Análise exploratória

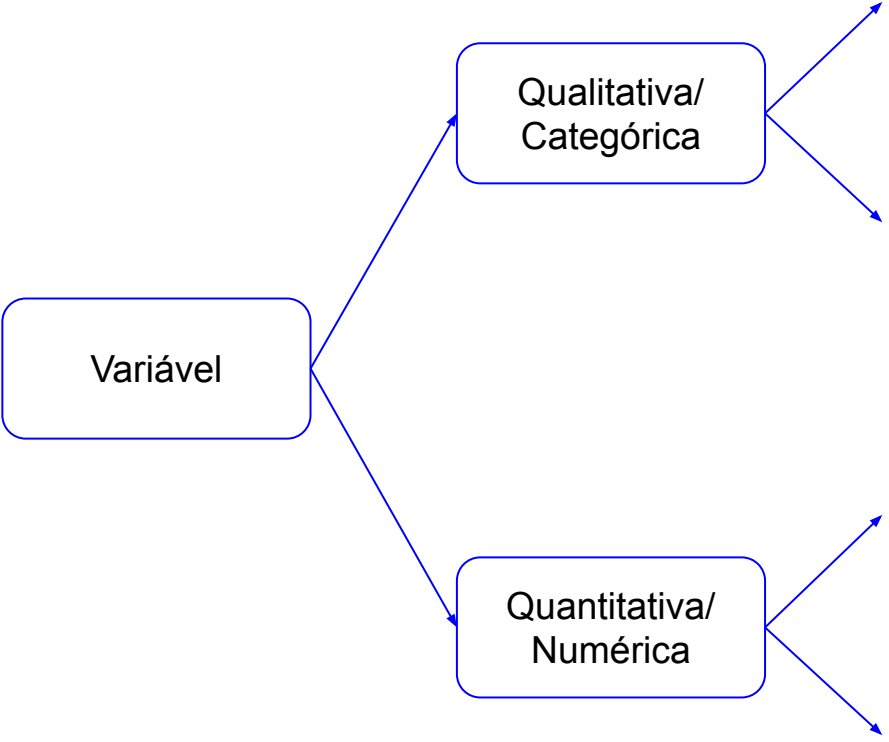
Descritiva univariável usando matplotlib

Agenda

- Tipos de Variáveis
- Tipos de gráficos
  - Linha, Dispersão, Barras, Colunas, Setores
- Análise Univariada
- Matplotlib
  - Demo
- Exercício

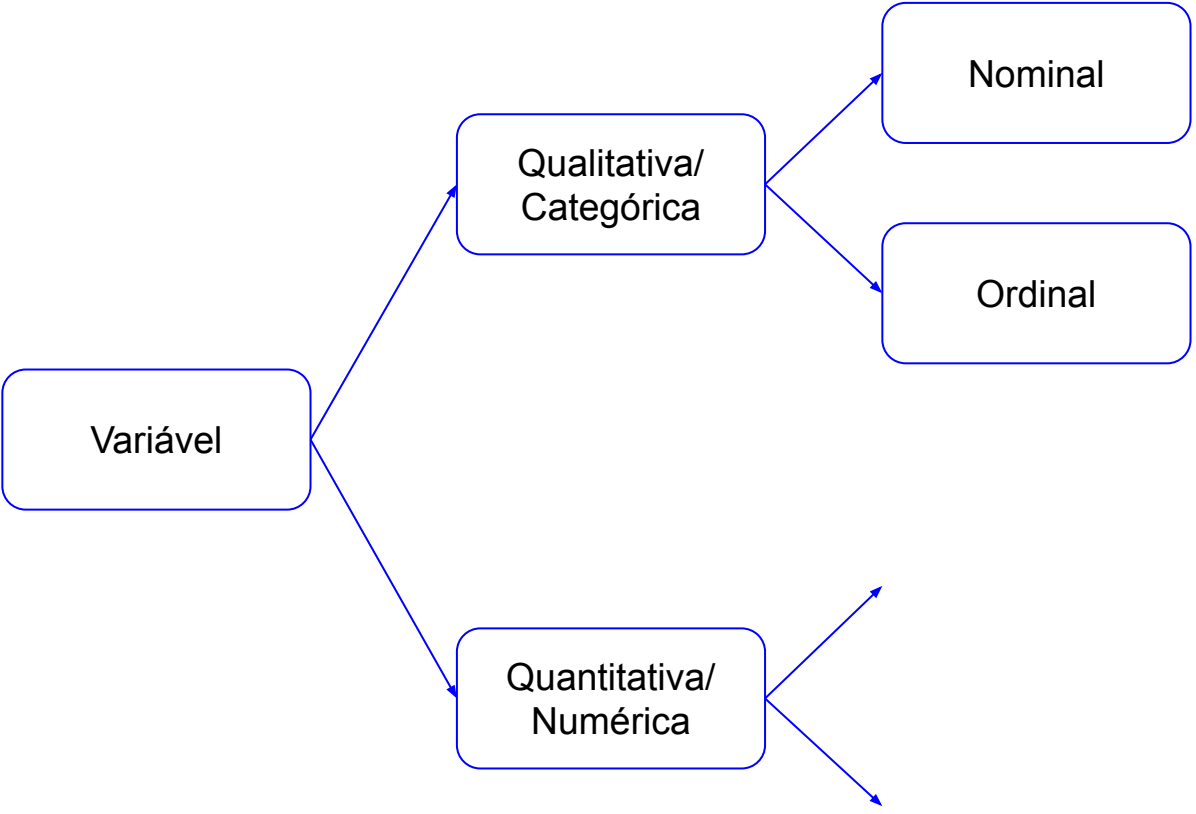
Variáveis

Variável: Característica comum a determinada população que pode assumir valores distintos



