Visualização de dados: Como escolher o melhor gráfico para um conjunto de dados?

Em qualquer projeto de ciência de dados, um dos passos mais essenciais ao explorar e interpretar resultados é visualizar os dados que temos. Visualizar dados é uma ferramenta muito importante, nomeadamente no início de um projeto, dado que nos ajuda a compreender melhor os dados, e encontrar padrões e correlações. No final do projeto, a visualização de dados é igualmente importante dado que uma visualização mais eficiente vai ajudar a comunicação dos resultados encontrados, para futuros leitores. Somos animais visuais por natureza: tudo faz mais sentido quando é representado de uma forma visual, especialmente se esta visualização for simples e intuitiva. É muito mais simples interpretar um gráfico de barras do que ler uma folha de Excel massiva, por exemplo. O método de visualização de dados é um fator importantíssimo em qualquer projeto, e pode definir quase totalmente a qualidade deste. Mesmo que façamos um projeto muito interessante, e tenhamos uma grande quantidade de resultados, com qualidade, se no final escolhermos o tipo errado de gráfico para apresentar os resultados, e não conseguirmos transmitir aos leitores as conclusões que retirámos, estes não irão interiorizar todo o trabalho e esforço que foi dedicado ao projeto, nem irão aprender os conceitos que aprendemos, com as conclusões retiradas. Há uma grande quantidade de diferentes tipos de gráficos. O processo de escolha pode ser avassalador e confuso. Este tutorial tem como objetivo dar a conhecer os diferentes principais tipos de gráficos, quando devem ou não ser aplicados, e no final possibilitar todos os leitores a escolher melhor a forma como apresentam os seus dados, e ter mais capacidade de interpretar dados de outros autores.

Como começar?

No processo de visualização de dados, antes de começar a ver diferentes tipos de gráficos, é necessário colocar 5 questões, que irão aumentar o nosso conhecimento sobre o conjunto de dados com que vamos trabalhar.

1. Qual é a história que o nosso conjunto de dados quer contar?

Dados são apenas uma história, contada com números.

A primeira coisa que devemos procurar saber sobre o conjunto de dados que temos, é: Que história estamos a tentar contar? Porque é que temos este conjunto de dados? Como é que obtivemos este conjunto de dados? Compreender a história do nosso conjunto de dados, e compreender o que estamos a tentar transmitir facilita muito a escolha do tipo de gráfico. Podemos estar à procura de correlações, comparar diferentes opiniões, demonstrar algum tipo de distribuição, etc. Há tipos de gráficos mais complexos, quando estamos a lidar com dados que contém muitas variáveis diferentes. No entanto, estes podem ser avassaladores, e transmitir demasiado de uma vez, é necessário encontrar um meio termo.

2. Quem é o público alvo?

Depois de termos a história definida, temos de perceber quem irá ver os nossos resultados. Se formos apresentar um conjunto de dados relativos à bolsa, à procura de tendências, vamos escolher um tipo de gráfico diferente se o formos apresentar a um bolsista com dezenas de anos de experiência, ou se formos apresentar a um profissional em início de carreira. Todo o propósito da visualização de dados é tornar a comunicação mais eficiente. E como qualquer tipo de comunicação, esta tem de ser adaptada ao recetor. Como tal, temos de ter em consideração quem irá ver e interpretar os nossos resultados, antes de decidir como os vamos apresentar.

3. Qual o tamanho do conjunto de dados?

A ordem de grandeza do nosso conjunto de dados vai ter um grande impacto no tipo de gráficos a escolher. Alguns tipos de gráficos não foram feitos para serem usados com um grande número de dados, enquanto que outros são perfeitos para estes casos. Por exemplo, um gráfico circular functiona para um conjunto de dados com um número de baixo de categorias. Para um número grande é impensável. É fundamental escolher um tipo de gráfico adequado à dimensão do conjunto de dados de modo a podermos apresentá-lo de forma clara, sem que fique demasiado saturado e amontoado.

