



# Audible Data Cleaning Project

**Audible is an American online audiobook and podcast service that allows users to purchase and stream audiobooks and other forms of spoken word content.**

**Presented by Thomas Sangala**

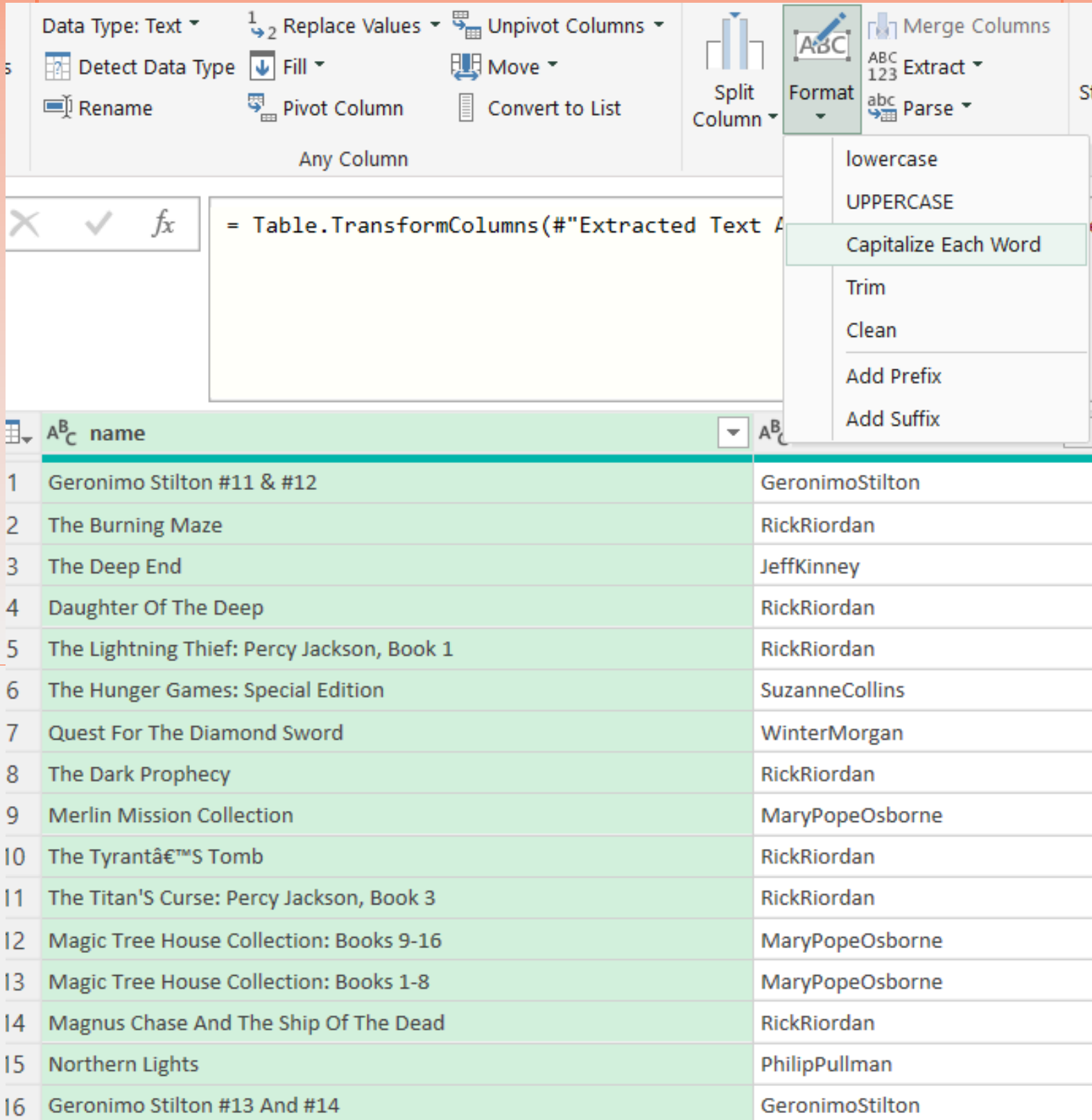


# 1. Standardize the name column to ensure consistent title casing

## Uncleaned Data

1	name
2	Blue Moon
3	The Suspicion
4	The Van Gogh Deception
5	Concealed
6	Ricky Ricotta'S Mighty Robot Vs. The Naughty Nightcrawlers From Neptune
7	Infinity Ring, Book 3: The Trap Door
8	Good Dog
9	Last Day On Mars
0	Dark Days
1	365 Days To Alaska
2	Everybunny Loves Magic
3	Better With Butter
4	The Doughnut King
5	Jennifer Lopez
6	The Greatest Treasure Hunt In History
7	The Middle School Rules Of Jamaal Charles
8	The Middle School Rules Of Charles Tillman: "Peanut"
9	Booker T. Washington
0	D-Day
1	Ulrich Zwingli: Shepherd Warrior

