WSCube tech

Cohort 5

Week 2: Case Study

Audible Data Cleaning Project

Binay Kumar Naik

24-04-2025

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

PROJECT TASK:

- Utilize Power Query Editor in Excel to clean and standardize an Audible dataset. Tasks include formatting columns, ensuring data consistency, and preparing the dataset for analysis.

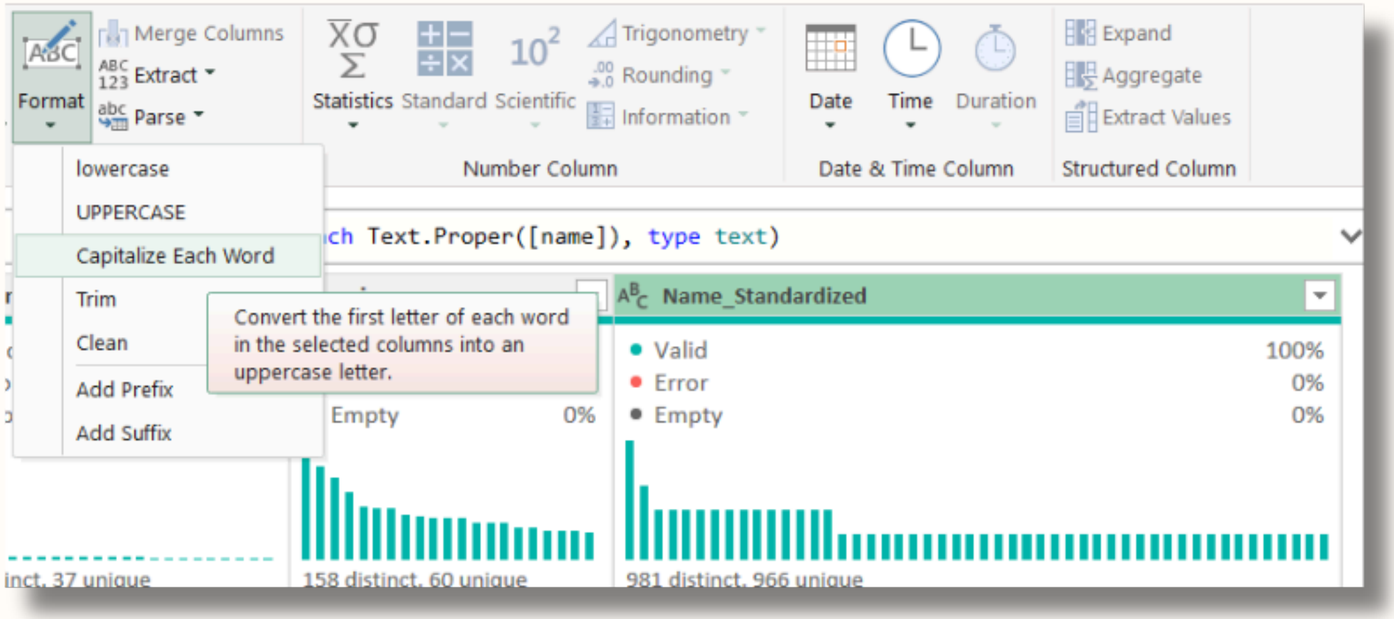
DATA CLEANING TASKS

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

DATA CLEANING TASKS

TASK 1

- **Standardize the name column to ensure consistent title casing.**

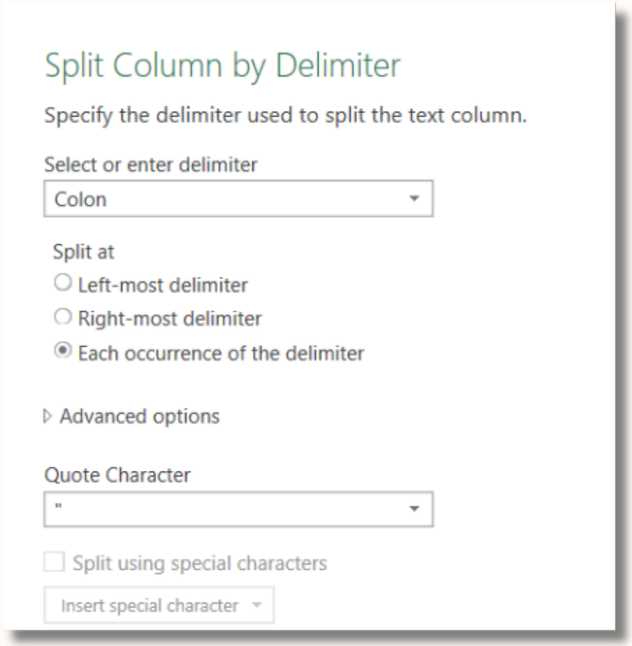


