Introdução à Ciência de Dados

Visualização de Dados

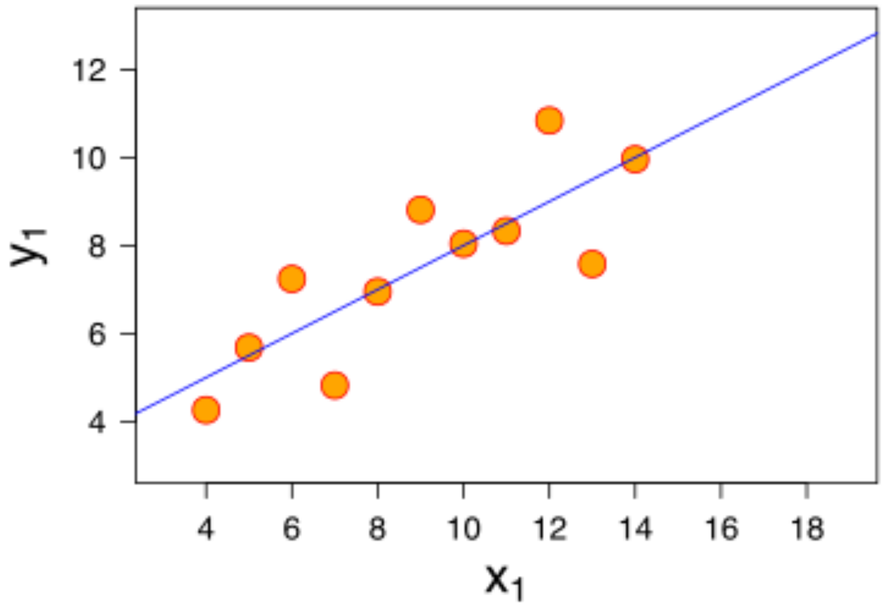
Yuri Malheiros
(yuri@ci.ufpb.br)

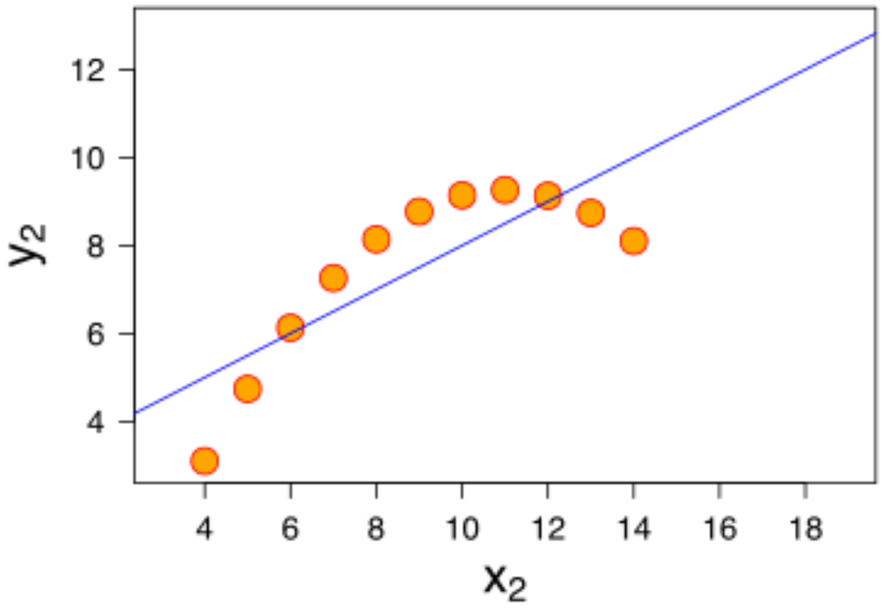
Visualização

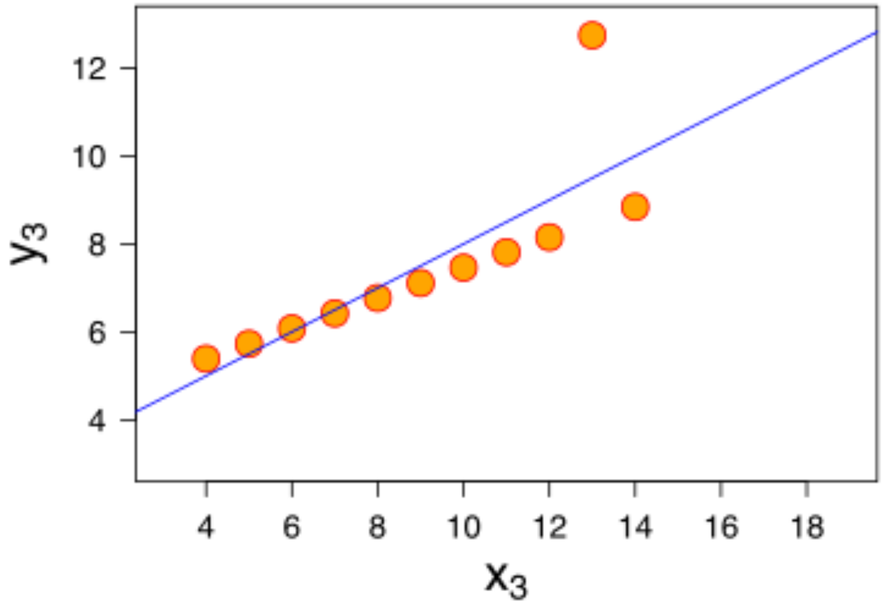
- Tabelas, números, medidas de centralidade e dispersão são úteis, mas eles podem não substituir uma boa visualização
- Vamos analisar a tabela a seguir