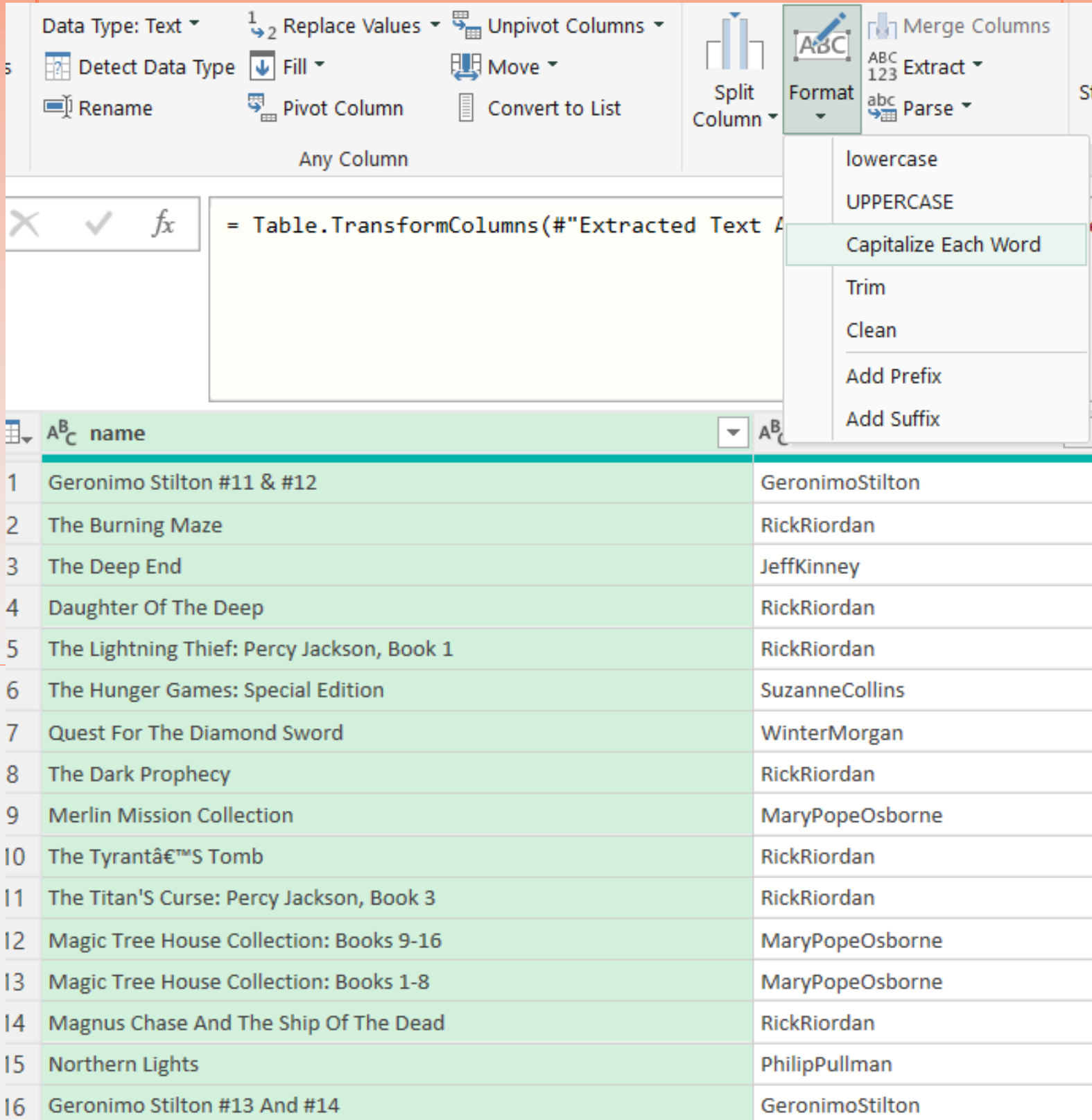
# Cleaned Data



The screenshot displays the Microsoft Excel interface. The 'Format' dropdown menu is open, showing options like 'lowercase', 'UPPERCASE', 'Capitalize Each Word' (which is highlighted), 'Trim', 'Clean', 'Add Prefix', and 'Add Suffix'. A large black arrow points to the 'Capitalize Each Word' option. The background shows a table with 16 rows of data, including book titles and authors.