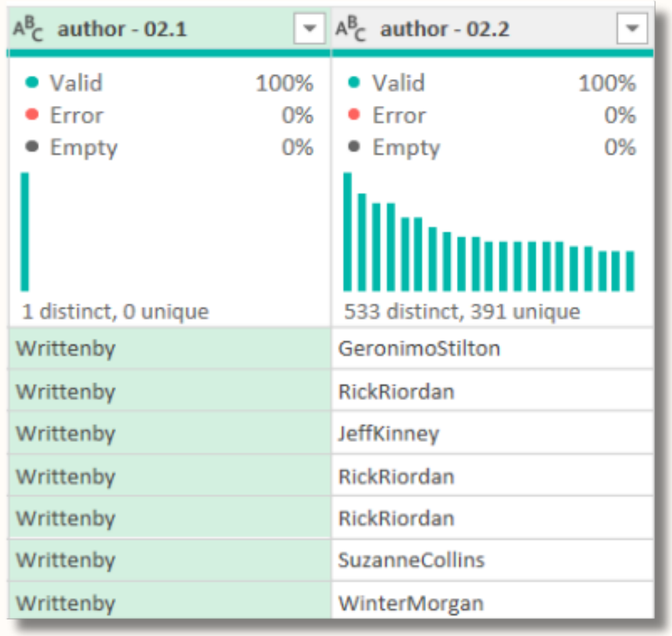
- By using format col. I have capitalized each word
- The new “**Name_Standardized**” col. has proper title casing

DATA CLEANING TASKS

TASK 2

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





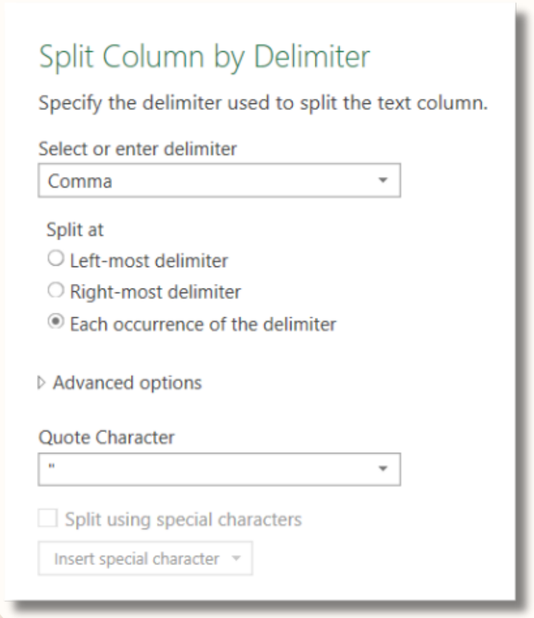
author - 02.1	author - 02.2
Valid 100%	Valid 100%
Error 0%	Error 0%
Empty 0%	Empty 0%
1 distinct, 0 unique	533 distinct, 391 unique
Writtenby	GeronimoStilton
Writtenby	RickRiordan
Writtenby	JeffKinney
Writtenby	RickRiordan
Writtenby	RickRiordan
Writtenby	SuzanneCollins
Writtenby	WinterMorgan

