# Enriquecimento dos dados

## Combinar 2 variáveis diferentes

a) Vamos combinar as colunas likes\_mean e dislikes\_mean calculando a razão de uma sobre a outra. Isso nos dará uma ideia da proporção entre likes e dislikes de um comentário.

b) Por ser uma divisão, essa variável ficou com alguns valores estranhos, como nulos e infinitos.

O nulo acontece quando dividimos 0 por 0. Já o infinito acontece quando temos algum like e 0 dislikes.

Calcule o valor de nulos e infinitos (np.inf).

c) Vamos preencher os nulos com o valor 1 que indica mesma qt. de likes e dislikes

d) Plote o histograma dessa nova variável

e) Como essa variável tá com alguns picos indesejados e uma grande quantidade de outliers, faz sentido nesse caso aplicarmos a discretização por equal-frequncy usando 3 bins.

Dica: para funcionar o algoritmo do KBins, antes substitua os infinitos por um valor alto

f) Verifique a nova distribuição

## Criação de uma nova dimensão

Vamos retornar aos datasets de comentários! Já que não sabemos se os comentários são positivos ou não, poderiamos calcular uma dimensão de comentários que indique se o autor daquele comentário chegou a sair da empresa ou não. Isso pode nos indicar que um funcionário que dê like em algum comentário de alguém que saiu pode ser mais propenso a churn também.

a) Faça um merge entre a df\_comments\_clean\_anonimized e a df\_churn\_enriq, trazendo somente a variável stillExists.

b) Agora faça um merge entre a df\_comment\_interactions e essa nova base, trazendo somente a variável stillExists.

c) Nosso objetivo agora é resgatar o conceito de pivot table e calcular a media do stillExists para cada <employee, company> separando em liked e disliked comments.

#### Importante!

Antes de realizarmos um merge com uma pivot table, precisamos dimencionar o nome das colunas da mesma:

d) Agora vamos fazer um merge com o dataset inicial!

## Tratamentos em uma única variável

a) Plote os histogramas de cada variável numérica. As distribuicões parecem normais?

b) Muitas variáveis do dataset estão com uma cauda muito longa na distribuição, vamos melhorar isso?

Aplique o Box-Cox Transformation para as variáveis de bx\_columns.

**Observação:** Valores negativos ou zeros não são aceitos pelo algoritmo!

d) Visualize novamente após a transformação

### Variáveis Categóricas

a) Faça um value\_counts() das variáveis categóricas

b) Aplique o One-Hot Encoding para cada uma delas, aplicando o drop\_first.

## Validação & Limpeza Final

a) Ver correlação entre variáveis e variável target

b) Verifique de novo as distribuições e os tipos das variáveis restantes

c) Há ainda algum valor nulo? Se sim, trate-os.

d) Vamos ver quais colunas manter