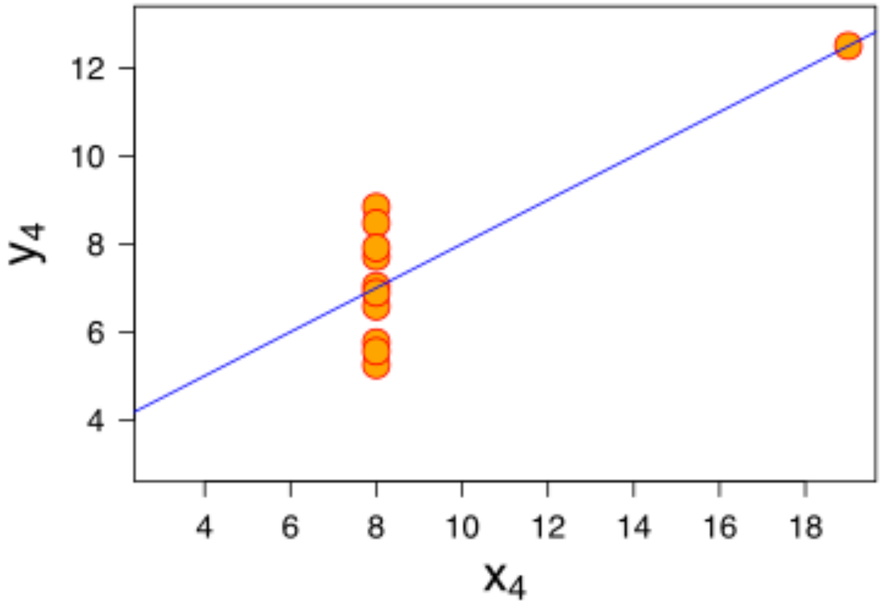
	I		II		III		IV	
	x	**y**	**x**	**y**	**x**	**y**	**x**	**y**
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	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	

- Estes conjuntos de dados são parecidos?


y1
x1
4
6
8
10
12
14
16
18


y2
x2
4
6
8
10
12
14
16
18


y3
x3
4
6
8
10
12
14
16
18


y4
x4
4
6
8
10
12
14
16
18

Visualização

- Existem seis princípios que devemos seguir para termos uma boa visualização
  - Maximizar a razão dados-tinta: a visualização deve mostrar os dados, não outra coisas
  - Minimizar o fator mentira: seus dados devem mostrar a verdade
  - Minimizar lixo: seu gráfico deve ser interessante por causa dos dados, não por causa de feitos gráficos

Visualização

- Existem seis princípios que devemos seguir para termos uma boa visualização
  - Usar escalas adequadas e rotular os dados claramente
  - Usar cores de forma efetiva
  - Explorar o poder da repetição para comparações

