04

Limpando os missings

Limpando os missings

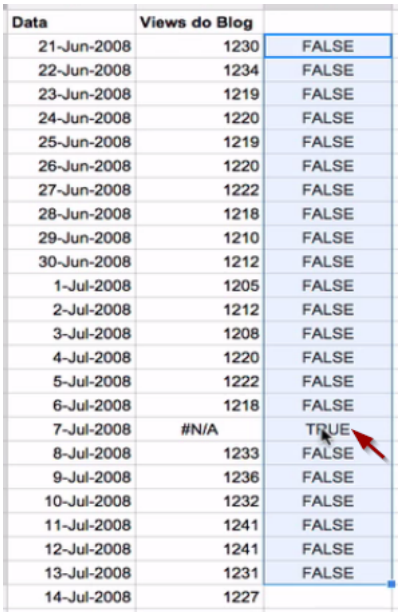
Imagine que deu tudo certo, você já apresentou as análises das informações para o seu chefe e ele ficou bastante satisfeito. Inspirado neste sucesso, você recebeu a proposta de fazer o mesmo tipo de análise para um blog sobre gatos e ver se os dados podem ajudar a Jumping Cats.

Então, você pede que os dados do blog sejam informados em um arquivo `.csv`, que é de fácil importação. Vamos importar o arquivo `blog.csv` e receber os dados que iremos analisar.

Data	Views do Blog
21-Jun-2008	1310
22-Jun-2008	1386
23-Jun-2008	1265
24-Jun-2008	1370
25-Jun-2008	1227
26-Jun-2008	1659
27-Jun-2008	1231
28-Jun-2008	1298
29-Jun-2008	1211
30-Jun-2008	1712
1-Jul-2008	1208
2-Jul-2008	1618
3-Jul-2008	1208
4-Jul-2008	1404
5-Jul-2008	1223
6-Jul-2008	1585
7-Jul-2008	#N/A
8-Jul-2008	1417
9-Jul-2008	1698
10-Jul-2008	1371
11-Jul-2008	1558
12-Jul-2008	1513
13-Jul-2008	1661
14-Jul-2008	1530
15-Jul-2008	1286
16-Jul-2008	1591
17-Jul-2008	1658
18-Jul-2008	1646

Os dados do blog são de 2008 até 2015. Observe que encontramos alguns `#N/A` s na nossa tabela. Por alguma razão, os dados destes dias não foram exportados. Minha recomendação é que estes dados não sejam apagados. Nestes casos, nós iremos **interpolar**, ou seja, baseado nos números em volta, nós iremos estimar esses números.

Podemos perceber que a série não tem grandes variações. Uma solução era substituir `#N/A` pela média. Faremos isto de uma maneira automática, usando a função `isna`. Ela irá nos indicar quais valores da tabela "não são um número" (*is not a number*).



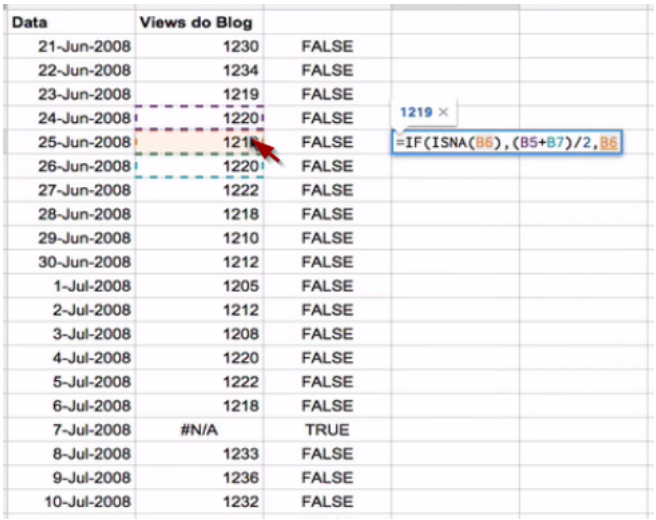
Data	Views do Blog	
21-Jun-2008	1230	FALSE
22-Jun-2008	1234	FALSE
23-Jun-2008	1219	FALSE
24-Jun-2008	1220	FALSE
25-Jun-2008	1219	FALSE
26-Jun-2008	1220	FALSE
27-Jun-2008	1222	FALSE
28-Jun-2008	1218	FALSE
29-Jun-2008	1210	FALSE
30-Jun-2008	1212	FALSE
1-Jul-2008	1205	FALSE
2-Jul-2008	1212	FALSE
3-Jul-2008	1208	FALSE
4-Jul-2008	1220	FALSE
5-Jul-2008	1222	FALSE
6-Jul-2008	1218	FALSE
7-Jul-2008	#N/A	TRUE
8-Jul-2008	1233	FALSE
9-Jul-2008	1236	FALSE
10-Jul-2008	1232	FALSE
11-Jul-2008	1241	FALSE
12-Jul-2008	1241	FALSE
13-Jul-2008	1231	FALSE
14-Jul-2008	1227	

