1. ANÁLISE DE DADOS EXPLORATÓRIA

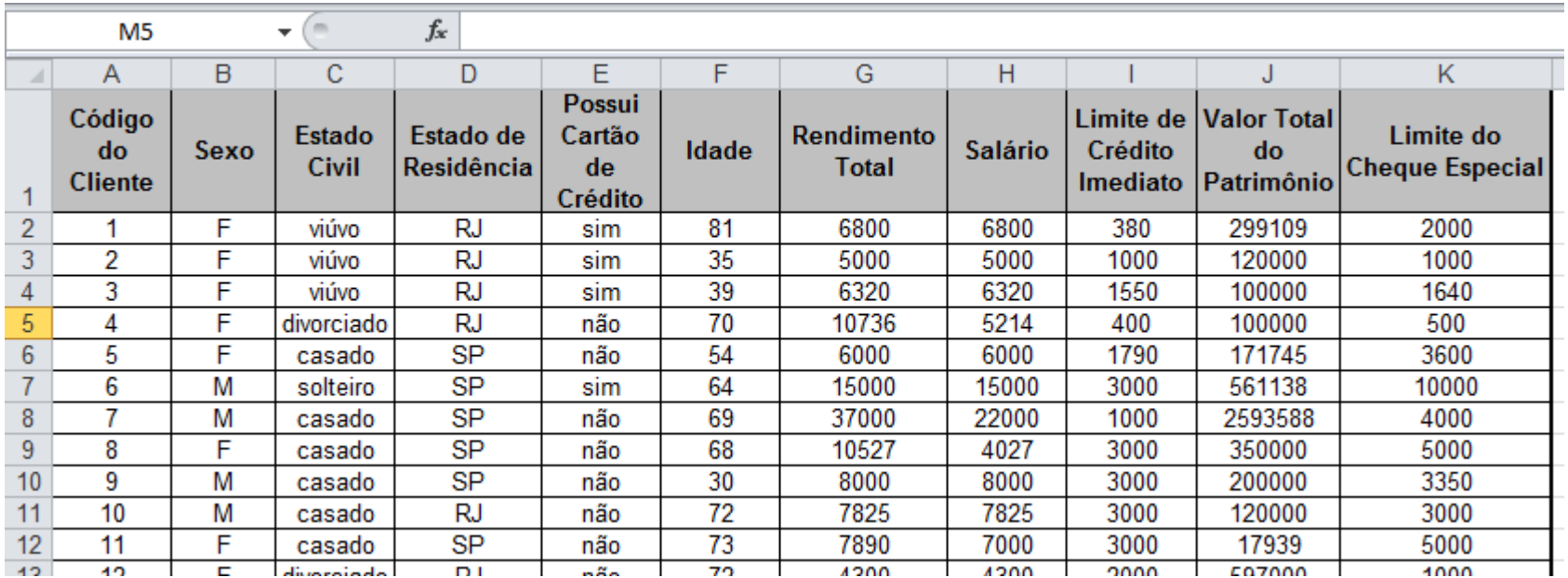
ADE (ANÁLISE DE DADOS EXPLORATÓRIA)

Tem objetivo de realizar análise preliminar do banco de dados por meio de gráficos, tabelas, medidas de posição e de dispersão

Extrair conhecimento dos dados

Bancos de dados

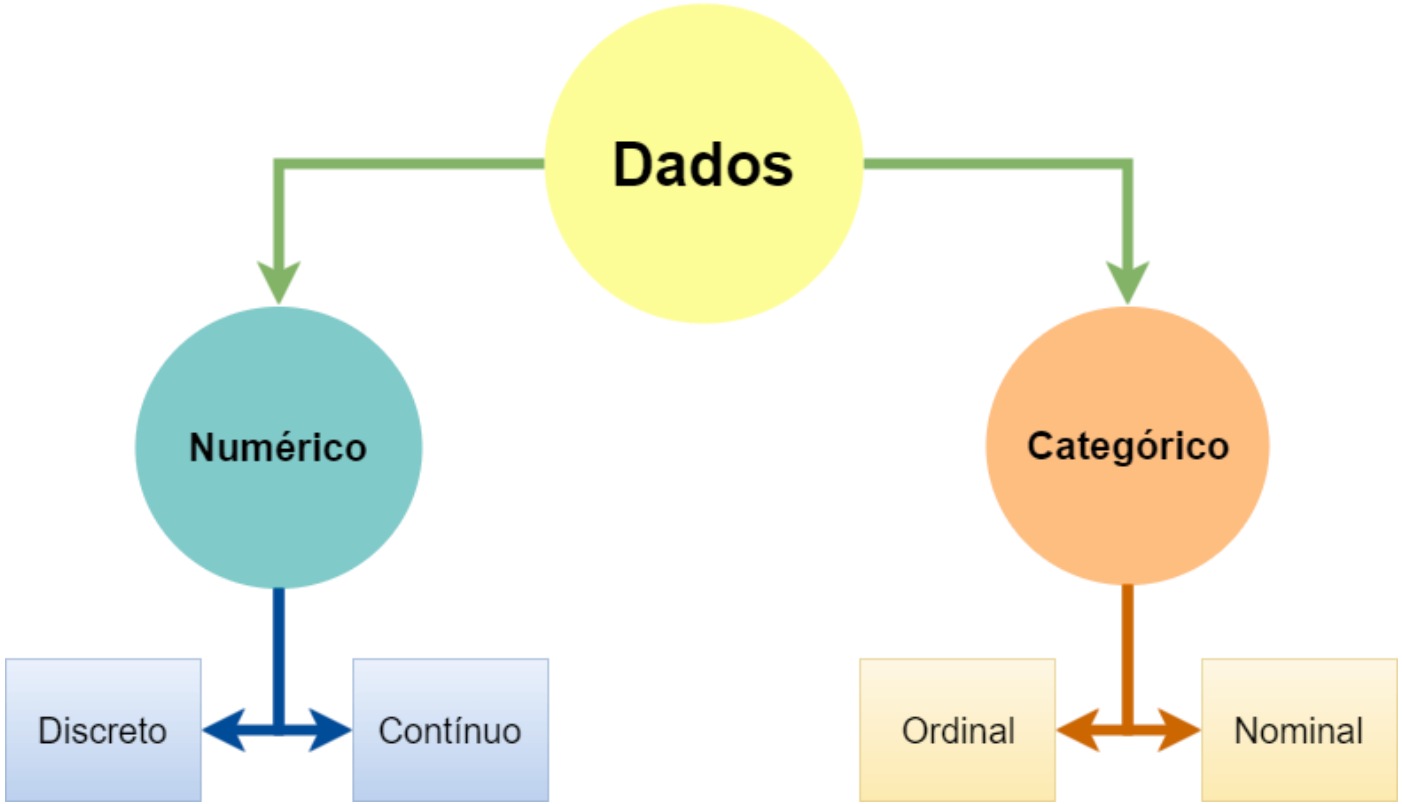
- Dados podem ter variáveis de diferentes tipos



M5

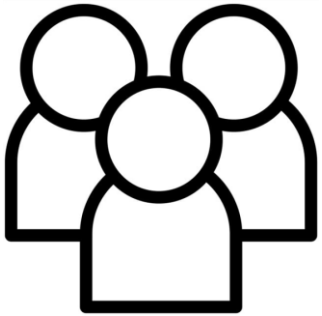
	A	B	C	D	E	F	G	H	I	J	K
1	Código do Cliente	Sexo	Estado Civil	Estado de Residência	Possui Cartão de Crédito	Idade	Rendimento Total	Salário	Limite de Crédito Imediato	Valor Total do Patrimônio	Limite do Cheque Especial
2	1	F	viúvo	RJ	sim	81	6800	6800	380	299109	2000
3	2	F	viúvo	RJ	sim	35	5000	5000	1000	120000	1000
4	3	F	viúvo	RJ	sim	39	6320	6320	1550	100000	1640
5	4	F	divorciado	RJ	não	70	10736	5214	400	100000	500
6	5	F	casado	SP	não	54	6000	6000	1790	171745	3600
7	6	M	solteiro	SP	sim	64	15000	15000	3000	561138	10000
8	7	M	casado	SP	não	69	37000	22000	1000	2593588	4000
9	8	F	casado	SP	não	68	10527	4027	3000	350000	5000
10	9	M	casado	SP	não	30	8000	8000	3000	200000	3350
11	10	M	casado	RJ	não	72	7825	7825	3000	120000	3000
12	11	F	casado	SP	não	73	7890	7000	3000	17939	5000

