

Métodos de Estatística Aplicada com Python

Aula 3

Carlos Góes¹

¹Pós-Graduação em Ciência de Dados
Instituto de Educação Superior de Brasília

2017

Sumário

1 Introdução à visualização de dados

- O que fazer?
- O que não fazer?
- Mentira vs. erros
- Melhores práticas em visualização de dados

2 Formas gráficas comuns para representação de dados

- Valores absolutos: barras, linhas e colunas
- Valores relativos: pizza e área
- Representação de dispersão: intervalo interquartil, boxplot e histograma
- Representação de associação: diagrama de dispersão

3 Introdução às ferramentas gráficas em Python

- Introdução
- Exemplos de gráficos

Sumário

1 Introdução à visualização de dados

- O que fazer?
- O que não fazer?
- Mentira vs. erros
- Melhores práticas em visualização de dados

2 Formas gráficas comuns para representação de dados

- Valores absolutos: barras, linhas e colunas
- Valores relativos: pizza e área
- Representação de dispersão: intervalo interquartil, boxplot e histograma
- Representação de associação: diagrama de dispersão

3 Introdução às ferramentas gráficas em Python

- Introdução
- Exemplos de gráficos

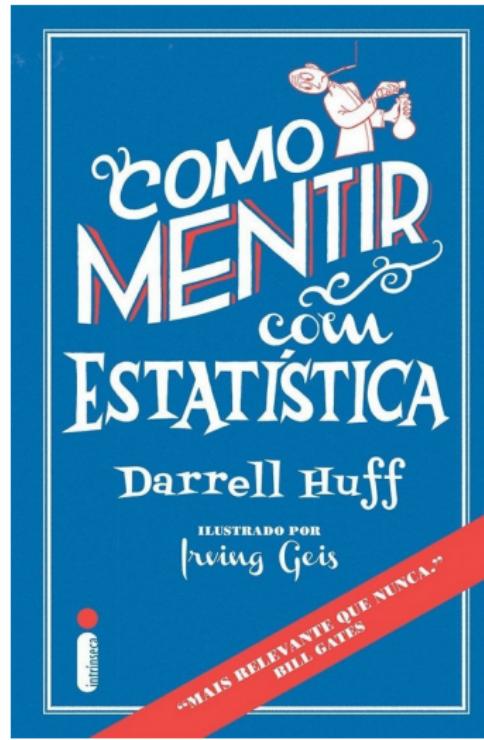
Introdução à visualização de dados

Qual é o propósito?

- “[O] PRINCIPAL PROPÓSITO *da visualização de dados* é *comunicar a informação de forma CLARA e EFICIENTE por meio gráfico* ...”
- “*para transmitir ideias de forma eficiente*, TANTO FUNCIONALIDADE QUANTO ESTÉTICA IMPORTAM, *porque é preciso organizar DADOS COMPLEXOS e comunicar seus aspectos principais de uma FORMA INTUITIVA*”
- Vitaly Friedman (2008). “Data Visualization and Infographics” Smashing Magazine.

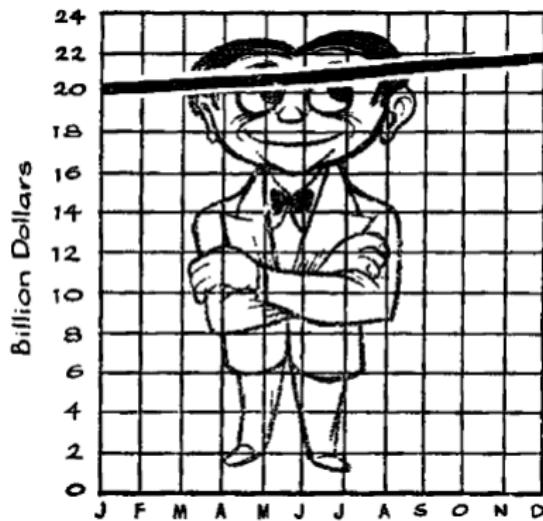
Introdução à visualização de dados

O que não fazer?



Introdução à visualização de dados

Ex: crescimento da receita de 10%



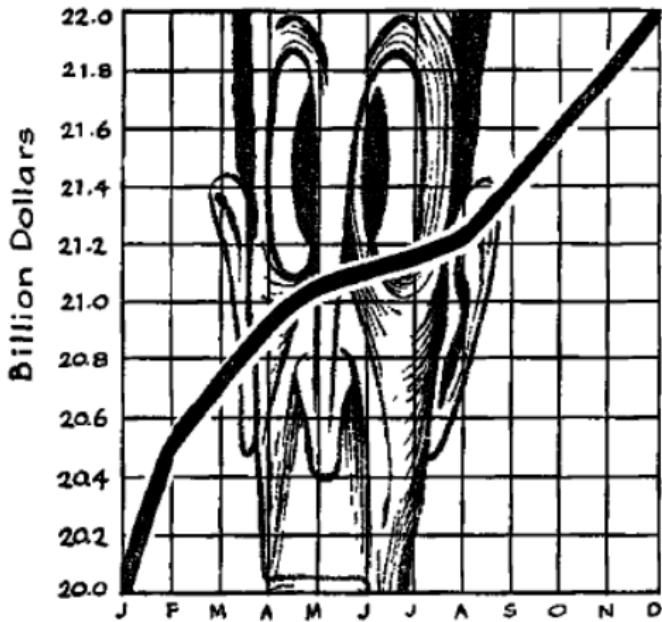
Introdução à visualização de dados

Ex: crescimento da receita de 10%!



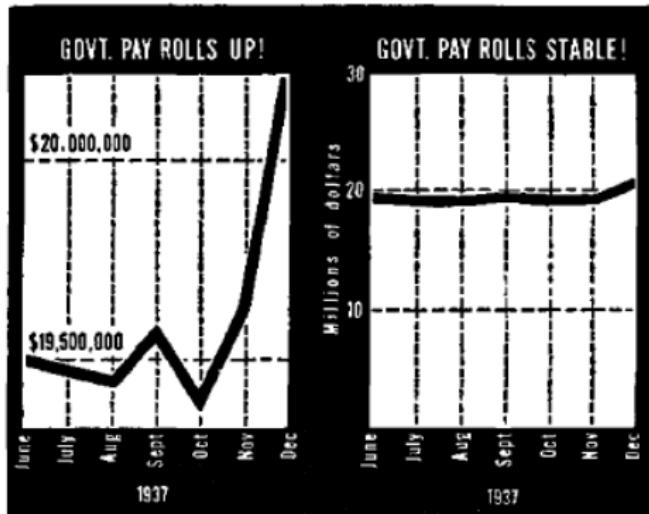
Introdução à visualização de dados

Ex: crescimento da receita de 10%!!!



Introdução à visualização de dados

Ex: gastos de pessoal do governo - crescendo ou estáveis?



