PUC RIO

AULA 2

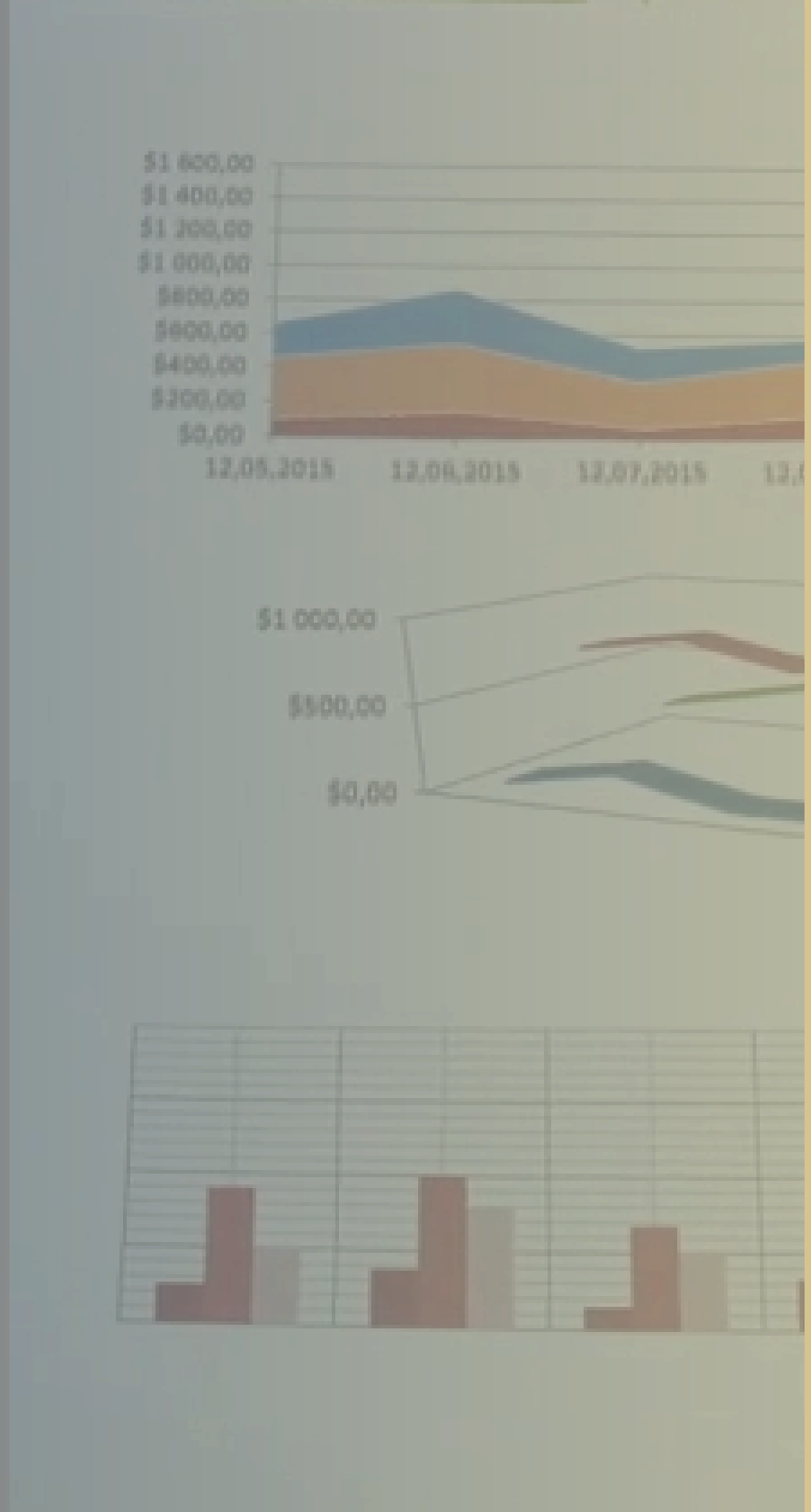
Construção de gráficos básicos para apoiar tarefas de análise

- Clients;
- Marketing;
- Organization and business process;
- Investments.

business unit, clients, distribution channel, products, geographic area; Development of sales and margins; Direct and indirect costs; Return of capital invested; Financial Expenses.

ess

tric
or changes



Objetivos de aprendizagem da aula

Ao final desta aula, você irá:

- Conhecer os tipos de visualizações comumente utilizados (barra, linha, *pizza*, etc.).
- Refletir sobre a adequação de certas visualizações a certas tarefas analíticas.

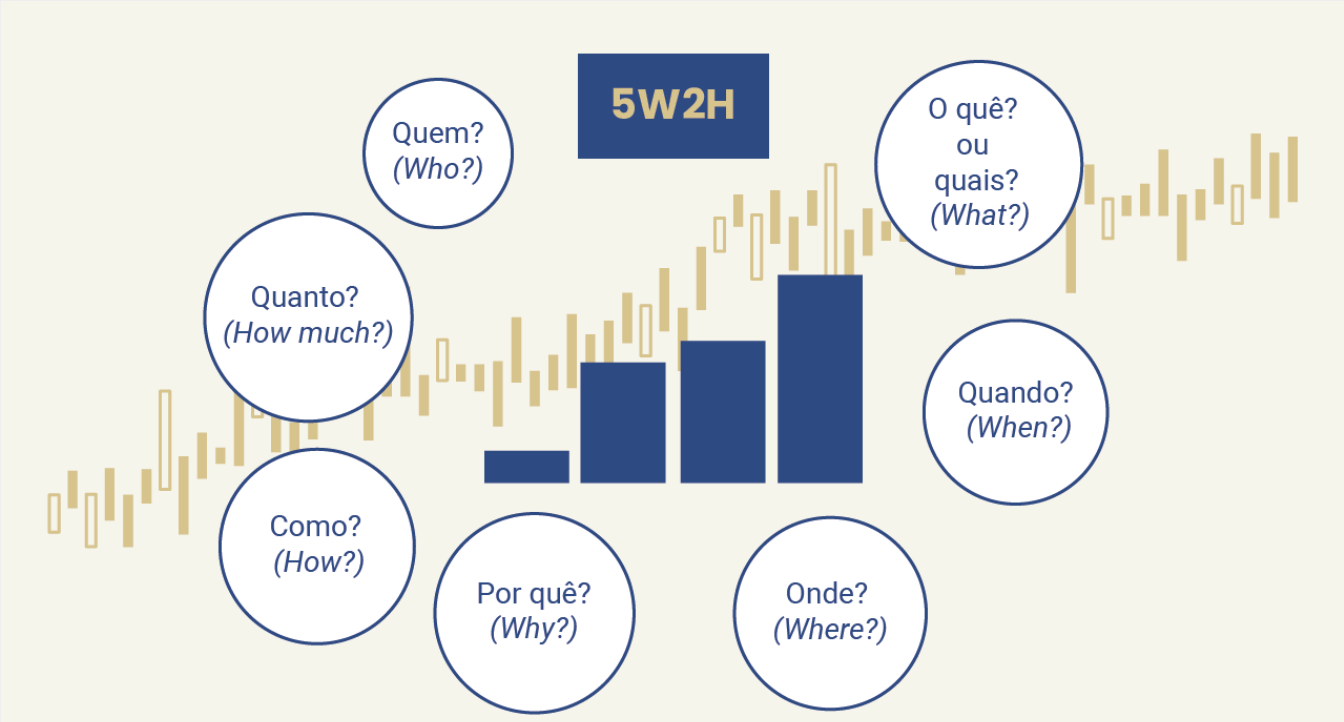
Nesta aula, apresentaremos algumas das visualizações mais comuns, indicando como foram construídas a partir do mapeamento de atributos de dados a variáveis visuais. Discutiremos, ainda, a adequação de visualizações a certas tarefas analíticas. Tudo desta aula foi preparado para que você tenha ótimos estudos. Vamos nessa?

Algumas pessoas acreditam que basta obter os dados para produzir boas visualizações. No entanto, antes de explorarmos os diversos tipos de visualizações, é importante pensarmos em quais perguntas podemos responder com cada uma. Para isso, podemos tomar como ponto de partida as perguntas 5W2H, comumente utilizadas por jornalistas e investigadores:

Que história é essa?

Algumas pessoas acreditam que basta obter os dados para produzir boas visualizações. No entanto, antes de explorarmos os diversos tipos de visualizações, é importante pensarmos em

quais perguntas podemos responder com cada uma. Para isso, podemos tomar como ponto de partida as perguntas 5W2H, comumente utilizadas por jornalistas e investigadores:

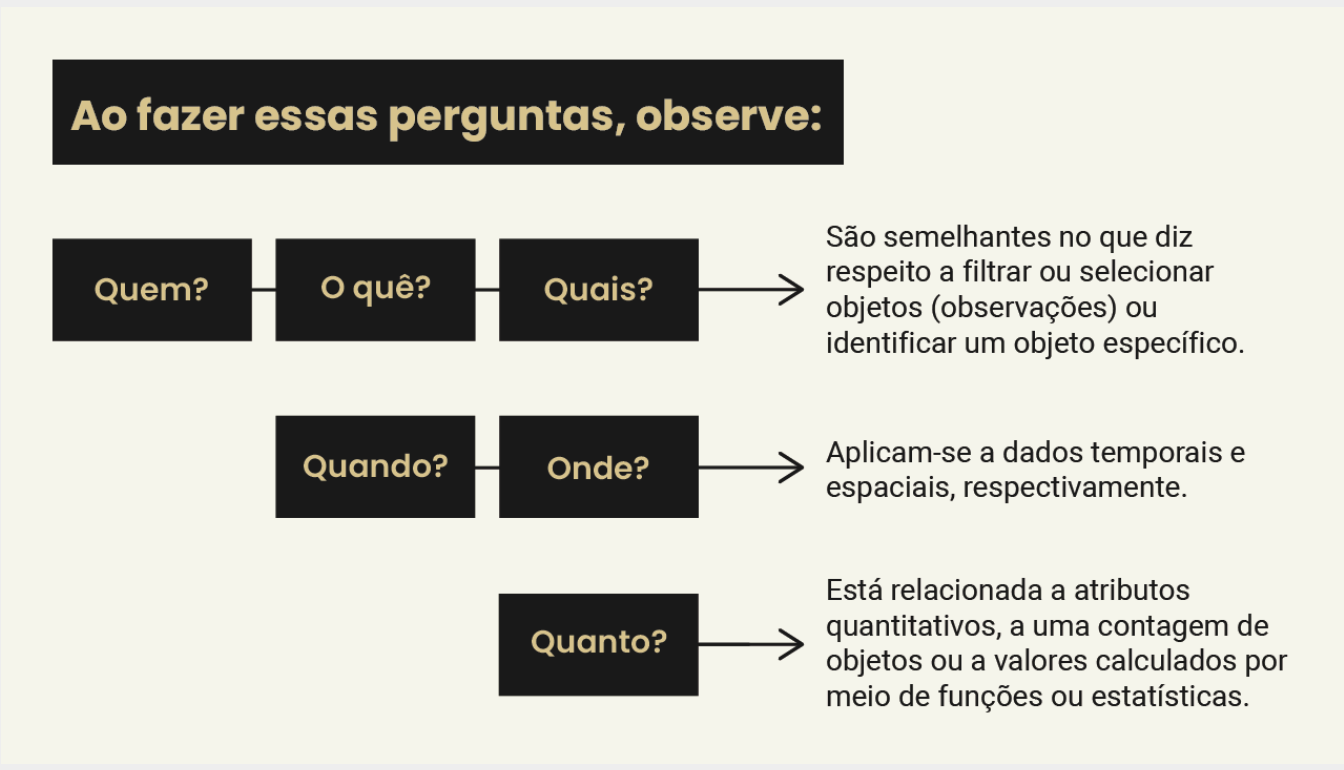


Quando combinadas com os tipos de dados que temos, essas perguntas determinam, em grande medida, o conjunto de visualizações que podemos considerar para respondê-las.

Fique ligado

Vale lembrar que as visualizações dirigidas por dados que vamos tratar aqui não nos permitem responder diretamente perguntas do tipo **por que?** ou **como?**
E, embora não possamos responder diretamente perguntas do tipo como?, temos a pergunta relacionada a tendências: **como varia/variou?**, geralmente associada a uma variável temporal ou mesmo a uma variável quantitativa.

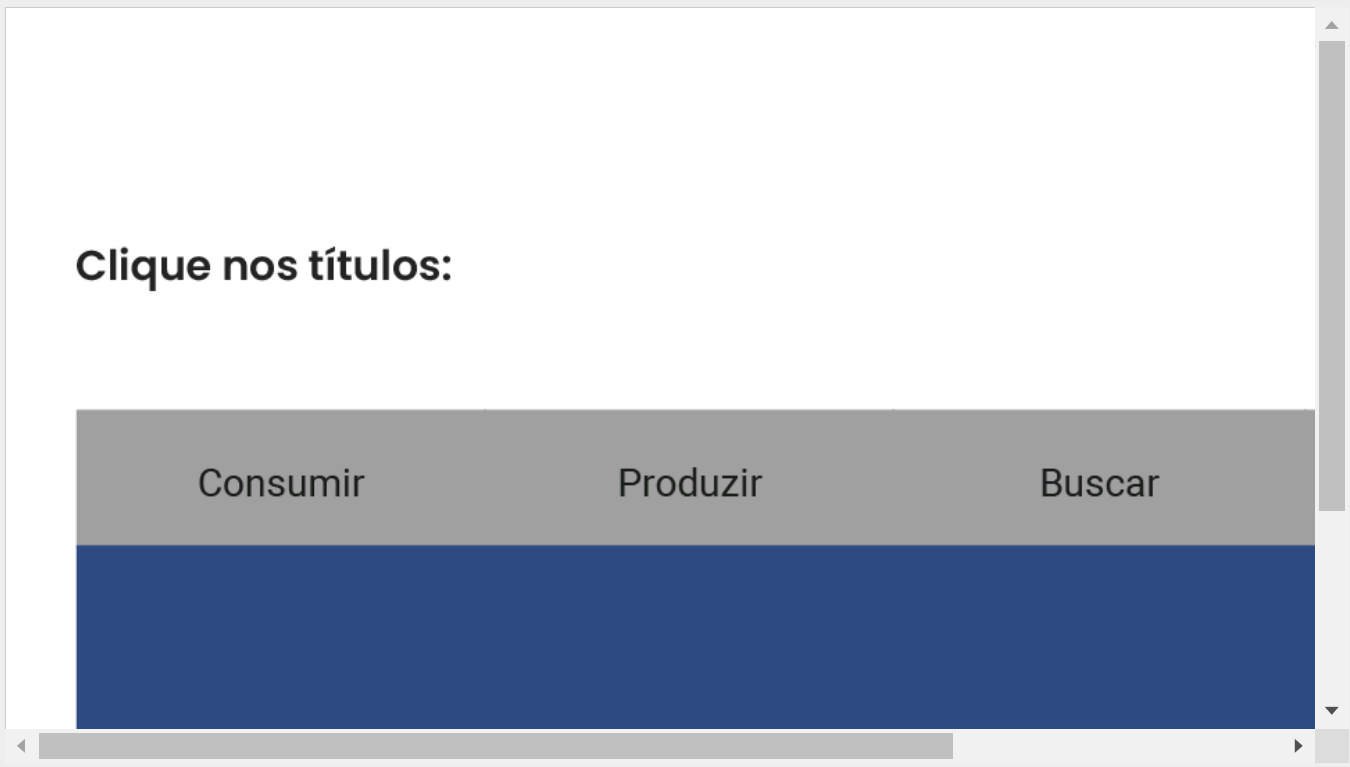
Para ajudá-lo nesta parte prática, explicamos detalhadamente o que significa cada pergunta e como respondê-la. Veja a seguir.



Ao examinarmos uma visualização, nem sempre temos uma pergunta específica em mente. A pergunta inicial pode ser bem geral, algo como "**O que temos aqui?**". Entretanto, logo em seguida, começamos a enfocar diferentes aspectos da visualização, levantar hipóteses e fazer perguntas mais específicas.

