Vizualização de dados com Matplotlib e Seaborn

Programa de Pós-Graduação em Engenharia Elétrica e de Computação (PPGEEC)
Introdução à Ciências de Dados - UFC *Campus* Sobral – 2023.1
Andressa Gomes Moreira – andressagomes@alu.ufc.br

Sumário

Visualização de Dados

- Uma parte fundamental do kit de ferramentas do cientista de dados é a visualização de dados.
- A visualização de dados é parte arte e parte ciência.
- Existem dois usos primários para a visualização de dados:
  - Para explorar dados
  - Para comunicar dados

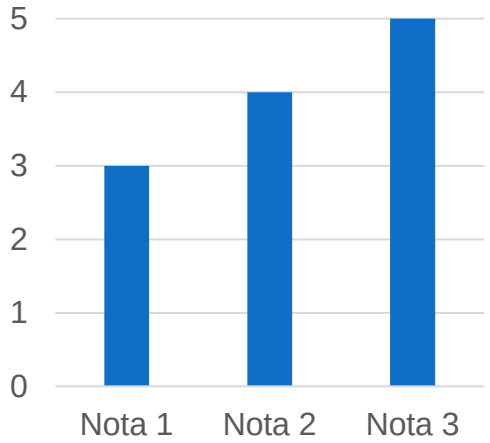
Visualização de Dados

- O que é mais importante: transmitir os dados com precisão ou a estética com que a informação é transmitida?

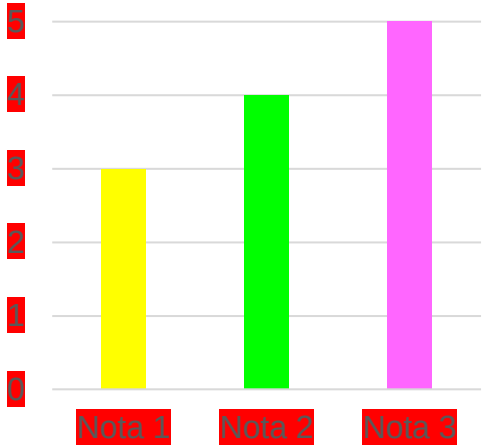
Visualização de Dados

- O que é mais importante: transmitir os dados com precisão ou a estética com que a informação é transmitida?
- Uma visualização de dados, antes de mais nada, deve transmitir os dados com precisão e ao mesmo tempo, uma visualização de dados deve ser esteticamente agradável.

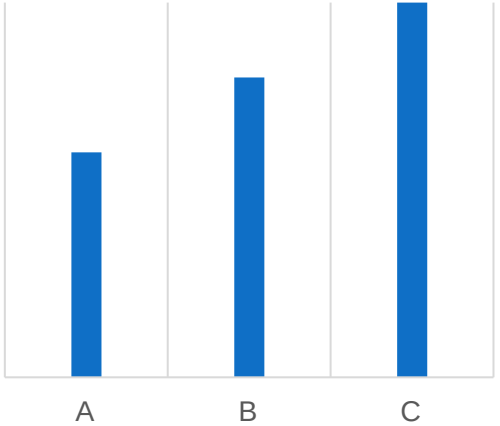
Visualização de Dados



(a) Gráfico 1



(b) Gráfico 2



(c) Gráfico 3

Gráficos e Plotagens

- Existem vários gráficos e plotagens comumente usados para visualizar dados:
  - **Visualizar valores**: Gráficos de Barras, Mapas de Calor;
  - **Distribuições**: Histogramas, Boxplots;
  - **Proporções**: Gráficos de Pizza;
  - **Relações x – y**: Gráficos de Dispersão, Gráficos de Linha.

Gráfico de Barras

- Visualização de quantidades
- Consiste em um conjunto de categorias e um valor quantitativo para cada categoria.
- Existem diversas variações do gráfico de barras: barras simples, barras agrupadas e empilhadas.
- A ordem em que as barras são dispostas é importante.