Introdução à visualização de dados

Ex: Resultado das Eleições na Venezuela



Introdução à visualização de dados

Ex: Resultado das Eleições na Venezuela

5 Graphics Lies, Misleading Visuals

109

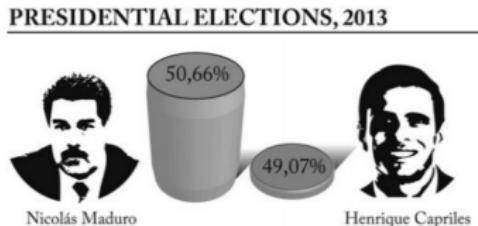


Fig. 5.3 Presidential election results in Venezuela, based on a graphic by Venezolana de Televisión. Notice the truncated Y-axis which greatly distorts the difference between the percentages of vote

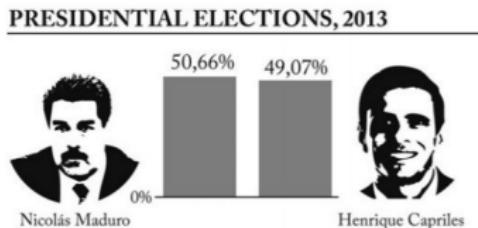
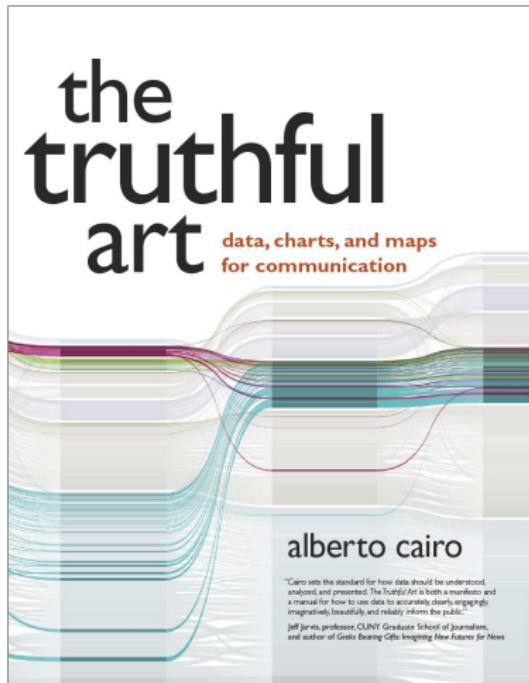


Fig. 5.4 An alternative version of the previous graphic in which a 0-baseline has been added, and the 3D effect has been removed

Introdução à visualização de dados

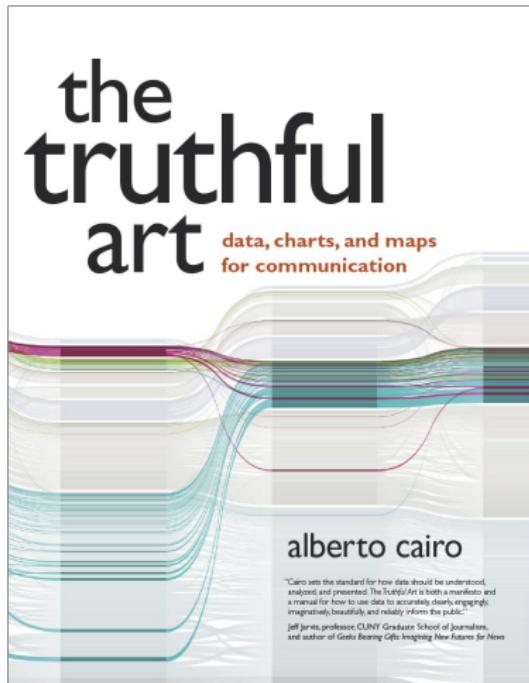
Mentira vs. erros



- Mentira é uma violação ética
 - Gráficos não mentem, são pessoas que mentem
- Nem todo erro, contudo, é uma violação ética.
 - Um gráfico pode distorcer a realidade por causa de erros de boa fé. Isso é éticamente neutro.

Introdução à visualização de dados

Mentira vs. erros



- Três formas de mentir com gráficos:
 - ➊ Esconder dados relevantes para mostrar o que nos beneficia;
 - ➋ Expor muitos dados para tornar a realidade incompreensível;
 - ➌ Usar formas gráficas inapropriadas (distorcer os dados).

Introdução à visualização de dados

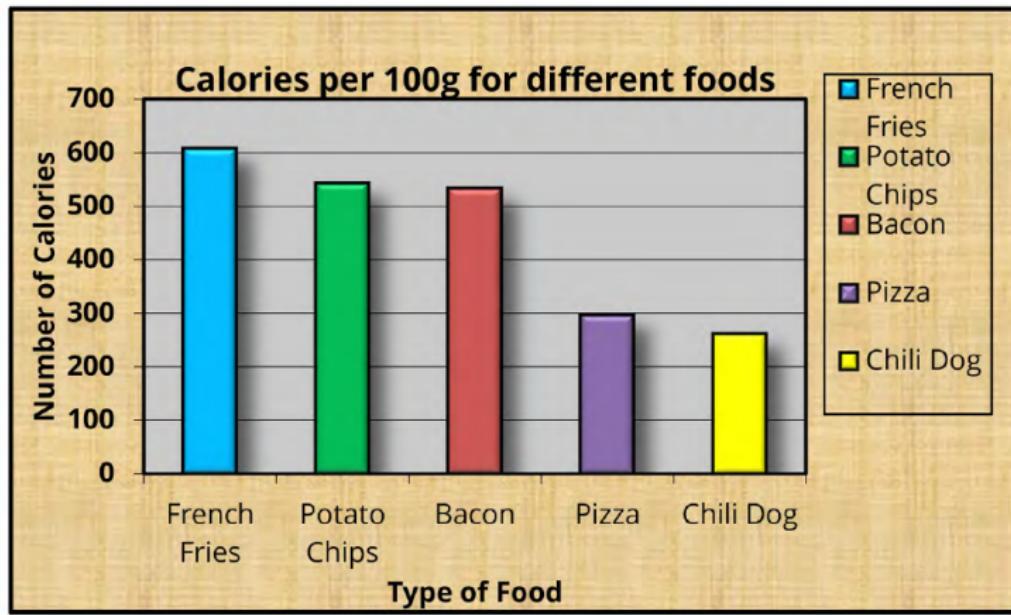
Melhores práticas em visualização de dados

- Retirar poluição para melhorar a visualização
- Dark Horse Analytics: menos é mais