Variáveis

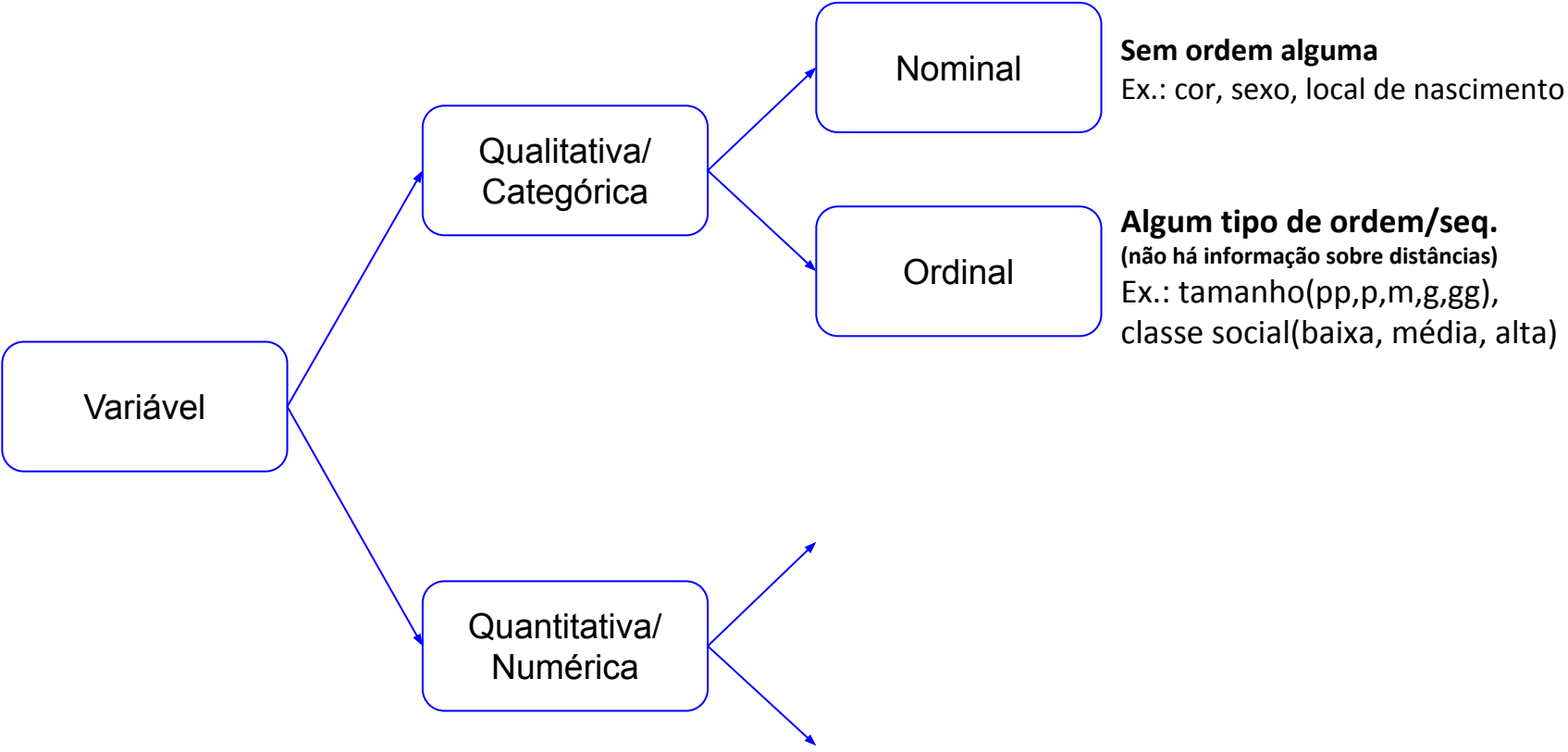
Variável: Característica comum a determinada população que pode assumir valores distintos



Sem ordem alguma
Ex.: cor, sexo, local de nascimento

Variáveis

Variável: Característica comum a determinada população que pode assumir valores distintos

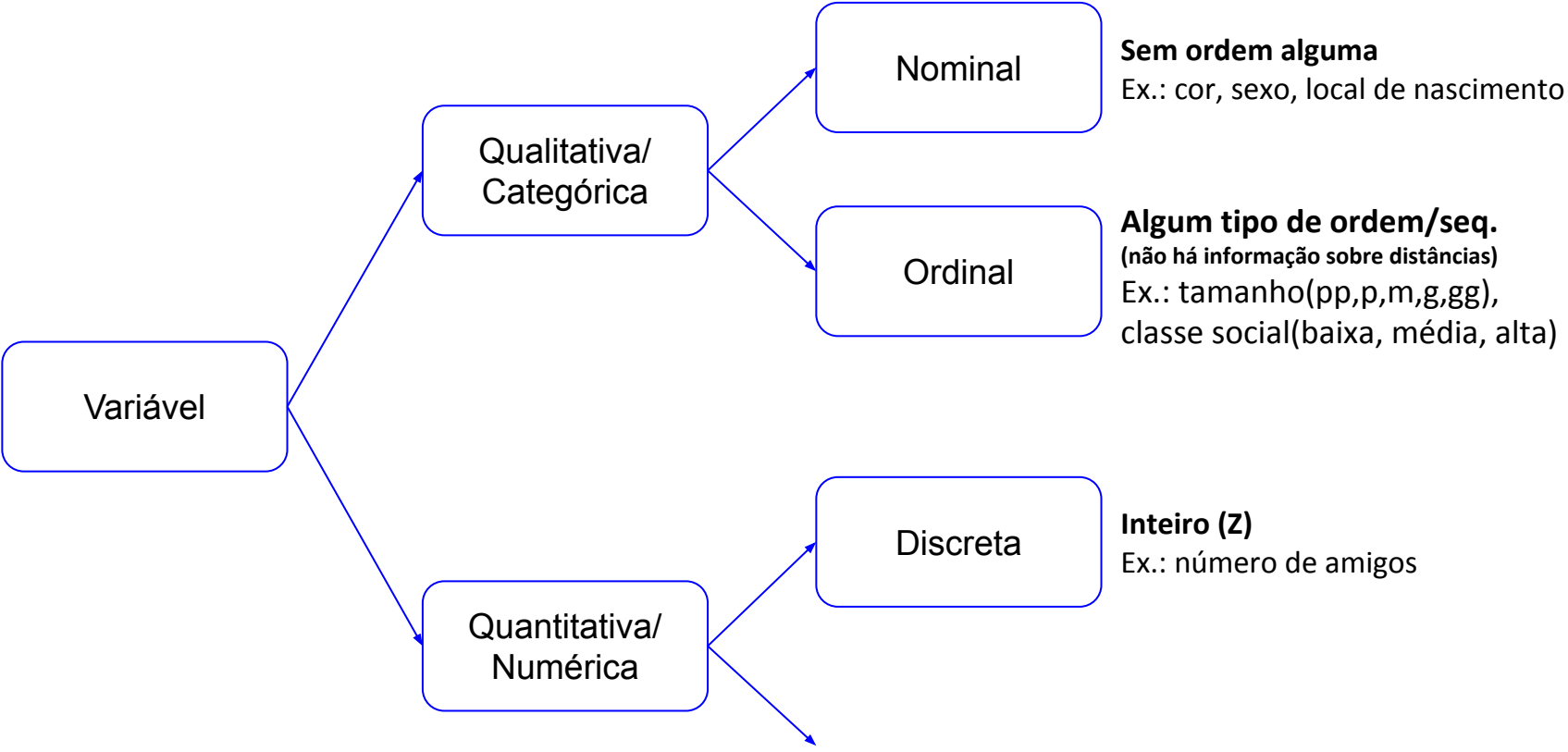


Sem ordem alguma
Ex.: cor, sexo, local de nascimento

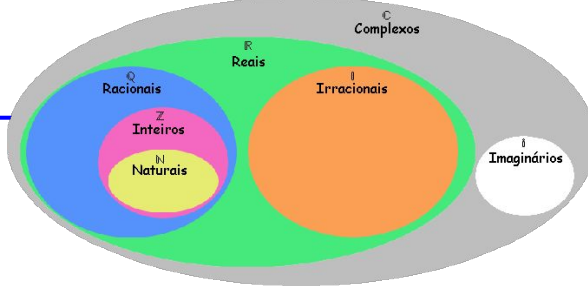
Algum tipo de ordem/seq.
(não há informação sobre distâncias)
Ex.: tamanho(pp,p,m,g,gg), classe social(baixa, média, alta)