Tarefas de visualização

Veja como Munzner (2014) classifica as tarefas que as pessoas realizam com visualizações em quatro grupos de **alto nível**:



Interativo

Descrição do interativo

Diversas outras classificações de tarefas vêm sendo propostas ao longo dos anos. Amar, Eagan e Stasko (2005) conduziram um estudo com alunos de um curso de visualização de informação, que geraram 196 tarefas de análise válidas. Eles classificaram essas tarefas em 10 tipos:

- ***Retrieve value*** **(recuperar valor)**
- ***Filter*** **(filtrar)**
- ***Compute derived value*** **(computar valor derivado)**
- ***Find extremum*** **(encontrar extremos)**
- ***Sort*** **(ordenar)**
- ***Determine range*** **(determinar intervalo)**
- ***Characterize distribution*** **(caracterizar distribuição)**
- ***Find anomalies*** **(encontrar anomalias)**
- ***Cluster*** **(agrupar)**
- ***Correlate*** **(correlacionar)**

Fique ligado

Podemos utilizar essas tarefas de **baixo nível** de forma composta, como, por exemplo, “Ordene os estados brasileiros de acordo com a média de renda de sua população”, o que envolve as tarefas ***computar valor derivado*** e ***ordenar***.

Vamos ver agora como realizar essas tarefas sobre diferentes tipos de dados?

Contagens e comparações entre categorias

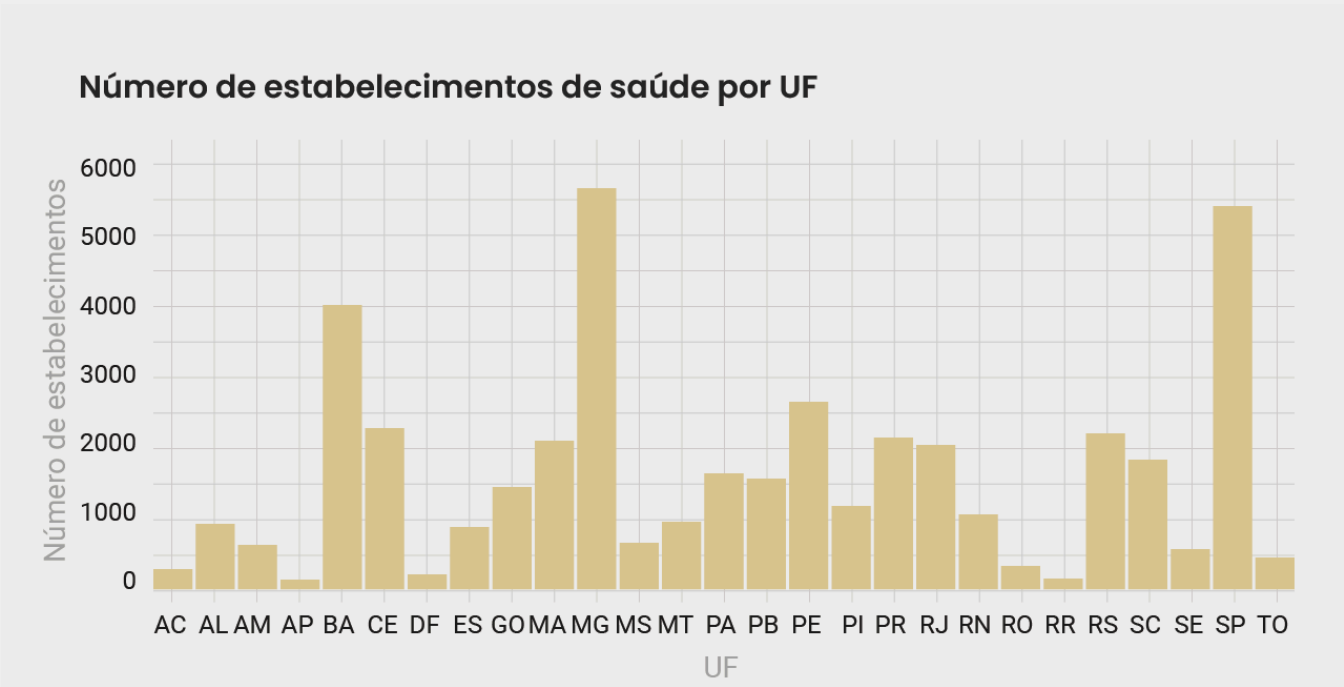
Quando temos dados categóricos, podemos contar ou comparar quantos objetos temos em cada categoria (tarefa *Recuperar valo*r), ranquear diferentes categorias (tarefa *Ordenar*), para encontrar as categorias com mais (ou menos) elementos (tarefa *Encontrar extremos*), e buscar identificar associações entre categorias. A tabela a seguir relaciona algumas perguntas e tipos de dados a visualizações usuais considerando esses objetivos de **contagem, comparações, ranqueamento, proporções e associações entre categorias**. A tabela utiliza as abreviações de tipos de dados apresentadas na aula anterior, e as subseções seguintes descrevem cada gráfico mencionado na tabela.

Pergunta	Tipos de gráficos
Quantos [objetos/observações] existem em cada **C0/T0**?	Gráfico de barras*
Qual é a [soma, média, mediana, ...] de **Q0** para cada **C0/T0**?	Gráfico de barras
Qual **C0/T0** tem mais (ou menos) [objetos/observações]?	Gráfico de barras ordenadas pela contagem de objetos (observações)
Qual **C0/T0** tem a maior (ou menor) [soma, média, mediana, ...] de **Q0**?	Gráfico de barras ordenadas pelo valor de **Q0**
Quantos [objetos/observações] existem em cada **C0**, por **C1/T0**?	Gráfico de barras agrupadas
Qual é a [soma, média, mediana, ...] de **Q0** para cada **C0**, por **C1/T0?**	Gráfico de barras agrupadas
Qual é a proporção de [objetos/observações] em cada **C0**?	Gráfico de *pizza; waffle chart;* gráfico de barras; gráfico de barras empilhadas
Qual é a proporção de [objetos/observações] em cada **C0**, por **C1/T0**?	Gráfico de barras empilhadas a 100% - pequenos múltiplos: gráficos de barras

*Sempre que se considerar uma variável temporal (T), deve-se preservar a associação da passagem do tempo com a leitura da esquerda para a direita. No caso de barras, por exemplo, devem-se usar barras verticais.

Gráfico de barras simples

Um dos objetivos mais comuns de um gráfico de barras simples é permitir contar o número de objetos (ou observações) em dada categoria. Vamos ver o exemplo de um gráfico de barras com a contagem de estabelecimentos de saúde por unidade federativa (UF). Note que a variável UF está mapeada na posição da barra no eixo X e que a variável de contagem de estabelecimentos está mapeada no comprimento da barra, conforme escala apresentada no eixo Y.



Fonte: https://dados.gov.br/ (setembro de 2022)

Como não existe uma ordenação natural entre as UFs, o gráfico apresenta as UFs em ordem alfabética. Tal ordenação possibilita localizar rapidamente a barra correspondente a uma UF.

Fique ligado

Embora o gráfico permita analisar e comparar os tamanhos das barras, isso não pode ser feito com muita precisão. Caso seja necessário identificar o tamanho exato das barras, pode ser útil acrescentar os valores associados a cada barra.

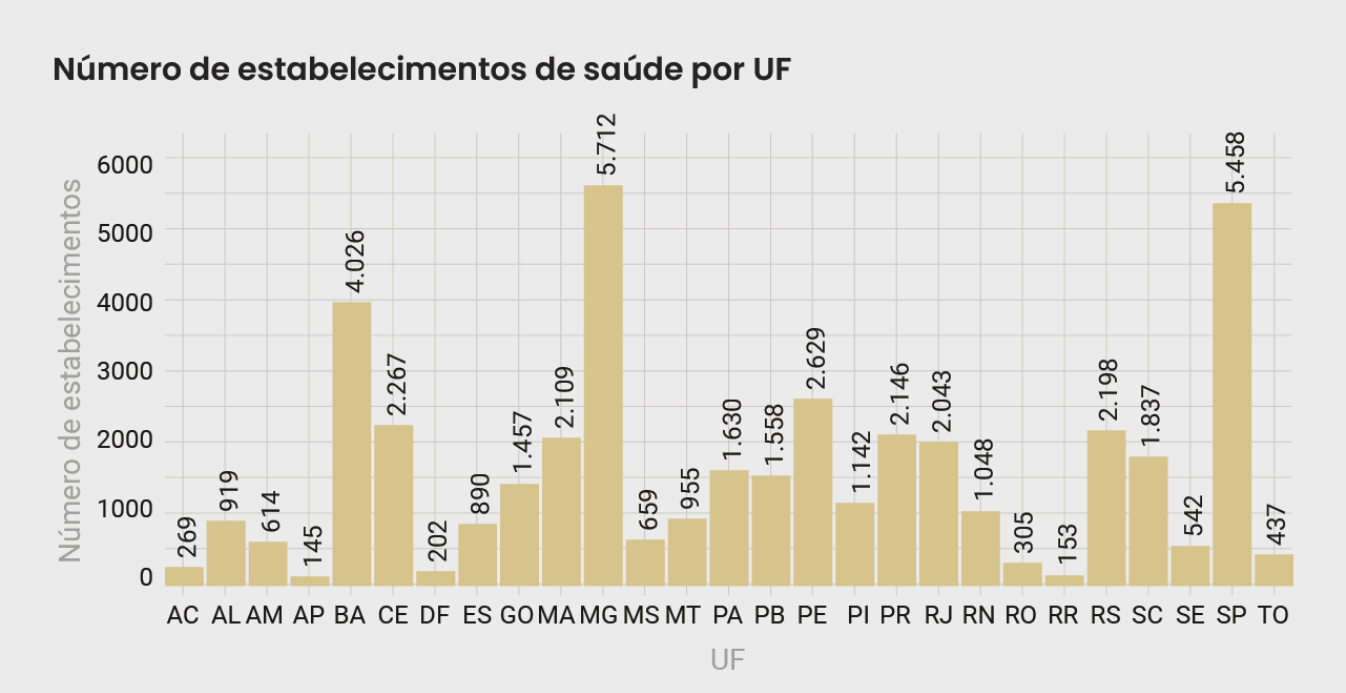
O vídeo a seguir apresenta um exemplo de código em Python para a construção de um gráfico de barras.

Gráfico de barras

Nesta aula, você vai aprender a construir um gráfico de barras utilizando a biblioteca gráfica Matplotlib.

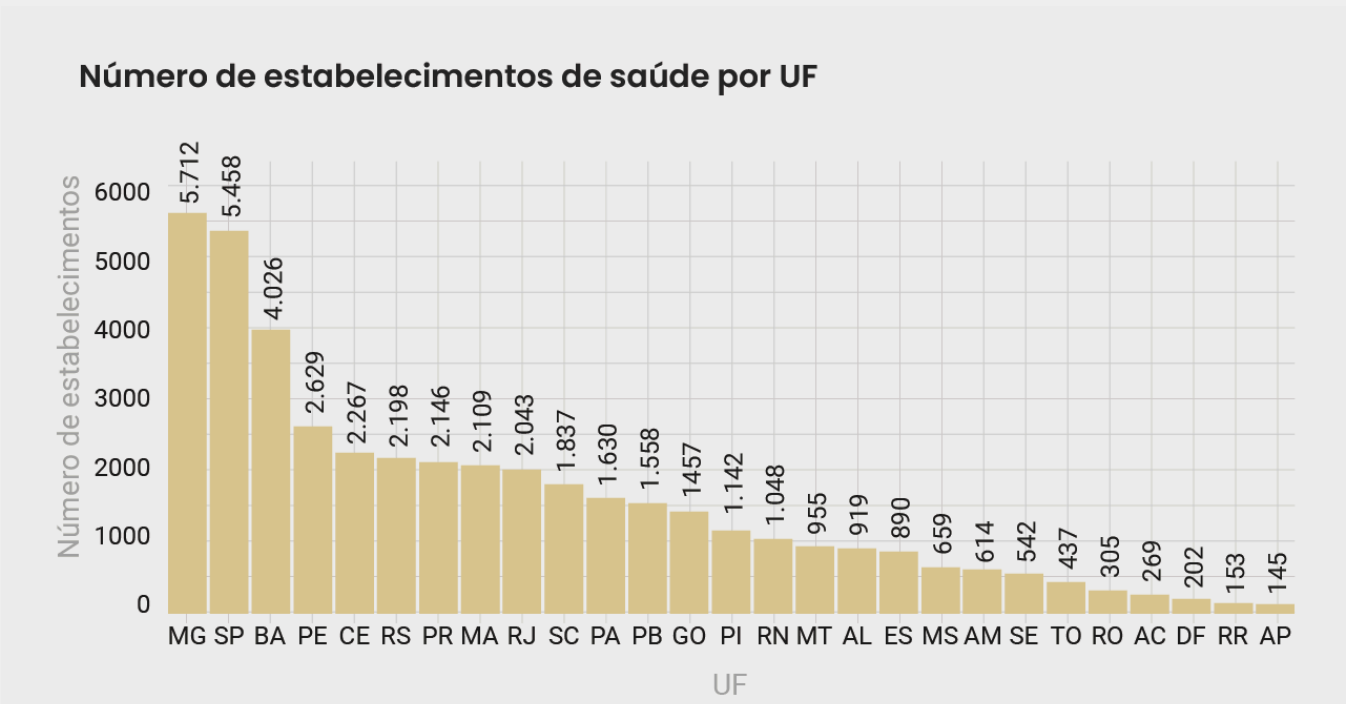
Assista o Vídeo

Agora, veja o gráfico a seguir:



Fonte: https://dados.gov.br/ (setembro de 2022)

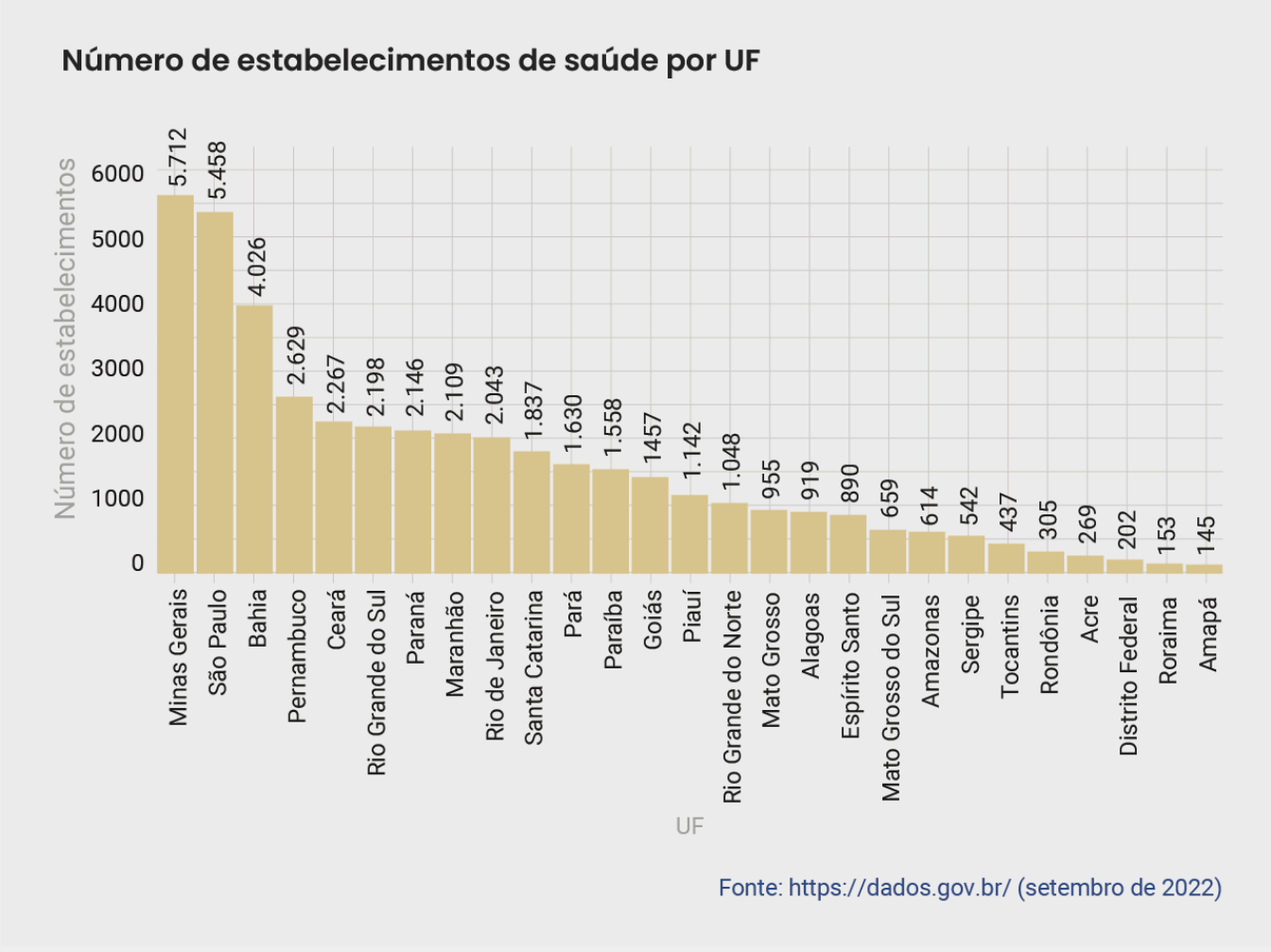
Lembre-se de que **a eficiência de um gráfico depende do objetivo de análise do leitor do gráfico**. Então, caso o objetivo seja descobrir quais UFs têm mais estabelecimentos de saúde, o gráfico da figura a seguir exige que o leitor compare os valores numéricos de barras de tamanhos semelhantes. Mas, atenção! Isso leva tempo e pode resultar em uma análise errada. Nesse caso, é melhor ordenar as barras pelo seu tamanho, como ilustrado a seguir.