Introdução à visualização de dados

Melhores práticas em visualização de dados

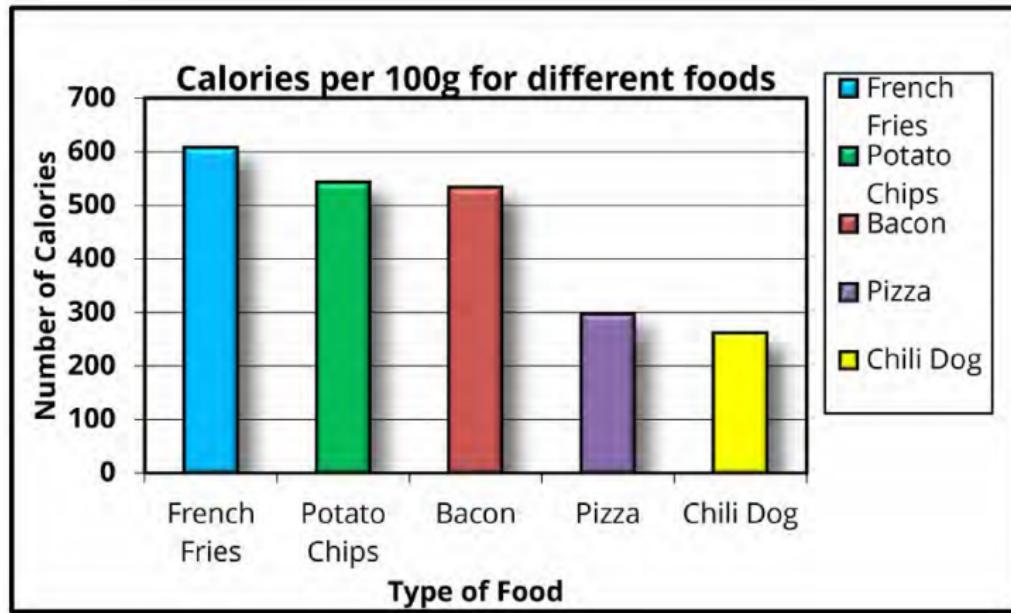
1. Remover coloração de fundo



Introdução à visualização de dados

Melhores práticas em visualização de dados

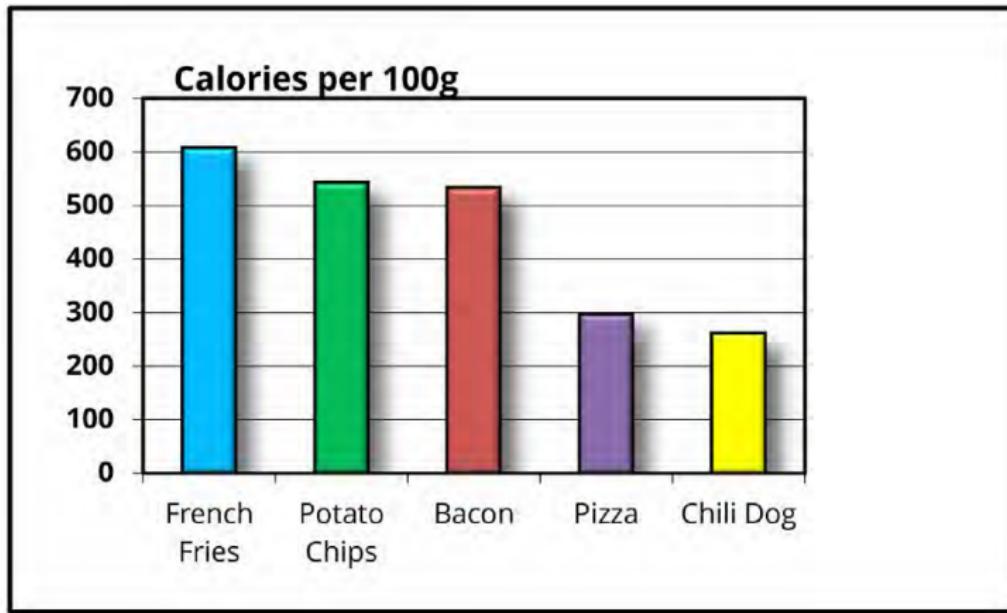
2. Remover redundâncias



Introdução à visualização de dados

Melhores práticas em visualização de dados

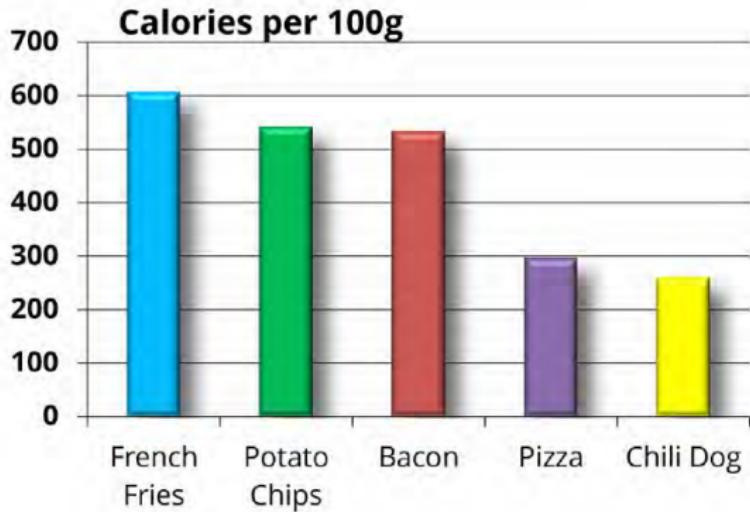
3. Remover bordas



Introdução à visualização de dados

Melhores práticas em visualização de dados

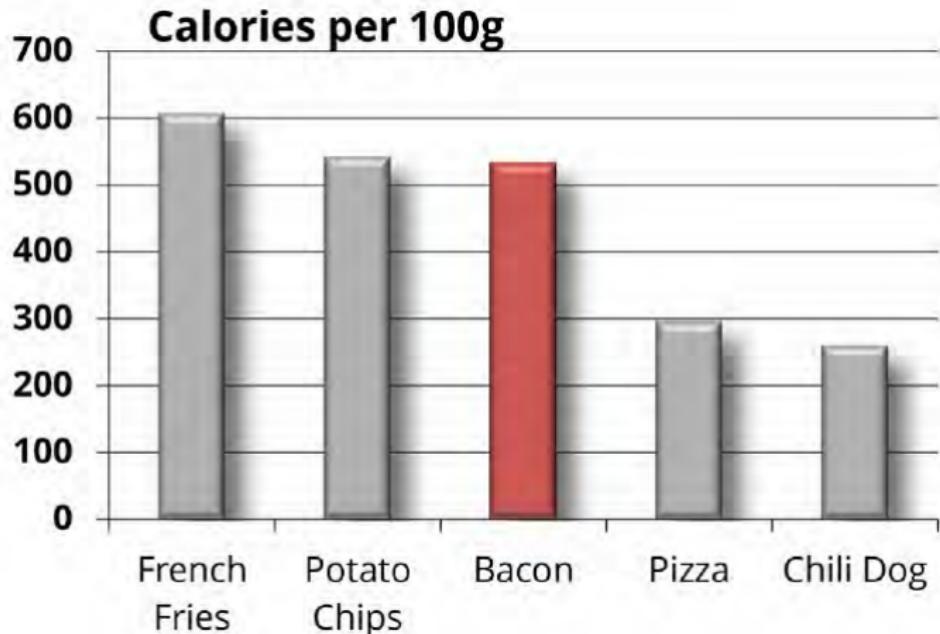
4. Reduzir o número de cores



Introdução à visualização de dados

Melhores práticas em visualização de dados

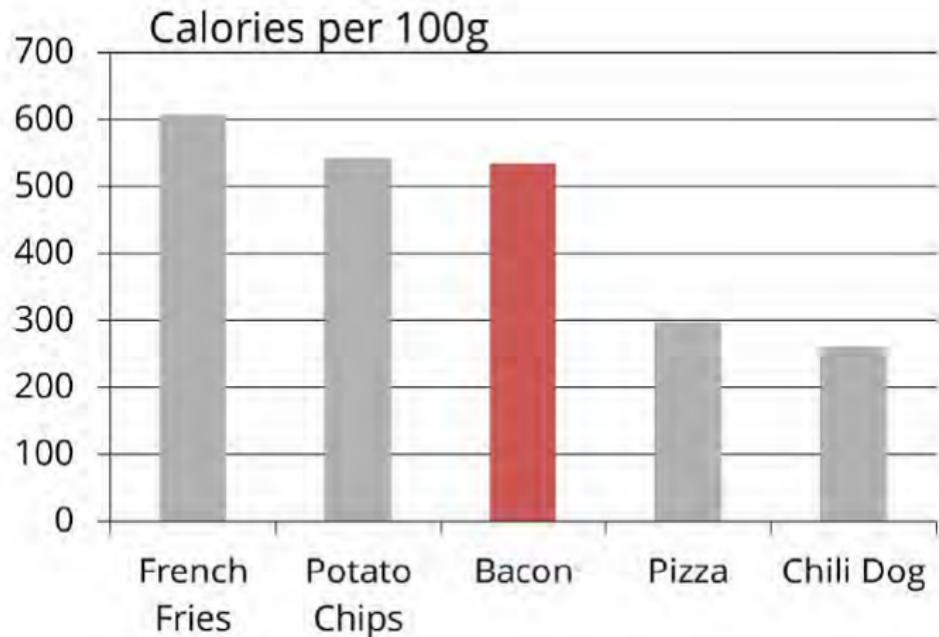
5. Remover efeitos especiais



Introdução à visualização de dados

Melhores práticas em visualização de dados

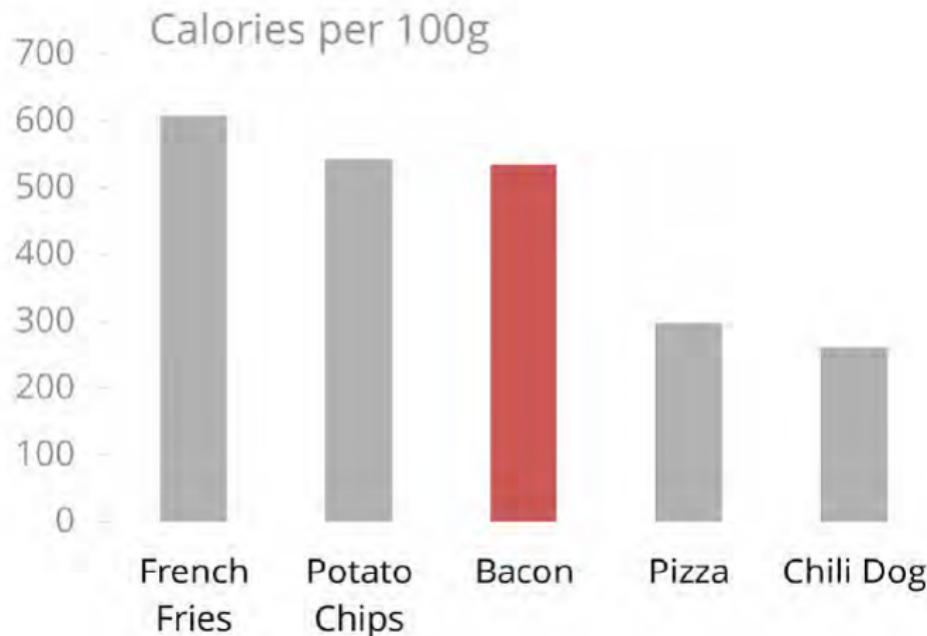
