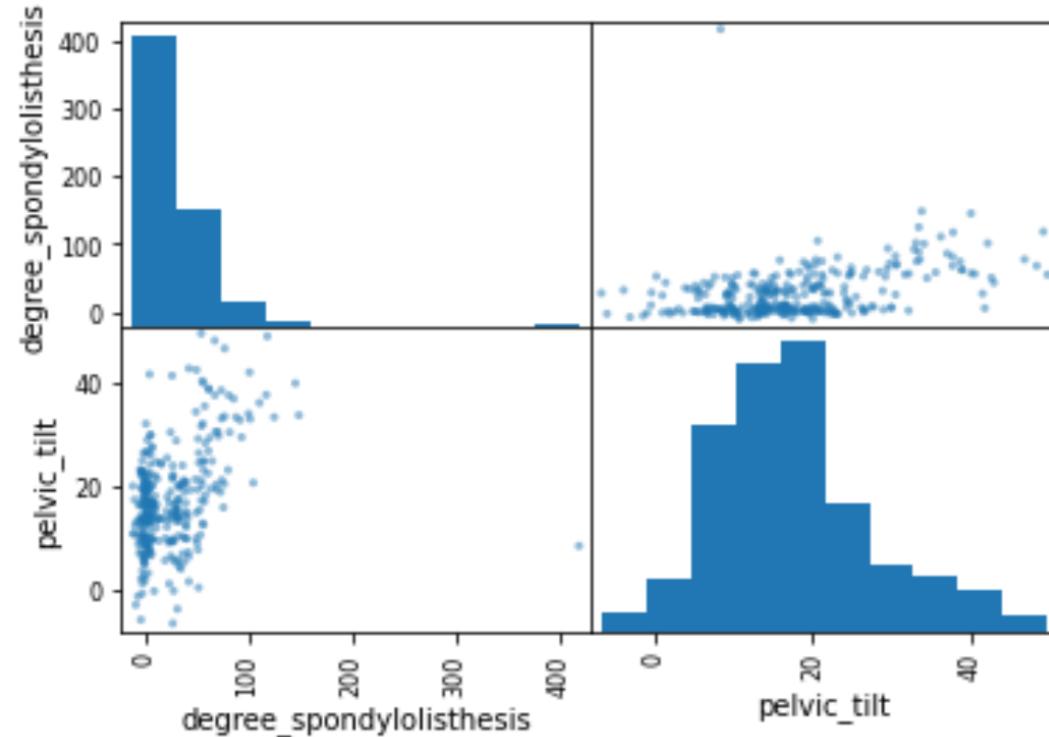


Plotagem de estatísticas multivariadas

Quando você tem mais de duas variáveis numéricas em um conjunto de dados de recurso, é recomendável observar seu relacionamento.

Um **gráfico de dispersão** é uma boa maneira de identificar quaisquer relacionamentos especiais entre essas variáveis.

```
pd.plotting.scatter_matrix(  
    df[['degree_spondylolisthesis',  
        'pelvic_tilt']])
```



Plotagem de estatísticas multivariadas

```
pd.plotting.scatter_matrix(df, figsize=(12,12))  
plt.show()
```

