



# Python Data Analysis Mastery

Prof. Nicksson Freitas



# Aula 04. Técnicas para preparação de dados

Prof. Nicksson Freitas

# Objetivos

- Conhecer a importância da preparação de dados em um projeto de ciência de dados
- Conhecer as tarefas para processamento, transformação e limpeza de dados

## Resumo

1. Introdução ao processamento de dados
2. Tarefas do processamento de dados

# 1. Introdução ao processamento de dados

# Por que precisamos processar os dados?

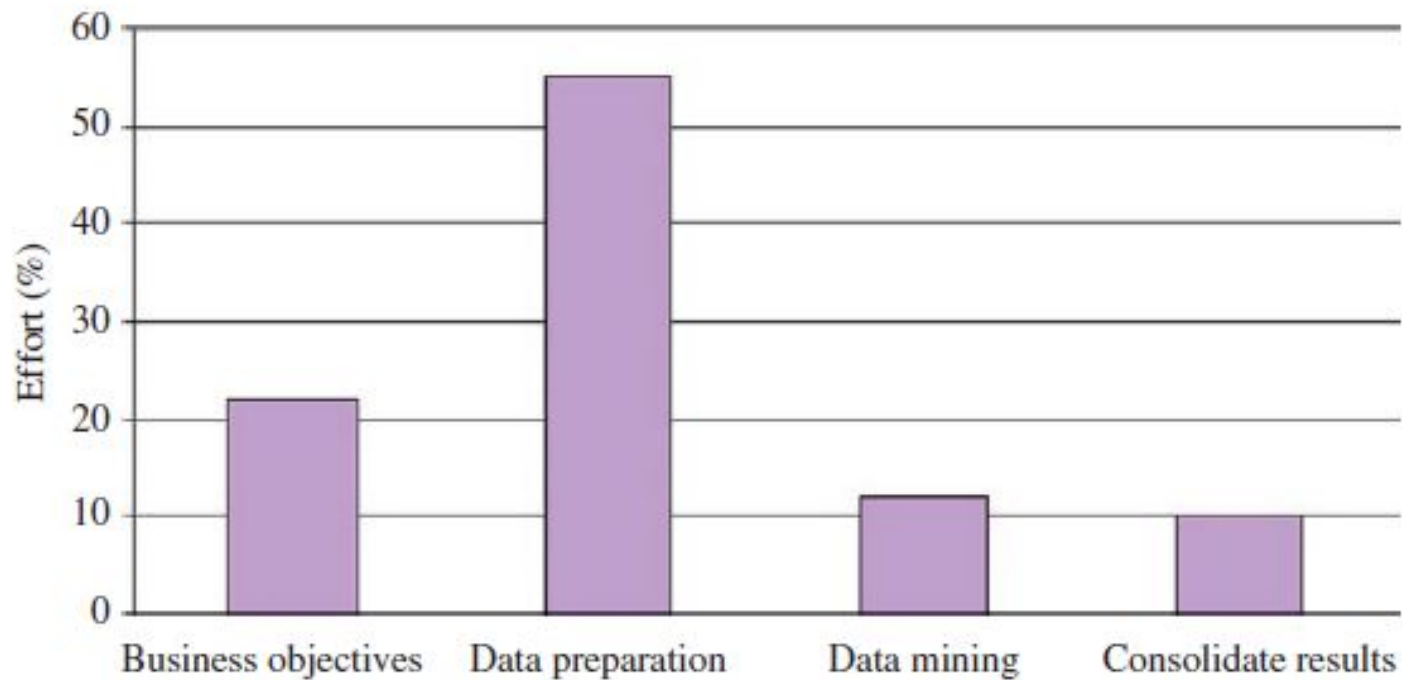
- Campos são obsoletos ou redundantes
- Existem valores faltantes
- Existem outliers
- Dados em formatos inadequado ou incompreensíveis
- Valores não condizem com a política ou bom senso

## Esforço da preparação dos dados

**Tabela 2.1** Cadastro de pessoas interessadas em obter um financiamento imobiliário

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal (\$)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