	name	
1	Geronimo Stilton #11 & #12	GeronimoStilton
2	The Burning Maze	RickRiordan
3	The Deep End	JeffKinney
4	Daughter Of The Deep	RickRiordan
5	The Lightning Thief: Percy Jackson, Book 1	RickRiordan
6	The Hunger Games: Special Edition	SuzanneCollins
7	Quest For The Diamond Sword	WinterMorgan
8	The Dark Prophecy	RickRiordan
9	Merlin Mission Collection	MaryPopeOsborne
10	The Tyrant's Tomb	RickRiordan
11	The Titan's Curse: Percy Jackson, Book 3	RickRiordan
12	Magic Tree House Collection: Books 9-16	MaryPopeOsborne
13	Magic Tree House Collection: Books 1-8	MaryPopeOsborne
14	Magnus Chase And The Ship Of The Dead	RickRiordan
15	Northern Lights	PhilipPullman
16	Geronimo Stilton #13 And #14	GeronimoStilton

# Cleaned Data



The screenshot displays the Microsoft Excel interface. The 'Format' dropdown menu is open, showing options like 'lowercase', 'UPPERCASE', 'Capitalize Each Word' (which is highlighted), 'Trim', 'Clean', 'Add Prefix', and 'Add Suffix'. A large black arrow points to the 'Capitalize Each Word' option. The background shows a table with 16 rows of data, including book titles and authors.

	name	
1	Geronimo Stilton #11 & #12	GeronimoStilton
2	The Burning Maze	RickRiordan
3	The Deep End	JeffKinney
4	Daughter Of The Deep	RickRiordan
5	The Lightning Thief: Percy Jackson, Book 1	RickRiordan
6	The Hunger Games: Special Edition	SuzanneCollins
7	Quest For The Diamond Sword	WinterMorgan
8	The Dark Prophecy	RickRiordan
9	Merlin Mission Collection	MaryPopeOsborne
10	The Tyrant™'S Tomb	RickRiordan
11	The Titan'S Curse: Percy Jackson, Book 3	RickRiordan
12	Magic Tree House Collection: Books 9-16	MaryPopeOsborne
13	Magic Tree House Collection: Books 1-8	MaryPopeOsborne
14	Magnus Chase And The Ship Of The Dead	RickRiordan
15	Northern Lights	PhilipPullman
16	Geronimo Stilton #13 And #14	GeronimoStilton

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

# Uncleaned Data

author
Writtenby:GeronimoStilton
Writtenby:RickRiordan
Writtenby:JeffKinney
Writtenby:RickRiordan
Writtenby:RickRiordan
Writtenby:SuzanneCollins
Writtenby:WinterMorgan
Writtenby:RickRiordan
Writtenby:MaryPopeOsborne
Writtenby:RickRiordan
Writtenby:RickRiordan
Writtenby:MaryPopeOsborne
Writtenby:MaryPopeOsborne
Writtenby:RickRiordan
Writtenby:PhilipPullman
Writtenby:GeronimoStilton

# Applied Steps

Source	⚙
Promoted Headers	⚙
Changed Type	
Extracted Text After Delimi...	⚙
Extracted Text After Delimi...	⚙
Capitalized Each Word	
Split Column by Delimiter	⚙
Changed Type1	
Removed Columns	
Replaced Value	⚙
Replaced Value1	⚙
Split Column by Character ...	
Merged Columns	⚙
Renamed Columns	
Split Column by Character ...	
Trimmed Text	
Split Column by Character ...	
Merged Columns1	⚙
Trimmed Text1	
Split Column by Character ...	
Merged Columns2	⚙
Trimmed Text2	
Split Column by Delimiter1	⚙

