*Ferramentas de visualização*

Visualização de dados

## Introdução

Os dados são uma **faca de dois gumes** . Quando se trata de aproveitá-los, eles podem ser usados para o bem, mas também para enganar e desinformar. É importante lembrar neste momento o Princípio de Peter Parker: “ **Com grandes poderes vêm grandes responsabilidades** ”. Além do que dizem os super-heróis, é verdade que temos que ser **fortemente responsáveis no uso dos dados** . Vejamos este conceito.

Antes de tudo, como pensar em responsabilidade se tudo o que estou fazendo é analisar dados? A resposta está na **saída do processo** . Os dados são fundamentalmente **entradas para a tomada de decisões** . Nesse sentido, eles podem se **comunicar diretamente** ou ser usados como **entrada para algoritmos** e desenvolvimentos de ciência de dados.

Para o primeiro caso, e em todos os casos em que devemos comunicar nossos resultados, devemos utilizar as práticas e ferramentas **de visualização de dados.** A visualização é muito mais do que apenas selecionar um tipo de gráfico aleatoriamente e colocá-lo em cima de uma planilha. É, sim, uma área do conhecimento orientada para a **preparação detalhada para a apresentação eficaz das informações recolhidas** . Um gráfico deve sempre **dizer muito com economia de elementos** , trazendo para o mundo real a conhecida máxima de que “uma imagem vale mais que mil palavras”. Certamente, um gráfico deve valer pelo menos a quantidade de dados que pretende exibir. E, claro, você deve sempre ser preciso e fiel a esses dados.

Para aprofundar sua compreensão da visualização de dados, recomendamos os excelentes livros de um dos mais importantes estudiosos do assunto, Edward Tufte, em especial "A exibição visual de informações quantitativas", todos disponíveis para venda em livrarias e online. seu site pessoal <https://www.edwardtufte.com/tufte/>

Para o que segue nesta aula, vamos ter em mente o que vimos na aula anterior, pois existe uma **conexão muito próxima entre a definição de uma variável e sua exibição** .

## Elementos de um gráfico

Um bom gráfico deve ser **claro e conciso** para comunicar efetivamente os dados que o compõem. Você deve sempre ter pelo menos os seguintes elementos:

* Um **título conciso e representativo** , que indique claramente quais informações devem ser mostradas com o gráfico
* Uma indicação concreta e precisa das **variáveis envolvidas** , especialmente qual o nome de cada variável e em quais unidades de medida ela é expressa.
* Se houvesse diferentes **categorias ou variáveis** juntas, uma enunciação e representação clara e bem diferenciada de cada uma

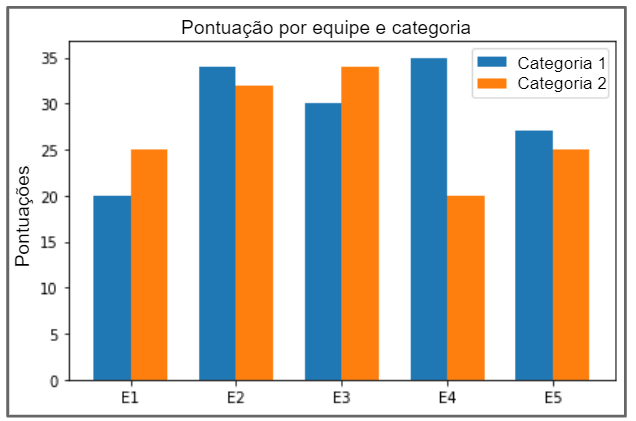
Com esses elementos, temos o terreno preparado para colocar os dados em si. Se os dados forem colocados em um gráfico adequado, servirão para comunicar efetivamente as informações coletadas.

Análise de dados básicos

## gráficos de barra

A operação mais básica que pode ser realizada com um conjunto de dados é **contá-los** , responder à pergunta “quantos são?”, considerando-os como um todo, ou com a possibilidade de desmembrá-los em categorias, se houver (“ quantos existem de cada tipo?"). Outra operação muito básica é simplesmente **exibir valores específicos** associados a uma categoria, por exemplo, o valor total de vendas por linha de produto.

Para esses fins, usamos **gráficos de barras** . Os gráficos de barras são utilizados com **dados categóricos** e consistem em um eixo horizontal onde estão localizadas as categorias e uma barra para cada uma delas cuja altura no eixo vertical **representa a quantidade ou valor** dos elementos da categoria em questão.