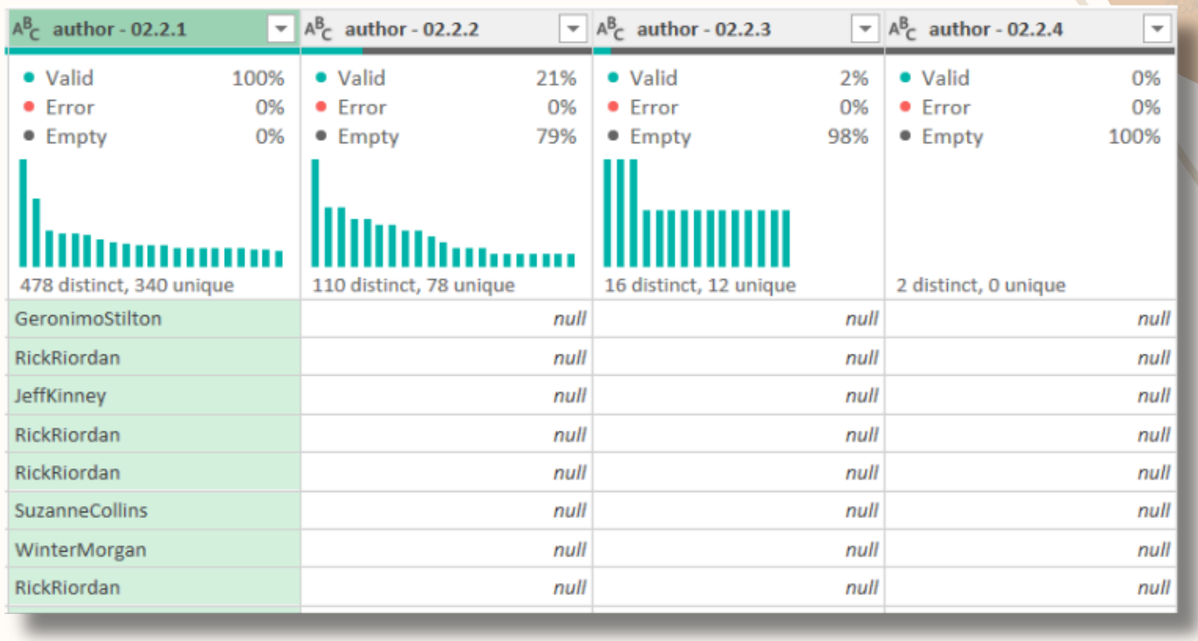
- **Step 1**: First I split the author col. using a colon delimiter.
- The author col. was then split into two halves, one with the text "**Writtenby**" and the other col. with the authors.

DATA CLEANING TASKS

TASK 2

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





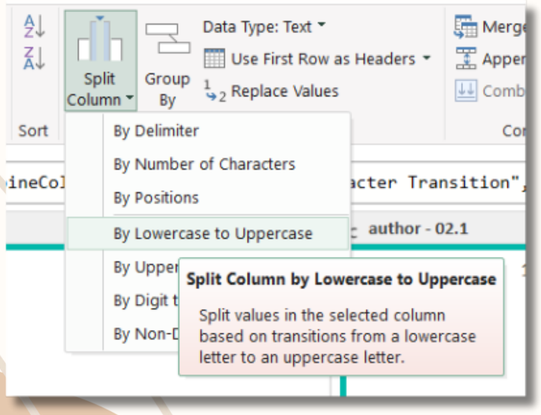
author - 02.2.1	author - 02.2.2	author - 02.2.3	author - 02.2.4
Valid 100%	Valid 21%	Valid 2%	Valid 0%
Error 0%	Error 0%	Error 0%	Error 0%
Empty 0%	Empty 79%	Empty 98%	Empty 100%
478 distinct, 340 unique	110 distinct, 78 unique	16 distinct, 12 unique	2 distinct, 0 unique
GeronimoStilton	null	null	null
RickRiordan	null	null	null
JeffKinney	null	null	null
RickRiordan	null	null	null
RickRiordan	null	null	null
SuzanneCollins	null	null	null
WinterMorgan	null	null	null
RickRiordan	null	null	null