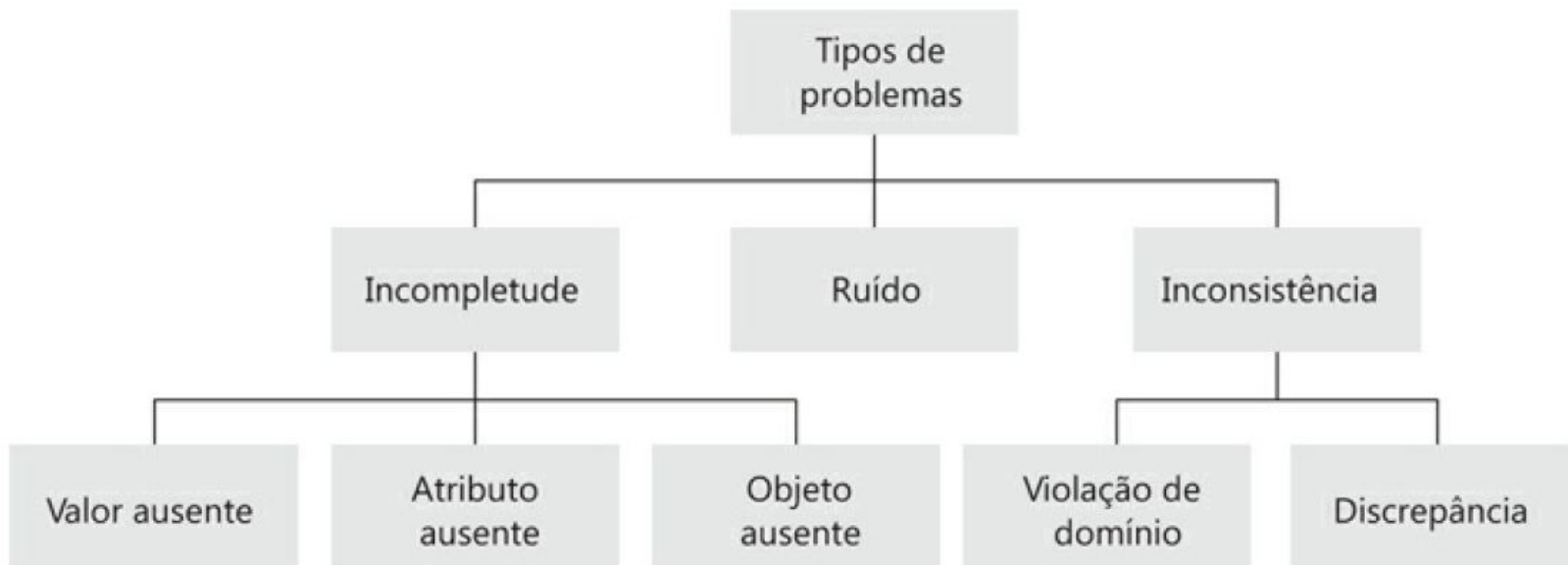
## Esforço da preparação dos dados





# Esforço da preparação dos dados

**Figura 2.1** Principais problemas com os dados

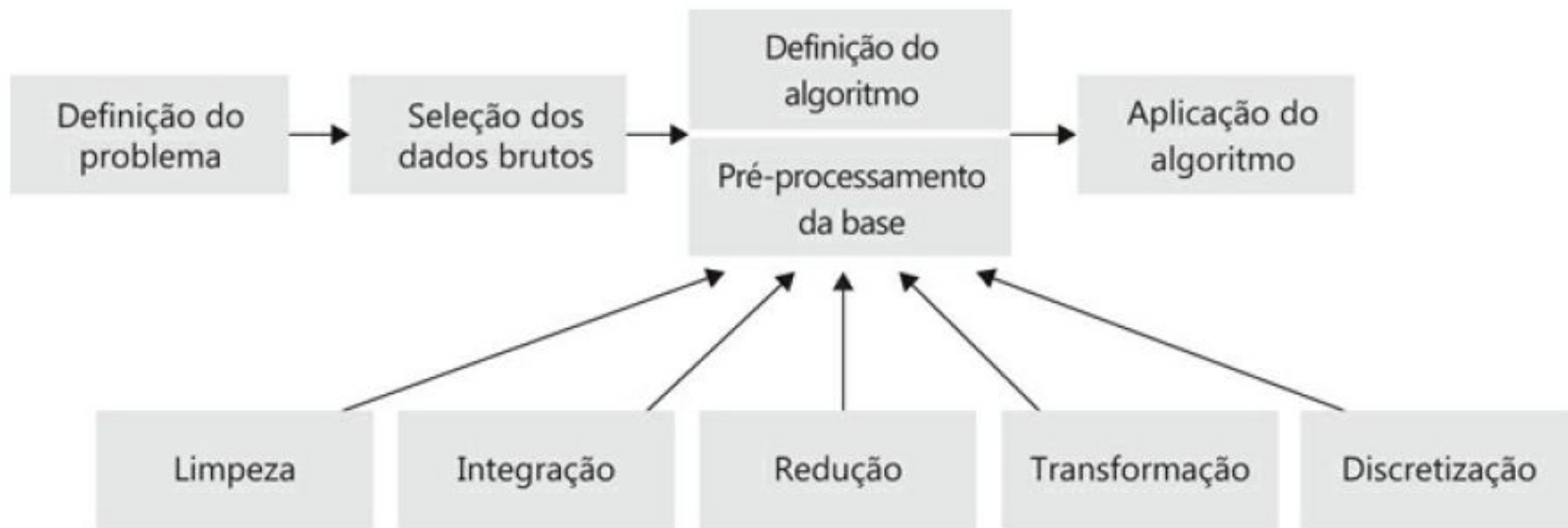


# Como avaliar a qualidade dos dados?

- Precisão (accuracy)
- Completude (completeness)
- Consistência (consistency)
- Pontualidade (timeliness)
- Credibilidade (believability)
  - Os dados são verdadeiros para o usuário?
- Interpretabilidade (interpretability)
  - Os dados são compreendidos facilmente?

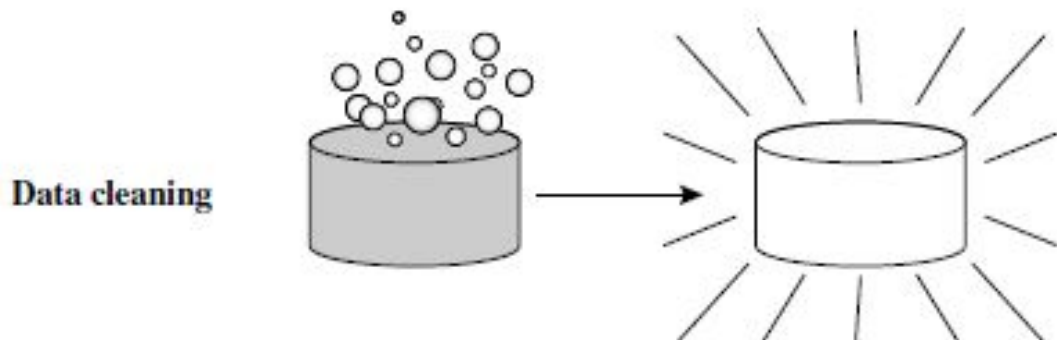
## 2. Tarefas de pré-processamento de dados

# Tarefas de pré-processamento de dados



# Limpeza de Dados

- Tem o objetivo de tratar valores ausentes suavizar ruídos, identificar outliers e corrigir inconsistências



# Limpeza de Dados

## Limpeza de Dados

ID	BI-RADS	Idade	Forma	Contorno	Densidade	Severidade
1	5	67	Lobular	Especulada	Baixa	Maligno
2	4	43	Redonda	Circunscrita	?	Maligno
3	5	58	Irregular	Especulada	Baixa	Maligno
4	4	28	Redonda	Circunscrita	Baixa	Benigno
5	5	74	Redonda	Especulada	?	Maligno
6	4	65	Redonda	?	Baixa	Benigno
7	4	70	?	?	Baixa	Benigno
8	5	42	Redonda	?	Baixa	Benigno
9	5	57	Redonda	Especulada	Baixa	Maligno
10	5	60	?	Especulada	Alta	Maligno

