Inteligência Artificial na Saúde

Exploração e mineração de dados

Júlio César Batista, M. Sc.
julio.batista@outlook.com

Agenda

- Aquisição de dados
- Limpeza de dados
- Similaridade
- Visualização de dados
- Prática de análise exploratória de dados

Aquisição de dados

Elaborando um experimento

Saber qual problema quer responder ajuda a desenhar um experimento que pode validar uma hipótese, o que vamos medir como critério de sucesso e quais dados vamos precisar coletar

- Exemplo 1: Existe diferença na qualidade de diagnóstico médico após o horário de almoço em relação às outras horas do dia?

- Exemplo 2: Diagnóstico auxiliado por computador melhora reduz a quantidade de erros de diagnóstico?

Raspagem de dados (web scraping)

- Um software (bot) que extrai informações de websites automaticamente
- Dessa forma o Google encontra as páginas na internet para o mecanismo de busca

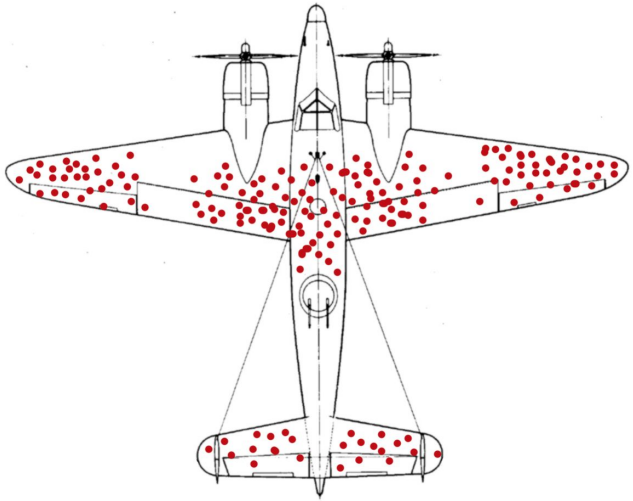
Exemplo: https://github.com/LaCAfe/Bulario2018PT-br

Bases de dados para pesquisa

- UCI
- Kaggle
- Google Datasets
- https://basedosdados.org/en/

Aquisição de dados

- Entrevista ou formulário para obter respostas
  - As perguntas devem ser neutras para não conduzir a respostas específicas
  - Busque uma amostragem de toda a população e não apenas um subconjunto (viés do sobrevivente)
  - Cuidado com os incentivos para participar do estudo
  - Respostas podem ser imprecisas



Viés do sobrevivente: Survivorship bias, Wikipedia.
Incentivos ruins: https://www.geckoboard.com/best-practice/statistical-fallacies/#cobra-effect

Aquisição de dados

- Sensores
  - Usuário pode ter se movido durante o CT
  - Não respeitou o período de jejum para exames de sangue
  - Medir a pressão após esforço e não estar em repouso