Variáveis

Variável: Característica comum a determinada população que pode assumir valores distintos

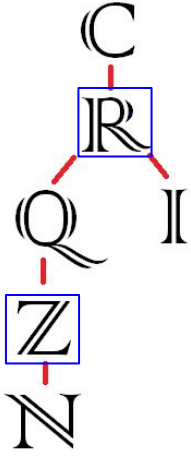


Variáveis



Variável: Característica comum a determinada população que pode assumir valores distintos

- Variável
  - Qualitativa/ Categórica
    - Nominal
      Sem ordem alguma
      Ex.: cor, sexo, local de nascimento
    - Ordinal
      Algum tipo de ordem/seq.
      (não há informação sobre distâncias)
      Ex.: tamanho(pp,p,m,g,gg), classe social(baixa, média, alta)
  - Quantitativa/ Numérica
    - Discreta
      Inteiro (Z)
      Ex.: número de amigos
    - Contínua
      Real (R)
      Ex.: peso, altura, preço



Discretização

Variáveis, como tempo, são contínuas. No entanto, em alguns casos, podemos classificá-las como discretas. Este processo de discretização pode ser utilizado quando:

- Não há perda de informação ao discretizar a variável, ou quando o detalhamento não é fundamental para o entendimento da variável;

- A variável é medida por um equipamento que não é capaz de mensurar com a precisão ideal. Exemplo: peso de um carro medido por uma balança que adota escala de 1 em 1 kg -- o peso seria considerado discreto, mesmo que saibamos que é contínuo.

Gráfico de Linha

- Utilizado para mostrar evolução, tendências nos dados em intervalos iguais (ex.: tempo)
- Plota dois grupos de números como uma série (ex.: temporal) de coordenadas xy

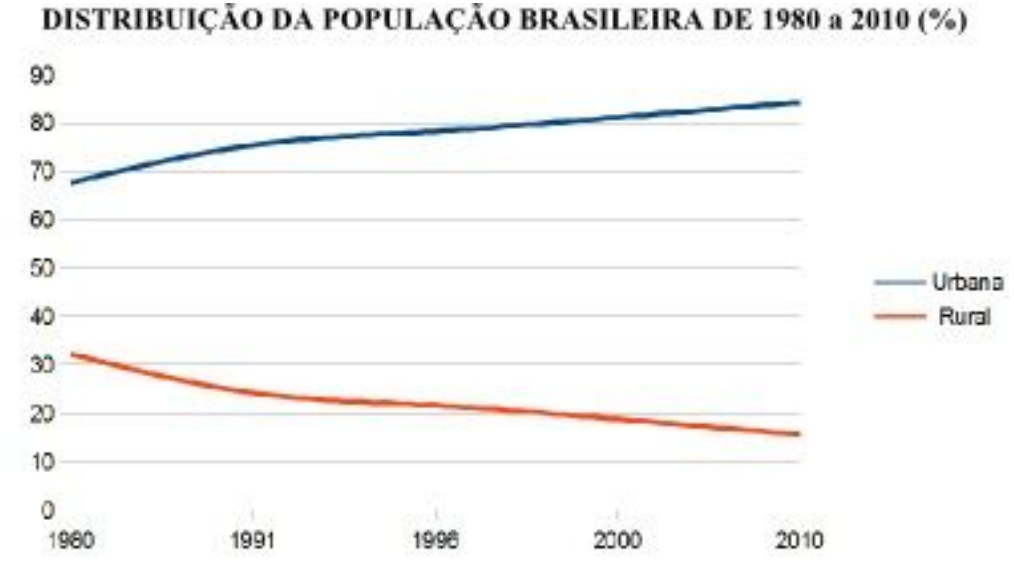
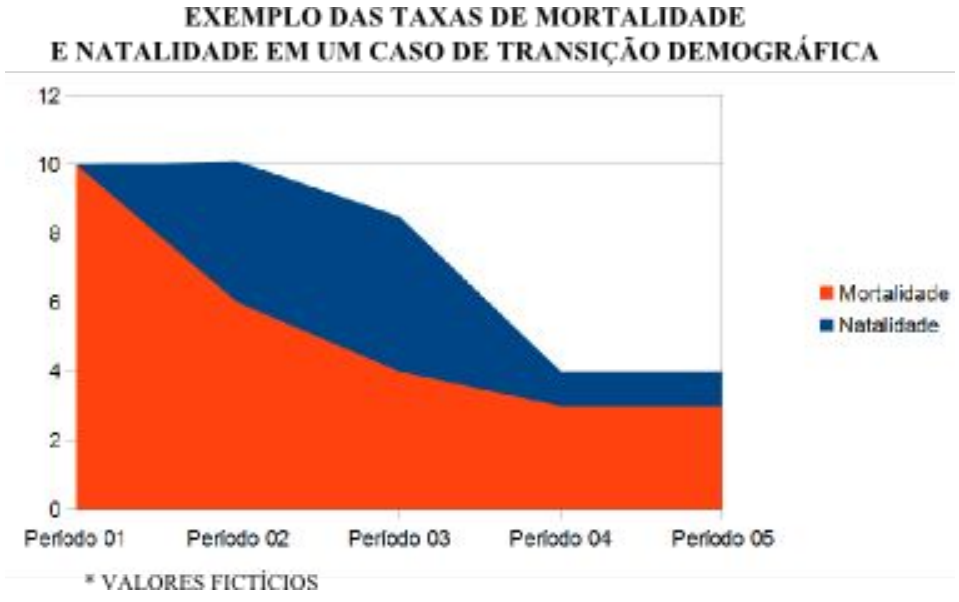