Na figura acima, temos as pontuações de 5 equipes, de E1 a E5, separadas por categorias, nomeadas como Categoria 1 na cor azul e Categoria 2 na cor laranja.

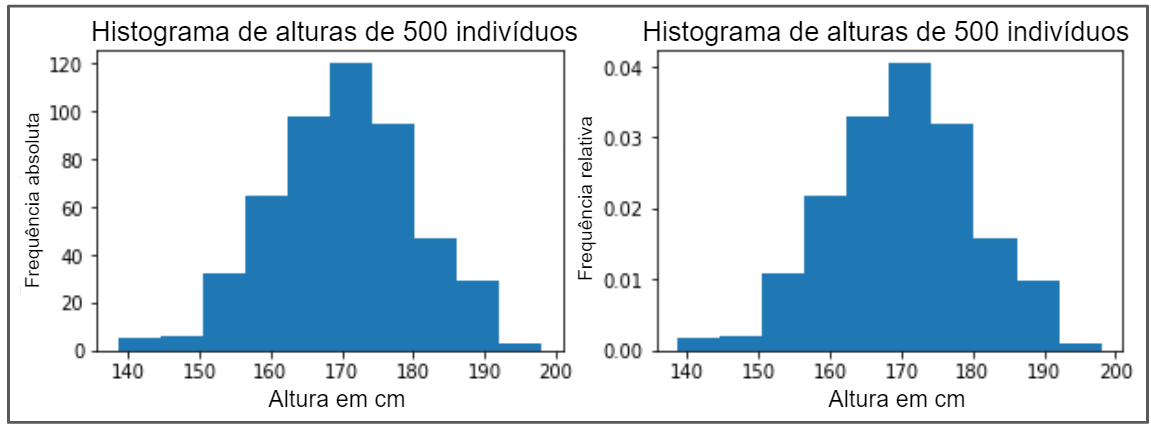
Recomendação: **Não use gráficos de barras para representar dados ao longo do tempo** , pois relacionamentos e tendências são mais difíceis de visualizar. Veremos como representar dados ao longo do tempo posteriormente na seção Séries Temporais.

## Histogramas

Veremos aqui o básico de uso de histogramas, que apresentamos brevemente como um método da biblioteca matplotlib na Aula 5. Os histogramas **mostram as frequências de aparecimento para cada intervalo de valores da variável** , por isso também são chamados **histogramas de frequência** . Os histogramas são muito semelhantes em forma aos gráficos de barras, mas muito diferentes no comportamento. Um histograma **mostra uma distribuição de dados** , como mencionamos na aula anterior.

Considere os dados para as alturas de 500 pessoas, medidas em centímetros. Para construir um histograma, vamos organizar as medidas de altura disponíveis em intervalos, também chamados de bins. Para trabalhar com intervalos, proceda da seguinte forma: pegue todos os valores da variável, do mínimo ao máximo, e divida o intervalo correspondente em intervalos de igual tamanho. Existem diferentes formas de escolher quantos intervalos usar, mas geralmente as linguagens de programação já fornecem um valor calculado por padrão que serve adequadamente para nossa análise. Em Python, podemos revisar a classe 5 para o método hist do matplotlib, onde o número de bins é definido com o parâmetro bins. Dessa forma, um histograma mostra **a importância de cada intervalo de valores em relação ao total de dados** .

Uma vez definidos os intervalos, estamos prontos para apresentar um histograma. Vamos ao nosso exemplo: a figura a seguir mostra os histogramas para as alturas de 500 indivíduos.



No eixo horizontal, é apresentada **a variável a ser analisada** . Cada intervalo de valores da variável possui uma altura que representa o número de observações nesse intervalo. Essa quantidade pode ser o resultado de uma **simples contagem: "Quantas unidades existem para cada intervalo da variável?"** , ou podem ser expressos como uma **proporção em relação ao total: “Quanto do total cada intervalo da variável representa?”** . Neste caso, **a quantidade de cada intervalo** **é dividido pelo número total de observações** , com as quais as proporções mencionadas são obtidas. No primeiro caso falaremos sobre **histogramas de frequências absolutas** , mostrados no gráfico à esquerda. No segundo caso, falaremos sobre **histogramas de frequência relativa** , mostrados no gráfico à direita. A forma é a mesma, o que muda é a interpretação dos resultados: no primeiro caso falamos do **número de observações para cada intervalo da variável** , no segundo caso falamos da **proporção de tempos de cada intervalo da variável em relação ao total de observações da variável** .