Gráfico de Barras

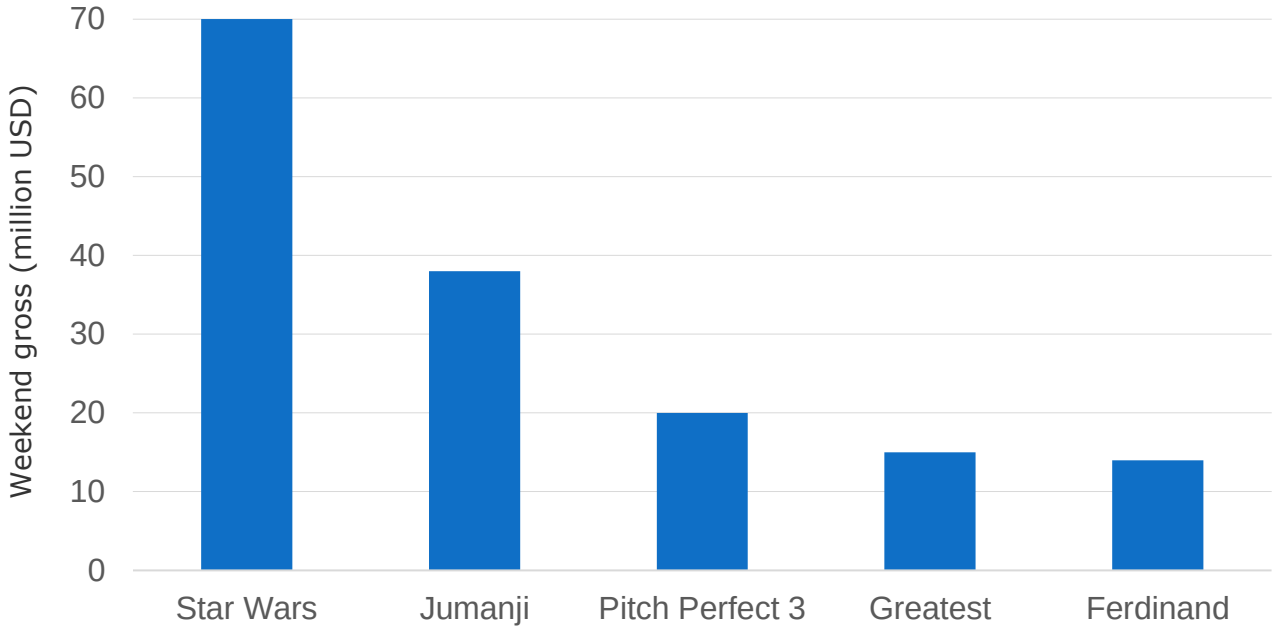


Figura - Filmes de maior bilheteria no fim de semana de 22 a 24 de dezembro de 2017

Mapa de Calor

- Representação gráfica de dados em que cada valor de uma matriz é representado por uma cor.
- Apresenta uma correlação entre todas as variáveis numéricas no conjunto de dados.
- Útil para destacar tendências mais amplas.