Quando a célula tiver um `#N/A` , irá aparecer ao lado um `TRUE` .

Vamos usar uma outra função chamada `IF` . e iremos definir uma expressão lógica. Nós queremos definir comportamentos diferentes quando a expressão for `TRUE` ou `FALSE` .

Iremos usar a seguinte fórmula, usando as células `B5` . `B6` e `B7` :

```
IF(ISNA(B6),(B5+B7)/2)
```



Data	Views do Blog	
21-Jun-2008	1230	FALSE
22-Jun-2008	1234	FALSE
23-Jun-2008	1219	FALSE
24-Jun-2008	1220	FALSE
25-Jun-2008	1219	FALSE
26-Jun-2008	1220	FALSE
27-Jun-2008	1222	FALSE
28-Jun-2008	1218	FALSE
29-Jun-2008	1210	FALSE
30-Jun-2008	1212	FALSE
1-Jul-2008	1205	FALSE
2-Jul-2008	1212	FALSE
3-Jul-2008	1208	FALSE
4-Jul-2008	1220	FALSE
5-Jul-2008	1222	FALSE
6-Jul-2008	1218	FALSE
7-Jul-2008	#N/A	TRUE
8-Jul-2008	1233	FALSE
9-Jul-2008	1236	FALSE
10-Jul-2008	1232	FALSE

Neste caso, trata-se de um número. Não haverá substituições. Mas vamos aplicar a regra as outras células.

Data	Views do Blog		
21-Jun-2008	1230	FALSE	
22-Jun-2008	1234	FALSE	
23-Jun-2008	1219	FALSE	
24-Jun-2008	1220	FALSE	
25-Jun-2008	1219	FALSE	1219
26-Jun-2008	1220	FALSE	1220
27-Jun-2008	1222	FALSE	1222
28-Jun-2008	1218	FALSE	1218
29-Jun-2008	1210	FALSE	1210
30-Jun-2008	1212	FALSE	1212
1-Jul-2008	1205	FALSE	1205
2-Jul-2008	1212	FALSE	1212
3-Jul-2008	1208	FALSE	1208
4-Jul-2008	1220	FALSE	1220
5-Jul-2008	1222	FALSE	1222
6-Jul-2008	1218	FALSE	1218
7-Jul-2008	#N/A	TRUE	1225.5
8-Jul-2008	1233	FALSE	

Na célula em que o resultado de `ISNA` foi `TRUE`, o `IF` calculou a média e inclui o valor. Se fizermos o cálculo manualmente, veremos que a resposta será também `1225,5`.

Em seguida, apagaremos a coluna do `ISNA`, porque ela foi criada apenas para demonstrar a fórmula. E vamos substituir os valores de `Views do Blog` pelos da coluna do `IF`, que preencheu todas as células que estão em branco.

Data	Views do Blog
21-Jun-2008	1230
22-Jun-2008	1234
23-Jun-2008	1219
24-Jun-2008	1220
25-Jun-2008	1219
26-Jun-2008	1220
27-Jun-2008	1222
28-Jun-2008	1218
29-Jun-2008	1210
30-Jun-2008	1212
1-Jul-2008	1205
2-Jul-2008	1212
3-Jul-2008	1208
4-Jul-2008	1220
5-Jul-2008	1222
6-Jul-2008	1218
7-Jul-2008	1225.5
8-Jul-2008	1233
9-Jul-2008	1236
10-Jul-2008	1232
11-Jul-2008	1241
12-Jul-2008	1241
13-Jul-2008	1231
14-Jul-2008	1227
15-Jul-2008	1228
16-Jul-2008	1238
17-Jul-2008	1231
18-Jul-2008	1221

