Audible Data Cleaning Project

Presented by- BHUVAN JARI

www.linkedin.com/in/bhuvan-jari

Introduction❗

Audible is an American online audiobook and podcast service that allows users to purchase and stream audiobooks and other forms of spoken word content. This content can be purchased individually or under a subscription model where the user receives "credits" that can be redeemed for content monthly and receive access to a curated on-demand library of content. Audible is the United States' largest audiobook producer and retailer. The service is owned by Audible, a wholly-owned subsidiary of Amazon.com, Inc., headquartered in Newark, New Jersey.

DON KATZ, is the founder and longtime CEO of Audible, Inc., which was founded in 1995 and headquartered in Newark, New Jersey. Which creates and sells audiobooks and other original content.

PROBLEMS
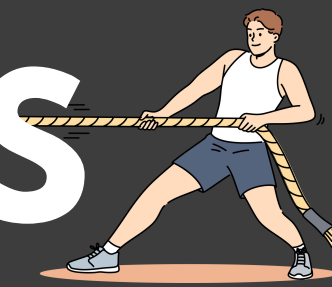
1. Standardize the name column to ensure consistent title casing.
2. Separate combined first and last names in the author column if they are currently combined.
3. Ensure all entries in the releasedate column follow a consistent date format (DD-MM-YYYY).
4. Convert the time column from text format to a duration format that Excel recognizes.
5. Ensure the price column is in a numeric format, and identify any non-numeric values.
6. Convert text ratings in the stars column to numeric values.
7. Split the narratedby column into multiple columns if multiple narrators are listed.
8. Merge the releasedate and language columns into a single new column named releaseinfo with the format "DD-MM-YYYY, Language."
9. Ensure all currency values in the price column are formatted consistently with two decimal places.

Standardize the name column to ensure consistent title casing.

Standardized the NAME COLUMN to ensure consistent title casing by using the CAPITAL EACH WORD function from FORMAT

	Home	Transform	Add Column	View										

Transpose — Data Type: Text — Replace Values — Unpivot Columns — Merge Columns
Reverse Rows — Detect Data Type — Fill — Move — Extract — Statistics Standard Scientific
Use First Row as Headers — Count Rows — Rename — Pivot Column — Convert to List — Format — Parse

Table — Any Column — Number Colum

Format dropdown:
- lowercase
- UPPERCASE
- Capitalize Each Word
- Trim
- Clean
- Add Prefix
- Add Suffix

`= Table.TransformColumnTypes(#"Promoted Headers",{{"name", type ... text}, {"narrator",`

	name	author	narrator	...me
1	Geronimo Stilton #11 & #12	Writtenby:GeronimoStilton	Narratedby:BillLobely	and 20 mins
2	The Burning Maze	Writtenby:RickRiordan	Narratedby:RobbieDa	s and 8 mins
3	The Deep End	Writtenby:JeffKinney	Narratedby:DanRusse	and 3 mins
4	Daughter of the Deep	Writtenby:RickRiordan	Narratedby:SoneelaNankani	11 hrs and 16 mins
5	The Lightning Thief: Percy Jackson, Book 1	Writtenby:RickRiordan	Narratedby:JesseBernstein	10 hrs
6	The Hunger Games: Special Edition	Writtenby:SuzanneCollins	Narratedby:TatianaMaslany	10 hrs and 35 mins
7	Quest for the Diamond Sword	Writtenby:WinterMorgan	Narratedby:LukeDaniels	2 hrs and 23 mins
8	The Dark Prophecy	Writtenby:RickRiordan	Narratedby:RobbieDaymond	12 hrs and 32 mins
9	Merlin Mission Collection	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	10 hrs and 56 mins
10	The Tyrant's Tomb	Writtenby:RickRiordan	Narratedby:RobbieDaymond	13 hrs and 22 mins
11	The Titan's Curse: Percy Jackson, Book 3	Writtenby:RickRiordan	Narratedby:JesseBernstein	8 hrs and 48 mins
12	Magic Tree House Collection: Books 9-16	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	5 hrs and 23 mins
13	Magic Tree House Collection: Books 1-8	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	6 hrs and 1 min
14	Magnus Chase and the Ship of the Dead	Writtenby:RickRiordan	Narratedby:MichaelCrouch	12 hrs and 58 mins
15	Northern Lights	Writtenby:PhilipPullman	Narratedby:PhilipPullman,fullcast,RuthWilson	11 hrs and 55 mins
16	Geronimo Stilton #13 and #14	Writtenby:GeronimoStilton	Narratedby:BillLobley	2 hrs and 25 mins
17	Magic Tree House Collection	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	5 hrs and 4 mins
18	Exile	Writtenby:ShannonMessenger	Narratedby:CaitlinKelly	14 hrs and 41 mins
19	Merlin Mission Collection	Writtenby:MaryPopeOsborne	Narratedby:MaryPopeOsborne	10 hrs and 18 mins
20	Neverseen	Writtenby:ShannonMessenger	Narratedby:CaitlinKelly	16 hrs and 42 mins
21	The Tower of Nero	Writtenby:RickRiordan	Narratedby:RobbieDaymond	12 hrs and 12 mins
22	Eldest: The Inheritance Cycle, Book 2	Writtenby:ChristopherPaolini	Narratedby:GerrardDoyle	23 hrs and 29 mins
23	Artemis Fowl	Writtenby:EoinColfer	Narratedby:GerryO'Brien	6 hrs and 46 mins