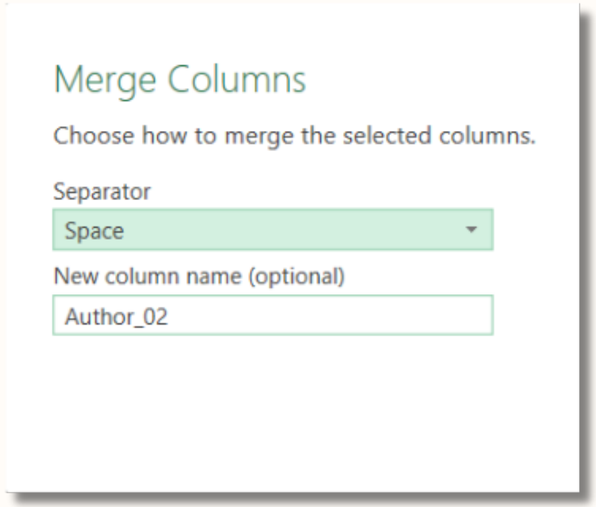
- **Step 2**: Then I split the **author 02.2** col. using a coma delimiter.
- The **author 02.2** col. was then split into three halves, with each different authors in each different col.

DATA CLEANING TASKS

TASK 2

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





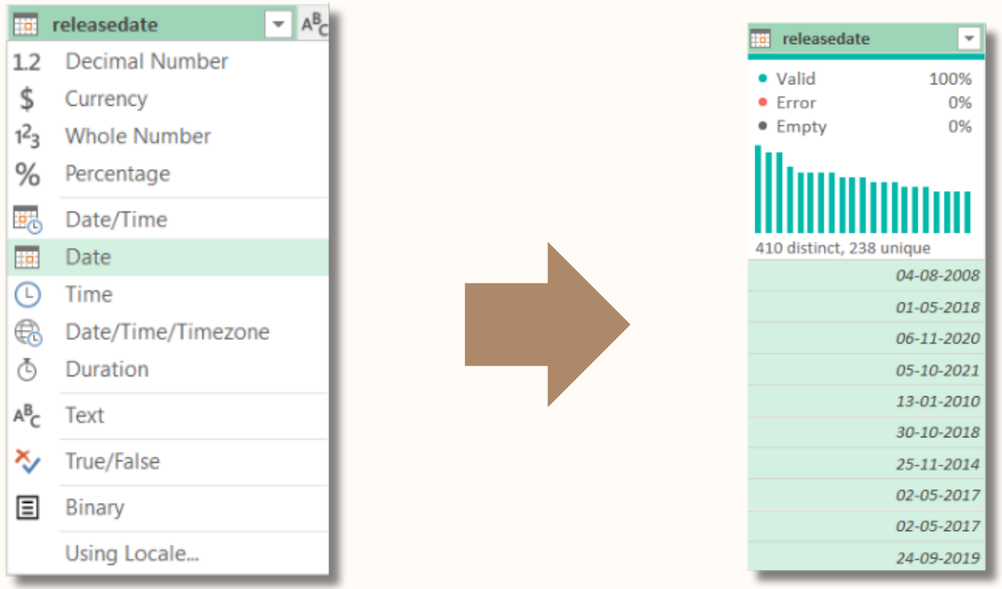
Author_01	Author_02	Author_03
Valid 100%	Valid 100%	Valid 100%
Error 0%	Error 0%	Error 0%
Empty 0%	Empty 0%	Empty 0%
1 distinct, 0 unique	1 distinct, 0 unique	1 distinct, 0 unique
Bertrand Fichou	Nora Thullin	Catherinede Lasa
M.G.Leonard	Sam Sedgman	Elisa Paganelli-illustrator
Mac Kenzie Cadenhead	Sean Ryan	Anke Albrecht-Übersetzer
Jaume Cabré	Queralt Armengol-ilustrador	Concha Cardeñoso-traductor
Jaume Cabré	Romina Martí-ilustrador	Concha Cardeñoso-traductor
Mac Kenzie Cadenhead	Sean Ryan	Anke Albrecht-Übersetzer
Sofie Forsman	Ida Kjellin	Charlotta Borelius

- **Step 3**: Then I split each author col. using a delimiter lowercase to uppercase. We split the First and last name by using this delimeiter
- The split author col. was then merged to form proper author names

DATA CLEANING TASKS

TASK 3

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



- From the drop-down menu of the col. I made sure the "**releasedate**" col was in (DD-MM-YYYY) format

