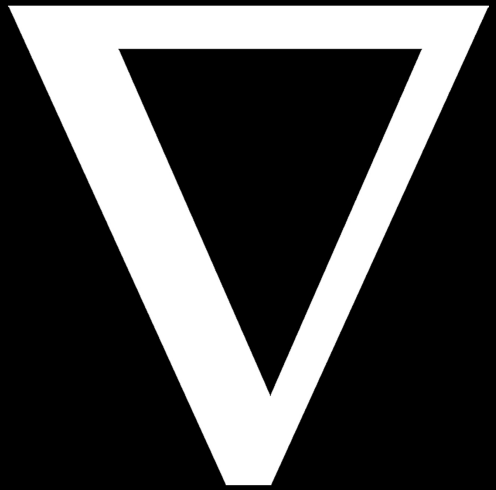


Análise Exploratória de Dados

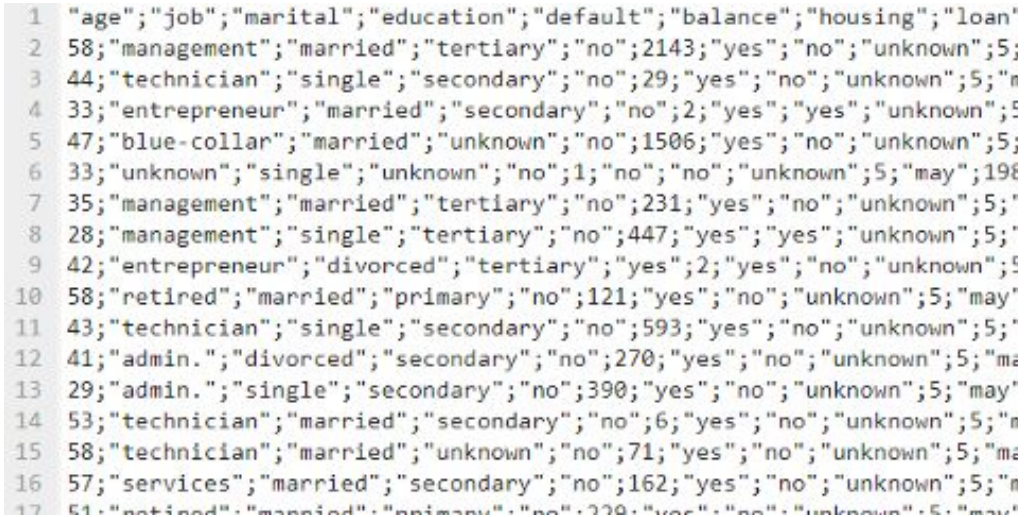
Aimê @itsaime

Conteúdo da aula

- Análise exploratória de dados
  - Análise intuitiva através da visualização de dados
  - Análise estatística
- Limpeza de dados
- Seleção de atributos

Introdução

- É importante compreender a complexidade dos dados que estamos lidando
- Precisamos mensurar a:
  - propriedades dos dados;
  - qualidade dos dados;
  - e o poder preditivo dos dados.



```
"age";"job";"marital";"education";"default";"balance";"housing";"loan"
58;"management";"married";"tertiary";"no";2143;"yes";"no";"unknown";5;
44;"technician";"single";"secondary";"no";29;"yes";"no";"unknown";5;"
33;"entrepreneur";"married";"secondary";"no";2;"yes";"yes";"unknown";5
47;"blue-collar";"married";"unknown";"no";1506;"yes";"no";"unknown";5;
33;"unknown";"single";"unknown";"no";1;"no";"no";"unknown";5;"may";198
35;"management";"married";"tertiary";"no";231;"yes";"no";"unknown";5;"
28;"management";"single";"tertiary";"no";447;"yes";"yes";"unknown";5;"
42;"entrepreneur";"divorced";"tertiary";"yes";2;"yes";"no";"unknown";5
58;"retired";"married";"primary";"no";121;"yes";"no";"unknown";5;"may"
43;"technician";"single";"secondary";"no";593;"yes";"no";"unknown";5;"
41;"admin.";"divorced";"secondary";"no";270;"yes";"no";"unknown";5;"ma
29;"admin.";"single";"secondary";"no";390;"yes";"no";"unknown";5;"may"
53;"technician";"married";"secondary";"no";6;"yes";"no";"unknown";5;"
58;"technician";"married";"unknown";"no";71;"yes";"no";"unknown";5;"ma
57;"services";"married";"secondary";"no";162;"yes";"no";"unknown";5;"
```

Análise estatística de dados

Esse passo é usado para qualquer dataset, não só os usados para ML. Iremos avaliar a qualidade e o poder preditivo dos dados.

Conteúdo:

- Análise univariada (descritiva)
- Análise bivariada (correlação)
  - Qualitativa
  - Quantitativa
- Análise contextual

Análise univariada (descritiva)

Atributo comum

- **Tipo do dado** (importante!!!)
- Valores faltantes (printar tabela com porcentagem de cada coluna)

Atributo categórico

- cardinalidade (baixa == muitos elementos duplicados)
- quantidade de valores únicos

Atributo numérico

- Quantis estatísticos (Q1, Q2, Q3, min, max, range)
- Estatística descritiva (mean, mode, sd, median, kurtosis, skewness)
- Histograma de distribuição

Atributo textual

- tokens
- document frequency (DF)/ term frequency (TF) sem palavras vazias (stop words)

Passo crucial.

Verificar atributos comuns: valores faltantes e tipo de dado das colunas.

	Missing Values	Percentage	Data Types
Id	0	0.000000	int64
MSSubClass	0	0.000000	int64
MSZoning	0	0.000000	object
LotFrontage	259	17.739726	float64
LotArea	0	0.000000	int64
...	...	...	...
MoSold	0	0.000000	int64
YrSold	0	0.000000	int64
SaleType	0	0.000000	object
SaleCondition	0	0.000000	object
SalePrice	0	0.000000	int64

81 rows × 3 columns

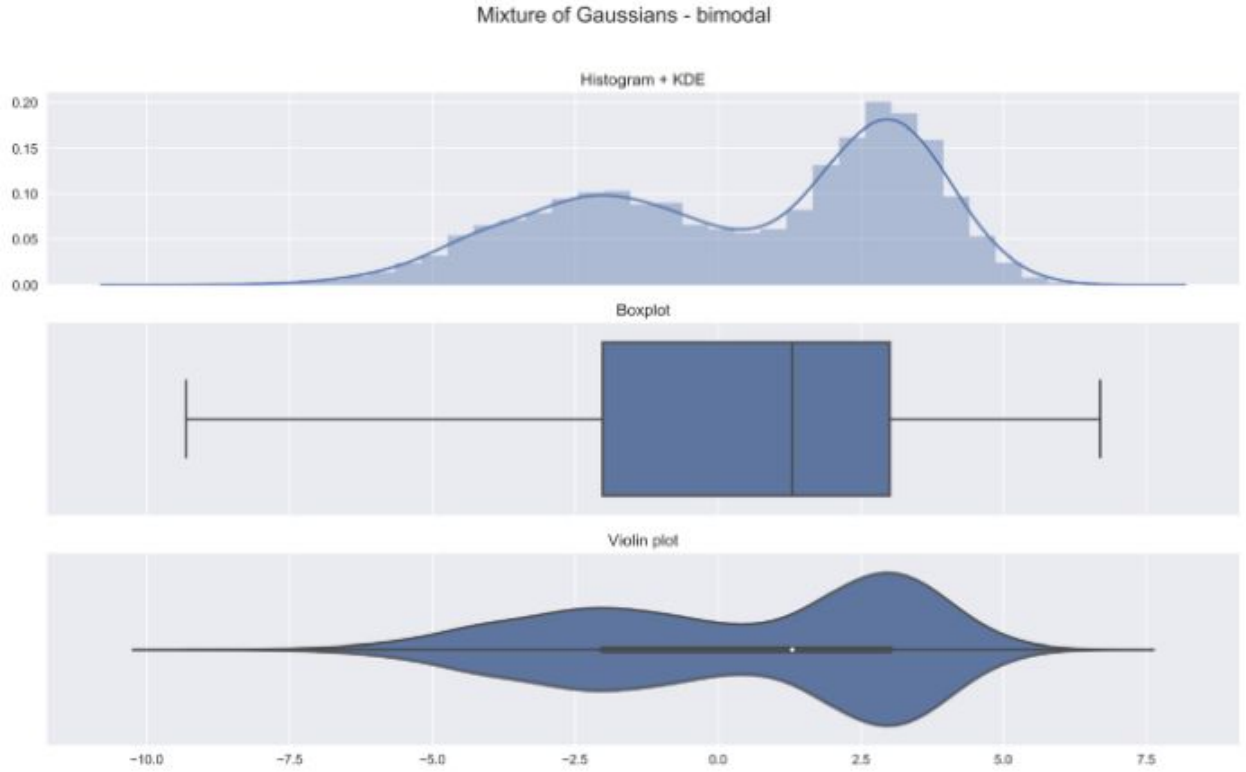
Limpeza dos dados

- Remover duplicados
- Remover dados irrelevantes
- Padronizar capitalização
- Converter tipos de dado
- Remover erros e inconsistências
- Todos dados na mesma linguagem
- Lidar com dados faltantes
- Lidar com outliers