Aquisição de dados

- Responder a seguinte pesquisa
  - https://forms.gle/7maLZtYQ5VWJa6Br6

Limpeza de dados

Preparar os dados para análise

- Padronizar unidades de medida

Altura
1.89m
189cm

Altura (cm)
189
189

Preparar os dados para análise

- Padronizar nomenclaturas

Bebida alcoólica	Fumante
V	Não
F	Sim

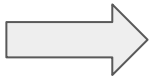
Bebida alcoólica	Fumante
V	F
F	V

Preparar os dados para análise

- Filtrar ou corrigir valores

Idade
150
-
-1
110

Idade
-
-
-
11

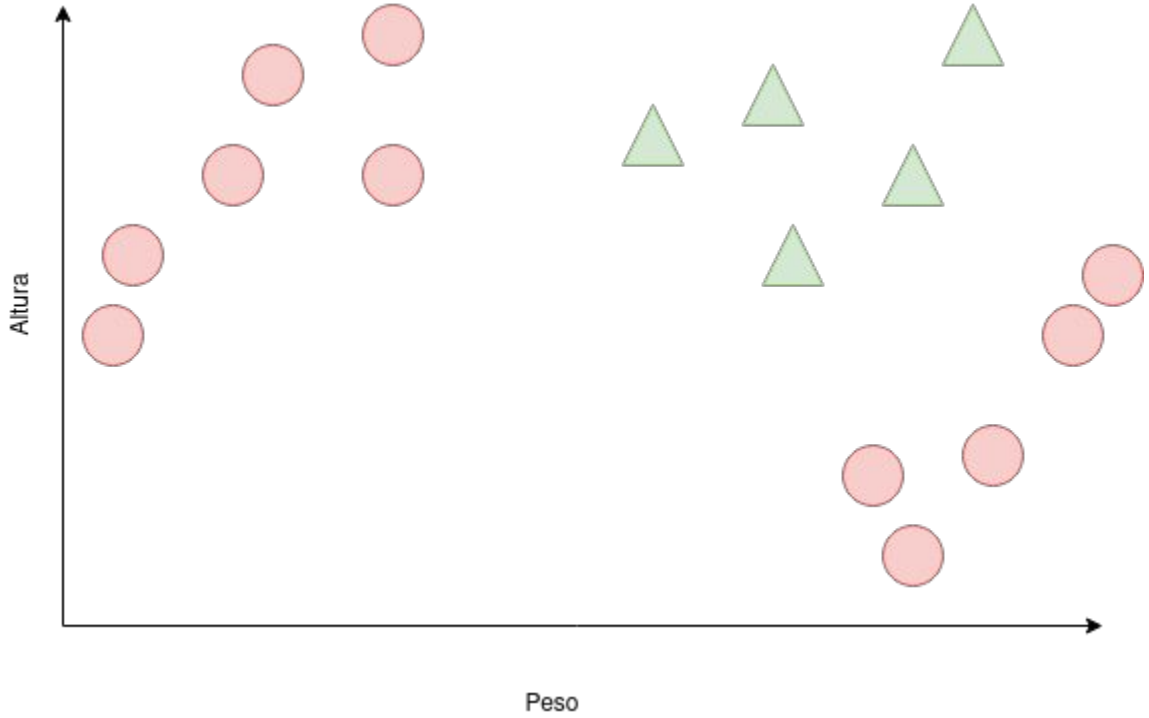


Idade
11

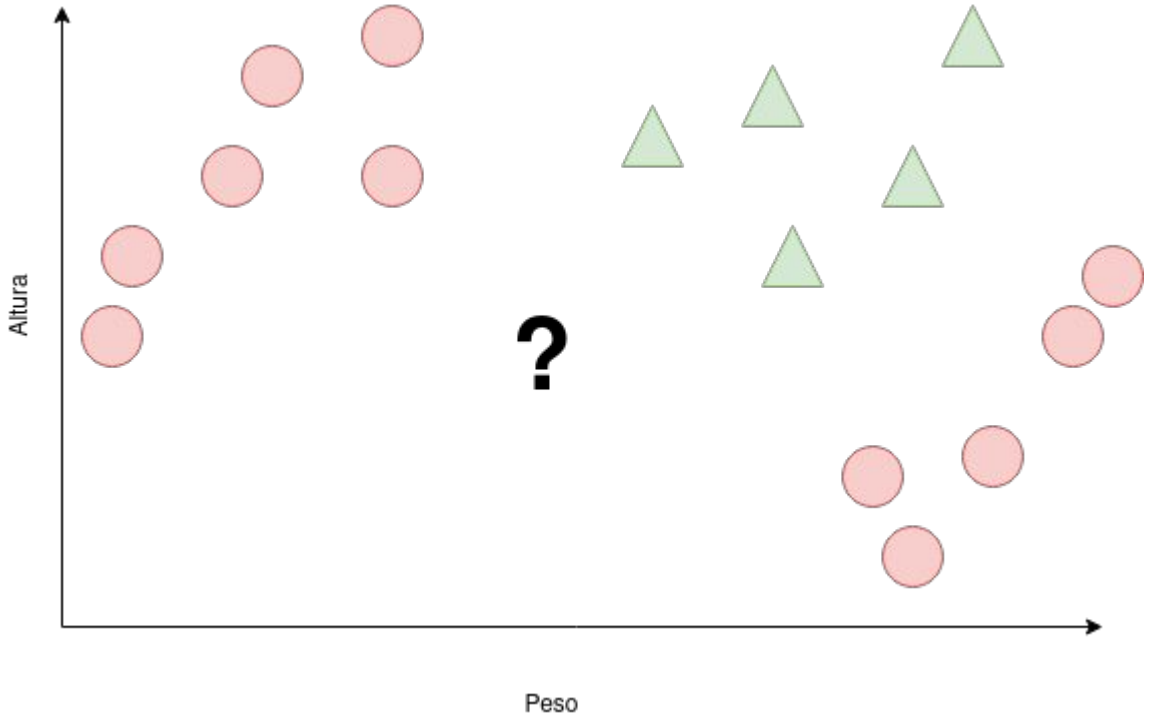
Problemas com datas

- 12/11/2021 é 12 de Novembro ou 11 de Dezembro?
  - Se possível, opte por trabalhar com o padrão ISO YYYY-MM-DD
  - 2021-11-12 é 12 de Novembro