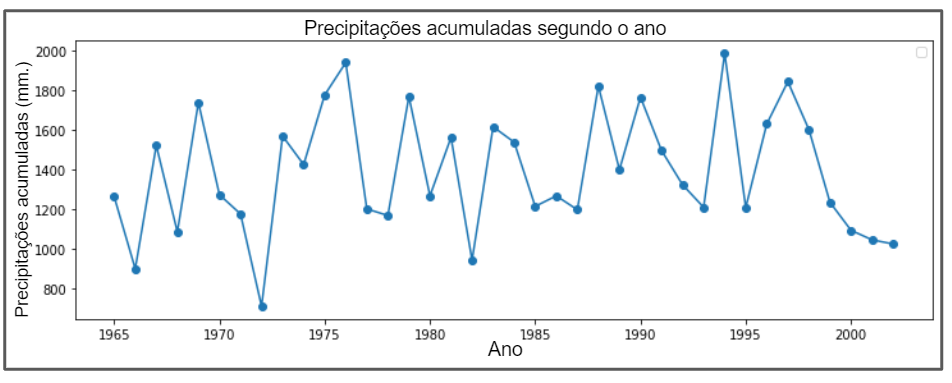
De acordo com o que foi visto na aula anterior, percebe-se que essa distribuição está muito próxima da normal, com os valores se acumulando em maior quantidade em torno de uma média, e em menor quantidade e mais dispersos à medida que nos afastamos do média (cerca de 170 centímetros). O histograma **dá uma ideia da forma da distribuição** , o que o ajudará a tirar conclusões úteis sobre seus dados. Recomendamos o uso do histograma como **uma das primeiras ferramentas de análise de variáveis quantitativas** .

Relações entre elementos

Muitas vezes, uma variável possui relações intrínsecas entre seus valores, ou queremos verificar se uma variável está relacionada a outra ou outras. Nesta seção, veremos duas ferramentas gráficas básicas para verificar relacionamentos: as séries temporais e os gráficos de dispersão.

## Série temporal

Sempre que tivermos uma variável que tenha **algum tipo de evolução ao longo do tempo** , ou **onde organizar seus valores ao longo do tempo faça sentido para análise** , podemos traçar uma **série temporal** . Para este tipo de gráfico, desenharemos um ponto cuja altura representará o valor da variável e cuja posição em relação ao eixo horizontal representará o momento em que essa variável é medida. Então vamos conectar os pontos sequencialmente. Considere o exemplo a seguir, retirado da Aula 5.

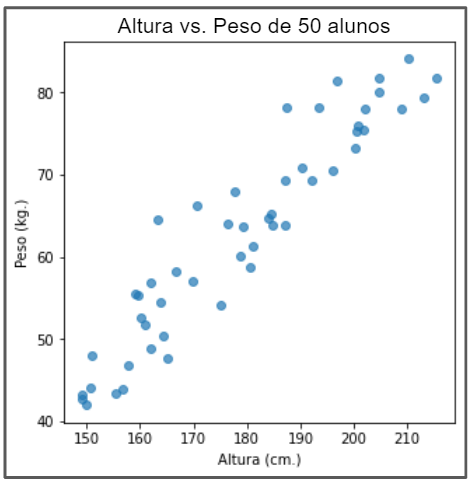


Lembre-se que este gráfico mostra a quantidade de chuva acumulada, medida em milímetros, por ano civil. A altura de cada ponto representa a quantidade de precipitação e sua localização no eixo horizontal representa seu ano de medição. As linhas que conectam os pontos nos ajudam a **perceber rapidamente se há alguma tendência, repetição de valores ou qualquer outra característica** para começarmos a conhecer o comportamento dos dados ao longo do tempo.

Recomendamos fortemente **o uso deste tipo de gráfico para mostrar a evolução de uma série de dados, faça sempre com pontos e não com barras, e que o tempo esteja sempre no eixo horizontal, movendo-se da esquerda para a direita** . Isso parece muito óbvio, mas nem sempre é levado em consideração e é essencial para boas práticas em dados afetados pelo tempo.