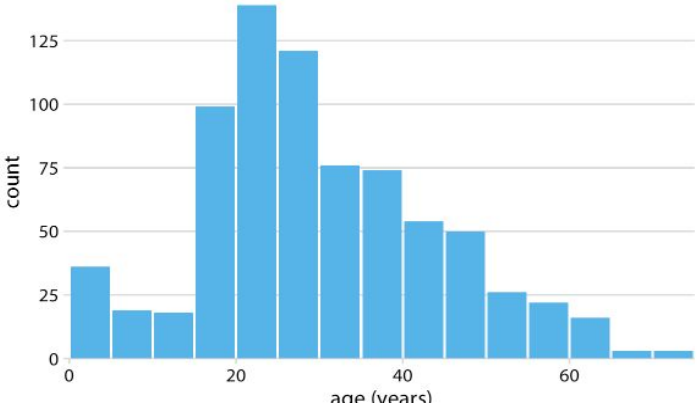
Atributos numéricos

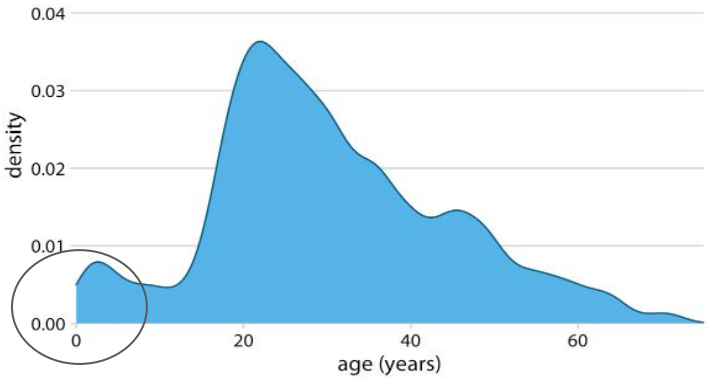
	max	range	IQR	mode	mad	kurtosis	skewness
age	90	73	20	36	11.189182	-0.166127	0.558743
functional_weight	1484705	1472420	119224	123011	77608.21854	6.218811	1.44698
education_num	16	15	3	9	1.903048	0.623444	-0.311676
capital_gain	99999	99999	0	0	1977.373437	154.799438	11.953848
capital_loss	4356	4356	0	0	166.462055	20.376802	4.594629
hours_per_week	99	98	5	40	7.583228	2.916687	0.227643

	mean	std	min	25%	50%	75%
age	38.581647	13.640433	17	28	37	48
functional_weight	189778.3665	105549.9777	12285	117827	178356	237051
education_num	10.080679	2.57272	1	9	10	12
capital_gain	1077.648844	7385.292085	0	0	0	0
capital_loss	87.30383	402.960219	0	0	0	0
hours_per_week	40.437456	12.347429	1	40	40	45



Podem ser representados por gráficos





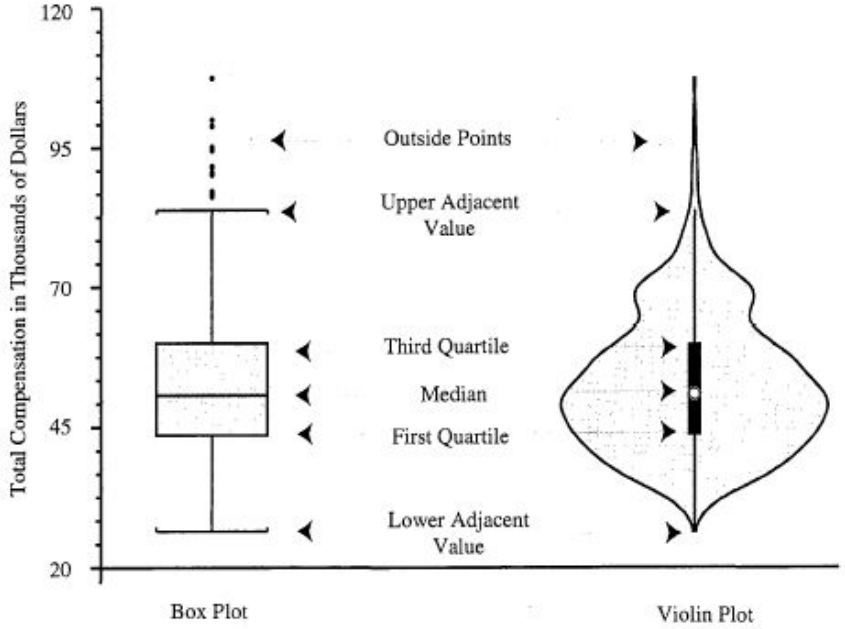


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

Plot de histograma, densidade, boxplot e violino.

Análise bivariada (de correlação)

Examina correlação entre variáveis, dois a dois. Determina variáveis redundantes e/ou que não são preditivas em relação ao valor alvo.

Qualitativa:

- 2 categóricas: tabela de contingência
- categórica vs. numérica: estatística descritiva ou histograma

Problemas: enviesamento (verificado com métodos de amostragem e design de experimentos)

Quantitativa:

Teste de Hipóteses, onde a hipótese nula é de que não há relação entre as duas variáveis. Usamos p-valor de 0.05.

Tabela de contingência

Item	Extra
Pizza	Cheese
Pizza	Sauce
Burger	Cheese
Burger	Cheese
Burger	Sauce

	Cheese	Sauce
Pizza	1	1
Burger	2	1

Sex \ Handed-ness	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

Teste de hipóteses

X	Y	
	Categorical	Numerical
Categorical	Chi-square test Information gain	Student T-test ANOVA Logistic regression Discretize *Y* (left column)
Numerical	Student T-test ANOVA Logistic regression Discretize *X* (row above)	Correlation Linear Regression Discretize *Y* (left column) Discretize *X* (row above)

Handedness / Sex	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

Handedness / Sex	Right-handed	Left-handed	Total
Male	100*0.52*0.87=45	100*0.52*0.13=7	52
Female	100*0.48*0.87=42	100*0.48*0.13=6	48
Total	**87**	**13**	**100**

Test the statistics.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(43-45)^2}{45} + \frac{(44-42)^2}{42} + \frac{(9-7)^2}{7} + \frac{(4-6)^2}{6} = 1.42$$

$$p_{value} = Pr_{DF=1}(\chi^2 > 1.42) = 0.233$$

Exemplo de Teste Chi-Quadrado

Análise Contextual

Entender o contexto no qual o dataset está inserido pode nos dar insights em relação aos dados.

Baseada em tempo:

- horas de rush
- quebra da bolsa de 2008
- datas de eleição
- black friday

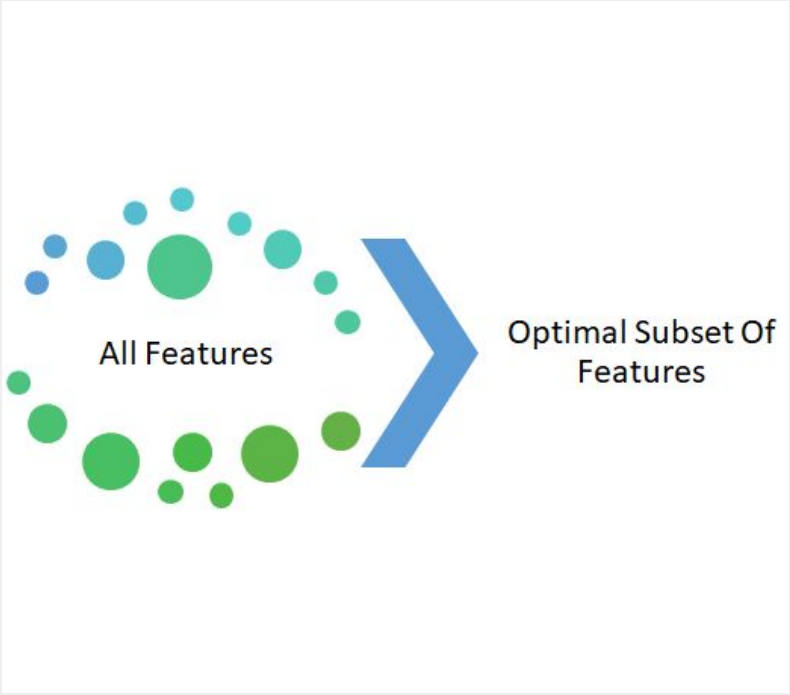
Baseado em usuários.