DATA CLEANING TASKS

TASK 4

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

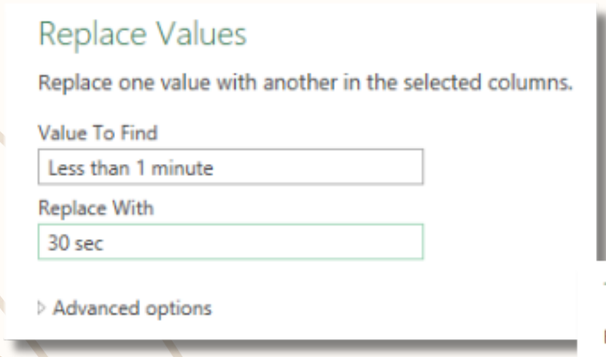
time - 02	time - 02 - Copy.1	time - 02 - Copy.2
Valid 100%	Valid 100%	Valid 67%
Error 0%	Error 0%	Error 0%
Empty 0%	Empty 0%	Empty 33%
455 distinct, 234 unique	82 distinct, 10 unique	60 distinct, 0 unique
2 hrs and 20 mins	2 hrs	20 mins
13 hrs and 8 mins	13 hrs	8 mins
2 hrs and 3 mins	2 hrs	3 mins
11 hrs and 16 mins	11 hrs	16 mins
10 hrs	10 hrs	*null*
10 hrs and 35 mins	10 hrs	35 mins
2 hrs and 23 mins	2 hrs	23 mins
12 hrs and 32 mins	12 hrs	32 mins
10 hrs and 56 mins	10 hrs	56 mins
13 hrs and 22 mins	13 hrs	22 mins

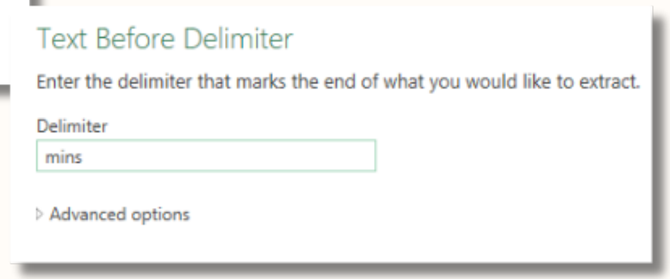
- **Step 1** : First, we try to separate both the **hour** and **minute** from the **time** col.

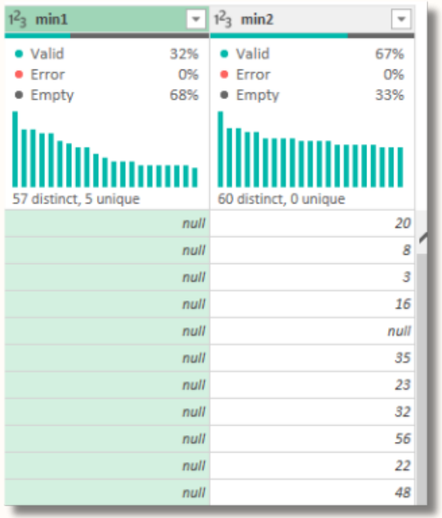
DATA CLEANING TASKS

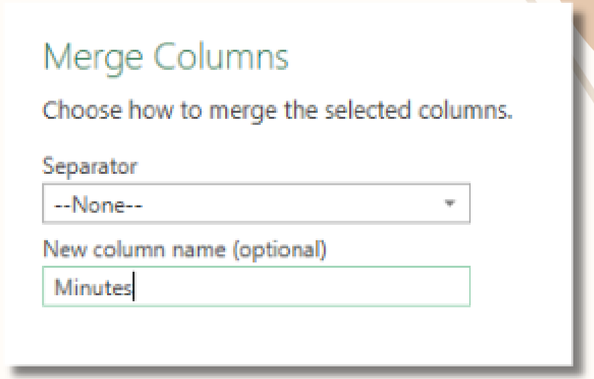
TASK 4

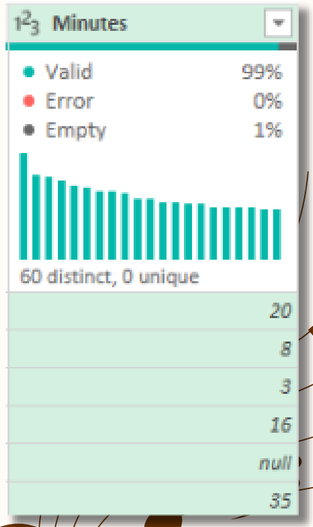
- **Convert the time column from text format to a duration format that Excel recognizes.**









