



# A Importância da Manipulação e Preparação de Dados para Análise

Antes de analisar, é preciso preparar

Profa. Daniela Trevisan

Prof. Diego Rodrigues

Prof. David Prata

# Análise de Dados

Dê os seus primeiros passos aprendendo sobre como os dados podem ser úteis para o trabalho do dia a dia e como extrair o máximo deste recurso

Aprenda três linguagens necessárias para a análise de dados

Introdução à Ciências de Dados

Análise estatística para Ciências de Dados

Linguagens de Programação para Ciências de Dados

Manipulação de Dados

Entenda os conceitos fundamentais de estatística para análise de dados e a interpretação de resultados

Faça análise exploratória, pré-processamento e entenda sobre visualização de dados, fases essenciais na ciência de dados



## AGOSTO 2025

DOM	SEG	TER	QUA	QUI	SEX	SAB
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

Curso de Manipulação de Dados

Quintas-Feiras – Manhã

Sextas-feiras – Manhã

Aula Síncrona

Aula Assíncrona

## SETEMBRO 2025

DOM	SEG	TER	QUA	QUI	SEX	SAB
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

Curso de Manipulação de Dados

Quintas-Feiras – Manhã

Sextas-feiras – Manhã

Aula Síncrona

Aula Assíncrona

Término do Curso

07-Dia da Independência

# Dado - Ativo Estratégico - Saúde Pública

**Em tempos de transformação digital, os dados são tão valiosos quanto equipamentos ou infraestrutura... ou mais!**

**Utilização adequada permite:**

- ✓ Monitorar surtos e epidemias
- ✓ Planejar recursos com eficiência
- ✓ Basear decisões clínicas e políticas públicas

**⚠ Sem análise, dados não geram conhecimento — apenas ocupam espaço**

# A Super Geração de Dados

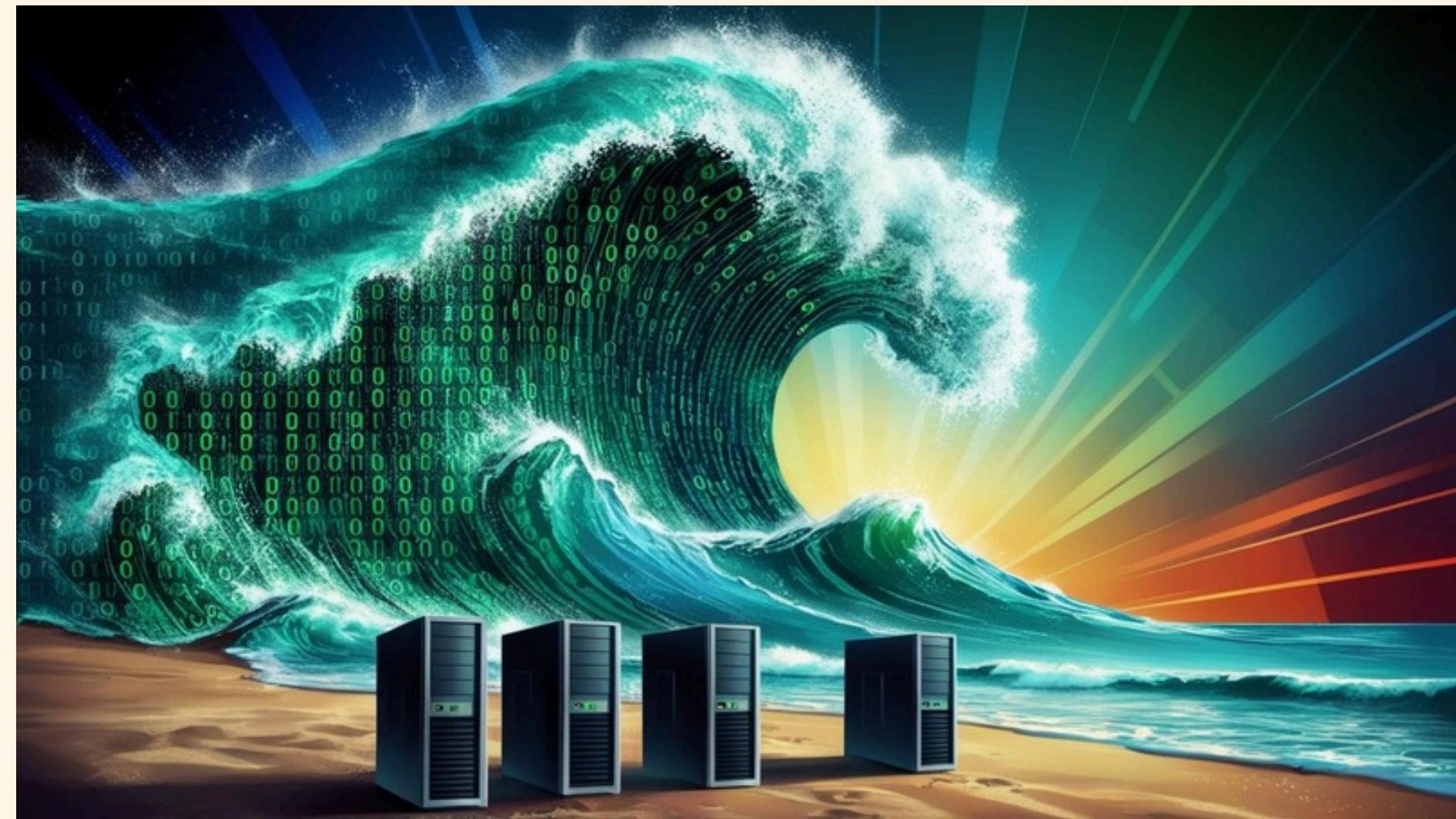
**2,5 quintilhões de bytes de dados são criados diariamente**