- verificar comportamento de cada usuário ou em conjunto
- achar anomalias

Seleção de atributos

Baseado em:

- Análise descritiva
- Análise de correlação
- Análise contextual



Seleção de atributos

Baseado em análise descritiva

- Alta porcentagem de números faltantes não constituindo informação
- Pouca variância. Remover esses dados pode diminuir complexidade da ML
- Pouca entropia de categorias pode enviesar ML
    - Desbalanço de classe pode ser resolvida com amostragem
- Distribuição distorcida com calda longa. Pode ser resultado de erros que geram outliers.
- Alta cardinalidade, ou seja, muitos valores únicos

Baseado em análise de correlação

- Correlação muito baixa pode significar que atributo não será útil para previsão desejada
- Alta correlação. Não acrescenta informação ao modelo e não é recomendada para modelos lineares.

Baseado em análise contextual

Leva em conta conhecimento do contexto em que os dados estão inseridos.

Visualização de dados

Tabelas

Guiar o leitor para a informação que queremos passar. Procurar não usar bordas, pois poluem a visão.

Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

LOW-HIGH

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

FIGURE 2.5 Two views of the same data

Heavy borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Light borders

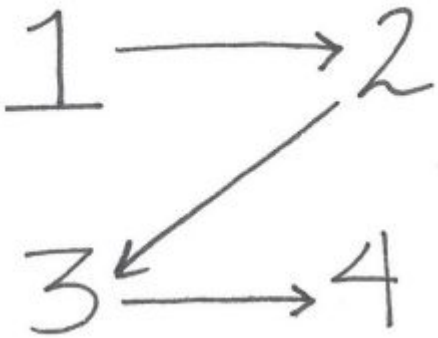
Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Minimal borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

FIGURE 2.4 Table borders

Escolhas sobre o design dos gráficos



Caminho do foco do leitor

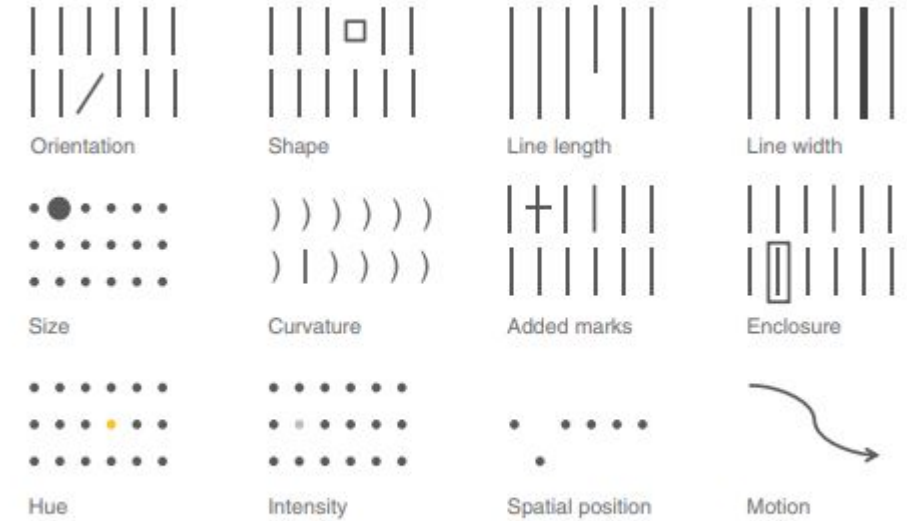


FIGURE 4.4 Preattentive attributes

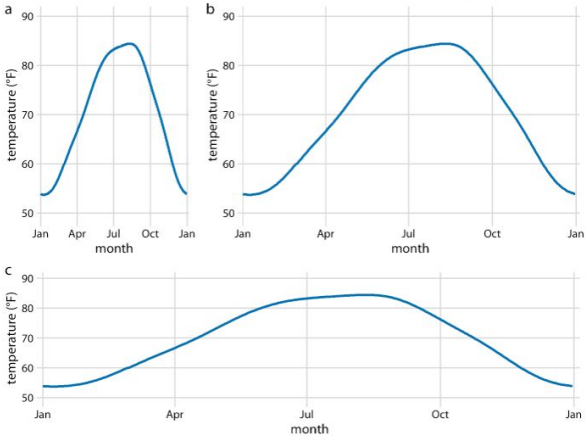
Source: Adapted from Stephen Few's *Show Me the Numbers*, 2004.

Recursos para destacar atributos.

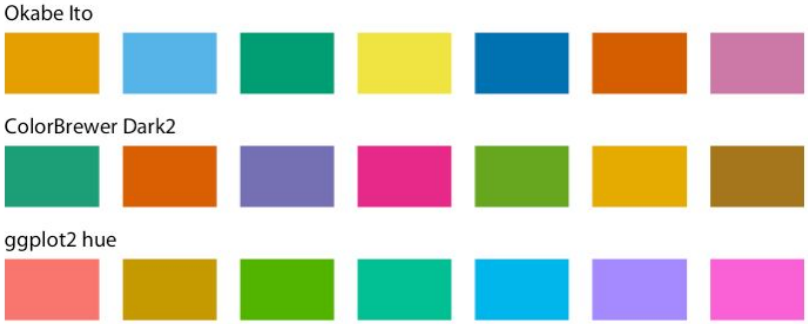
Escolhas sobre o design dos gráficos

FIGURE 2.14 Bar width

Barras do gráfico

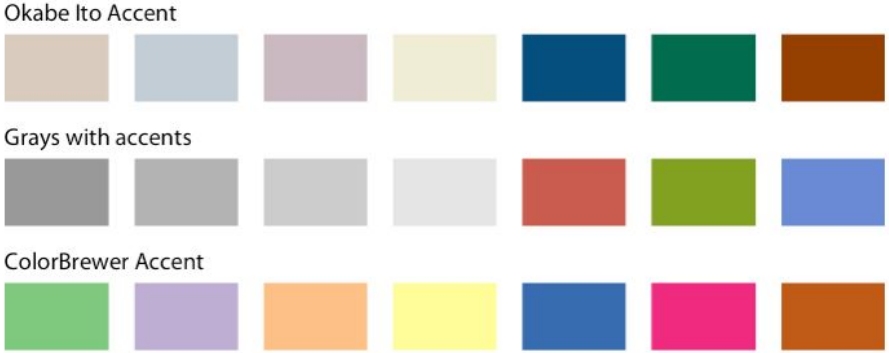


Escala do gráfico



Categórico

Quantitativo



Destaque

Amigável a daltônicos

Círculo cromático

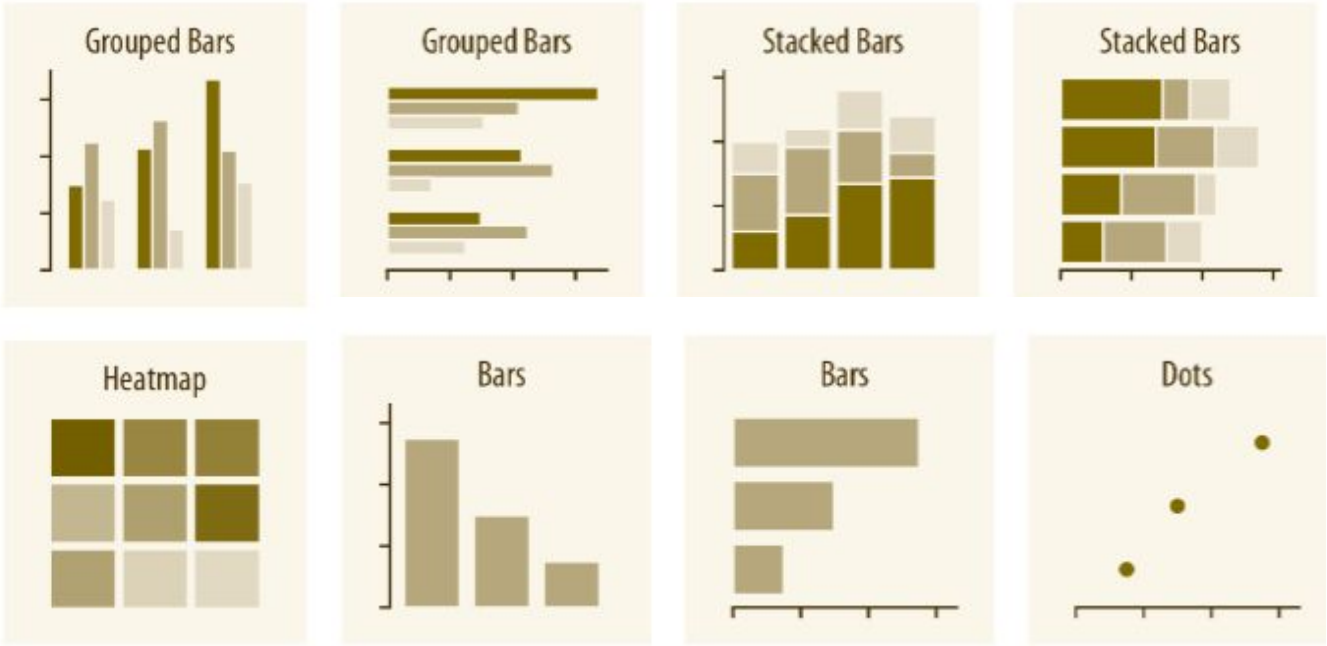
Que informação devemos apresentar?