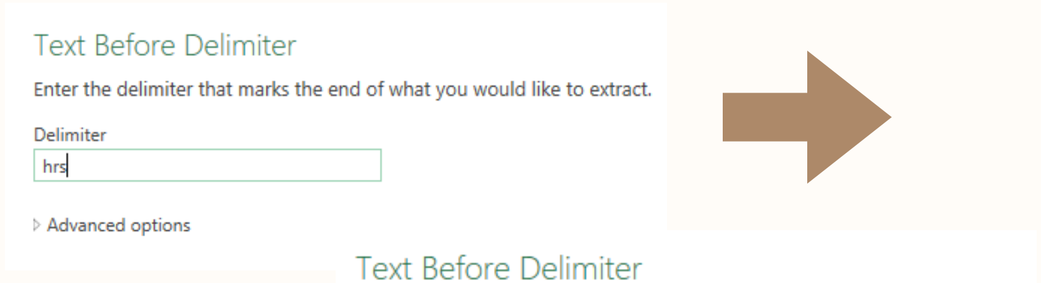


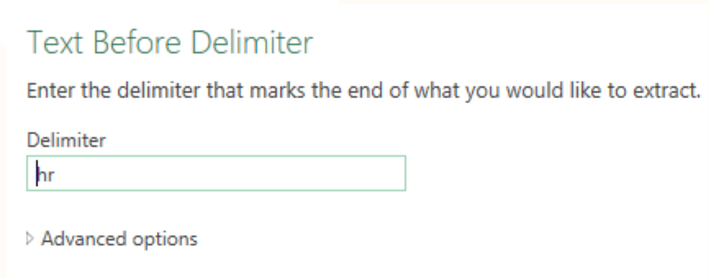
- **Step 2**: Replace the "**Less than 1 minute**" text to "**30 sec**".
- Extracted the "mins" and "min" from the hour col. using text before delimeter then converted the entire col. to the whole number.
- All the "**hr**" and "**hrs**" change into error which is then replaced with 0 then to null.
- Finally merge both "**min 1**" and "**min 2**" col to form minutes