## Diagramas de dispersão. Noções de correlação

Quando queremos analisar a **relação entre duas variáveis** , queremos ver o que acontece quando as olhamos juntas. Por exemplo, considere o caso da Classe 5 sobre pesos em quilogramas e alturas em centímetros para um conjunto de 50 alunos.



Cada ponto representa um aluno. Para cada aluno podemos ver a intersecção do eixo horizontal, que indica as alturas, com o eixo vertical, que indica os pesos.

Este é um gráfico de **dispersão** , onde os pontos são colocados em duas variáveis e **as relações entre elas podem ser analisadas** . Para este caso, podemos ver que maiores alturas correspondem a maiores pesos e vice-versa.

Deve-se ter muito **cuidado ao expressar essas conclusões na forma de relações simples e não em ligações do tipo causa-efeito** . Dizer que maiores alturas correspondem a maiores pesos (o que equivale a dizer que maiores pesos correspondem a maiores alturas) não é o mesmo que dizer que se um indivíduo é alto, então pesará mais. Para poder afirmar esta última frase, é preciso verificar que uma variável depende da outra (ou seja, afirmar que um indivíduo com grande estatura necessariamente terá um peso maior). Esta verificação não pode ser feita com o gráfico de dispersão.

Embora a diferença seja sutil, não deixa de ser importante, e **neste caso falaremos da relação entre variáveis** . Apresentaremos a noção de relação como resultado da ocorrência conjunta dos valores das variáveis em estudo. É importante ter em mente que **quando falamos de relacionamento não estamos necessariamente falando da dependência de uma variável em relação à outra** . A dependência é o objeto de estudo da análise de Regressão, que veremos mais adiante.

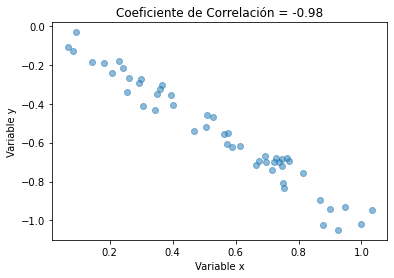
### Correlação

Se a relação **tiver alguma forma definida ,** será considerada como uma **correlação entre as variáveis** . Falaremos comumente de **correlação linear** , que é a mais prática para fins de análise em Data Science. A correlação linear pode ser medida através de um indicador chamado **coeficiente de correlação** . O coeficiente de correlação pode ter valores entre -1 e 1, e nós o interpretamos da seguinte forma:

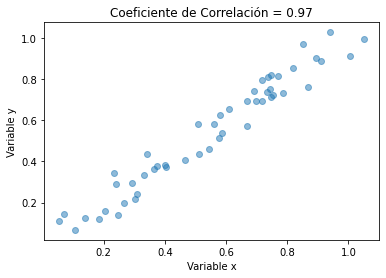
* Se o valor da correlação estiver **próximo de 1** , significa que quando os valores de uma variável são altos, os da outra variável também são altos, e da mesma forma com valores baixos.
* Se o valor da correlação estiver **próximo de -1,** significa que quando os valores de uma variável são baixos, os valores da outra variável são altos, e da mesma forma com valores altos.
* Se o valor de correlação estiver **próximo de 0** , significa que **não há correlação linear forte** entre as variáveis.

Exemplos dos três valores de correlação são mostrados nas figuras abaixo. O coeficiente de correlação de 0,97 é próximo de 1 e **corresponde ao primeiro caso,** **Falamos de correlação positiva** . O coeficiente de correlação de -0,98 é próximo de -1 e **corresponde ao segundo caso, falamos de correlação negativa** . O coeficiente de correlação de 0,12 é próximo de 0 e **corresponde ao terceiro caso, e falamos de correlação nula ou sem correlação** .

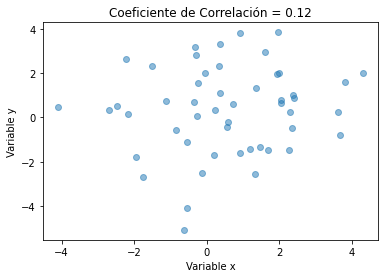
Coeficiente de correlação = -0,98



Coeficiente de correlação = 0,97



Coeficiente de correlação = 0,12



Análise de distribuições de dados

## Entenda a forma dos dados

Como mencionamos na seção sobre histogramas e nas aulas anteriores, é importante entender a forma dos dados. Para isso, utilizaremos principalmente duas ferramentas gráficas: por um lado, o histograma de frequências e, por outro, o box-and-whisker plot.