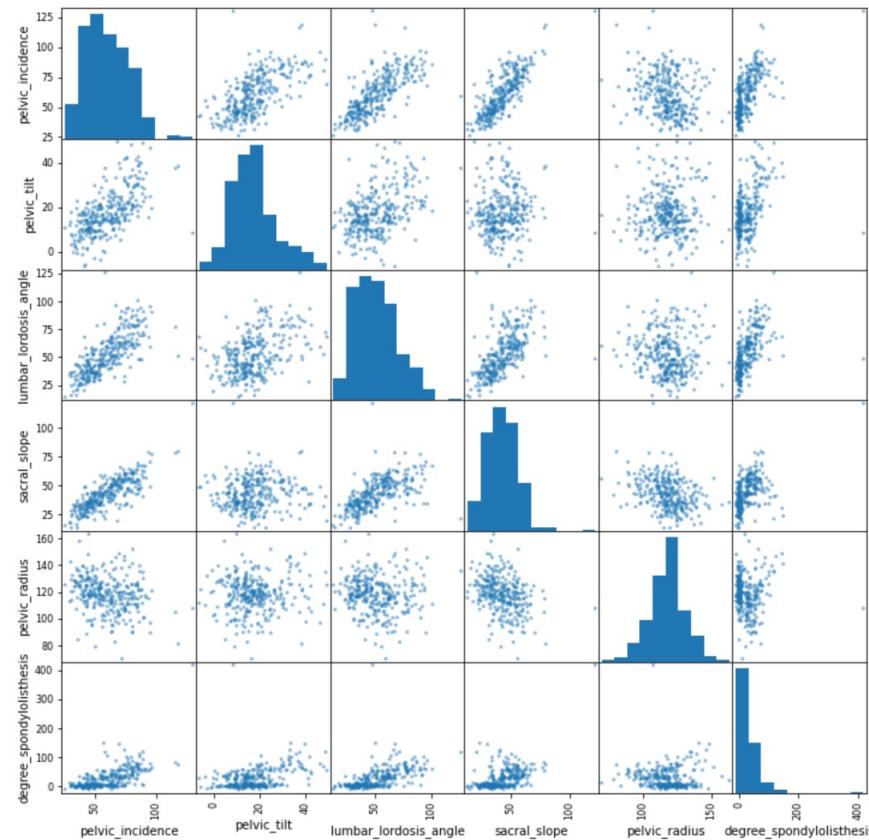
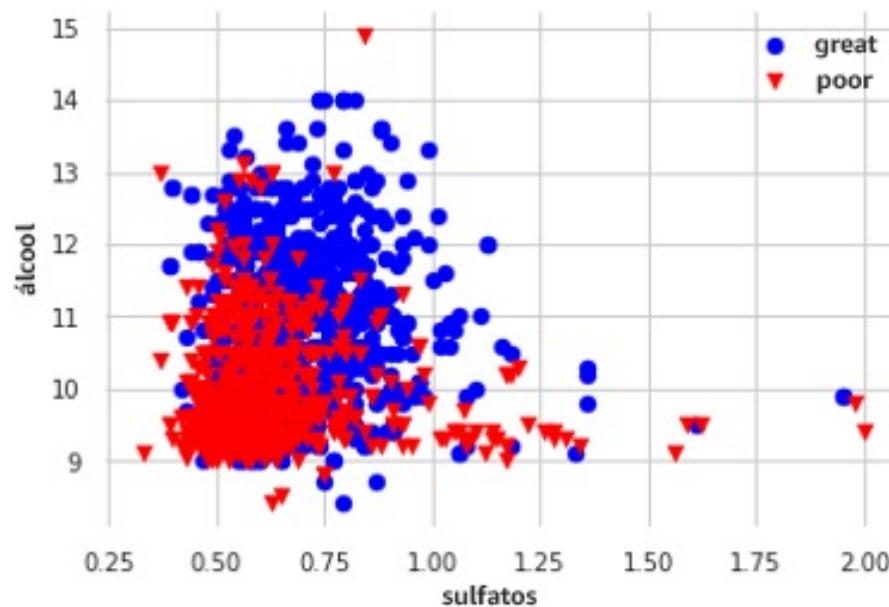


Gráfico de dispersão com identificação

```
high = df_wine[['sulphates','alcohol']][df_wine['quality']>5]
low = df_wine[['sulphates','alcohol']][df_wine['quality']<=5]

plt.scatter(high['sulphates'],high['alcohol'],s=50,c='blue',marker='o',label='great')
plt.scatter(x=low['sulphates'],y=low['alcohol'],s=50,c='red',marker='v',label='poor')
```



Matriz de correlação

Uma *matriz de correlação* é uma boa ferramenta nessa situação, pois ela transmite as relações lineares fortes e fracas entre variáveis numéricas.

