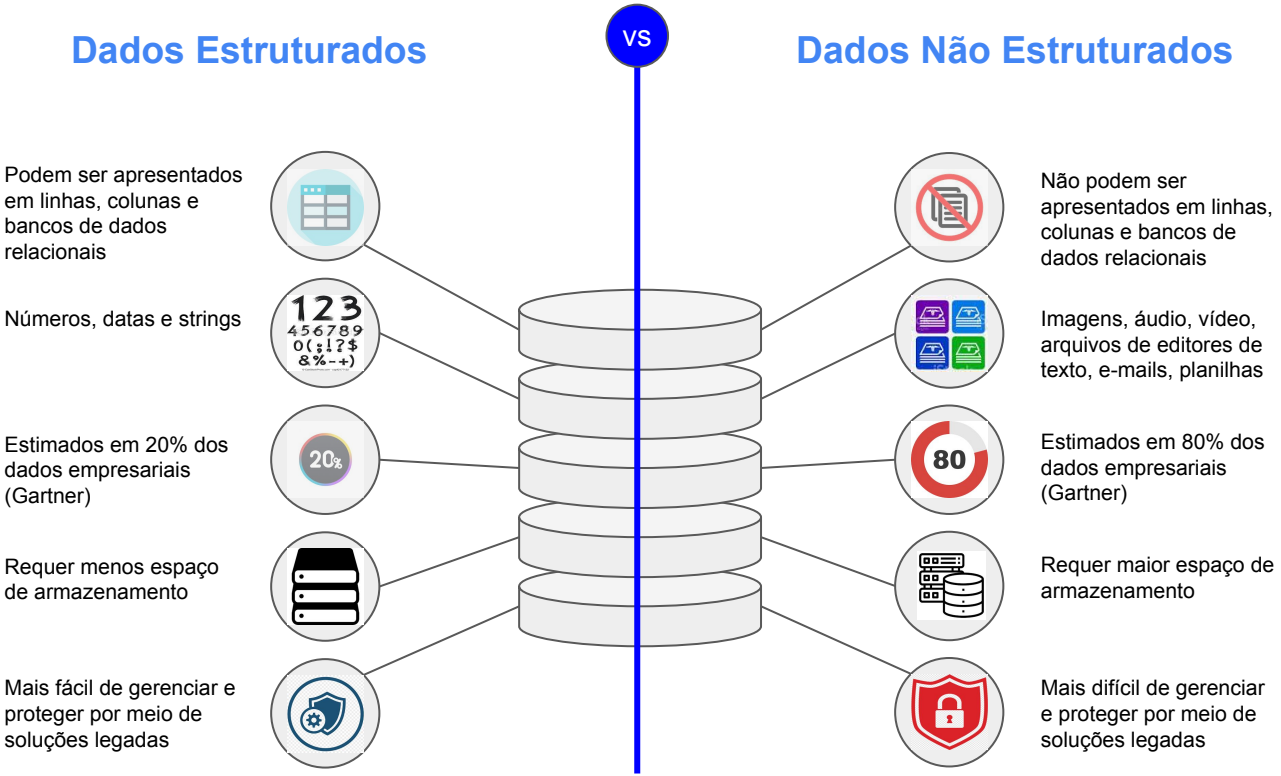
Aula 4
Preparação de Dados
Conceitos fundamentais

Adaptado dos slides do Prof Samuel Botter

Análise Exploratória

Refere-se ao processo de investigação de dados (de um fenômeno ou de um comportamento observado) para a descoberta de *insights*, padrões, anomalias, para testar hipóteses e verificar suposições, empregando-se técnicas e conceitos estatísticos e representações visuais/gráficas.

Tipos de Dados



Dados Estruturados

Registro, Exemplo, Observação, Amostra(*)

Característica (*feature*), Atributo, Variável, Entrada

	DATA INICIAL	DATA FINAL	MÊS	ANO	UF	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO DE REVENDA
0	2004-05-09	2004-05-15	5	2004	DF	127	R$/l	1,288
1	2004-05-09	2004-05-15	5	2004	GO	387	R$/l	1,162
2	2004-05-09	2004-05-15	5	2004	MT	192	R$/l	1,389
3	2004-05-09	2004-05-15	5	2004	MS	162	R$/l	1,262
4	2004-05-09	2004-05-15	5	2004	AL	103	R$/l	1,181
5	2004-05-09	2004-05-15	5	2004	BA	408	R$/l	1,383

Índice (geralmente numérico, mas pode ser texto)

Data Frame, Tabela, Dados retangulares

Dataset → Gas Prices in Brazil: https://www.kaggle.com/matheusfreitag/gas-prices-in-brazil

Dados Estruturados

	DATA INICIAL	DATA FINAL	MÊS	ANO	UF	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO DE REVENDA
0	2004-05-09	2004-05-15	5	2004	DF	127	R$/l	1,288
1	2004-05-09	2004-05-15	5	2004	GO	387	R$/l	1,162
2	2004-05-09	2004-05-15	5	2004	MT	192	R$/l	1,389
3	2004-05-09	2004-05-15	5	2004	MS	162	R$/l	1,262
4	2004-05-09	2004-05-15	5	2004	AL	103	R$/l	1,181
5	2004-05-09	2004-05-15	5	2004	BA	408	R$/l	1,383

Série (vetor de dados, vetor de características)

DATA INICIAL	2004-05-09
DATA FINAL	2004-05-15
MÊS	5
ANO	2004
UF	DF
NÚMERO DE POSTOS PESQUISADOS	127
UNIDADE DE MEDIDA	R$/l
PREÇO MÉDIO DE REVENDA	1,288

Série (vetor de dados)

0	1,288
1	1,162
2	1,389
3	1,262
4	1,181
5	1.383

Nome: PREÇO MÉDIO DE REVENDA, dtype: float64

Dados Estruturados

Variáveis independentes

Variável dependente (Saída, Output, Target, Resposta)

	DATA INICIAL	DATA FINAL	MÊS	ANO	UF	NÚMERO DE POSTOS PESQUISADOS	UNIDADE DE MEDIDA	PREÇO MÉDIO DE REVENDA
0	2004-05-09	2004-05-15	5	2004	DF	127	R$/l	1,288
1	2004-05-09	2004-05-15	5	2004	GO	387	R$/l	1,162
2	2004-05-09	2004-05-15	5	2004	MT	192	R$/l	1,389
3	2004-05-09	2004-05-15	5	2004	MS	162	R$/l	1,262
4	2004-05-09	2004-05-15	5	2004	AL	103	R$/l	1,181
5	2004-05-09	2004-05-15	5	2004	BA	408	R$/l	1,383

Terminologia