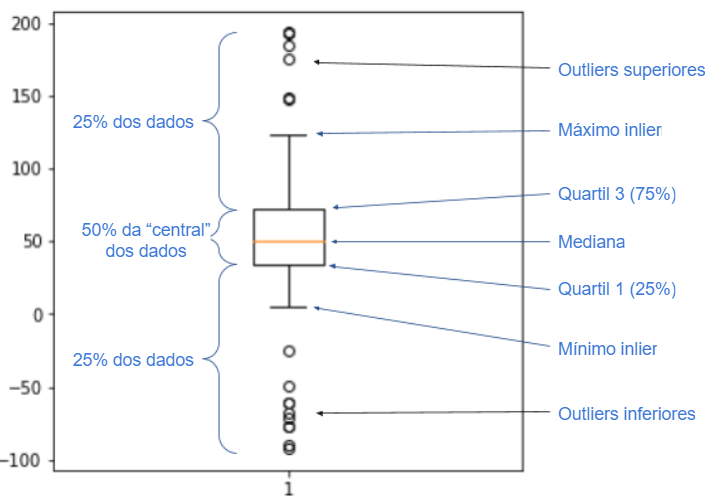
O que estamos procurando quando analisamos a forma dos dados? A resposta é longa, mas podemos resumi-la nos seguintes pontos:

* A distribuição é **simétrica ou assimétrica** ?
* Em qual parte da distribuição está localizada **a maior concentração de dados** ?
* Os dados estão bastante **dispersos ou concentrados em torno de algum valor** ?
* Será que existem dados **muito longe do resto** ?

Como vimos anteriormente, o histograma dá uma boa ideia sobre a **simetria** da distribuição. Também dá uma ideia sobre o acúmulo ou **concentração** dos dados, ou seu grau de **dispersão** . Mas às vezes **pode ser confuso, ou a maneira como você vê no gráfico não é muito conclusiva** . Portanto, precisamos de outra ferramenta para complementar nossa análise.

## Gráficos de caixa e bigode.

Devido ao exposto, para complementar esta análise e incorporar a mediana e os quartis como indicadores para medir a concentração dos dados, o **box-and-whisker plot é muito útil** . Este diagrama consiste precisamente numa caixa com dois “bastões” ou “bigodes”, num desenho muito simples mas ao mesmo tempo muito eficaz. Vejamos um exemplo onde desenvolveremos seus elementos.



A figura mostra o gráfico de caixa e bigode, neste caso verticalmente. As linguagens de programação também oferecem a possibilidade de grafá-lo horizontalmente. Uma escala para a variável analisada aparece no eixo vertical. Abaixo detalhamos os elementos do gráfico:

* A linha laranja dentro da caixa representa **a mediana da distribuição** . Isso significa que 50% das observações estão abaixo do valor da variável marcada por essa linha, e os 50% restantes estão acima desse valor.
* **A borda inferior da caixa marca o valor do quartil 1 (25%) e a borda superior da caixa marca o valor do quartil 3 (75%)** . Isso significa que a caixa contém 75-25 = 50% dos dados. Outra maneira de dizer isso é que **a concentração mais “central” de 50% dos dados é delimitada pela caixa** .
* Isso também significa que **25% dos valores estão abaixo do limite inferior da caixa e 25% dos valores estão acima do limite superior da caixa.**

Desta forma temos uma ideia detalhada sobre a concentração e distribuição dos dados. Vamos agora falar sobre **outliers** , que serão abordados com mais detalhes na próxima aula. Outliers são valores **“muito” distantes da massa central de dados** . Se os valores não forem outliers, ou seja, estiverem **“mais próximos” do centro dos dados** , vamos chamá-los **de inliers** . Existem muitos critérios para delimitar o espaço onde residem os inliers, e fora do qual teremos os outliers. O box plot usa um em particular que veremos em detalhes na próxima aula.

Suponha que tenhamos essa barreira que separa os inliers dos outliers. **A barreira em questão não está representada no gráfico** , isso é importante porque pode gerar confusão. Então vamos ver o resto dos elementos:

* Os bigodes representam **os valores dos últimos inliers** , ou seja, os últimos valores que estão distantes do centro dos dados, mas ainda não são outliers.
* Por fim, **os outliers, se existirem, são representados pelos círculos localizados fora dos bigodes** .