Fonte: https://dados.gov.br/ (setembro de 2022)

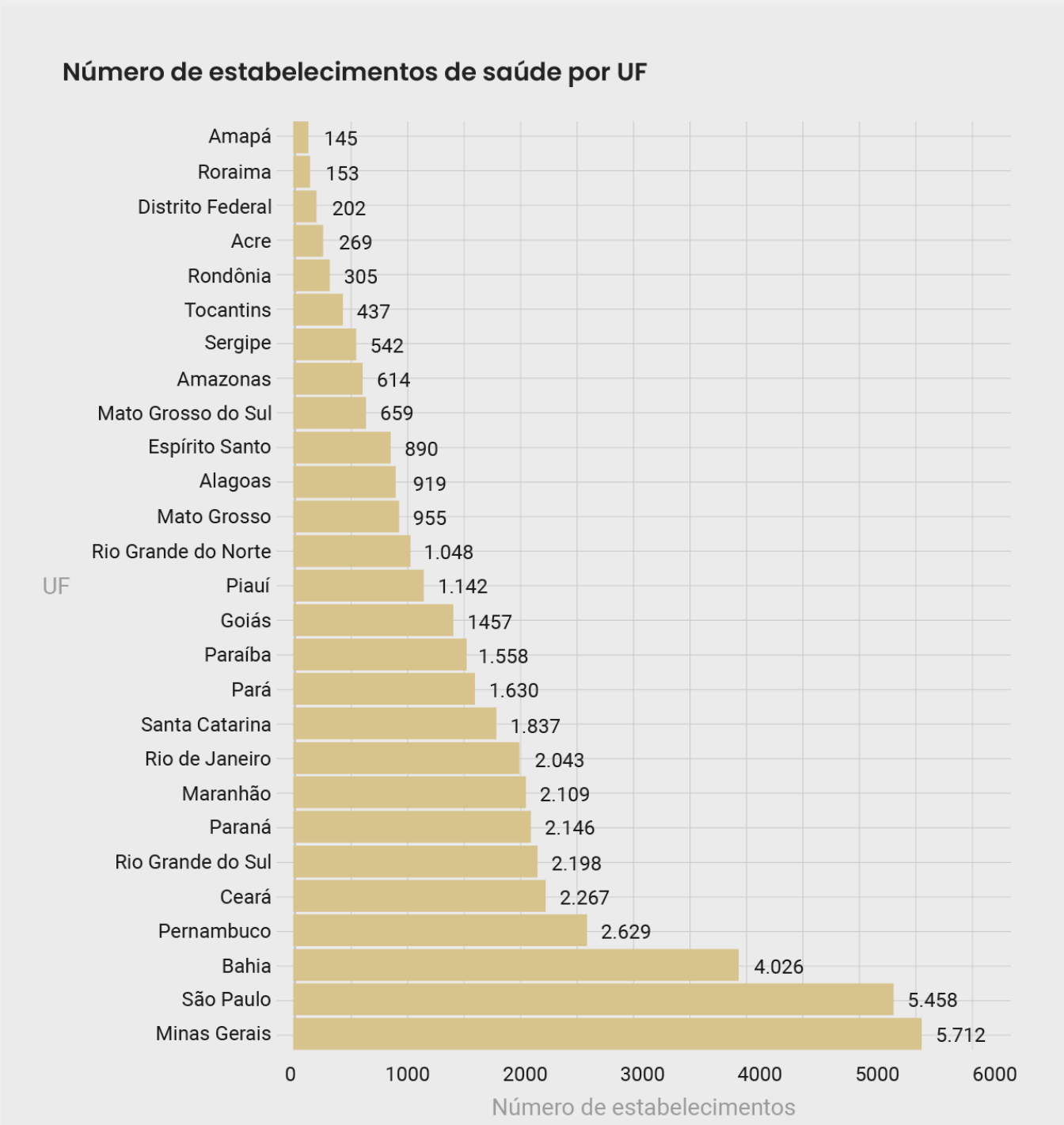
Como você pode observar, até aqui, os gráficos apresentaram barras na vertical. Como as UFs são siglas de apenas dois caracteres, **não foi necessário alterar o ângulo de leitura.**

Caso os rótulos sejam maiores, como no caso do nome do estado, podemos seguir duas abordagens. Na primeira abordagem alteramos o ângulo de leitura dos rótulos das barras:



Fonte: https://dados.gov.br/ (setembro de 2022)

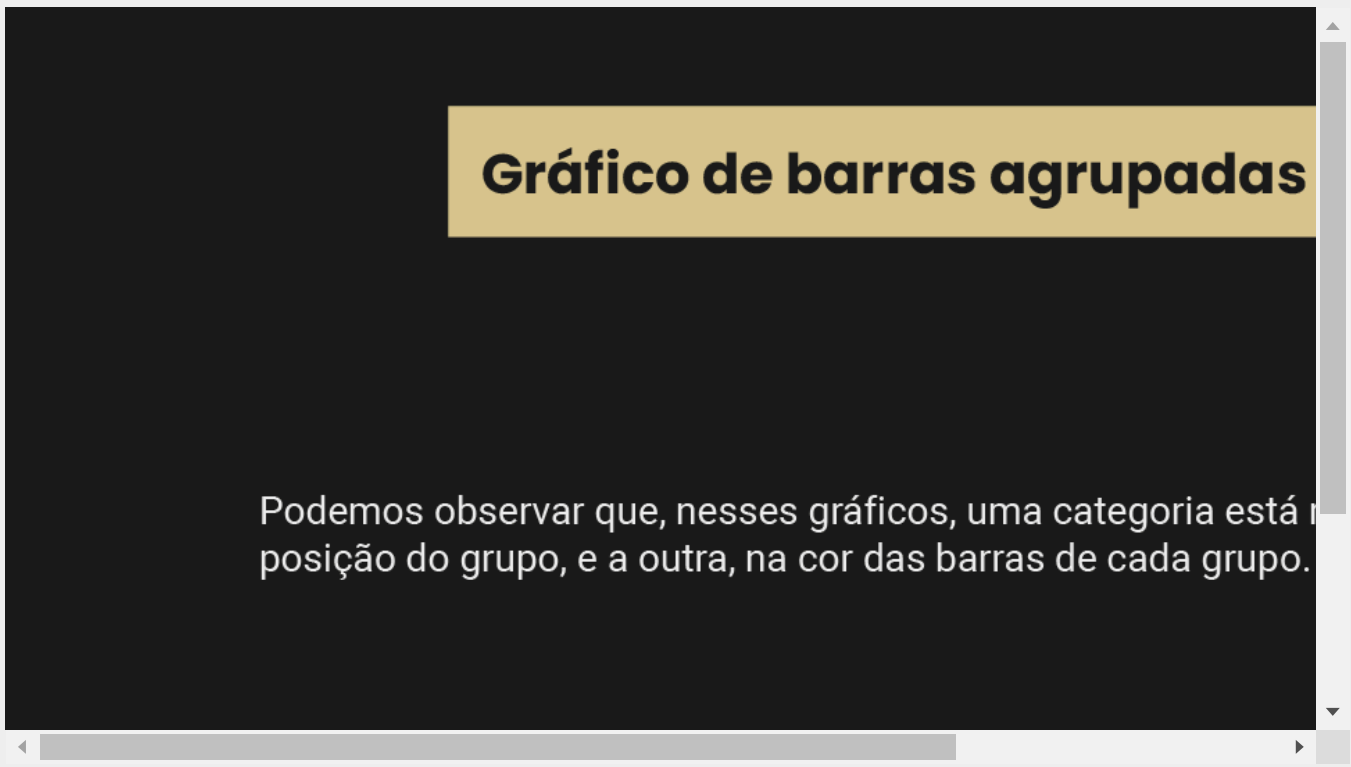
Na segunda abordagem alteramos as próprias barras para uma orientação horizontal, veja como fica:



Fonte: https://dados.gov.br/ (setembro de 2022)

Gráfico de barras agrupadas e pequenos múltiplos

Quando queremos comparar o número de objetos entre diferentes categorias, podemos utilizar gráficos de barras agrupadas, como iremos analisar a seguir.



Interativo

Descrição do interativo

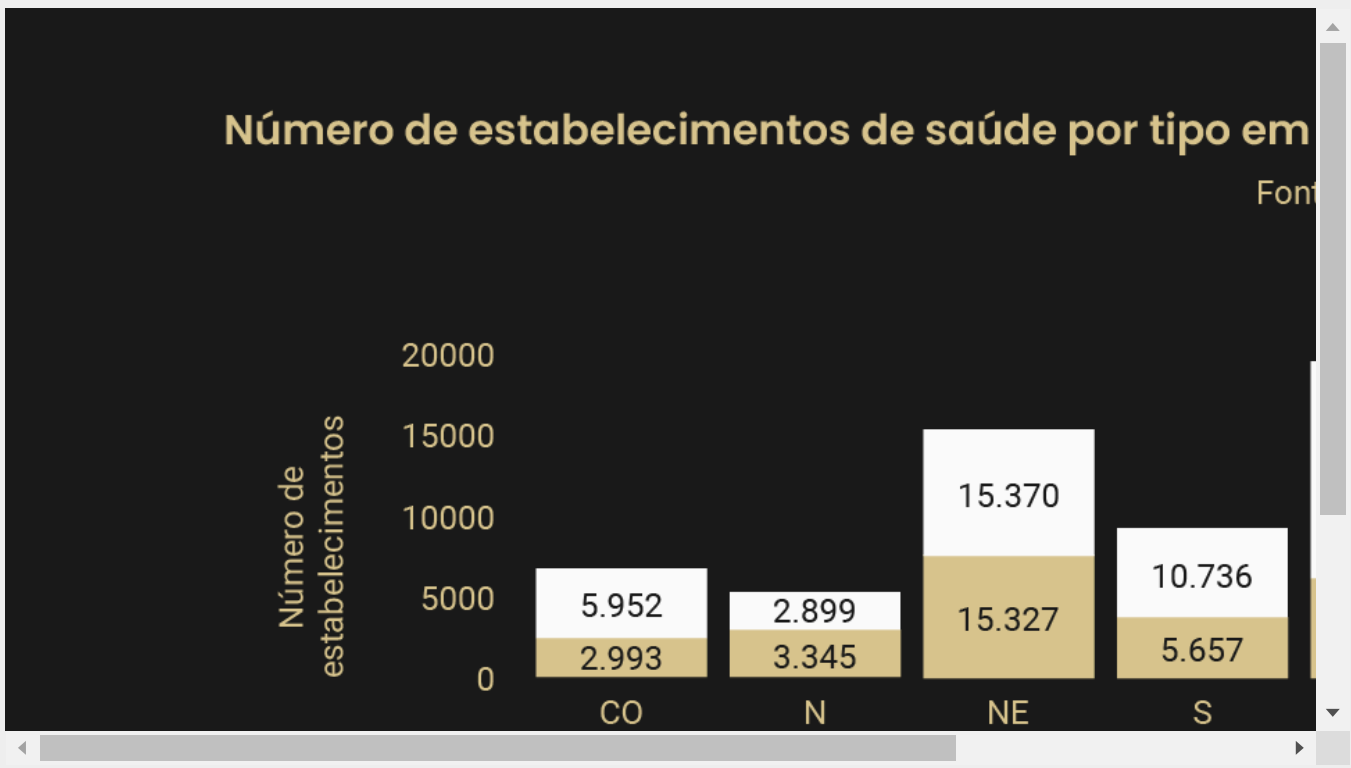
Ao utilizar pequenos múltiplos, você deverá tomar uma decisão importante a respeito dos eixos. Na figura anterior, do gráfico sobre o número de estabelecimentos de cada tipo por região, o eixo X dos dois lados tem a mesma escala, de 0 a 25.000. Isso permite comparar os comprimentos das barras. Se alterássemos as escalas dos eixos, os gráficos poderiam se tornar enganosos, pois barras de tamanhos semelhantes (visualmente) corresponderiam a valores diferentes.

Fique ligado

Caso as diferenças entre os grupos sejam muito grandes, e se torne necessário alterar as escalas, devemos tornar isso muito evidente no gráfico, seja por meio de anotações textuais, seja por meio de texturas diferentes nas barras de cada grupo ou outro marcador visual bem enfatizado.

Gráfico de barras empilhadas

Ainda considerando duas variáveis categóricas, como, por exemplo, região do país e tipo de estabelecimento de saúde, **quando queremos entender o quanto cada grupo contribui para o todo, podemos utilizar barras empilhadas.** Nesse tipo de gráfico, em vez de os segmentos ficarem lado a lado, eles são empilhados. Quando é importante termos a noção da soma dos segmentos, utilizamos valores absolutos, e as barras resultantes do empilhamento podem ter comprimentos diferentes, como você pode observar na figura a seguir.

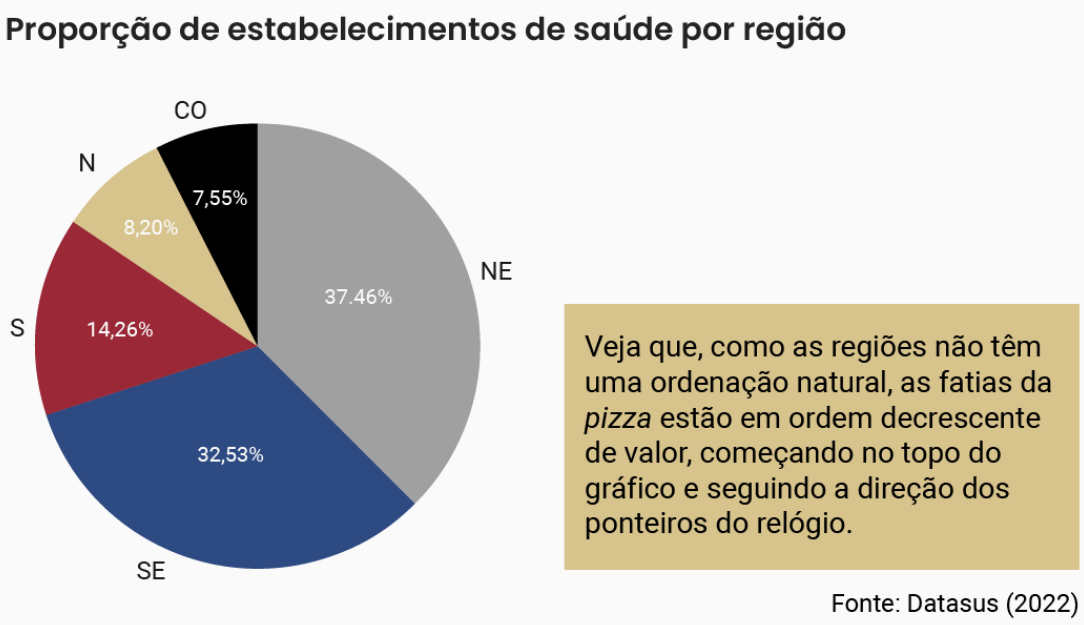


Interativo

Descrição do interativo

Gráfico de *pizza*

Para analisar a **proporção de casos ou valores de uma única variável**, como, por exemplo, estabelecimentos de saúde em diferentes regiões do país, com frequência vemos gráficos de *pizza* (*pie charts*).
Veja que, como as regiões não têm uma ordenação natural, as fatias da *pizza* estão em ordem decrescente de valor, começando no topo do gráfico e seguindo a direção dos ponteiros do relógio.



