

A Report on Audible Uncleaned Dataset

Issues with Audible Uncleaned Dataset

1. Redundant Data issue

‘Writtenby:’ and ‘Narratedby:’ in the [author] and [narrator] columns next to actual data is a kind of redundant data that is appearing along the whole column. These can cause ineffective results or analysis in the future.

2. Data-type conversion issue

Columns [price] and [stars] contribute to numeric data in general, but in the uncleaned dataset it contains a mix of numbers and characters. A numeric analysis cannot be done for ones whose data type is non-numeric.

Techniques to overcome the issues above mentioned

1. Solving redundant data problems

We can remove redundant information by using either Delimiter or Find & Replace tool which is followed along entire column like ‘Writtenby:’ in [Author] column. I have used Text After Delimiter from power query to extract the author name from column [author].



author
Writtenby:GeronimoStilton
Writtenby:RickRiordan
Writtenby:JeffKinney

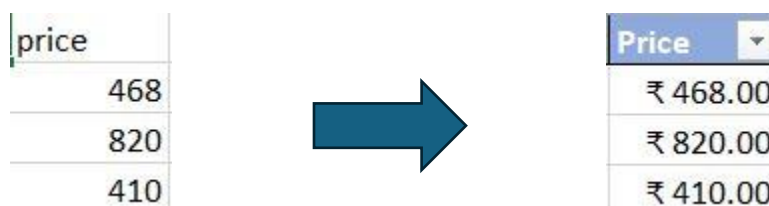
Author
Geronimo Stilton
Rick Riordan
Jeff Kinney

Or else Find and Replace tool can also be used instead of Text After Delimiter, we have to replace ‘Writtenby:’ or ‘Narratedby:’ with none.

2. Solving data conversion problem

a. Type conversion of column [price]

In the uncleaned dataset, price column is holding a numerical data type instead of currency. Just converting the numerical data type of whole column [price] to currency datatype will give us the direct desired result.




price
468
820
410

Price
₹ 468.00
₹ 820.00
₹ 410.00

There was an exception while converting the data type, data of 'Free' was found which is logically ₹0. So, the data 'Free' could be replaced with ₹0 by using Find & Replace Tool.

b. Type conversion of column [rating]

There are two different kinds of data in the same column [rating], one is rating and, the other number of ratings. We need to use split columns to make 2 columns for each kind of data.



stars		Rating (out of 5)	No.of Ratings
5 out of 5 stars34 ratings		5	34
4.5 out of 5 stars41 ratings		4.5	41
4.5 out of 5 stars38 ratings		4.5	38

After splitting the columns, Text After Delimiter was used again to clear the buffer ' out of ', ' stars' and ' ratings' to make the columns a pure numerical value.

Assignment by:

Varun Kumar Kolloju

Matriculation number 95675739