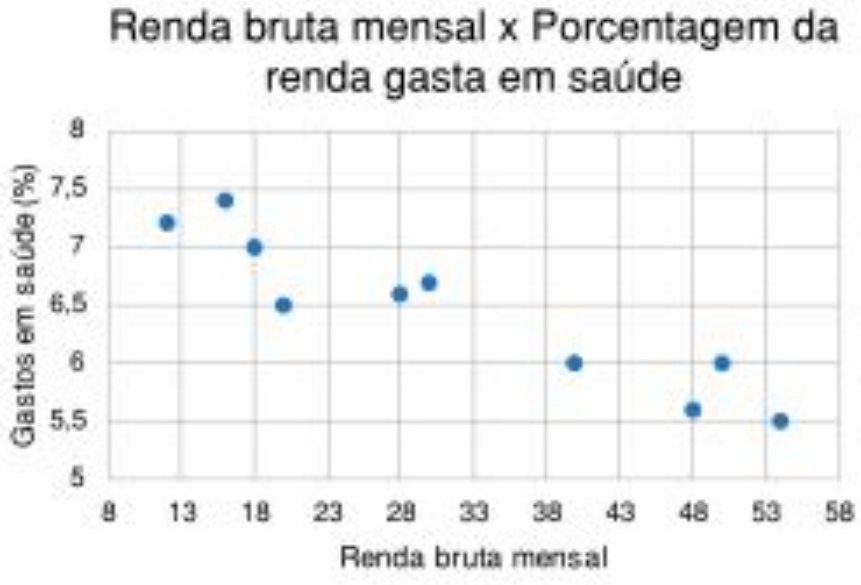
Gráfico de Área

- Utilizado para mostrar evolução, tendências nos dados em intervalos iguais (ex.: tempo)
- Evidencia uma noção de proporção sobre o todo



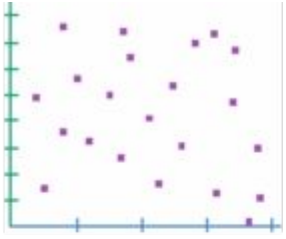
Dispersão XY

- Usados para examinar a associação entre duas medidas
- Permitem enxergar padrões, outliers, agrupamentos etc.

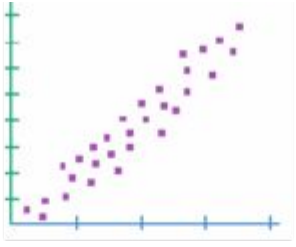


Dispersão XY

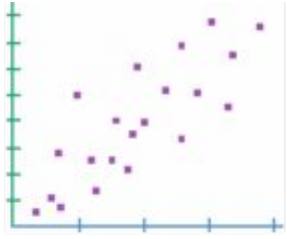
- Usados para examinar a associação entre duas medidas
- Permitem enxergar padrões, outliers, agrupamentos etc.



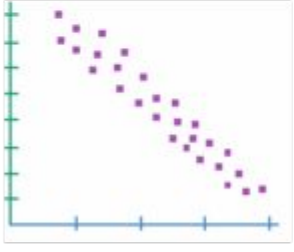
Sem correlação



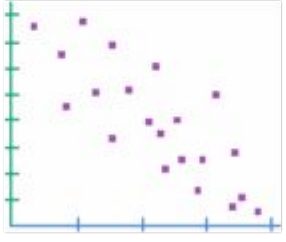
Correlação positiva forte



Correlação positiva média



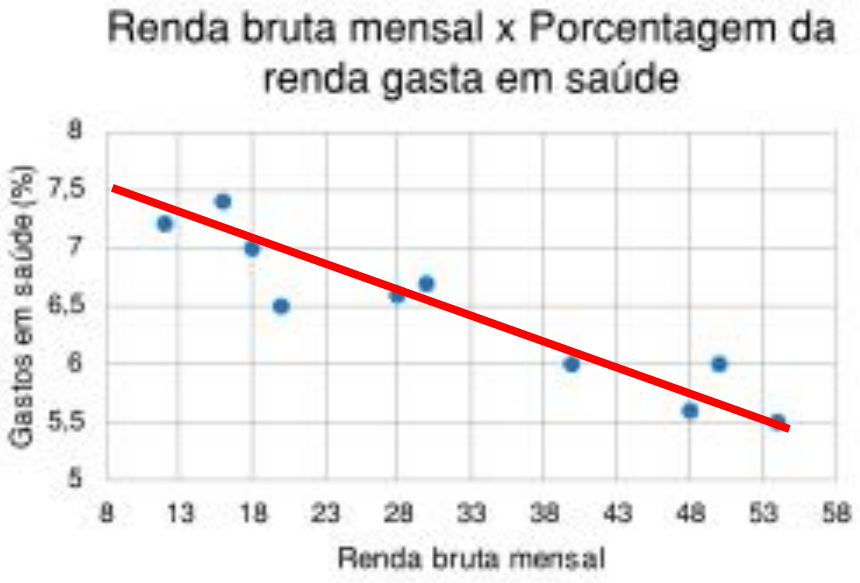
Correlação negativa forte



Correlação negativa média

Dispersão XY

- Usados para examinar a associação entre duas medidas
- Permitem enxergar padrões, outliers, agrupamentos etc.



Barras/colunas

- Ilustra comparações entre itens individuais
- As categorias e valores são organizados com o objetivo de facilitar a comparação (menos ênfase ao tempo, por exemplo)

Setores (pizza)

- Busca dar visão proporcional a (poucos ou muito discrepantes) itens
- Mostra somente uma única série de dados
- Útil quando você deseja dar ênfase a um elemento importante

Análise Univariada

- Dados devem estar no formato de matriz
  - Cada linha da matriz corresponde a uma unidade experimental: Elemento da população ou amostra no qual observamos as variáveis
  - Cada coluna da matriz corresponde a uma variável
- Para cada uma das variáveis individualmente:
  - Classificar a variável quanto a seu tipo: qualitativa (nominal ou ordinal) ou quantitativa (discreta ou contínua)
  - Obter tabelas, gráficos e/ou medidas que resumem a variável

Análise Univariada

- Dados devem estar no formato de matriz
  - Cada linha da matriz corresponde a uma unidade experimental: Elemento da população ou amostra no qual observamos as variáveis
  - Cada coluna da matriz corresponde a uma variável
- Para cada uma das variáveis individualmente:
  - Classificar a variável quanto a seu tipo: qualitativa (nominal ou ordinal) ou quantitativa (discreta ou contínua)
  - Obter tabelas, gráficos e/ou medidas que resumem a variável
- A partir dos resultados, montar um resumo geral dos dados;
- Objetivo do cientista é: conhecer o **comportamento** da variável e entender as características das ocorrência de suas possíveis realizações.
  - Distribuições de frequência: **principal** recurso para resumir uma única variável.