6. Tornar rótulos mais leves



Introdução à visualização de dados

Melhores práticas em visualização de dados

7. Rotular diretamente (sem eixos)

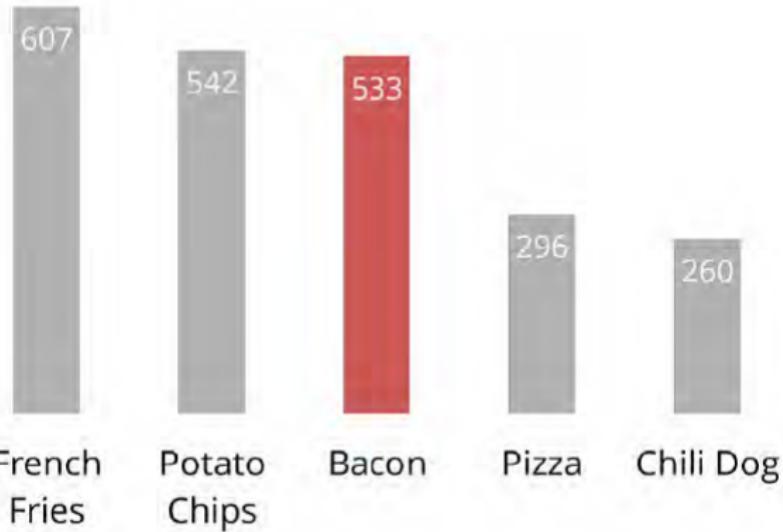


Introdução à visualização de dados

Melhores práticas em visualização de dados

8. Final

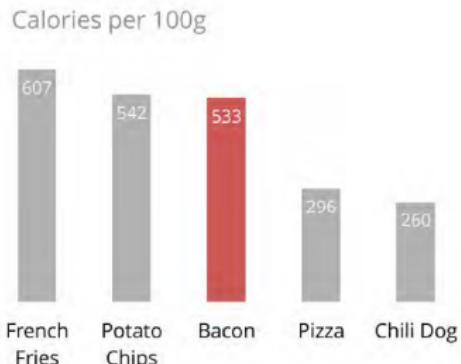
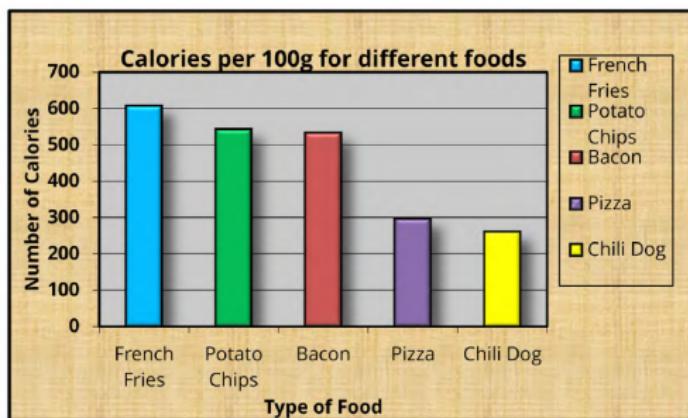
Calories per 100g



Introdução à visualização de dados

Melhores práticas em visualização de dados

9. Comparação



Sumário

1 Introdução à visualização de dados

- O que fazer?
- O que não fazer?
- Mentira vs. erros
- Melhores práticas em visualização de dados

2 Formas gráficas comuns para representação de dados

- Valores absolutos: barras, linhas e colunas
- Valores relativos: pizza e área
- Representação de dispersão: intervalo interquartil, boxplot e histograma
- Representação de associação: diagrama de dispersão