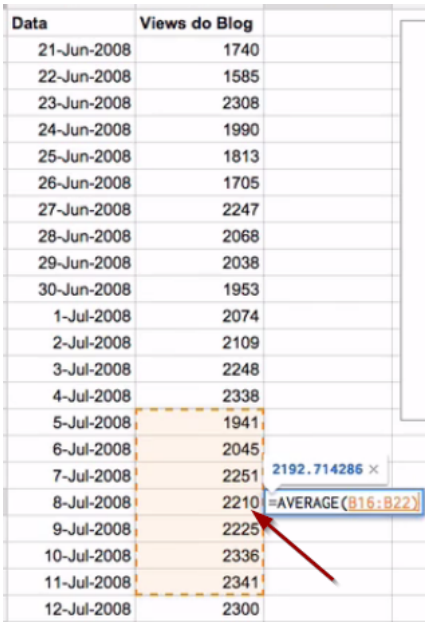
Iremos também criar o gráfico e ver quais informações conseguimos extrair dele. Em seguida iremos analisar os dados do gráfico.

Analisando os dados do blog

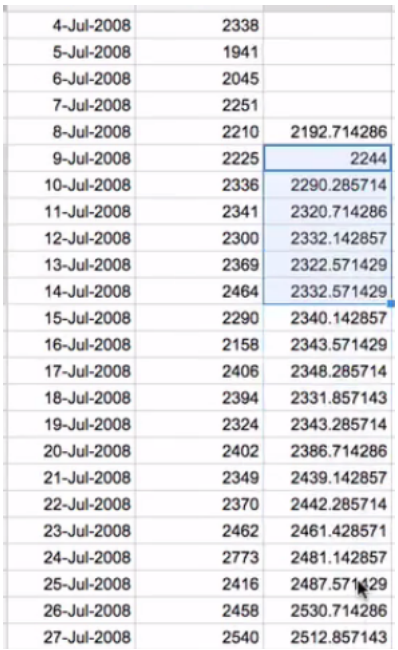
Primeiramente, podemos observar é a grande variação do gráfico. Vemos um "sobe e desce" constante. Até aqui, vimos em exemplos anteriores curvas mais simples, mas o gráfico atual é mais próximo da realidade. Encontraremos no nosso cotidiano o que chamamos de **variâncias**, estas subidas e descidas no gráfico.

Então, podemos perceber que o gráfico de *views* do blog tem mais variância que os anteriores. Isto não é um problema. Analisando o gráfico podemos dizer que o número de visitas tem uma variância de 600 acessos. Isto significa se fizermos uma campanha e tiver um aumento de acesso de 400 pessoas, o comportamento não foi excepcional.

Mas caso exista o desejo de criar uma curva suave, temos uma forma de fazer isto, usando o `AVERAGE` do Google Spreadsheets, que irá nós fornecer uma média de um determinado período. No nosso caso, iremos começar tirando a média de três dias anteriores a 8 de Julho de 2008 e três dias posteriores.



A média da semana será `2.192,714286`. Podemos fazer o mesmo processo com as demais células, por isso chamamos de **médias móveis**, porque conseguimos gerar uma média que "escorrega" junto com os valores.



Vamos completar a coluna da Média Móvel de 7 dias.

Data	Views do Blog	Media Móvel - 7
21-Jun-2008	1740	
22-Jun-2008	1585	
23-Jun-2008	2308	
24-Jun-2008	1990	1912.571429
25-Jun-2008	1813	1959.428571
26-Jun-2008	1705	2024.142857
27-Jun-2008	2247	1973.428571
28-Jun-2008	2068	1985.428571
29-Jun-2008	2038	2027.714286
30-Jun-2008	1953	2105.285714
1-Jul-2008	2074	2118.285714
2-Jul-2008	2109	2100.142857
3-Jul-2008	2248	2101.142857
4-Jul-2008	2338	2143.714286
5-Jul-2008	1941	2163.142857
6-Jul-2008	2045	2179.714286
7-Jul-2008	2251	2192.285714
8-Jul-2008	2210	2192.714286
9-Jul-2008	2225	2244
10-Jul-2008	2336	2290.285714
11-Jul-2008	2341	2320.714286
12-Jul-2008	2300	2332.142857
13-Jul-2008	2369	2322.571429
14-Jul-2008	2464	2332.571429
15-Jul-2008	2290	2340.142857
16-Jul-2008	2158	2343.571429
17-Jul-2008	2406	2348.285714
18-Jul-2008	2394	2331.857143