Quem: a quem iremos falar

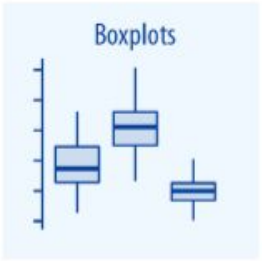
O que: ideia que queremos passar

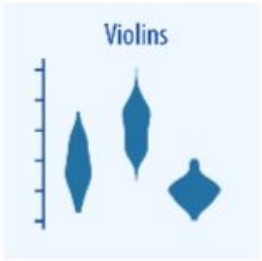
Como: escolha da informação apresentada

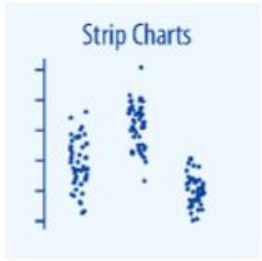
Visualizando quantidades

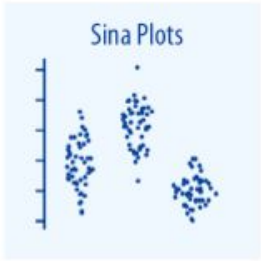


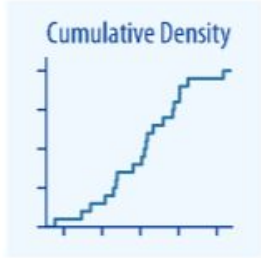
Visualizando distribuições

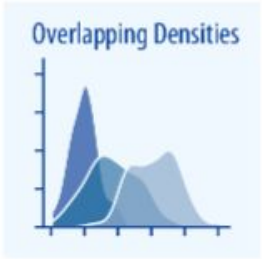


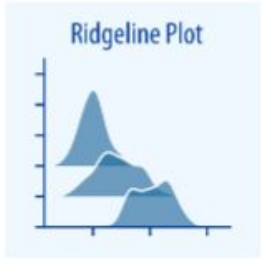


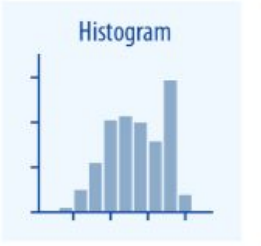


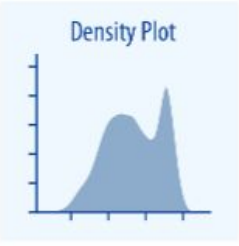








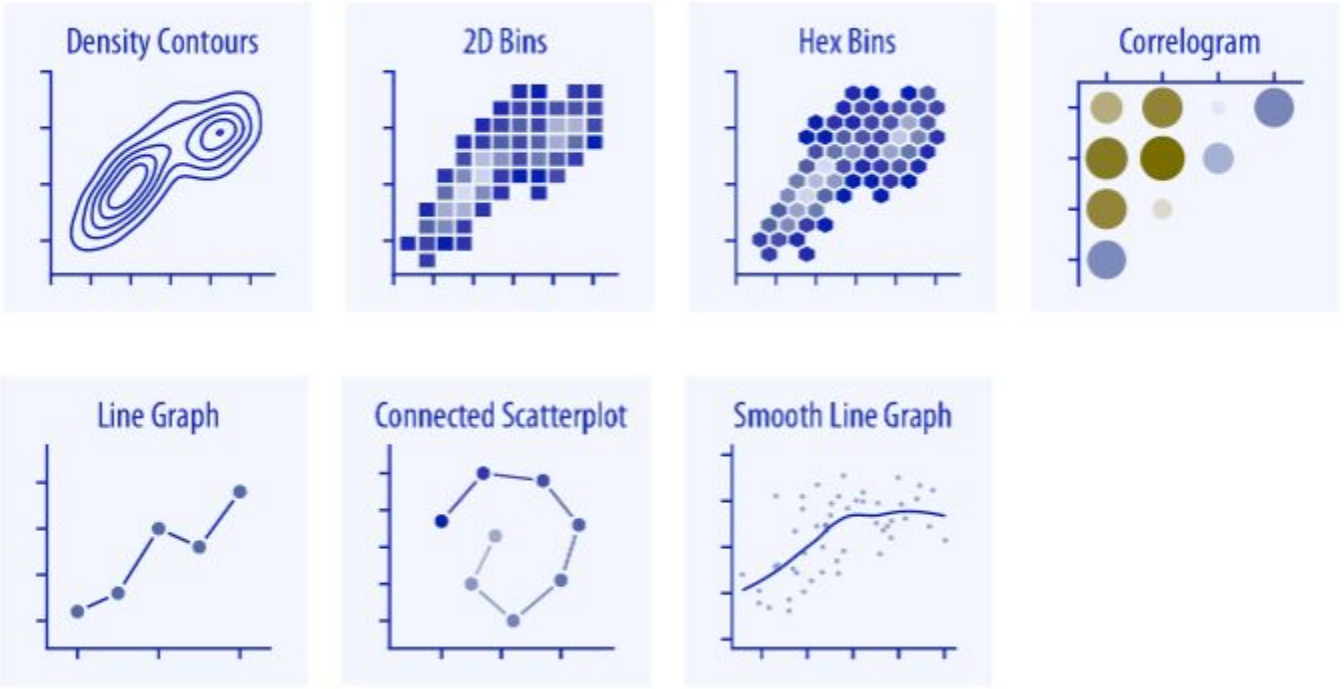




Visualizando proporções

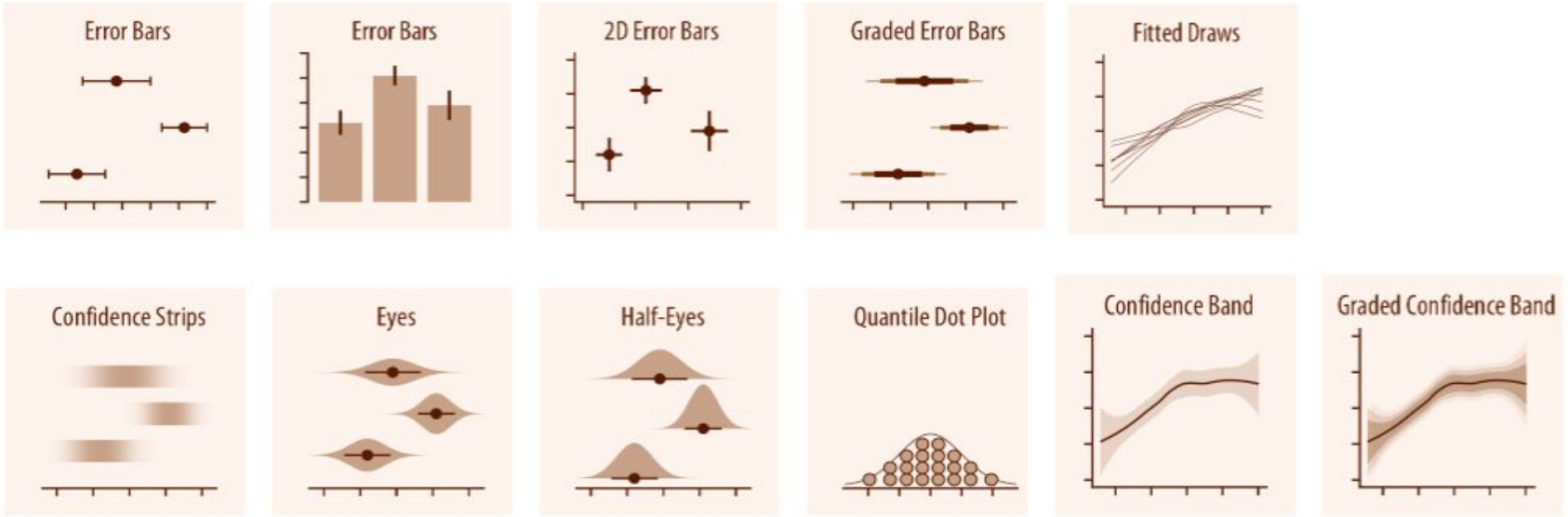
Visualizando proporções

Relações entre duas variáveis



Dado geoespacial

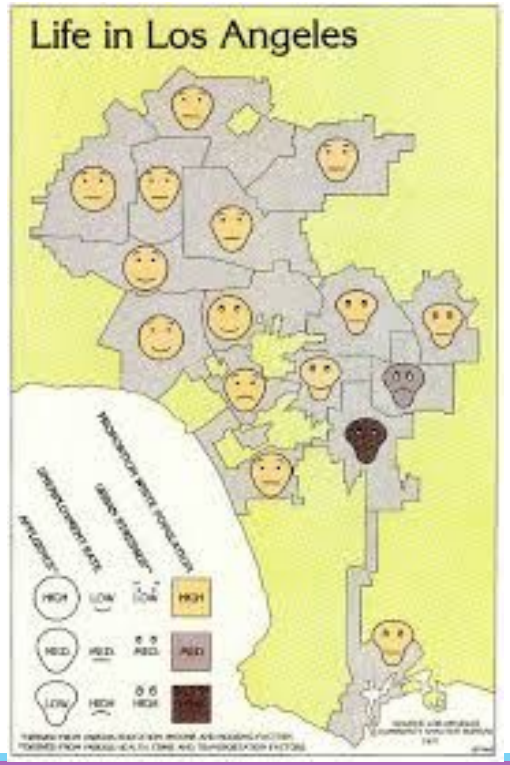
Visualizando incerteza



Séries temporais

Figure 14.1: Daily closing values of the Dow Jones Industrial Average for the year 2009. Data source: Yahoo! Finance