Tipos de variáveis



Variáveis numéricas (ou quantitativas)

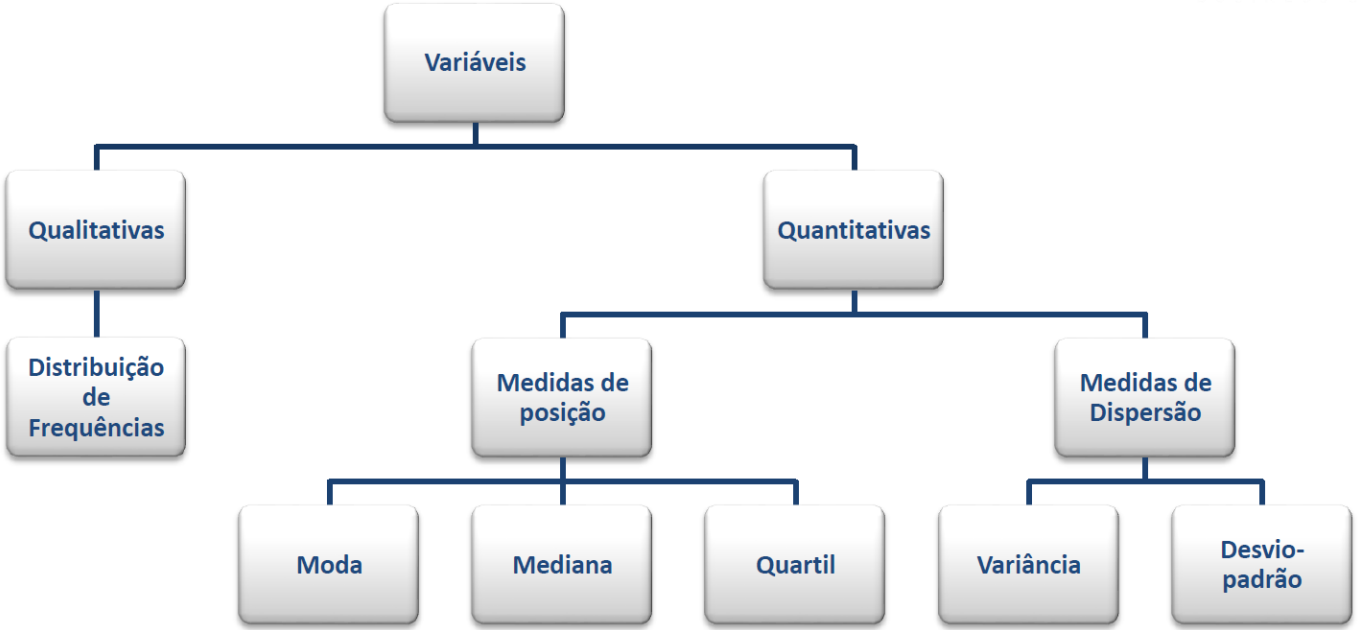
- **Discretos:**
  - Número de filhos
  - Número de dias para pagamento
  - Número de clientes
  - Número de unidades vendidas
- **Contínuas:**
  - Salário
  - Temperatura
  - Faturamento
  - Cotação do dólar



Variáveis categóricas (ou qualitativas)

- **Ordinal:**
  - Escolaridade
  - Tamanho
  - Risco
- **Nominal:**
  - Sexo
  - Estado brasileiro

Dependendo do tipo de variável, analisamos de diferentes maneiras



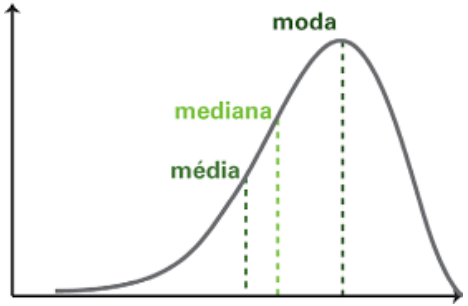
Analisando variáveis categóricas

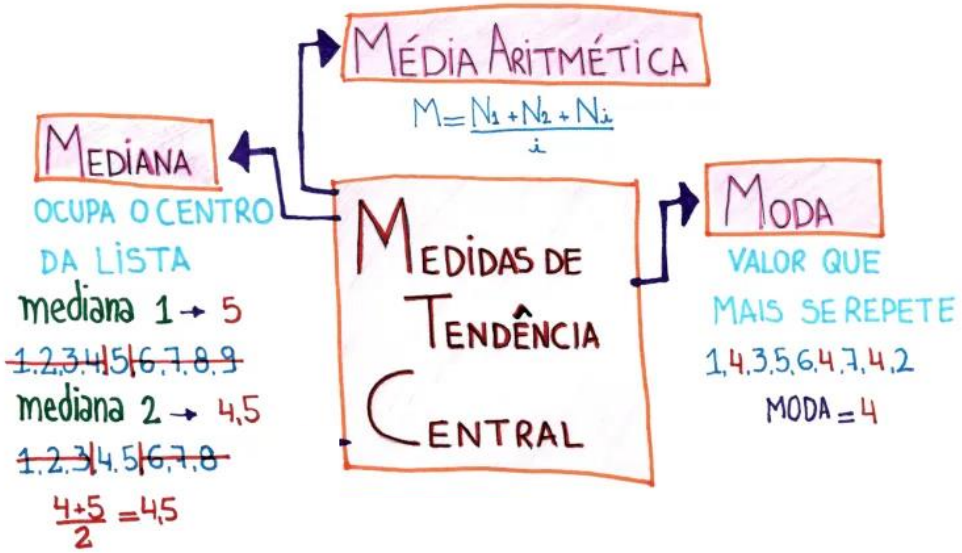
- Tabela de frequência absoluta
- Tabela de frequência relativa

Categoria	Frequência Absoluta	Frequência relativa
Castanhos	10	0,50
Pretos	7	0,35
Azuis	2	0,10
Verdes	1	0,05
Total	20	1.00

Analisando variáveis numéricas

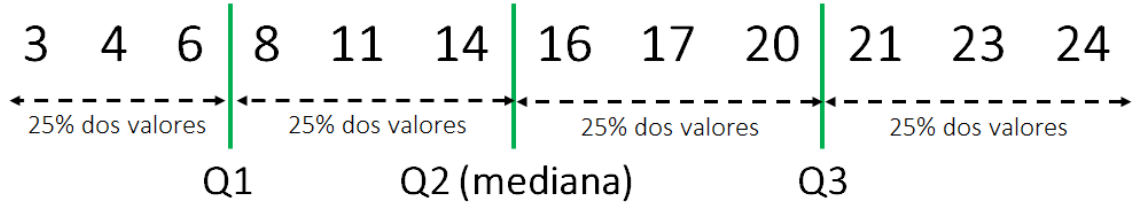
- **Medidas de posição e suas definições**
  - Média
    - Soma de todos elementos dividido pelo número de elementos
  - Moda
    - Valor que mais se repete
  - Mediana
    - Valor central
    - Deve-se ordenar a base de dados para obter a mediana

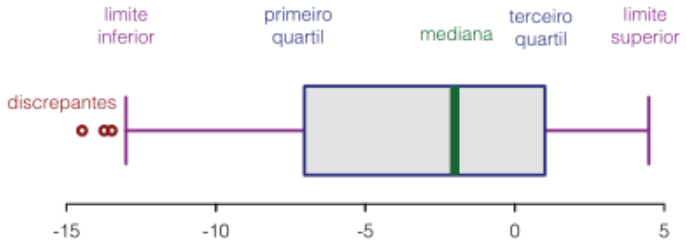




Analisando variáveis numéricas

- **Quartil**
  - Informa limiares que contém 25%, 50% e 75% dos valores
- **Boxplot**
  - Sumariza distribuição da variável
- **Valores discrepantes (*outliers*)**
  - Valores que se destacam da maioria