IoT: sensores, monitores cardíacos, dispositivos vestíveis

Apps: glicemia, passos, sono

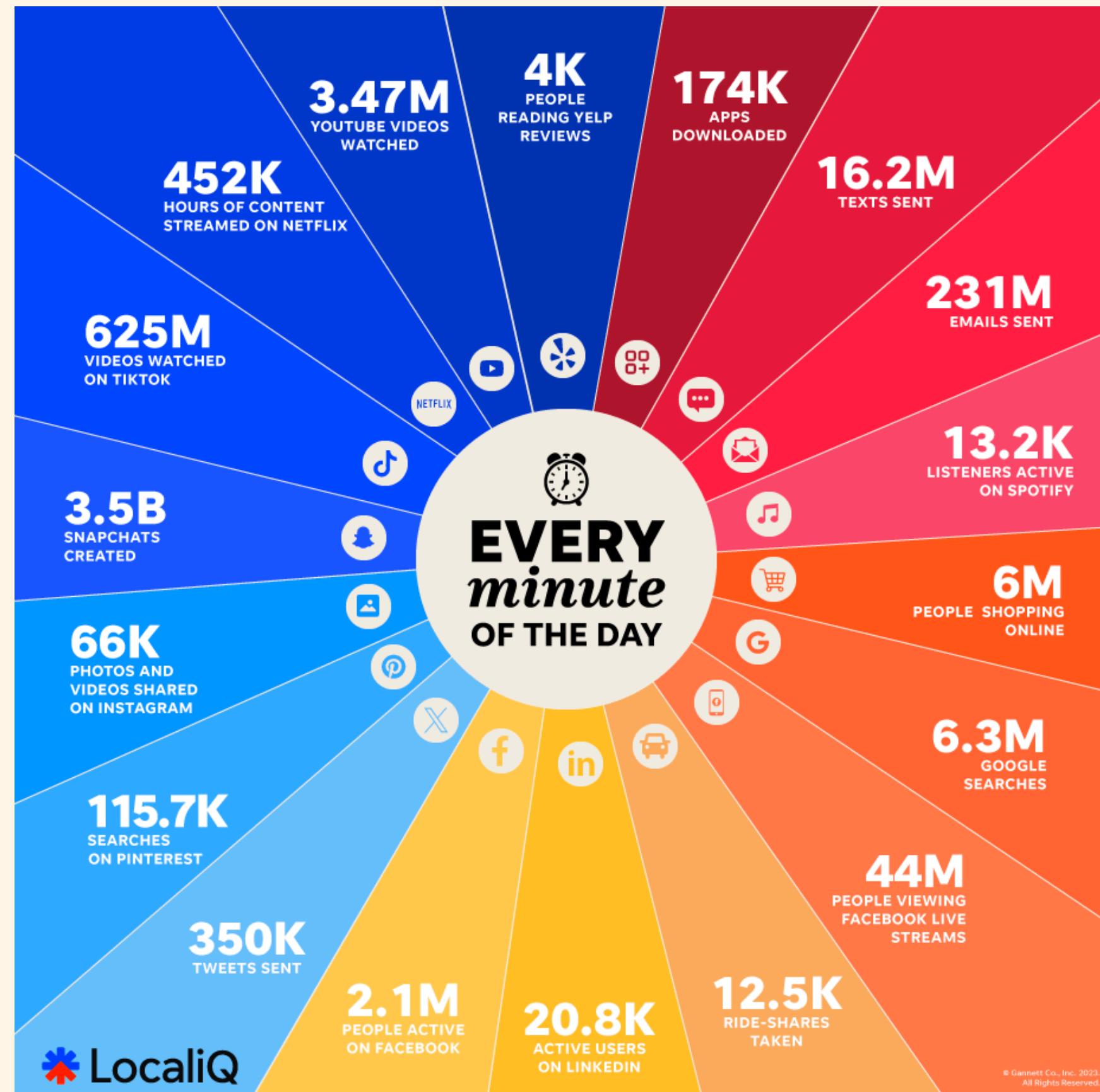
Sistemas oficiais: e-SUS, SINAN, SIVEP-Gripe

