Relatório do primeiro trabalho de Trabalhando com grande volumes de dados - Big Data

Annanda Dandi 112105859

Pergunta a ser respondida a partir do dataset: Qual a porcentagem dos personagens que morreram no mesmo livro em que foram apresentados?

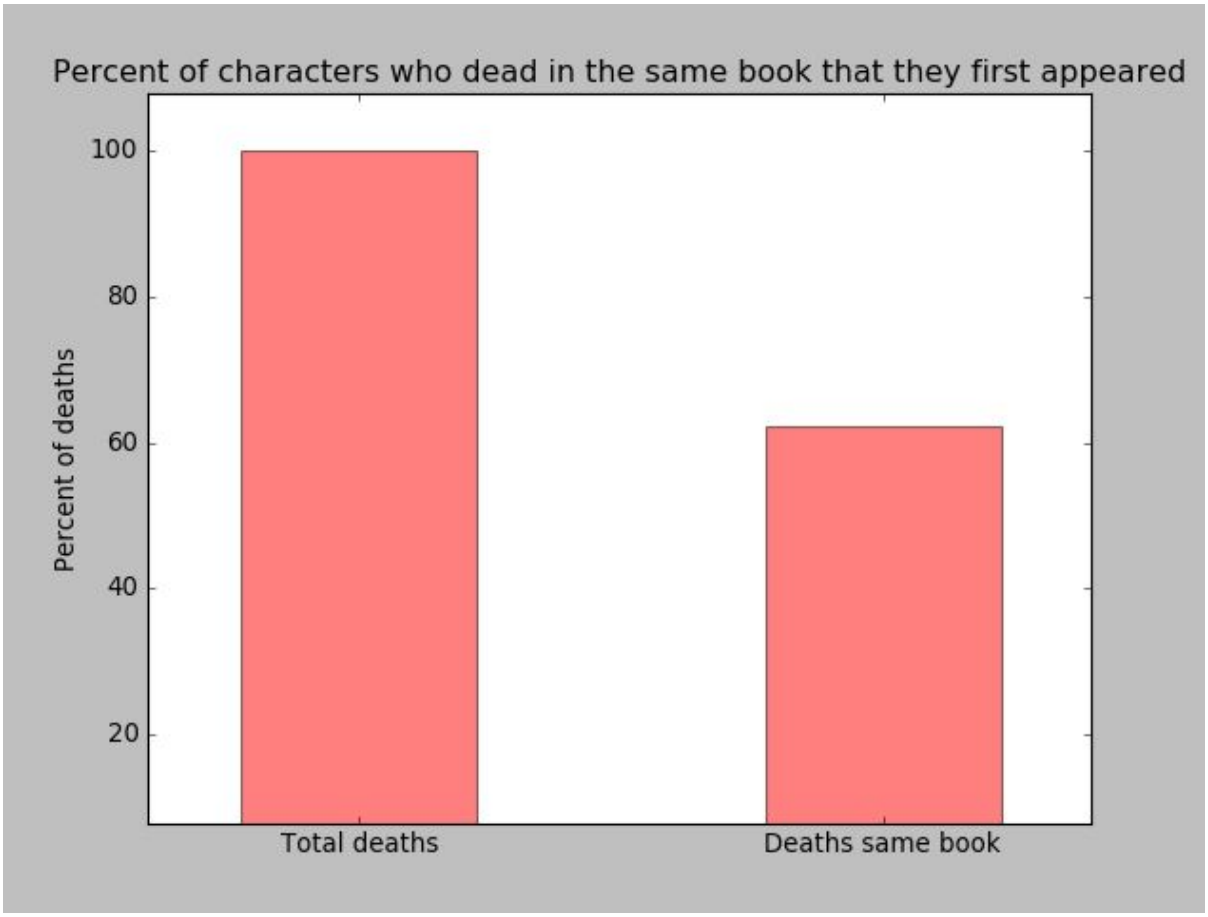
Para responder a essa pergunta eu preciso dos seguintes dados: o número total de mortes e o número de personagens que morreram no mesmo livro em que foram apresentados.

Para saber o número total de mortes eu percorri todas as linhas do dataset e quando na quarta coluna (de índice 3) houvesse um valor indicando o livro que o personagem morreu, eu somava o contador `total_deaths`.

Para saber se o personagem morreu no mesmo livro que apareceu pela primeira vez, eu fiz uma função que retorna o livro que o personagem apareceu a primeira vez `first_book_appeared`().
Se o retorno dessa função for igual ao valor da quarta coluna do dataset (de índice 3, que contém o livro em que o personagem morreu) eu somaria o valor do contador `total_deaths_same_book_appeared`.

Então para fazer a porcentagem eu dividi o valor de `total_deaths_same_book_appeared` por `total_deaths`.

O resultado que obtive foi 62.21%. Ou seja, mais da metade dos personagens que morreram, foram introduzidos no mesmo livro em que morreram.