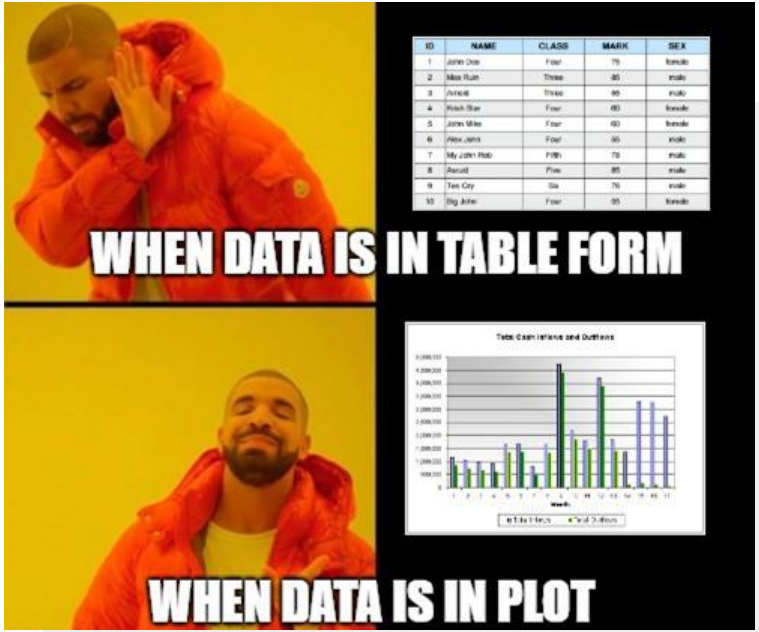




2. VISUALIZAÇÃO DE DADOS

Visualização de dados

- Dados são complexos de serem compreendidos
  - Possuem muitas dimensões
  - Podem não ter relações óbvia entre si
- Ferramentas de visualizações de dados auxiliam no seu entendimento



Visualização de dados é parte arte, parte ciência

- ✔ O desafio de comunicar informações através de gráficos é comum ao dia a dia na maior parte das empresas
- ✔ O maior desafio é passar informação de maneira clara e sem distorções para o público-alvo

Entender o **contexto**

Eliminar **ruídos**

Contar uma **história**

Escolher um **visual efetivo**

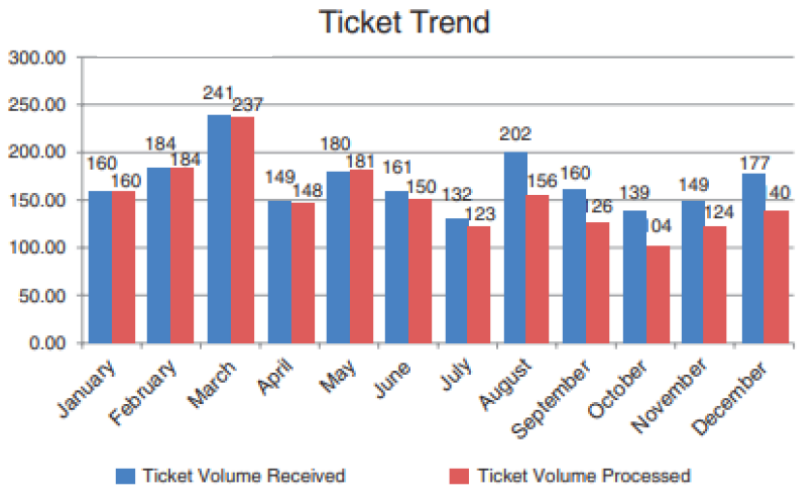
Focar **atenção**

DATAVIZ

Não existe certo ou errado em visualização de dados, mas existem algumas boas práticas

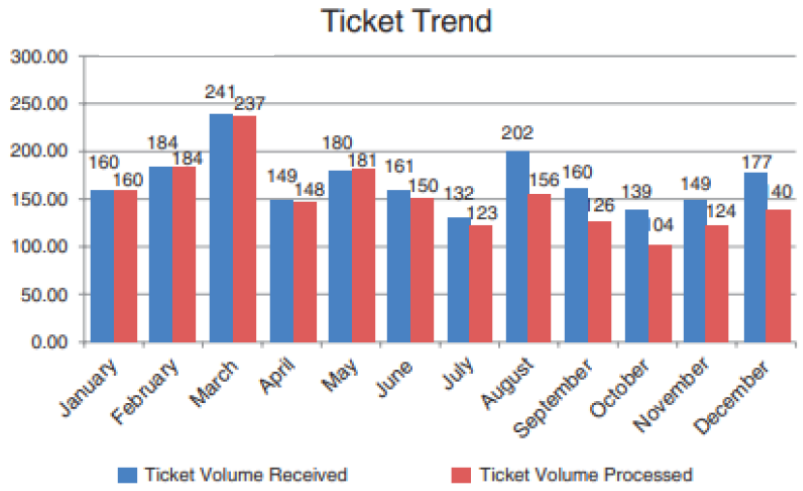
É comum encontrarmos gráficos como esse

- Aparentemente, não parece ter nenhum problema
- Você identifica algo de errado?



É comum encontrarmos gráficos como esse

- Alguns pontos de atenção:
  - Redundância no eixo y (rótulo + grid)
  - O que significam as cores?
  - Qual mensagem estou tentando transmitir?

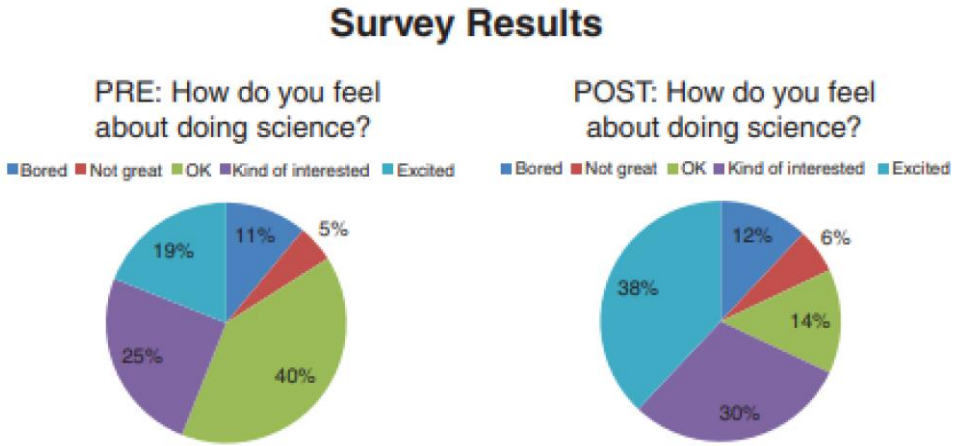


Agora os mesmos dados com outra visualização

- Algumas melhorias:
  - Gráfico de linhas são mais apropriados que barras para mostrar evolução temporal
  - Além disso, elas destacam a diferença de crescimento entre tickets processados e recebidos
  - A linha em maio destaca a diferença entre os 2 grupos
  - Gráfico se torna auto explicativo

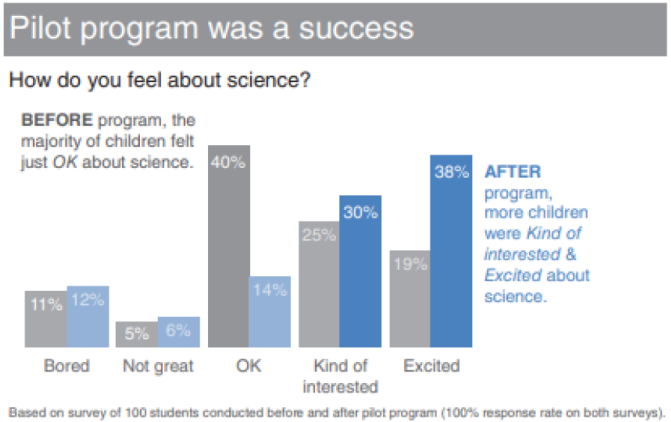
Outro exemplo que pode ser melhorado

- Para exercitar o que pode ser melhorado, podemos nos perguntar:
  - O que eu mudaria nesse gráfico?
  - O que está faltando?
  - Qual mensagem dessa visualização?