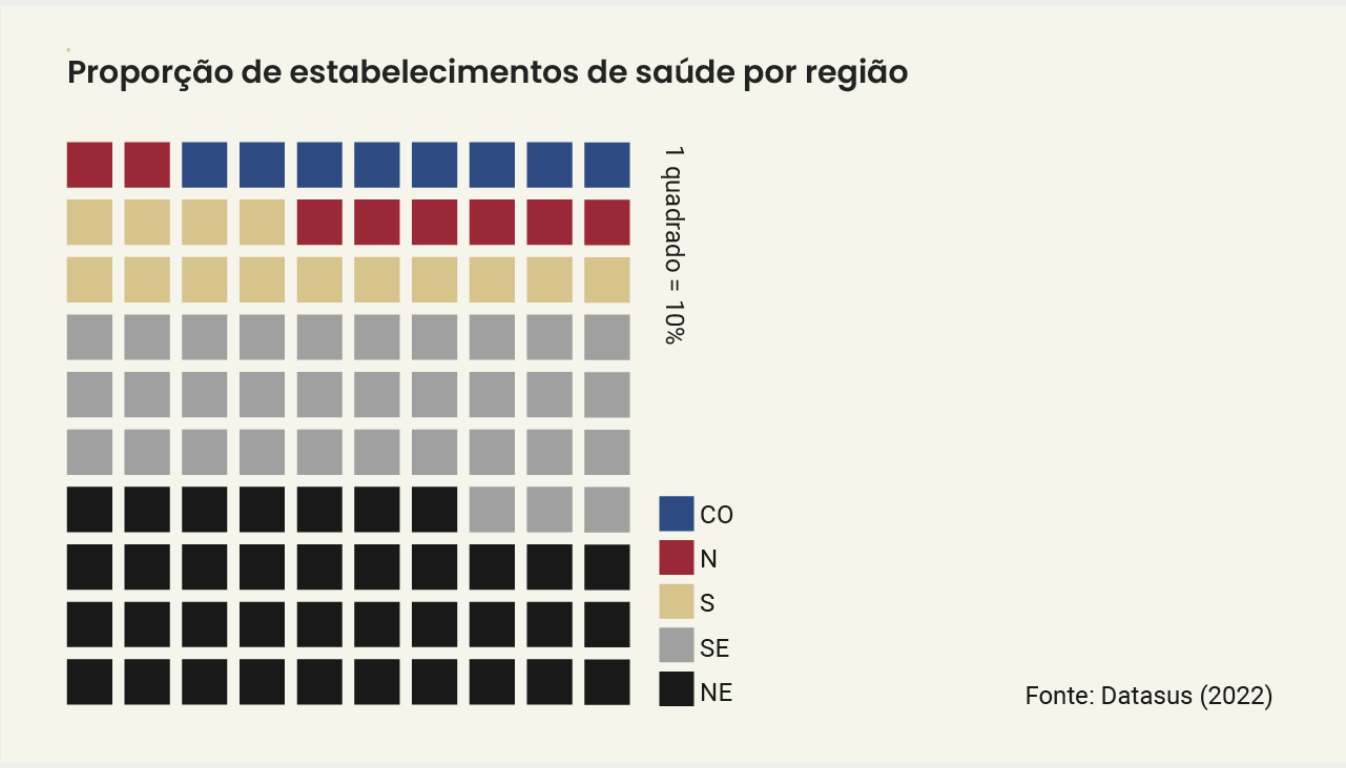
Você consegue distinguir ou comparar visualmente algumas fatias do gráfico? Tarefa difícil, não? Para lidar com essa limitação dos gráficos de *pizza*, geralmente incluímos nas fatias os valores numéricos. Além disso, observe que, embora cada fatia tenha uma cor, não é necessário incluir uma legenda indicando qual cor corresponde a qual categoria. Em vez disso, podemos colocar o nome da categoria ao lado da fatia correspondente.

Fique ligado

Você pode estar pensando: o que fazer quando as fatias são tão finas que não é possível colocar os nomes das categorias ao lado de cada uma?
Nesses casos, vale a pena avaliar se um gráfico de *pizza* é o mais adequado ou se seria melhor utilizar um gráfico de barras, mesmo que com valores percentuais.

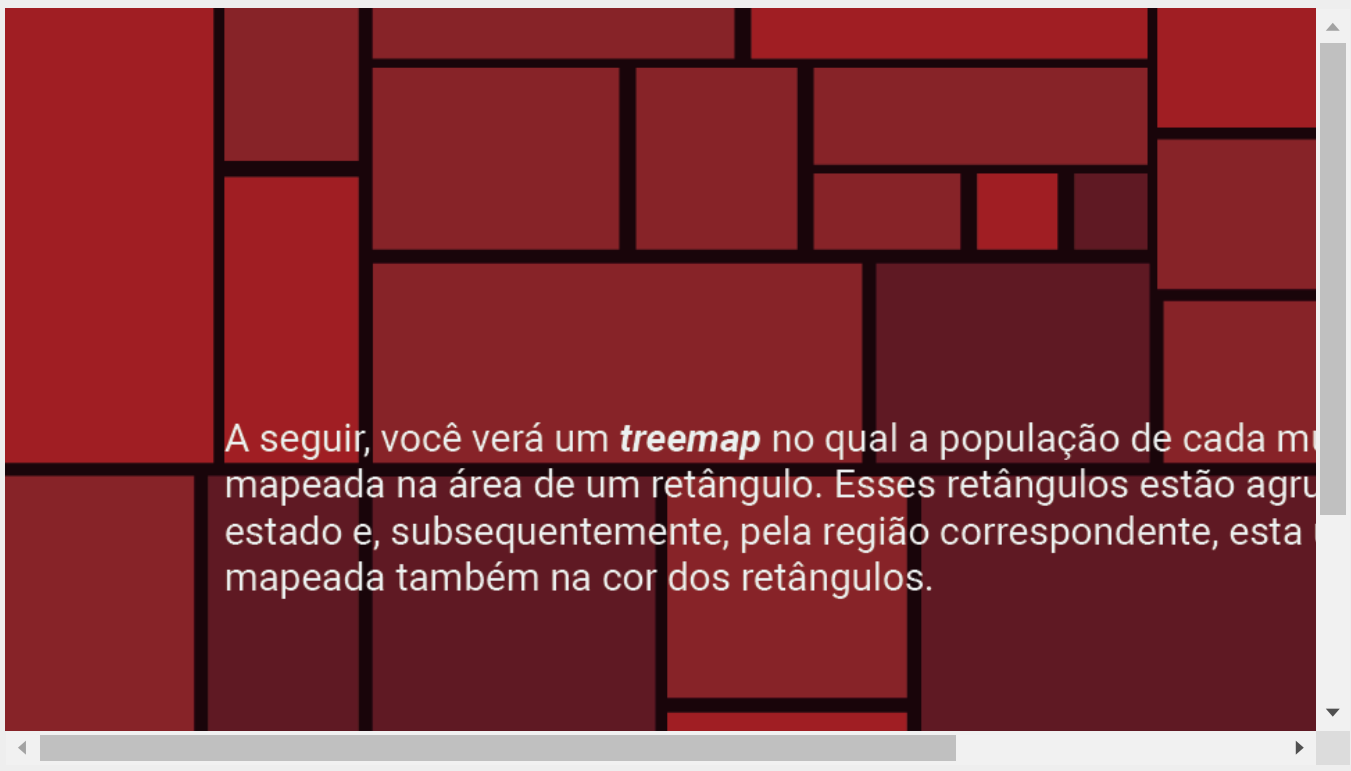
Waffle chart

Uma variação do gráfico de *pizza* é o *waffle chart*, que se trata de um *grid* de elementos (geralmente quadrados) que representam as proporções de cada categoria com respeito ao número de objetos ou a uma variável quantitativa.
Em geral, utiliza-se um **grid de 10 X 10** para facilitar a identificação dos valores, que somariam, então, **100%**.



Visualizando hierarquias

Algumas variáveis categóricas têm uma relação de hierarquia. **Uma das representações mais comuns de dados hierárquicos é uma visualização em árvore. Outra visualização comum é a de *treemap*, que procura maximizar o espaço ocupado com dados.**



Interativo

Descrição do interativo

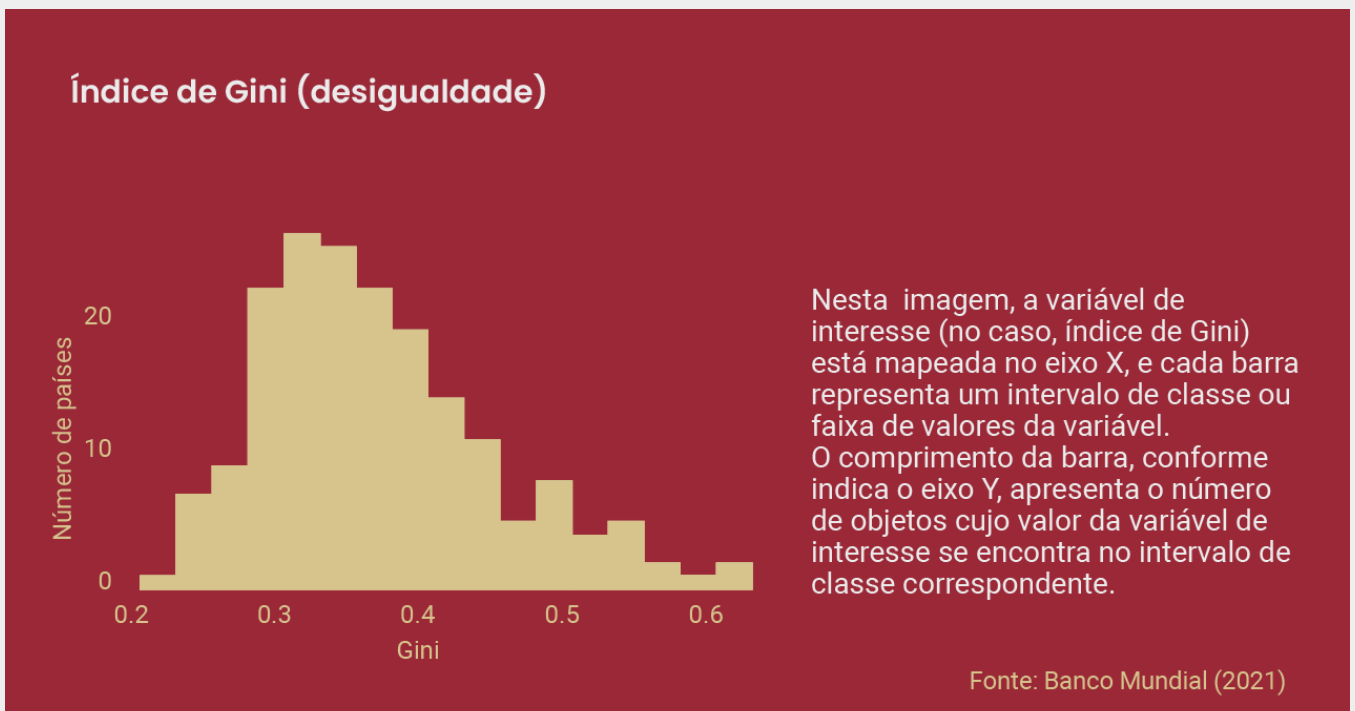
Identificação de distribuições

Até aqui, vimos visualizações relacionadas principalmente com variáveis categóricas. Nesta seção, veremos visualizações relacionadas com variáveis quantitativas.
A tabela a seguir relaciona algumas perguntas e tipos de dados a visualizações usuais considerando o objetivo de **identificação de distribuições.**

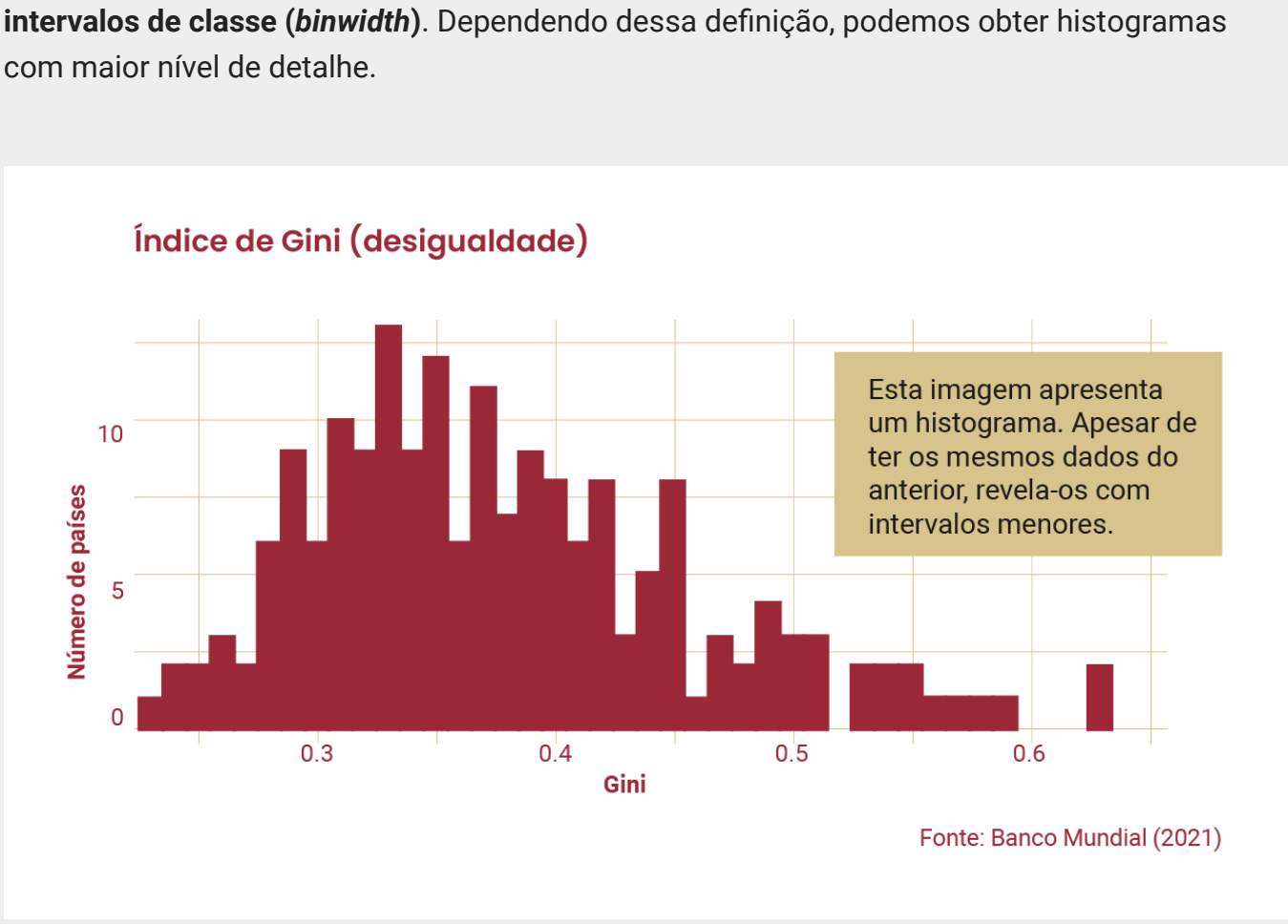
Pergunta	Tipos de gráficos
Qual é a distribuição dos valores de **Q0**?	Histograma; curva de densidade; *strip plot*
Qual é a distribuição dos valores de **Q0** para cada **C0/T**?	*Boxplot; violin plot; strip plot* – pequenos múltiplos: histogramas; curvas de densidade; *strip plots*

Histograma

Para visualizar a **distribuição de uma variável quantitativa**, podemos utilizar um histograma. Observe que, diferentemente de um gráfico de barras, não há espaço entre as barras de um histograma, o que reforça a natureza contínua do dado sendo visualizado.