Matriz de correlação

A correlação pode ser tão alta quanto um ou tão baixa quanto menos um. Quando a correlação é um, isso significa que esses dois recursos numéricos estão perfeitamente correlacionados entre si. É como dizer que Y é *proporcional a* X .

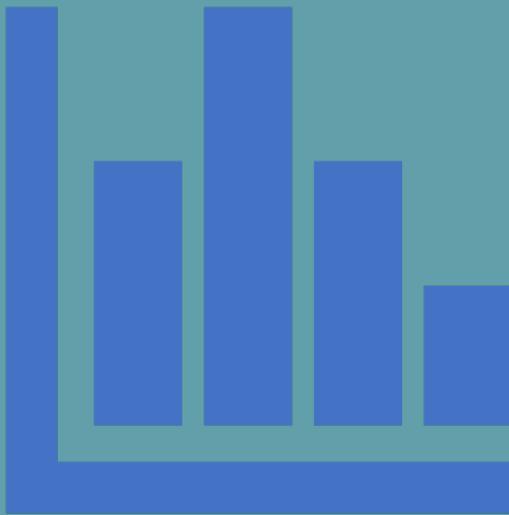
Quando a correlação dessas duas variáveis é menos um, é como dizer que Y é *proporcional a menos* X . Qualquer relação linear entre elas pode ser quantificada usando a correlação. Se a correlação for zero, isso significa que não existe nenhuma relação linear, mas *não* significa que nenhuma relação existe. É apenas uma indicação de que as duas variáveis não têm relação linear.

Mapa de calor da matriz de correlação

No entanto, examinar um número nem sempre é simples. Muitas vezes, é mais fácil visualizar os números se eles estiverem representados por cores. Agora, considere um *mapa de calor*. Você tem o número mais alto (que é 1) em verde escuro e -1 em marrom escuro. A cor fornece as direções positiva e negativa e a intensidade das correlações.



Lembrando a Estatística



Média e *Mediana* são duas medidas diferentes que descrevem a extensão em que seus dados são agrupados em torno de algum valor ou posição.

Média pode ser um método útil para entender seus dados quando os dados são simétricos.

No entanto, seus dados podem estar distorcidos ou conter discrepâncias. Nesse caso, a mediana tende a fornecer a melhor métrica para entender seus dados em relação à tendência central. Por exemplo, seus dados podem conter discrepâncias com valores grandes. Grandes discrepâncias podem distorcer a média de uma maneira para que ela não sirva como uma representação precisa de onde seus valores estão realmente centralizados. As discrepâncias não afetam a mediana da mesma maneira.