3 Introdução às ferramentas gráficas em Python

- Introdução
- Exemplos de gráficos

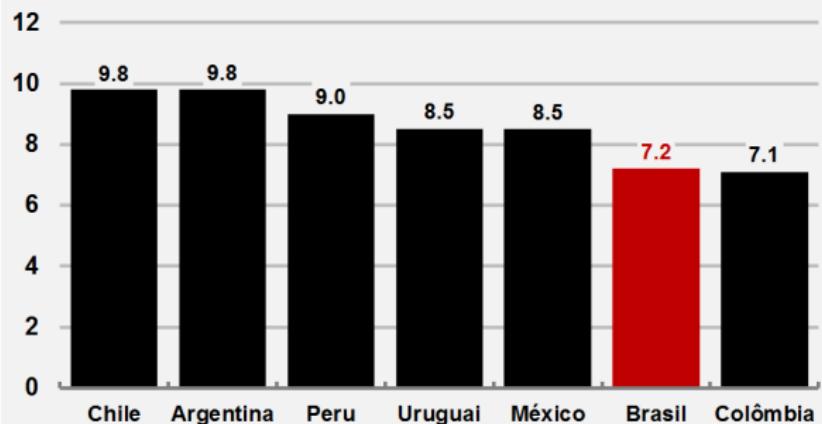
Formas gráficas comuns para representação de dados

Barras, linhas e colunas

Barras/colunas simples

FIGURA 3. AMÉRICA LATINA: ANOS DE ESTUDO DA POPULAÇÃO

(Média de anos de estudo da população com mais de 25 anos, dados de 2013)



Fonte: Programa das Nações Unidas para o Desenvolvimento.



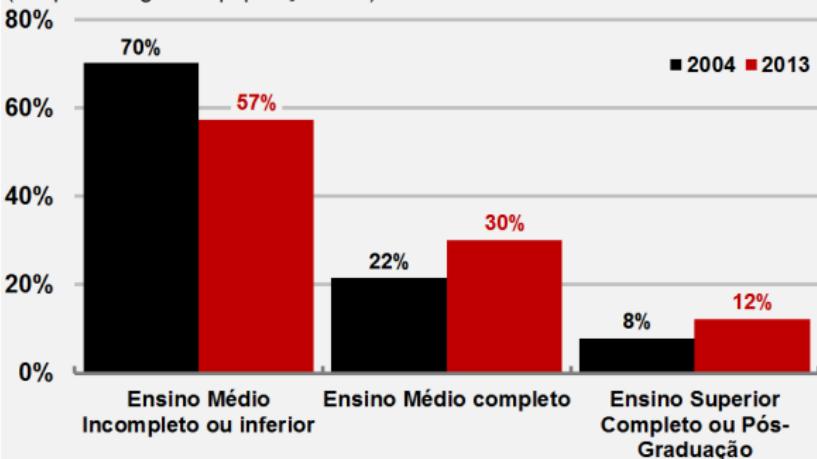
Formas gráficas comuns para representação de dados

Barras, linhas e colunas

Barras/colunas múltiplas

FIGURA 1. BRASIL: NÍVEL EDUCACIONAL DA POPULAÇÃO

(Em porcentagem da população total)



Fontes: IBGE, Pesquisas Nacional de Amostra de Domicílios (PNADs) de 2004 e 2013.

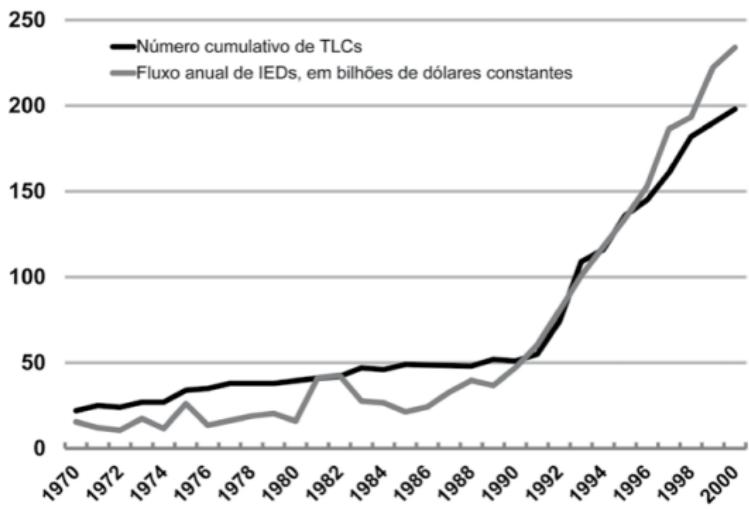


Formas gráficas comuns para representação de dados

Barras, linhas e colunas

Quando há muita informação: linhas (Imagine esse mesmo gráfico com barras múltiplas)

Países em Desenvolvimento: Número de Acordos de Livre Comércio e Fluxos de IEDs
(Número cumulativo de tratados e fluxos anuais em bilhões de dólares constantes)



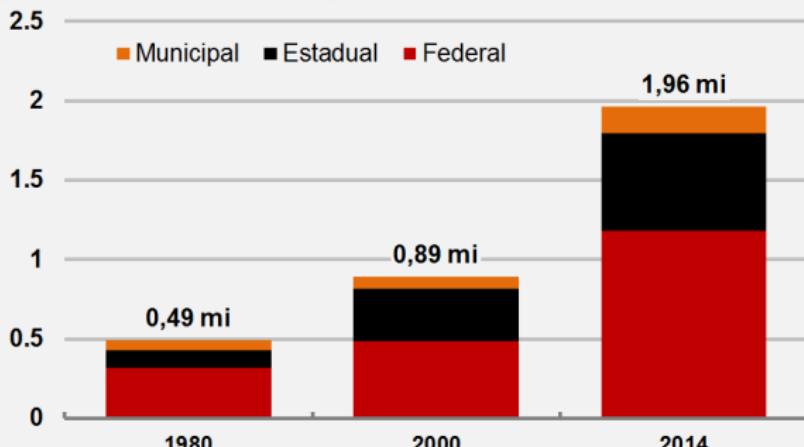
Formas gráficas comuns para representação de dados

Barras, linhas e colunas

Barras empilhadas

FIGURA 5. BRASIL: VAGAS EM UNIVERSIDADES PÚBLICAS

(Em milhões de matrículas no ano)



Fonte: Ministério da Educação / INEP.



Formas gráficas comuns para representação de dados

Pizza e área

Gráfico de pizza (relativo a 100%)

BRASIL: DESTINAÇÃO DAS RECEITAS FEDERAIS, 2013 (em porcentagem das receitas totais)

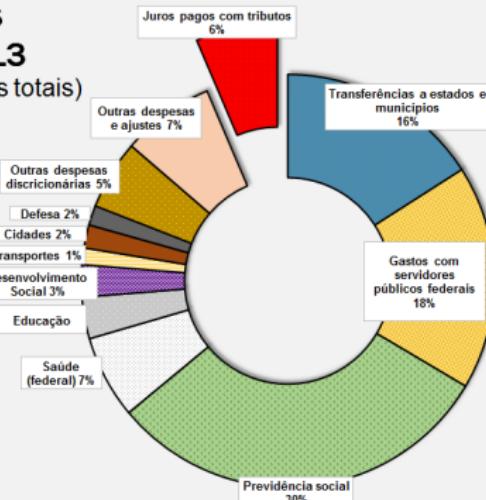
Segundo a Auditoria Cidadã da Dívida, o governo destina quase 50% do Orçamento federal pra financiamento da dívida pública. Isso é distorção deliberada do que é a verdade. Eles aglomeraram duas coisas diferentes na mesma conta: (a) rolagem da dívida; e (b) serviço e amortização.