Mapa de Calor

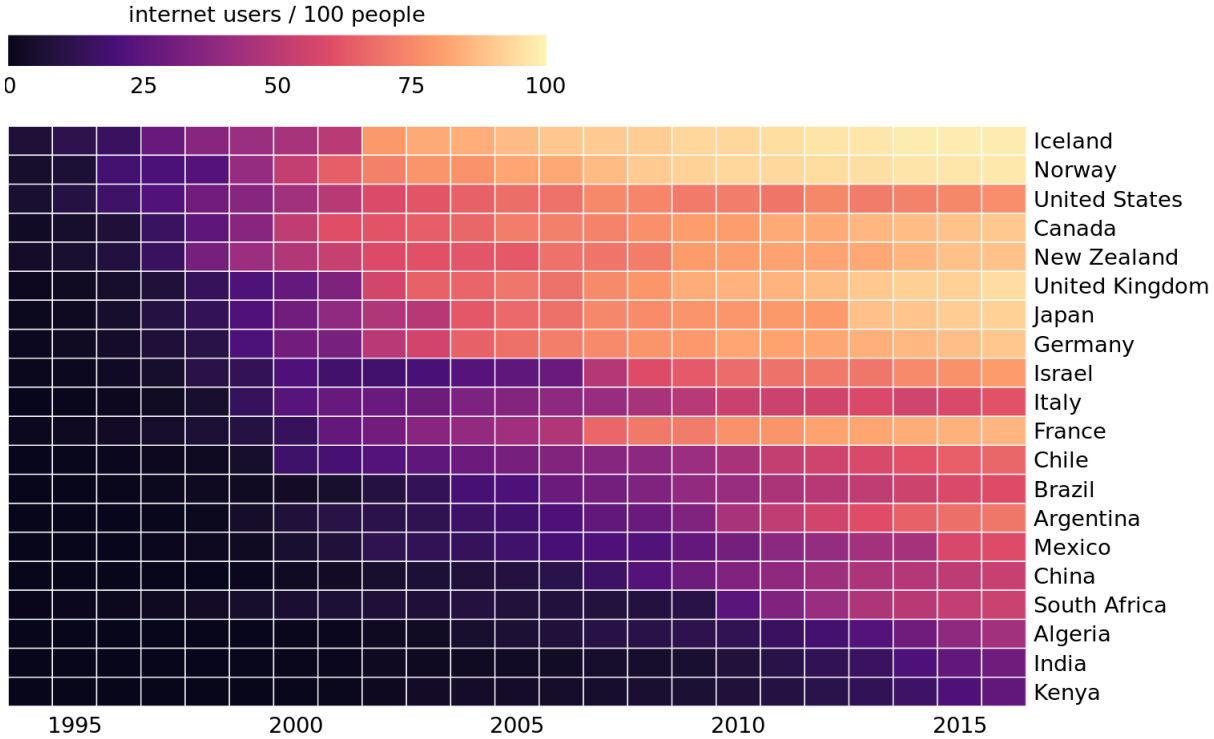


Figura - Porcentagem de usuários da Internet para o respectivo país e ano.

Histograma

- Demonstra a distribuição de frequências de uma variável em um conjunto de dados
- A base de cada uma das barras representa uma classe e a altura representa a quantidade ou frequência absoluta com que o valor de cada classe ocorre.

Histograma

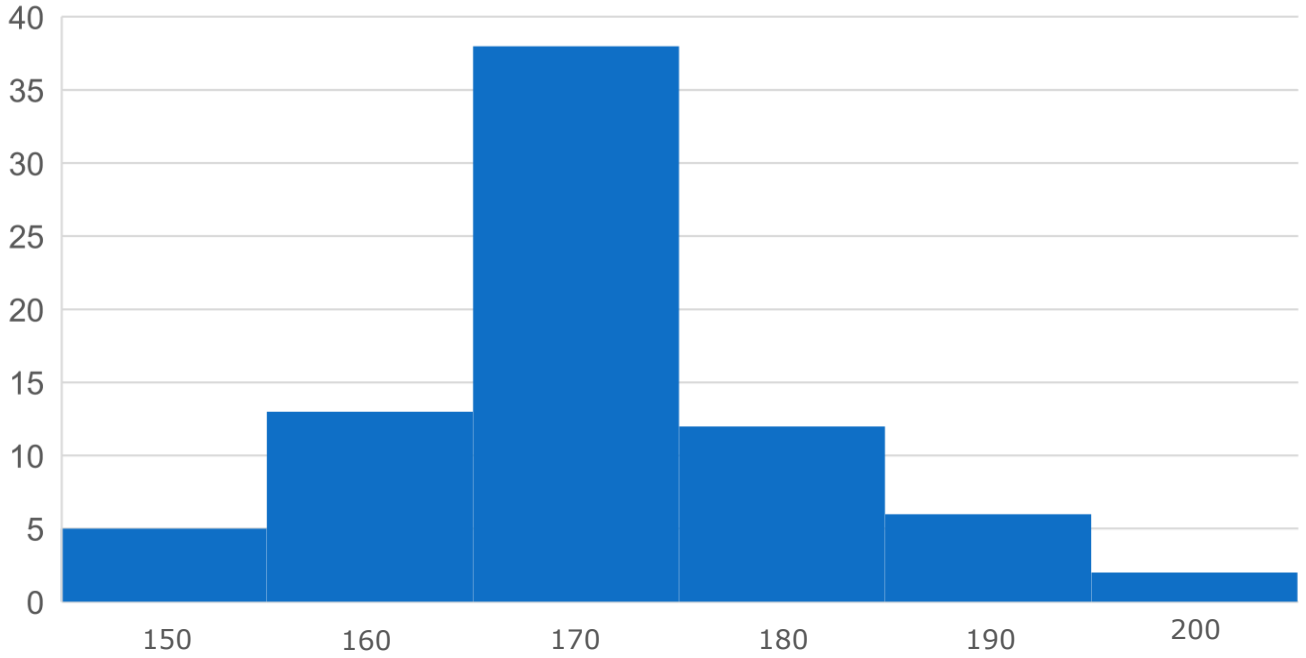
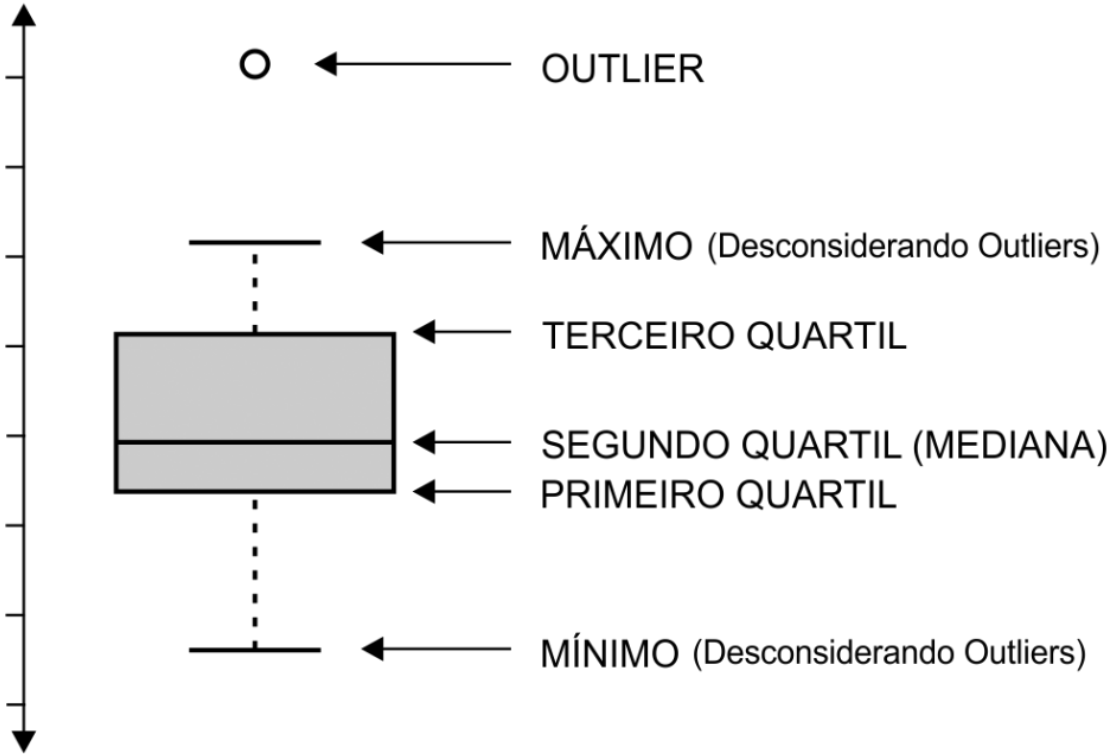


Figura – Histograma da altura (cm) de alunos.

Boxplot

- Diagrama de Caixas
- Utilizado para visualizar a distribuição e valores discrepantes dos dados.
- Um boxplot divide os dados em quartis e os visualiza de maneira padronizada.

Boxplot


OUTLIER
MÁXIMO (Desconsiderando Outliers)
TERCEIRO QUARTIL
SEGUNDO QUARTIL (MEDIANA)
PRIMEIRO QUARTIL
MÍNIMO (Desconsiderando Outliers)

Boxplot

Idade
18
19
21
21
21
22
22
22
23
23
24
27

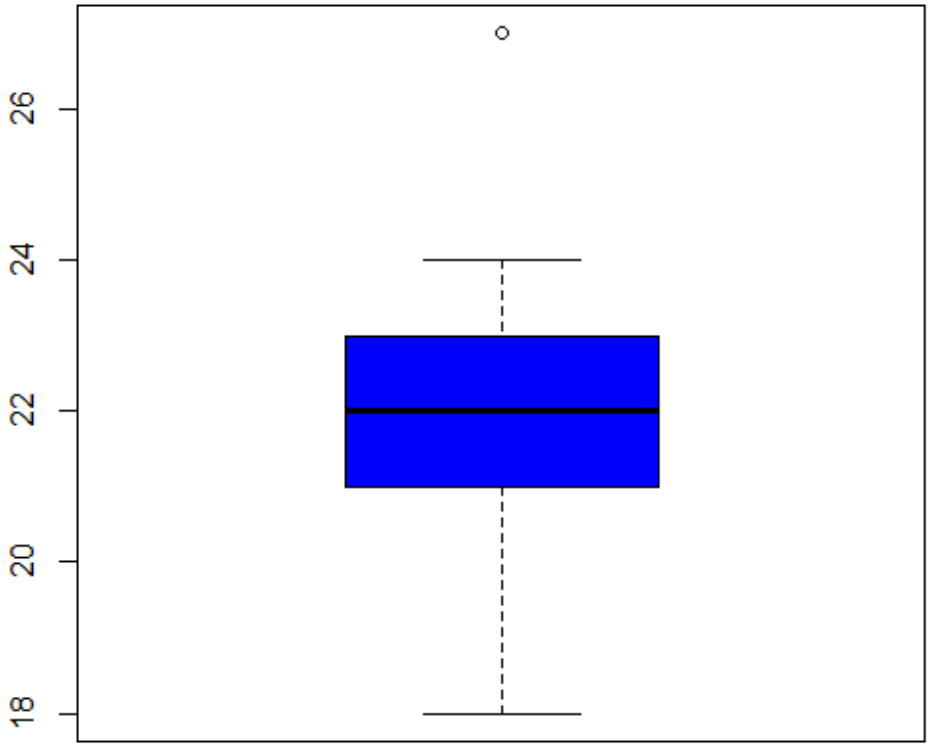


Gráfico de Dispersão

- A dispersão indica o quão "espalhados" ou variados são os valores de uma amostra.
- Os gráfico de dispersão são úteis para mostrar como duas variáveis se relacionam entre si.
- Servem para avaliar a consistência dos dados e identificar possíveis padrões ou outliers.
- Cada par (x,y) é um ponto no gráfico.

Gráfico de Dispersão

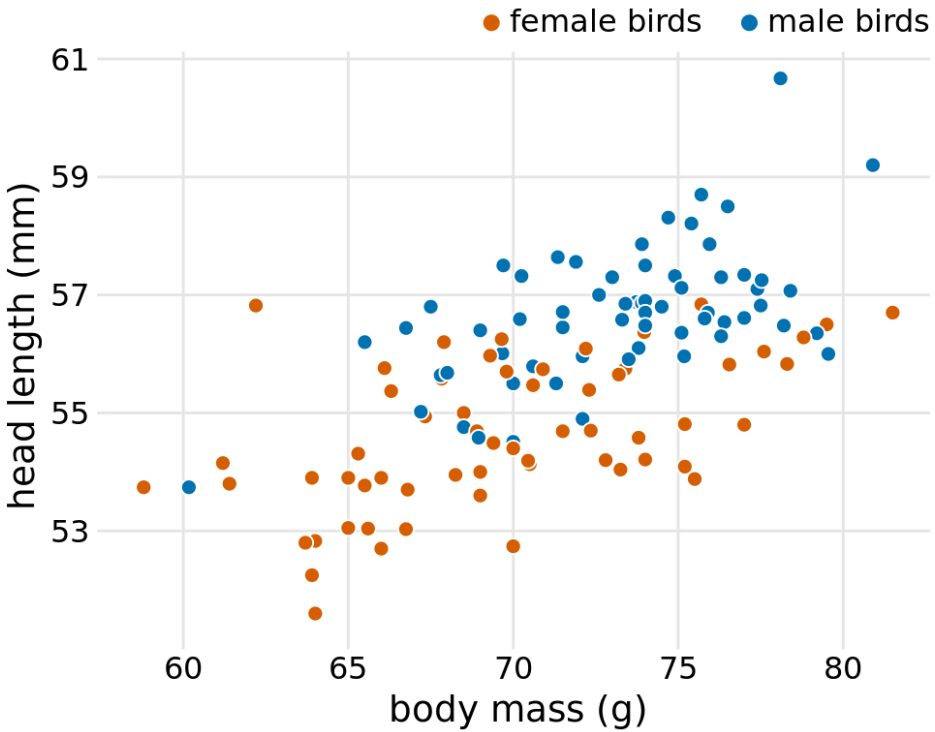


Figura - Comprimento da cabeça (mm) versus massa corporal (g) de pássaros gaios-azuis.

Gráfico de Linha

- Representa de tendências ao longo de um período de tempo.
- São apropriados sempre que uma variável impõe uma ordenação nos dados.

Gráfico de Linha

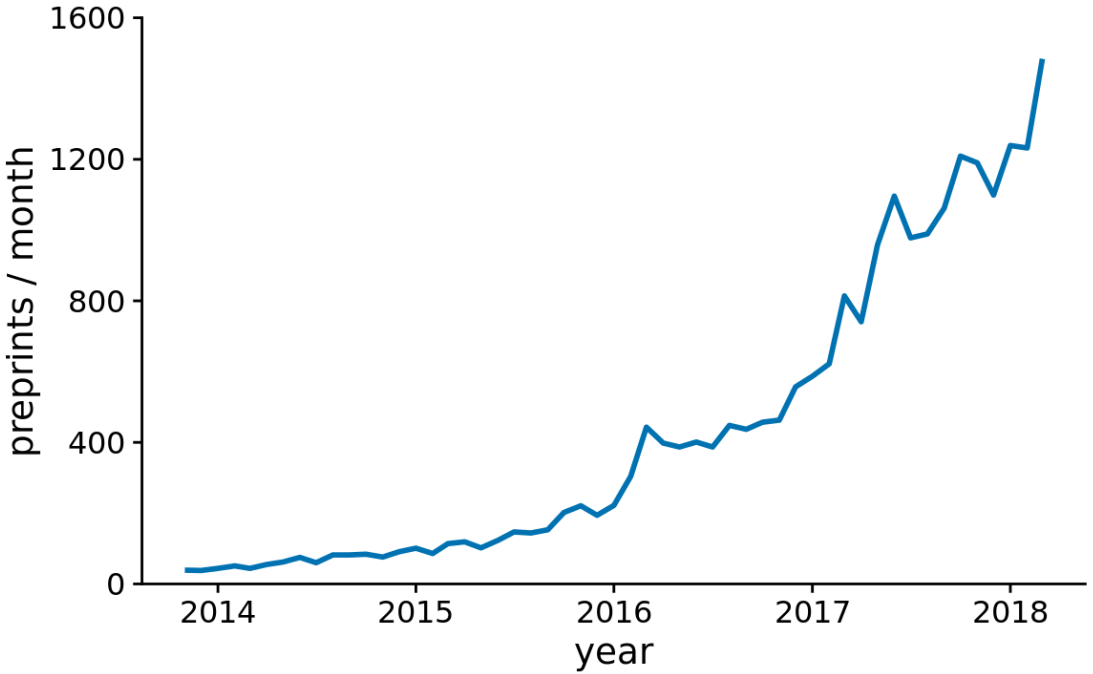


Figura - Envios mensais para o servidor de pré-impressão bioRxiv

Principais Pacotes

matplotlib

seaborn

1

Matplotlib

http://matplotlib.org/

Matplotlib

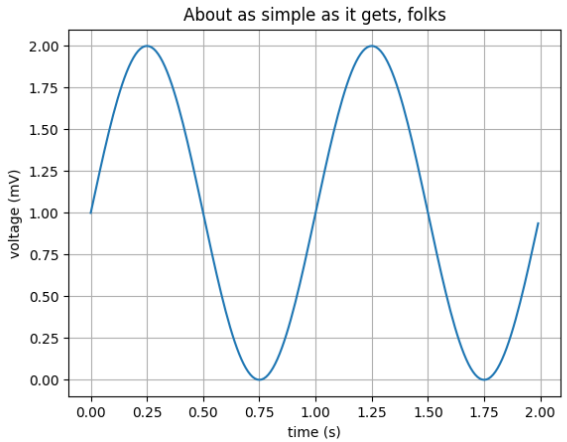
- O Matplotlib é uma biblioteca abrangente para criar visualizações estáticas, animadas e interativas em Python.
- Uma das ferramentas de plotagem em Python mais utilizadas.
- O módulo mais usado do Matplotlib é o *matplotlib.pyplot*.

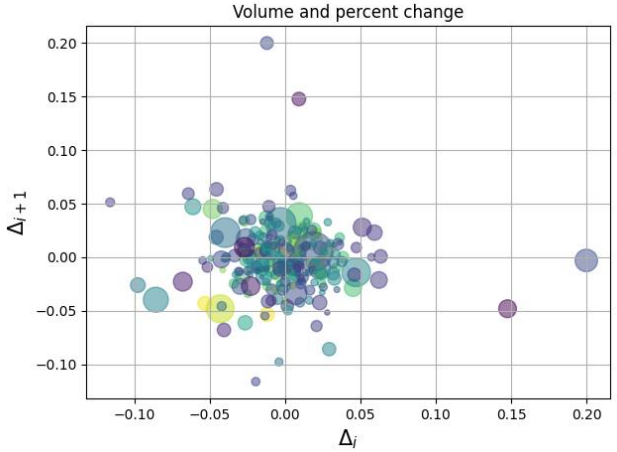
Tipos de Plotagem

- Existem diversos comandos de plotagens no Matplotlib:
  - Básicos: Linha, Barras, Dispersão;
  - Gráficos Estatísticos: Histograma, Boxplot;
  - 3D: Superfície 3D, Dispersão 3D.

Tipos de Plotagem

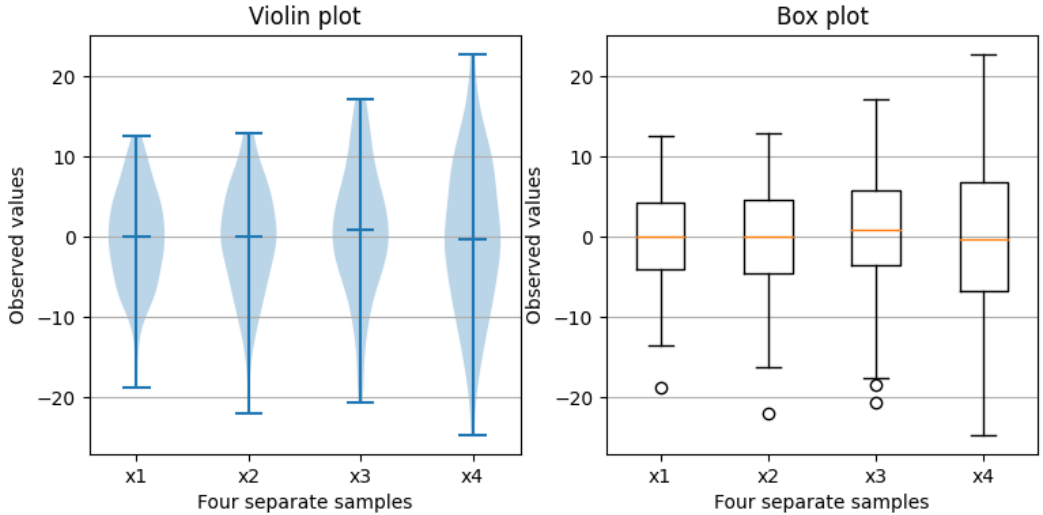
- Básicos





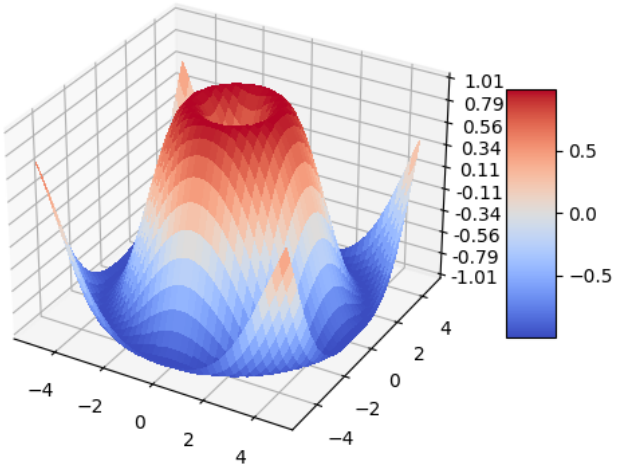
Tipos de Plotagem

- Estatísticos



Tipos de Plotagem

- 3D



2

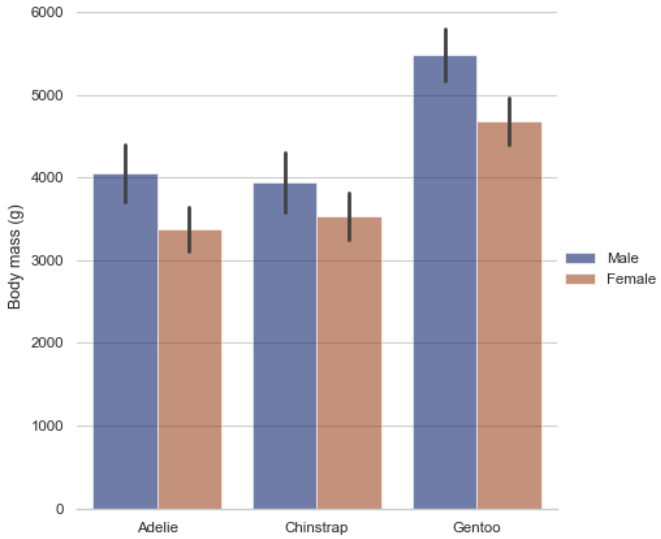
Seaborn

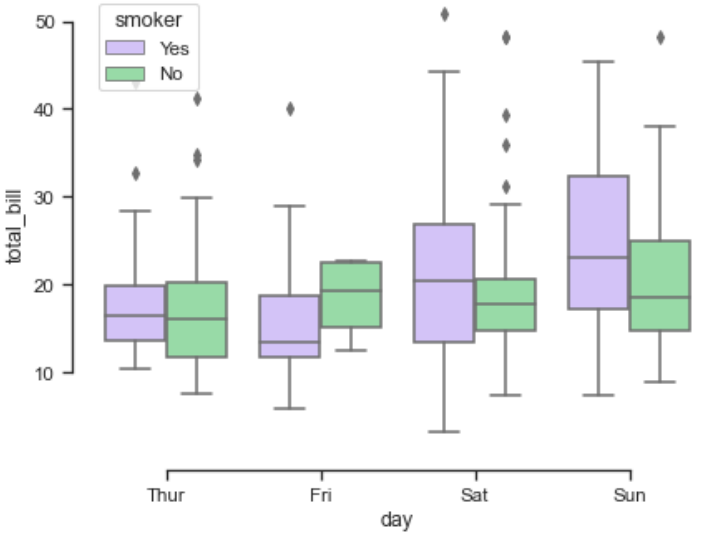
https://seaborn.pydata.org/

Seaborn

- Seaborn é uma biblioteca de visualização de dados Python baseada em Matplotlib e integra-se facilmente com as estruturas de dados do pandas.
- Fornece uma interface de alto nível para fazer gráficos estatísticos atraentes e informativos.

Seaborn


6000
5000
4000
3000
2000
1000
0
Body mass (g)
Adelie
Chinstrap
Gentoo
Male
Female


smoker
Yes
No
50
40
30
20
10
total_bill
Thur
Fri
Sat
Sun
day

Matplotlib x Seaborn

- O Matplotlib é ferramenta de visualização útil e popular, mas existem reclamações acerca do Matplotlib.
  - Fazer visualizações sofisticadas é possível, mas requer um código específico;
  - O Matplotlib não foi projetado para uso com Pandas DataFrames.
- Seaborn fornece uma API sobre o Matplotlib.
  - Define funções simples para estilos de plotagem e padrões de cores;
  - Integra-se com as funcionalidades fornecidas pelo Pandas.

Aplicação *Jupyter*

- Dataset Irís:

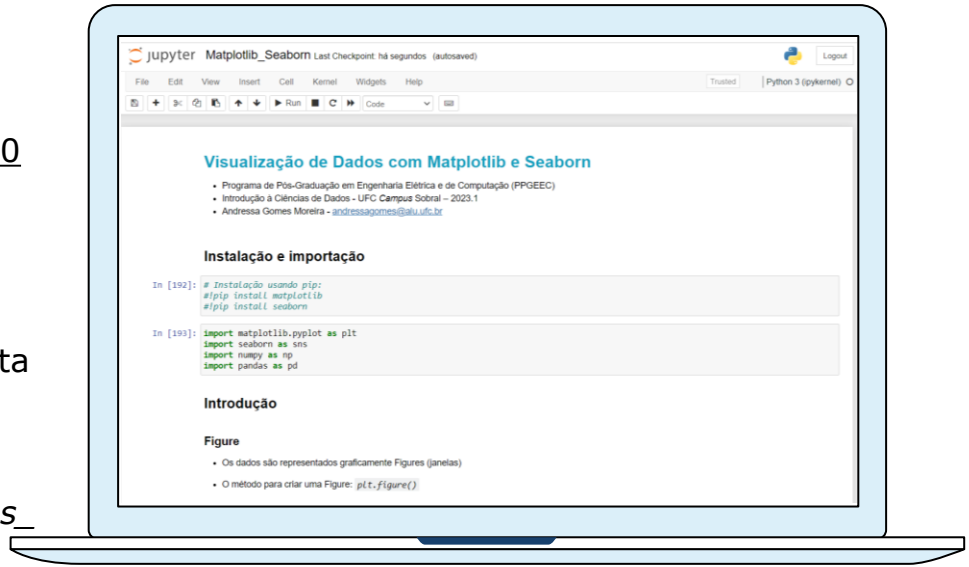
https://www.kaggle.com/datasets/saurabh00007/iriscsv

- Dataset Titanic:

https://www.kaggle.com/c/titanic-dataset/data

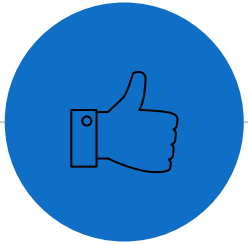
- Código – *git clone*

https://github.com/andressagomes26/ciencias_dados_estagio-docencia.git



Referências

- Documentação Matplotlib: ***https://pandas.pydata.org/***
- Documentação Seaborn***: https://seaborn.pydata.org/***
- Python Data Sience Handbook: ***http://ww12.allitebooks.org/***
- Data Science do Zero: ***https://altabooks.com.br/***
- Fundamentals of Data Visualization: ***http://oreilly.com***



Obrigada!

Alguma dúvida ?

- andressagomes@alu.ufc.br