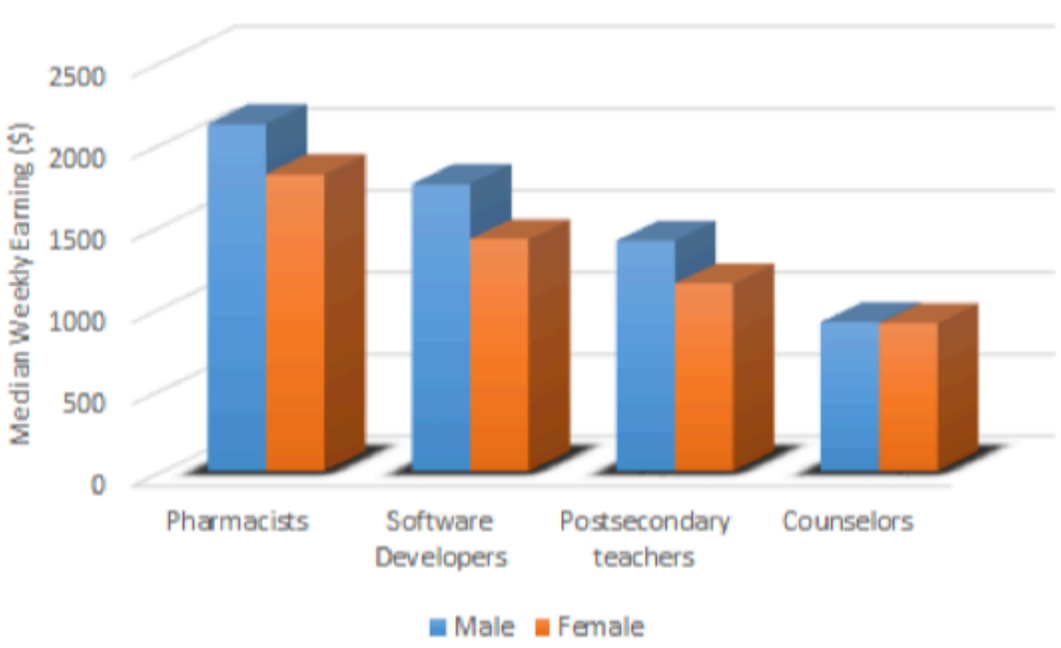
Visualização

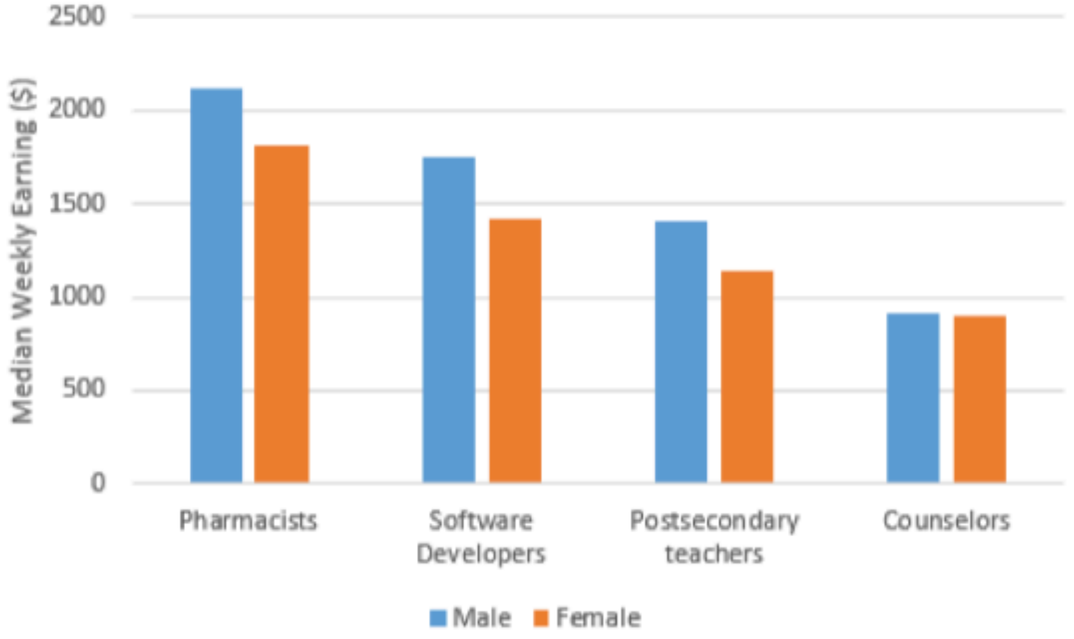
- Maximizar a razão dados-tinta

$$\text{Razão dados-tinta} = \frac{\text{Tinta usada para os dados}}{\text{Total de tinta usada}}$$

