Secretaria
de Planejamento, Gestão
e Desenvolvimento
Regional

GOVERNO DE PERNAMBUCO
ESTADO DE MUDANÇA

Adulando os dados

Curso Ciência de Dados para a Gestão Pública

Júlia Barrêto

13 de Maio de 2025

Problema x ferramenta

- Está usando a ferramenta certa para resolver seu problema?
- Compreender o problema e as condições atuais da organização é fundamental para o processo de escolha da ferramenta
- Sair da zona de conforto é necessário

Condicionamento

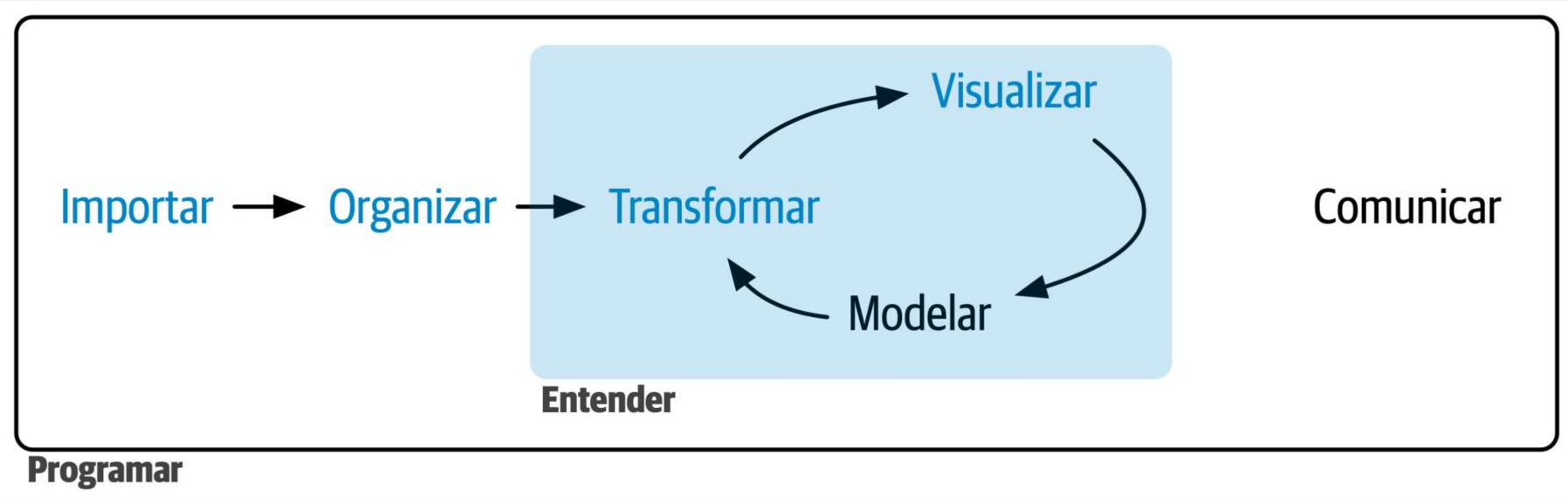
- Questionar os comportamentos
- Propensão que os seres humanos têm de não questionar as coisas e simplesmente aceitar e repetir
- O experimento demonstra como o medo e a busca por evitar o desconforto leva ao conformismo, mesmo quando a causa original do medo já não existe

Adulando dados

1. Extração, transformação e leitura (ETL).
2. Higienização.
3. Casos ausentes: eventos ocultos, matriz sombra, imputação.

Adulando dados

- Deixar os dados prontos, bonitos e confiáveis para análise.