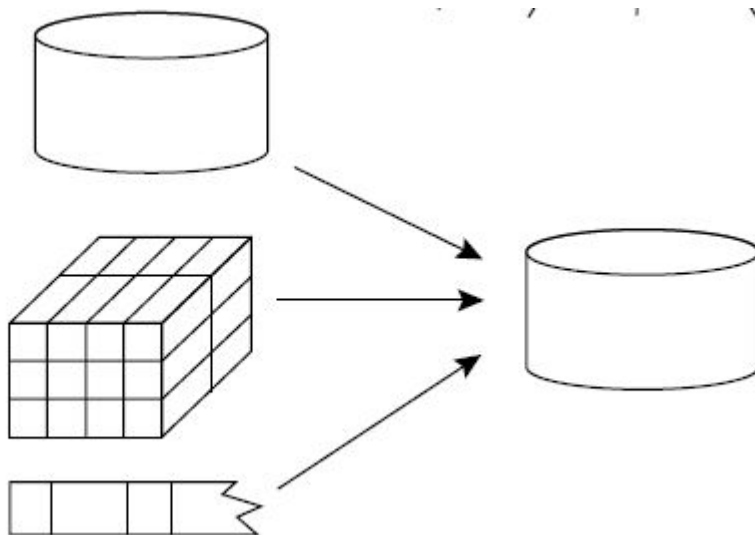


ID	BI-RADS	Idade	Forma	Contorno	Densidade	Severidade
1	5	67	Lobular	Especulada	Baixa	Maligno
2	4	43	Redonda	Circunscrita	Baixa	Maligno
3	5	58	Irregular	Especulada	Baixa	Maligno
4	4	28	Redonda	Circunscrita	Baixa	Benigno
5	5	74	Redonda	Especulada	Baixa	Maligno
6	4	65	Redonda	Circunscrita	Baixa	Benigno
7	4	70	Redonda	Circunscrita	Baixa	Benigno
8	5	42	Redonda	Circunscrita	Baixa	Benigno
9	5	57	Redonda	Especulada	Baixa	Maligno
10	5	60	Irregular	Especulada	Alta	Maligno

# Integração de Dados

- Enriquecer a base de dados com múltiplos recursos e tratar redundância, duplicidade e conflitos

**Data integration**



# Integração de Dados

Vendas		
ID	Produto	Quantidade
1	1	2
2	2	1
3	1	3
4	3	2
5	3	2
...	...	...

Produtos		
ID	Nome	Valor
1	Mouse	50,00
2	Teclado	100,00
3	Monitor	800,00

JOIN DAS TABELAS

Vendas JOIN Produtos					
ID	Produto	Quantidade	Nome	Valor	Total da Venda (Cálculo)
1	1	2	Mouse	50,00	100,00
2	2	1	Teclado	100,00	100,00
3	1	3	Mouse	50,00	150,00
...	...	...	...	...	...

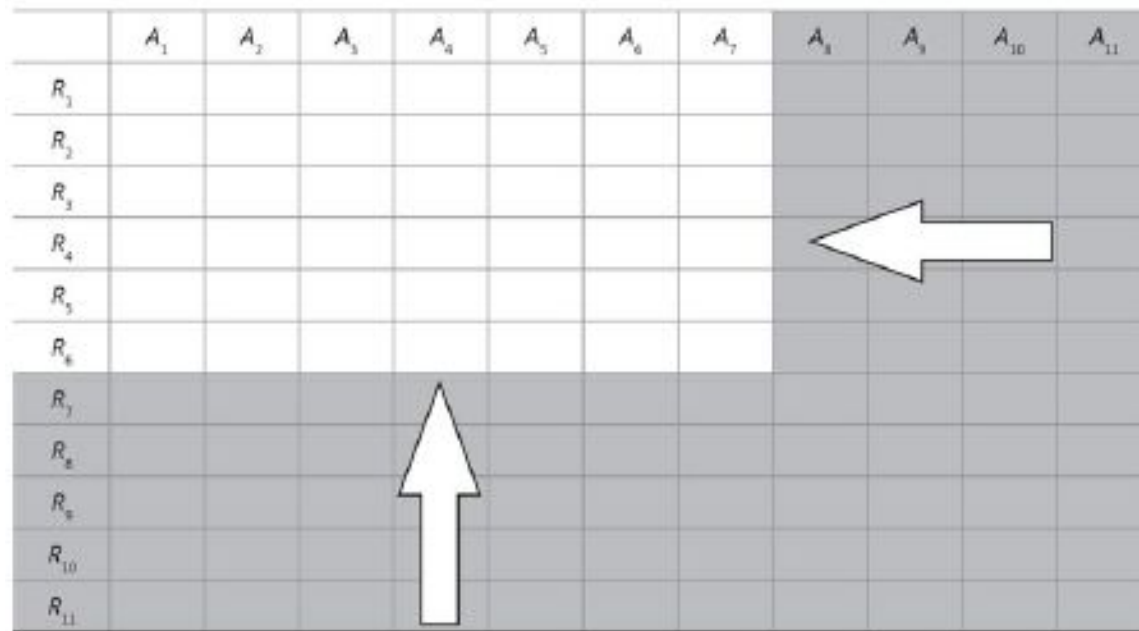


# Redução de Dados

- Métodos para redução de dados
  - Seleção de dados
  - Compressão de atributos

# Redução de Dados - seleção

	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$	$A_6$	$A_7$	$A_8$	$A_9$	$A_{10}$	$A_{11}$
$R_1$											
$R_2$											
$R_3$											
$R_4$											
$R_5$											
$R_6$											
$R_7$											
$R_8$											
$R_9$											
$R_{10}$											
$R_{11}$											