Observações

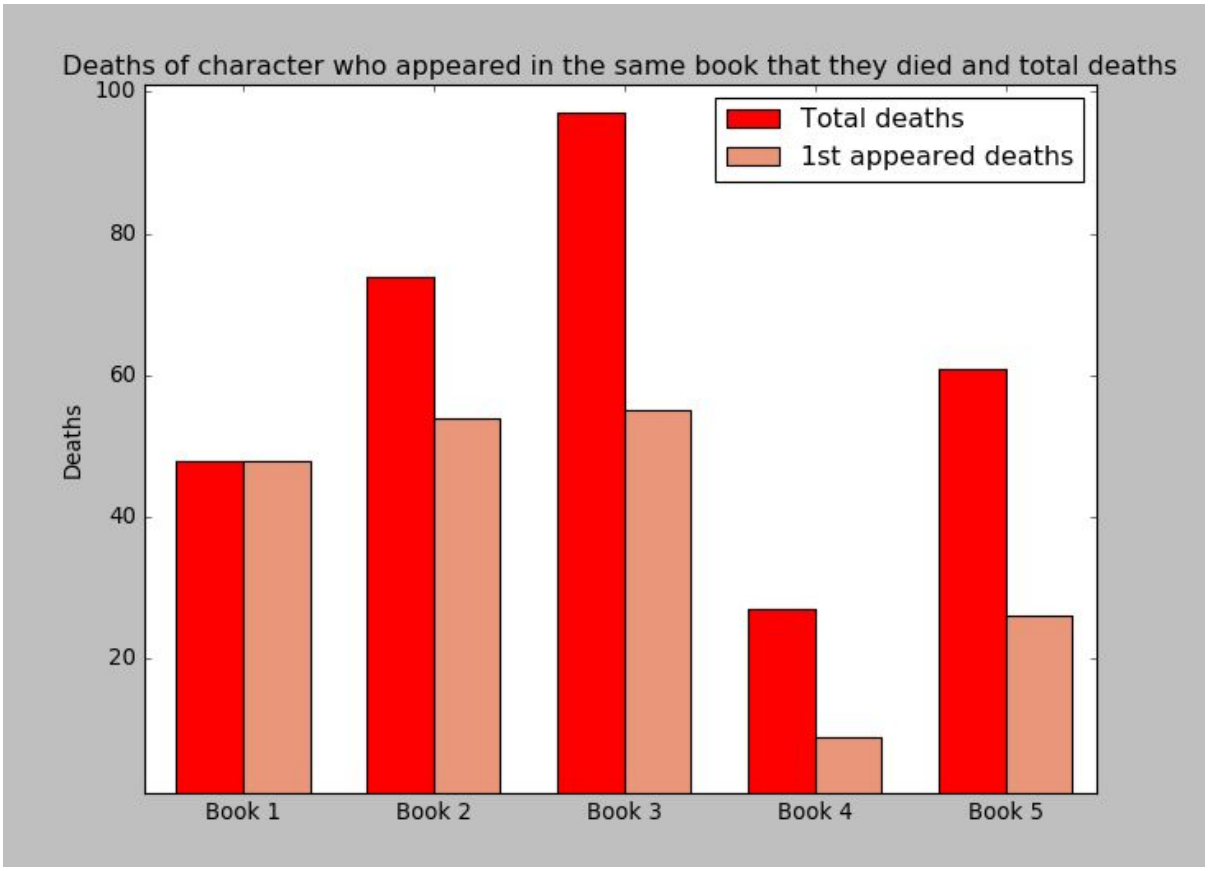
Há um erro no dataset. Pois ao contabilizar todas as mortes do livro 1, 49 mortes, esse também deveria ser o número de mortes dos personagens que foram introduzidos nesse livro, afinal de contas, não tem um livro anterior em que o personagem poderia ter aparecido a primeira vez.
Porém não é o que acontece, o número contabilizado de mortes no livro 1 de personagens que apareceram no livro 1 é igual a 48, um a menos do que deveria ser.

Ao investigar, vi que a linha do dataset que gerava o problema é a 157:
['Cressen', 'Baratheon', '299', '1', '0', '0', '1', '0', '0', '1', '0', '0', '0']

Ao observá-la nota-se que diz que o personagem morreu no livro 1, mas não apareceu no livro 1 e apareceu no livro 2.
Quem conhece a história poderia até pensar que o personagem ressuscitou no livro 2, mas como poderia ele ter morrido no livro 1 sem aparecer no mesmo?
Pode ter sido um erro de definição do que significa aparecer no livro. Ou poderia ser um erro no dataset.

Após uma pesquisa no Google verifiquei que o erro está no dado do livro em que o personagem morreu. Pois ele morreu no livro 2, e não no livro 1.

Então corrigindo a porcentagem, 62.54% dos personagens que morreram, foram introduzidos no mesmo livro em que morreram.



Acima está o gráfico comparando o número total de mortes em cada livro, e o número de mortes de personagens que tinham sido apresentados no livro em que morreram.