Como apontamos no início, o gráfico de caixa e bigode é muito simples em seu design, mas **muito poderoso devido a todas as informações importantes que contém** para conhecer a distribuição dos dados.

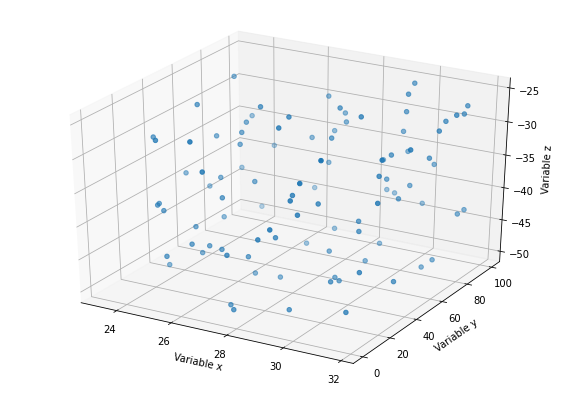
Gráficos para ver mais de duas dimensões

## mais de duas dimensões

Quando temos que comparar mais de duas variáveis estamos em um problema. Isso ocorre porque um gráfico em papel ou tela não pode exibir mais de duas variáveis ou dimensões. Para trabalhar em mais dimensões podemos usar perspectiva, ou outros recursos visuais que nos permitem analisar os dados. Essas novas ferramentas, embora não forneçam informações completas sobre os dados, permitem uma nova forma de análise que muitas vezes ajuda a entendê-los melhor. Veremos alguns deles a seguir.

## Gráficos tridimensionais.

A maneira mais básica de representar dados, quando temos **três variáveis** , é fazer um **gráfico de dispersão tridimensional** . Desta forma, fazemos uso da perspectiva e muitas vezes podemos ver claramente formas e comportamentos nos dados. Vejamos um exemplo na figura abaixo.



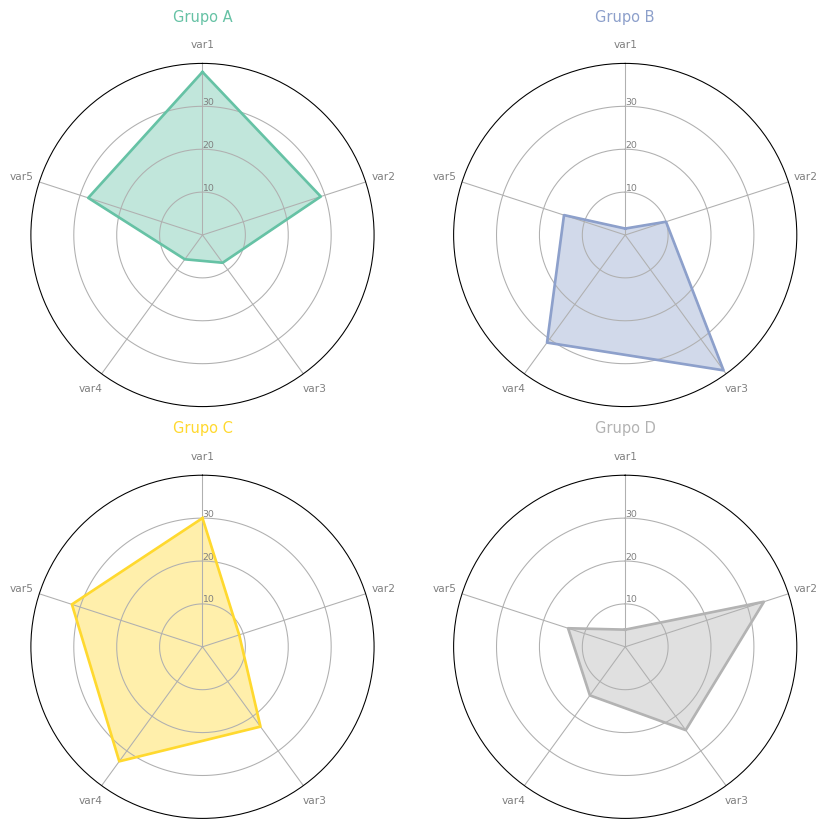
Neste caso as variáveis x, y e z estão localizadas nas arestas de um cubo, e os pontos estão localizados no espaço interior do cubo. Cada ponto representa **a interseção dos valores das variáveis x, y e z** .