Inteligência Artificial: depende de grandes volumes de dados

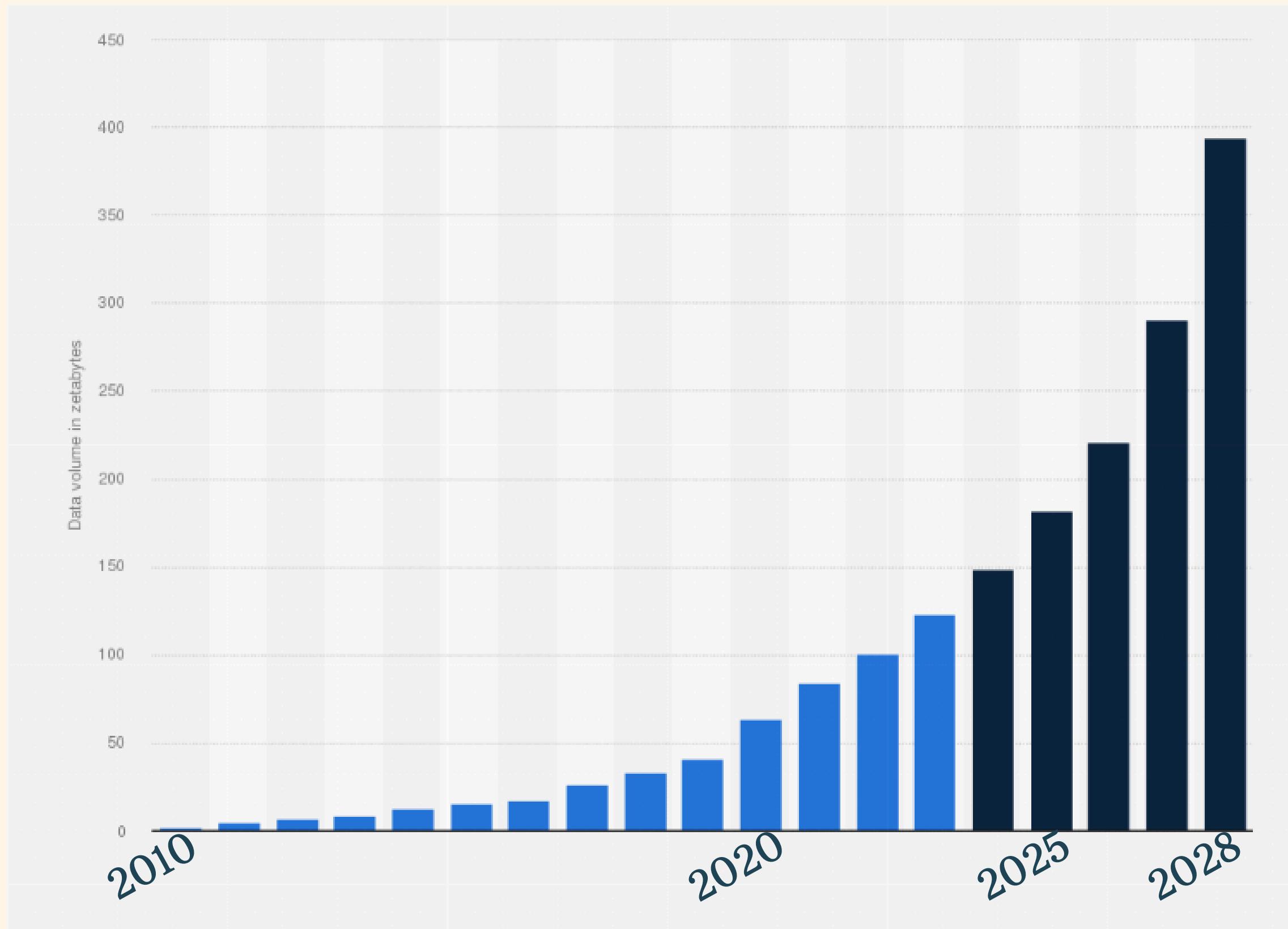


**Acumular dados não é o mesmo que compreender**

# A Super Geração de Dados

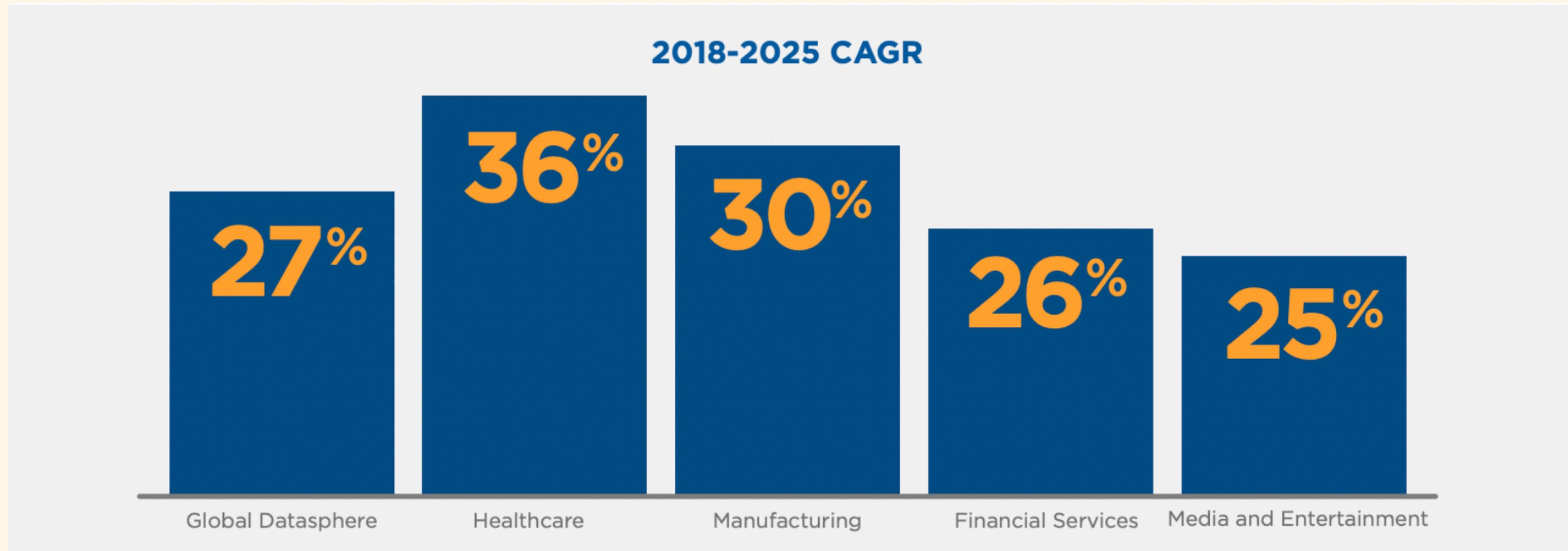


# A Super Geração de Dados

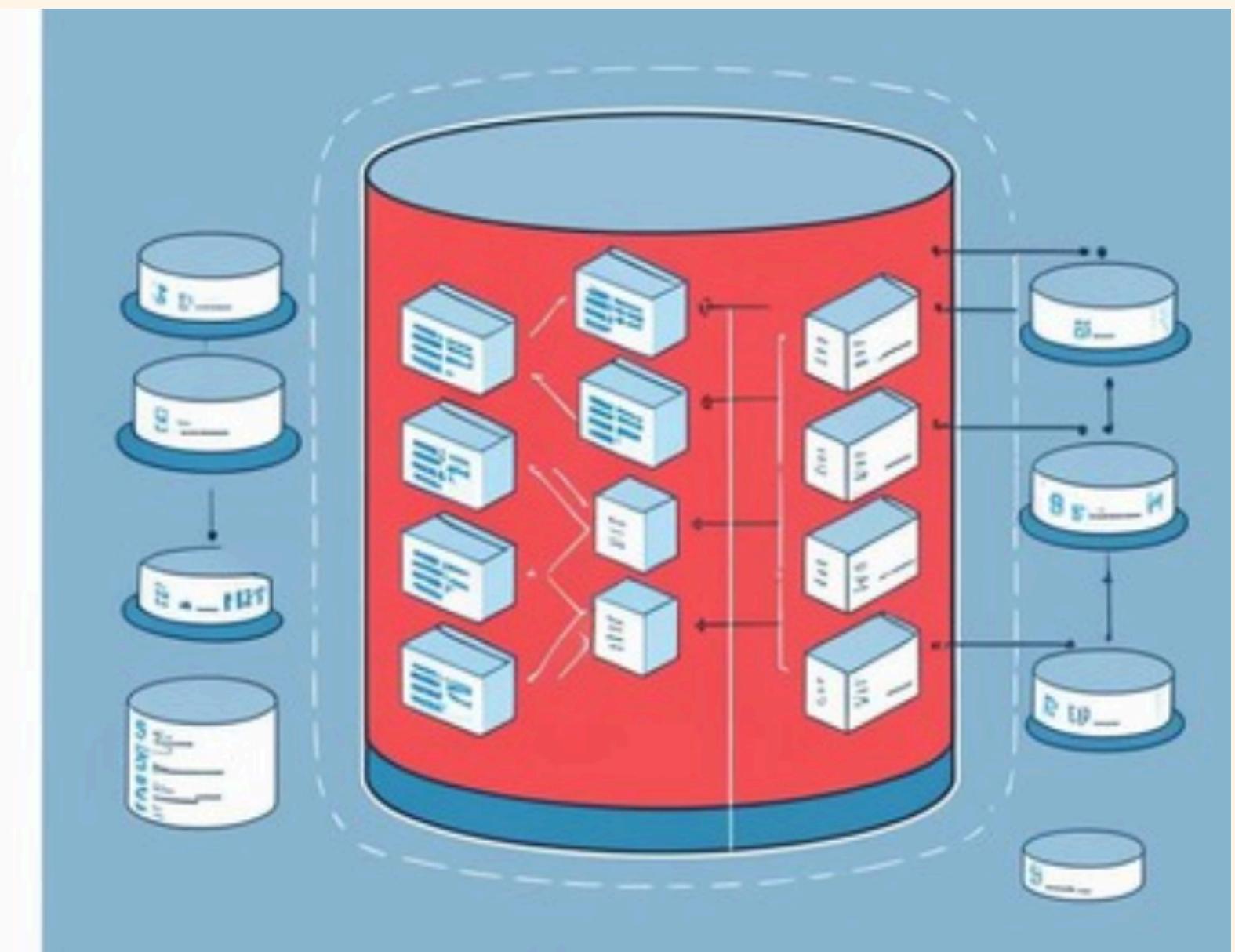
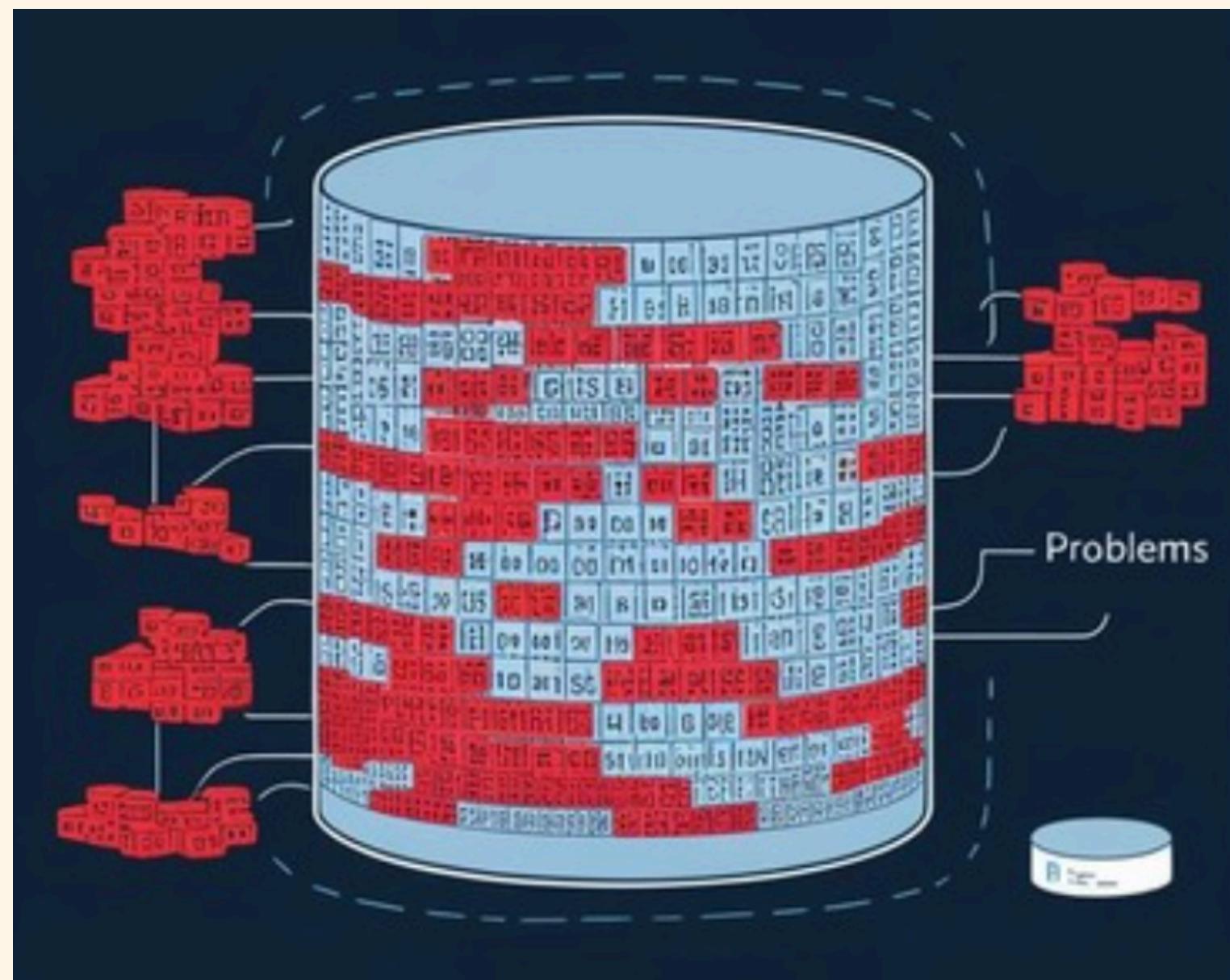


Fonte: Statista

# A Super Geração de Dados



# O que fazemos com esse volume todo?



# Por que Adotar uma Cultura Data-Driven?

## O que é:

-  Decisões com base em dados
-  Políticas públicas baseadas em evidências
-  Planejamento com indicadores reais
-  Eficiência no uso de recursos

## Principais Benefícios:

-  Tomada de decisão baseada em evidências
-  Eficiência operacional
-  Inovação e competitividade
-  Personalização ao cliente
-  Gestão de riscos

**Maturidade em dados = decisões melhores, menos desperdício!**

# Como evoluir para uma organização Data-Driven



A maturidade em dados é uma jornada

# Por que ser Data-Driven transforma a gestão em saúde?

-  Redução de filas e tempo de espera
-  Identificação precoce de surtos e padrões anormais
-  Planejamento de recursos com base em demanda real
-  Monitoramento do uso de medicamentos e insumos
-  Acompanhamento de metas e indicadores assistenciais
-  Mais precisão na vigilância e nas auditorias

**Ex: “Reduzimos filas sem gastar mais” – esse é o impacto real da cultura orientada por dados**

# Por que Preparar Dados é Fundamental?



## Dados Brutos

Dados brutos são frequentemente inconsistentes e desorganizados, exigindo tratamento para se tornarem úteis.