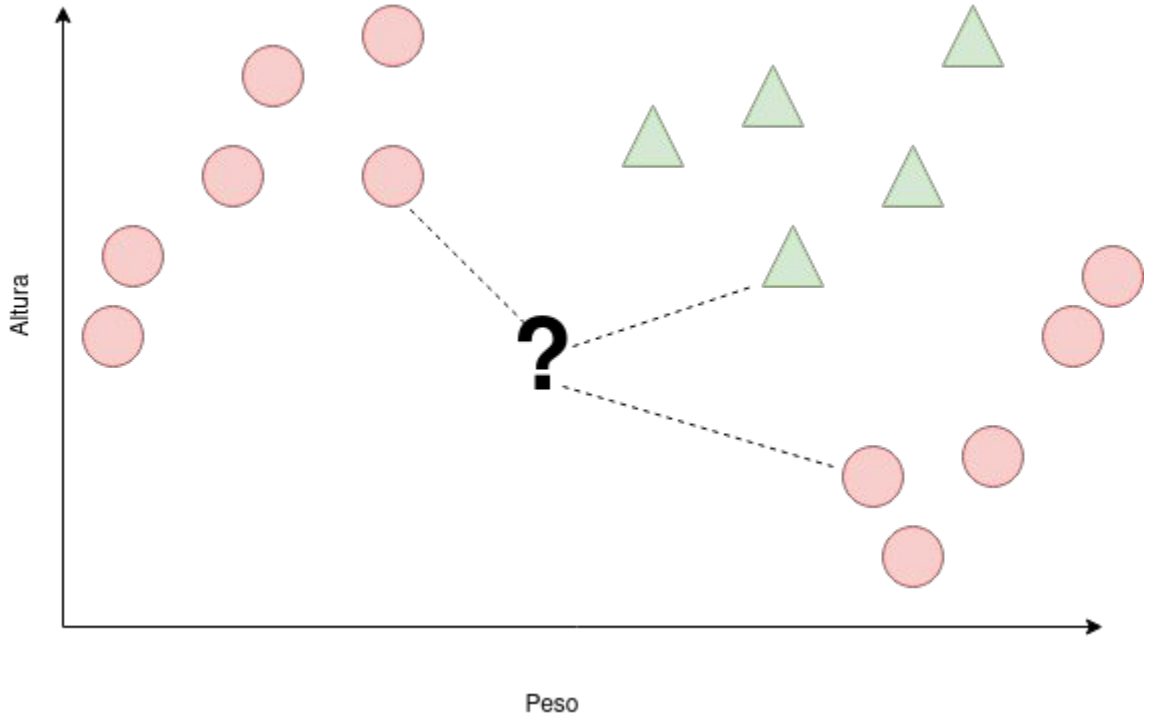
Similaridade



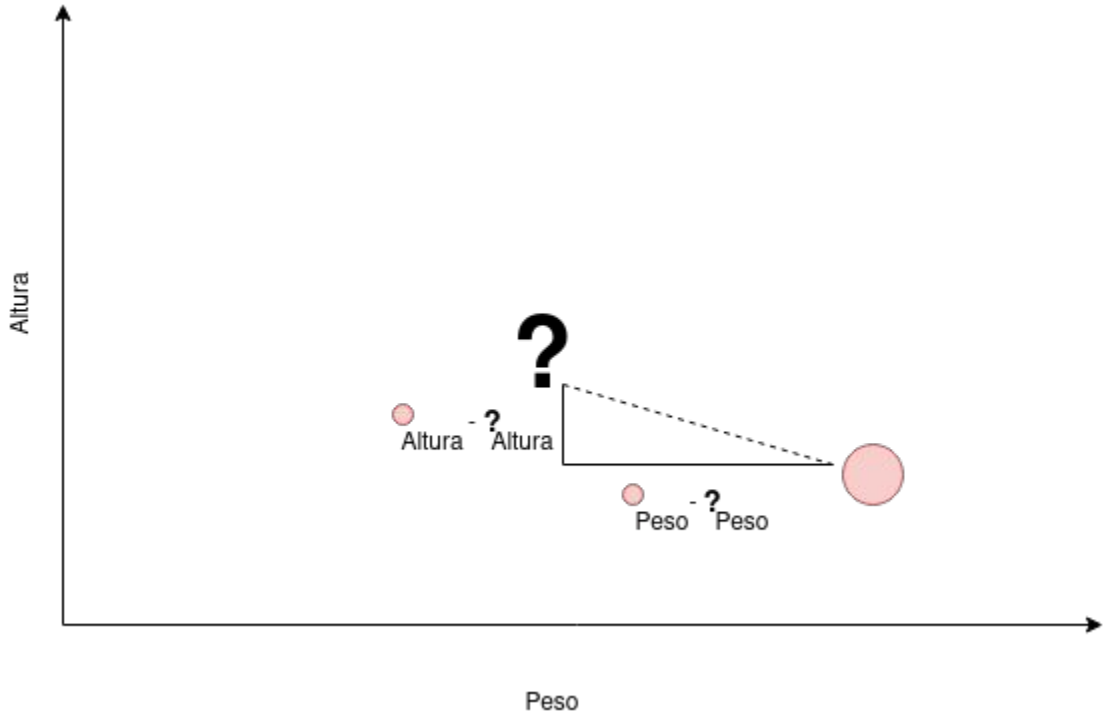
Similaridade



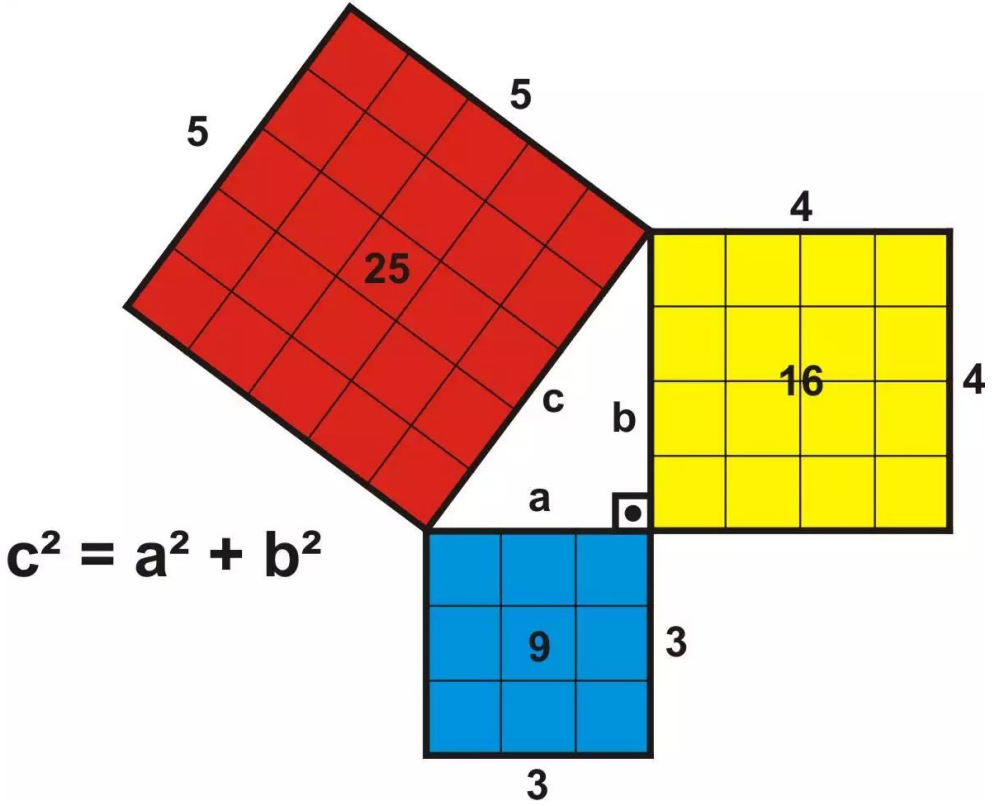
Similaridade



Similaridade



Pitágoras



Fonte: https://www.estudokids.com.br/teorema-de-pitagoras/

Trabalhando com texto: tokenização

Documento 1: O paciente está esperando a consulta.

Tokens: [“O”, “paciente”, “está”, “esperando”, “a”, “consulta”, “.”]

Documento 2: O paciente recebeu o diagnóstico.

Tokens: [“O”, “paciente”, “recebeu”, “o”, “diagnóstico”, “.”]

Trabalhando com texto: Matriz Documentos-Termos

Documento 1: O paciente está esperando a consulta.

Documento 2: O paciente recebeu o diagnóstico.

	o	paciente	está	esperando	a	consulta	recebeu	diagnóstico
Doc 1	1	1	1	1	1	1	0	0
Doc 2	2	1	0	0	0	0	1	1

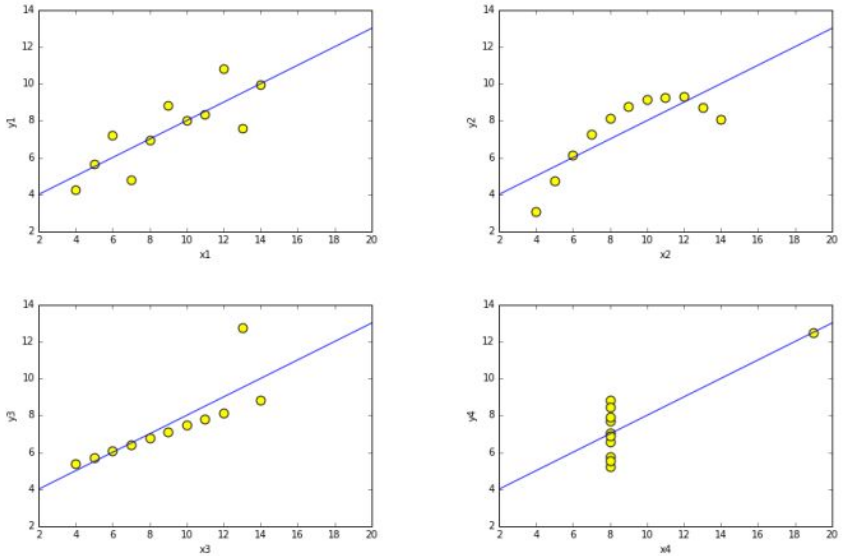
Bag of Words
&
TF-IDF

Visualização

Por quê?

- Exploração/Descoberta de padrões (EDA, rápido/simples)

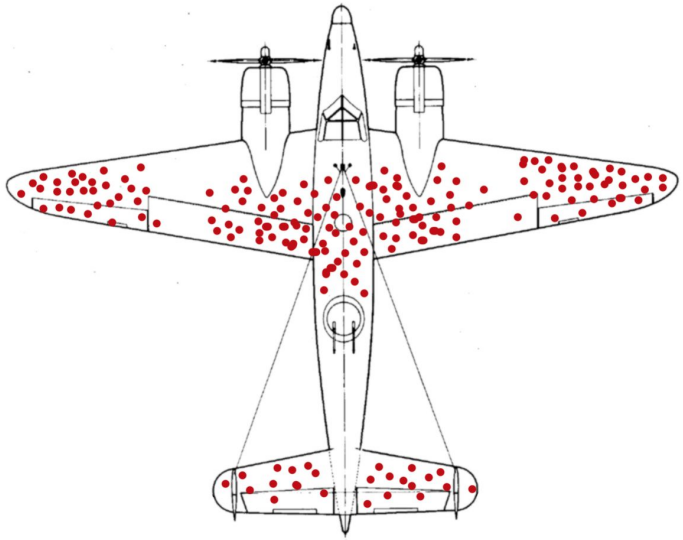
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.31	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Corr.	0.816		0.816		0.816		0.816	



Ascombe's quartlet. Fonte: Skiena, Steven. The Data Science Design Manual, p. 160-161.

Por quê?

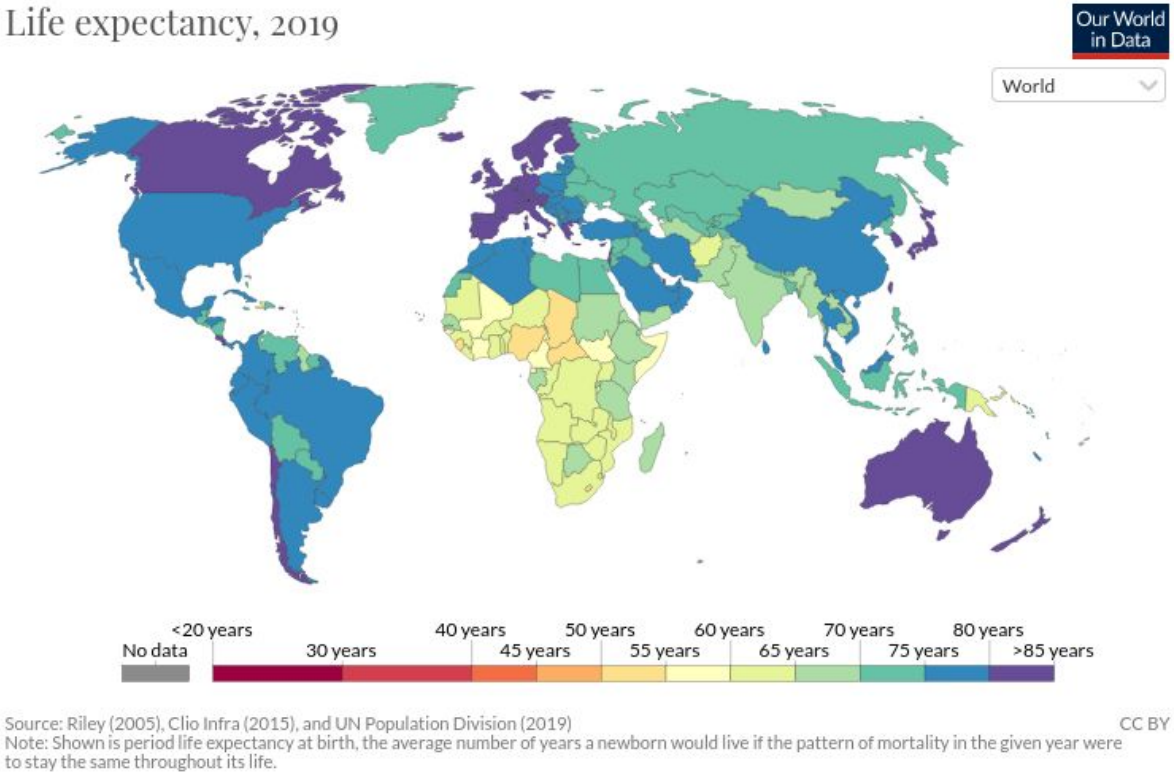
- Exploração/Descoberta de padrões (EDA)



Fonte: Survivorship bias, Wikipedia.