Faces de Chernoff



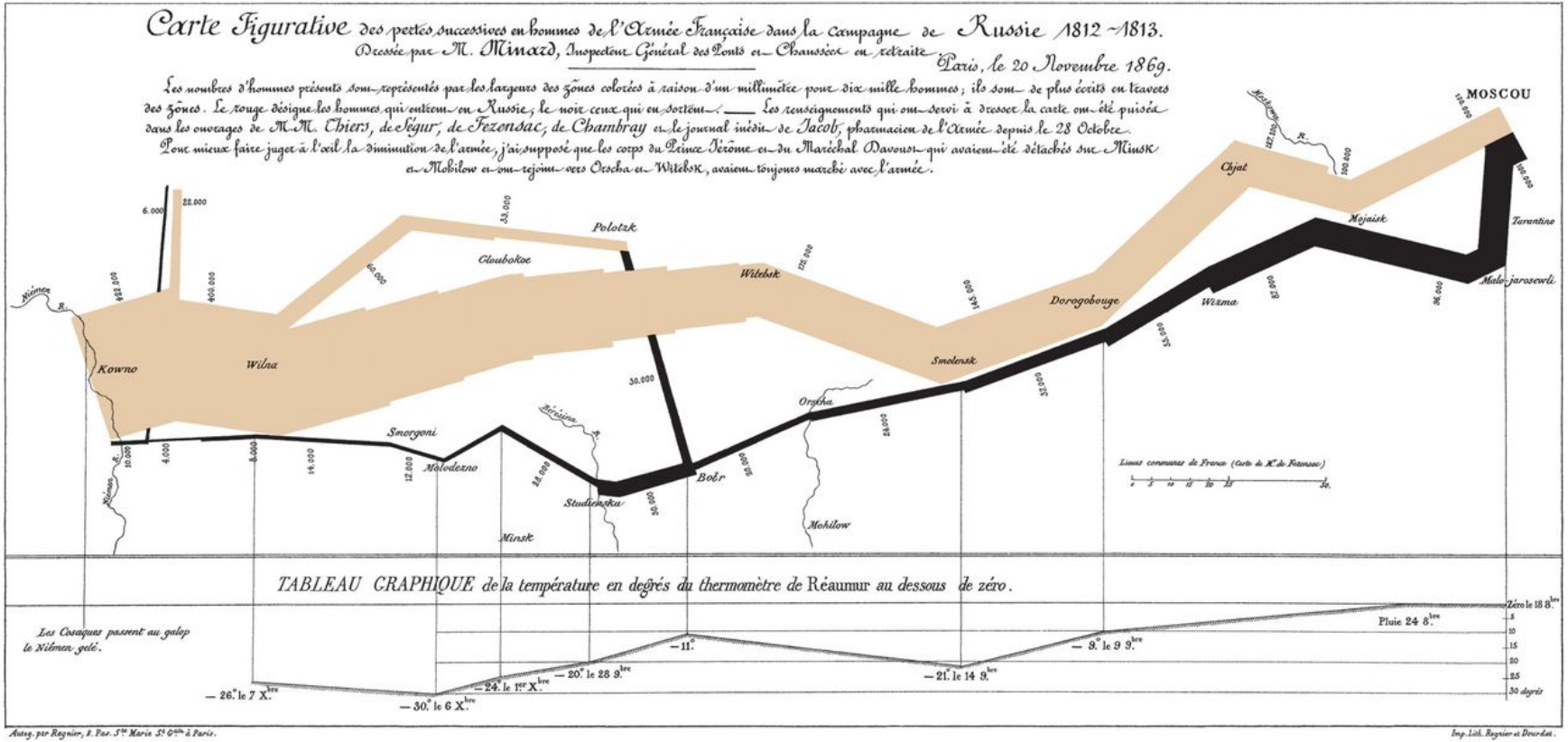


Diagrama de Sankey

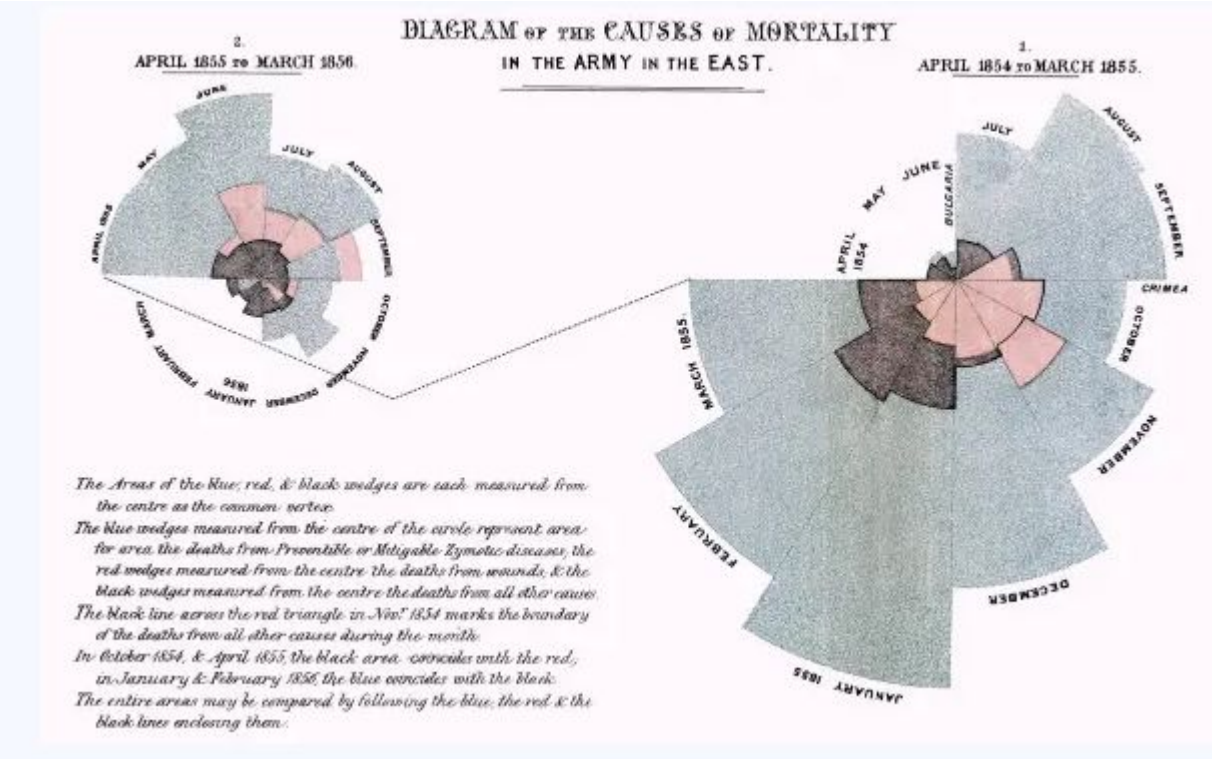
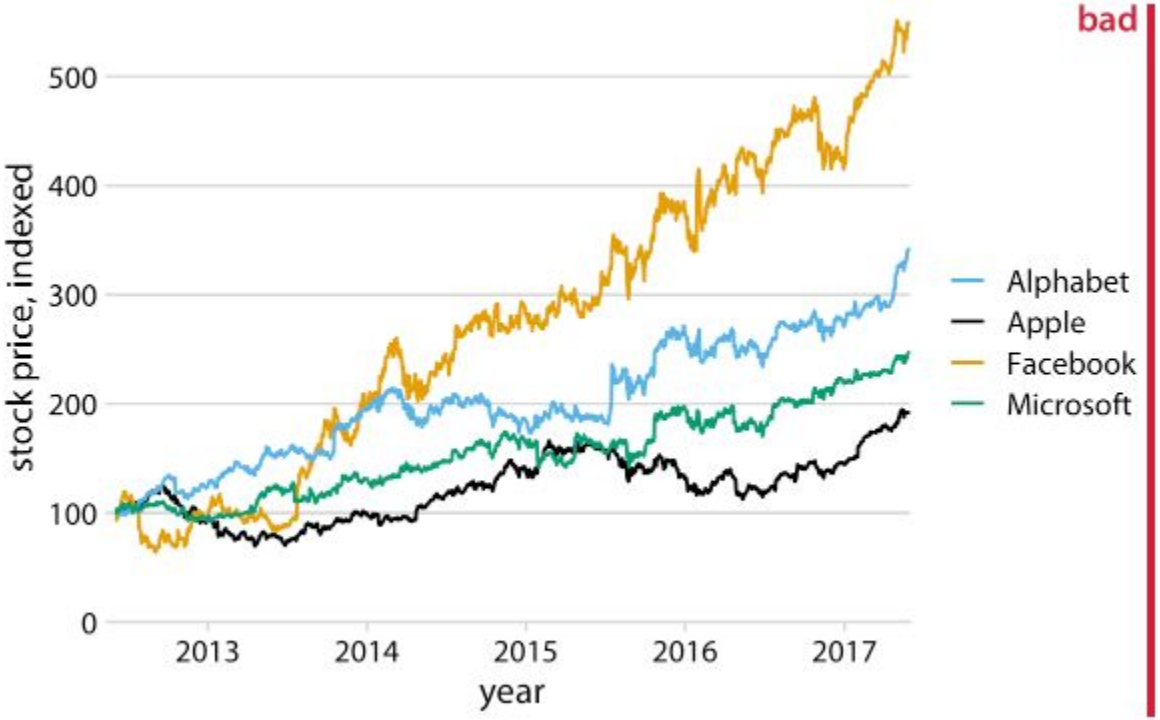
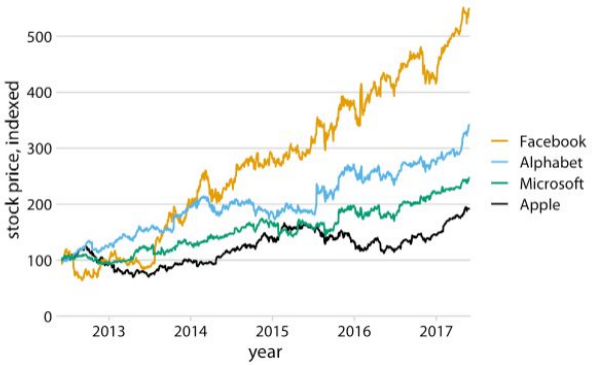