Visualização

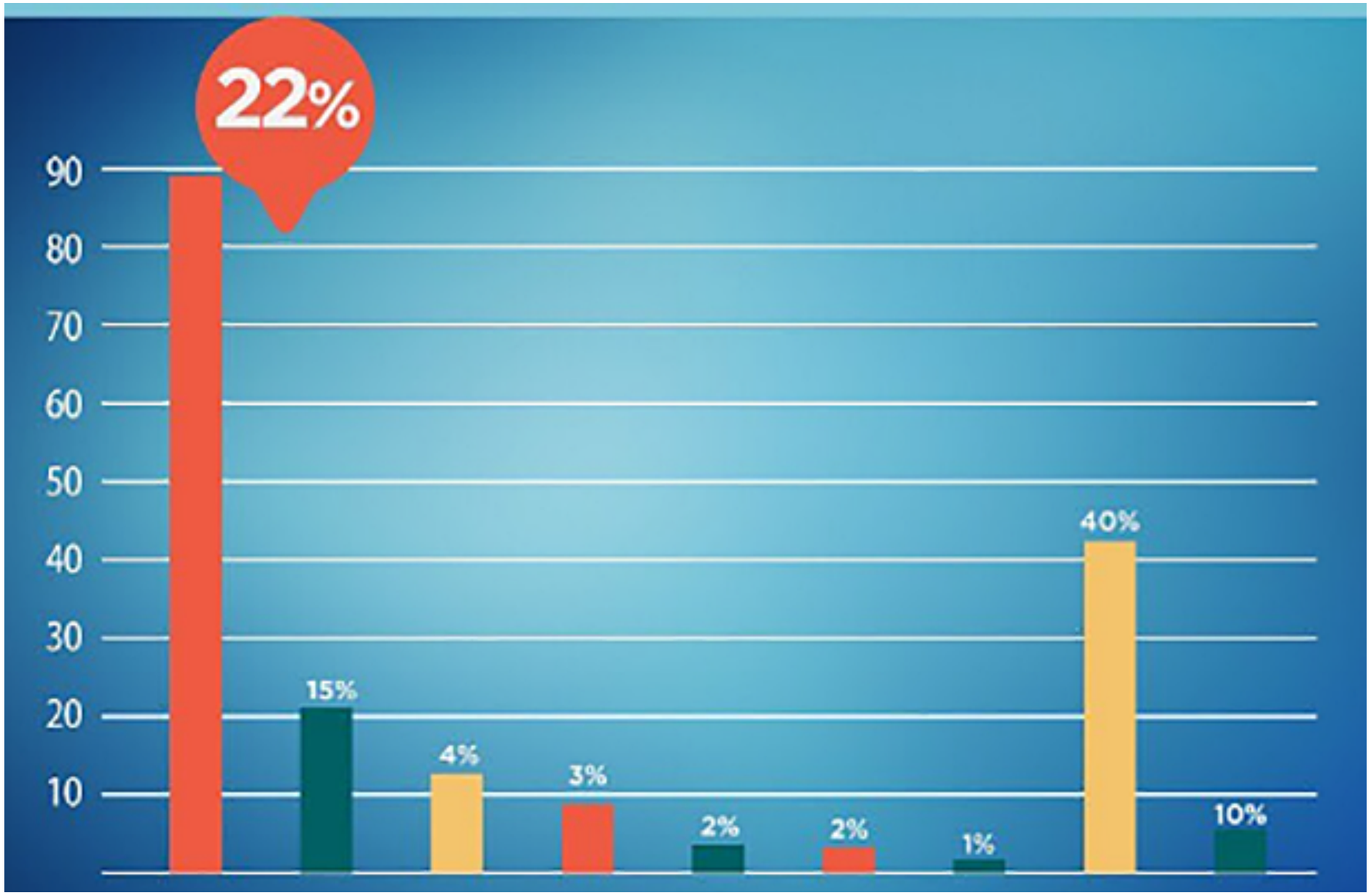
- Maximizar a razão dados-tinta



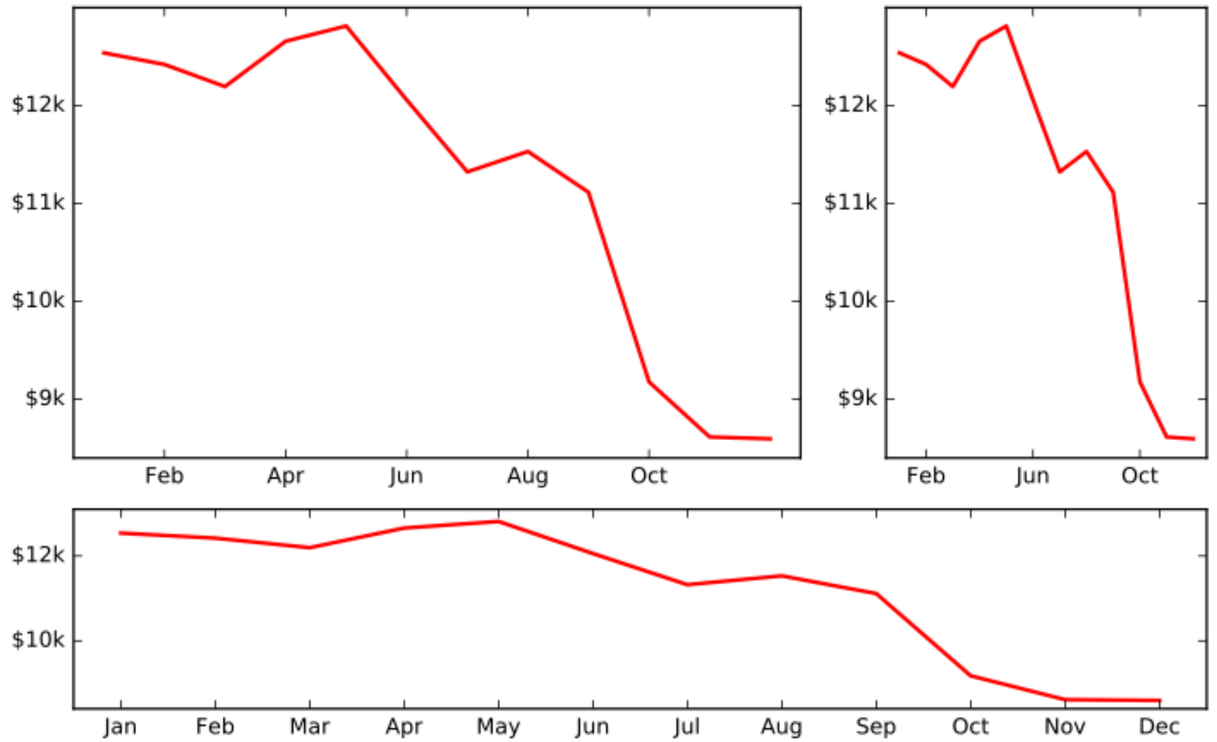


Visualização

- Minimizar o fator mentira
  - Evite apresentar médias sem a variância
  - Evite esconder a origem do gráfico
  - Evite distorções na escala