Fonte: R4DC

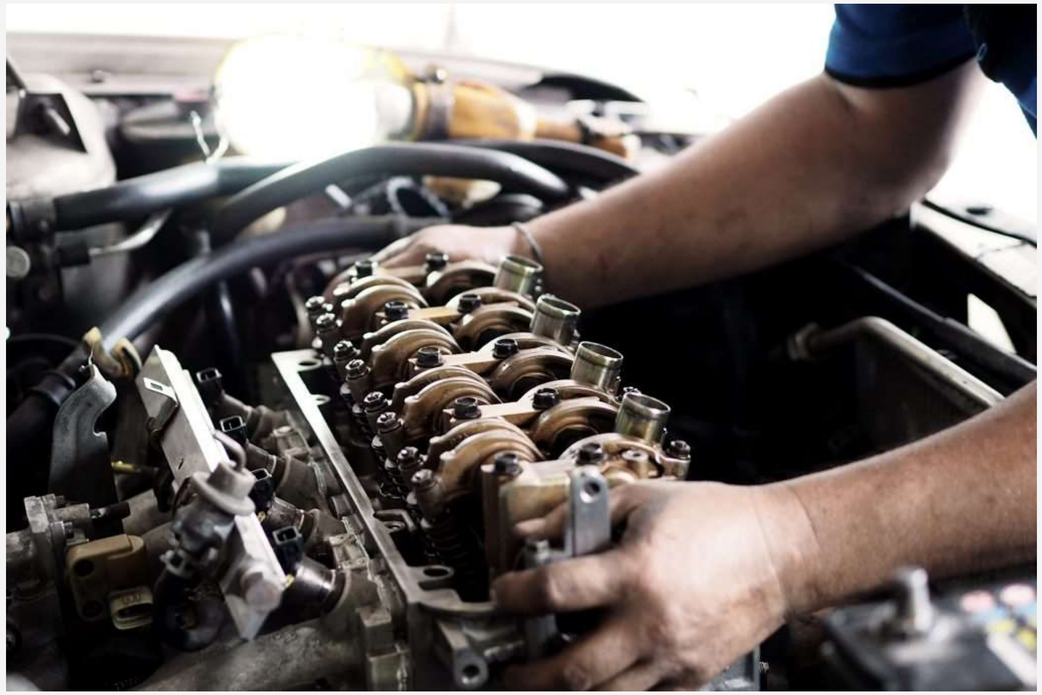
Ferramentas

- Python, R e SQL
- Escolha depende do contexto, orçamento, time e necessidades do negócio.

Uso da linguagem R

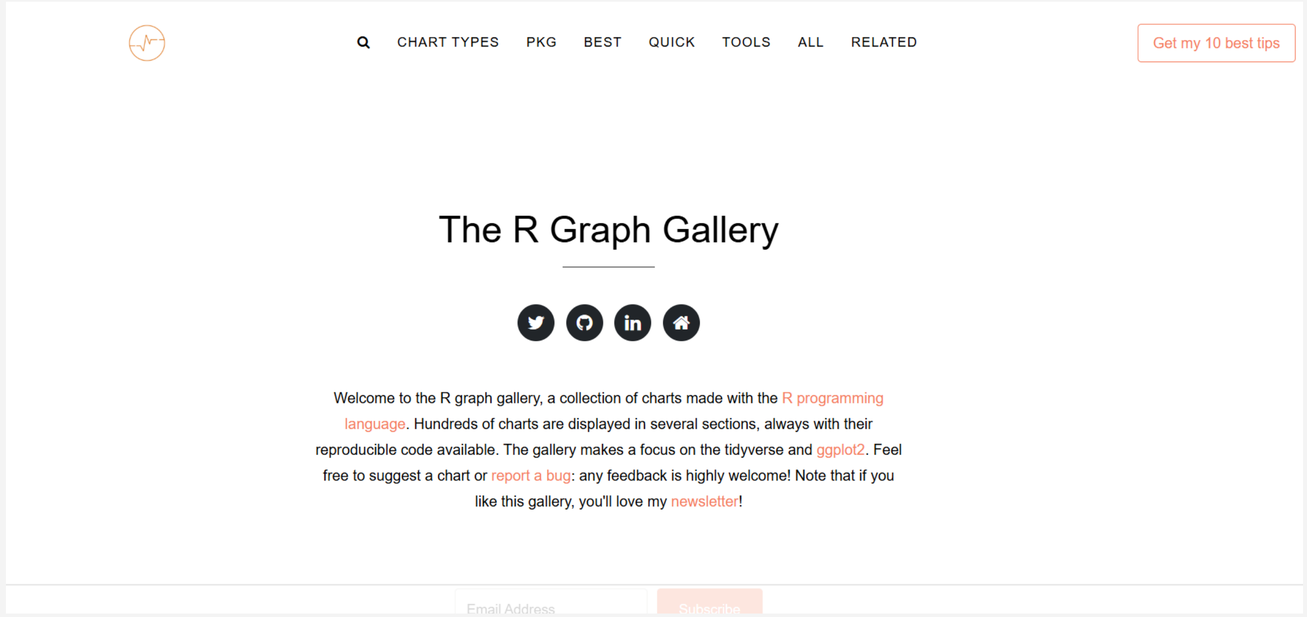
- É gratuita
- Open Source
- Criada por estatísticos para estatística
- Excelente para visualização
- Linguagem alto nível (mais próxima da linguagem humana)
- Ajuda na manipulação, análise e visualização de dados, sendo atualmente considerada uma das melhores ferramentas para essa finalidade.