E podemos sugerir melhoria do destaque do sucesso do piloto

- Agora sabemos do que se trata esse gráfico
- Se referem a pesquisas realizadas antes e depois de um programa piloto de estímulo à ciência
- As cores foram utilizadas pra enfatizar os resultados positivos
- Textos de apoio são essenciais pra passar uma mensagem coesa

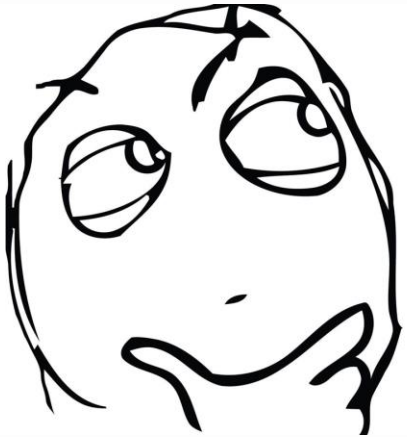


Tá ok, então como passar uma boa mensagem com dataviz?

- Uma boa visualização deve ter como companhia uma boa história
- Ou seja, cada elemento deve ter uma razão de existir
- Alguns guias:

Menos é mais

Cuidado com redundância, excesso de cores e de informações



Quem é a audiência?

Provavelmente o nível de detalhe para um diretor executivo e para um time técnico será diferente

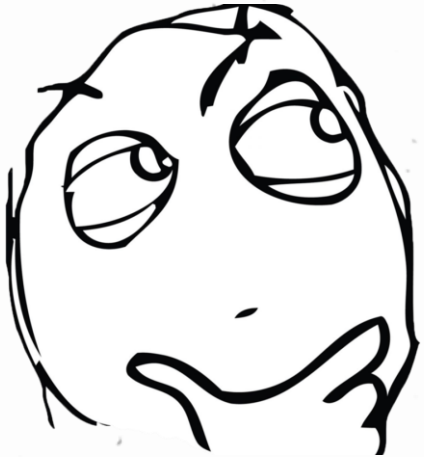
Saiba sua mensagem

Sua visualização deve ter um significado e como os gráficos transmitem a mensagem

ALÉM DISSO, PODEMOS EVITAR ALGUNS TIPOS DE GRÁFICOS

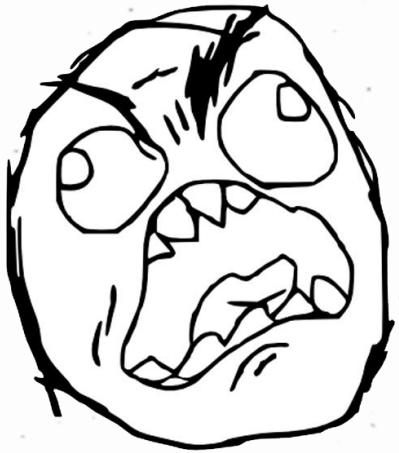
Humanos são naturalmente ruins em interpretar áreas

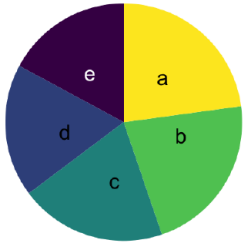
- Olhe o gráfico ao lado e tente ordenar as partes por tamanho



Não está convencido?

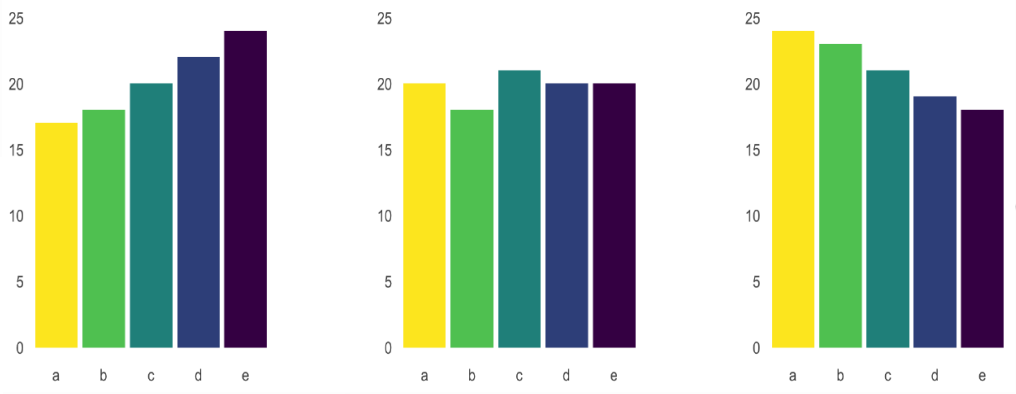
- Agora compare os 3 gráficos e tente entender a evolução dos valores entre eles



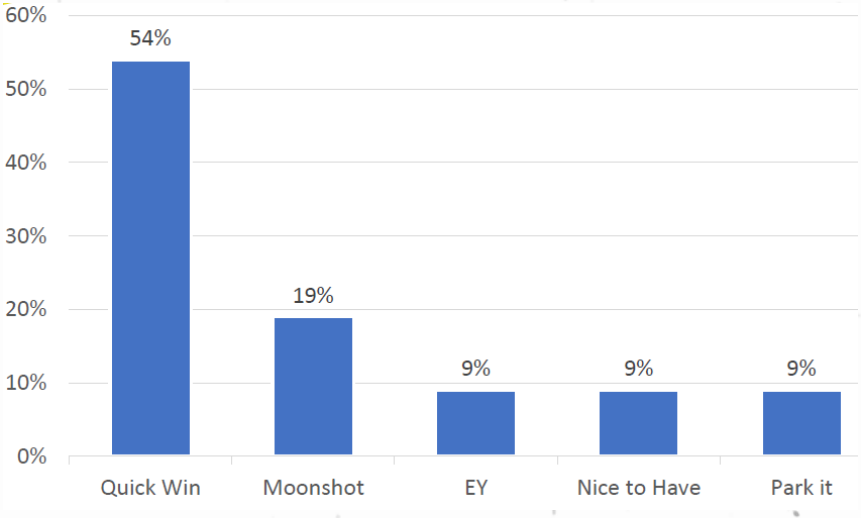


E se trocarmos por um gráfico de barras?

- Agora compará-los ficou mais fácil, né?!

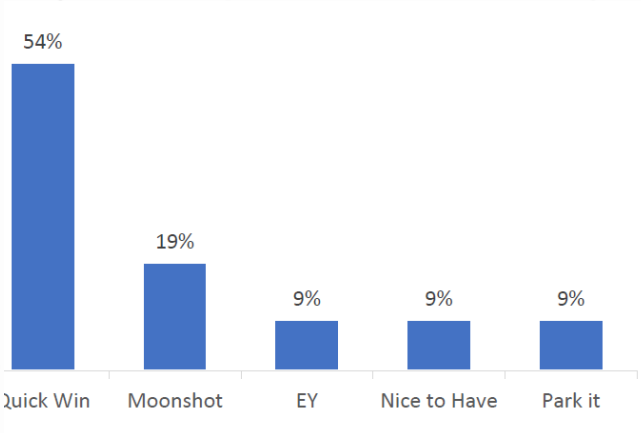


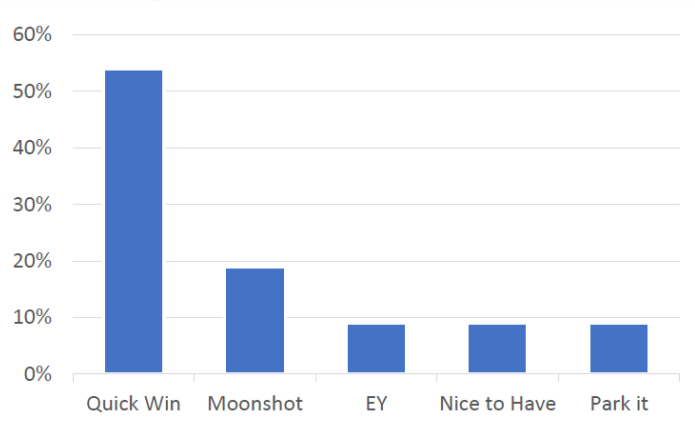
Gráficos de barras permitem comparação direta de alturas