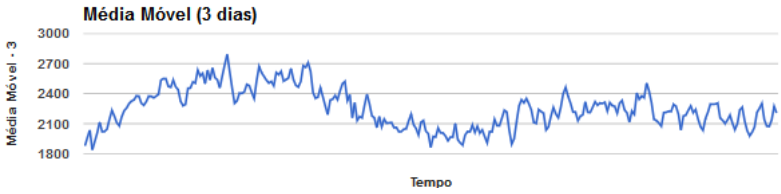
Nós estamos interessados no formato do nosso gráfico, por isso é irrelevante que os dados dos três primeiros dias não estejam preenchidos. Também criaremos o gráfico das médias móveis.

Vemos que este gráfico tem uma curva mais suave que o anterior.

Mais adiante, testaremos como seria o nosso gráfico, se fizessemos a média móvel de 3 dias, em que consideraremos a data, o dia anterior e o posterior.

Data	Views do Blog	Media Móvel - 7	Média Móvel - 3
21-Jun-2008	1740		
22-Jun-2008	1585		1877.666667
23-Jun-2008	2308		1961
24-Jun-2008	1990	1912.571429	2037
25-Jun-2008	1813	1959.428571	1836
26-Jun-2008	1705	2024.142857	1921.666667
27-Jun-2008	2247	1973.428571	2006.666667
28-Jun-2008	2068	1985.428571	2117.666667
29-Jun-2008	2038	2027.714286	2019.666667
30-Jun-2008	1953	2105.285714	2021.666667
1-Jul-2008	2074	2118.285714	2045.333333
2-Jul-2008	2109	2100.142857	2143.666667
3-Jul-2008	2248	2101.142857	2231.666667
4-Jul-2008	2338	2143.714286	2175.666667
5-Jul-2008	1941	2163.142857	2108
6-Jul-2008	2045	2179.714286	2079
7-Jul-2008	2251	2192.285714	2168.666667
8-Jul-2008	2210	2192.714286	2228.666667
9-Jul-2008	2225	2244	2257
10-Jul-2008	2336	2290.285714	2300.666667
11-Jul-2008	2341	2320.714286	2325.666667
12-Jul-2008	2300	2332.142857	2336.666667
13-Jul-2008	2369	2322.571429	2377.666667
14-Jul-2008	2464	2332.571429	2374.333333
15-Jul-2008	2290	2340.142857	2304
16-Jul-2008	2158	2343.571429	2284.666667
17-Jul-2008	2406	2348.285714	2319.333333
18-Jul-2008	2394	2331.857143	2374.666667

Veremos que agora, a curva no gráfico ficará menos suave.



No gráfico de 7 dias, diminuímos a variância. O mesmo aconteceria se fizéssemos um gráfico de 9 dias. Porém, existe um problema em fazer uma média móvel de um período muito longo. Vamos testar a média de 21 dias.

Data	Views do Blog	Media Móvel - 7	Média Móvel - 3	Média Móvel - 21
21-Jun-2008	1740			
22-Jun-2008	1585		1877.666667	
23-Jun-2008	2308		1961	
24-Jun-2008	1990	1912.571429	2037	
25-Jun-2008	1813	1959.428571	1836	
26-Jun-2008	1705	2024.142857	1921.666667	
27-Jun-2008	2247	1973.428571	2006.666667	
28-Jun-2008	2068	1985.428571	2117.666667	
29-Jun-2008	2038	2027.714286	2019.666667	
30-Jun-2008	1953	2105.285714	2021.666667	
1-Jul-2008	2074	2118.285714	2045.333333	2074.52381
2-Jul-2008	2109	2100.142857	2143.666667	2101.190476
3-Jul-2008	2248	2101.142857	2231.666667	2138.52381
4-Jul-2008	2338	2143.714286	2175.666667	2145.952381
5-Jul-2008	1941	2163.142857	2108	2160.238095
6-Jul-2008	2045	2179.714286	2079	2176.666667
7-Jul-2008	2251	2192.285714	2168.666667	2210.047619
8-Jul-2008	2210	2192.714286	2228.666667	2217.047619
9-Jul-2008	2225	2244	2257	2229.238095
10-Jul-2008	2336	2290.285714	2300.666667	2246.571429
11-Jul-2008	2341	2320.714286	2325.666667	2265.428571
12-Jul-2008	2300	2332.142857	2336.666667	2279.52381
13-Jul-2008	2369	2322.571429	2377.666667	2296.333333
14-Jul-2008	2464	2332.571429	2374.333333	2321.333333
15-Jul-2008	2290	2340.142857	2304	2325.047619
16-Jul-2008	2158	2343.571429	2284.666667	2349.666667
17-Jul-2008	2406	2348.285714	2319.333333	2373.238095
18-Jul-2008	2394	2331.857143	2374.666667	2380.047619