Separate combined first and last names in the author column if they are currently combined.

We separated the combined first name and last name columns of authors using these "APPLIED STEPS"

◢ APPLIED STEPS

Source	⚙
Promoted Headers	⚙
Changed Type	
Capitalized Each Word	
Extracted Text After Delimiter	⚙
Trimmed Text	
Split Column by Delimiter	⚙
Changed Type1	
Removed Columns	
Split Column by Character Transi...	
Merged Columns	⚙
Trimmed Text1	
Split Column by Character Transi...	
Merged Columns1	⚙
Split Column by Character Transi...	
Merged Columns2	⚙
Trimmed Text2	
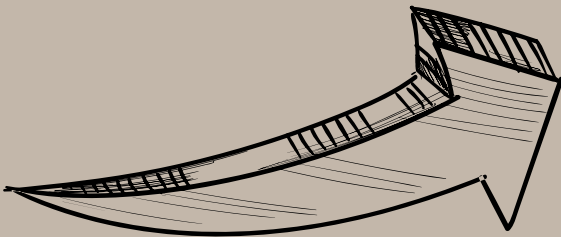
$^{A^B}_C$ Author-1 Full name ▼	$^{A^B}_C$ Auther-2 Full name ▼	$^{A^B}_C$ Auther-3 Full name ▼
Francisco Díaz Valladares		
Mattel		
Vegetta777	Willyrex	
Roshani Chokshi		
Sophie Schoenwald	Nadine Reitz-Illustrator	
Holly Rivers		
Leah Cypess		
Bertrand Fichou	Nora Thullin	Catherinede Lasa
Orianne Lallemand		
Adam Blade		
Jessica Khoury		
Leah Cypess		
Dav Pilkey		
Yoon Ha Lee		
Jonathan Stroud		

Ensure all entries in the releasedate column follow a consistent date format (DD-MM-YYYY).

We Ensured that all the entries in the release date column were in consistent date format (DD-MM-YYY).

	releasedate		
1.2	Decimal Number		
$	Currency		
1²₃	Whole Number		
%	Percentage		
	Date/Time		
	Date		
	Time		
	Date/Time/Timezone		
	Duration		
ᴬᴮ C	Text		
✗	True/False		
	Binary		
	Using Locale...		

24-06-2021

Convert the time column from text format
to a duration format that Excel recognizes.

Used custom column to
CONVERT TIME INTO HOURSE

Custom Column

Add a column that is computed from the other columns.

New column name

hrs

Custom column formula ⓘ

```
= if Text.Contains([time],"hr") then Number.FromText
(Text.BeforeDelimiter([time],"hr"))
    else 0
```

Available columns

name
Author-1 Full name
Auther-2 Full name
Auther-3 Full name
narrator
time
mins

<< In

Custom Column

Add a column that is computed from the other columns.

New column name

mins

Custom column formula ⓘ

```
= if Text.Contains([time],"minute") then 1

    else if Text.Contains([time],"min") and not Text.Contains(
    [time],"and") then Number.FromText(Text.BeforeDelimiter(
    [time],"min"))

    else if Text.Contains([time],"min") then
    Number.FromText(Text.BeforeDelimiter(Text.AfterDelimiter(
    [time],"and ")," min"))
    else 0
```