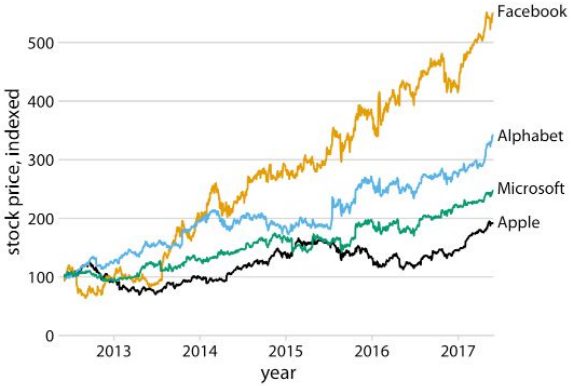


Gráfico de Florence Nightingale (Coxcomb Plot)

Atenção às legendas







Importância da escolha do gráfico

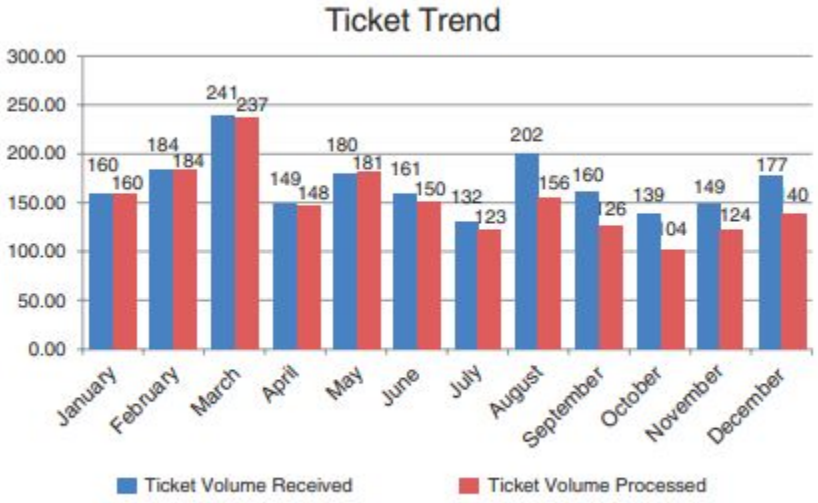
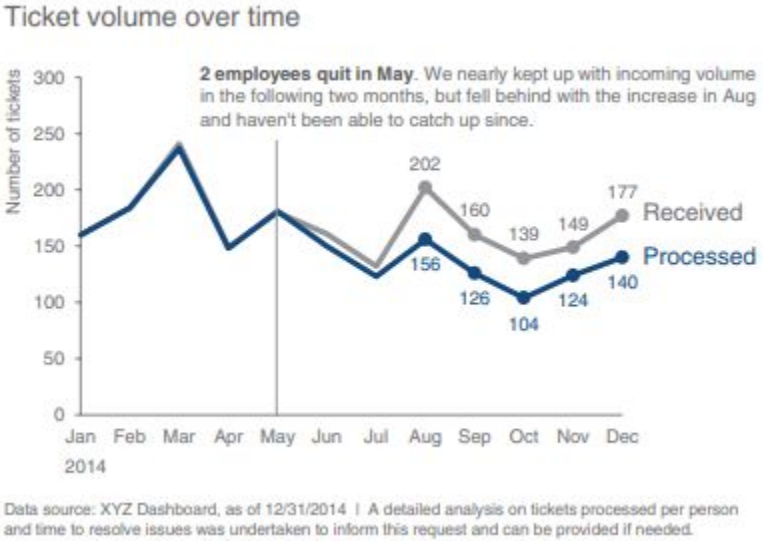