## Outras maneiras de exibir dados em mais dimensões

E se tivermos dados em mais dimensões? Aqui temos que apelar para a **engenhosidade e recursos que permitem ao nosso cérebro “pensar” em mais dimensões** . Mostraremos duas ferramentas nesta seção

### gráfico de radar

O gráfico de radar surge para mostrar observações **em diversas variáveis** , e tentar compará-las visualmente. Um radar ou "teia de aranha" é desenhado, onde cada eixo representa uma variável. Em seguida, o valor correspondente a cada variável é plotado no eixo e a área delimitada pelos pontos é plotada. **Observações semelhantes terão áreas semelhantes e, da mesma forma, com observações diferentes** . Vejamos um exemplo

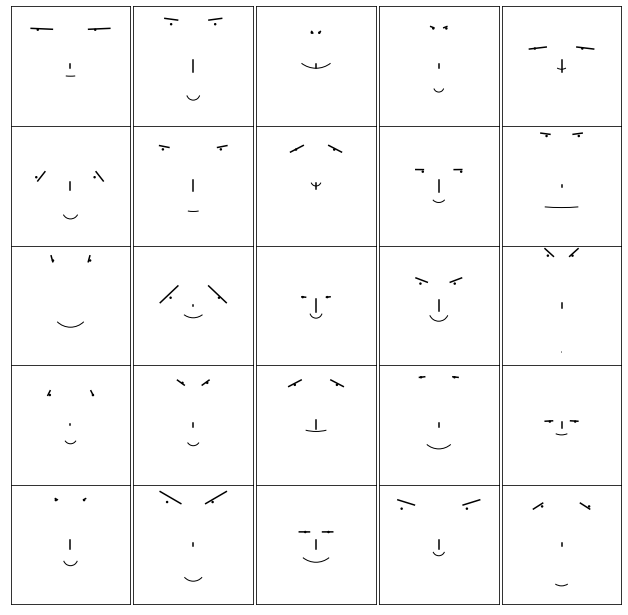


Este gráfico **permite comparar rapidamente várias qualidades em algumas unidades de observação** . Neste exemplo, poderíamos dizer que os grupos A e C possuem as maiores áreas, portanto suas variáveis geralmente possuem valores maiores. Além disso, percebe-se que o grupo A é muito diferente do grupo B, pois seus valores se destacam em diferentes variáveis (1, 2 e 5 para o grupo A; 3 e 4 para o grupo B).

### Rostos de Chernoff

Os rostos de Chernoff são baseados no mesmo princípio dos gráficos de radar, mas neste caso eles **apelam para o funcionamento do nosso cérebro** e como ele reage ao reconhecimento facial. Geralmente tendemos a agrupar rostos com expressões faciais semelhantes. Com base nesse fato, os rostos de Chernoff **associam cada variável a uma característica do rosto** : formato do rosto, formato do nariz, expressão das sobrancelhas, expressão da boca, formato das orelhas, etc. Você também pode adicionar acessórios, como chapéus ou cores, embora neste último caso você tenha que levar em consideração possíveis destinatários com problemas de daltonismo, então você deve ter cuidado ao escolher a paleta de cores.

O exemplo a seguir mostra rostos de forma simplificada (como figuras de palito) para cada observação.



Aqui está uma grade de 25 observações. Cada observação é representada por um conjunto de sobrancelhas, olhos, nariz e boca. Apesar de muito rudimentares, **as expressões podem ser claramente inferidas** : há rostos “tristes”, rostos “entediados” e rostos “raivosos”, por exemplo. **Faces com expressões semelhantes possuem elementos semelhantes** , portanto os valores das variáveis são semelhantes. Análogo aos gráficos de radar, os rostos de Chernoff são muito úteis, mas **apenas quando as observações não são muitas** . Isso é lógico, pois embora possamos reconhecer expressões faciais, essa tarefa se torna mais difícil quando há muitos rostos para analisar, como quando estamos diante de uma multidão.

Para ver um rosto Chernoff interativo, mostrando como a mudança de uma variável afeta a expressão de um rosto (desenhado como um boneco), você pode visitar <https://gramener.com/faces/>e brincar movendo os controles deslizantes.