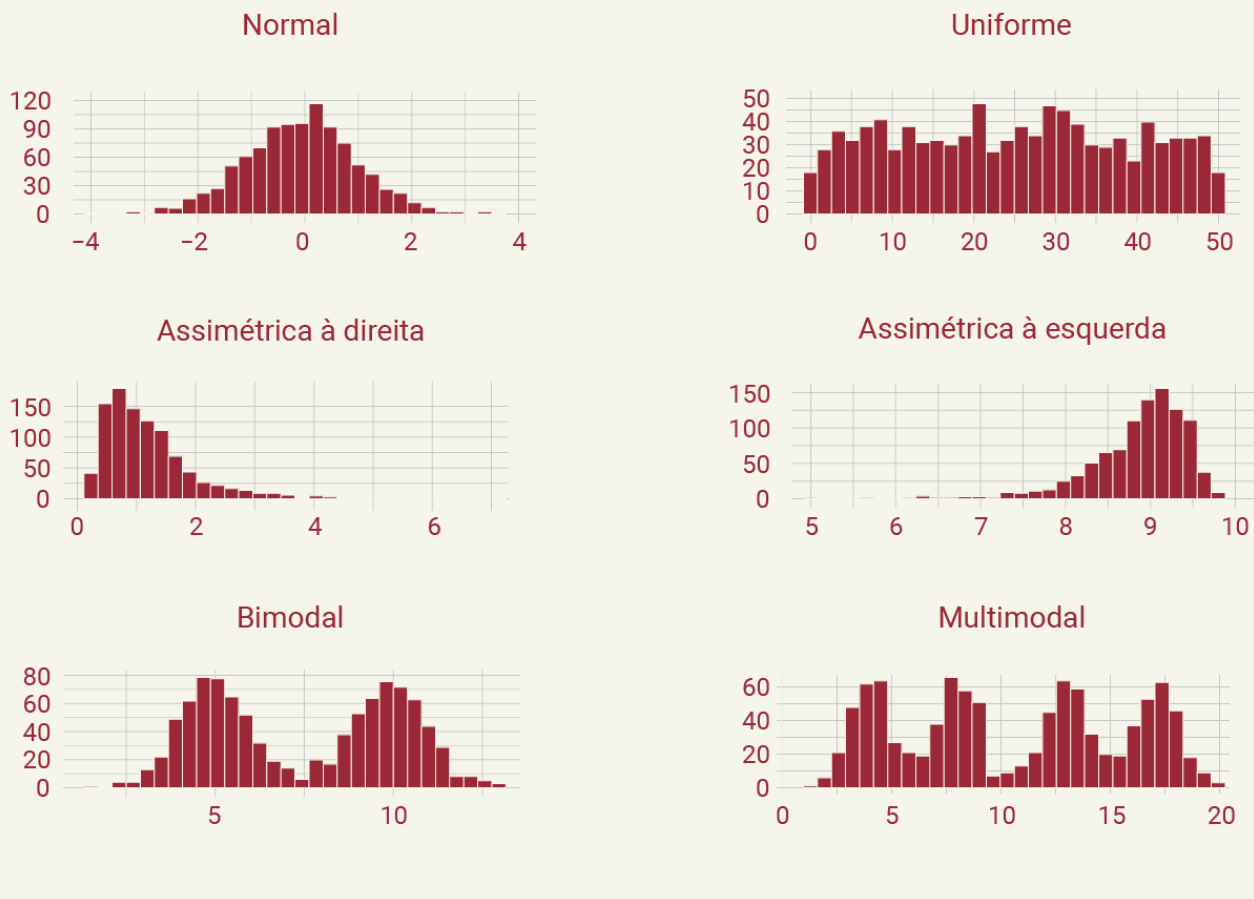
Ao construir um histograma, podemos definir o número de **barras (*bins*)** ou o tamanho dos **intervalos de classe (*binwidth*)**. Dependendo dessa definição, podemos obter histogramas com maior nível de detalhe.



A forma de um histograma permite que você identifique algumas características da distribuição, como **centralidade, amplitude (ou dispersão) e simetria**. Vamos analisar, agora, seis histogramas.

Histogramas de algumas distribuições aleatórias

A partir das imagens gráficas, é fácil identificar, por exemplo, a simetria da distribuição normal, bem como estimar que tem média e mediana próximo de zero e desvio padrão próximo de 1.



Strip plot

A distribuição de uma variável pode ser visualizada também em um *strip plot*, que é um gráfico de dispersão unidimensional, ou seja, de pontos dispostos em uma linha reta. Para tornar a concentração de pontos mais evidente, utilizamos algum grau de transparência no desenho dos pontos. Os trechos da distribuição com mais objetos com valores semelhantes aparecerão com a cor mais forte, devido à sobreposição dos pontos correspondentes.

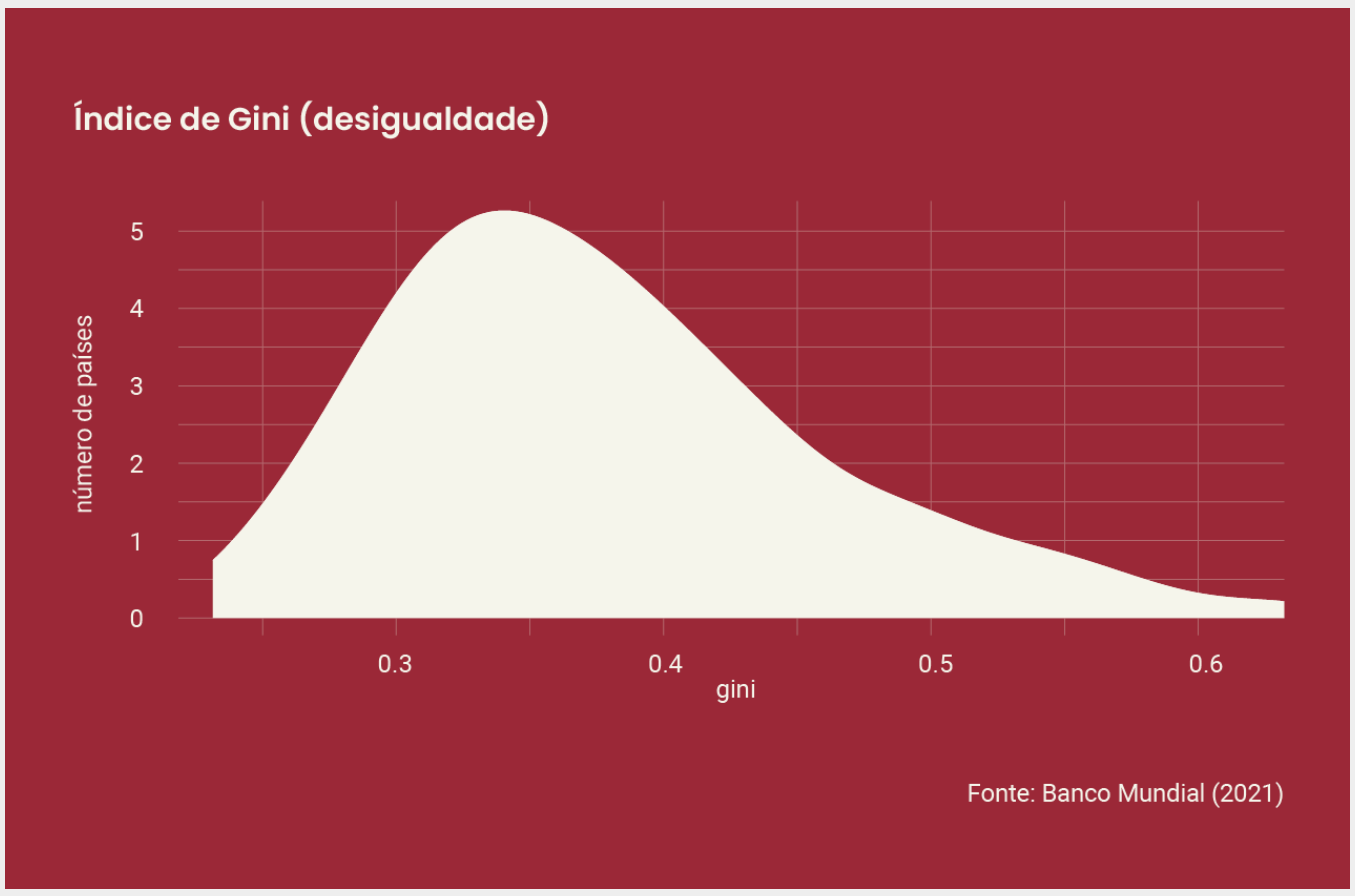
Índice de Gini (desigualdade)

Para esse tipo de visualização, é importante ajustar a opacidade dos pontos, para que a sua sobreposição se torne mais evidente e dê indícios da densidade de pontos de cada valor.

Fonte: Banco Mundial (2021)

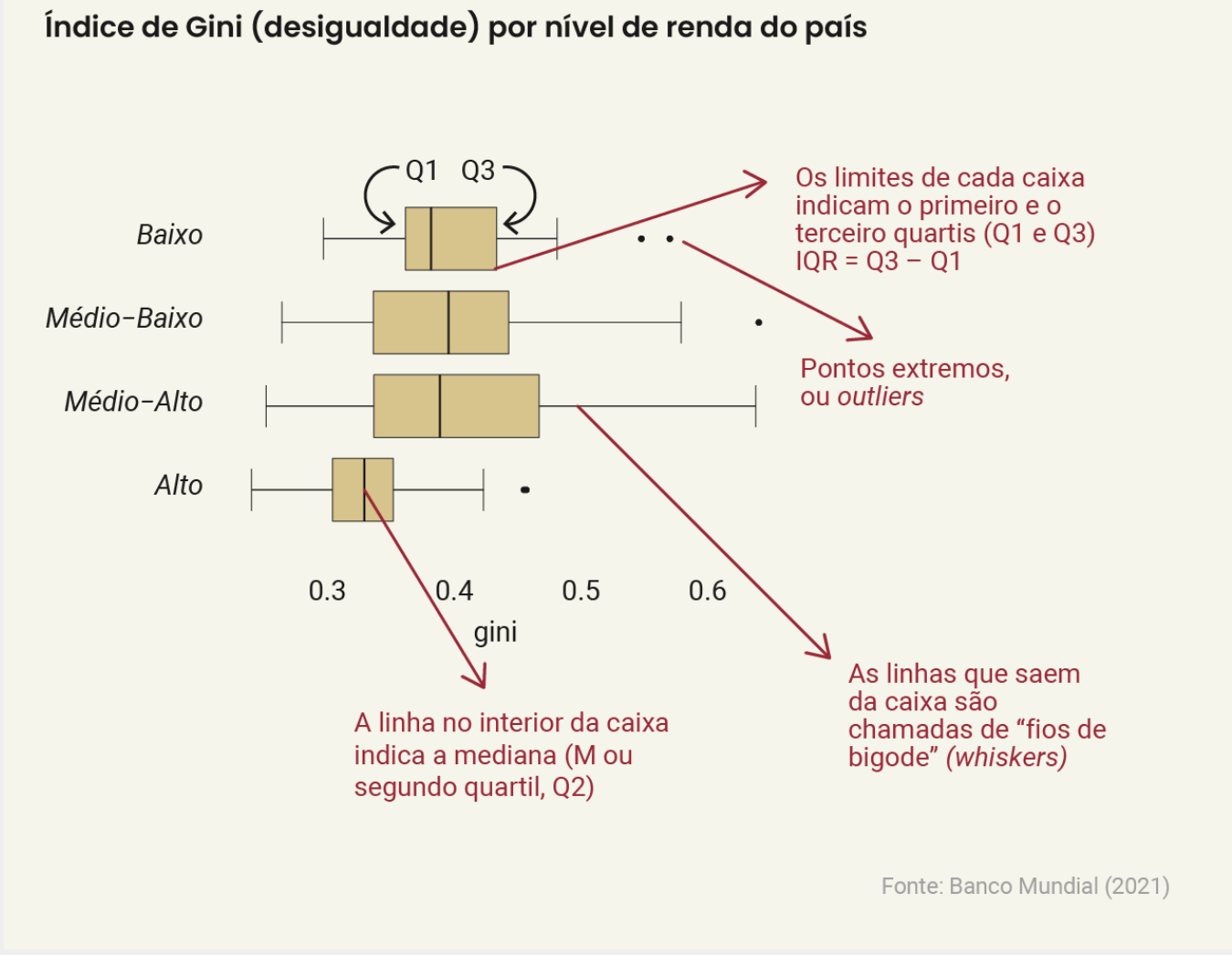
Curva de densidade

Uma distribuição também pode ser visualizada por meio de uma curva de densidade, na qual vemos a porcentagem de objetos cujos valores se encontram em determinada faixa de valores. Uma curva de densidade é análoga a um histograma com intervalos de classe infinitesimais. Veja como a curva de densidade permite analisar a forma da distribuição equivalente aos histogramas sobre o índice de Gini, apresentado anteriormente.



Boxplot (diagrama de caixa)

Quando queremos **comparar diferentes distribuições**, um gráfico comumente utilizado é o *boxplot* (ou diagrama de caixa). Nesse *boxplot* do índice de Gini, cada caixa representa a distribuição do índice em um conjunto diferente de países, conforme seu nível de renda.



Os limites de cada caixa indicam o primeiro e o terceiro quartis (Q1 e Q3), e a linha no interior da caixa indica a mediana (M ou segundo quartil, Q2). O intervalo interquartil (IQR=Q3-Q1, *interquartile range*) é representado pelo comprimento da caixa.

Mantendo o foco

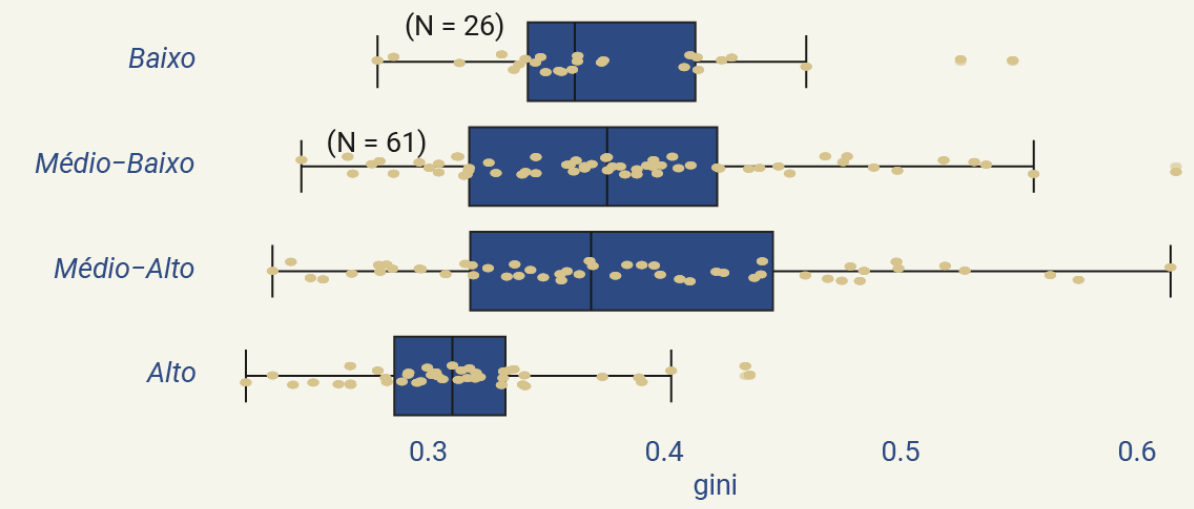
As linhas que saem da caixa são chamadas de “fios de bigode” (*whiskers*). Em geral, cada uma tem comprimento máximo de 1.5 x IQR, limitada pelo último ponto dentro do intervalo correspondente [Q1-1.5 x IQR, Q1] ou [Q3, Q3+1.5 x IQR].