A gente organizou os dados para ver qual era a estatística verdadeira. Em 2013, cerca de 6% dos seus impostos foram usados para saldar juros da dívida. E um bocado, mas **não é** quase 50% do orçamento, como alguns tentam fingir.

A dívida pública e suas consequências pra gerações futuras merecem um debate sério. Mas um debate sério deve ser calculado em números sérios - e não em fantasia estatística para apoiar um discurso político.

Fontes: MPOG e Banco Central.

Nota: Inclui a fração da conta de juros e amortização paga com receita primária, ou seja, com impostos.



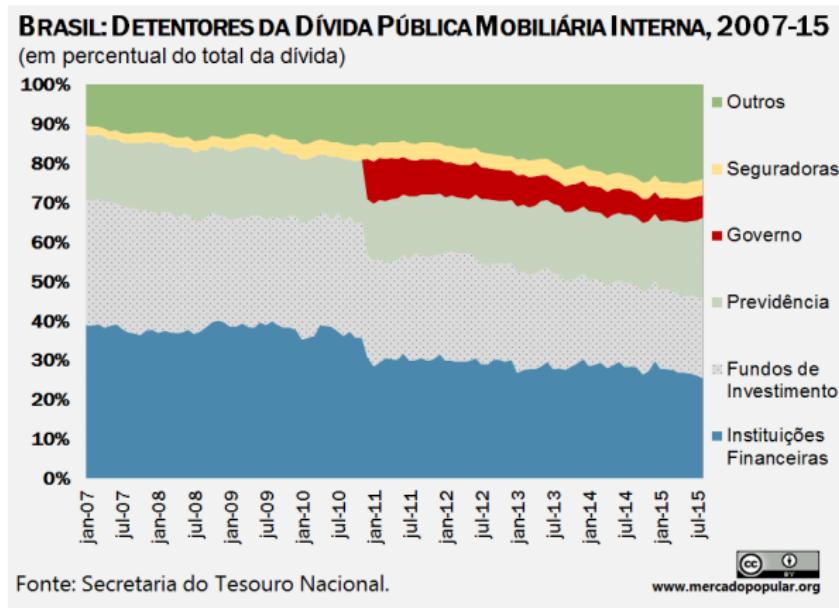
www.mercadopopular.org



Formas gráficas comuns para representação de dados

Pizza e área

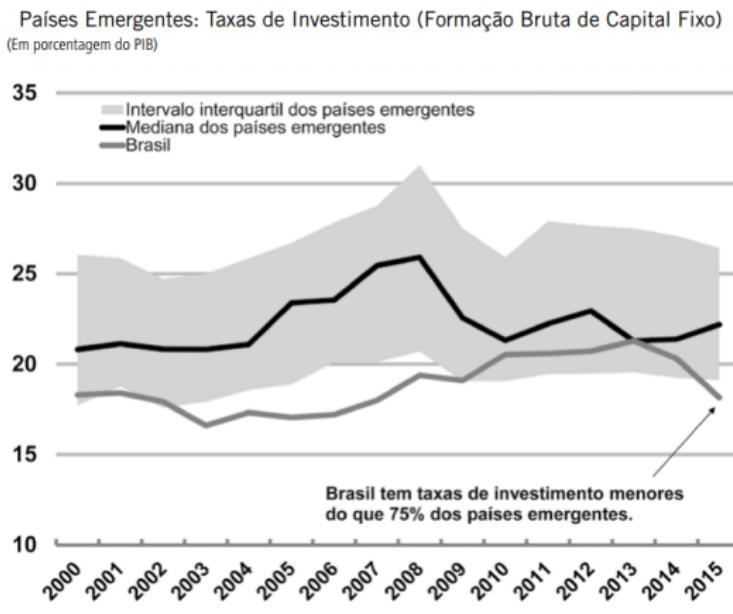
Gráfico de área (relativo a 100%)



Formas gráficas comuns para representação de dados

Intervalo interquartil, boxplot e histograma

Linhas com intervalo interquartil



Fonte: Cálculos do autor dos dados de Haver Analytics e World Economic Outlook/FMI. Inclui 53 países considerados como mercados emergentes pelo FMI.

Formas gráficas comuns para representação de dados

Intervalo interquartil, boxplot e histograma

Boxplot (Ex 1)

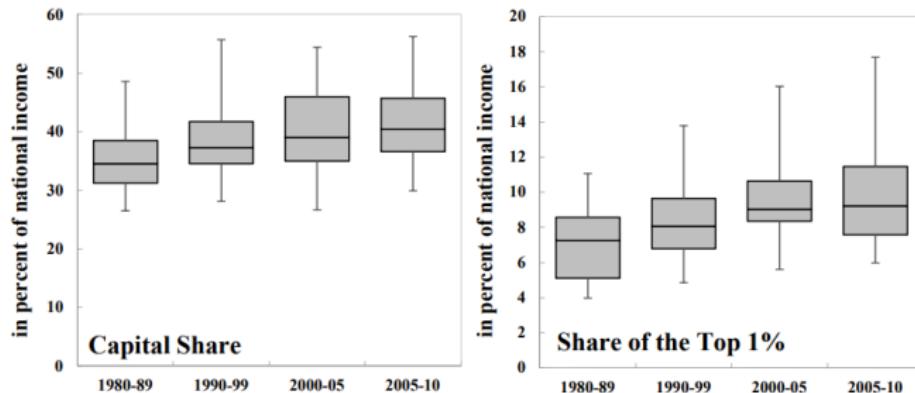
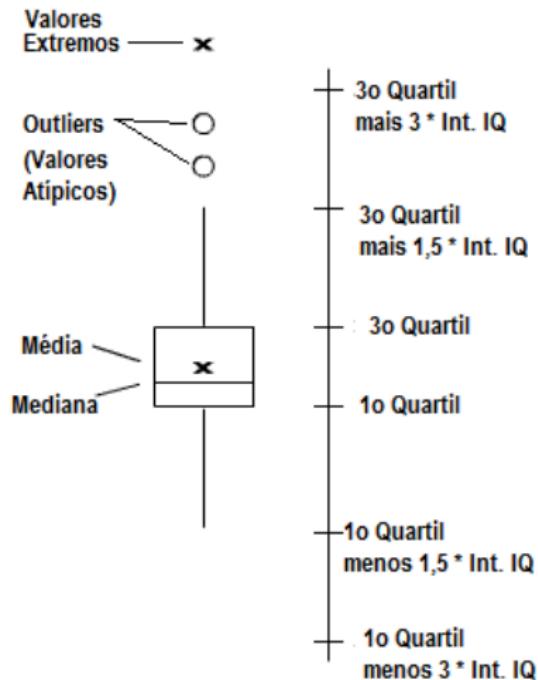


Figure 1: **Distribution of capital share and share of the top 1% over time.** Y-axis in percent, x-axis represents period averages. The sample refers to an unbalanced panel of 19 advanced economies ranging from 1981-2010. Boxplots show interquartile ranges and medians. Whiskers show minimums and maximums.