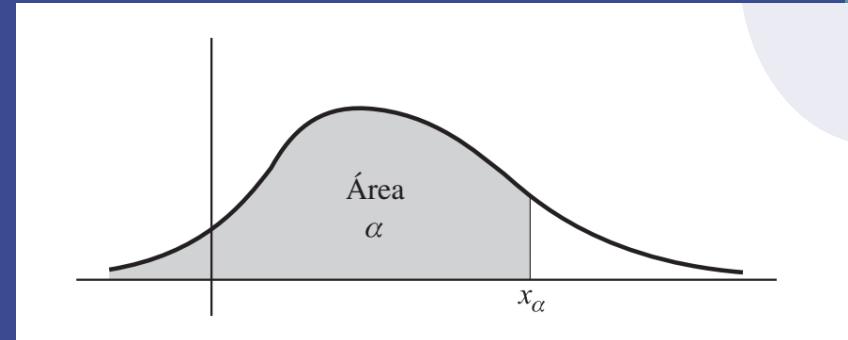
Desvio Padrão

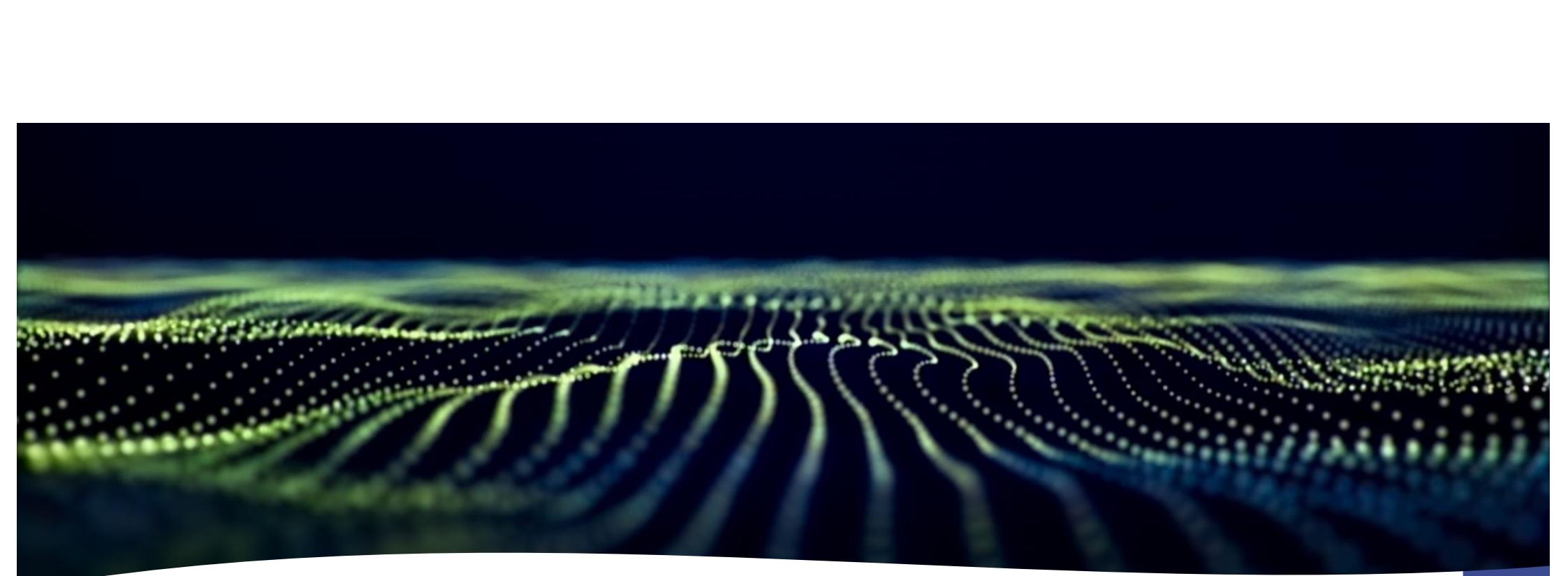
O desvio padrão é uma medida que expressa o grau de dispersão de um conjunto de dados. Ou seja, o desvio padrão indica o quanto um conjunto de dados é uniforme. Quanto mais próximo de 0 for o desvio padrão, mais homogêneo são os dados.

Percentis

Geralmente é conveniente subdividir a área sob uma curva de densidade pelo uso de ordenadas tal que a área à esquerda da ordenada é um percentual da área total. Os valores correspondentes a tais áreas são chamados de valores percentis, ou de forma abreviada percentis.

A área à esquerda de $x_{0,10}$ seria 0,10 ou 10% e $x_{0,10}$ seria chamado de o 10º percentil (também chamado de o primeiro decil). A mediana seria o 50º percentil (ou o quinto decil).



The background of the slide features a dark blue header and a white rectangular content area. The content area is framed by two vertical blue bars on the left and right sides. Inside this frame, there is a large, abstract digital wave pattern composed of numerous small, glowing green and yellow dots arranged in a wavy, undulating form.

Gerenciar
~~Excluir~~ os recursos
criados nesta aula.

“That’s all folks!”

