02

Limpando os missings

Transcrição

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	

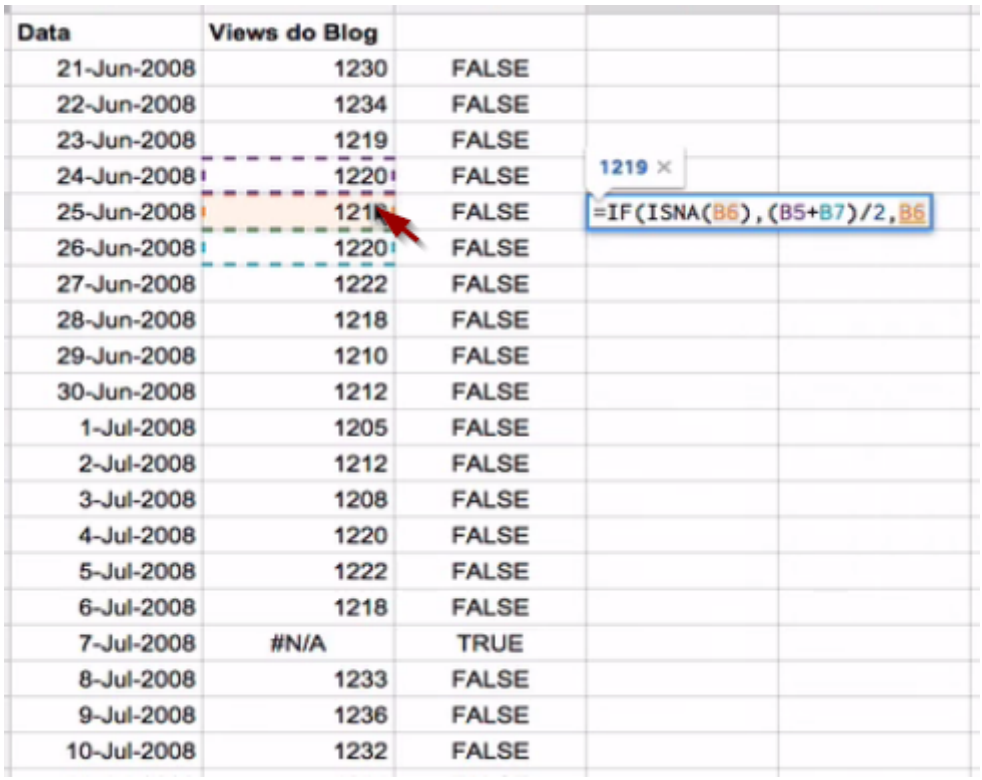
Quando a célula tiver um `#N/A` , irá aparecer ao lado um `TRUE` .

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` :

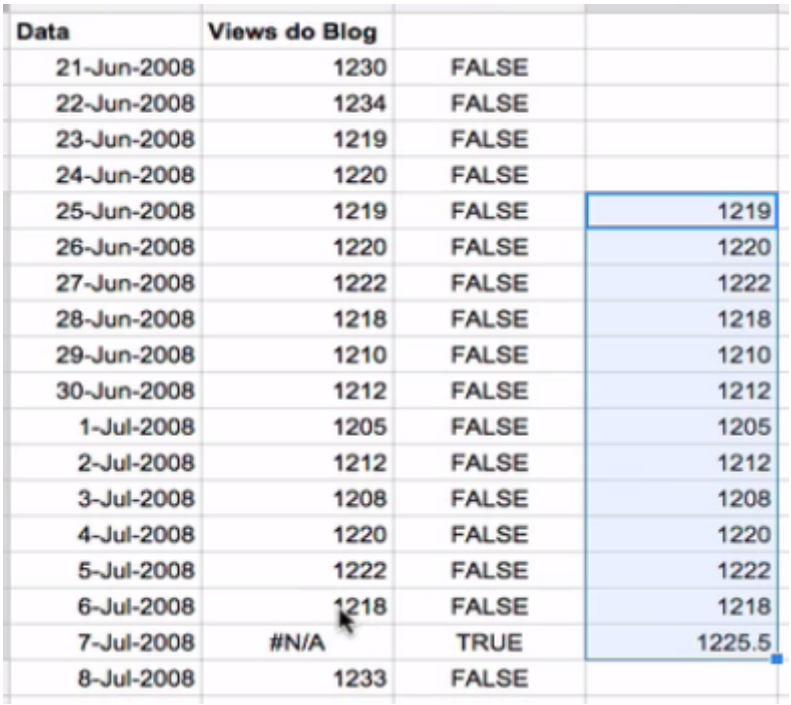
```
IF(ISNA(B6),(B5+B7)/2)
```

COPIAR CÓDIGO



Data	Views do Blog		
21-Jun-2008	1230	FALSE	
22-Jun-2008	1234	FALSE	
23-Jun-2008	1219	FALSE	
24-Jun-2008	1220	FALSE	1219 ×
25-Jun-2008	121	FALSE	=IF(ISNA(B6),(B5+B7)/2,B6
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.