Formas gráficas comuns para representação de dados

Intervalo interquartil, boxplot e histograma

Boxplot (Ex 2)

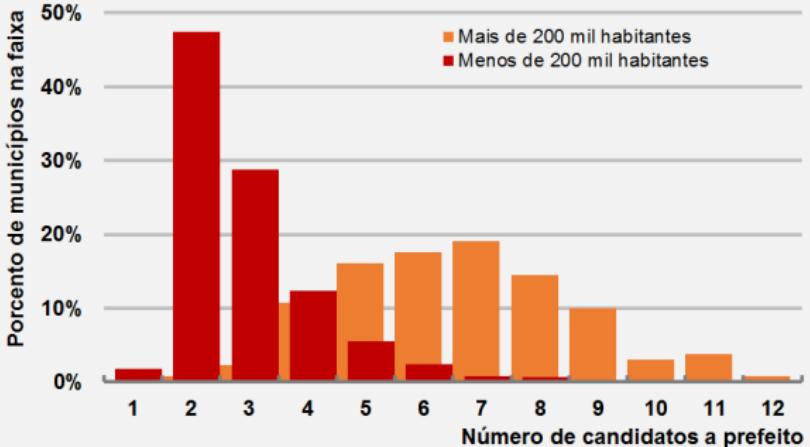


Formas gráficas comuns para representação de dados

Intervalo interquartil e boxplot

Histograma

FIGURA 6. ELEIÇÕES 2016: DISTRIBUIÇÃO DE CANDIDATURAS A PREFEITO (Distribuição percentual de municípios por número de candidaturas)

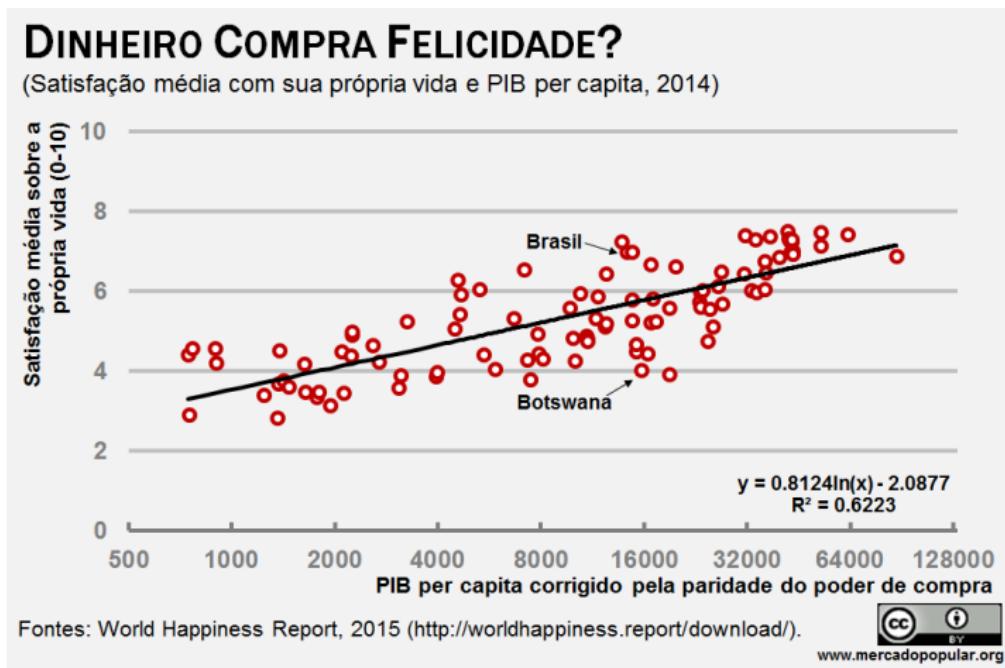


Fonte: Cálculos dos autores com microdados do Tribunal Superior Eleitoral e do IBGE.

Formas gráficas comuns para representação de dados

Diagrama de dispersão

Diagrama de dispersão (scatterplot)



Sumário

1 Introdução à visualização de dados

- O que fazer?
- O que não fazer?
- Mentira vs. erros
- Melhores práticas em visualização de dados

2 Formas gráficas comuns para representação de dados

- Valores absolutos: barras, linhas e colunas
- Valores relativos: pizza e área
- Representação de dispersão: intervalo interquartil, boxplot e histograma
- Representação de associação: diagrama de dispersão

3 Introdução às ferramentas gráficas em Python

- Introdução
- Exemplos de gráficos

Introdução às ferramentas gráficas em Python

Introdução ao matplotlib e seaborn

- matplotlib e seaborn são duas das principais ferramentas em Python!
- Para quem usa R, há também uma versão de ggplot2, com a mesma sintaxe (grammar of graphics), em Python!
- A primeira coisa a fazer é importar essas duas ferramentas:

```
import matplotlib.pyplot as plt  
import seaborn as sns
```

Introdução às ferramentas gráficas em Python

Introdução ao matplotlib e seaborn

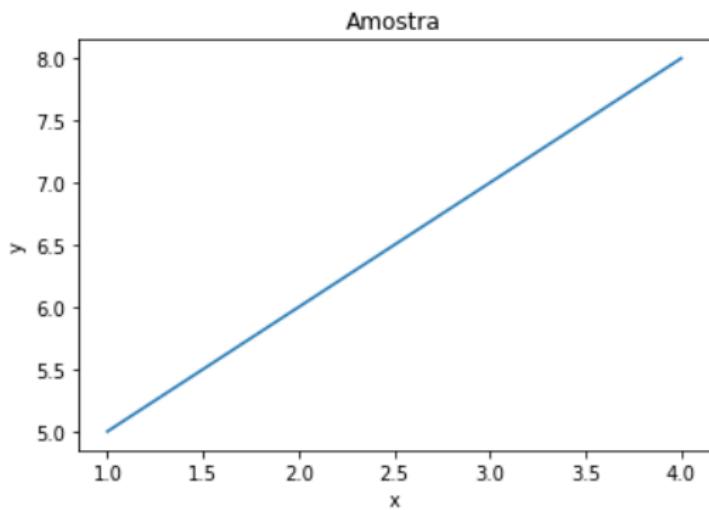
- A ideia fundamental do matplotlib: código é rápido, gráficos são demorados.
- Por isso, a forma como ele funciona é que primeiro definimos todas as características do gráfico e somente depois pedimos para máquina criá-la:

```
x = [1,2,3,4]
y = [5,6,7,8]
plt.plot(x, y)
plt.xlabel('x')
plt.ylabel('y')
plt.title('Amostra')
plt.show()
```

Introdução às ferramentas gráficas em Python

Introdução ao matplotlib e seaborn

Desenho da máquina



Exemplos de gráficos

Linha

- Vamos construir um gráfico com duas linhas no `matplotlib`.
- Primeiro, precisamos criar nossos dados.

```
import numpy as np
anos = np.linspace(2001, 2010, 10)
y1 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
y2 = [i+10 for i in y1]
```

Exemplos de gráficos

Linha

- Depois, adicionamos as linhas:

```
plt.plot(anos, y1, label="y1", color="red")
plt.plot(anos, y2, label="y2", color="black")
```

- Os rotulos:

```
plt.legend(loc="upper left")
plt.ylabel("Valores")
plt.xlabel("Anos")
plt.title("Duas linhas")
```