Caso haja valores fora desses intervalos, eles são considerados pontos extremos, ou *outliers*, e aparecem como pontos no gráfico.
Podemos observar na figura que, em sua maioria, os países com nível de renda alto têm índices de Gini menores, ou seja, são menos desiguais. Além disso, seus índices são mais concentrados (pelo comprimento mais curto da caixa). Já os países com nível de renda médio a alto têm a maior amplitude (diferença entre o valor mínimo e o máximo) de todos os grupos e mediana semelhante à dos países de menor nível de renda.

Embora *boxplots* permitam analisar quartis, medianas e *outliers*, não permitem comparar o número de objetos representados em cada grupo.

Agora, para compreender melhor, vamos entender como fica se sobrepomos os pontos às caixas?
Então, caso você queira saber esses números com precisão, é possível colocar o número exato nos rótulos de cada grupo (p. ex., baixo [N = 26], médio-baixo [N = 61], etc.).

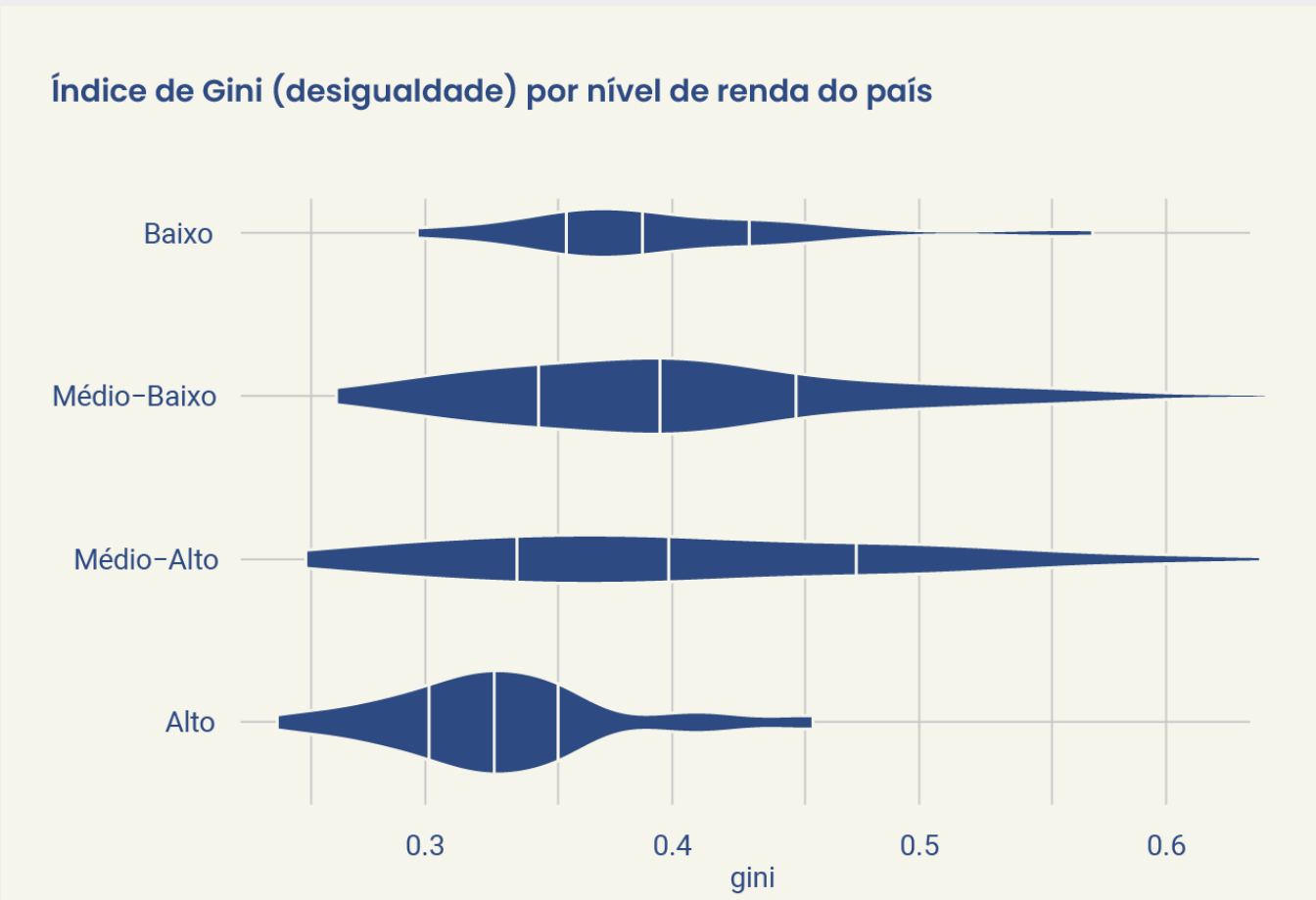
Índice de Gini (desigualdade) por nível de renda do país



Banco Mundial (2021)

Violin plot

Outro gráfico que apresenta diversas distribuições e que permite visualizar melhor as formas das distribuições é o *violin plot* (gráfico de violino), ilustrado na figura a seguir. Ele se assemelha a curvas de densidade espelhadas em um eixo central.

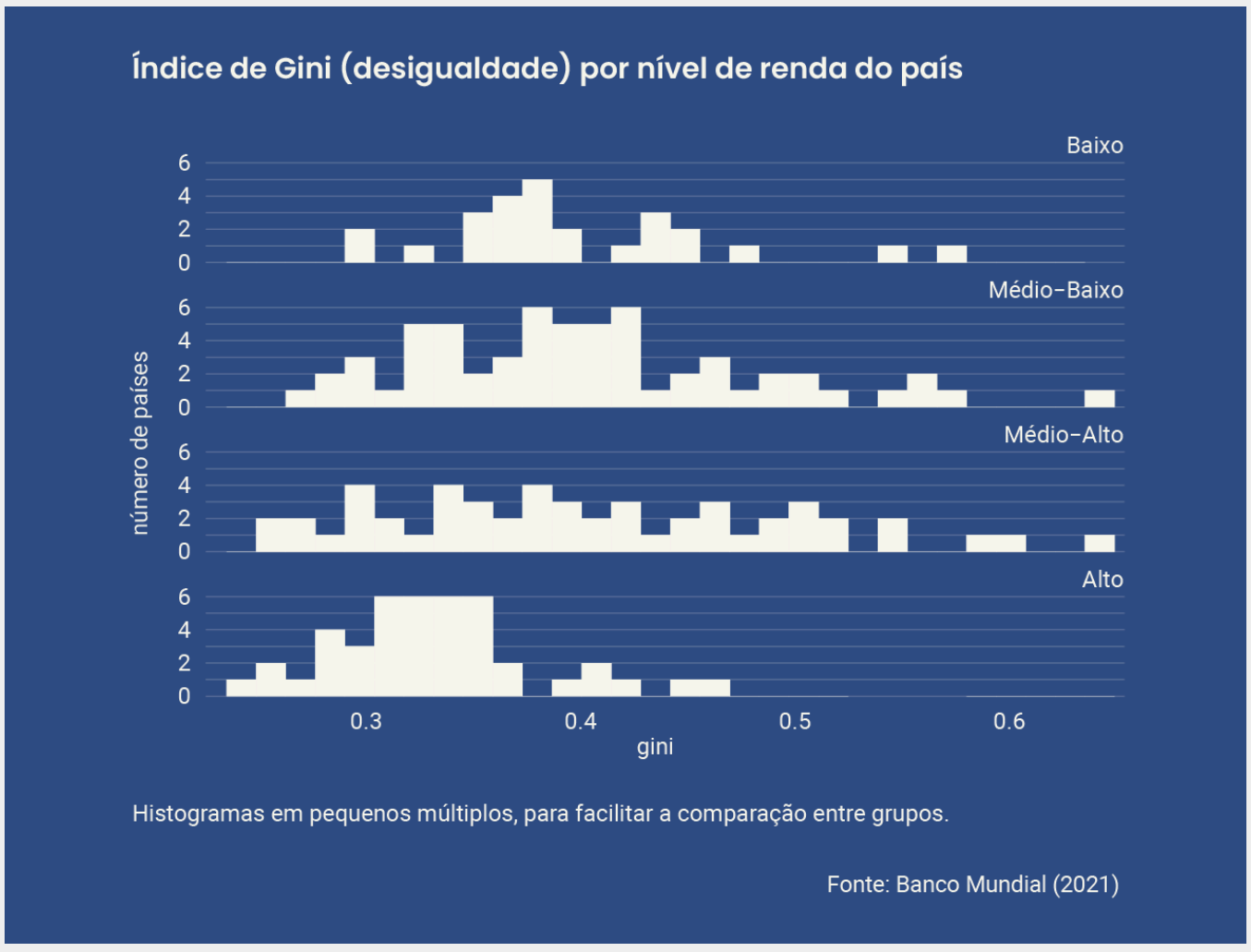


Fonte: Banco Mundial (2021)

Nesta imagem, definimos a área de cada forma como proporcional ao número de observações do grupo e incluímos também as linhas correspondentes aos quartis, para facilitar comparações.

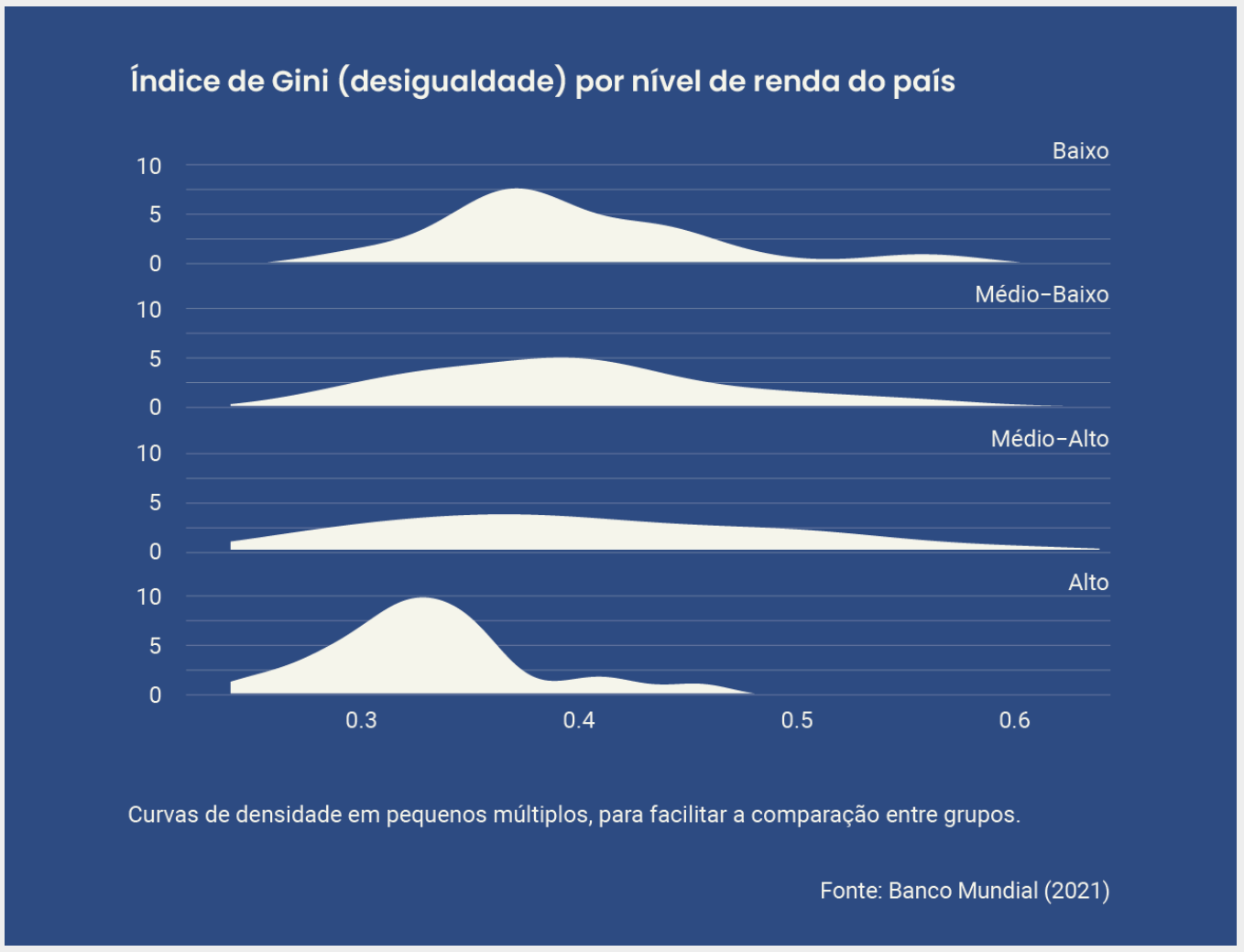
Visualizando distribuições por meio de pequenos múltiplos

Assim como visto na seção sobre gráficos de barras, também podemos utilizar pequenos múltiplos de histogramas, curvas de densidade e *strip plots* para comparar as distribuições de uma variável em diferentes grupos. Agora, vamos analisar esses gráficos para o mesmo conjunto de dados.



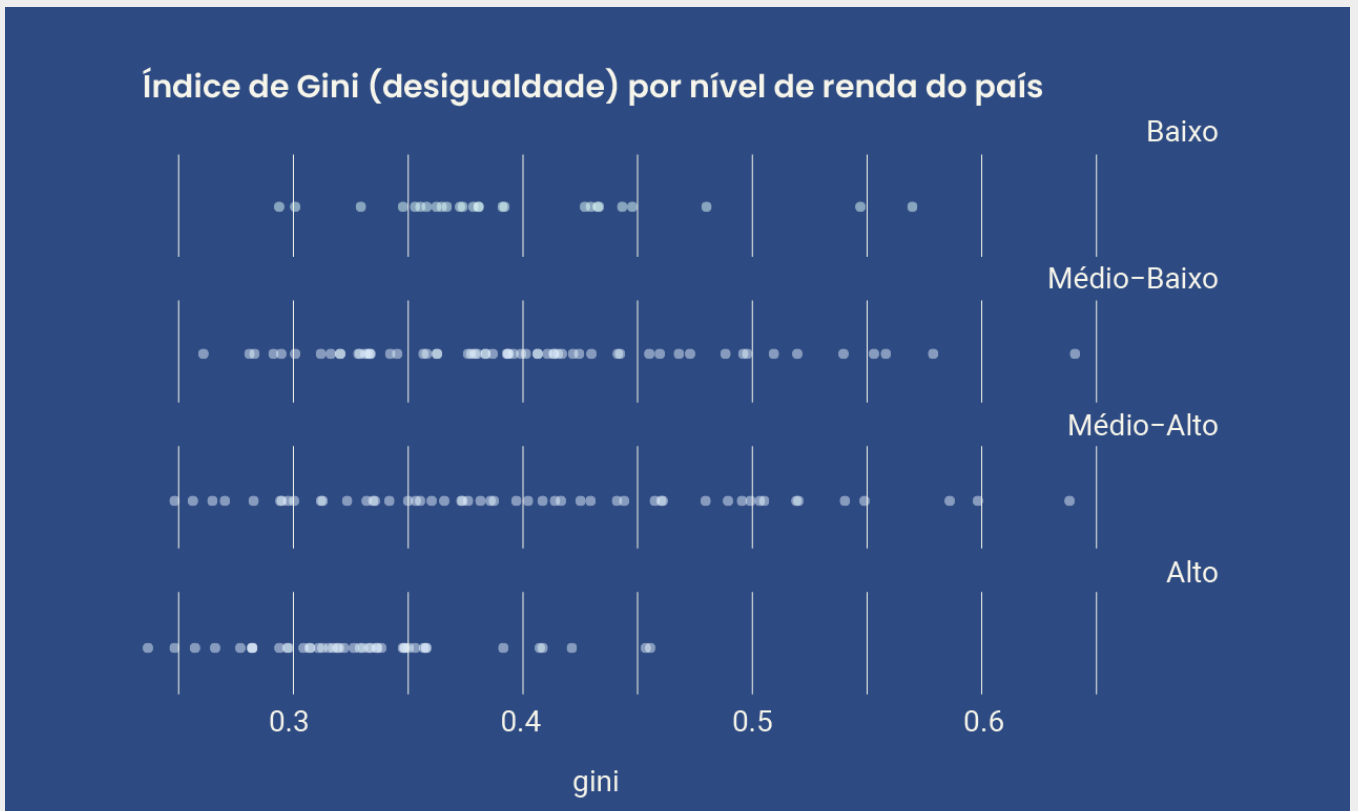
Histogramas em pequenos múltiplos, para facilitar a comparação entre grupos.