# Cleaned Data








A <sup>B</sup> <sub>C</sub> author.1	A <sup>B</sup> <sub>C</sub> author.2	A <sup>B</sup> <sub>C</sub> author.3
Vegetta777	Willyrex	Not Applicable
Roshani Chokshi	Not Applicable	Not Applicable
Sophie Schoenwald	Nadine Reitz-Illustrator	Not Applicable
Holly Rivers	Not Applicable	Not Applicable
Leah Cypess	Not Applicable	Not Applicable
Bertrand Fichou	Nora Thullin	Catherinede Lase
Orianne Lallemand	Not Applicable	Not Applicable
Adam Blade	Not Applicable	Not Applicable
Jessica Khoury	Not Applicable	Not Applicable
Leah Cypess	Not Applicable	Not Applicable
Dav Pilkey	Not Applicable	Not Applicable
Yoon Ha Lee	Not Applicable	Not Applicable

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

#### Uncleaned Data

releasedate
4/8/2008
1/5/2018
6/11/2020
5/10/2021
13-01-10
30-10-18
25-11-14
2/5/2017
2/5/2017
24-09-19
14-01-10
24-08-11
27-09-11

#### Cleaned Data

releasedate		
1.2	Decimal Number	04-08-2008
\$	Currency	01-05-2018
1 <sup>2</sup> <sub>3</sub>	Whole Number	06-11-2020
%	Percentage	05-10-2021
	Date/Time	13-01-2010
	Date	30-10-2018
	Time	25-11-2014
	Date/Time/Timezone	02-05-2017
	Duration	02-05-2017
A <sup>B</sup> <sub>C</sub>	Text	24-09-2019
	True/False	14-01-2010
	Binary	24-08-2011
	Using Locale...	27-09-2011
		03-10-2017
		24-06-2021

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

### Uncleaned Data

time
2 hrs and 20 mins
13 hrs and 8 mins
2 hrs and 3 mins
11 hrs and 16 mins
10 hrs
10 hrs and 35 mins
2 hrs and 23 mins
12 hrs and 32 mins
10 hrs and 56 mins
13 hrs and 22 mins
8 hrs and 48 mins
5 hrs and 23 mins
6 hrs and 1 min
12 hrs and 58 mins
11 hrs and 55 mins

### Formula

#### Custom Column

Add a column that is computed from the other columns.

New column name

hours

Custom column formula ⓘ

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

# Formula

Custom column formula ⓘ

```
= 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
```

# Applied Steps

Split Column by Character ...

Merged Columns5



Replaced Value2



Trimmed Text5

Replaced Value3



✕ Added Custom



Reordered Columns

Multiplied Column



Renamed Columns1

Added Custom1



Inserted Addition



Renamed Columns2

Changed Type3

Replaced Value4



Changed Type4










# Cleaned Data

1.2 duration-mins
788
123
676
600
635
143
752
656
802
528
323
361
778

converted from text format to  
Duration format



duration-mins	
1.2	Decimal Number
\$	Currency
1 <sup>2</sup> <sub>3</sub>	Whole Number
%	Percentage
	Date/Time
	Date
	Time
	Date/Time/Timezone
	Duration
A <sup>B</sup> <sub>C</sub>	Text
	True/False
	Binary
	Using Locale...

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

Uncleaned Data

price
468
820
410
615
820
656
233
820
1,256.00
820
820
1,206.00
1,206.00
820
1,093.00
467
1,206.00

Clear Filter from

Filter By Color

Number Filter

Search

☒ 3,416.00

☒ 3,571.00

☒ 4,185.00

☒ 4,783.00

☒ 6,194.00

☒ 7,198.00

☒ Free

Cleaned Data

price	
820	13 h
410	2 hrs
615	11 h
820	10 h
656	10 h

Replace Values

Replace one value with another in the selected columns.

Value To Find

free

Replace With

0

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

Uncleaned Data

stars
5 out of 5 stars34 ratings
4.5 out of 5 stars41 ratings
4.5 out of 5 stars38 ratings
4.5 out of 5 stars12 ratings
4.5 out of 5 stars181 ratings
5 out of 5 stars72 ratings
5 out of 5 stars11 ratings
5 out of 5 stars50 ratings
5 out of 5 stars5 ratings
5 out of 5 stars58 ratings
4.5 out of 5 stars130 ratings
5 out of 5 stars6 ratings

# Formula

New column name

ratings

Custom column formula ⓘ

```
= if Text.Contains([stars],"ratings") then Number.FromText
  (Text.BeforeDelimiter(Text.AfterDelimiter([stars],"stars"),
    "ratings"))

else null
```