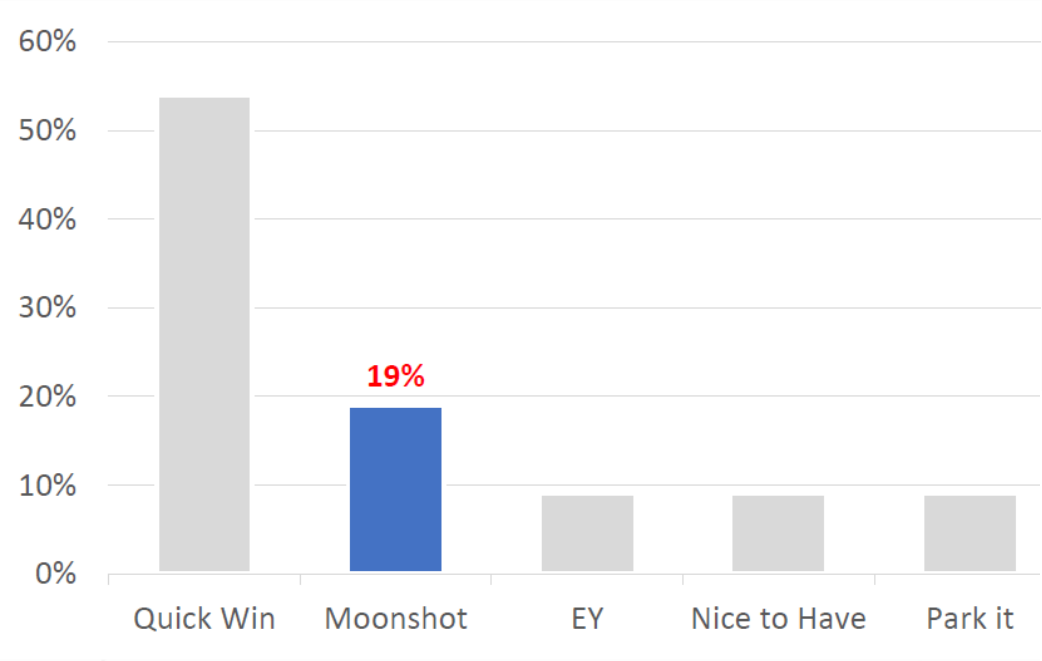
- No entanto, ainda podemos melhorá-lo

Rótulos de dados e linhas de fundo são redundantes. Podemos escolher um só



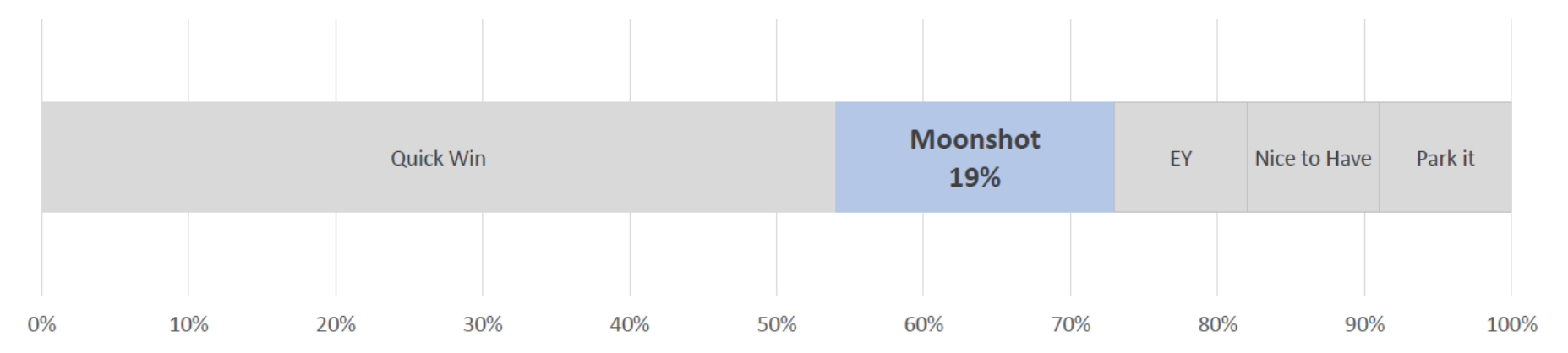


É possível usar cores pra destacar o dado mais importante

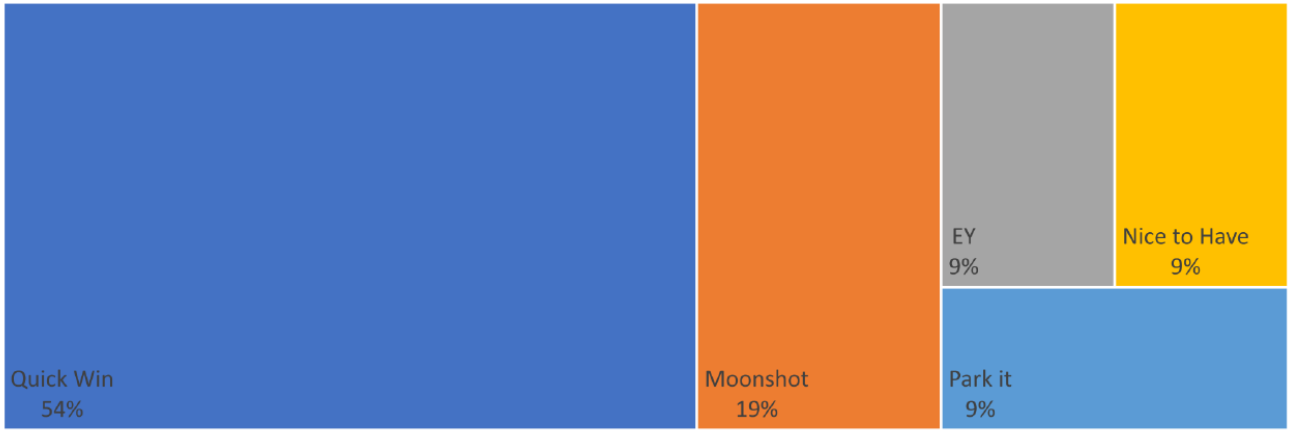


- Porém, gráficos de barra não servem sempre
- Não permite enxergar os dados como **parte de um todo**

Outra opção é usar um gráfico de barras empilhadas

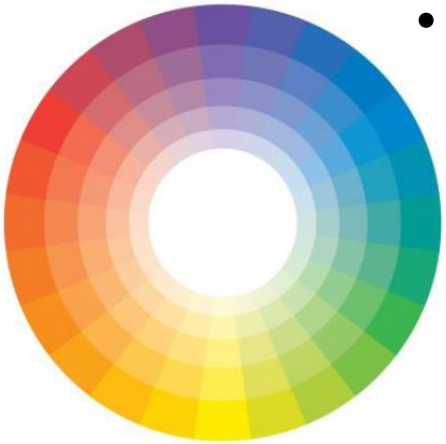


Ou o treemap pra enxergarmos proporções



OK, E SOBRE CORES?