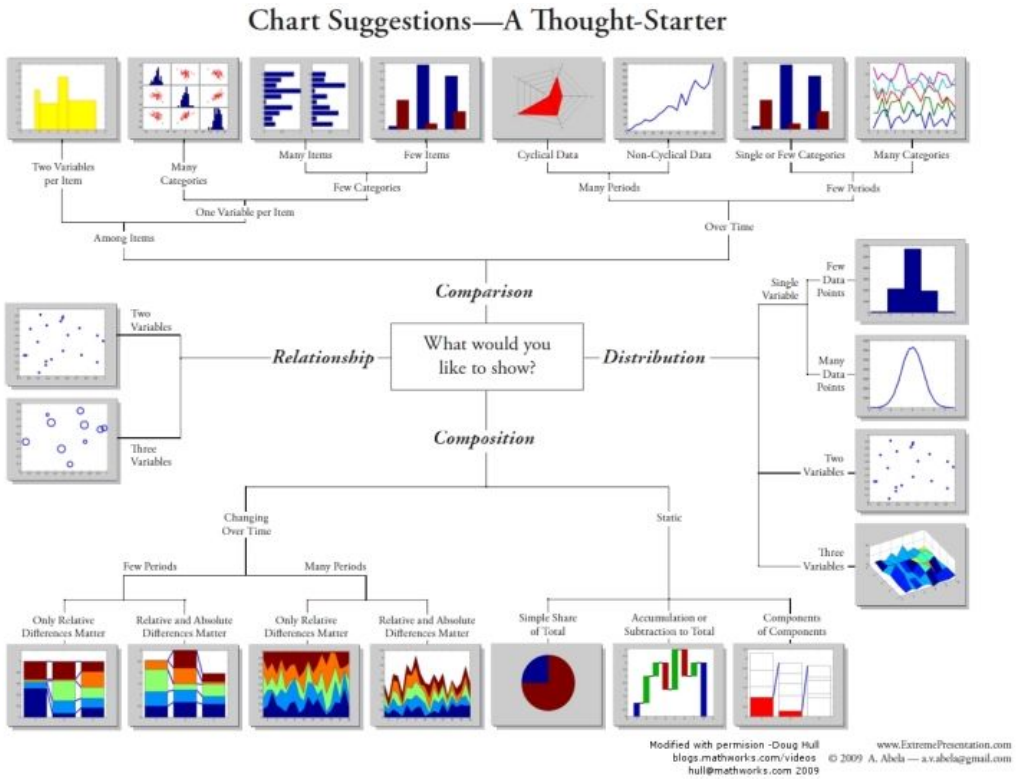
Por quê?

- Comunicação



Fonte: Roser, Max; Ortiz-Ospina, Esteban; Ritchie, Hannah. Life Expectancy. Our World in Data

Qual Gráfico?



https://datavizblog.com/2013/04/29/andrew-abelas-chart-chooser/

Tabelas

Country	Population	Area	Density	Mortality	GDP	Birth Rate
Afghanistan	31,056,997	647,500	47.96	**163.07**	700	36.0
Australia	20,264,082	7,686,850	2.64	4.69	29,000	**100.0**
Burma	47,382,633	678,500	69.83	67.24	1,800	85.3
China	**1,313,973,713**	9,596,960	136.92	24.18	5,000	90.9
Germany	82,422,299	357,021	230.86	4.16	27,600	99.0
Israel	6,352,117	20,770	305.83	7.03	19,800	95.4
Japan	127,463,611	377,835	**337.35**	3.26	28,200	99.0
Mexico	107,449,525	1,972,550	54.47	20.91	9,000	92.2
New Zealand	4,076,140	268,680	15.17	5.85	21,600	99.0
Russia	142,893,540	**17,075,200**	8.37	15.39	8,900	99.6
Tajikistan	7,320,815	143,100	51.16	110.76	1,000	99.4
Tanzania	37,445,392	945,087	39.62	98.54	600	78.2
Tonga	114,689	748	153.33	12.62	2,200	98.5
United Kingdom	60,609,153	244,820	247.57	5.16	27,700	99.0
United States	298,444,215	9,631,420	30.99	6.50	**37,800**	97.0

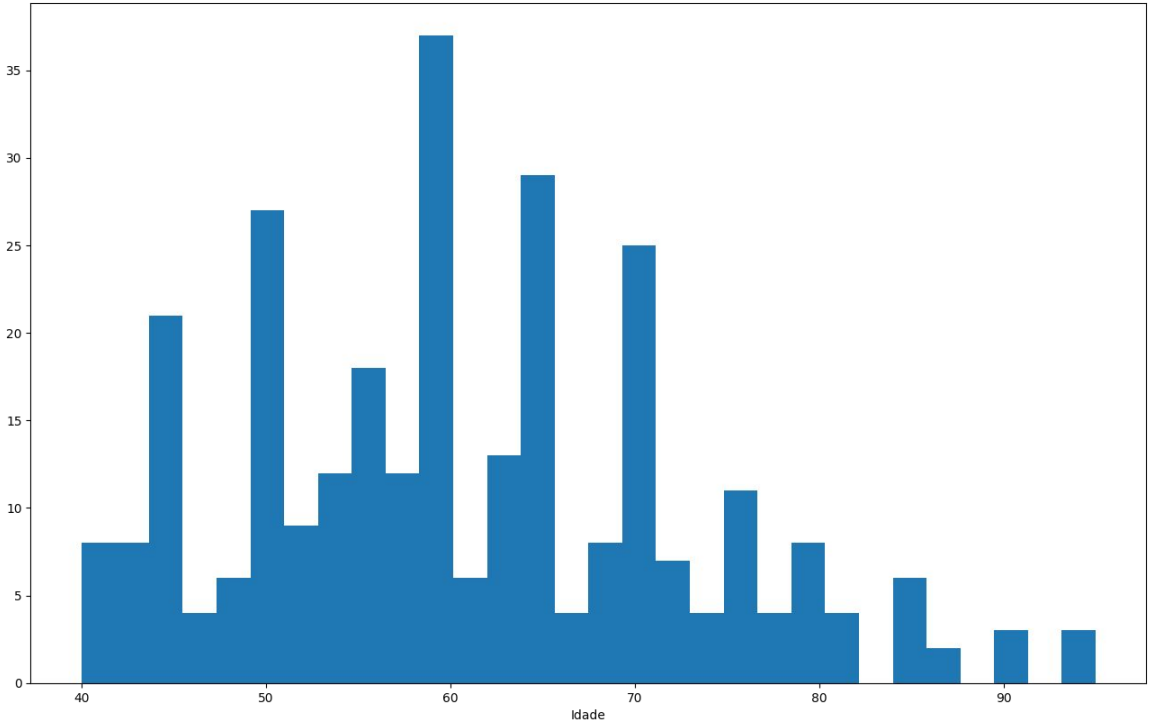
Fonte: Skiena, Steven. The Data Science Design Manual, p. 173.