## Informação Confiável

A preparação transforma dados brutos em informação confiável, essencial para análises precisas.



## Decisões Assertivas

Com dados preparados, é possível tomar decisões estratégicas mais assertivas, minimizando riscos.

# O Que é Preparação de Dados?

A preparação de dados é o **coração** do processo de análise - é onde **damos vida** aos dados brutos, transformando-os em informações valiosas e confiáveis.

Isso envolve:

- 1 Remover duplicatas
- 2 Corrigir erros
- 3 Padronizar formatos



# Benefícios Diretos da Preparação de Dados



## Confiança

Dados limpos e precisos elevam a segurança, permitindo decisões estratégicas com base em informações totalmente confiáveis.



## Agilidade

Agilize suas análises e otimize o tempo. Dados bem estruturados garantem um fluxo de trabalho mais rápido e resultados imediatos.



## Descoberta

Desvende padrões ocultos e novas oportunidades. Dados preparados são a chave para insights profundos que impulsionam a inovação.



## Economia

Reduza custos e minimize retrabalho. Com dados organizados, você economiza recursos e evita erros, garantindo maior eficiência operacional.

# Etapas Essenciais da Preparação de Dados



## Coleta

Reunir dados de múltiplas fontes, garantindo abrangência.



## Limpeza

Corrigir erros, preencher dados faltantes e remover inconsistências.



## Transformação

Normalizar, codificar, agregar e criar variáveis derivadas para análise.



## Integração

Combinar dados de diferentes fontes em um conjunto unificado.