Variável Qualitativa Nominal

- Ex.: Estado civil
  - Uma tabela de frequências (absolutas e/ou relativas)
  - Um gráfico de barras ou de setores (pizza, circular)

casado	0.555556
solteiro	0.444444
soma	1

Freq. relativa

casado	20
solteiro	16
soma	36

Freq. absoluta

Frequência absoluta (f_i): número total de elementos em cada classe
Frequência relativa (fr_i): razão entre cada valor da frequência absoluta e o total de observações

$$fr_i = \frac{f_i}{\sum f_i}$$

Frequência percentual (fp_i): frequência relativa em porcentagem

$$fp_i = fr_i \times 100$$

Variável Qualitativa Nominal

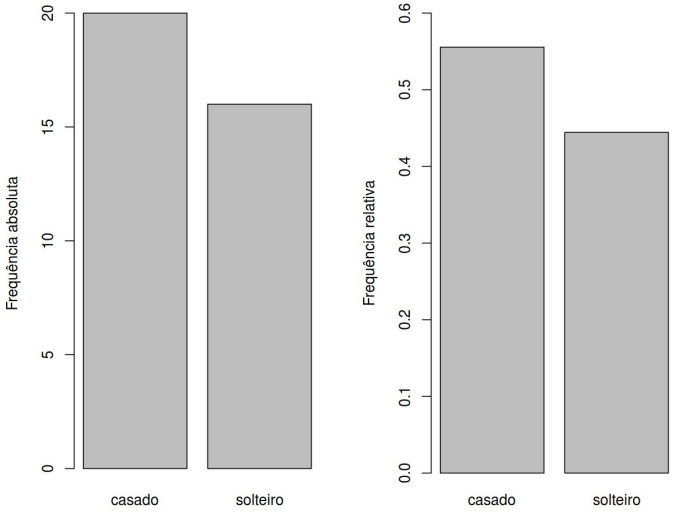
- Ex.: Estado civil
  - Uma tabela de frequências (absolutas e/ou relativas)
  - Um gráfico de barras ou de setores (pizza, circular)

casado	0.555556
solteiro	0.444444
soma	1

Freq. relativa

casado	20
solteiro	16
soma	36

Freq. absoluta



Variável Qualitativa Nominal

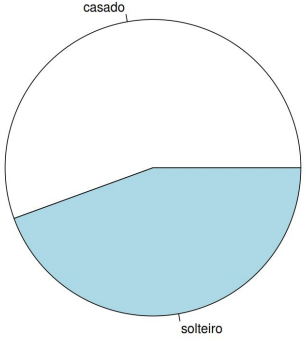
- Ex.: Estado civil
  - Uma tabela de frequências (absolutas e/ou relativas)
  - Um gráfico de barras ou de setores (pizza, circular)

casado	0.555556
solteiro	0.444444
soma	1

Freq. relativa

casado	20
solteiro	16
soma	36

Freq. absoluta



Observações:

Gráficos Circulares não são bons para entendimento de variáveis qualitativas com muitas categorias

A moda é definida como o valor mais frequente na amostra. No caso de variáveis qualitativas, a moda é a categoria que apresenta maior frequência.

Variável Qualitativa Ordinal

- Ex.: Escolaridade
  - Uma tabela de frequências (absolutas e/ou relativas)
  - Um gráfico de barras

Freq. absoluta

1o grau	12
2o grau	18
Superior	6
soma	36

Freq. relativa

1o grau	0.33333
2o grau	0.50000
Superior	0.16666
soma	1

Variável Qualitativa Ordinal

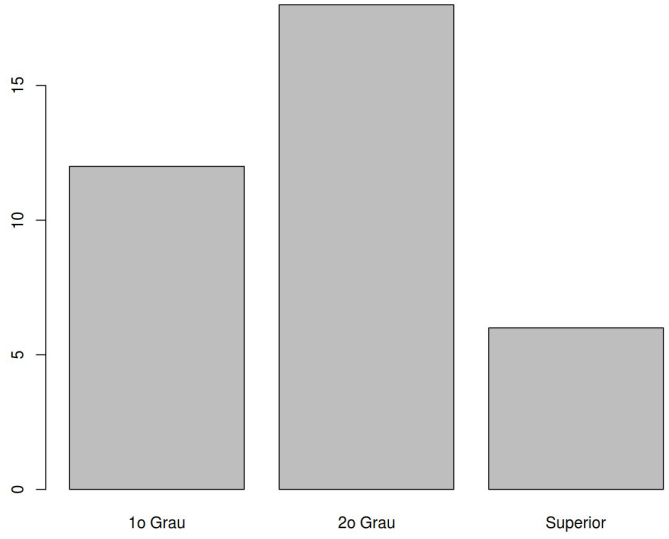
- Ex.: Escolaridade
  - Uma tabela de frequências (absolutas e/ou relativas)
  - Um gráfico de barras

Freq. absoluta

1o grau	12
2o grau	18
Superior	6
soma	36

Freq. relativa

1o grau	0.33333
2o grau	0.50000
Superior	0.16666
soma	1

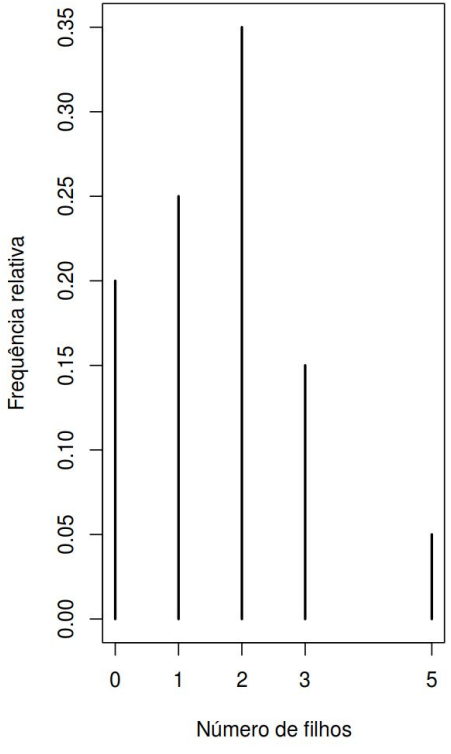


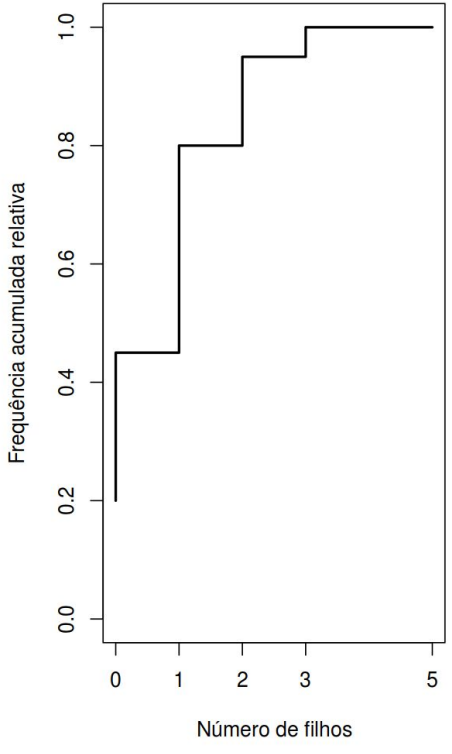
Observações:

Gráficos Circulares não expressam a ordem intrínseca a esse tipo de variável

Variável quantitativa discreta

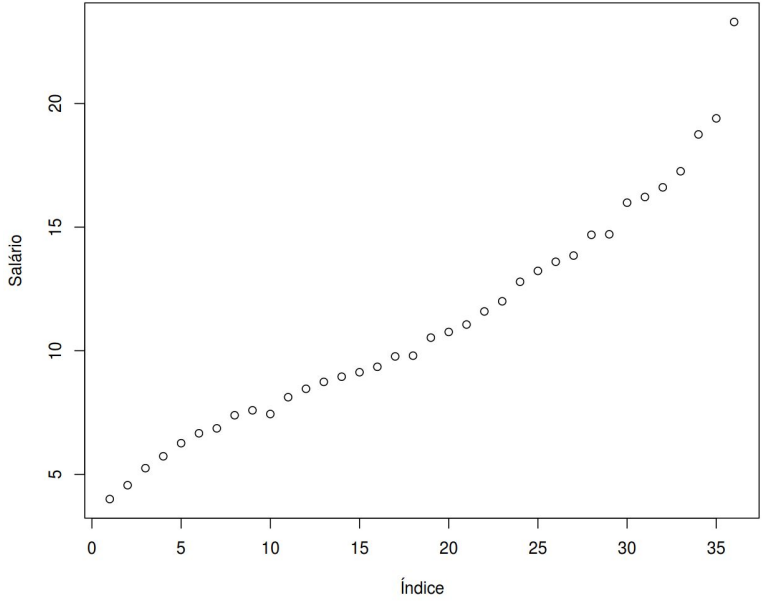
- Ex.: Número de Filhos
  - Uma tabela de frequências (cenários com pouca variação)
  - Frequência acumulada (df.cumsum())
  - Um gráfico de barras ou linhas

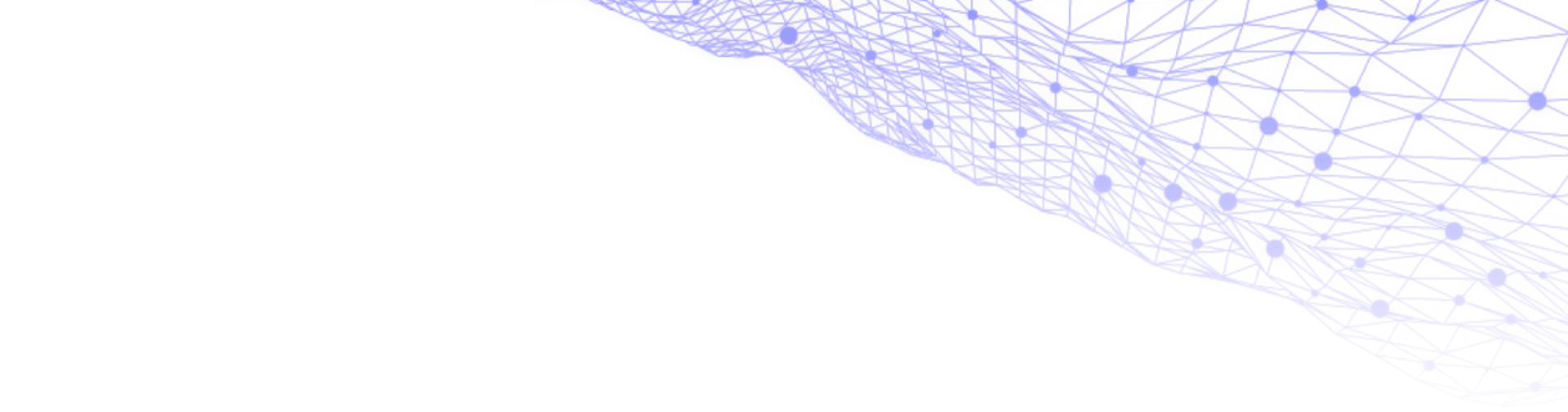




Variável quantitativa contínua

- Ex.: Salário
  - Tabela de frequências de uma variável contínua: agrupar os dados em classes pré-estabelecidas
  - Gráfico de Dispersão e Histograma
  - Permite a variação dos intervalos de classe
    - aleatoriamente ou baseado-se em algumas regras (amplitude, desvio padrão)