Vamos conceitualizar os tipos de cores/paletas e a finalidade de cada uso

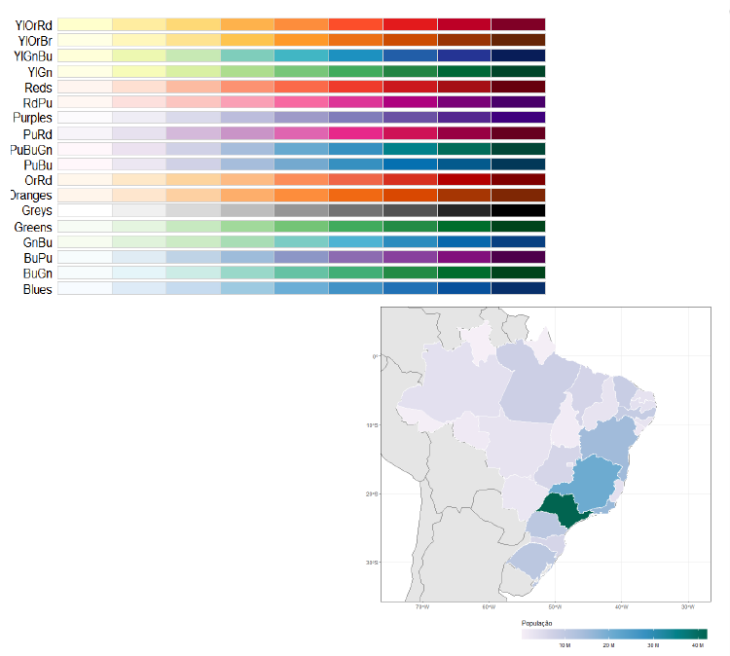


- **Tipos:**
  - Sequencial
  - Divergente
  - Qualitativa

- **Finalidade:**
  - Distinguir
  - Representar
  - Destacar

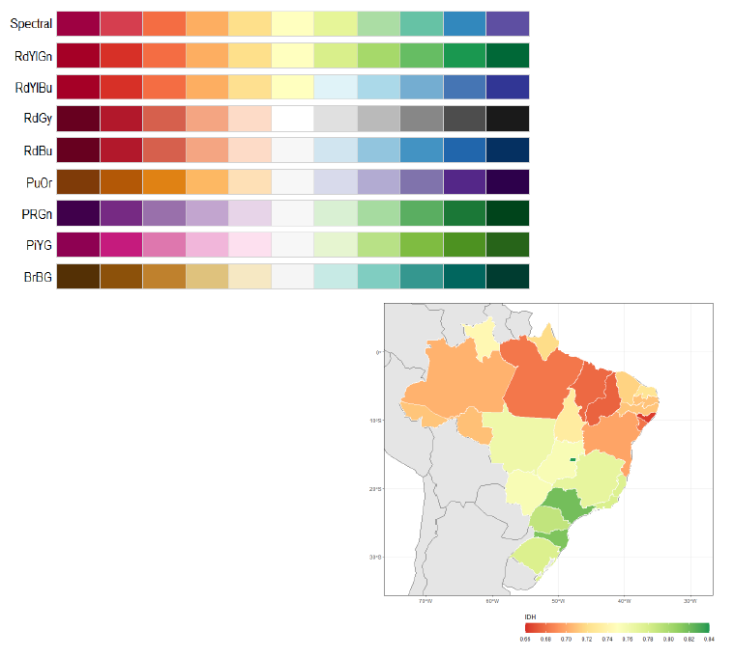
Qual a natureza dos seus dados?

- **Paletas sequenciais:**
  - Dados ordenados que progridem
  - Apenas um dos extremos merece destaque
  - Usar em casos: quanto maior, melhor



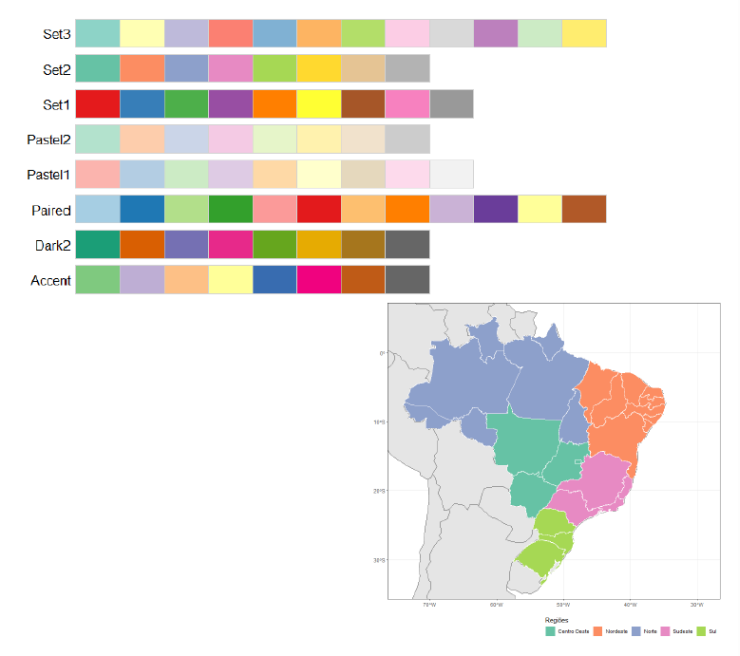
Qual a natureza dos seus dados?

- **Paletas divergentes:**
  - Os 2 extremos merece destaque com mesma ênfase
  - Usar em casos: ruim x bom

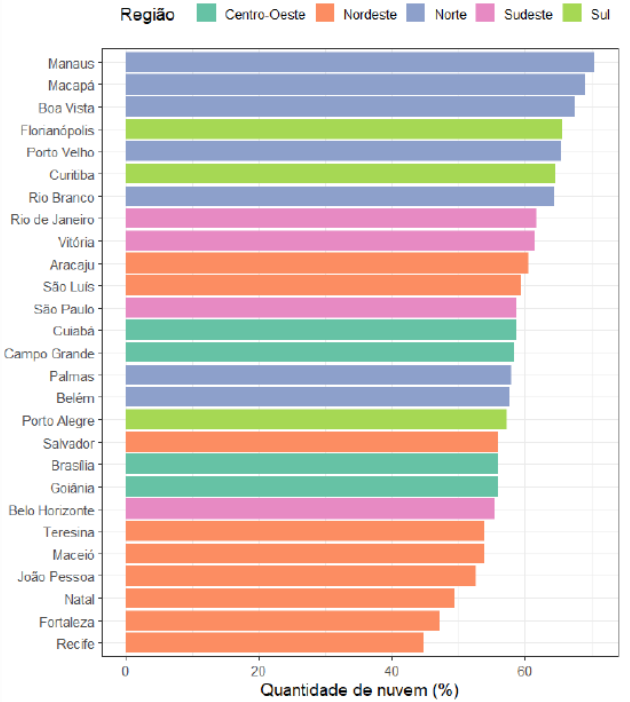


Qual a natureza dos seus dados?

- **Paletas qualitativas:**
  - Não induz ordem nem grandeza
  - Distinção de atributos
  - Usar em casos: n diferentes grupos

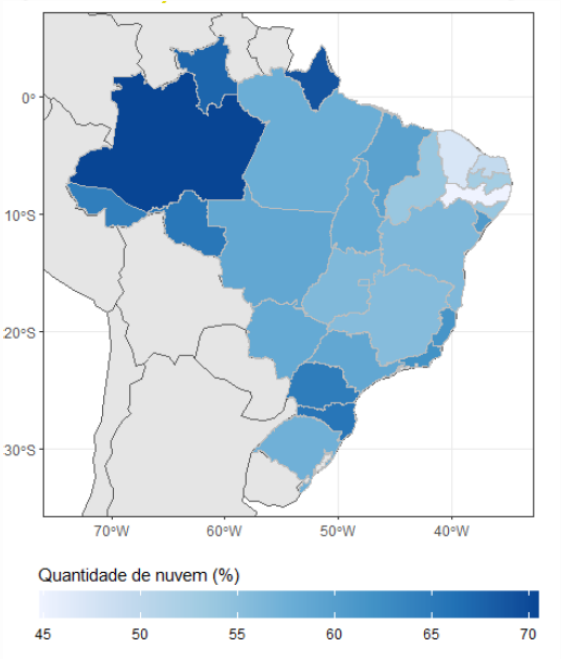


Qual a finalidade do gráfico?