## Validação

Garantir consistência e qualidade final dos dados preparados.

# Os 5 passos da manipulação de dados na prática

Etapa	O que é na prática	Exemplo real aplicado à saúde
<b>1. Coleta</b>	Reunir dados de diferentes fontes	Bases do e-SUS, SIVEP-Gripe, CNES ou planilhas
<b>2. Limpeza</b>	Corrigir valores inválidos, nulos ou inconsistentes	Sexo = “feminino”, “F”, “f” → padronizar
<b>3. Transformação</b>	Calcular novos campos, categorizar, normalizar	Criar “tempo de internação” ou “faixa etária”
<b>4. Integração</b>	Cruzar informações de diferentes arquivos ou sistemas	Unir dados de internação com dados de vacinação
<b>5. Validação</b>	Verificar coerência, duplicações, tipos e padrões	Conferir datas invertidas, colunas com muitos NaN

# Exemplos de Manipulação de Dados

- **Integração de Sistemas:** Combinar dados de diferentes sistemas (CRM, ERP, Vendas) com formatos distintos para uma visão unificada do cliente.
- **Extração Não Estruturada:** Extrair informações cruciais de documentos não estruturados, como contratos em PDF ou e-mails de feedback, para análise de sentimentos.
- **Padronização Manual:** Corrigir dados inseridos manualmente, como nomes de cidades ou produtos, para evitar duplicidades e variações inconsistentes que distorcem resultados.
- **Tratamento de Valores Ausentes:** Decidir como lidar com dados faltantes, seja preenchendo-os com valores médios, mediana ou moda, ou removendo as entradas incompletas.
- **Conversão de Tipos de Dados:** Alterar o formato de dados de texto para numérico, ou de data para um formato padronizado, para permitir cálculos e análises corretas.
- **Agregação de Dados:** Resumir grandes volumes de dados em métricas mais gerenciáveis, como calcular a receita total mensal a partir de transações individuais.
- **Filtragem e Seleção:** Remover linhas ou colunas irrelevantes para focar apenas nos dados pertinentes à análise, reduzindo ruído e melhorando a performance.



# Do caos à clareza: exemplos de problemas e como resolvê-los

● Problema: Coluna "sexo" com valores como M, F, feminino, Masculino, 3

Solução: Padronização com valores únicos e consistentes (F, M)

● Problema: Idade = 999 ou data de nascimento = 1900-01-01

Solução: Tratamento de outliers e correção de datas inválidas

● Problema: Paciente aparece duas vezes com mesmo ID e data

Solução: Remoção de duplicatas reais ou reconciliação de registros

● Problema: Diagnóstico como “Outros” ou “Não informado”

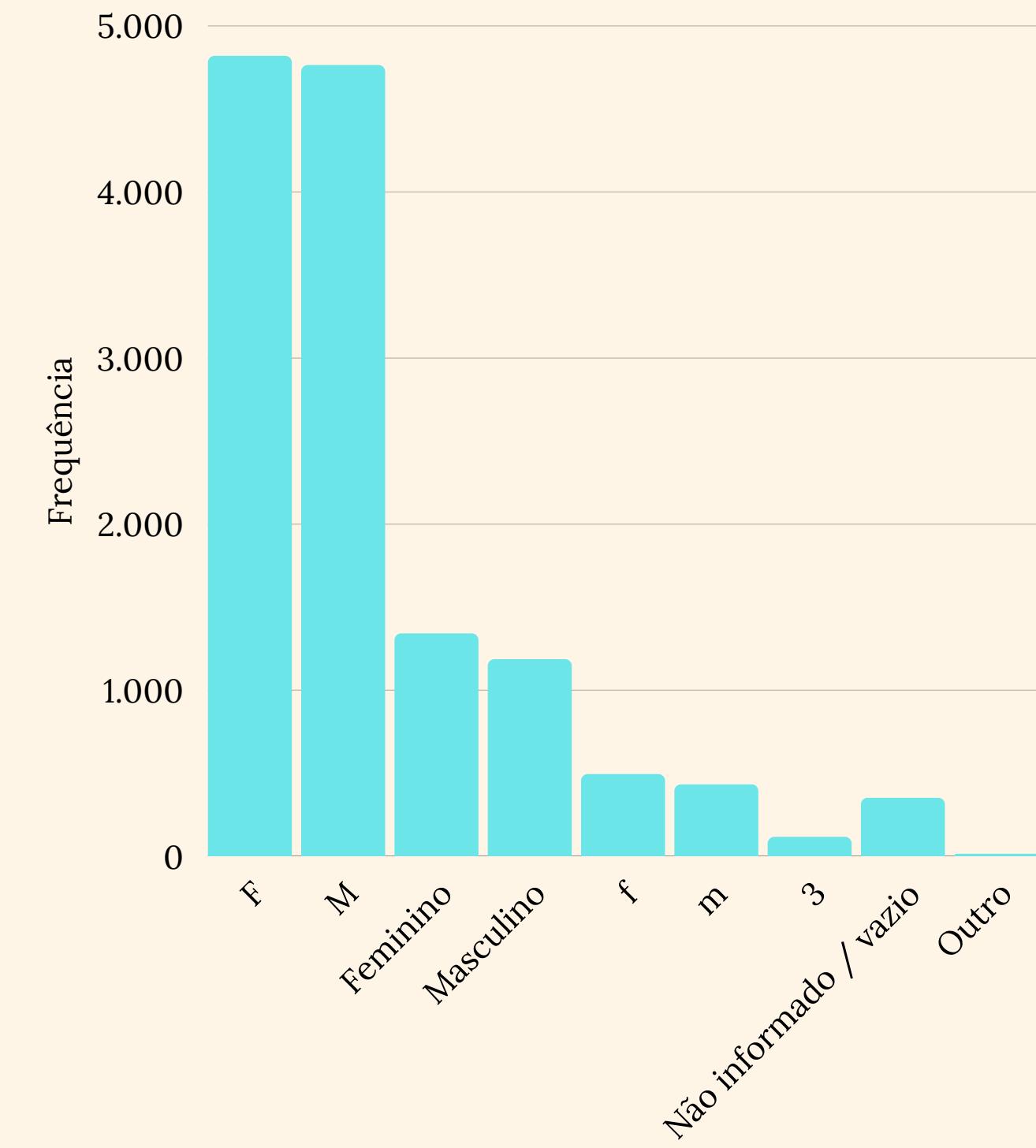
Solução: Revisão da codificação e categorização mais informativa

● Problema: Datas em formatos diferentes (pt-BR, ISO, texto)