Fonte: Banco Mundial (2021)



Curvas de densidade em pequenos múltiplos, para facilitar a comparação entre grupos.

Fonte: Banco Mundial (2021)



Strip plots em pequenos múltiplos, para facilitar a comparação entre grupos.

Fonte: Banco Mundial (2021)

Tendências e relacionamentos

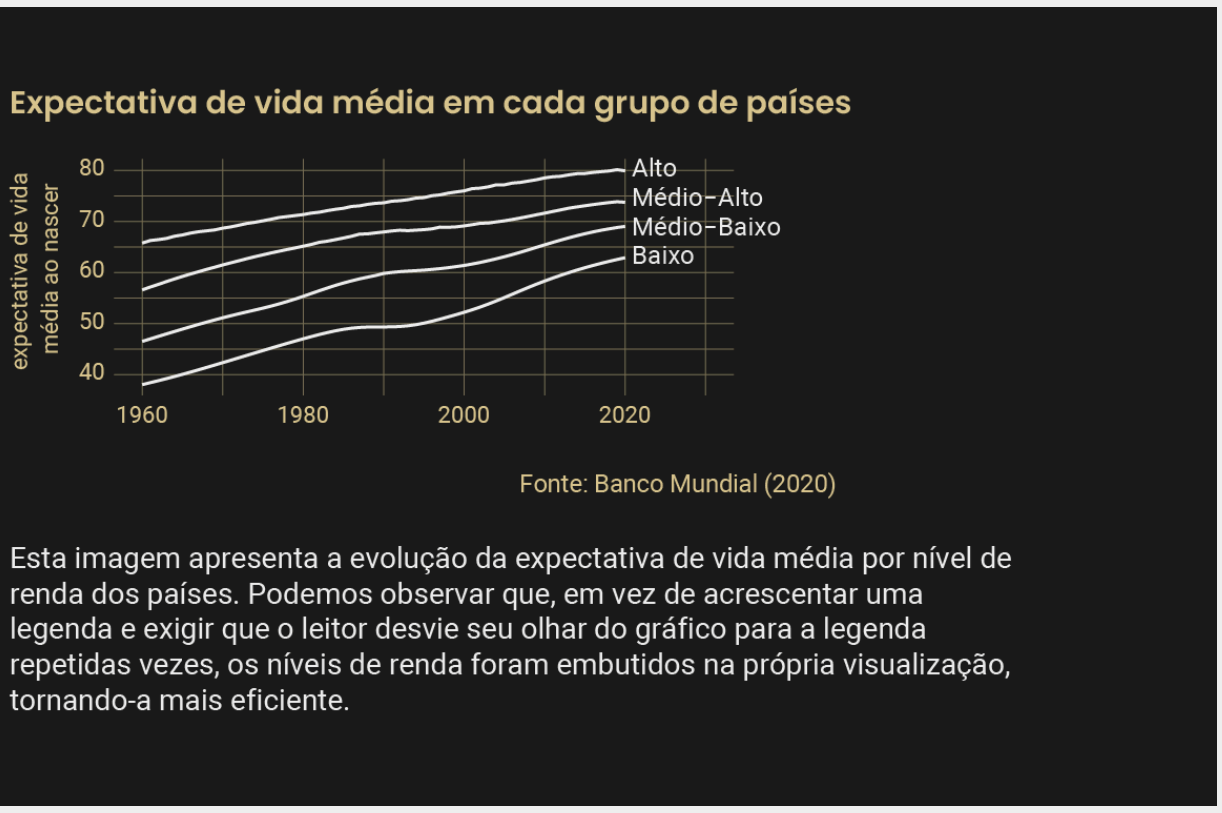
Vamos agora conhecer e pontuar algumas perguntas a serem feitas a determinados tipos de dados a visualizações usuais considerando os objetivos de **identificação de tendências e relacionamentos**.

Pergunta	Tipos de gr
Como variou o número de [objetos] ao longo de **T0**?	Gráfico de linhas
Como variou o número de [objetos] ao longo de **T0**, por **C0**?	Gráfico de linhas múltipl
Como variou a [soma, média, mediana, ...] de Q0 ao longo de **T0**?	Gráfico de linhas
Como variou a [soma, média, mediana, ...] de Q0 ao longo de **T0**, por **C0**?	Gráfico de linhas múltipl
Qual é a (cor) relação entre **Q0** e **Q1**?	Gráfico de dispersão
Qual é a (cor) relação entre **Q0** e **Q1**, considerando **C0**?	Gráfico de dispersão cor símbolos

Interativo

Descrição do interativo

Nesses casos, pode valer a pena agregar os dados e produzir um gráfico com menos linhas e selecionar menos linhas para visualizar ou acrescentar mecanismos de interação para que o leitor do gráfico decida quais estados quer visualizar a cada momento.



O vídeo a seguir apresenta um exemplo de código em Python para a construção de um gráfico de linhas.

Gráfico de linhas

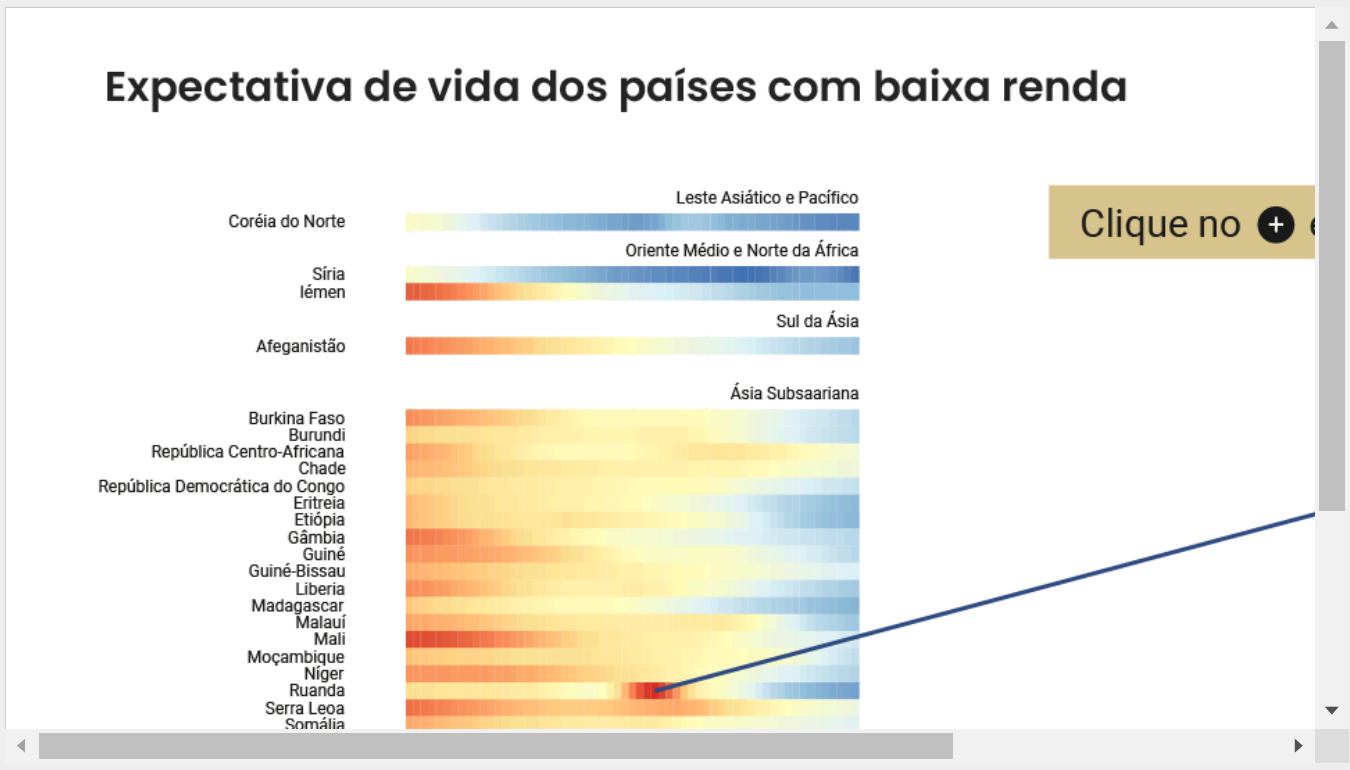
Veja, nesta aula, a construção de um gráfico de linhas utilizando a biblioteca Matplotlib.

Assista o vídeo

Heatmap (mapa de calor)

Heatmaps (mapas de calor) permitem visualizar a associação de uma variável quantitativa (contínua ou discreta) com outras duas variáveis discretas (sejam categóricas, quantitativas ou temporais).

Um *heatmap* também pode ser utilizado para analisar a evolução ao longo do tempo de muitos grupos ao mesmo tempo, como na figura a seguir. Além da evolução de cada país ao longo do tempo, a figura agrupa ainda os países com nível baixo de renda de acordo com suas regiões, provendo mais uma dimensão de comparação.

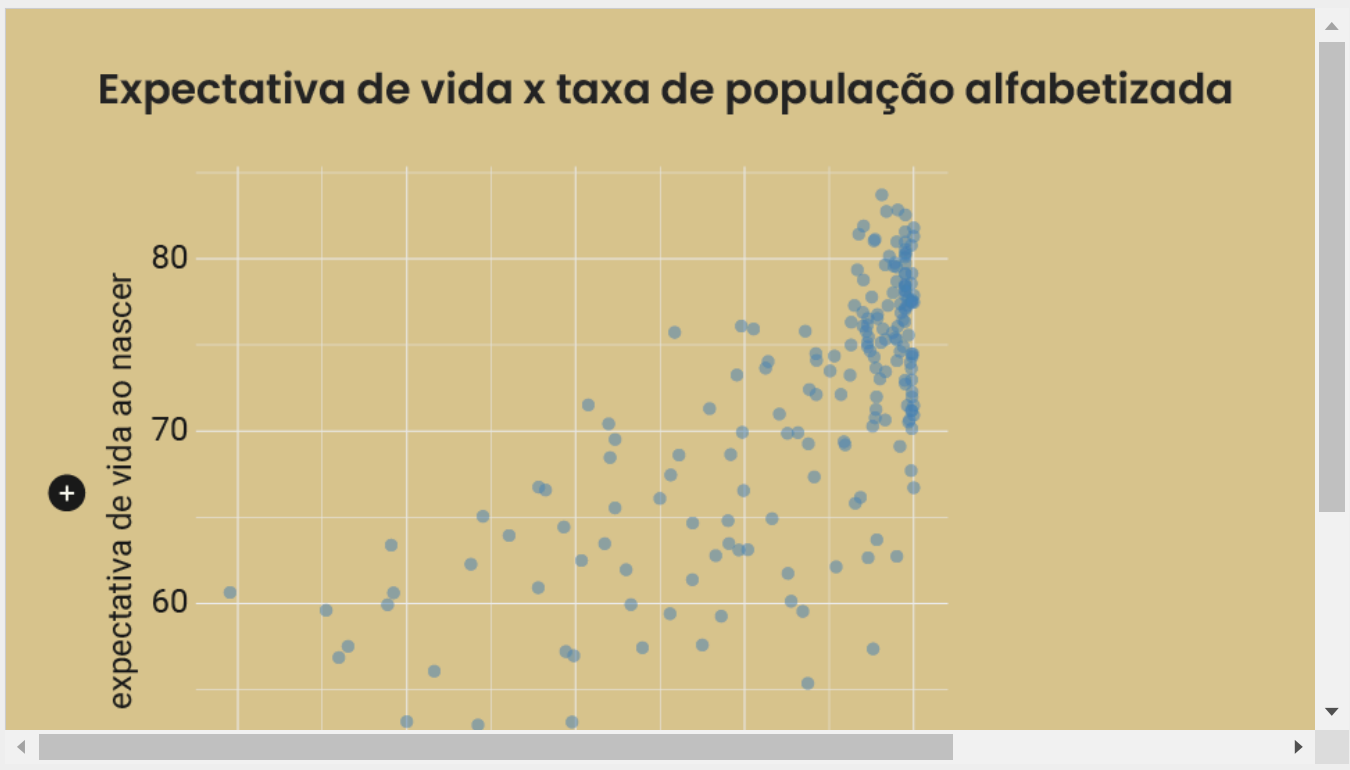


Interativo

Descrição do interativo

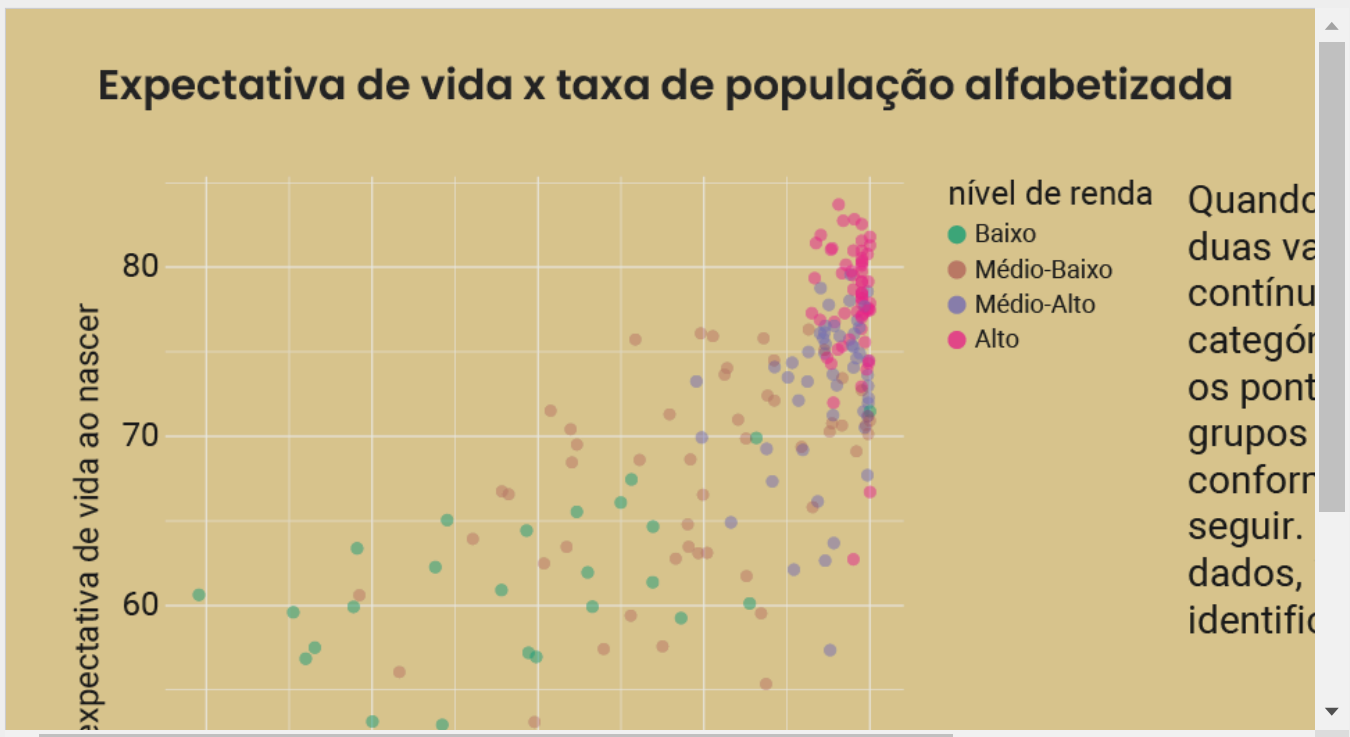
Scatterplot (gráfico de dispersão)

Para analisarmos relações entre duas variáveis quantitativas contínuas, utilizamos com frequência *scatterplots* (gráficos de dispersão).