# Redução de Dados - compressão de atributos

- As técnicas aplicam uma codificação ou transformação para que uma representação compacta dos dados ou atributos seja obtida
- Métodos para redução de dimensionalidade
  - Análise de Componentes Principais (PCA)
  - Kernel PCA
  - Locally-Linear Embedding (LLE)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
  - Single Value decomposition (SVD)

# Transformação de dados

- Dados são poucos padronizados
  - Dados sem padronização (ex: nomes em maiusculo ou minusculo)
  - Dados em escalas diferentes
- Métodos para transformação de dados
  - Padronização
  - Normalização
  - Discretização de Dados

# Transformação de dados - padronização

- Tem o objetivo de resolver as diferenças de unidades e escala dos dados
- Métodos
  - Capitalização (padronizar dados maiúsculas e minúsculas)
  - Caracteres especiais (normalizar acentuação e caracteres especiais)
  - Padronização de Formatos (definir o formato adequado DD/MM/AAAA)
  - Conversão de unidade (Velocidade em km/h para m/s)

# Transformação de dados - normalização

- O processo de transformar dados que objetiva torná-los mais apropriados à aplicação de algoritmos de mineração
  - Ex: redes neurais ou métodos baseado em distância

**Tabela 2.11** Amostra da base de dados Mamo com valores normalizados

ID	BI-RADS	Idade	Forma	Contorno	Densidade	Severidade
1	0,83	0,63	0,67	1,00	0,67	1,00
2	0,67	0,23	0,00	0,00	0,67	1,00
3	0,83	0,63	1,00	1,00	0,67	1,00
4	0,67	0,00	0,00	0,00	0,67	0,00
5	0,83	0,81	0,00	1,00	0,67	1,00
6	0,67	0,63	0,00	0,00	0,67	0,00
7	0,67	0,81	0,00	0,00	0,67	0,00
8	0,83	0,23	0,00	0,00	0,67	0,00
9	0,83	0,63	0,00	1,00	0,67	1,00
10	0,83	0,63	1,00	1,00	0,00	1,00

# Transformação de dados - normalização

- Métodos
  - Normalização Max-Min
  - Normalização pelo escore-Z
  - Normalização pelo escalonamento deciona
  - Normalização pelo range interquartil

# Transformação de dados - discretização

- Alguns algoritmos de mineração de dados operam apenas com atributos categóricos
- Métodos de discretização
  - Binning
  - Análise de histograma
  - Agrupamento e discretização baseada em entropia



# Transformação de dados - discretização

ID	Idade
1	30
2	35
3	50
4	76

Discretização  
de dados



ID	Intervalo de idade
1	[18, 31]
2	[31, 44]
3	[44, 57]
4	[70, 83]

# Considerações Finais

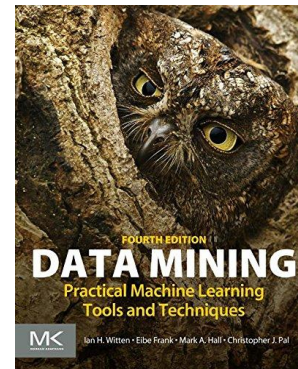
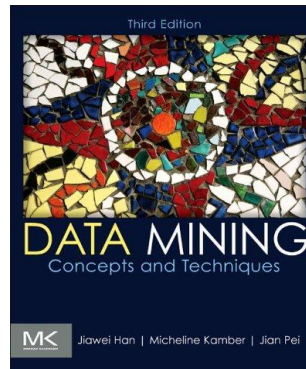
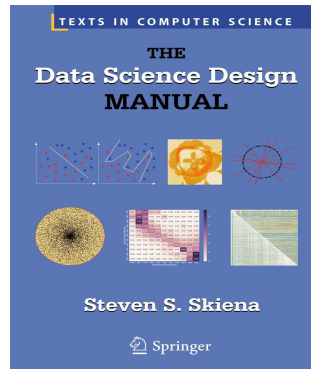
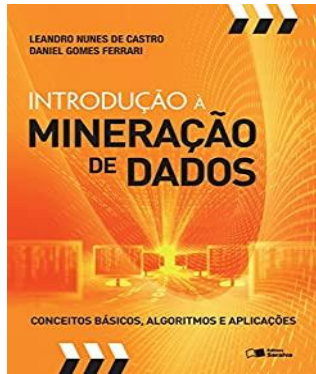
# Considerações finais