Solução: Conversão padronizada para formato aaaa-mm-dd ou datetime

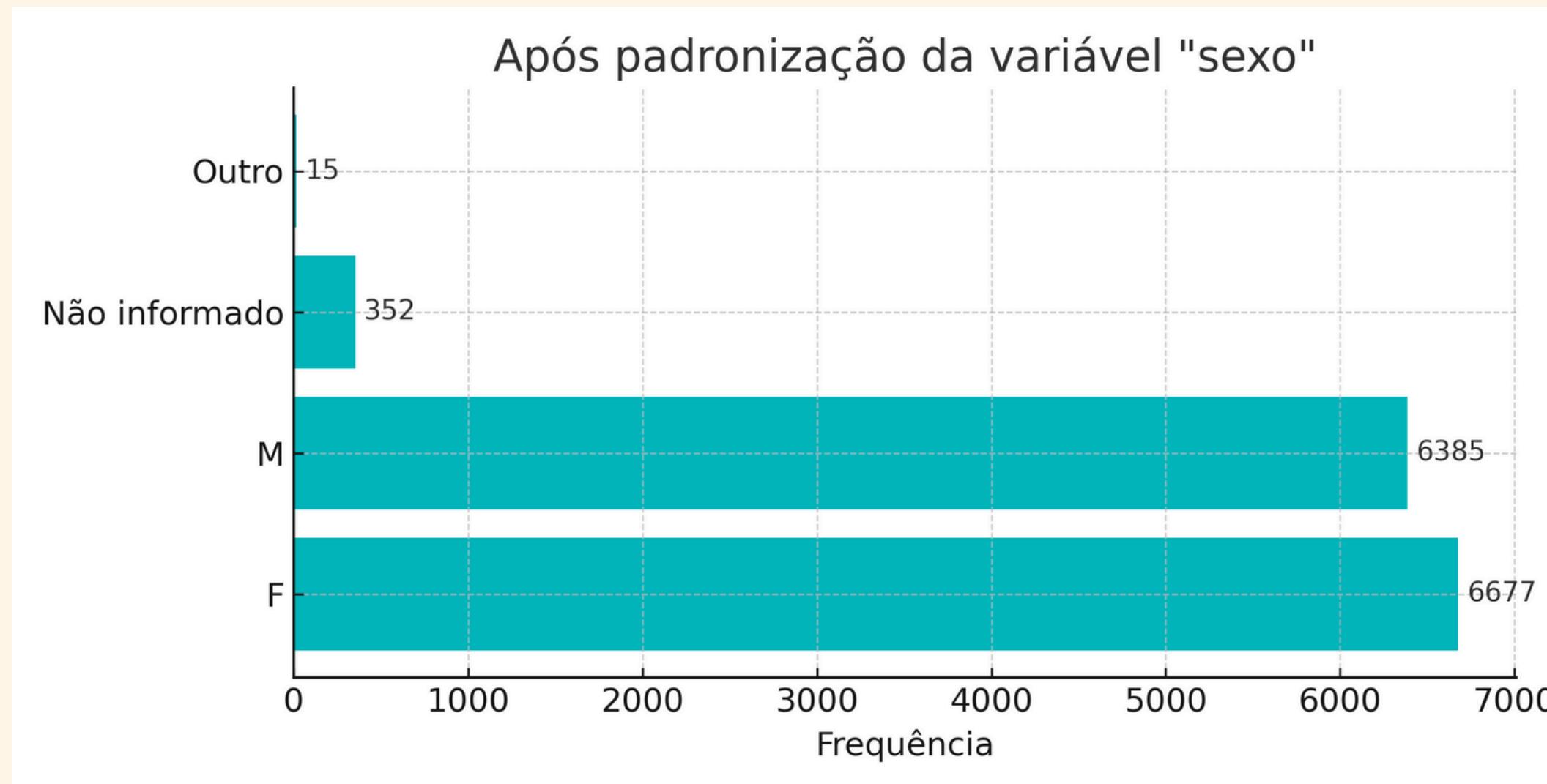
# O problema da despadronização: variável "sexo" mal preenchida

	Frequência
F	4.820
M	4.765
Feminino	1.342
Masculino	1.187
f	495
m	433
3	117
Não informado	352
Outro	15



# Depois da limpeza: dados organizados e prontos para análise

Valor padronizado	Originais incluídos	Frequência total
<b>F</b>	F, f, Feminino	$4.820 + 495 + 1.342 = \mathbf{6.657}$
<b>M</b>	M, m, Masculino	$4.765 + 433 + 1.187 = \mathbf{6.385}$
<b>Não informado</b>	Vazio, "", Não informado	<b>352</b>
<b>Outro</b>	Outro	<b>15</b>



# Checklist de qualidade dos dados

**Antes de analisar, pergunte-se:**

- 🔍 Completos? → Há dados ausentes ou campos obrigatórios em branco?
- ♻️ Consistentes? → Os valores seguem o mesmo padrão (ex: F/feminino/FEM)?
- 🧽 Limpos? → Há erros de digitação, duplicidades ou códigos inválidos?
- 📅 Temporais coerentes? → Datas fazem sentido? Não há alta antes da internação?
- 🔄 Sem redundância? → Há registros duplicados ou repetições desnecessárias?
- 👤 Identificáveis? → Os dados permitem distinguir indivíduos ou eventos corretamente?

# Um erro simples com um impacto complexo

## Base de dados pública de internações hospitalares.

Um relatório apontava que a média de permanência hospitalar de mulheres com 75 anos era de 139 dias – muito acima do esperado.

Após análise da base, descobriu-se:

- Diversos registros com alta após o óbito
- Datas invertidas: alta antes da internação
- Campo "data de alta" não preenchido (ou com “00/00/0000”)

O resultado?

Cálculo da média distorcido, levando gestores a acreditar que havia problemas graves de permanência prolongada.

# Impacto na Análise e Modelagem

## Aumento da Precisão

Dados preparados aumentam a precisão de modelos estatísticos e de Machine Learning, levando a previsões mais acuradas.

## Redução de Vieses

Minimizam vieses e erros que podem comprometer os resultados e a validade das conclusões.

## Análises Confiáveis

Permitem análises preditivas e diagnósticas mais robustas e confiáveis, suportando decisões críticas.