22%
90
80
70
60
50
40
30
20
10
15%
4%
3%
2%
2%
1%
40%
10%

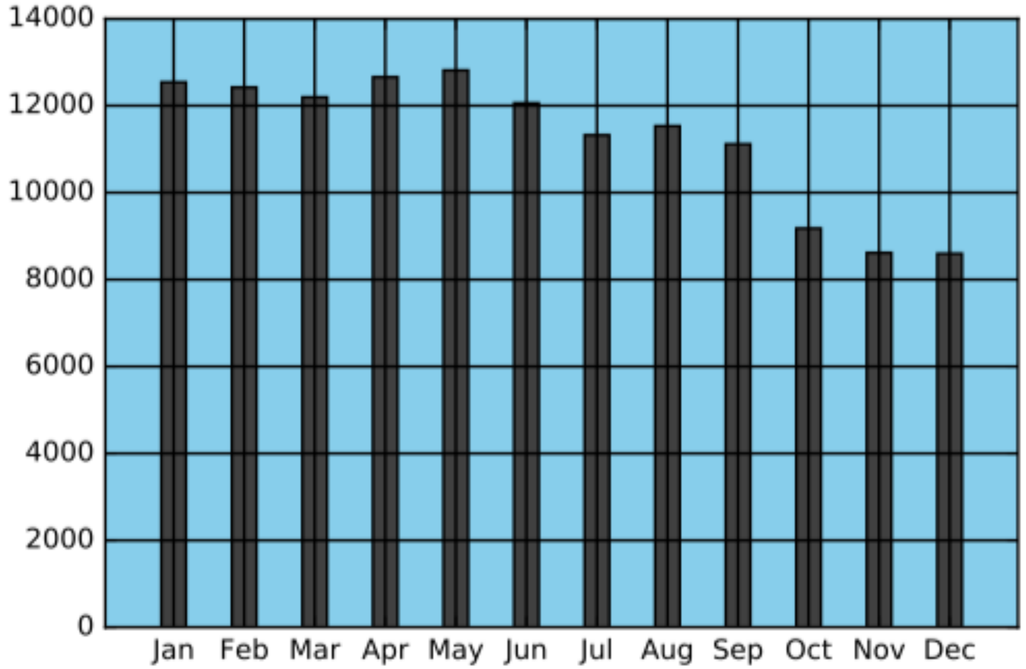
Visualização



- Mesmo gráfico, escalas diferentes

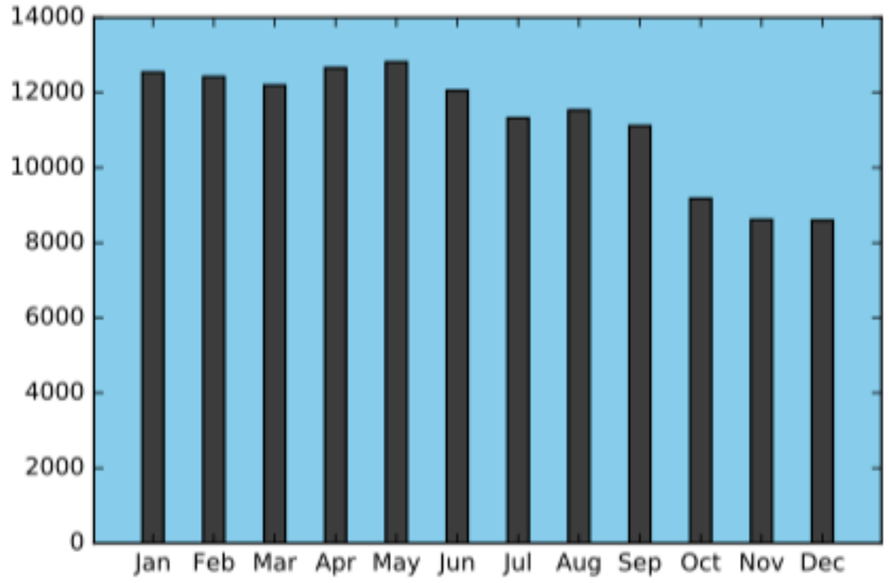
Visualização

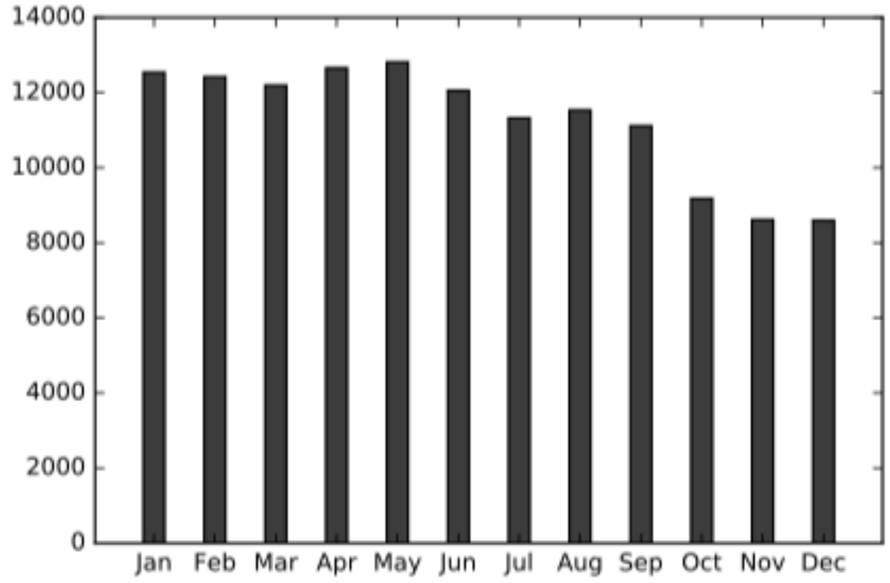
- Minimizar lixo