Available columns

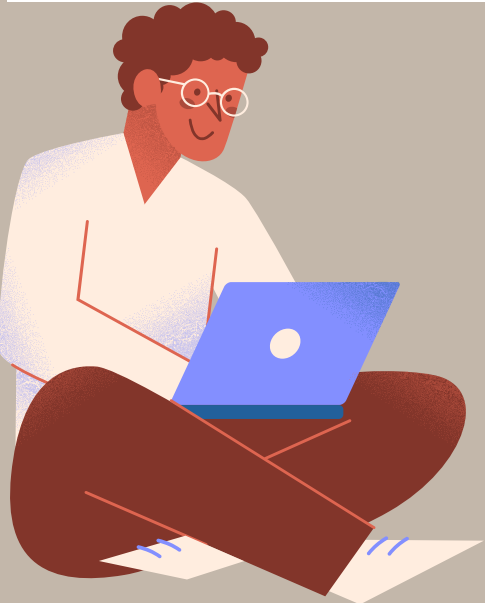name
Author-1 Full name
Auther-2 Full name
Auther-3 Full name
narrator
time
releasedate

<< In

converted the TIME COLUMN from text format to DURATION FORMAT by applying these steps

Added Custom ⚙
Reordered Columns
Added Custom1 ⚙
Reordered Columns1
Multiplied Column ⚙
Reordered Columns2
Inserted Addition ⚙
Reordered Columns3
Renamed Columns
✕ Removed Columns1

Ensure the price column is in a numeric format, and identify any non-numeric values.

There were non-numeric values in the price column after identifying them, we got some free values >for that we used the 'REPLACE VALUES' function and replaced "0" in the place of FREE value this is how non-numeric values are converted into numeric format.

A^B_C stars	A^B_C price
5 out of 5 stars34 ratings	468.00
4.5 out of 5 stars41 ratings	820.00
4.5 out of 5 stars38 ratings	410.00
4.5 out of 5 stars12 ratings	615.00
4.5 out of 5 stars181 ratings	820.00
5 out of 5 stars72 ratings	656.00
5 out of 5 stars11 ratings	233.00
5 out of 5 stars50 ratings	820.00
5 out of 5 stars5 ratings	1,256.00

PROPERTIES

Name

audible_uncleaned

All Properties

APPLIED STEPS

Source

Promoted Headers

✕ Replaced Value1

Changed Type

Replace Values

Replace one value with another in the selected columns.

Value To Find

Free

Replace With

0

▷ Advanced options

OK   Cancel

Convert text ratings in the stars column to numeric values. 🖍️

We used these applied steps and converted the "TEXT RATINGS" into "NUMERIC VALUES".

Split Column by Delimiter1	⚙
Changed Type2	
Replaced Value	⚙
Changed Type3	
Split Column by Character Tra...	
Removed Columns2	
Extracted Text Before Delimiter	⚙
Changed Type4	
✕ Renamed Columns1	
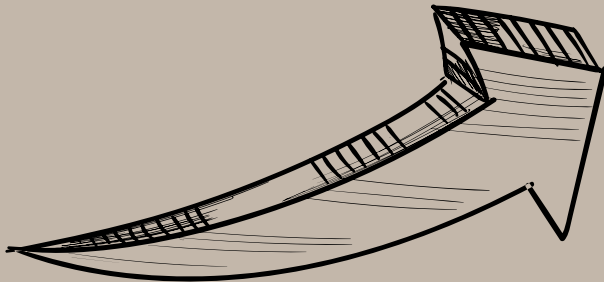
1.2 stars.1	1²3 no of people rated
5	34
4.5	41
4.5	38
4.5	12
4.5	181
5	72
5	11
5	50
5	5
5	58

Split the narratedby column into multiple columns if multiple narrators are listed.

There were multiple narrators in the list to split them we used these APPLIED STEPS given in the image.

Applied Steps	
Extracted Text After Delimiter1	⚙
Trimmed Text3	
Split Column by Delimiter2	⚙
Changed Type6	
Filtered Rows	
Removed Columns3	
Split Column by Character Transition4	
Merged Columns3	⚙
Trimmed Text4	
Split Column by Character Transition5	
Merged Columns4	⚙
Split Column by Character Transition6	
✕ Merged Columns5	⚙
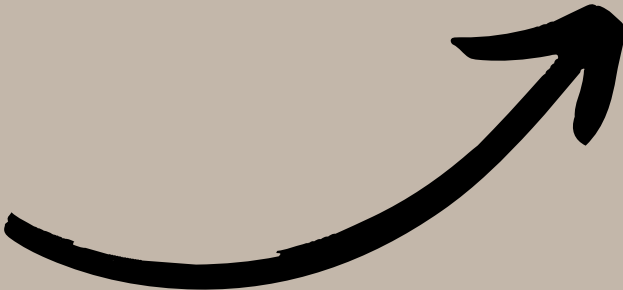
A^B_C narrator-1 full name ▼	A^B_C Narrator-2 full name ▼	A^B_C Narratoe-3 full name ▼
Bill Lobely		
Robbie Daymond		
Dan Russell		
Soneela Nankani		
Jesse Bernstein		
Tatiana Maslany		
Luke Daniels		
Robbie Daymond		
Mary Pope Osborne		
Robbie Daymond		
Jesse Bernstein		
Mary Pope Osborne		
Mary Pope Osborne		
Michael Crouch		
Philip Pullman	fullcast	Ruth Wilson
Bill Lobley		
Mary Pope Osborne		
Caitlin Kelly		