FIGURE 0.2 Example 1 (before): showing data



Importância da escolha do gráfico

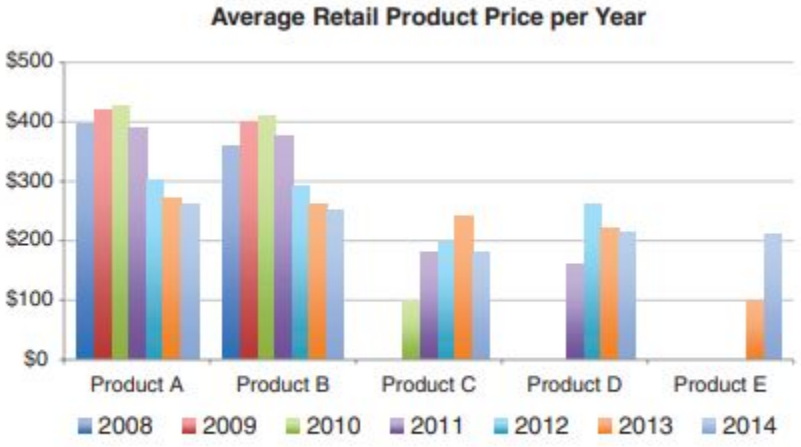


FIGURE 0.6 Example 3 (before): showing data

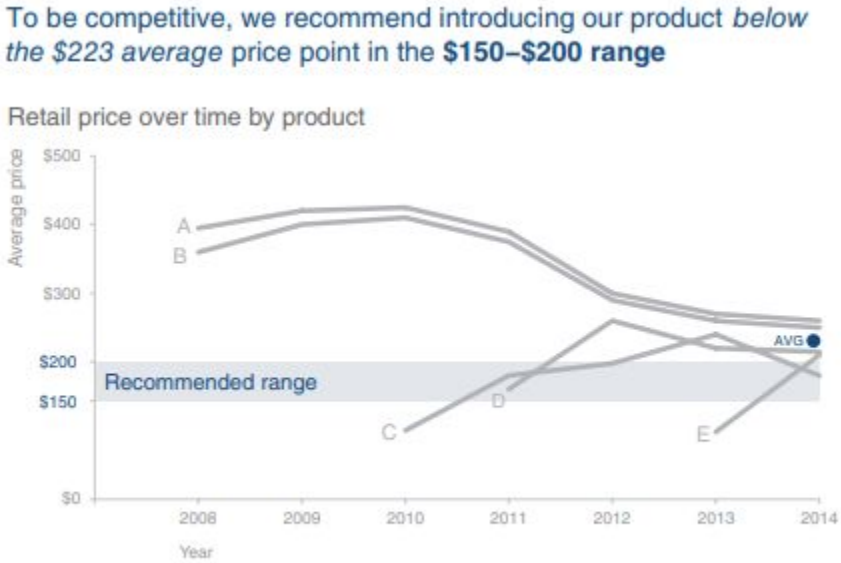


FIGURE 0.7 Example 3 (after): storytelling with data

Importância da escolha do gráfico

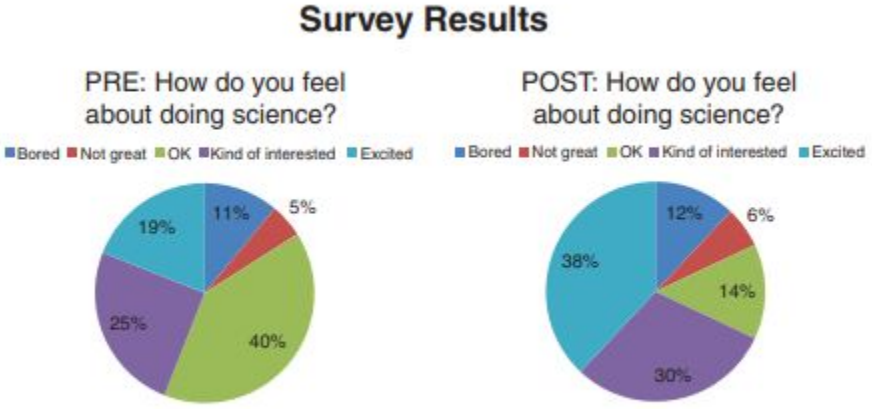


FIGURE 0.4 Example 2 (before): showing data

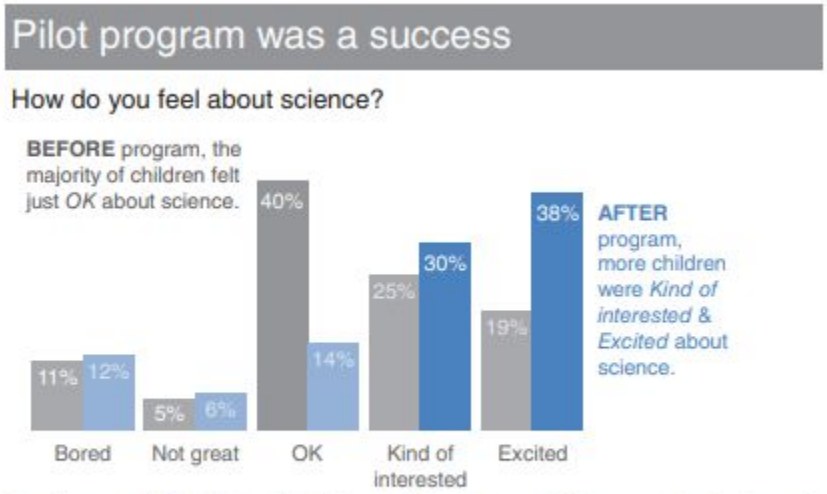
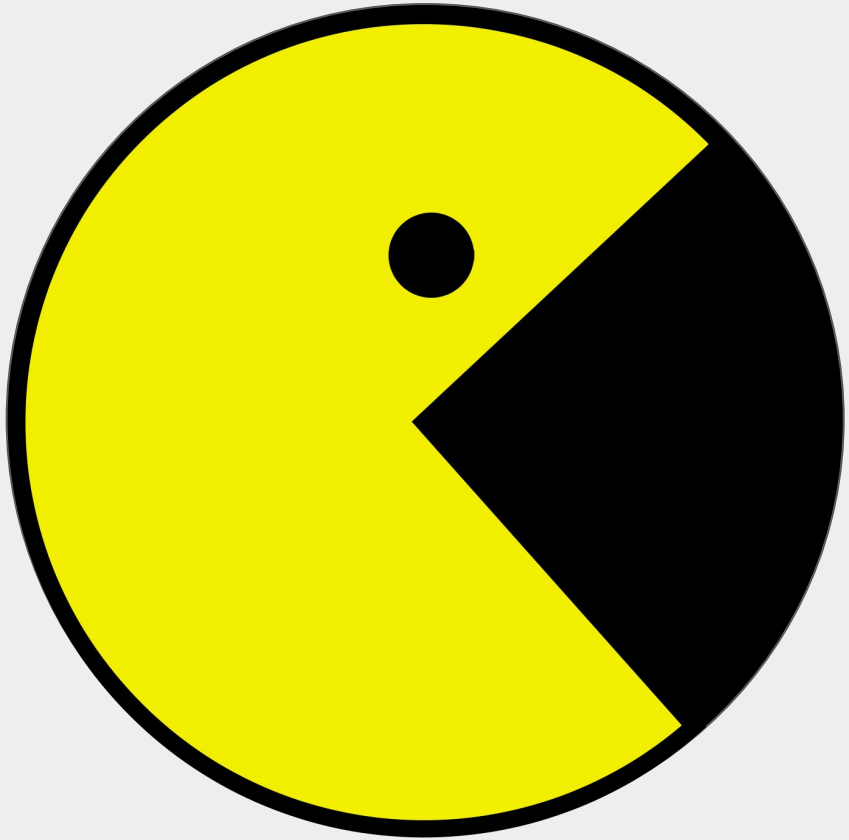


FIGURE 0.5 Example 2 (after): storytelling with data

Gráfico de pizza

Evite o gráfico de pizza. Se fizer questão, saiba que não passa muita informação. Use então a regra do Pacman.



75%

Muito mais impactante que um gráfico de pizza.