Tabelas

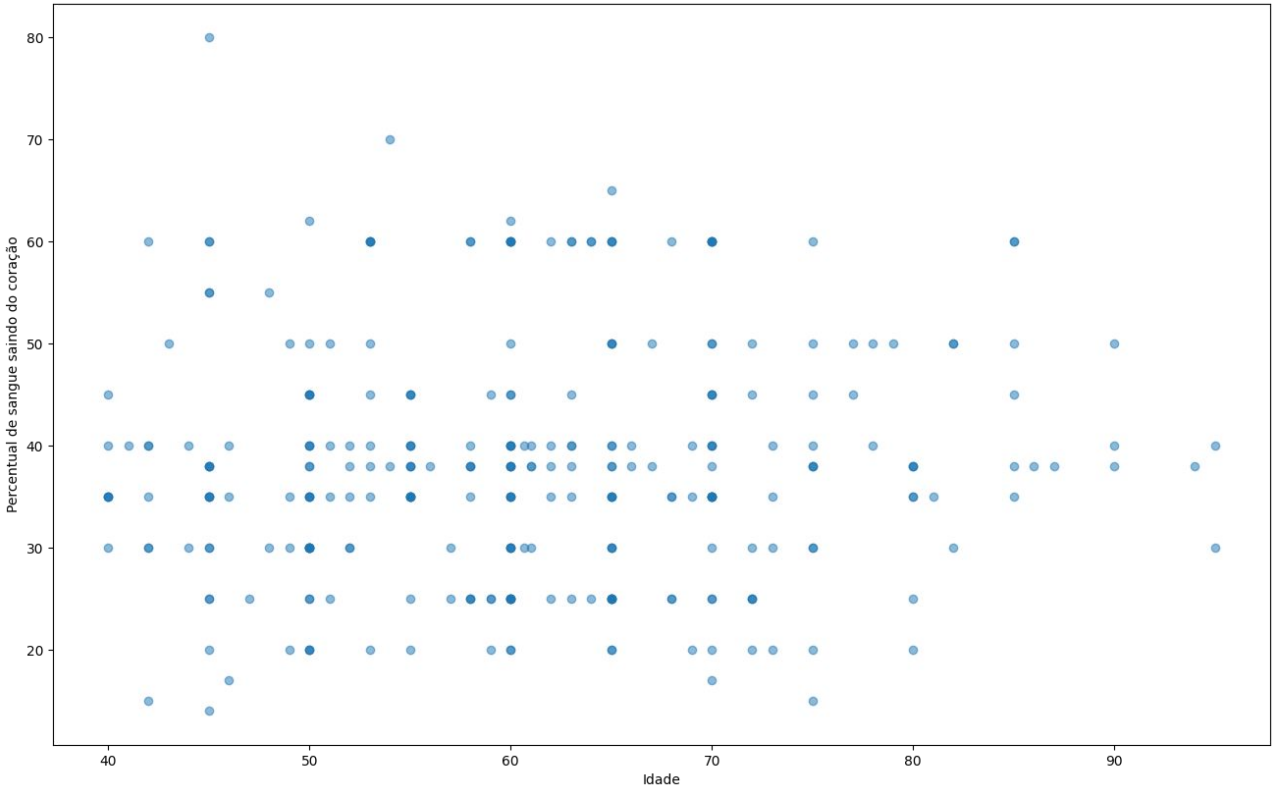
Country	Area	Density	Birthrate	Population	Mortality	GDP
Russia	17075200	8.37	99.6	142893540	15.39	8900.0
Mexico	1972550	54.47	92.2	107449525	20.91	9000.0
Japan	377835	337.35	99.0	127463611	3.26	28200.0
United Kingdom	244820	247.57	99.0	60609153	5.16	27700.0
New Zealand	268680	15.17	99.0	4076140	5.85	21600.0
Afghanistan	647500	47.96	36.0	31056997	163.07	700.0
Israel	20770	305.83	95.4	6352117	7.03	19800.0
United States	9631420	30.99	97.0	298444215	6.5	37800.0
China	9596960	136.92	90.9	1313973713	24.18	5000.0
Tajikistan	143100	51.16	99.4	7320815	110.76	1000.0
Burma	678500	69.83	85.3	47382633	67.24	1800.0
Tanzania	945087	39.62	78.2	37445392	98.54	600.0
Tonga	748	153.33	98.5	114689	12.62	2200.0
Germany	357021	230.86	99.0	82422299	4.16	27600.0
Australia	7686850	2.64	100.0	20264082	4.69	29000.0

Fonte: Skiena, Steven. The Data Science Design Manual, p. 173.

Histograma



Pontos


Percentual de sangue saindo do coração
Idade
80
70
60
50
40
30
20
40
50
60
70
80
90

Dicas

https://eagereyes.org/basics/banking-45-degrees

Dicas

https://eagereyes.org/basics/baselines

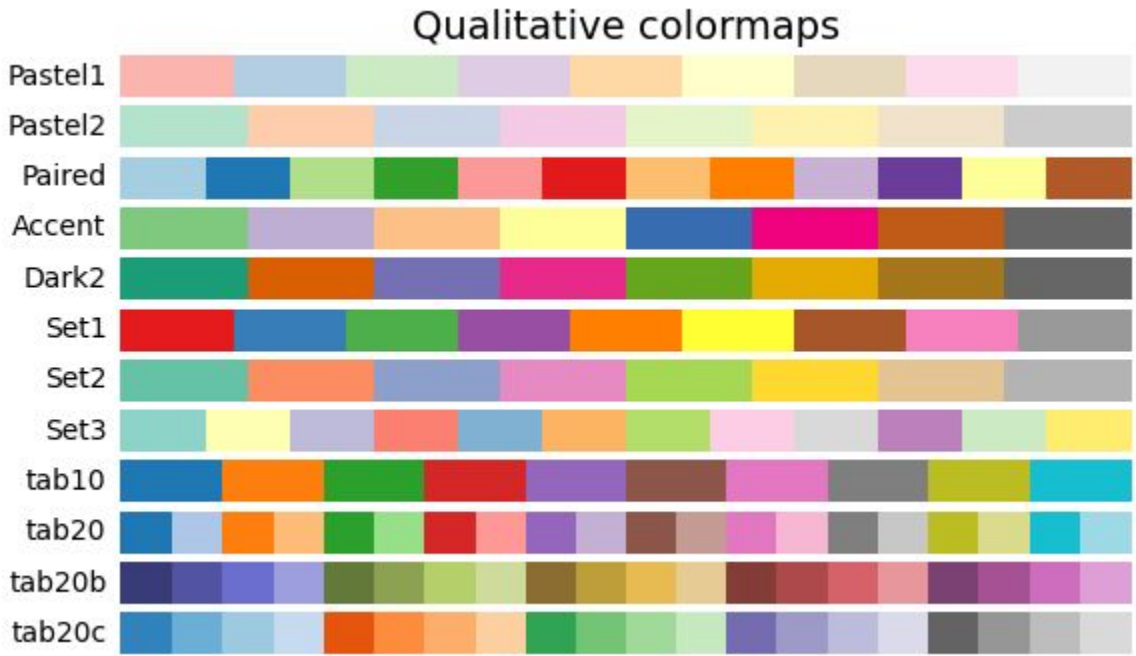
Dicas

- Evite 3D
  - Se necessário, prefira gráficos de pontos com cores/tamanhos
- Evite elementos desnecessários
  - Linhas guias/grades
- Menos é mais: simplicidade, direto ao ponto

Data visualization tips
https://www.geckoboard.com/best-practice/data-visualization-tips/