MERGE THE RELEASE-DATE AND LANGUAGE COLUMNS INTO A SINGLE NEW COLUMN NAMED RELEASE-INFO WITH THE FORMAT "DD-MM-YYYY, LANGUAGE."
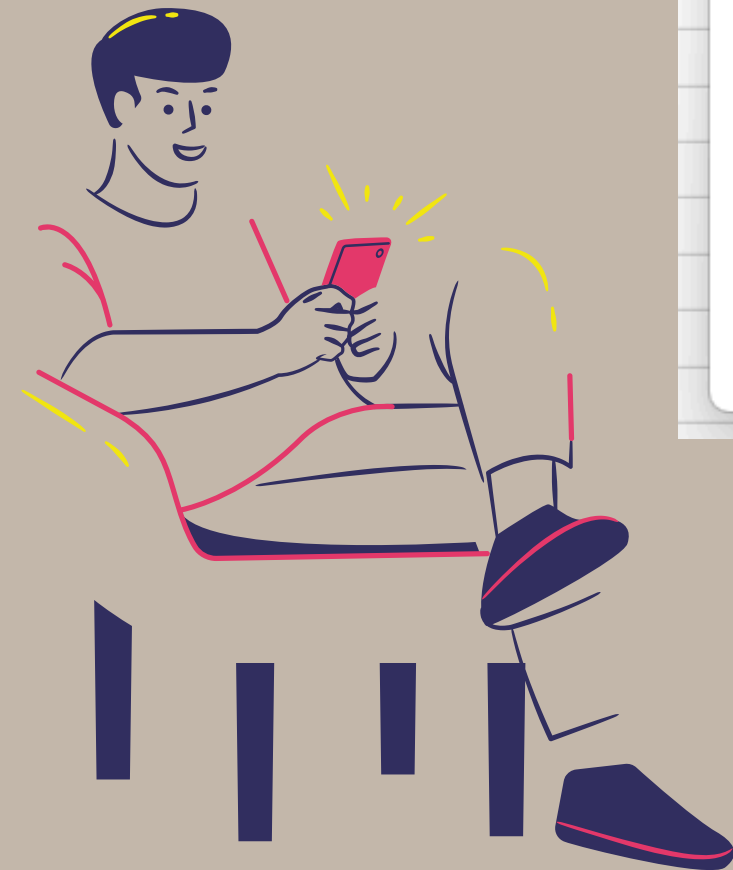
	releasinfo
468	04-08-2008, English
820	01-05-2018, English
410	06-11-2020, English
615	05-10-2021, English
820	13-01-2010, English
656	30-10-2018, English

PROP

Name

audib

All Pro

APPLI

Merge Columns

Choose how to merge the selected column

Separator

--Custom--

,

New column name (optional)

releasinfo

Using the merge column function we merged the date and language then Given the headers name as "RELEASE INFO".

releasinfo
04-08-2008, English
01-05-2018, English
06-11-2020, English
05-10-2021, English
13-01-2010, English
30-10-2018, English
25-11-2014, English
02-05-2017, English
02-05-2017, English
24-09-2019, English
14-01-2010, English
24-08-2011, English
27-09-2011, English

Ensure all currency values in the price column are formatted consistently with two decimal places.

We ensured that all the currency values in the price column were formatted in the two decimal places.

1.2	price	▼	ᴬᴮ𝐂
1.2	Decimal Number		
$	Currency		
1²₃	Whole Number		
%	Percentage		
	Date/Time		
	Date		
🕐	Time		
	Date/Time/Timezone		
🕐	Duration		
ᴬᴮ𝐂	Text		
✗	True/False		
	Binary		
	Using Locale...		

1093	24-(
467	08-(
1206	26-1

Thank you!

IF IT IS HELPFUL LIKE AND FOLLOW FOR MORE

LIKE   SHARE   FOLLOW

Click here to get the Link of CLEANED AUDIBLE DATASET.