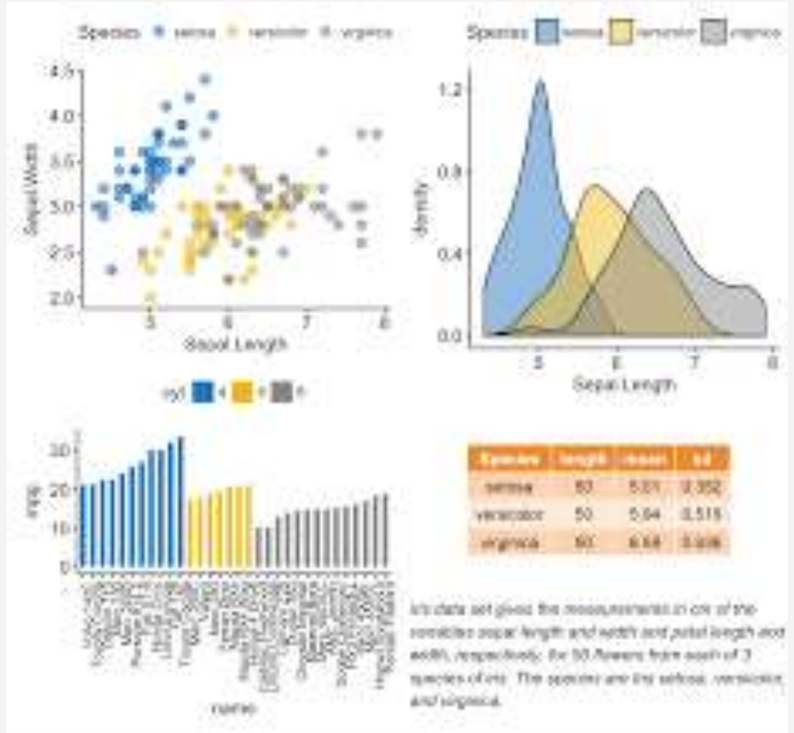
R e Rstudio



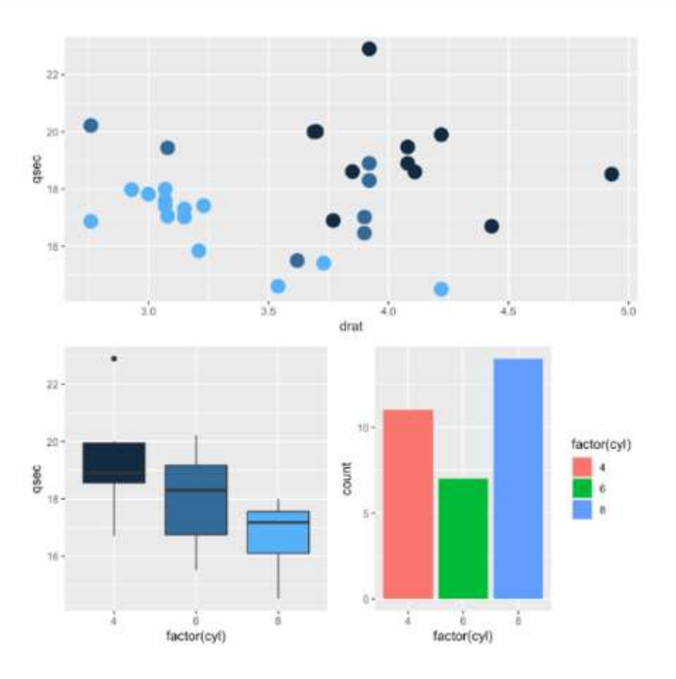
Visualização de dados

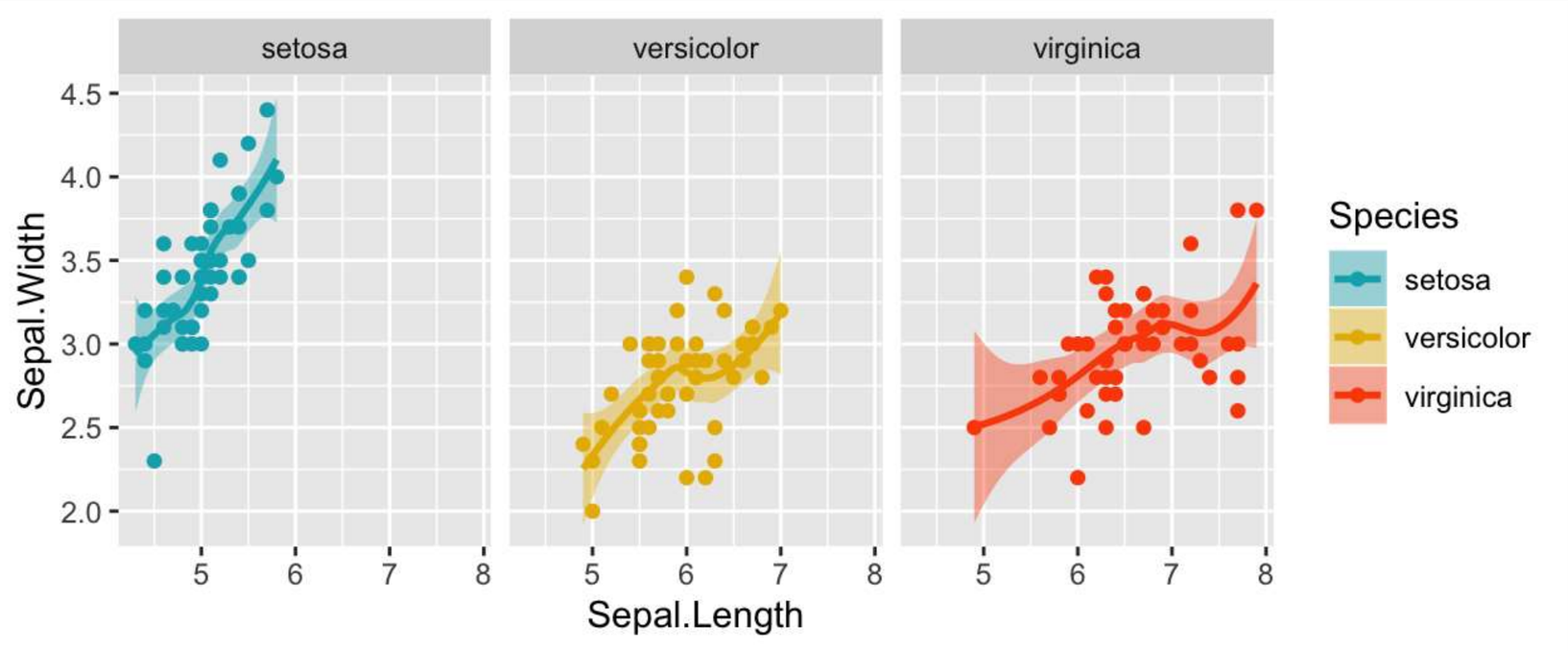
- O ggplot2 implementa a gramática dos gráficos, um sistema coerente para descrever e construir gráficos.





Visualização de dados





ETL

- Processo inicial para preparar os dados
- Extração (Extract): obter dados de fontes distintas -> bancos de dados, APIs, arquivos (CSV, JSON, Excel), web scraping, etc. (Exemplo: Baixar dados do IBGE ou de uma API pública de saúde)
- Transformação (Transform): Limpeza, filtragem, padronização e enriquecimento dos dados. (Exemplos:Corrigir nomes (ex: “BRASIL”, “brasil” → “Brasil”); Criar variáveis novas (ex: idade a partir de data de nascimento); Unir bases com join.
- Leitura/Carga (Load): Salvar os dados transformados em um novo local -> banco de dados, arquivo .parquet, .csv, data warehouse, etc. (Objetivo: Facilitar análises futuras com dados limpos)
- Parte mais demorada do trabalho com dados

Higienização de Dados

- É o coração da transformação. Envolve tornar os dados consistentes, confiáveis e utilizáveis.
- Padronização de texto: remover acentos, letras maiúsculas/minúsculas, espaços.
- Conversão de tipos: texto para data, número para fator etc.
- Tratamento de duplicatas: remoção de registros redundantes.
- Remoção de ruído: dados irrelevantes, erros de digitação, etc.

Casos Ausentes: Eventos Ocultos

- Eventos Ocultos (dados faltantes que representam uma ação ou omissão significativa)
- Exemplo: ausência de um registro pode significar que algo não aconteceu (ex: não houve atendimento médico), ou que não foi registrado.
- Solução: precisa de interpretação contextual; nem todo NA é igual.