Matplotlib

Ferramentas/Tecnologias/Plataformas

Seaborn

plotly

matplotlib

plotly

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

python

GitHub

kaggle

jupyter

kaggle

binder

docker

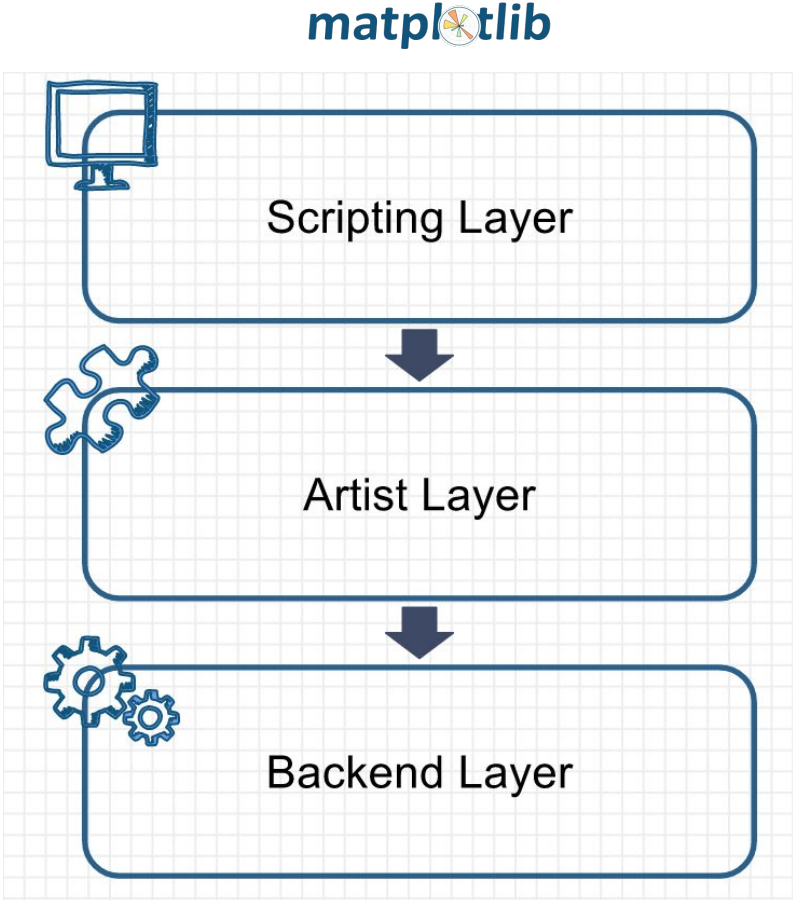
Matplotlib

- Começou como uma alternativa viável ao Matlab
- Os termos usados (Eixo, Figura, Gráficos) são semelhantes aos usados no MATLAB
- Provavelmente o pacote Python mais usado para gráficos 2D
- Fornece uma maneira muito rápida de visualizar dados do Python, mas não é a mais interativa
- É referência para outros pacotes. (Sabe matplotlib, não vai ter dificuldade com outras)
- Documentação gigantesca: matplotlib.org/Matplotlib.pdf

Matplotlib - Principais elementos

- Figura: o recipiente principal de um gráfico
- Axes (plural de axis): a área de “plotagem”, uma figura pode conter múltiplos eixos
- Objetos gráficos: linhas, retângulos, texto
- As funções são usadas para criar e manipular figuras, eixos, linhas etc. como scripts

Matplotlib - Arquitetura

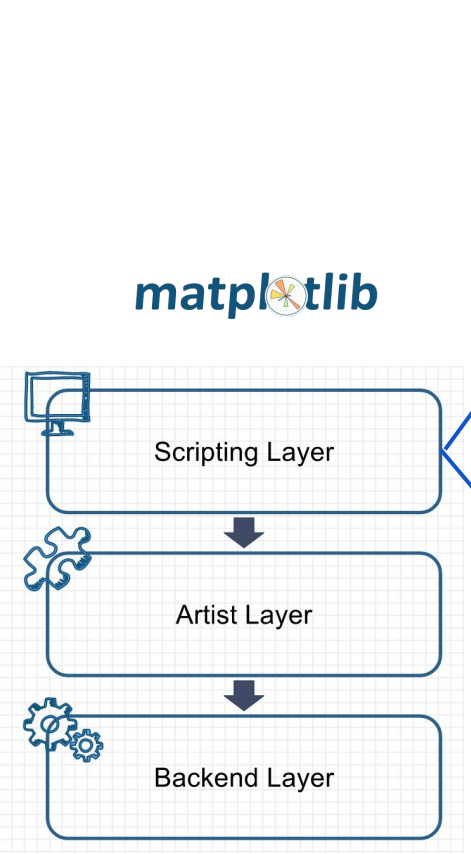


Simplifica o acesso às camadas Artista e Backend

Responsável por contêineres como Figuras, Subfiguras, Eixos, e primitivas de desenho como linhas, retângulos etc.

Lida com a renderização de gráficos em uma tela ou em arquivos

Matplotlib - Interfaces de uso



.pyplot

Interface tipo shell (iterativa) para o Matplotlib,

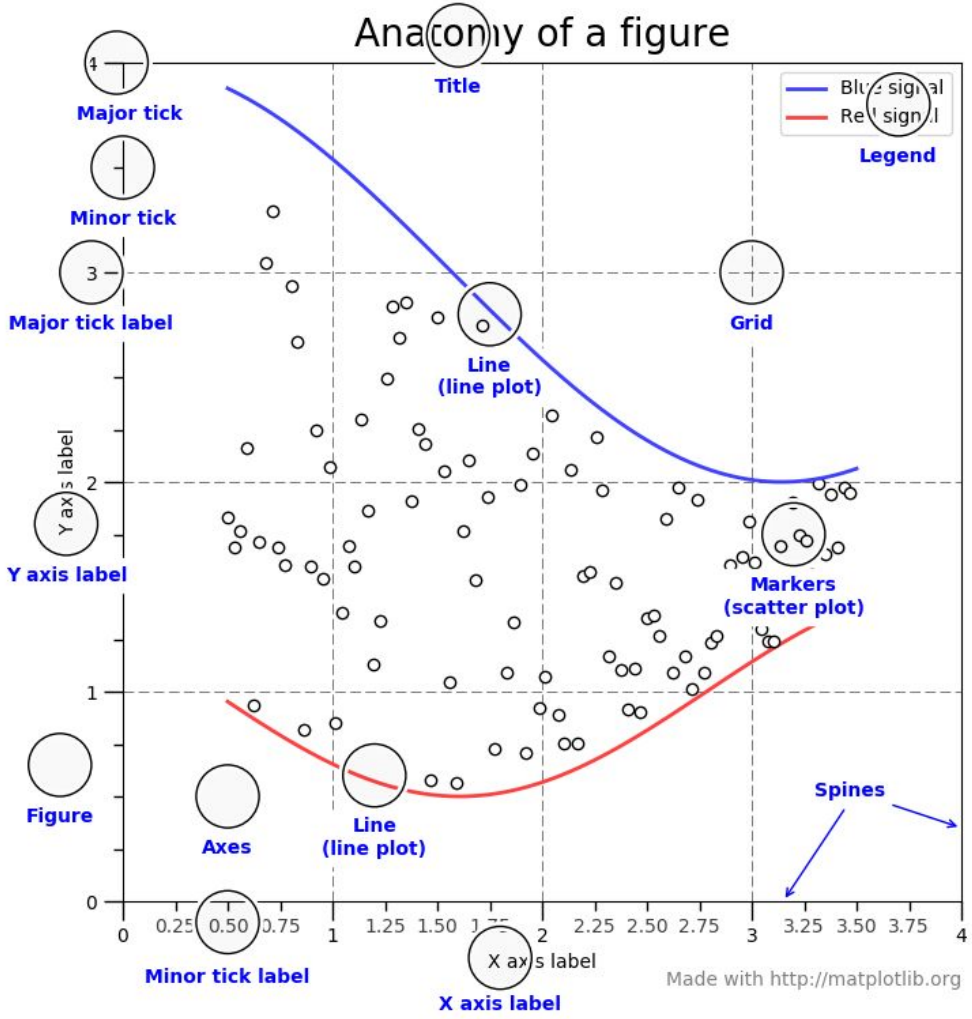
- Mais fácil de usar
- Mantém o estado entre as chamadas
- Útil para uso em notebooks Jupyter/IPython