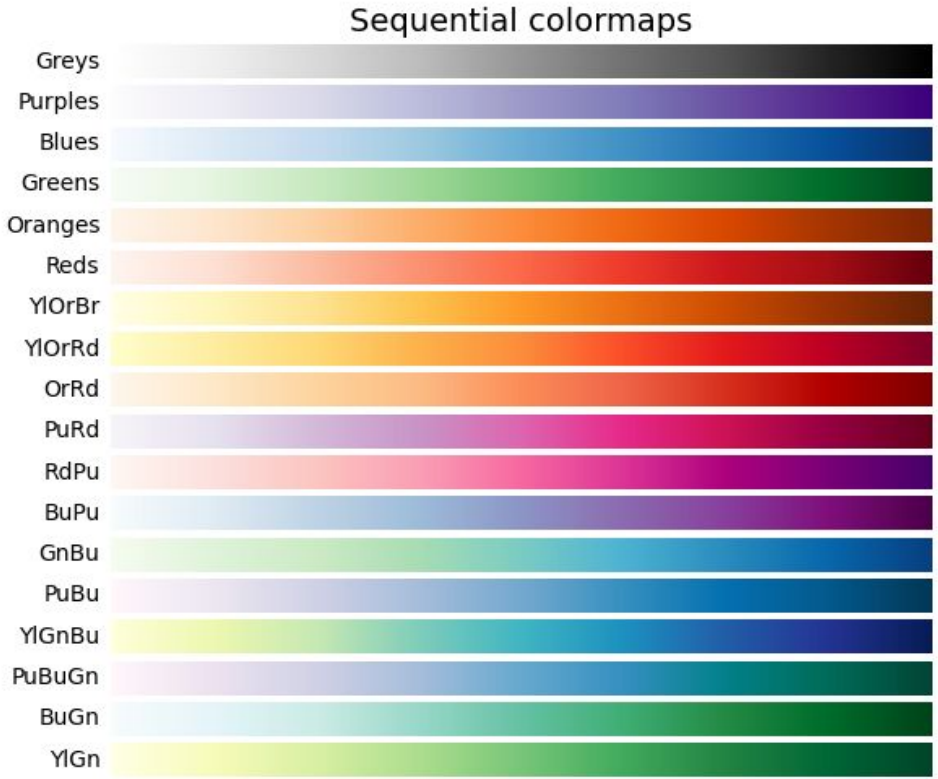
Basics https://eagereyes.org/section/basics

Quais cores usar?



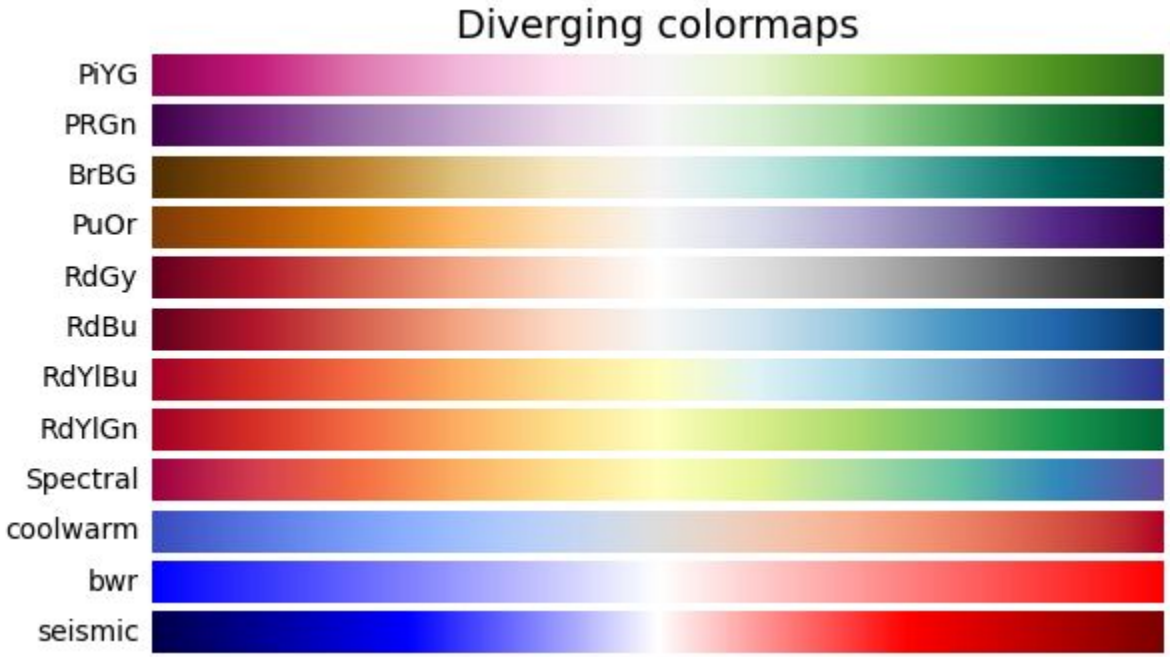
https://matplotlib.org/stable/tutorials/colors/colormaps.html

Quais cores usar?



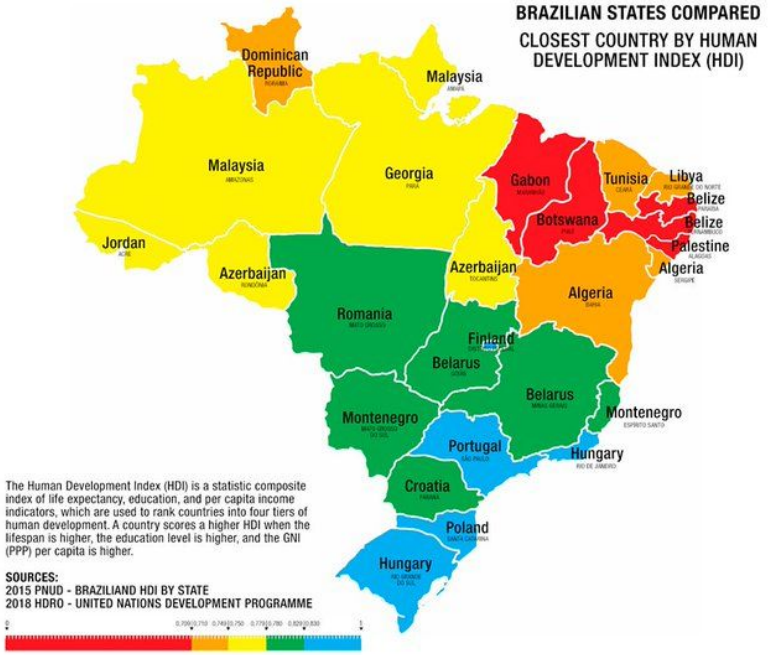
https://matplotlib.org/stable/tutorials/colors/colormaps.html

Quais cores usar?



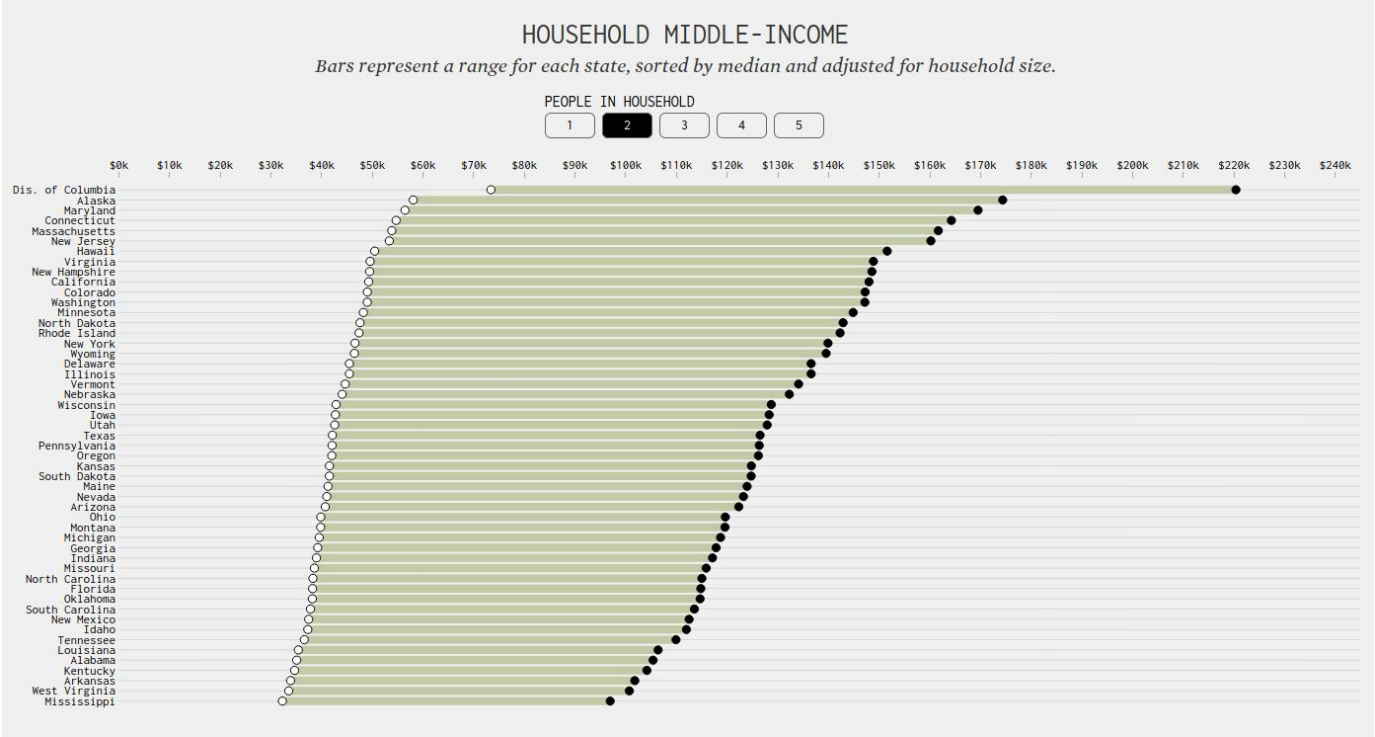
https://matplotlib.org/stable/tutorials/colors/colormaps.html