Casos Ausentes: Matriz Sombra

- Matriz Sombra (shadow matrix): técnica usada para diagnosticar, visualizar e entender padrões de ausência (missing values) em conjuntos de dados.
- Ela é particularmente útil quando você quer ir além de simplesmente saber “quantos NAs existem” e começar a responder perguntas como: “Os dados ausentes estão concentrados em algum grupo específico?”; “Há uma relação entre a ausência em uma coluna e valores em outra?”; “Algumas variáveis sempre faltam juntas?”
- Técnica de auditoria e diagnóstico de dados ausentes.
- Cria colunas binárias indicando se o valor original estava presente ou ausente.
- Ajuda a: identificar padrões de ausência (ex: "idade" ausente só quando "sexo" = "Feminino"?)

Casos Ausentes: Matriz Sombra

- Às vezes, a ausência em si tem significado! Exemplo: se "tempo de internação" está ausente, pode indicar que o paciente não foi internado.
- A matriz sombra ajuda a detectar quando os dados ausentes são:
- MCAR (Missing Completely At Random) – ausência aleatória
- MAR (Missing At Random) – ausência depende de outras variáveis
- MNAR (Missing Not At Random) – ausência depende do valor ausente (mais difícil)
- A matriz sombra é uma ferramenta poderosa para entender os dados ausentes com profundidade, indo além da contagem de NAs. Pode ser usada tanto para diagnóstico visual quanto como variável explicativa em modelos. Permite decisões melhores sobre imputação ou exclusão de dados.

Casos Ausentes: Imputação de Dados

- Substituição por média/mediana/moda
- Interpolação (para séries temporais)
- Modelos preditivos (ex: regressão, KNN)
- Multiple Imputation (mais avançado, para incerteza estatística)

Conceitos fundamentais

- Variável é uma quantidade, qualidade ou propriedade que você pode medir (colunas)
- Um valor é o estado de uma variável quando você a mede (“célula”)
- Uma observação é um conjunto de medições feitas em condições semelhantes (linhas)
- Dados tabulares são um conjunto de valores, cada um associado a uma variável e uma observação. Os dados tabulares estarão no formato tidy (arrumado) se cada valor estiver em sua própria “célula”, cada variável em sua própria coluna e cada observação em sua própria linha.

Exemplo

Nome	Disciplina	Nota	Status	Colégio
Priscila Silva	Português	7,85	Aprovado	Costa Monteiro
João Augusto	Português	8,95	Aprovado	Costa Monteiro
Maria Alice	Matemática	4,36	Reprovado	Costa Monteiro
Mateus Rocha	Matemática	10,0	Aprovado	Costa Monteiro
Alana Barbosa	Matemática	6,32	Reprovado	Costa Monteiro
Francisco Alves	Matemática	9,87	Aprovado	Costa Monteiro
Ana Laura Feitosa	Matemática	7,41	Aprovado	Costa Monteiro

Tipos de Variáveis

- Variável é a característica de interesse que é medida em cada elemento da amostra ou população.
- Qualitativas (categóricas): qualificam o objeto/indivíduo
  - Nominal: não existe hierarquia (cores, cor dos olhos, preso/solto, sexo, etc)
    - Variável Dummy (binária)
  - Ordinal: há hierarquia (grau de instrução, meses, patente, regime prisional)
- Quantitativas
  - Discretas: características mensuráveis que assumem valores inteiros (número de pessoas, número de bactérias, número de filhos, etc)
  - Contínuas: características mensuráveis que assumem valores em uma escala contínua. Valores fracionais são aceitos (peso, altura, tempo, idade, etc).

Base de dados

- Criação de uma base de dados
- Escrita das variáveis

Data de Nascimento
01 Data de Nascimento
Data_de_Nascimento

dt_nascimento
dt_nascimento_1

Operadores lógicos

Símbolo	Nome do Operador	Exemplo	Significado
>	Maior que	x > y	x é maior que y?
>=	Maior ou igual	x >= y	x é maior ou igual a y ?
<	Menor que	x < y	x é menor que y?
<=	Menor ou igual	x <= y	x é menor ou igual a y ?
==	Igualdade	x == y	x é igual a y?
!=	Diferente de	x != y	x é diferente de y?

Joins

left_join()

right_join()

inner_join()

full_join()

GOVERNO DE
PERNAMBUCO
ESTADO DE MUDANÇA