- Você viu a importância da etapa de processamento em um projeto de Ciência de Dados
- Você conheceu diversas tarefas de processamento, transformação e limpeza de dados

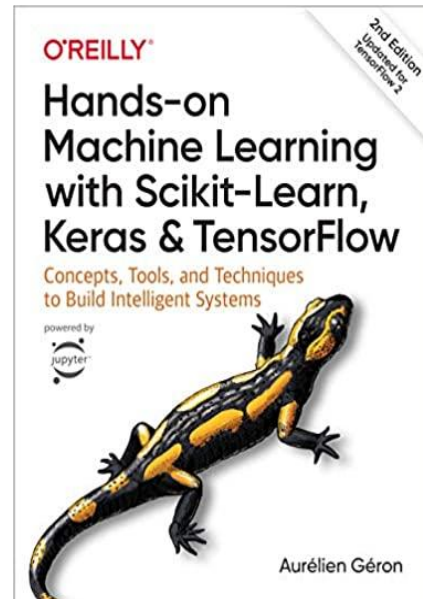
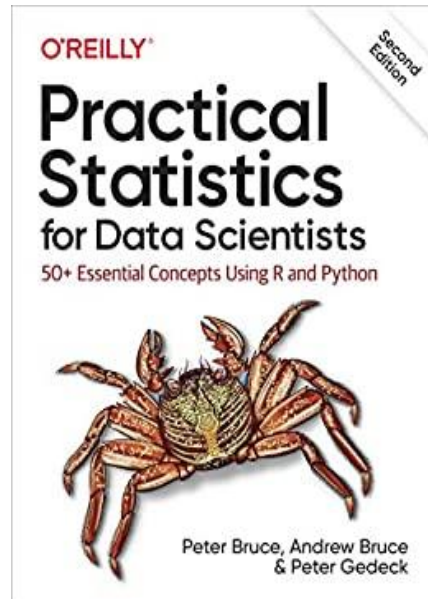
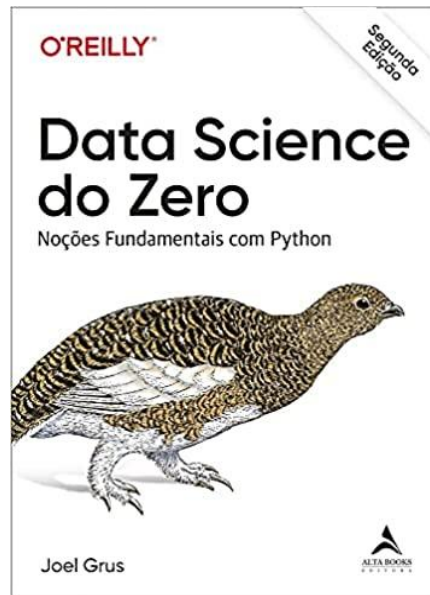
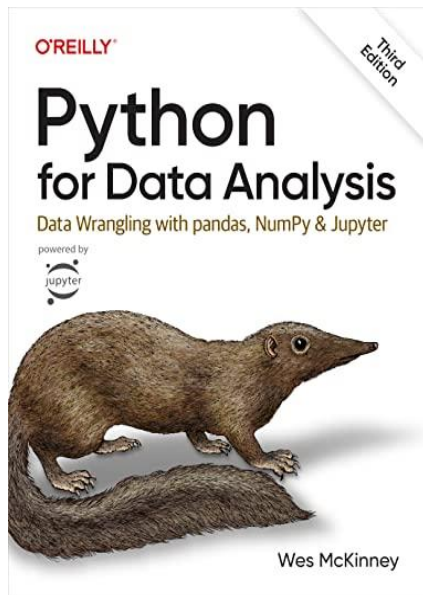
# Referências



# Bibliografia fundamental



# Bibliografia técnica



**OBRIGADO**