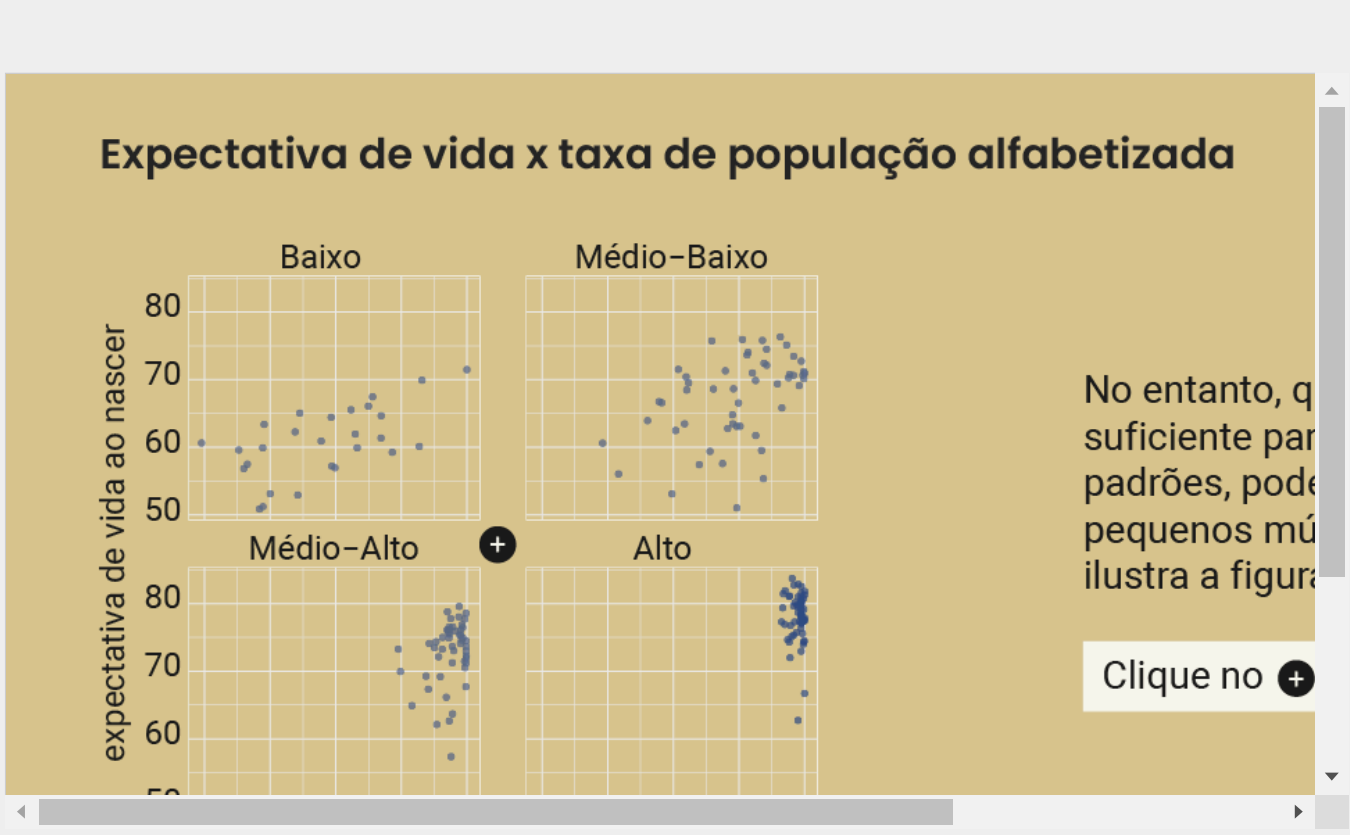
Interativo

Descrição do interativo



Interativo

Descrição do interativo



Interativo

Descrição do interativo

O vídeo a seguir apresenta um exemplo de código em Python para a construção de um gráfico de dispersão.

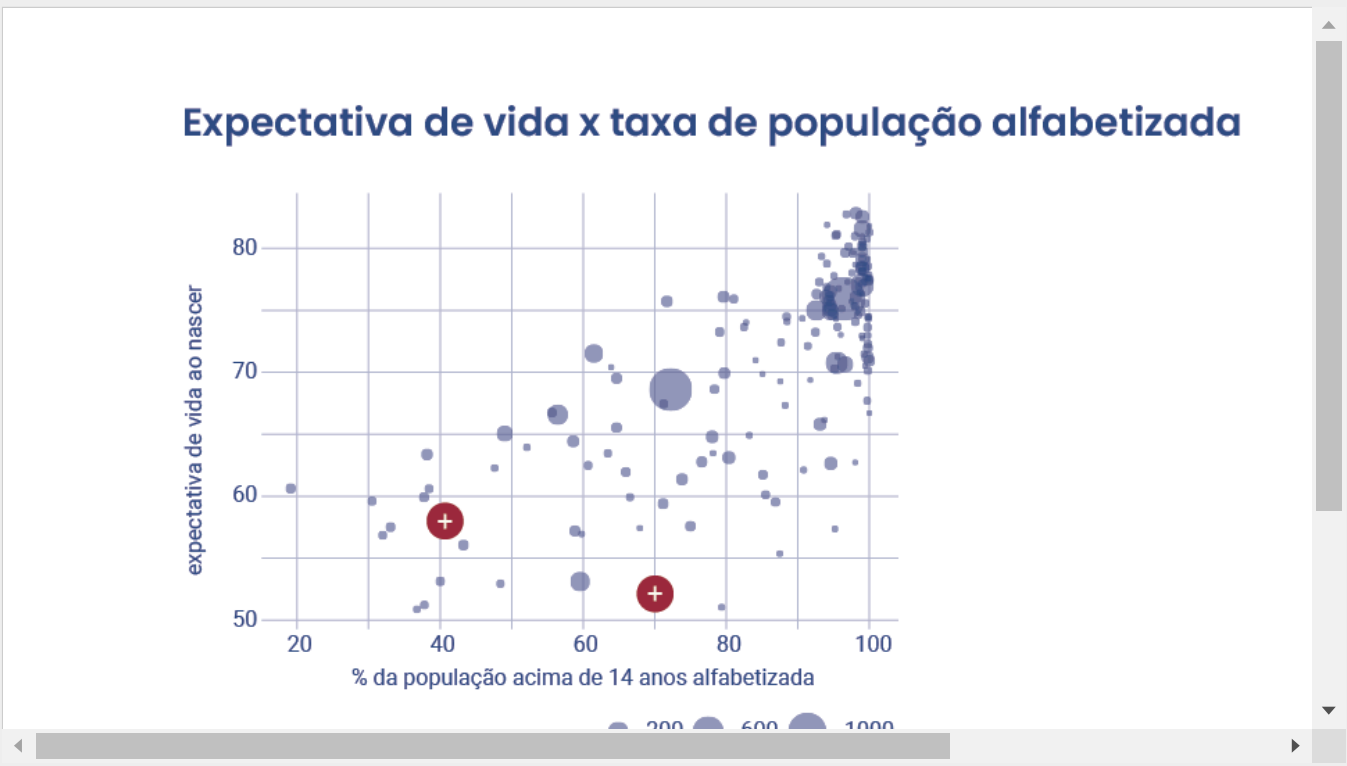
Gráfico de dispersão (*scatterplot* e *bubbleplot*)

Veja, nesta aula, a construção de um gráfico de dispersão utilizando a biblioteca Matplotlib.

Assista o vídeo

Bubbleplot (gráfico de bolhas)

Para relacionarmos três variáveis quantitativas contínuas, podemos utilizar o *bubbleplot*, ou gráfico de bolhas. Assim como em um *scatterplot*, mapeamos duas variáveis nas posições X,Y do plano cartesiano. Além disso, mapeamos a terceira variável no tamanho do símbolo, geralmente um círculo, de onde vem a denominação “bolha.” Trata-se de um gráfico de leitura mais difícil, e nem sempre conseguimos estabelecer visualmente as relações com a variável mapeada no tamanho.



Interativo

Descrição do interativo

Assim como no *scatterplot*, podemos acrescentar a um *bubbleplot* uma variável categórica mapeada em cores.

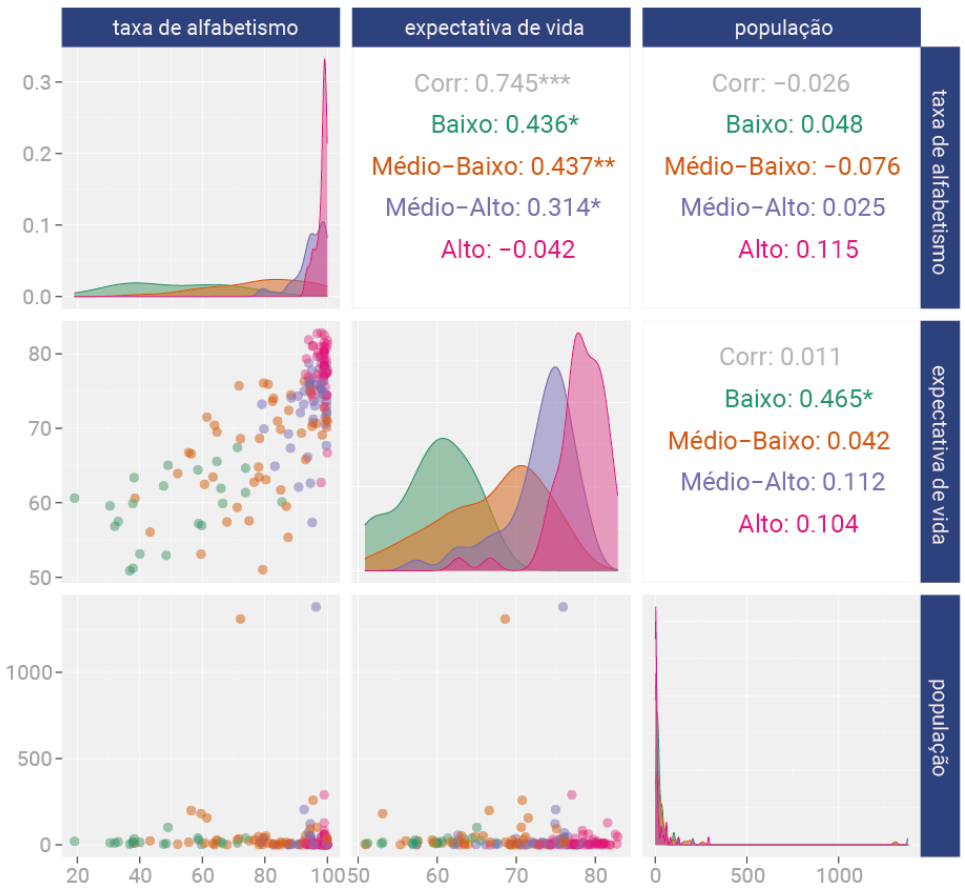
Interativo

Descrição do interativo

Como se torna mais difícil analisar a relação entre mais do que duas variáveis quantitativas contínuas, muitas vezes optamos por substituir ou complementar um gráfico de bolhas com uma matriz de *scatterplots*, conforme ilustra a figura a seguir.

A matriz de *scatterplots* da figura está estruturada em três partes. Na triangular inferior, a matriz apresenta os *scatterplots* de cada par de variáveis, conforme os nomes no topo e à direita da matriz. Na diagonal, a matriz apresenta os gráficos de densidade da variável correspondente. E, na triangular superior, a matriz apresenta o coeficiente de correlação linear entre as duas variáveis, tanto de forma agregada ("Corr") como por grupo conforme variável categórica mapeada na cor.

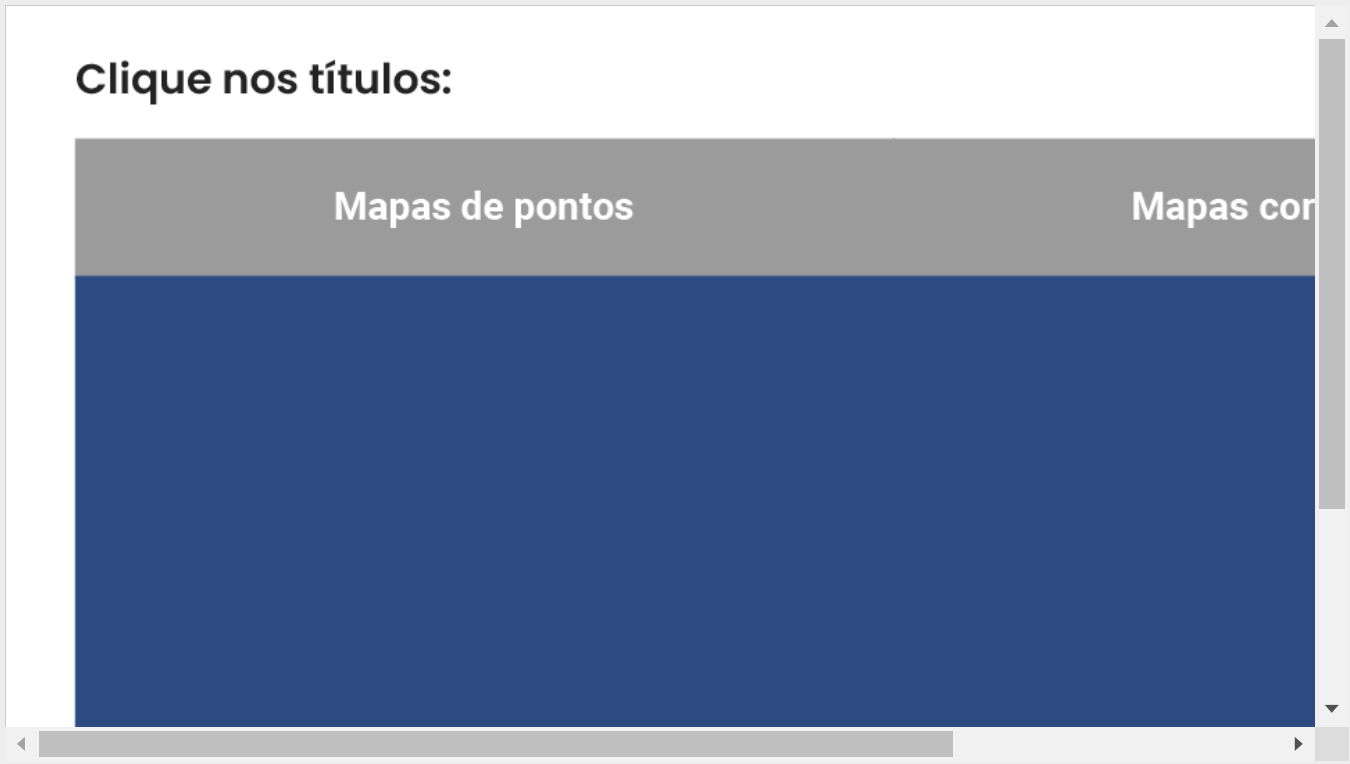
Relações entre taxa de alfabetismo, expectativa de vida e população



Fonte: Banco Mundial (2020)

Dados espaciais

Você sabia que com esses dados espaciais podemos localizar elementos pontuais em um mapa, traçar caminhos e caracterizar polígonos? A seguir, veremos os mapas de pontos e coropléticos.



Interativo

Descrição do interativo

Analise

Muitas vezes, um cliente nos fornece alguns dados e nos pede para criarmos uma ou mais visualizações para eles, sem nos dizer quais são os objetivos da visualização ou quais são as decisões que precisarão ser tomadas com base nessa visualização. Quais são os riscos de construirmos visualizações a partir dos dados apenas? Como mitigar esses riscos?

Referências

Clique aqui para acessar as referências e créditos desta aula.

Ir para Técnica aplicada