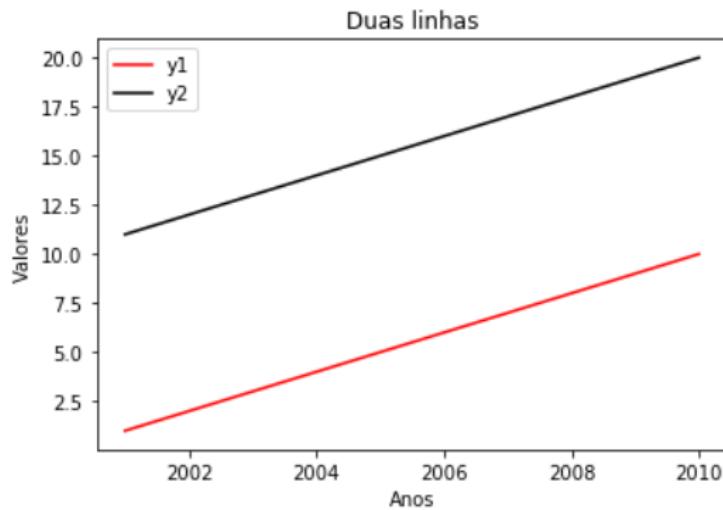
- E plotamos:

```
plt.show()
```

Exemplos de gráficos

Linha

Desenho da máquina:



Exemplos de gráficos

Linha

- Vamos fazer a mesma coisa no seaborn matplotlib.
- Primeiro, precisamos levar nossos dados para um pandas DataFrame.

```
import pandas as pd
data = {'y1': y1, 'y2': y2}
df = pd.DataFrame(data, index=anos)
df = df.unstack().reset_index()
df.columns = ['variavel', 'data', 'valor']
```

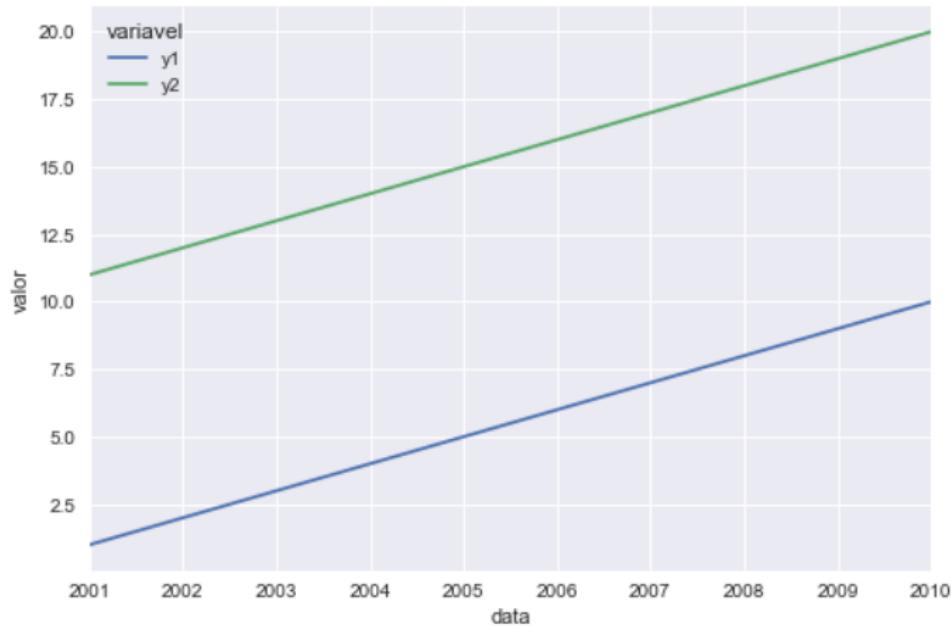
- Depois, conseguimos plotar em uma linha:

```
sns.tsplot(df, time='data', unit='variavel',
            condition='variavel', value='valor')
```

Exemplos de gráficos

Linha

Desenho da máquina:



Exemplos de gráficos

Pizza

- Carregamos nossos dados.

```
rotulos = ['Feijão', 'Arroz', 'Carne',
           'Farofa', 'Batata Frita', 'Outros']
respostas = [200, 300, 100,
            150, 200, 100]
```

- Depois, criamos o grafico:

```
plt.pie(respostas, labels=rotulos,
         autopct='%1.1f%%')
plt.title('O que está no almoço brasileiro?')
plt.show()
```

Exemplos de gráficos

Pizza

Desenho da máquina:



Exemplos de gráficos

Barras

- Carregamos nossos dados.

```
rotulos = ['Feijão', 'Arroz', 'Carne',
           'Farofa', 'Batata Frita', 'Outros']
respostas = [200, 300, 100,
            150, 200, 100]
```

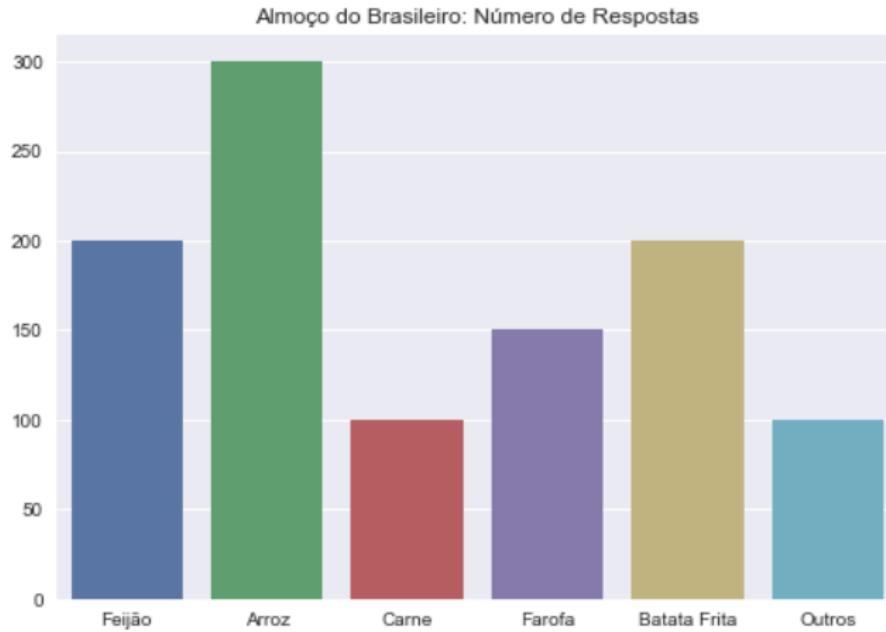
- Depois, criamos o gráfico:

```
sns.barplot(rotulos, respostas).
set_title('Almoço do Brasileiro: Número de Respostas')
```

Exemplos de gráficos

Barra

Desenho da máquina:



Exemplos de gráficos

Boxplot

- Carregamos nossos dados.

```
url = "https://raw.githubusercontent.com/ \
omercadopopular/cgoes/ \
master/piketty/fdatabasetax.csv"
```

- Depois, criamos o gráfico:

```
plt.xticks(rotation=45)
plt.title('Países Avançados: Fração da Renda Nacional
apropriada pelo 1% mais rico')
sns.boxplot(x='year', y='top1', data=piketty)
```

Exemplos de gráficos

Barra

Desenho da máquina:

