

Python: Análise de Dados com Pandas III

Poliana N Ferreira

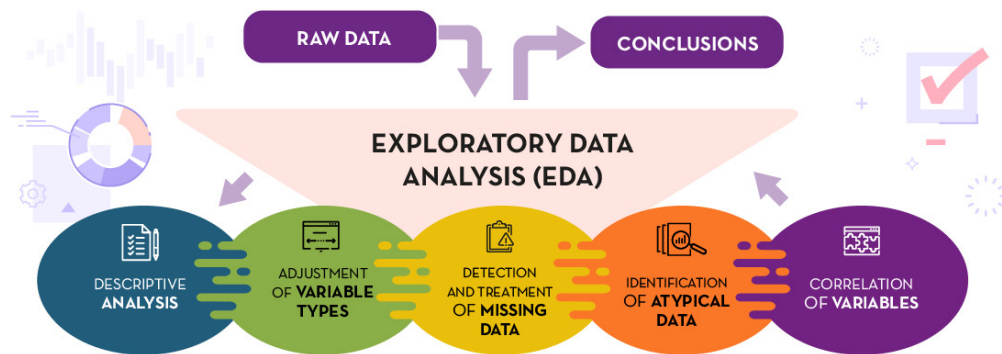
EDA – Análise de Dados exploratória



O que é?

Exploração dos dados

- Exploratory Data Analysis (EDA) – Análise exploratória dos dados.
- Serve para conhecer melhor os dados, descrevendo-os e preparando-os para análise preditiva.
- Já pode gerar informações úteis para o negócio



Estatística descritiva



Entendendo a estatística descritiva

O que é?

- A estatística descritiva é um ramo da estatística que se concentra em **resumir e descrever as características** de um conjunto de dados.
- Seu objetivo é fornecer uma compreensão clara dos dados, destacando padrões, tendências e características chave.



Importância da estatística descritiva

Ela serve como base para:

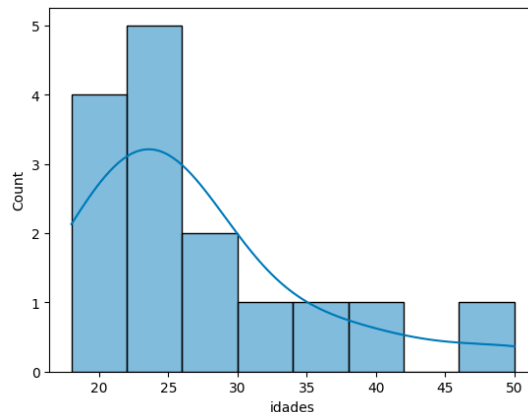
- **Compreensão Inicial dos Dados:** A estatística irá nos ajudar a descrever os dados. É muito difícil entender os dados só olhando pros seus valores.

Dados:

`idades = [20, 21, 24, 18, 24, 25, 26, 26, 24, 30, 19, 35, 40, 24, 50]`

Informações:

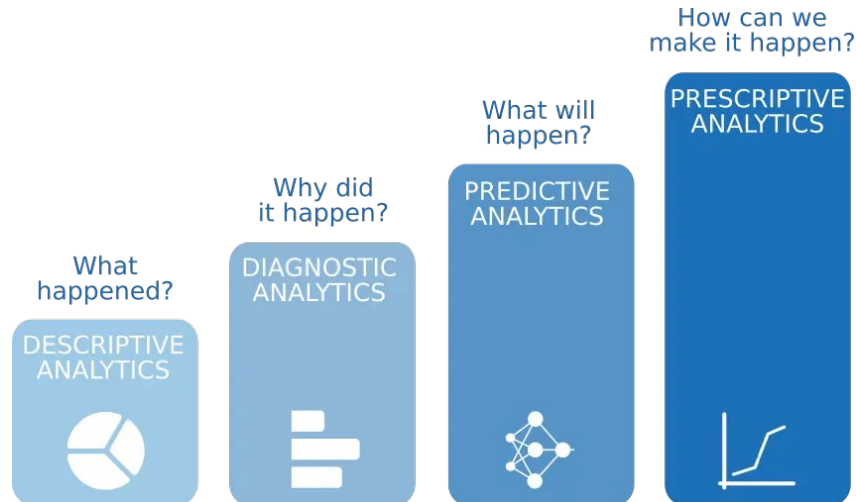
Idades variam entre **18 e 50**.
Média das idades é **27**.
Desvio padrão das idades é **8,3**.



Importância da estatística descritiva

Ela serve como base para:

- **Base para Análise Estatística Avançada:** Prepara o terreno para métodos estatísticos mais complexos, garantindo que análises subsequentes sejam realizadas em dados bem compreendidos e corretamente interpretados.

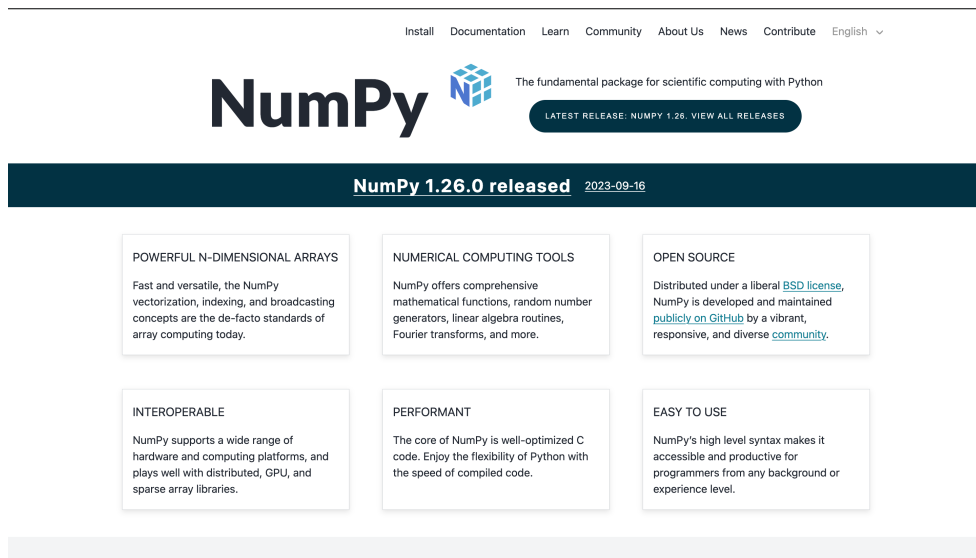


NumPy

Numerical Python

- O NumPy é uma biblioteca para a linguagem Python com funções para se trabalhar com computação numérica
- Baseada em arrays (listas) uni e multidimensionais

<https://numpy.org/>



The screenshot shows the NumPy website homepage. At the top, there is a navigation bar with links: Install, Documentation, Learn, Community, About Us, News, Contribute, and English. Below the navigation bar is the NumPy logo, which consists of the word "NumPy" in a large, bold, black font and a small 3D cube icon to its right. To the right of the logo is the tagline "The fundamental package for scientific computing with Python". Below the tagline is a dark blue button with white text that says "LATEST RELEASE: NUMPY 1.26. VIEW ALL RELEASES". Below this is a dark blue banner with white text that says "NumPy 1.26.0 released" and "2023-09-16". Below the banner are six white boxes with black text, each describing a feature of NumPy: "POWERFUL N-DIMENSIONAL ARRAYS", "NUMERICAL COMPUTING TOOLS", "OPEN SOURCE", "INTEROPERABLE", "PERFORMANT", and "EASY TO USE". Each box contains a brief description of the feature. At the bottom of the page, there is a light gray footer bar with the text "— . . . —".

Install Documentation Learn Community About Us News Contribute English

NumPy The fundamental package for scientific computing with Python

LATEST RELEASE: NUMPY 1.26. VIEW ALL RELEASES

NumPy 1.26.0 released 2023-09-16

POWERFUL N-DIMENSIONAL ARRAYS
Fast and versatile, the NumPy vectorization, indexing, and broadcasting concepts are the de-facto standards of array computing today.

NUMERICAL COMPUTING TOOLS
NumPy offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more.

OPEN SOURCE
Distributed under a liberal [BSD license](#), NumPy is developed and maintained [publicly on GitHub](#) by a vibrant, responsive, and diverse [community](#).

INTEROPERABLE
NumPy supports a wide range of hardware and computing platforms, and plays well with distributed, GPU, and sparse array libraries.

PERFORMANT
The core of NumPy is well-optimized C code. Enjoy the flexibility of Python with the speed of compiled code.

EASY TO USE
NumPy's high level syntax makes it accessible and productive for programmers from any background or experience level.

— . . . —

Medidas de tendência central

O que são?

- As medidas de tendência central são usadas para resumir os dados em torno de um ponto central.

Média (soma os valores e divide pela quantidade)

27

$$\bar{X} = \frac{(x_1 + x_2 + x_3 + \dots + x_n)}{n}$$

Media = (20+21+24+18+24+25+26+26+24+30+29+35+40+24+50)/15

Media = 27

Mediana (ordena os valores e pega o valor do meio)

24

Ordenado = [18, 19, 20, 21, 24, 24, 24, **24**, 25, 26, 26, 30, 35, 40, 50]

Mediana = 24

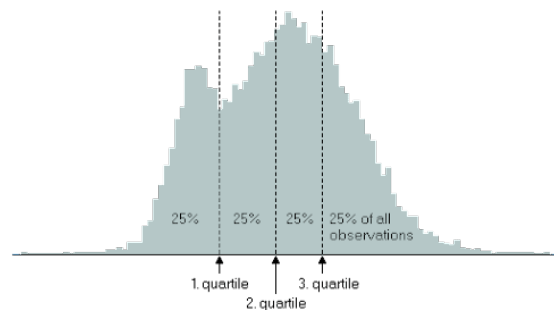
Moda (valor mais comum)

24

Moda = 24 (aparece 4x)

Quadris e percentis

- Valores que dividem o conjunto de dados em um certo percentual



Ordenado = [18, 19, 20, 21, 24, 24, 24, **24**, 25, 26, 26, 30, 35, 40, 50]

25% percentil
Q1

50% percentil
Q2

75% percentil
Q3

Medidas de dispersão

O que são?

- Cruciais para entender o quão espalhados ou concentrados estão os valores em um conjunto de dados.
- **Amplitude:** Diferença entre o valor mais alto e o mais baixo.
- **Desvio Padrão e Variância:** Medem a dispersão dos dados em relação à média; quanto maior, mais os dados estão espalhados.
 - A variância dá uma ideia do espalhamento enfatizando valores mais extremos, enquanto o desvio padrão nos dá uma ideia mais direta.

Medidas de Dispersão

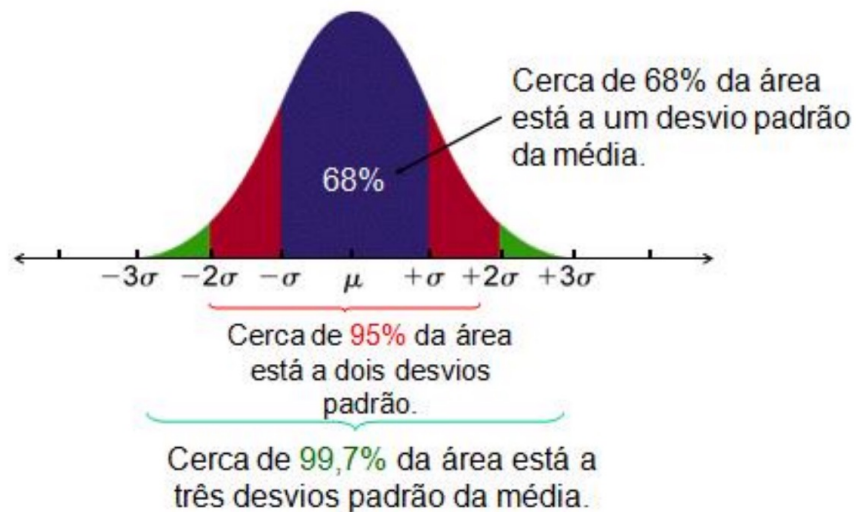
Fórmulas

	Populacional	Amostral
Variância	$\sigma^2 = \frac{\sum (X_i - \bar{X})^2}{n}$	$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$
Desvio Padrão	$\sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}}$	$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$

Medidas de dispersão

Explicando melhor o desvio padrão

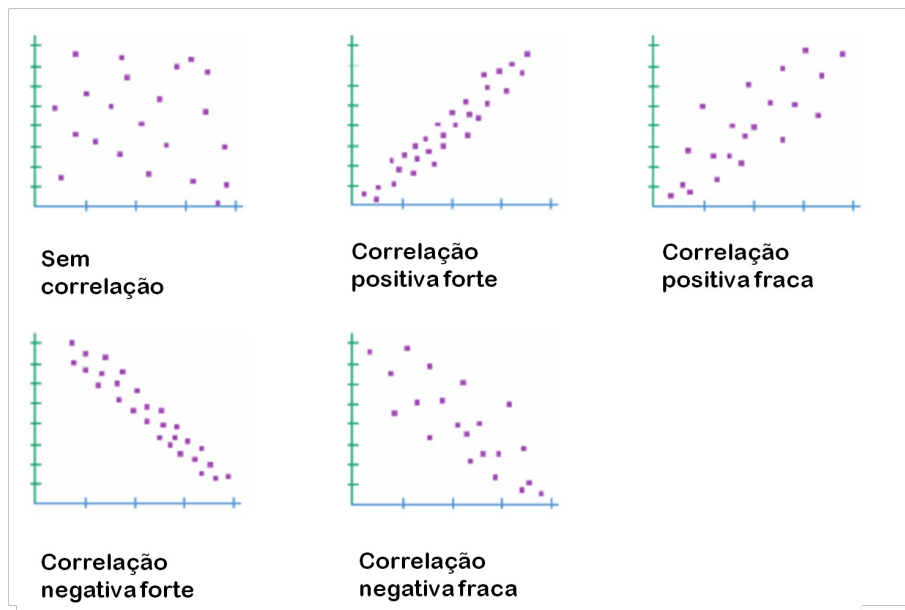
Regra empírica



Medidas de relação entre atributos

Covariância e correlação

- Quantifica o quanto um atributo varia quando o outro varia.



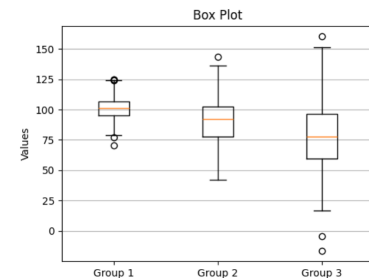
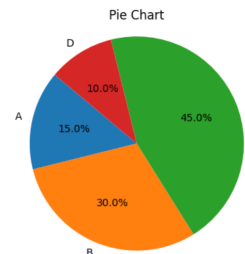
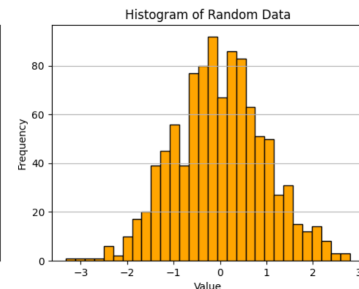
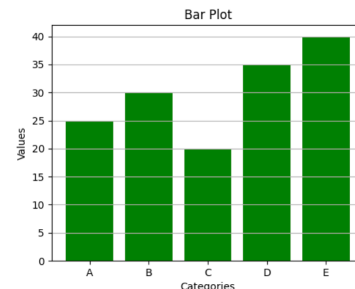
Visualização de Dados



O que é uma visualização de dados?

A prática de apresentar dados em formato gráfico ou visual.

- Utiliza vários elementos visuais, como tabelas, gráficos, mapas e infográficos para **representar conjuntos de dados complexos de forma eficaz**.
- Integra princípios de estatística, design e ciência cognitiva para criar representações de dados significativas e perspicazes.

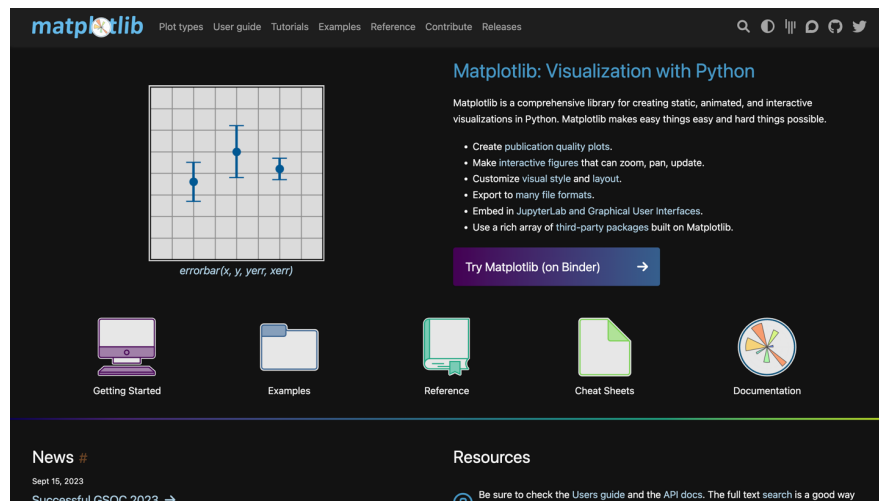


Matplotlib

Gráficos

- Matplotlib é uma biblioteca de software para criação de gráficos e visualizações de dados em geral, feita para e da linguagem de programação Python e sua extensão de matemática NumPy

<https://matplotlib.org/>



Seaborn

O que é e porque usar

- Uma biblioteca de visualização de dados Python baseada no Matplotlib, focada em estatísticas.
- Simplifica a criação de gráficos estatísticos com estilos e paletas atraentes. Ideal para análises exploratórias com menos código.
- Oferece uma API mais intuitiva e designs melhores com menos esforço.
- Enquanto o Matplotlib se destaca na personalização, o Seaborn brilha na facilidade de uso e visualizações estatísticas.



Plotly

O que é?

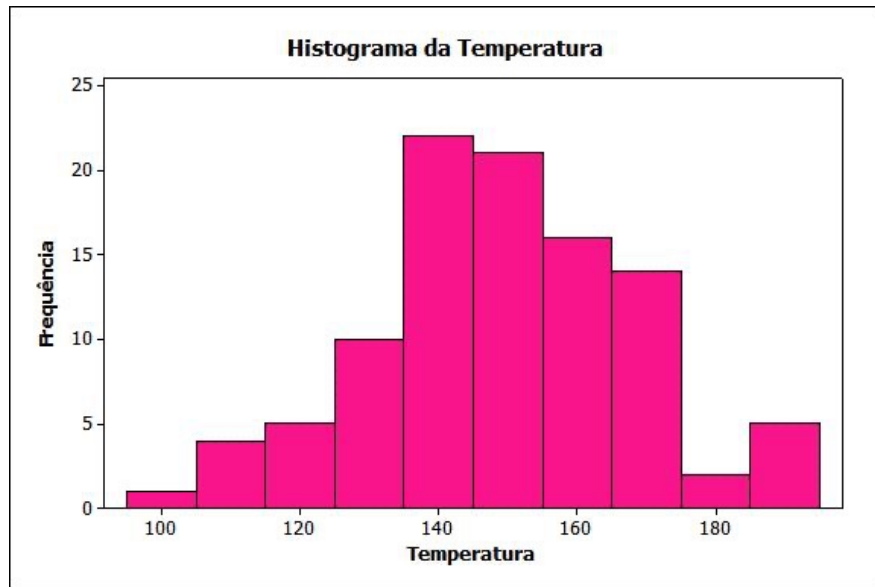
- Uma biblioteca de visualização de dados que tem em seu cerne a interatividade!
- Tem mais de 40 tipos de gráficos diferentes, desde gráficos estatísticos, até financeiros, geográficos e 3D.



Visualização de Dados

Histograma

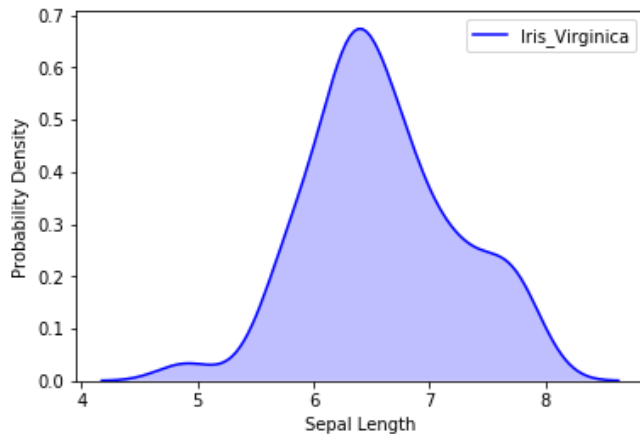
- Dados univariados
- Mostra a distribuição de uma variável contínua, ajudando a entender a forma e a dispersão dos dados.



Visualização de Dados

KDE – Gráfico de Densidade

- Dados univariados
- Distplot combina histogramas e KDE (Estimativa de Densidade Kernel) para visualizar a distribuição de uma variável contínua. KDE Plot é uma versão suavizada do histograma, mostrando a densidade de probabilidade.



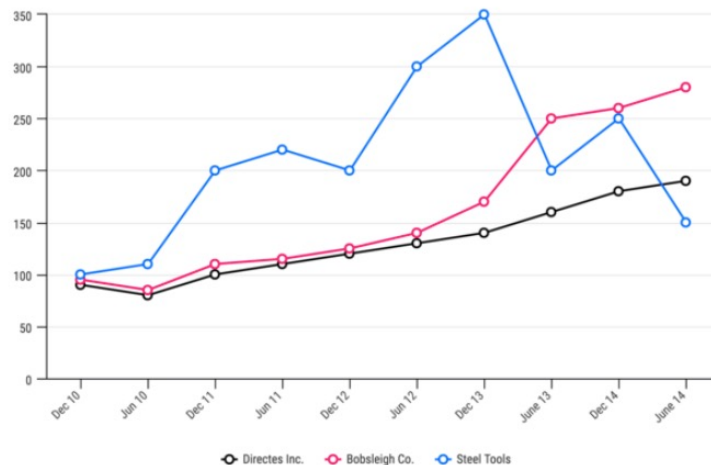
Visualização de Dados

Gráfico de Linha

- Dados univariados
- Utilizado para visualizar dados ao longo do tempo, mostrando tendências ou mudanças contínuas.

our year in review.

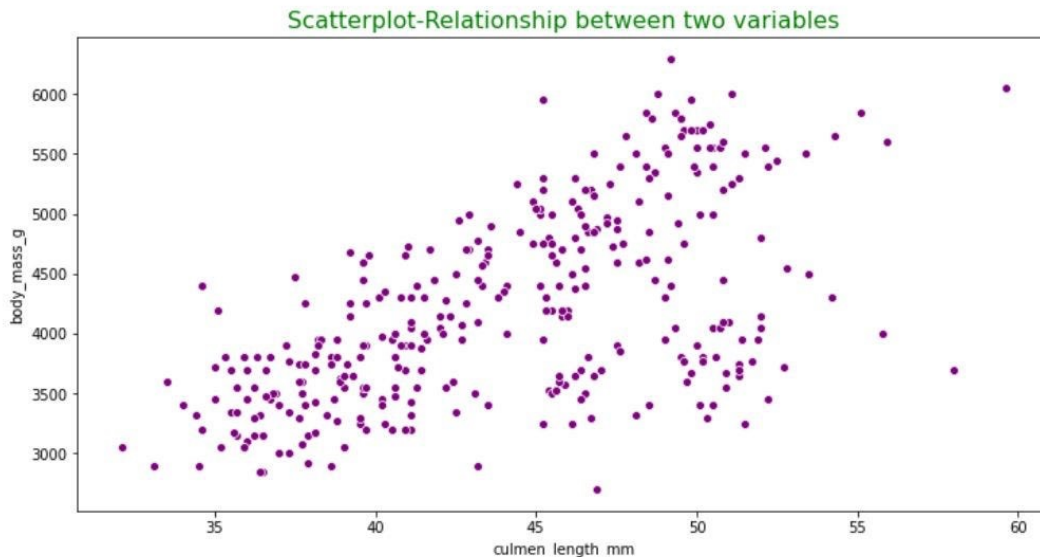
Steel Tool's Year Review & Performance



Visualização de Dados

Gráfico de Dispersão – Scatter Plot

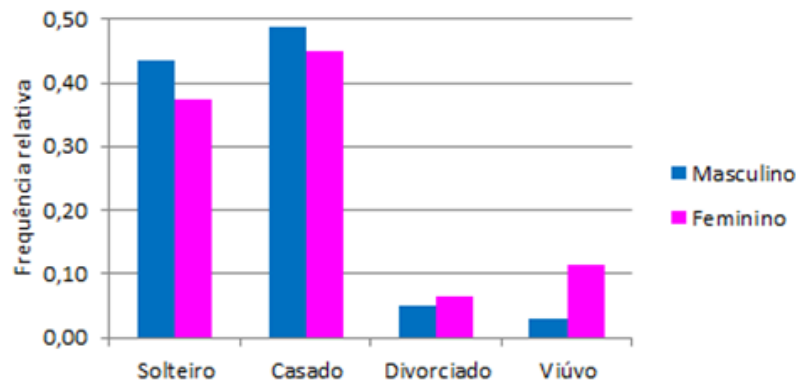
- Dados multivariados
- Utilizado para observar relações ou correlações entre duas variáveis contínuas.



Visualização de Dados

Gráfico de Barras / Categorias

- Dados multivariados
- Barplot e Catplot
- Ideal para comparar quantidades entre diferentes categorias.
- Permite visualizar a relação entre uma variável numérica e uma ou mais variáveis categóricas, usando barras para mostrar medidas de tendência central, por exemplo.

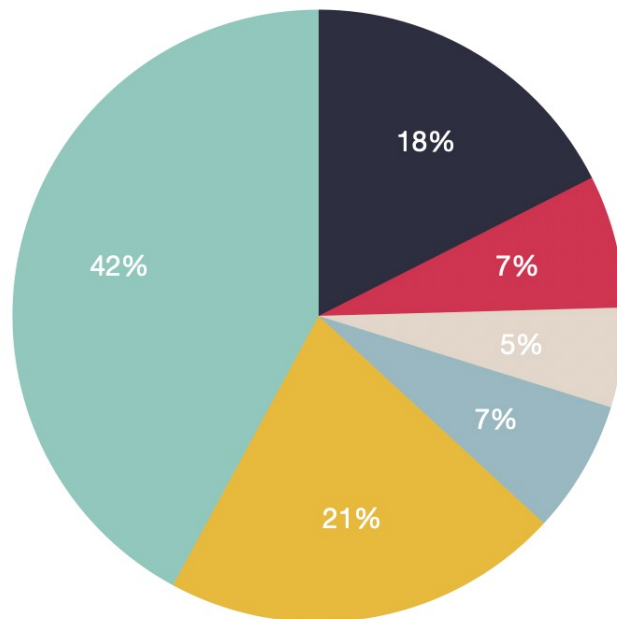


Visualização de Dados

Gráfico de Pizza

- Dados multivariados

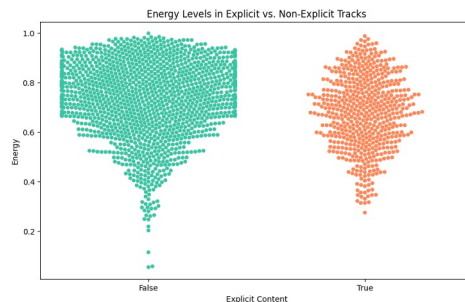
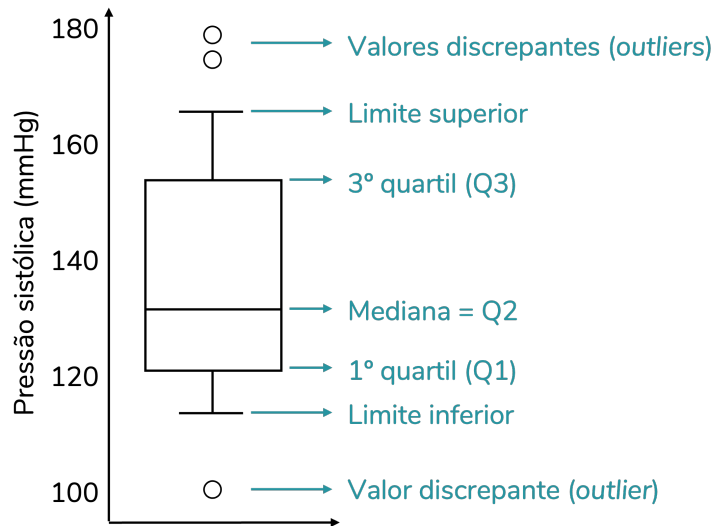
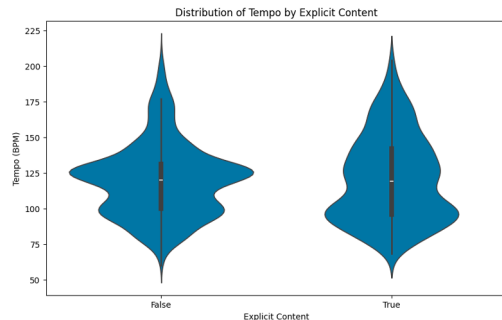
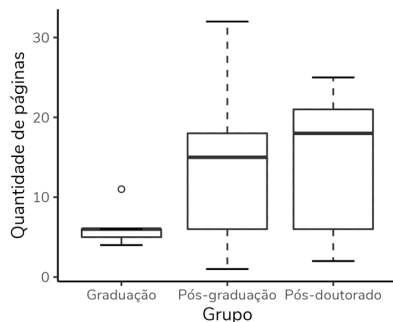
● Economias ● Transporte ● Seguro ● Entretenimento ● Comida
● Casa



Visualização de Dados

Gráfico BoxPlot

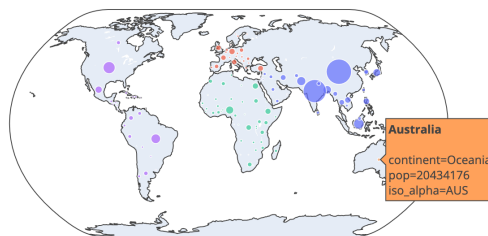
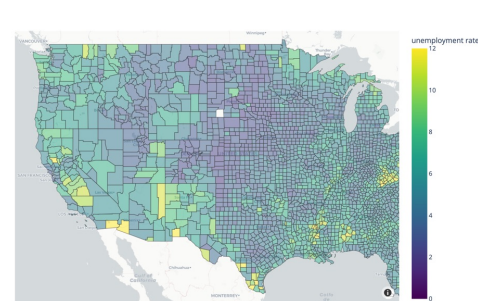
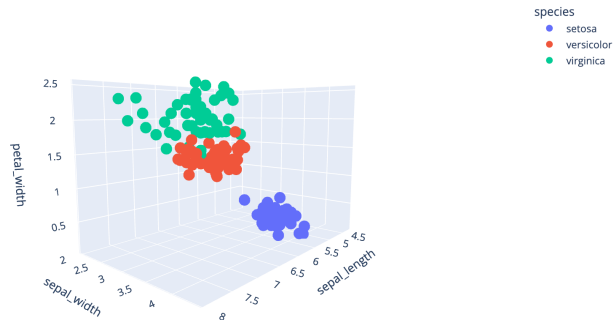
- Dados uni ou multivariados
- Visualiza a distribuição de dados através de quartis, destacando a mediana e identificando possíveis outliers.
- Similares: violinplot e swarmplot



Gráficos Avançados

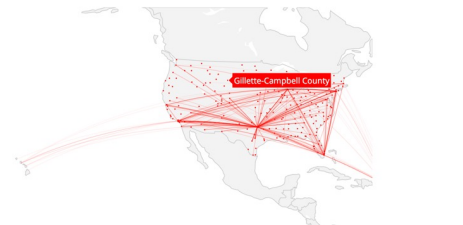
Possibilidade de desenvolvimento de gráficos mais complexos

- 3D interativo
- Mapas
 - Cloropeth
 - Scatter
 - Line



- continent
- Asia
 - Europe
 - Africa
 - Americas
 - Oceania

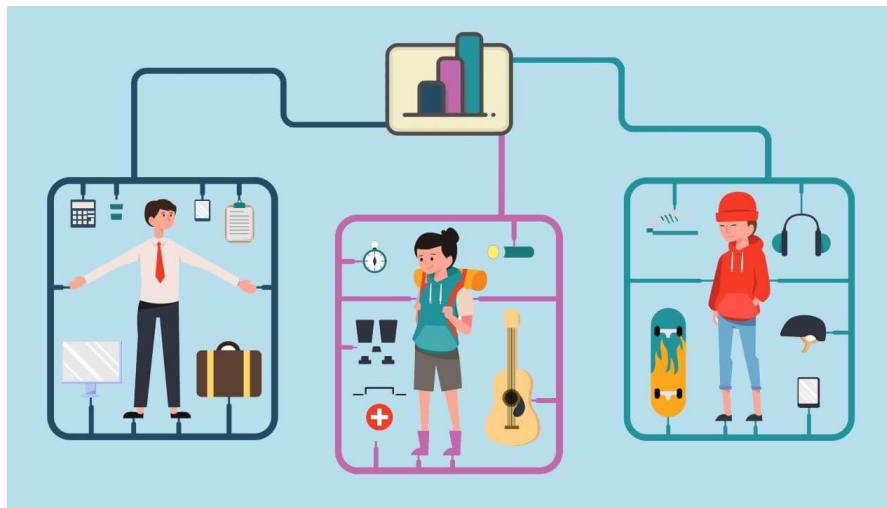
Feb. 2011 American Airline flight paths
(Hover for airport names)



<https://plotly.com/python/3d-scatter-plots/>

Personalização de gráficos

Cores, estilos, legendas e títulos



Salvar e persistência dos dados



Salvar dados - Pandas

- Podemos salvar para um csv ou para uma base de dados, como fizemos no bigquery.
- To_csv

```
df.to_csv("save.csv", sep=',', index=False, encoding='utf-8')
```

Agora vamos avaliar a aula?



Obrigada!