4. Qual é o tipo do conjunto de dados?

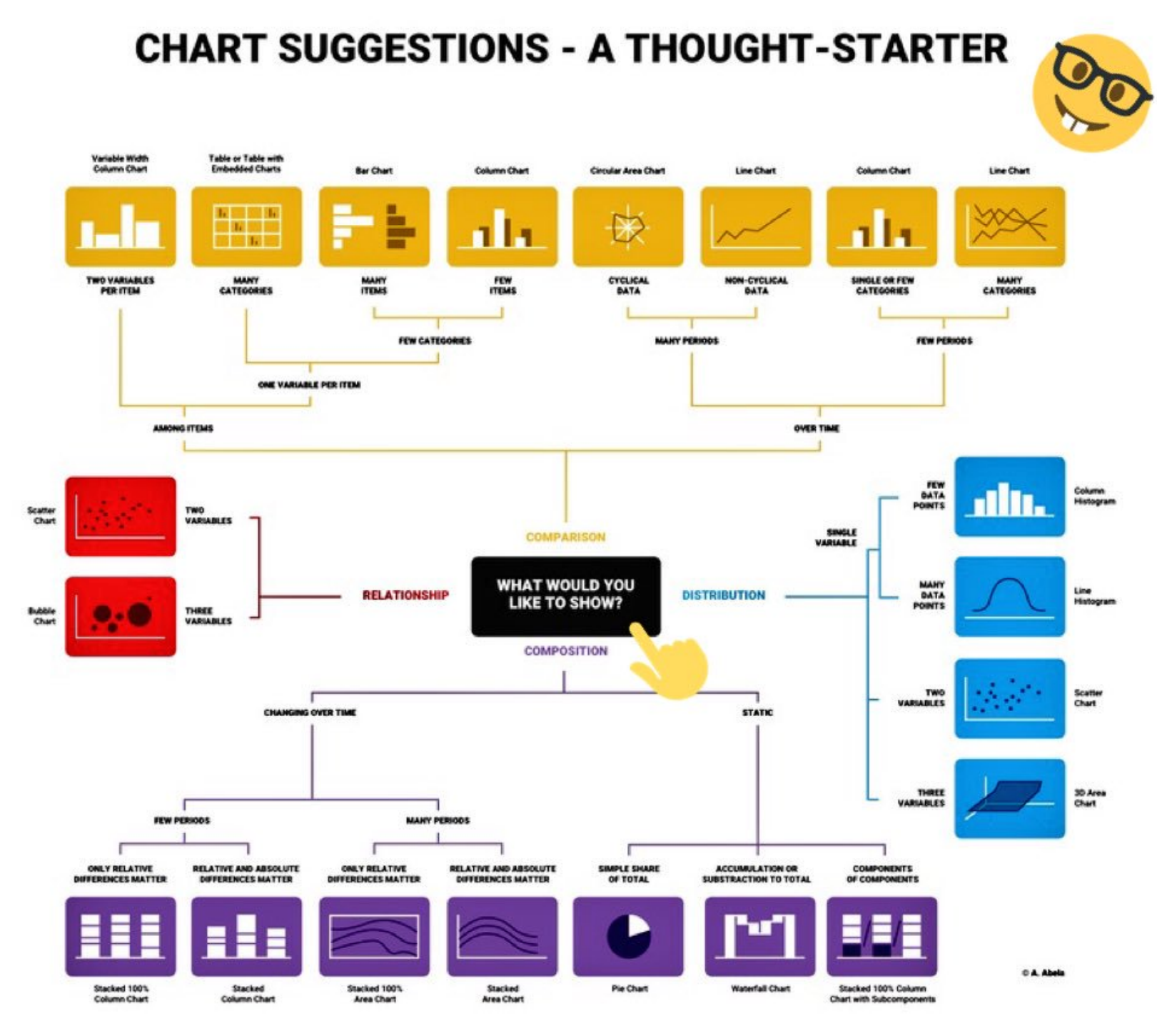
Há vários tipos de dados: contínuos, discretos, qualitativos, categóricos, etc. Ao encontrar o tipo do conjunto de dados, podemos logo eliminar alguns tipos de gráficos. Por exemplo, se tivermos um tipo de dados contínuo, não vamos escolher um gráfico de barras, um gráfico de linhas será, em princípio, uma escolha mais adequada. Do mesmo modo, se tivermos um tipo de dados categórico, será sensato escolher um gráfico de barras, ou circular. Neste caso não faz sentido escolher um gráfico de linhas, por definição não existem categorias contínuas. O número de categorias é sempre finito, e discreto, por definiçao.

5. De que modo é que os diferentes elementos do conjunto de dados se relacionam entre si?

Por fim, é necessário perceber de que modo os dados se correlacionam. Os dados podem ser ordenados com base em algum fator, como tempo, tamanho, tipo, etc. O conjunto de dados pode ser uma série temporal, ou seja, dados que se alteram com o tempo, ou algum tipo de distribuição. As correlaçoes entre os diferentes elementos num conjunto de dados facilita a escolha do tipo de gráfico.

In [1]:
```python
from IPython.display import Image
Image(filename='chart_suggestions.jpeg')
```

Out[1]:



Tipos de gráficos e de dados

Como é possível ver na imagem acima, podemos dividir os diferentes tipos de gráficos em 4 categorias:

Comparação

Gráficos de comparação são utilizados para comparar um ou mais conjuntos de dados. Estes permitem comparar diferentes items, ou mostrar diferenças ao longo do tempo

Relações

Gráficos de relações são utilizados para mostrar correlações e tendências entre duas ou mais váriaveis.

Composição

Gráficos de composição são utilizados para mostrar as alterações de um ou mais dados, ao longo do tempo.

Distribuição

Gráficos de distribuição são utilizados para mostrar como estão distribuídos os dados, ajudam a encontrar outliers e tendências como média e moda.

Durante todo o resto do documento, os tipos de conjunto de dados irão ser referidos como discretos, contínuos, qualitativos, e quantitativos. De modo a compreender totalmente a explicação de cada tipo de gráfico, além de quando os usar ou evitar, é necessário compreender estes quatro conceitos.

Dados quantitativos/qualitativos

Dados quantitativos são medições de valores contáveis, dados estes que são expressos por números. Por exemplo, número de ações trocadas num dia. Por outro lado, valores qualitativos são dados medidos em tipos, e podem ser representados por um nome, símbolo, código, etc.

Dados contínuos/discretos

Dados discretos são dados em que os únicos valores que estes podem tomar são definidos, e pertencem a um conjunto finito e enumerável. Por exemplo, número de refeições consumidas pode ter valores de 1, 2, 3, 4, etc; no entanto, um valor de 1,32 refeições não tem siginificado. Dados contínuos, por outro lado, podem ter qualquer tipo de valor associado. Neste caso, os valores que estes dados podem tomar são infinitos, e não enumeráveis. Será, por exemplo, a capacidade de um certo recipiente: 0,33; 0,20, etc.

Selecionar o tipo adequado de gráfico

Temos de decidir quantas variáveis vamos apresentar, quantos pontos, e a escala do nosso gráfico. Gráficos de linhas, colunas e barras representam alterações ao longo do tempo. Gráficos em pirâmide, e circulares representam percentagens de partes de um todo. Scatter plots e treemaps são úteis se quisermos mostrar um grande número de dados.

Os 7 tipos de gráficos mais utilizados

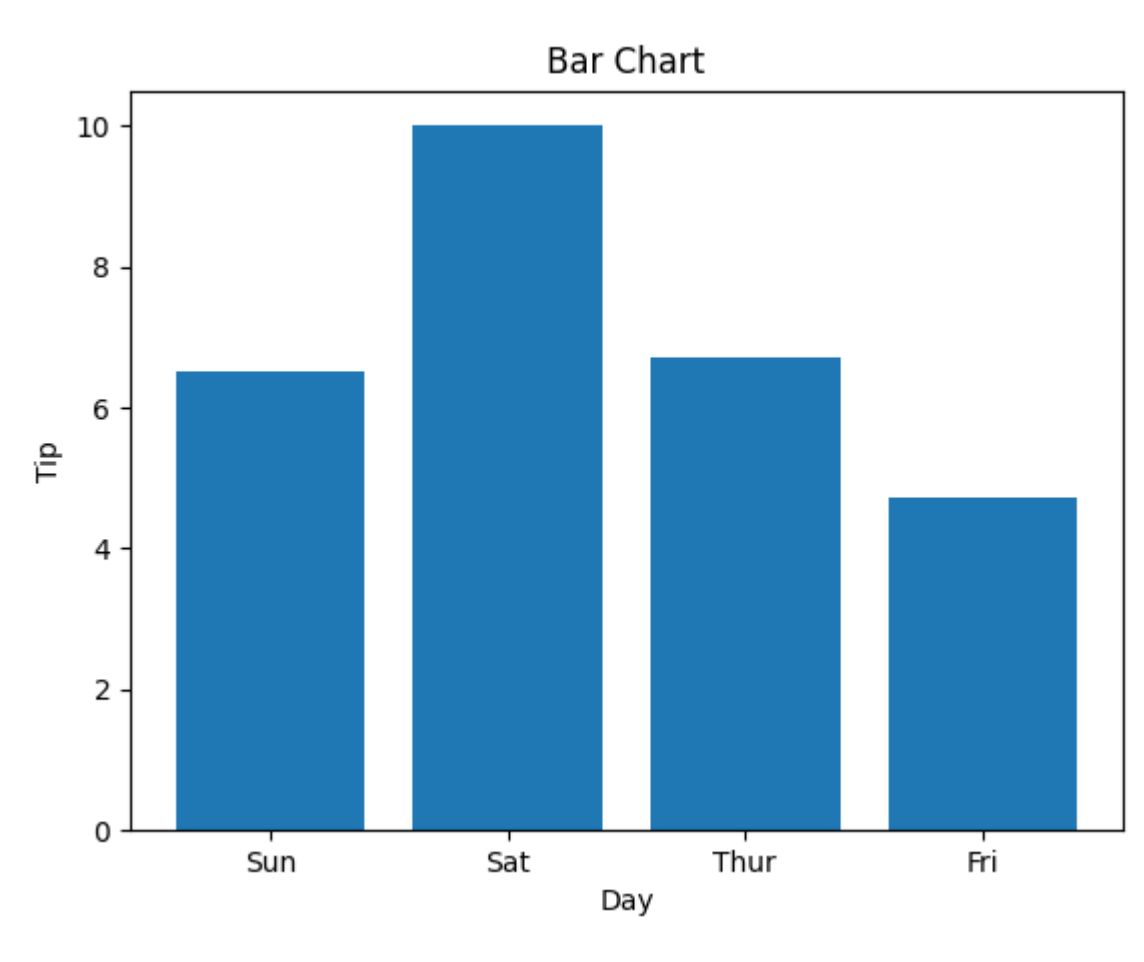
Apesar de haver mais de 40 tipos diferentes de gráficos, alguns são mais comuns do que outros, dado que são mais fáceis de construir e interpretar. De seguida vamos apresentar os 7 tipos de gráficos mais utilizados, e quando usar cada um deles.

Gráfico de barras

Gráficos de barras e colunas são utilizados para comparar diferentes items. Estes são geralmente utilizados para evitar aglomeramentos de dados, nos casos em que podemos ter, por exemplo, mais de 10 categorias a apresentar. São gráficos simples de criar, e de interpretar.

In [2]:
```python
Image(filename='barv.png')
```

Out[2]:



Quando utilizar:

- Comparar partes de um conjunto de dados maior, realçar diferentes categorias, ou mesmo mostrar alterações ao longo do tempo
- Quando há categorias de nomes extensos, este tipo de gráfico fornece mais espaço
- Quando queremos mostrar tanto valores positivos, como negativos, de um conjunto de dados.

Quando evitar:

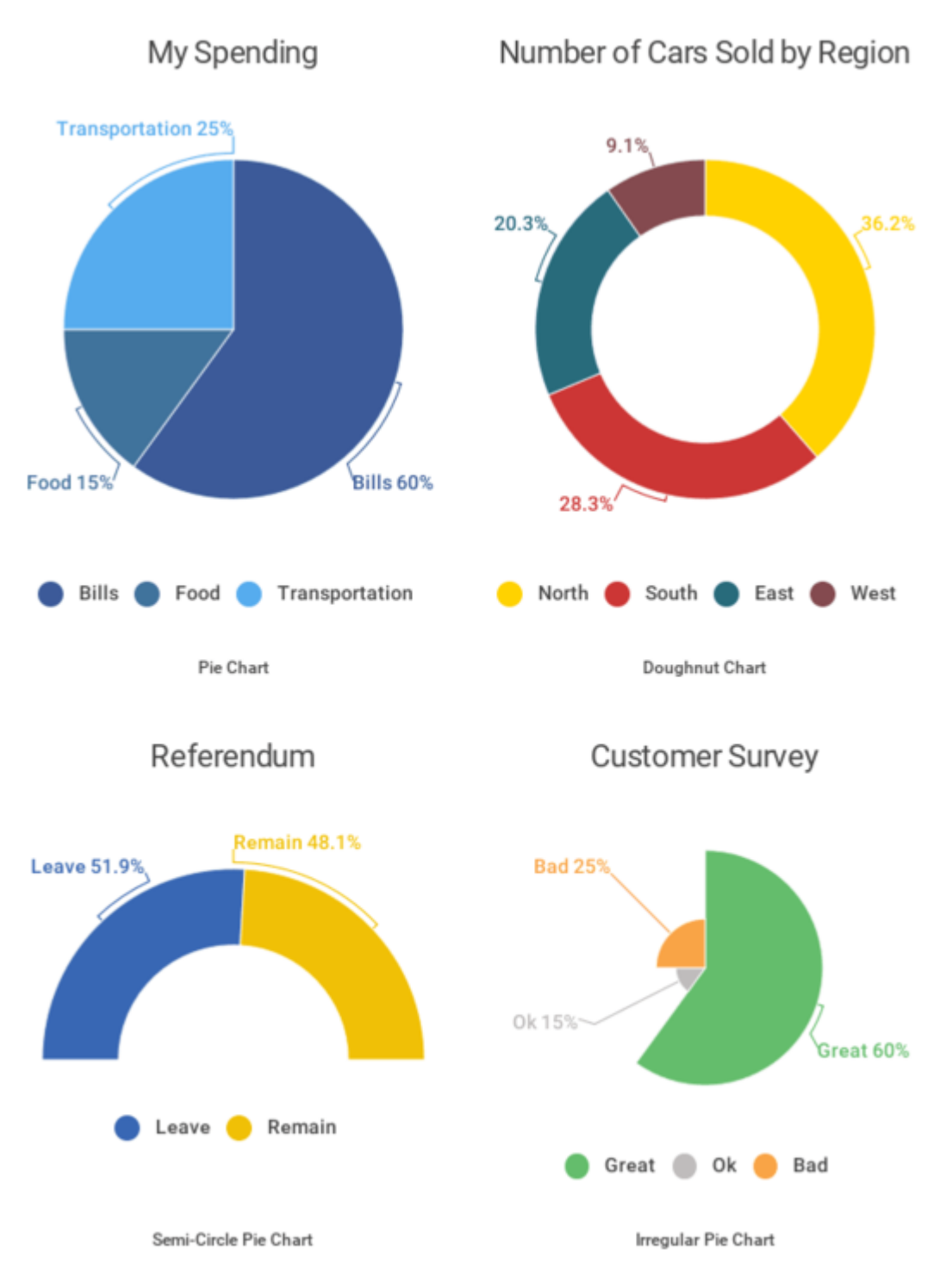
- Quando temos multiplos data points
- Quando temos muitas subcategorias. Este tipo de gráficos não deve ser demasiado denso.

Gráfico circular

Este é um dos tipos de gráficos mais controversos. Têm como objetivo representar partes de um todo, sendo que a soma destas partes deve igualar o todo. Este pode ser um gráfico muito intuitivo, se utilizado corretamente. No entanto, com um número de categorias próximo dos dois digitos, ou partes demasiado pequenas, pode tornar-se difícil de interpretar.

In [3]:
```python
Image(filename='pie.png')
```

Out[3]:



Quando utilizar

- Quando queremos mostrar proporções relativas, e precentagens da totalidade de um conjunto de dados
- Ideal para conjuntos de dados pequenos
- Quando queremos comparar o efeito de um fator em diferentes categorias
- Se temos, no máximo, 6 categorias

Quando evitar:

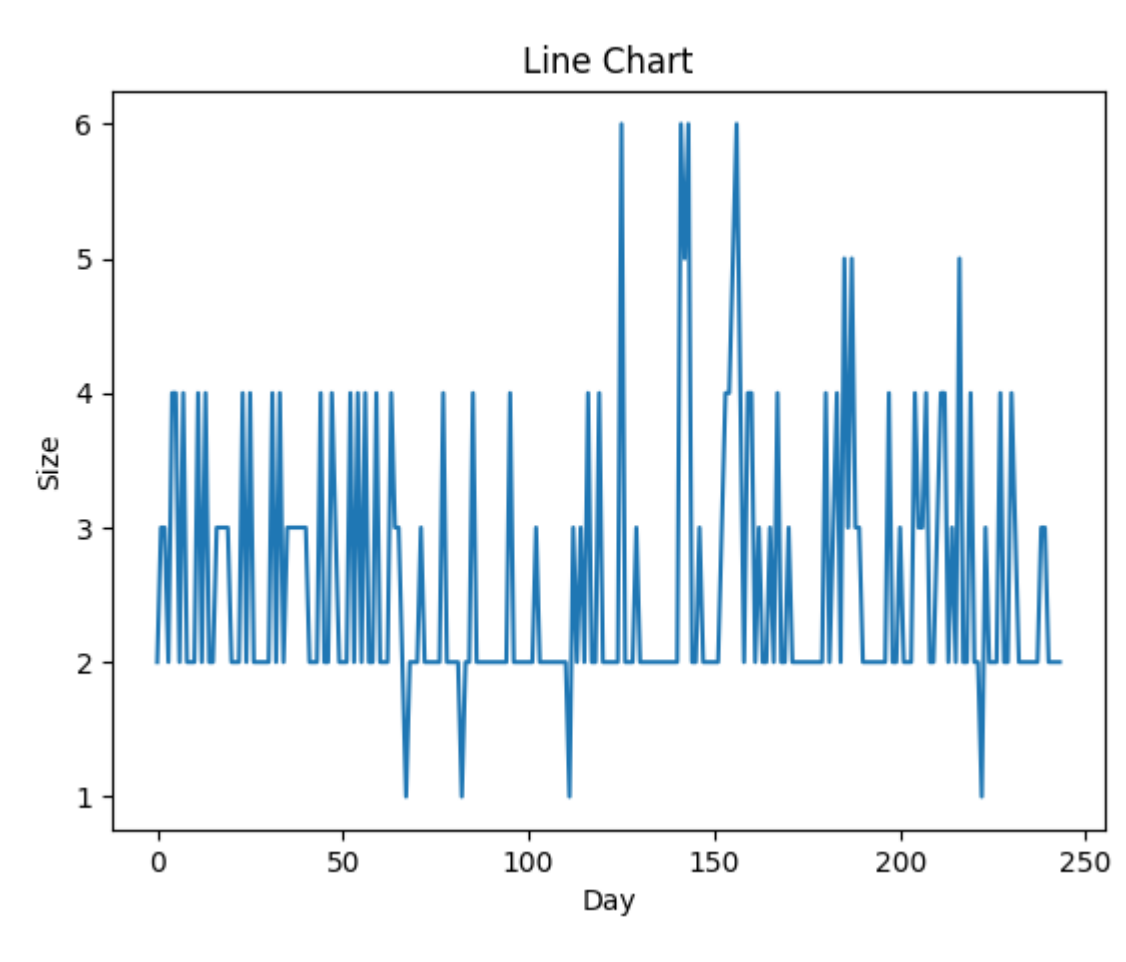
- Se o conjunto de dados é muito grande
- Se queremos fazer uma comparação absoluta, ou detalhada, de diferentes partes dos dados

Gráfico de linhas

Um gráfico de linhas revela tendências, ou alterações ao longo do tempo. Estes podem ser utilizados para encontrar relações dentro de um conjunto de dados contínuo, e pode ser aplicado a um grande leque de categorias, tal como número de visitantes diários de um site, ou variações de preços de ações.

In [4]:
```python
Image(filename='line_size.png')
```

Out[4]:



Quando utilizar:

- Quando o dataset é demasiado grande para um gráfico de barras.
- Quando temos um conjunto de dados contínuo, que se altera ao longo do tempo
- Quando queremos mostrar múltiplas séries para o mesmo espaço de tempo
- Quando queremos visualizar tendências em vez de valores exatos

Quando evitar:

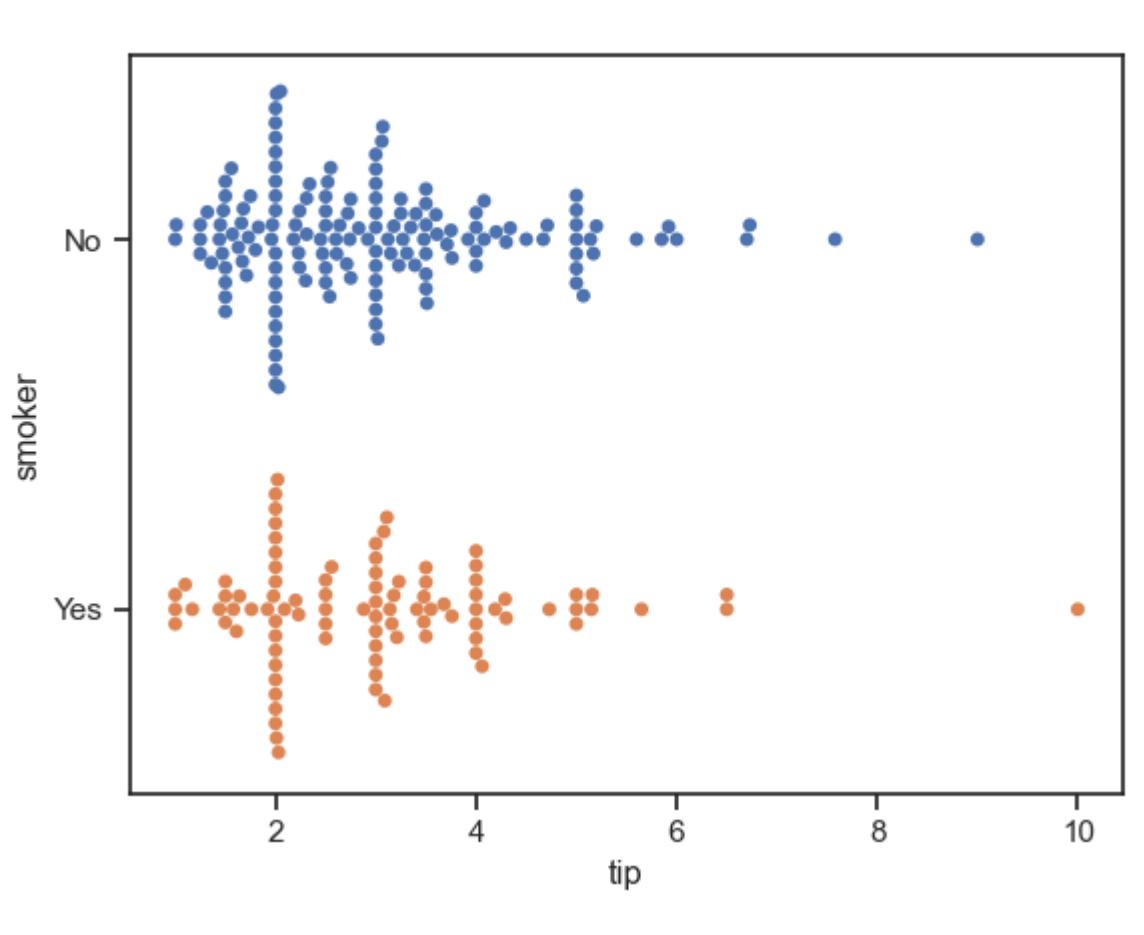
- Este tipo de gráficos funciona melhor com conjuntos de dados grandes. Se for pequeno, devemos utilizar um gráfico de barras

Scatter Plot

Um scatter plot usa pontos para representar valores para duas ou mais variáveis numéricas. A posiçaao de cada ponto indica o valor, em cada um dos eixos, para uma entrada no conjunto de dados. Estes são interessantes para se observar relações entre variáveis.

In [5]:
```python
Image(filename="scatterplot.png")
```

Out[5]:



Quando utilizar:

- Se o conjunto de dados contém pares de valores
- Para mostrar correlações e aglomerados em conjuntos de dados grandes
- Se a ordem dos pontos não é relevante

Quando evitar:

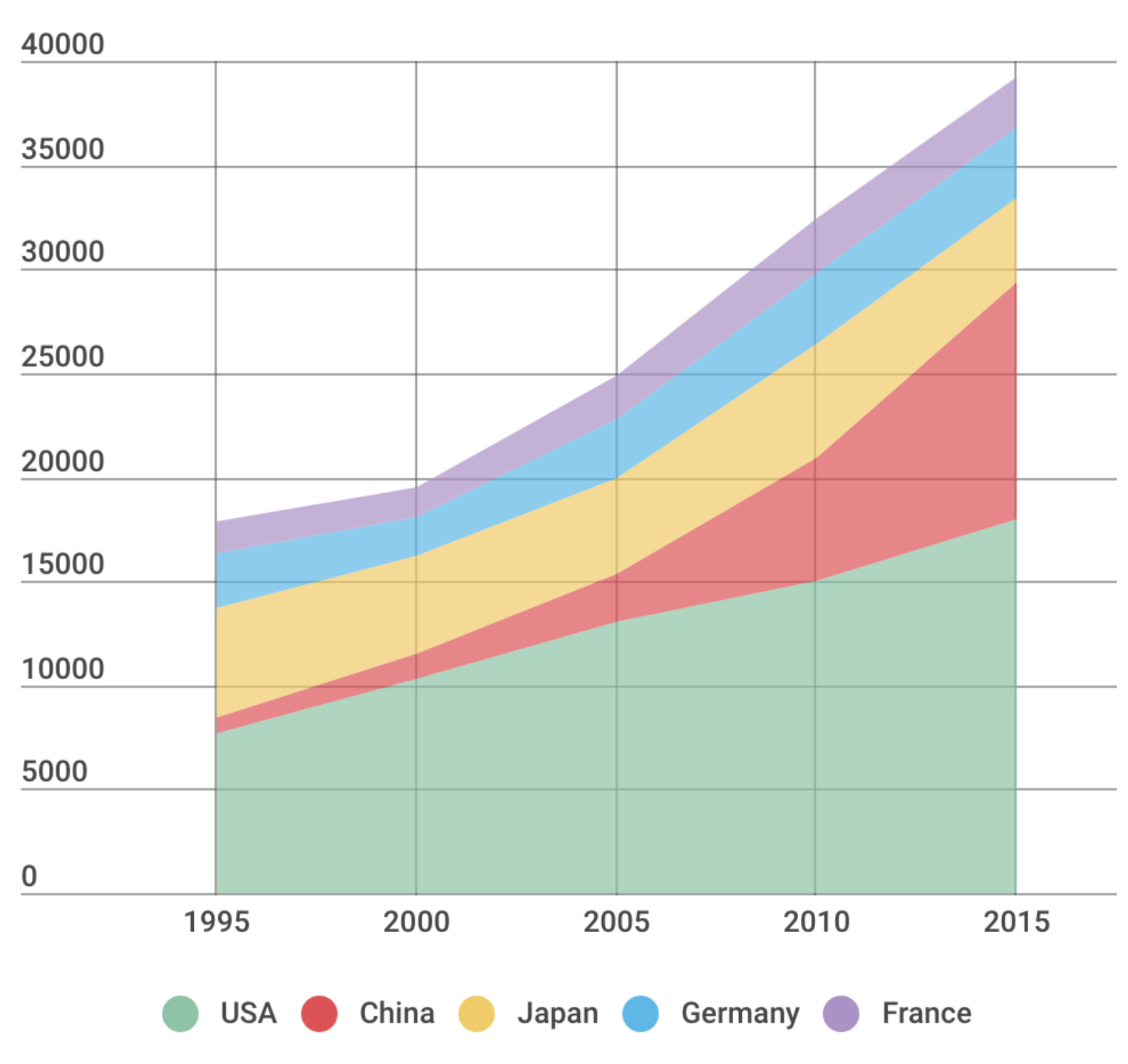
- Se tivermos um conjunto de dados pequeno
- Se os valores não estão relacionados.

Gráfico de área

Gráficos de área são muito semelhantes a gráficos de linhas, com algumas diferenças. Ambos permitem mostrar alterações ao longo do tempo, tendências, e continuidade ao longo de um conjunto de dados. No entanto, este tipo de gráficos acrescenta uma dimensão de informação, dado que o espaço entre as linhas é preenchido, indicando um volume.

In [6]:
```python
Image(filename="area.png")
```

Out[6]:



Quando utilizar:

- If you want to show part-to-whole relations.
- Quando queremos mostrar relações de partes para um todo, ao longo do tempo
- Quando queremos mostrar o volume do conjunto de dados

Quando evitar:

- Quando temos um conjunto de dados discreto.

Gráfico de bolhas

Este tipo de gráficos mostra três dimensões de dados. Cada ponto é um triplo de dados associados, onde a sua posição em x,y representa dois dados, e a sua área um terceiro. Estes facilitam a compreensão de valores socias, económicos, médicos, tal como um grande número de relaçoes científicas.

In [7]:
```python
Image(filename="bubble.png")
```

Out[7]:

Quando utilizar:

- Quando queremos comparar valores independentes
- Quando queremos demonstrar distribuiçoes.

Quando evitar:

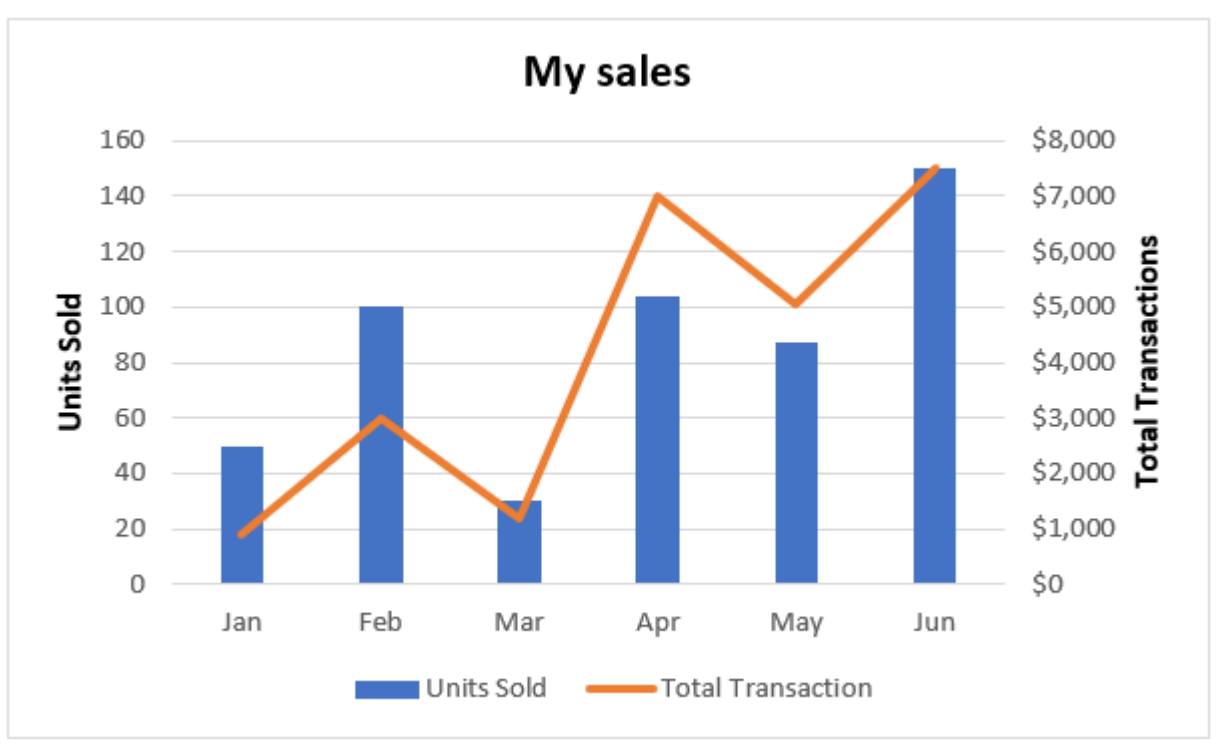
- Se temos um conjunto de dados pequeno

Gráficos compostos

Gráficos compostos são um tipo de visualização que combina um gráfico de linhas, com um de barras. Este mostra a data utilizando barras, e uma linha, que pode representar uma categoria em particular. A combinação destes pode ser útil para comparar valores em diferentes categorias, dado que mostra claramente que categoria é maior, ou mais pequena. Um exemplo disto pode ser visto ao utilizar este tipo de gráficos para estimativas de vendas, comparado com as vendas reais, para um dado período de tempo.

In [8]:
```python
Image(filename="combined.png")
```

Out[8]:



Quando utilizar:

- Quando queremos comparar valores para diferentes escalas.
- Quando os valores estão na mesma gama temporal

Quando evitar:

- Se queremos mostrar mais do que 2/3 tipos de gráficos. Nestes casos, o ideal será separar em gráficos independentes

Dicas finais de escolha de gráficos

Sempre que queremos escolher o nosso tipo de visualização de dados, devemos utilizar estas boas práticas, de modo a tornar a nossa escolha mais fácil, eficiente, e correta.

- Se temos dados categóricos, devemos utilizar um gráfico de barras, quando temos muitas categorias, e um gráfico circular, quando temos poucas
- Se temos dados nominais, devemos utiliar gráficos de barras, ou histogramas se os dados forem discretos. Gráficos de linhas/área são ideais para dados contínuos.
- Se queremos mostrar relações entre diferentes valores, devemos usar scatter plot, gráfico de bolhas, ou linhas.
- Se queremos comparar valores, devemos usar gráficos circulares para comparações relativas, e gráficos de barras para comparações mais precisas
- Se queremos comparar volumes, devemos utilizar um gráfico de área ou bolhas
- Se queremos mostrar tendências e padrões, devemos utilizar um gráfico de linhas, barras, ou scatter plot

Conclusão

Antes de escolhermos o tipo de gráfico a utilizar, devemos conhecer melhor o nosso conjunto de dados, a história que este quer contar, e quem vai ver os nossos resultados. Sempre que criamos um gráfico, devemos utilizar cores e tipos de letra simples. Uma visualização simples é geralmente melhor do que uma mais complexa. O principal objetivo da visualização de dados é torná-lo mais fácil de compreender e ler. Como tal, devemos evitar aglomerar muitos dados, torná-los demasiado densos e de difícil compreensão. Na maior parte dos casos, é melhor ter muitos gráficos simples, do que um demasiado complexo.