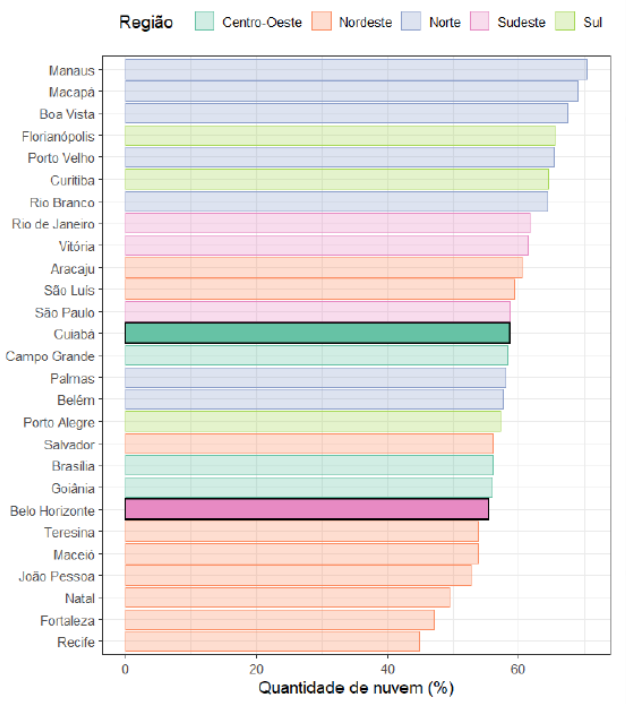
- **Distinguir grupos:**
  - Usar paletas qualitativas
  - Cuidado com a quantidade
  - Grupos podem ter cores temáticas

Qual a finalidade do gráfico?



- **Representar valores:**
  - Usar paletas sequenciais e divergentes
  - Escolher tonalidade de acordo com a intenção do gráfico (ex: azul = bom e vermelho = ruim)
  - Legenda pode ser contínua ou em classes

Qual a finalidade do gráfico?

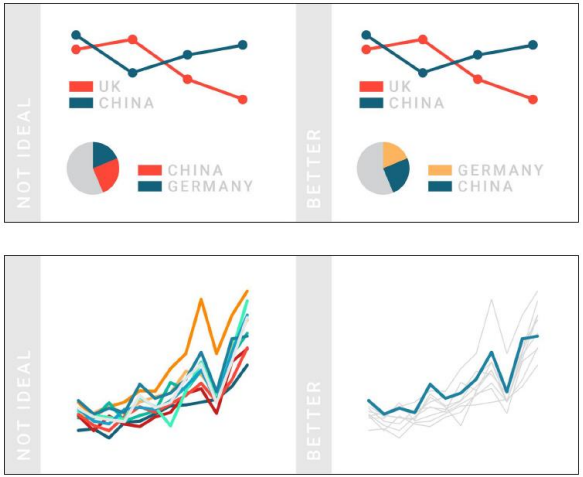


- **Ferramenta para destacar:**
  - Usar paletas qualitativas sobre cores neutras
  - Usar cor destaque sobre transparência ou cores neutras

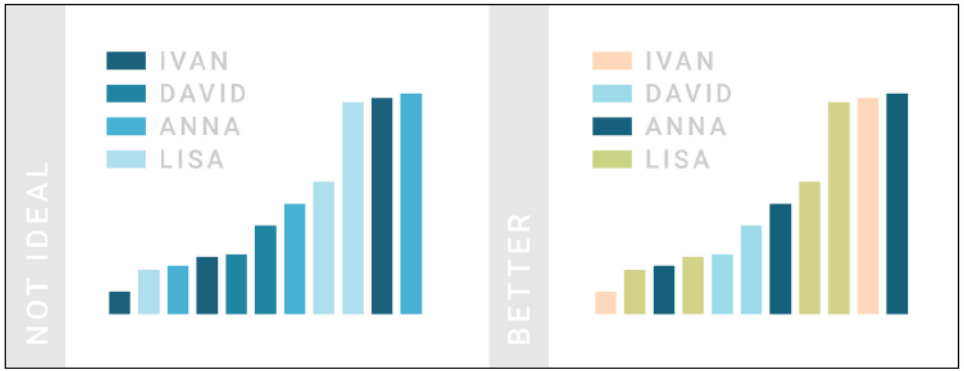
DICAS PRA GUARDAR NO S2

- Avalie se é melhor colocar o foco na escala de cor ou num dos eixos
- Se houver muitos grupos, evite a legenda de cor

- Mantenha a consistência na escolha das cores para as variáveis e grupos
- Menos é mais, lembra? Destaque o que é importante e faça outro gráfico para mostrar o contexto geral, se necessário



- Cores intuitivas ajudam na interpretação e associação
- Não use paletas com gradiente para representar grupos



Seja inclusivo!

- **Lembre-se que o coleguinha pode ser daltônico**
  - 8% dos homens e 0,5% das mulheres são daltônicos

Dê preferência por paletas *colorblind safe*!