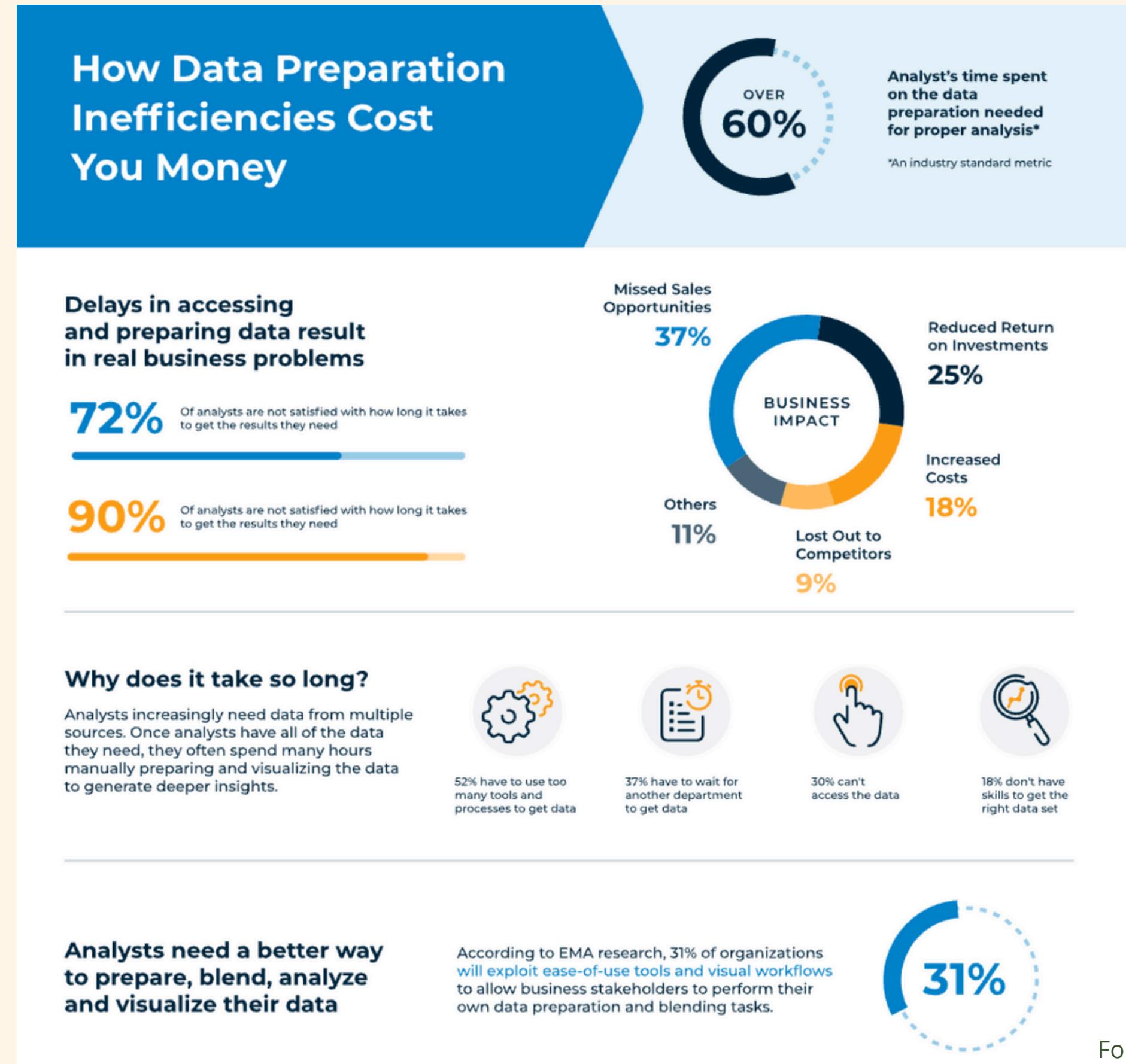
## Melhora na Performance

Dados limpos e organizados otimizam o desempenho de algoritmos, resultando em modelos mais eficientes e velozes.

## Novas Perspectivas

Permitem a descoberta de insights e padrões complexos que seriam obscurecidos por dados brutos ou inconsistentes.

# Preparação ineficiente = alto custo



Fonte: [alteryx.com](http://alteryx.com)

# Preparação ineficiente = alto custo

## 1. Tempo gasto na preparação de dados

- Pesquisas indicam que cientistas e analistas de dados ainda passam cerca de 60% a 80% do seu tempo coletando, limpando e organizando os dados — e apenas o restante em análise propriamente dita
- Uma revisão acadêmica recente destaca que a preparação de dados (ou data wrangling) continua sendo uma tarefa laboriosa e consome a maior parte do tempo operacional desses profissionais

## 2. Problemas de qualidade de dados

- Um relatório apontou que 57% dos entrevistados consideram a qualidade dos dados como um dos três maiores desafios durante a preparação de dados.

A má qualidade de dados gera prejuízos financeiros significativos — por exemplo, a Gartner estima perdas médias de US\$ 12,9 milhões por ano por organização.

# Preparação ineficiente = alto custo

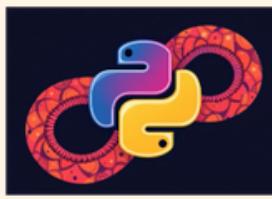
## 3. Crescimento do mercado de preparação de dados

- O mercado global de ferramentas de data preparation movimentou cerca de US\$ 6–7 bilhões em 2024/2025, com previsão de crescer para até US\$ 14–27 bilhões entre 2030 e 2033 (CAGR de ~15-18%).
- Essa expansão é impulsionada pela demanda por ferramentas de autoatendimento, automação com IA/ML e governança de dados.

## 4. Inovações em automação e assistentes de dados

- Um estudo de abril de 2025 introduziu um sistema chamado Auto-Prep, capaz de prever mais de 70% dos passos de transformação e junção de dados em projetos de Business Intelligence, superando modelos genéricos como o GPT-4 [arXiv](#).
- Técnicas de IA assistida estão sendo empregadas para automatizar tarefas manuais de limpeza e transformação, tornando o processo mais eficiente [arXiv](#).

# Ferramentas e Técnicas Comuns



As ferramentas modernas facilitam muito a preparação:

- **Planilhas Inteligentes:** Uso de planilhas como Excel com validação de dados para evitar erros na entrada manual.
- **Softwares Automatizados:** Ferramentas como Python (com bibliotecas Pandas, NumPy) e Power Query (Excel/Power BI) para automação de limpeza e transformação.
- **Pipelines de Dados:** Implementação de pipelines que automatizam as etapas de preparação, garantindo repetibilidade e eficiência em larga escala.

# O Papel da Visualização na Preparação



A visualização de dados não é apenas para o final da análise; ela é uma etapa crucial na preparação:

- **Identificação de Inconsistências:** Gráficos e dashboards ajudam a identificar rapidamente outliers, valores faltantes ou padrões anômalos que precisam de correção.
- **Comunicação Clara:** Facilita a comunicação dos resultados da preparação e das qualidades dos dados para stakeholders.
- **Exploração Acelerada:** A visualização interativa acelera a exploração e o entendimento dos dados, revelando insights que seriam difíceis de ver em tabelas.

# A IA como copiloto na manipulação de dados

**Ferramentas como ChatGPT, Copilot e Gemini  
são capazes de:**

- Sugerir códigos de limpeza e transformação
- Detectar padrões e inconsistências
- Gerar funções personalizadas para tarefas repetitivas
- Atuar como tutor no uso de bibliotecas (ex: Pandas, NumPy)



**A IA não faz o trabalho por você — ela acelera o caminho até ele.**

# Exemplo: automatizar transformações

## Prompt enviado à IA:

“Estou com uma base de dados da vacinação contra COVID. A coluna sexo está com valores como f, F, feminino, Masculino, m, M, Outro, e alguns campos em branco.