# Applied Steps

Changed Type5

✕ Added Custom2 ⚙

Extracted Text Before Deli... ⚙

Replaced Value5 ⚙

Replaced Value6 ⚙



# Cleaned Data

ABC stars	ABC ratings
5	34
4.5	41
4.5	38
4.5	12
4.5	181
5	72
5	11
5	50
5	5
5	58
4.5	130
5	6
5	7
5	41

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

Uncleaned Data

narrator
Narratedby:BillLobely
Narratedby:RobbieDaymond
Narratedby:DanRussell
Narratedby:SoneelaNankani
Narratedby:JesseBernstein
Narratedby:TatianaMaslany
Narratedby:LukeDaniels
Narratedby:RobbieDaymond
Narratedby:MaryPopeOsborne
Narratedby:RobbieDaymond
Narratedby:JesseBernstein
Narratedby:MaryPopeOsborne
Narratedby:MaryPopeOsborne
Narratedby:MichaelCrouch
Narratedby:PhilipPullman,fullcast,RuthWilson

# Applied Steps

## APPLIED STEPS

- ✕ Split Column by Delimiter1 ⚙
- Changed Type2
- Removed Columns1
- Trimmed Text3
- Split Column by Character ...
- Split Column by Character ...
- Merged Columns3 ⚙
- Merged Columns4 ⚙
- Trimmed Text4
- Split Column by Character ...
- Merged Columns5 ⚙
- Replaced Value2 ⚙
- Trimmed Text5
- Replaced Value3 ⚙
- Added Custom ⚙

## Replace Values

Replace one value with another in the selected column

Value To Find

null

Replace With

Not Applicable

> Advanced options

# Cleaned Data

<div><div><div>A<sup>B</sup><sub>C</sub></div></div>narrator.1</div> <div>▼</div>	<div><div><div>A<sup>B</sup><sub>C</sub></div></div>narrator.2</div> <div>▼</div>	<div><div><div>A<sup>B</sup><sub>C</sub></div></div>narrator.3</div>
Shannon Mc Manus	Not Applicable	Not Applicable
Imelda Staunton	Not Applicable	Not Applicable
Mary Pope Osborne	Not Applicable	Not Applicable
E.B.Stevens	Not Applicable	Not Applicable
Caitlin Kelly	Not Applicable	Not Applicable
Luke Daniels	Not Applicable	Not Applicable
Len Forgione	AshtonSundholm	Jaden Rogers
Michael Goldstrom	Not Applicable	Not Applicable
Michael Goldstrom	Not Applicable	Not Applicable
Stacy Gonzalez	Not Applicable	Not Applicable
Kristin Atherton	Not Applicable	Not Applicable
Derek Steel	Not Applicable	Not Applicable
Shannon Mc Manus	Not Applicable	Not Applicable
Shannon Mc Manus	Not Applicable	Not Applicable
David Pittu	Not Applicable	Not Applicable
Michael Goldstrom	Not Applicable	Not Applicable
Tim Gregory	Not Applicable	Not Applicable



8. Merge the releasedate and language columns into a single new column named releaseinfo with the format "DD-MM-YYYY, Language."

## Uncleaned Data

releasedate	language
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
03-10-2017	English
24-06-2021	English
08-02-2008	English

# Steps

releasedate	language
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

## Merge Columns

Choose how to merge the selected columns.

Separator

Comma

New column name (optional)

release info



# Cleaned Data








release info
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
03-10-2017,English
24-06-2021,English
08-02-2008,English
26-12-2004,English
06-11-2018,English

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

# Uncleaned Data

123 price
468
820
410
615
820
656
233
820
1256
820
820
1206
1206
820

# Steps

123 price		
1.2	Decimal Number	468
\$	Currency	820
123	Whole Number	410
%	Percentage	615
	Date/Time	820
	Date	656
	Time	233
	Date/Time/Timezone	820
	Duration	1256
A <sup>B</sup> <sub>C</sub>	Text	820
	True/False	820
	Binary	1206
	Using Locale...	820
		1093
		467



# Cleaned Data

\$ price	
	468.00
	820.00
	410.00
	615.00
	820.00
	656.00
	233.00
	820.00
	1,256.00
	820.00
	820.00
	1,206.00
	1,206.00
	820.00
	1,093.00
	467.00
	1,206.00
	836.00

# Thank You



[Follow me on LinkedIn](#)



**[Drive file link](#)**