Scatterplot

FIGURE 2.6 Scatterplot

FIGURE 2.7 Modified scatterplot

Slopegraph

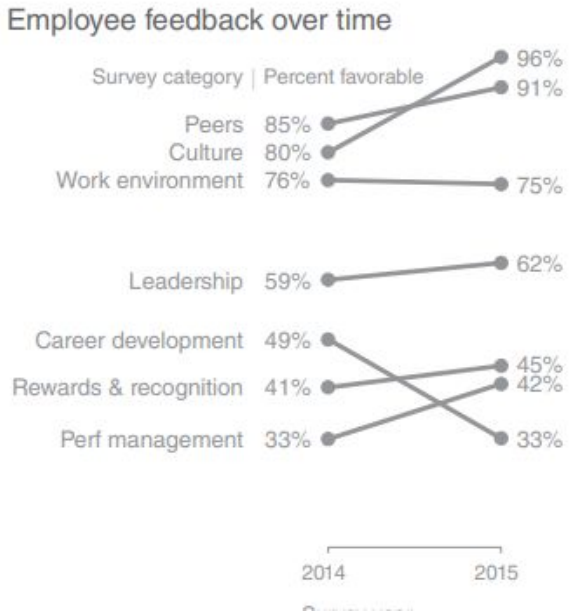


FIGURE 2.11 Modified slopegraph

Elimine a poluição visual

FIGURE 3.24 Before-and-after

Foco direcionado

FIGURE 4.11 First, push everything to the background

FIGURE 4.14 Data labels used sparingly help draw attention

Acréscimo de informações no gráfico

FIGURE 4.14 Data labels used sparingly help draw attention

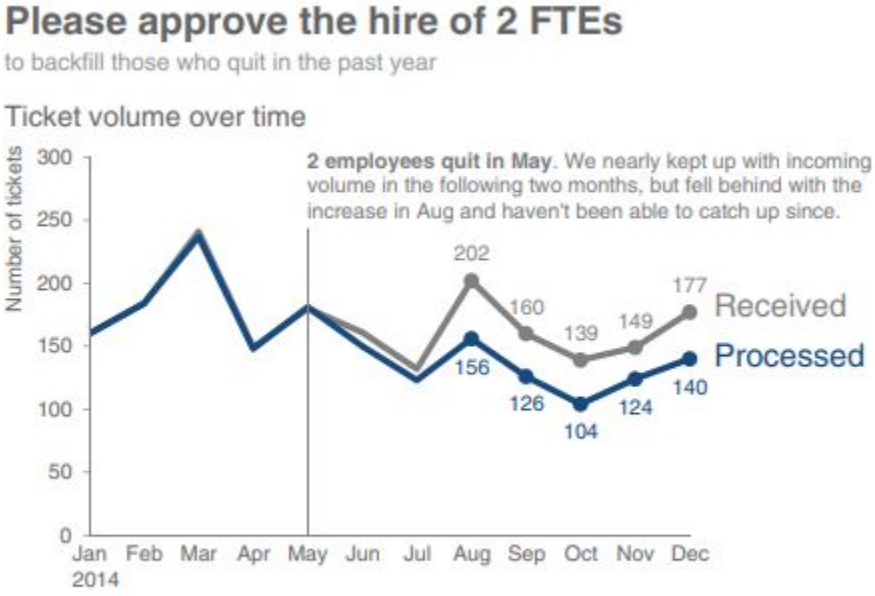
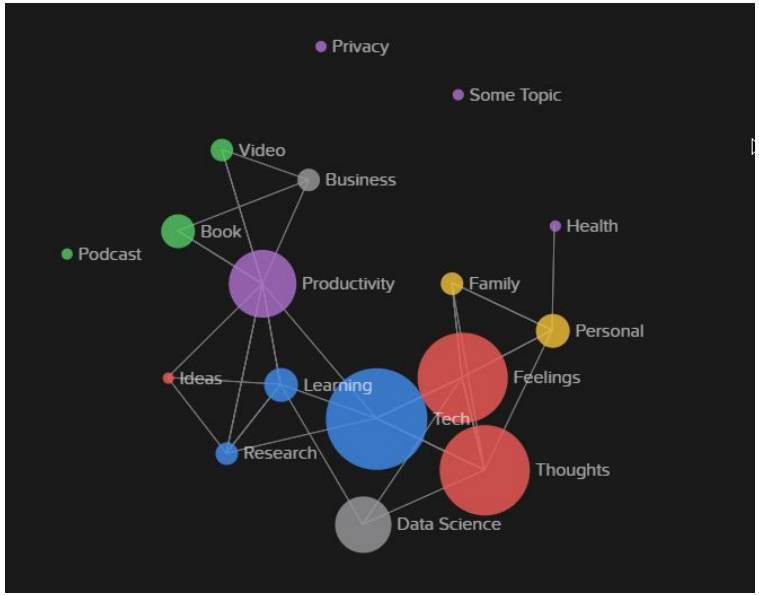


FIGURE 5.10 Add action title and annotation

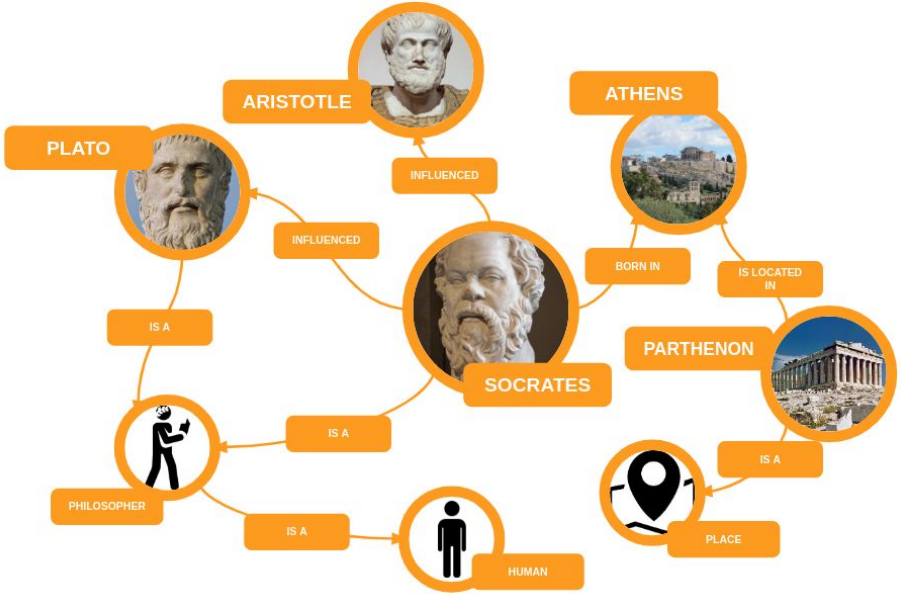
Exemplo

FIGURE 8.20 Before-and-after

Extra: Visualização informação através de grafos

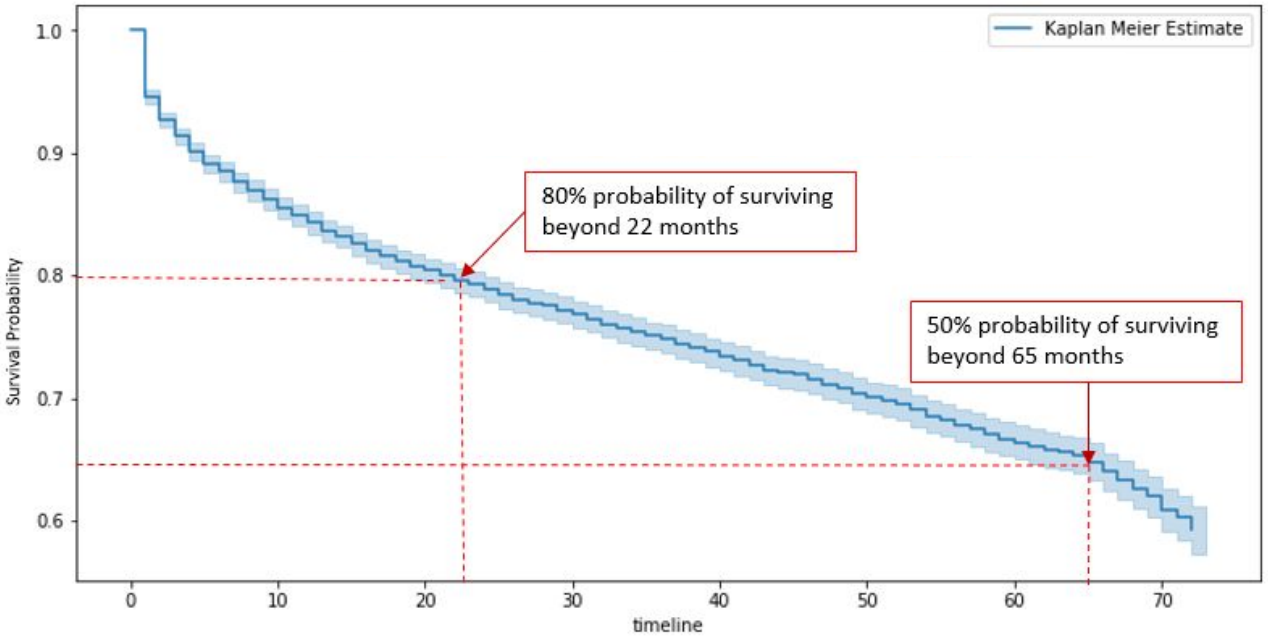


Grafo do aplicativo Obsidian



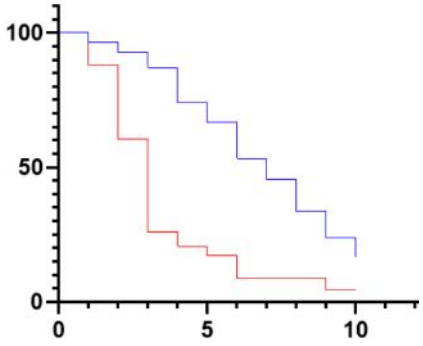
Grafo de Conhecimento

Extra: Análise de sobrevivência



Biblioteca python: lifelines

Mais a fundo: usar R



Ferramentas

- Python
  - Seaborn
  - Matplotlib
  - Plotly
  - geoplotlib
  - Folium
- R
  - Gráficos mais específicos de estatística, como QQ plot
  - Gráficos interativos

Obs: olhar sempre a documentação.

Ferramentas de DashBoard

- Python
- Data Studio (Google)
- PowerBI
- Tableau
- InDesign
- Slides (Google)

Inspiração

- Pinterest
- Behance (Adobe)
- Data Studio Report Gallery

Bibliografia

Fundamentals of Data Visualization - Claus O. Wilke

Exploratory Data Analysis for Feature Selection in Machine Learning - Google Cloud

https://datavizcatalogue.com/

Storytelling with data - Cole Nussbaumer Knaflic

dúvidas?