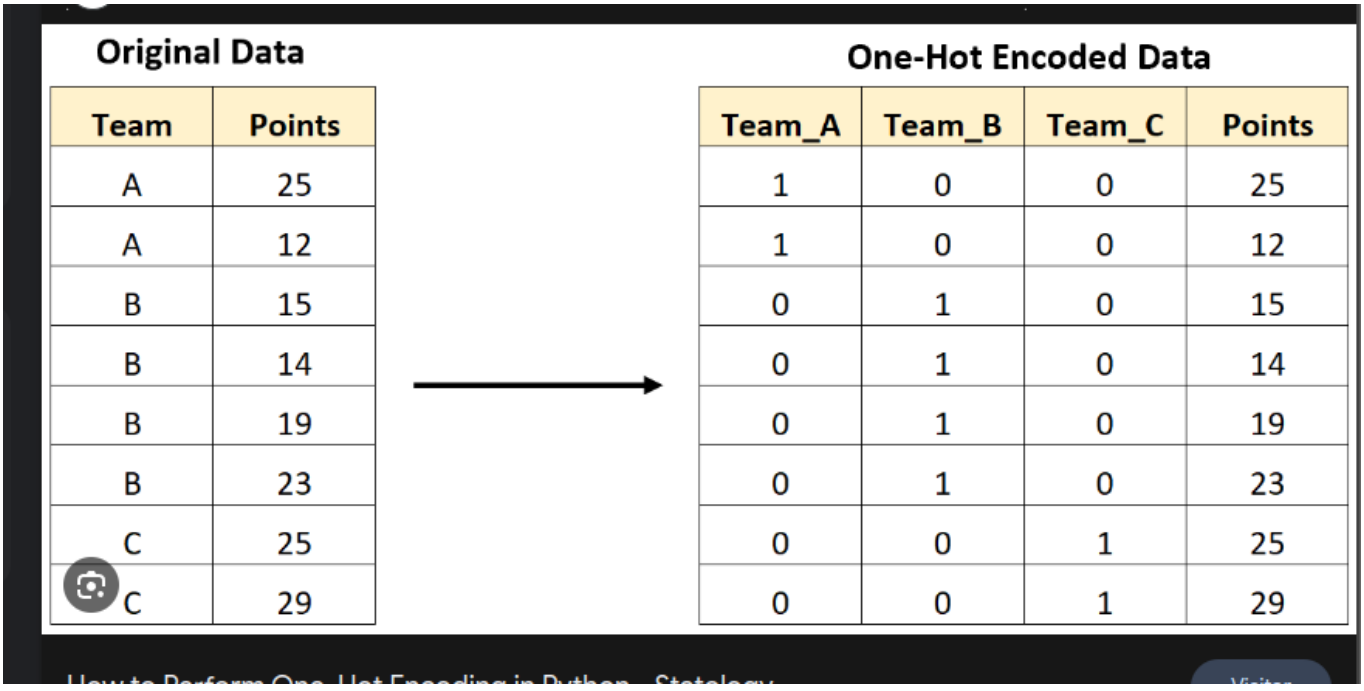
One hot encoding

Quando temos uma variável categórica e desejamos mantê-la como feature no dataset, devemos tomar muito cuidado ao apenas transformá-la em um número correspondente, por exemplo, se quisermos manter uma variável Bairro de um dataset, podemos ter N bairros, e ao tentar convertê-los para número, teremos 1, 2, 3, 4, 5 ... N. E cada bairro teria seu identificador correspondente. Porém isso não está certo, pois o bairro correspondente ao ID 16 teria duas vezes mais peso que o bairro de ID 8 para determinado modelo, ou seja, a abordagem de referenciar o bairro apenas com um número não funcionaria.

Para resolver esse problema podemos utilizar o One hot encoding, que transforma cada bairro em uma coluna do banco de dados, onde somente os valores 0 ou 1 podem ser atribuídos, ou seja, dado uma amostra que pertence ao bairro 1, todas as colunas bairro estariam com valor 0 e somente o bairro 1 teria o valor 1, isso serve para todas as amostras. E dessa forma, conseguimos classificar a qual bairro aquela amostra pertence sem atribuir um peso diferente para cada bairro. Também podemos verificar a correlação que cada bairro teria com a variável target.

Exemplo de one hot encoded data:



Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Cada valor único da coluna Team acaba virando uma coluna e tem seus valores binários se a amostra pertencer a um time A, B ou C. Não é possível pertencer a mais de um time ao mesmo tempo.