Visualização

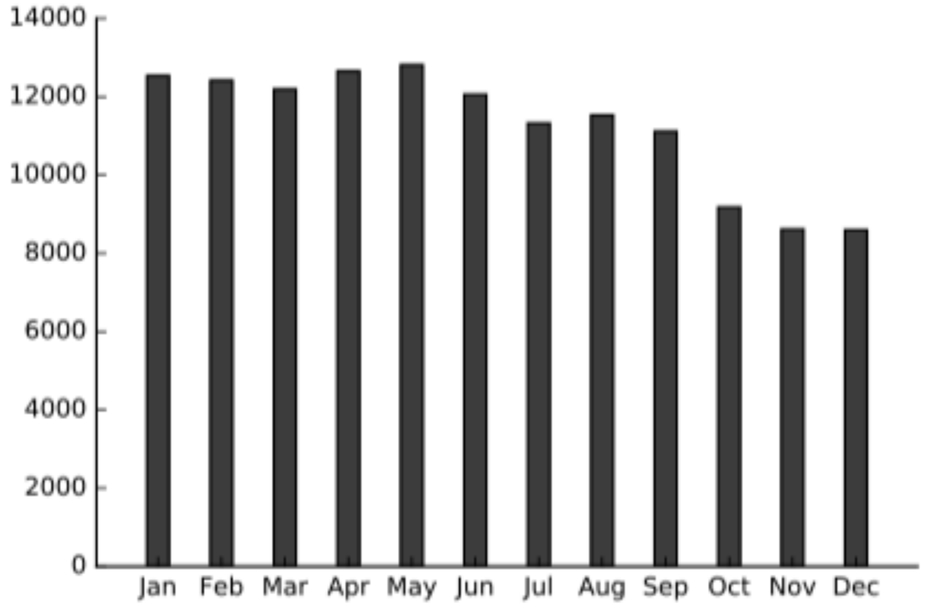
- Minimizar lixo

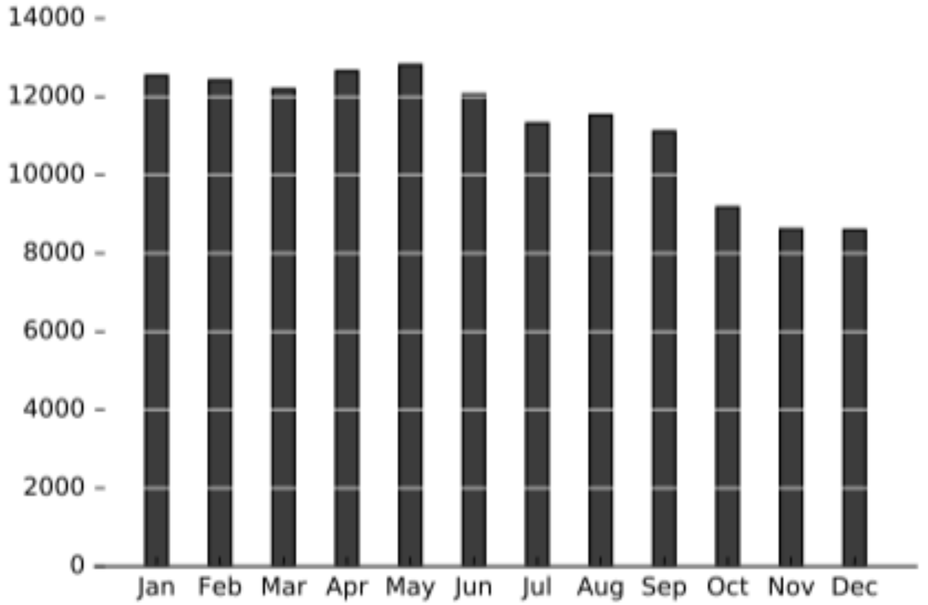




Visualização

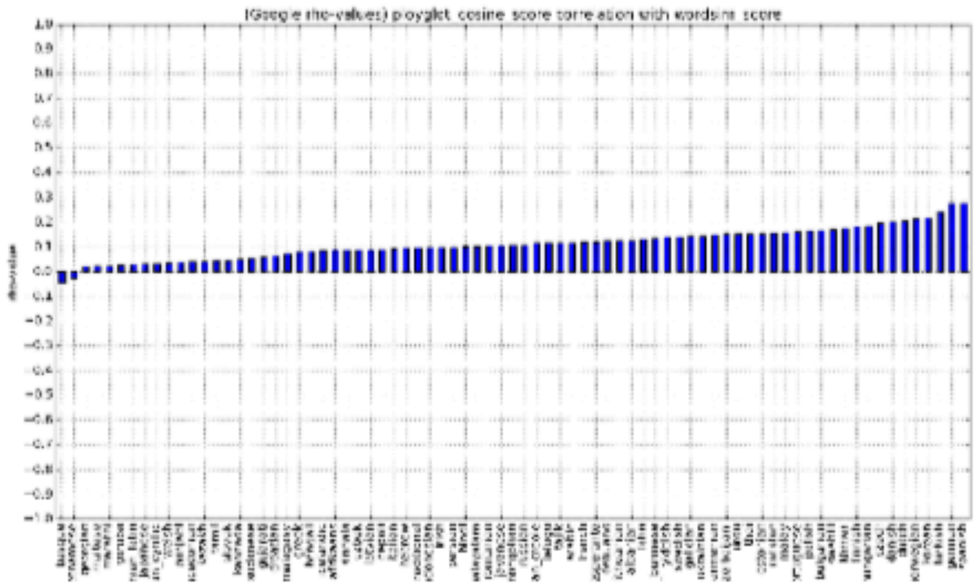
- Minimizar lixo

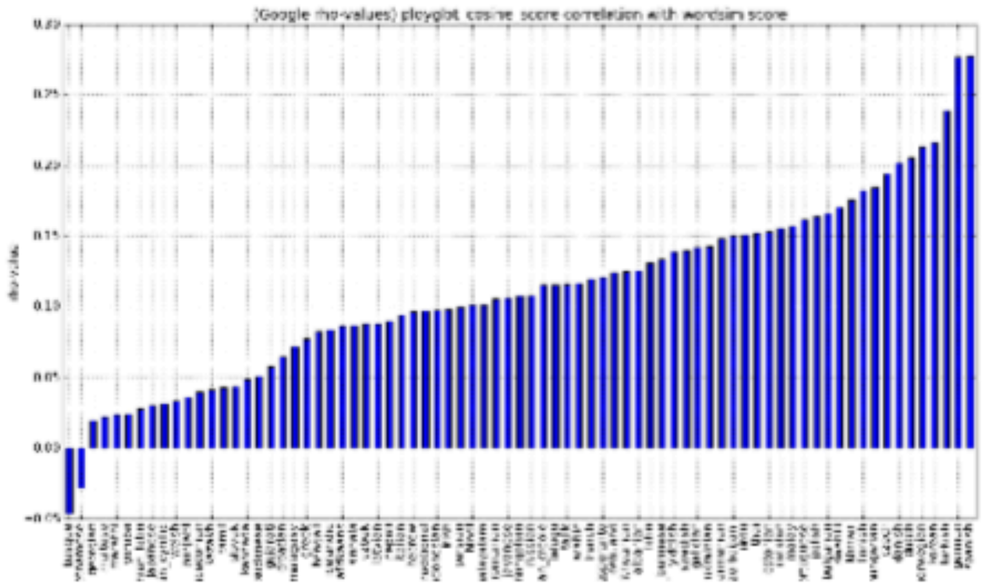




Visualização