- **Evite estas combinações de cores**

Vermelho & verde

Verde & marrom

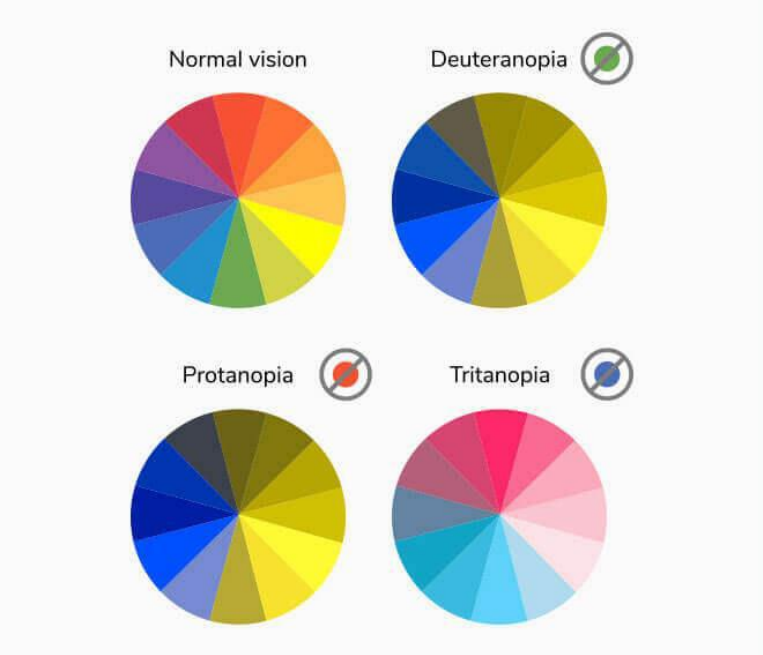
Verde & azul

Azul & cinza

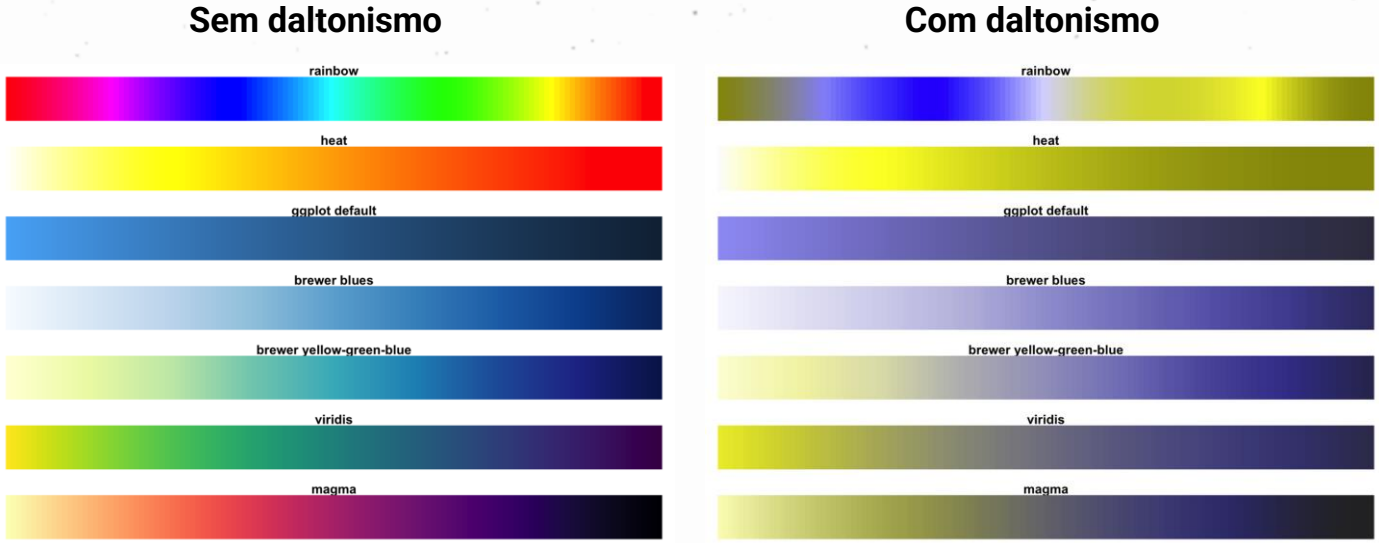
Azul & roxo

Verde & cinza

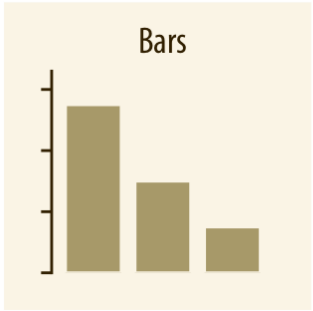
Verde & preto

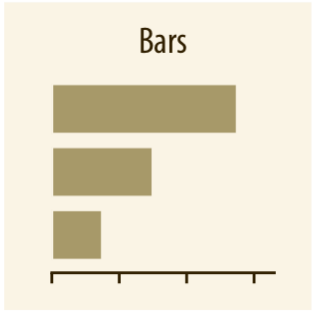


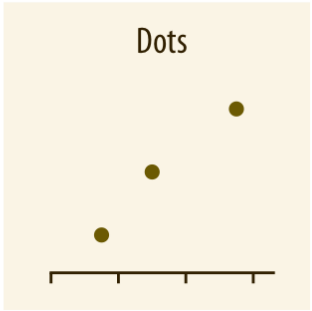
Um exemplo de paleta segura é “viridis”

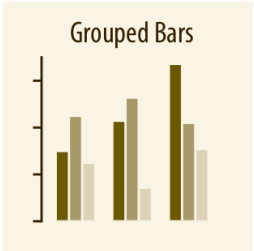


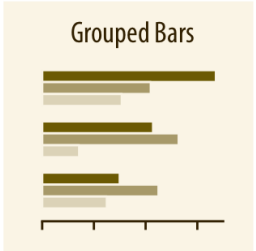
Visualizando quantidades

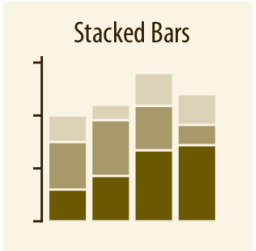


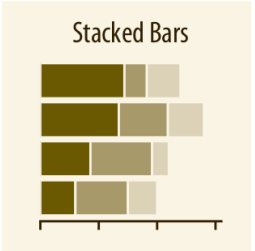








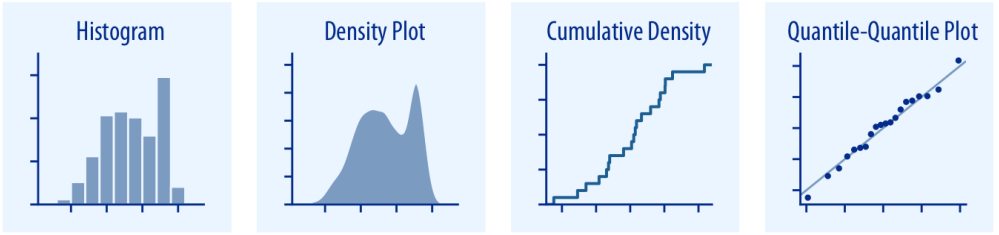




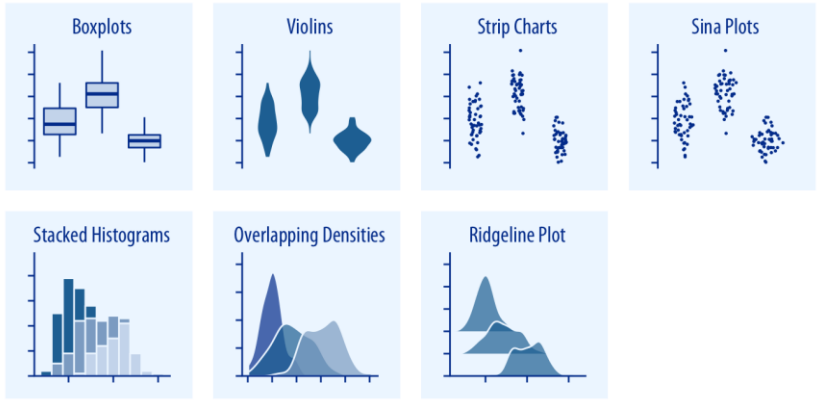
- Barras são a maneira mais comum de representar quantidades
- A direita estão as melhores formas quando existe uma classe

- E essas são mais comuns quando há mais de uma classe

Visualizando distribuições



- Histogramas e gráficos de densidade são mais intuitivos pra visualizar distribuições



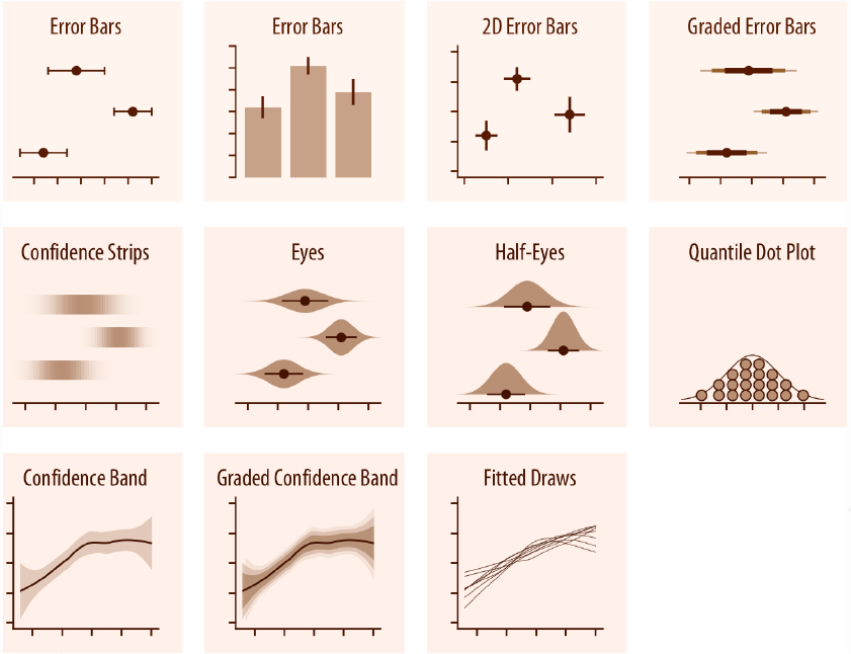
- E à direita são maneiras de visualizar distribuições de várias classes

Visualizando proporções

- Gráficos de barras empilhadas enfatizam partes de um tudo
- Gráficos de barras facilitam a comparação entre partes individuais
- Quando proporções são dadas de acordo com múltiplos agrupamentos, podemos usar gráficos de mosaicos, treemap e conjunto de paralelos

Visualizando incerteza

- Barras de erro indicam a variação esperada de alguma estimativa ou medição
- Para enfatizar a incerteza, podemos avaliar a distribuição das probabilidades
- Para gráficos de linha, o equivalente à barra de erro é o intervalo de confiança



Existem diversas ferramentas de visualizações disponíveis