Gere um código em Python para padronizar essa coluna.”

Processo tradicional	Com apoio da IA
Demandava conhecimento técnico	Sugestão pronta, explicada
Exige tempo de teste e ajuste	Pronto para rodar com leve revisão
Pode gerar insegurança inicial	IA atua como guia e acelera rotina

“Estou com uma base de dados da vacinação contra COVID. A coluna sexo está com valores como f, F, feminino, Masculino, m, M, Outro, e alguns campos em branco.

Gere um código em Python para padronizar essa coluna.”

Claro! Abaixo está um código em Python usando **Pandas** que padroniza a coluna `sexo` para os valores:

“Feminino”, “Masculino”, “Outro” ou `NaN` (caso esteja em branco ou não reconhecido):

```
python

import pandas as pd
import numpy as np

# Exemplo de carregamento de dados
# df = pd.read_csv('sua_base.csv') # substitua pelo caminho real do seu arquivo

# Função para padronizar a coluna sexo
def padronizar_sexo(valor):
    if pd.isna(valor) or str(valor).strip() == "":
        return np.nan
    valor = str(valor).strip().lower()
    if valor in ['f', 'feminino']:
```

# Exemplo: padronizar valores de idade inválida

## Prompt enviado à IA:

“Tenho uma coluna chamada idade. Alguns valores estão como 999 ou negativos. Como padronizar e substituir por NaN no Python?”

Tenho uma coluna chamada idade. Alguns valores estão como 999 ou negativos. Como padronizar e substituir por NaN no Python?

Você pode padronizar os valores da coluna `idade` no Python utilizando o `pandas`, substituindo os valores inválidos (como `999` e negativos) por `NaN` da seguinte forma:

python

[Copiar](#) [Editar](#)

```
import pandas as pd  
import numpy as np
```

```
# Exemplo: carregando seu DataFrame  
# df = pd.read_csv('seuarquivo.csv') # caso esteja lendo de um arquivo
```

```
# Suponha que já exista um DataFrame com a coluna 'idade'  
# Substituir valores 999 e negativos por NaN  
df['idade'] = df['idade'].apply(lambda x: np.nan if x == 999 or x < 0 else x)
```

Alternativa com `where` (mais eficiente com grandes bases):

python

[Copiar](#) [Editar](#)

```
df['idade'] = df['idade'].where((df['idade'] != 999) & (df['idade'] >= 0), np.nan)
```

Com uma linha de código, o dado ganha coerência.

# Exemplo: padronizar valores de idade inválida

Antes e depois da limpeza da idade

Situação	Exemplo na coluna idade
Antes da limpeza	32, 999, -1, 54, 0, 121
Depois da limpeza	32, NaN, NaN, 54, NaN, NaN

Com uma linha de código, o dado ganha coerência.

# Exemplo: Converter datas e calcular intervalo

## Prompt enviado à IA:

“Tenho colunas data\_entrada e data\_alta como texto. Quero convertê-las para datas e calcular o tempo de internação”

Tenho colunas data\_entrada e data\_alta como texto. Quero convertê-las para datas e calcular o tempo de internação

Para converter as colunas `data_entrada` e `data_alta` de texto para datas e calcular o tempo de internação em dias, você pode usar o seguinte código com `pandas`:

python

Copiar Editar

```
import pandas as pd

# Exemplo: convertendo as colunas para formato de data
df['data_entrada'] = pd.to_datetime(df['data_entrada'], errors='coerce')
df['data_alta'] = pd.to_datetime(df['data_alta'], errors='coerce')

# Calculando o tempo de internação em dias
df['tempo_internacao'] = (df['data_alta'] - df['data_entrada']).dt.days
```

## Explicação rápida:

- `pd.to_datetime(..., errors='coerce')` converte o texto para data; se tiver erro, vira `NaT`.
- A subtração entre datas resulta em um `timedelta`, e `.dt.days` extrai o número de dias.
- O resultado `tempo_internacao` será um número inteiro (ou `NaN` se alguma data estiver faltando ou inválida).

Se quiser, posso adicionar tratamento para casos em que `data_alta` vem antes de `data_entrada`.

Deseja isso também?

A IA também alerta para datas incoerentes, como alta antes da entrada.

# Exemplo: Preencher valores ausentes com mediana

## Prompt enviado à IA:

"Como posso preencher valores ausentes da coluna pressao\_sistolica com a mediana no Python?"

Como posso preencher valores ausentes da coluna pressao\_sistolica com a mediana no Python?

Você pode preencher os valores ausentes ( `Nan` ) da coluna `pressao_sistolica` com a **mediana** da própria coluna usando:

python

Copiar Editar

```
df['pressao_sistolica'] = df['pressao_sistolica'].fillna(df['pressao_sistolica'].median())
```

## Explicação:

- `df['pressao_sistolica'].median()` calcula a mediana da coluna.
- `.fillna(...)` preenche os valores ausentes com esse número.

Esse método é comum quando você quer imputar valores ausentes preservando a distribuição central dos dados.

A IA pode explicar o uso de mediana vs. média e quando cada um é mais indicado.

# Exemplo: Limpando a base

Temos uma base com registros de idade de pacientes. Alguns registros estão incorretos — idade = 999 foi usada como preenchimento automático.

Se não tratarmos esse erro, a análise estatística será distorcida.

Situação	Valores usados para o cálculo	Média calculada
Antes da limpeza	35, 52, 999, 60, 999, 40, 72, 29, 999, 88, 64, 999, 55, 999	392,1 anos ✗
Depois da limpeza	35, 52, 60, 40, 72, 29, 88, 64, 55	55,0 anos ✓

Dados sujos distorcem médias, tendências e inferências.

# Exemplo de fluxo completo com IA copiloto

- ✓ Prompt: “Padronize a coluna sexo, limpe idade inválida e crie coluna com tempo de internação”
- 🤖 A IA gera código com map(), np.nan, to\_datetime(), .dt.days
- 📝 Você testa, valida, ajusta
- 📈 Resultado: base coerente, pronta para análise



**IA como copiloto: você pilota, ela sugere rotas mais curtas.**

# Então... Preparar Dados é Preparar o Sucesso!

**A qualidade da análise depende fundamentalmente da qualidade dos dados.**

Investir tempo e recursos em manipulação e preparação de dados não é um custo, mas sim um **investimento estratégico** que garante decisões mais seguras e eficazes.

Comece hoje a transformar seus dados brutos em **insights valiosos** e impulsionone o sucesso da sua análise!