- Use escalas adequadas e rotule os dados claramente



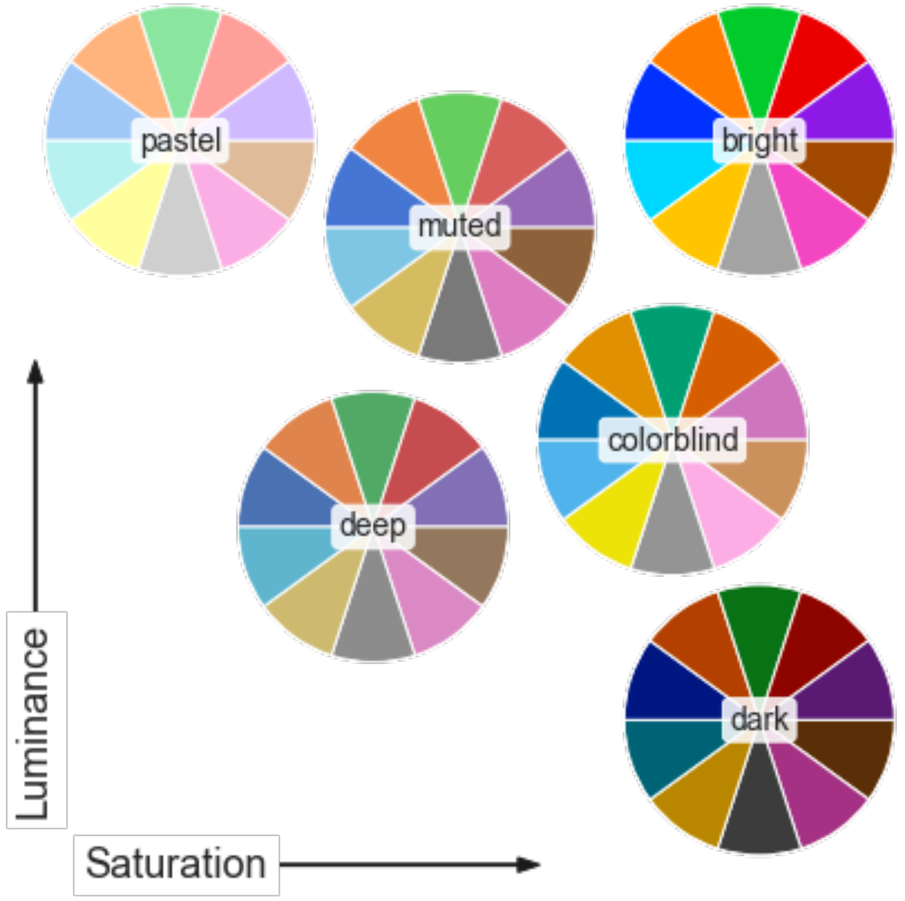


Visualização

- Use cores de forma efetiva
- A biblioteca seaborn traz diversas paletas de cores que podem ser usadas
- Elas podem ser escolhidas através do método **`sns.set_palette`**