Alguns exemplos bons



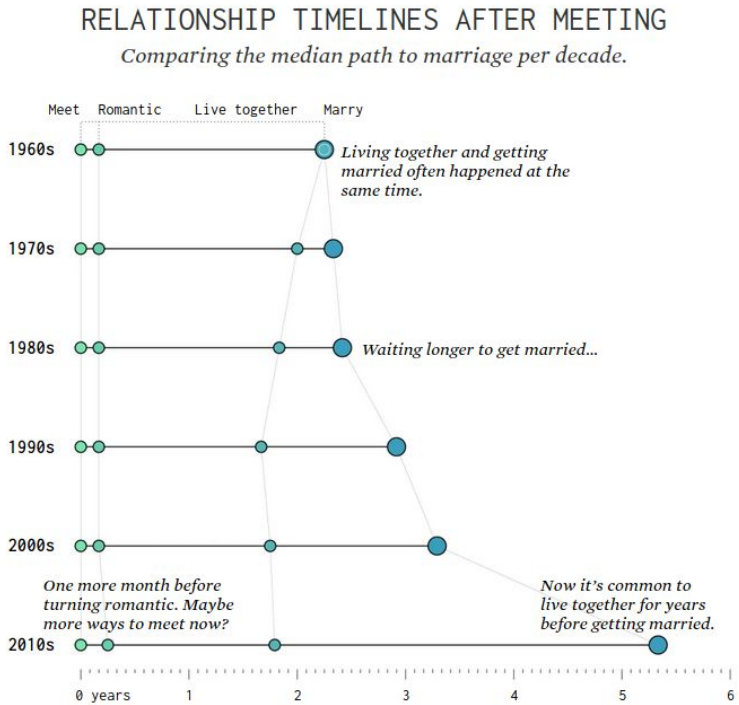
https://twitter.com/nanyhill/status/1211668861738782721

Alguns exemplos bons



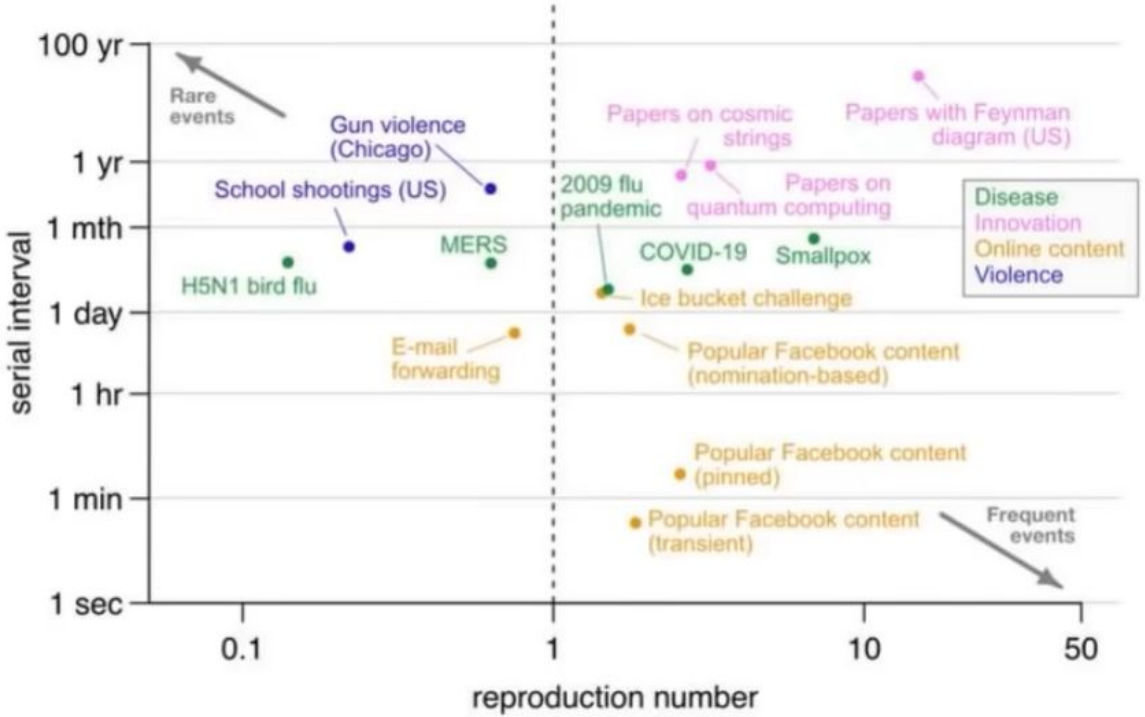
https://flowingdata.com/2019/04/12/middle-income-ranges-in-each-state/

Alguns exemplos bons



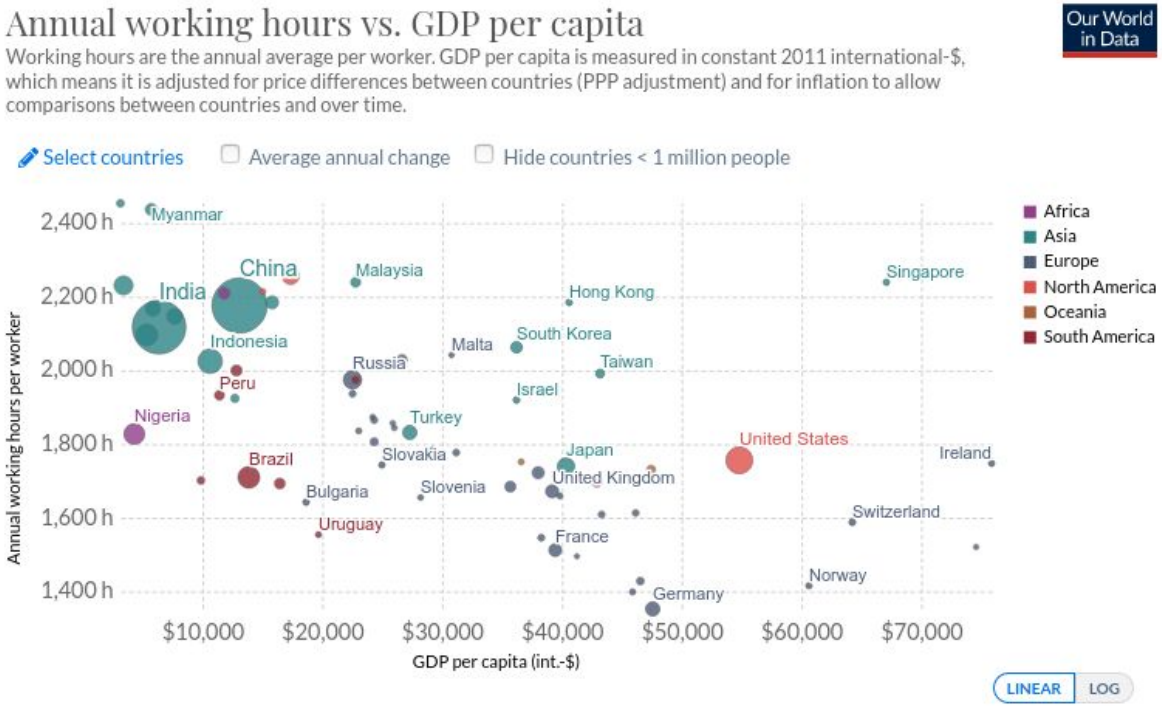
https://flowingdata.com/2019/03/19/the-relationship-timeline-continues-to-stretch/

Alguns exemplos bons



[The Maths of Contagion: Why Things Spread and Why They Stop - with Adam Kucharski, min 23.]

Alguns exemplos bons

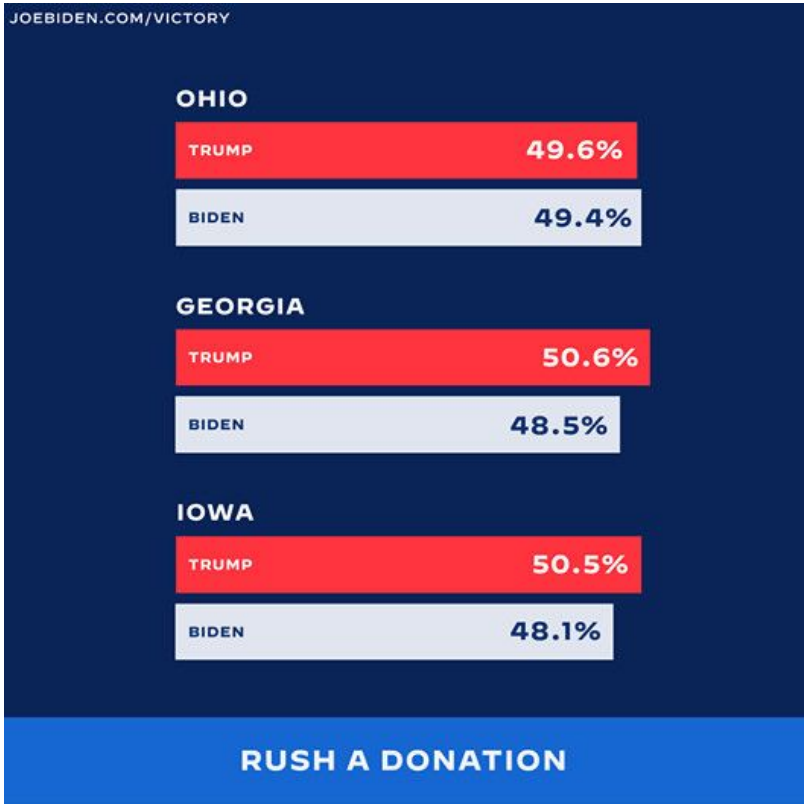


https://ourworldindata.org/working-hours

Alguns exemplos ruins

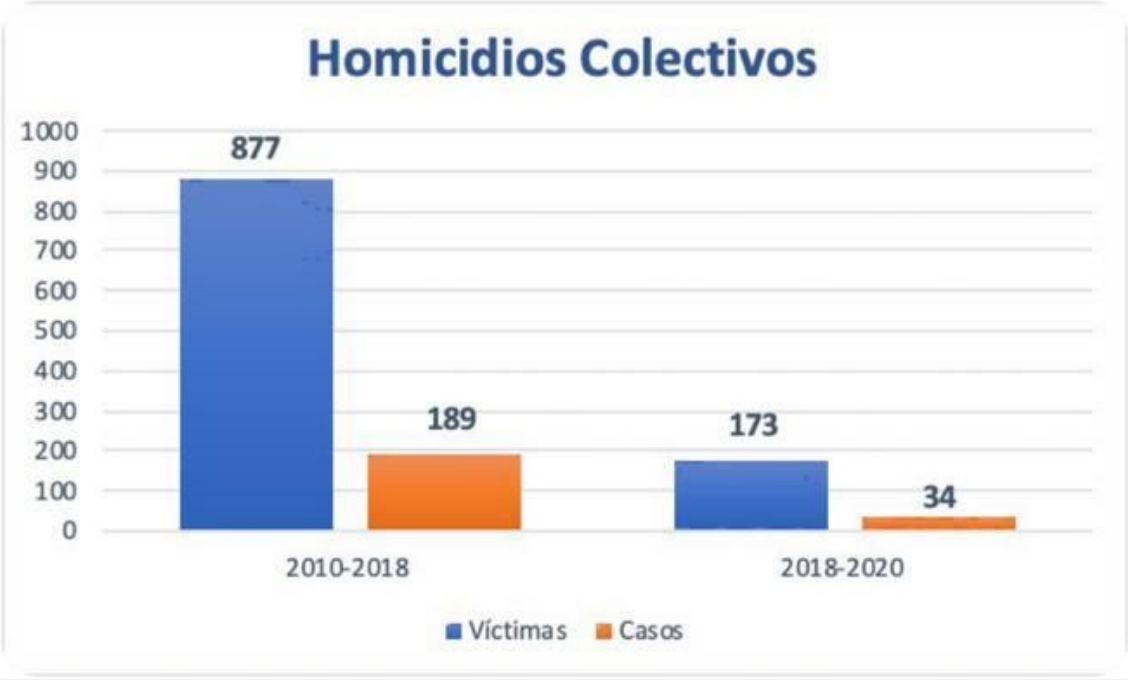
https://twitter.com/Birdbassador/status/1246439050426363906

Alguns exemplos ruins



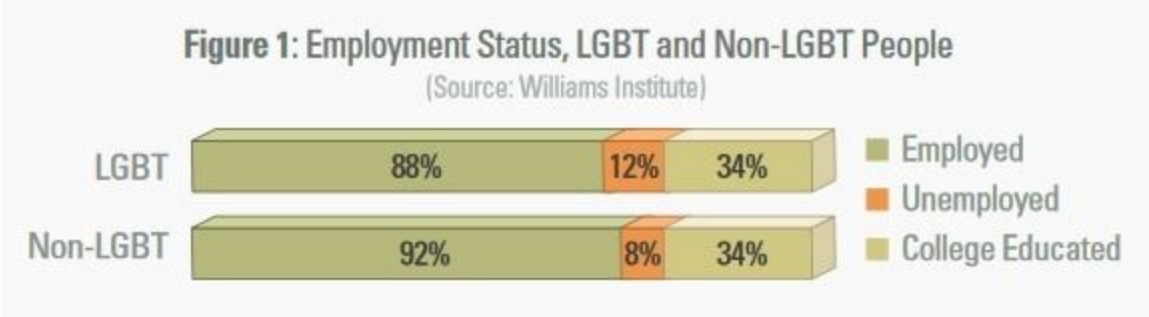
https://viz.wtf/image/631069183407489024

Alguns exemplos ruins



https://viz.wtf/image/627535888462856192

Alguns exemplos ruins



https://viz.wtf/image/182610975453

Alguns exemplos ruins

https://viz.wtf/image/182454529715

Leituras complementares

- https://www.geckoboard.com/best-practice/data-claim-checklist/
- https://www.geckoboard.com/blog/how-to-analyze-data/
- https://www.geckoboard.com/best-practice/data-visualization-tips/
- https://www.geckoboard.com/best-practice/statistical-fallacies/
- https://www.geckoboard.com/best-practice/data-science-glossary/
- Rápido e devagar, por Daniel Kahneman
- O Sinal e o Ruído, por Nate Silver