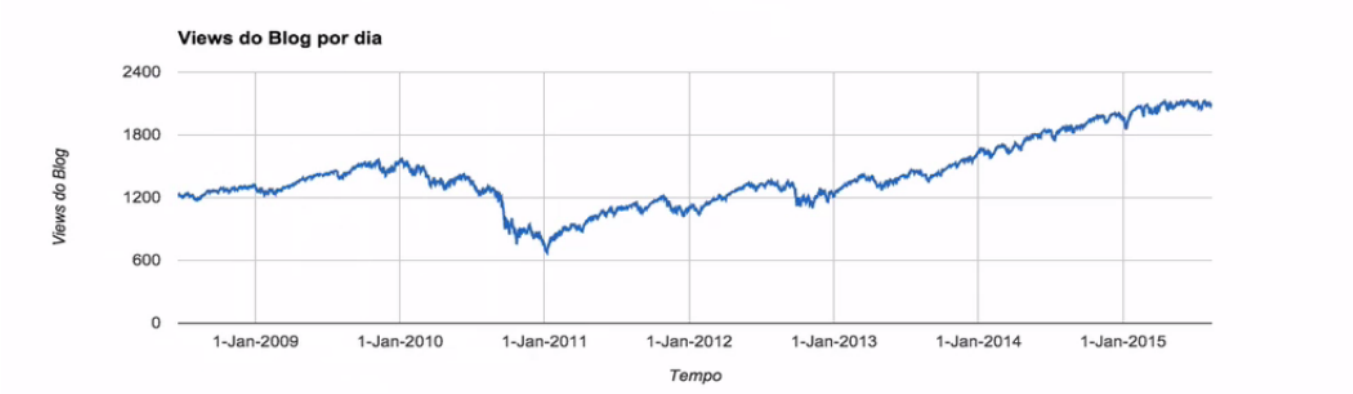
Vemos que o desenho do gráfico é o mais suave de todos, principalmente quando comparado com o primeiro. No entanto, a desvantagem é que perdemos a chance de analisar detalhes e não conseguimos analisar comportamentos diferenciados em determinados períodos.

Desta forma, variações pontuais, como picos em um dia, irá desaparecer. Por outro lado, enxergamos melhor as variações de um mês. Então, a forma correta de escolher o recorte de tempo é se perguntar: "o que eu quero responder sobre um determinado período?"

Análise de Dados é um grande quebra-cabeça. Cada gráfico irá trazer uma nova informação. E com cada um deles, nós preenchemos um peça que faltava, e compreendemos melhor o todo da nossa empresa. Com isto, conseguimos tomar melhores decisões para o negócio, que no fim é o objetivo da análise de dados: auxiliar e fundamentar decisões de negócios empresariais.

Interpretando corretamente os gráficos

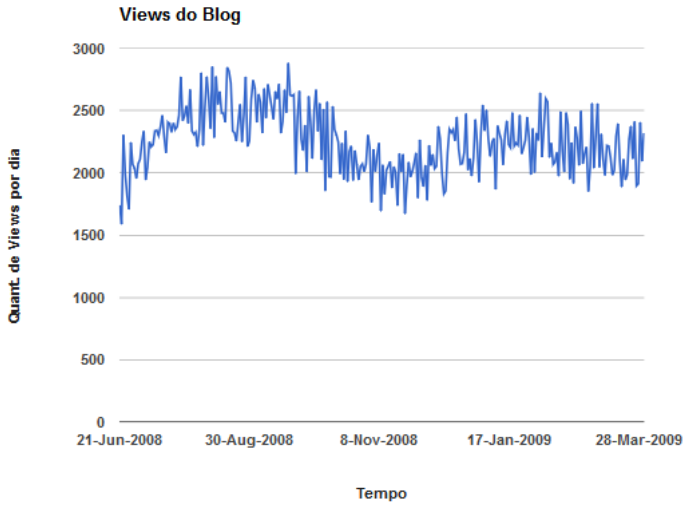
Vamos ressaltar algo que foi dito anteriormente, mas não foi muito destacado. Observe novamente o gráfico de *views* do blog:



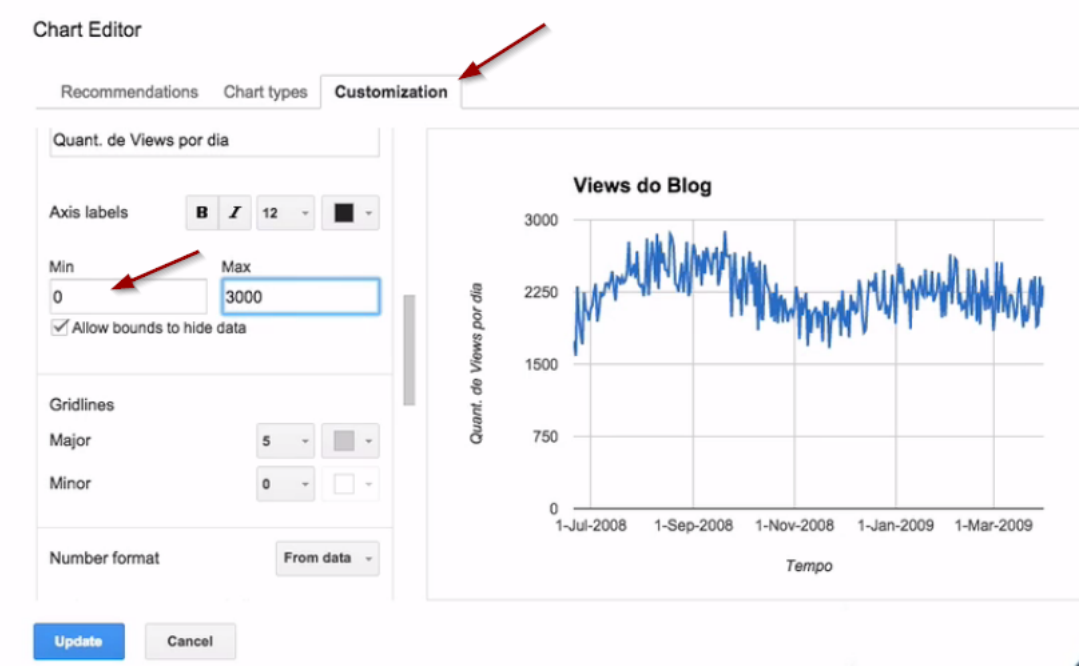
O primeiro item que reparamos, quando analisamos um gráfico e ono formato geral da linha, o quanto ela ocupa a tela. Em geral, inicialmente, não lemos a escala. Por exemplo, um observador não irá ler que o gráfico vai de `1500` a `3000`. Logo, os pontos mais baixos do gráfico podem ser considerados superficial, e avaliados como resultados ruins. É uma percepção errada, que ocorre naturalmente.

No momento em que vamos interpretar um gráfico é preciso ser cauteloso sobre isso, para que não resulte em análises e decisões errôneas. Como poderíamos criar um gráfico que gere menos dúvidas, facilitando a interpretação?

A maneira correta de se fazer isto é **zerando** a escala, fazer com que ela vá de `0` a `3000`.



Podemos fazer este tipo de configuração, nos parâmetros do editor do gráfico.



Se aumentarmos o topo para um número muito alto, como `8000`, por exemplo, vamos esmagar o gráfico. Então, usaremos o valor máximo que foi vendido: `3000`. Com as alterações do gráfico, facilitamos uma análise mais clara sobre os números e a percepção do mínimo de visualizações. Da maneira como estava antes, parecia que o número chegava a `0`, o que não condiz com a verdade.

Por isso, **atenção** quando for montar o gráfico, seja **cuidadoso com as escalas**, porque isto pode gerar uma má interpretação das informações.