Visualização

- Para dados categóricos temos:



Visualização

- Dados sequenciais são dados que variam de um valor baixo e não interessante até valores altos que são mais relevantes

- Para esse tipo, algumas paletas disponíveis são:

Blues

Reds

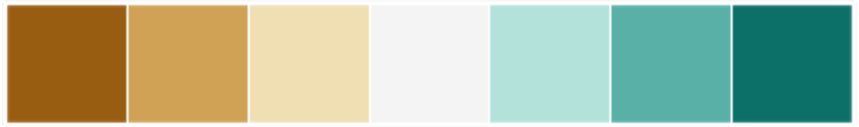
Cubehelix

Visualização

- Dados divergentes são dados que tanto os valores máximos quanto os mínimos são relevantes

- Para esse tipo, algumas paletas disponíveis são:

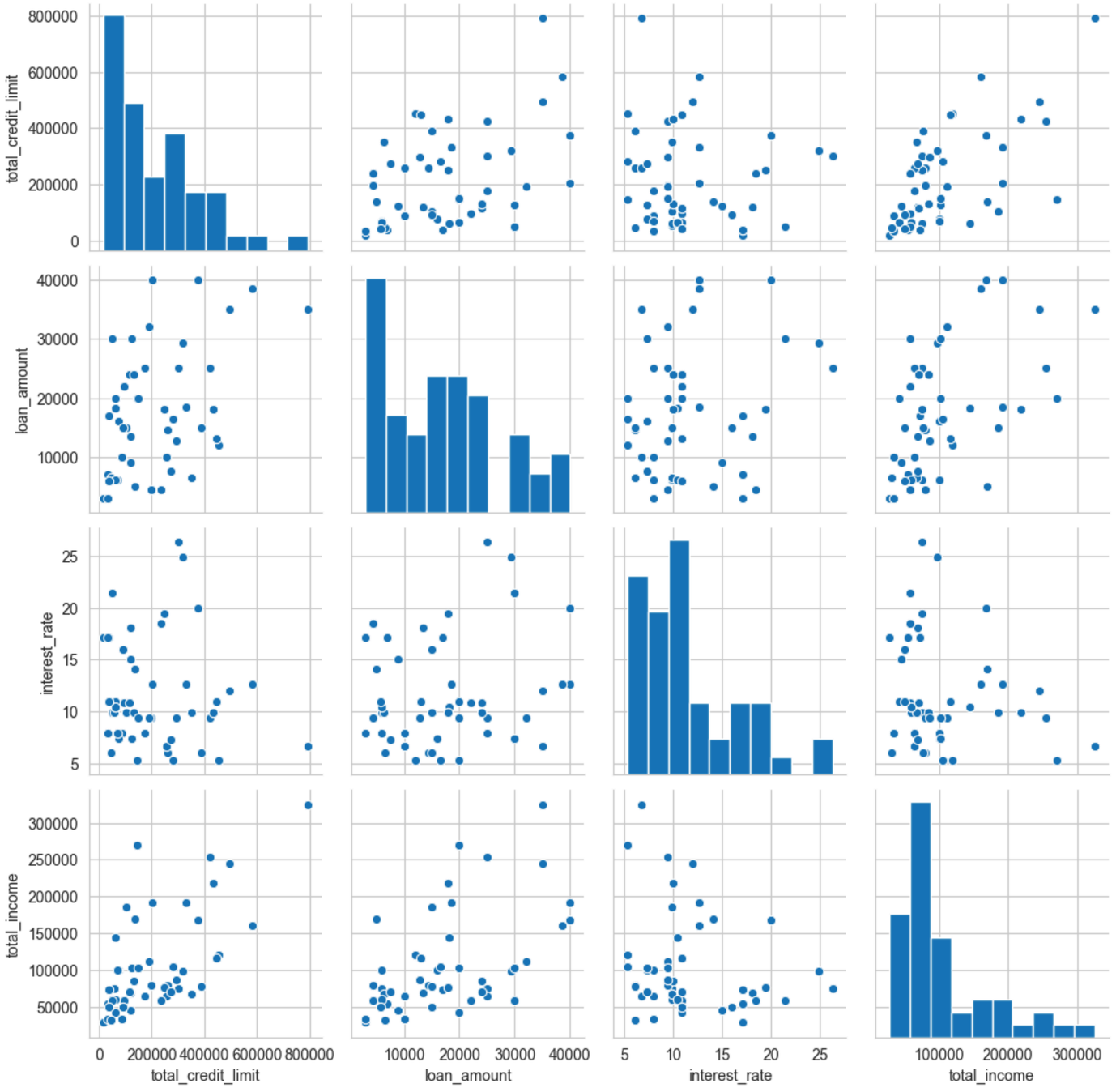
BrBG



Coolwarm

Visualização

- Explore o poder da repetição para comparações
- Múltiplos gráficos são excelentes para visualizar dados com muitas variáveis


total_credit_limit
loan_amount
interest_rate
total_income
0
200000
400000
600000
800000
10000
20000
30000
40000
5
10
15
20
25
50000
100000
150000
200000
250000
300000