.pylab

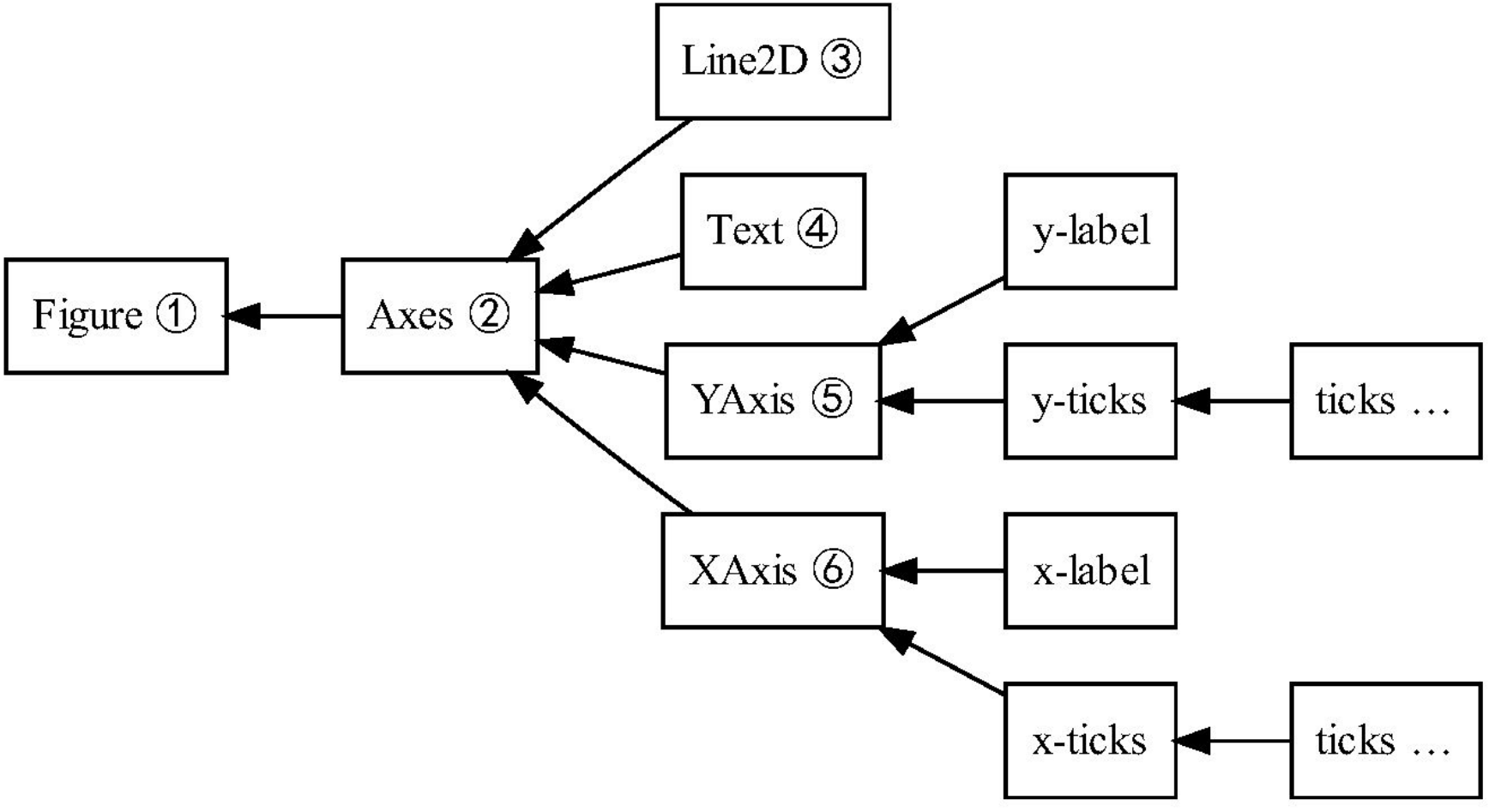
Interface “econômica” para o Matplotlib

- combina os namespaces PyPlot e NumPy em um único (para evitar a necessidade de importar dois namespaces)
- Possui integração entre os dois pacotes
- Uso do Pylab agora é desencorajado.

Matplotlib - Anatomia de um gráfico matplotlib

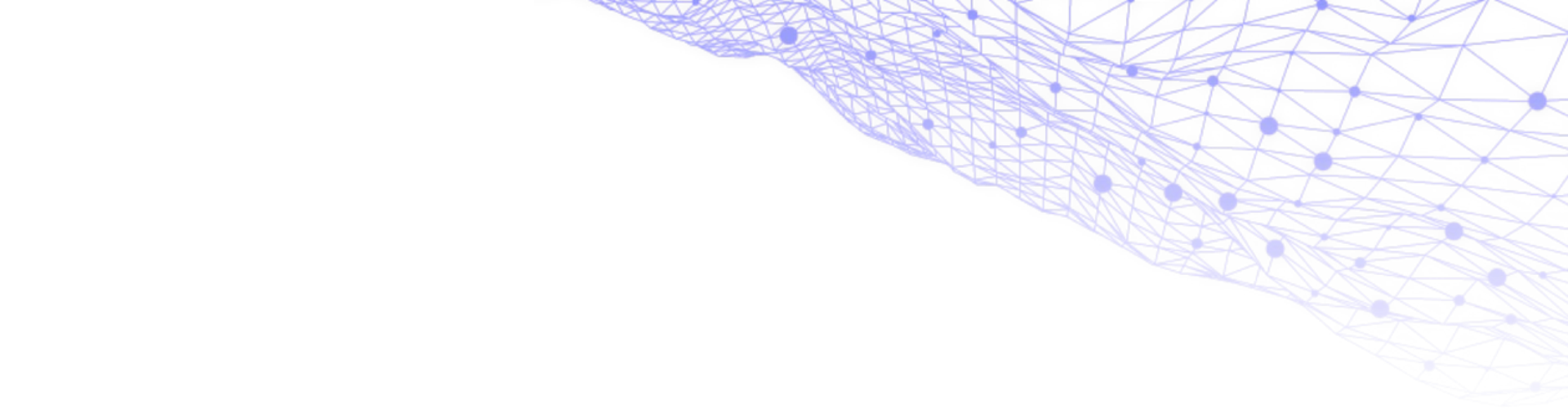


Matplotlib - Relacionamento entre objetos



Matplotlib - Atenção

- Todas as funções de plotagem esperam "np.array" ou "np.ma.masked_array" como entrada
- Classes como tipo *Array*, como objetos do Pandas e numpy.matrix, podem não funcionar como planejado. É melhor convertê-los em objetos numpy.array antes de plotar



Hora de colocar a mão na massa!

Demo2