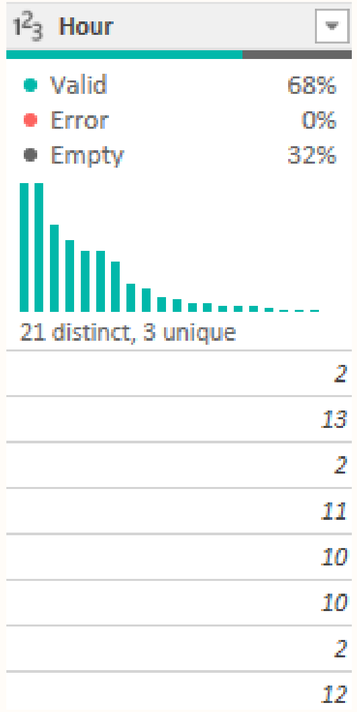
DATA CLEANING TASKS

TASK 4

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





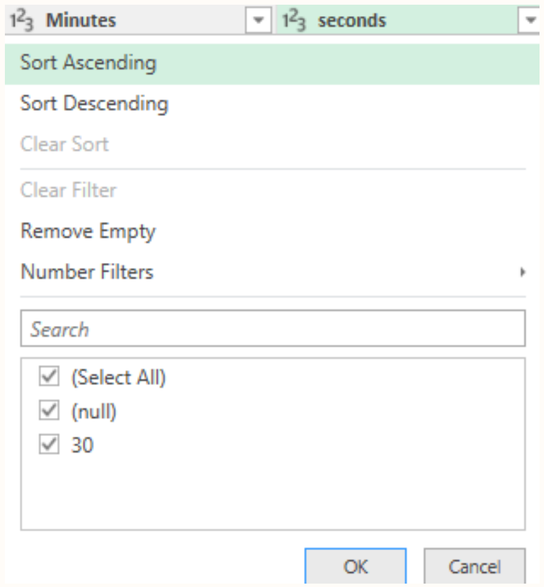


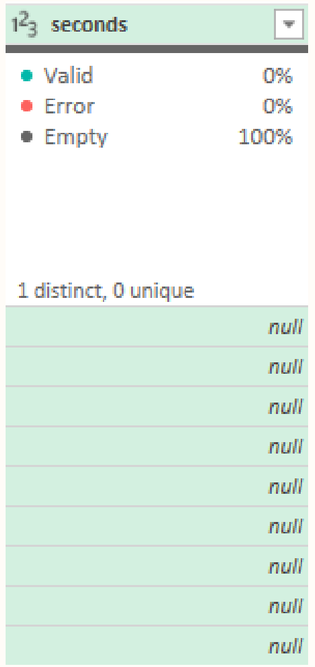
- **Step 3**: In the splitted col with hr I extracted the "**hr**" and "**hrs**". Then converted into whole number found only the hour. There were some errors which were then converted to 0 then to null.

DATA CLEANING TASKS

TASK 4

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



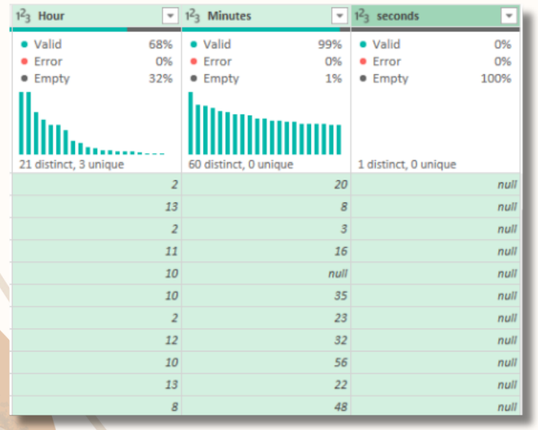


- **Step 4**: The converted “Less than 1 minute” to 30 sec is separated to one specific col.

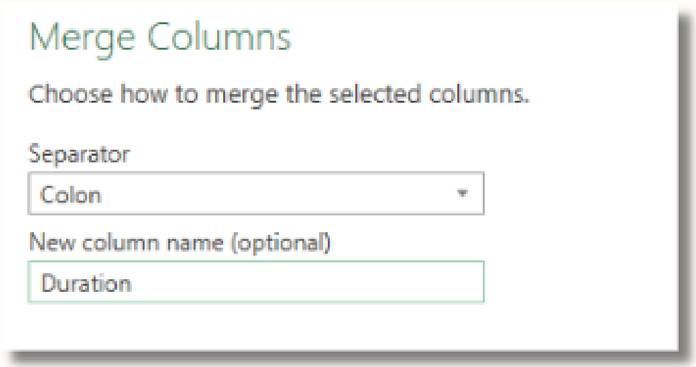
DATA CLEANING TASKS

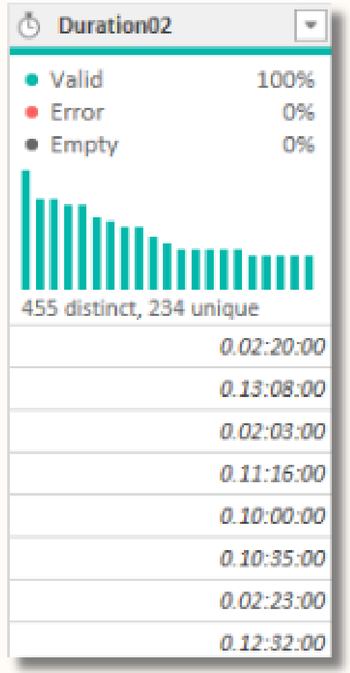
TASK 4

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



Hour	Minutes	seconds
Valid 68%	Valid 99%	Valid 0%
Error 0%	Error 0%	Error 0%
Empty 32%	Empty 1%	Empty 100%
21 distinct, 3 unique	60 distinct, 0 unique	1 distinct, 0 unique
2	20	null
13	8	null
2	3	null
11	16	null
10	null	null
10	35	null
2	23	null
12	32	null
10	56	null
13	22	null
8	48	null





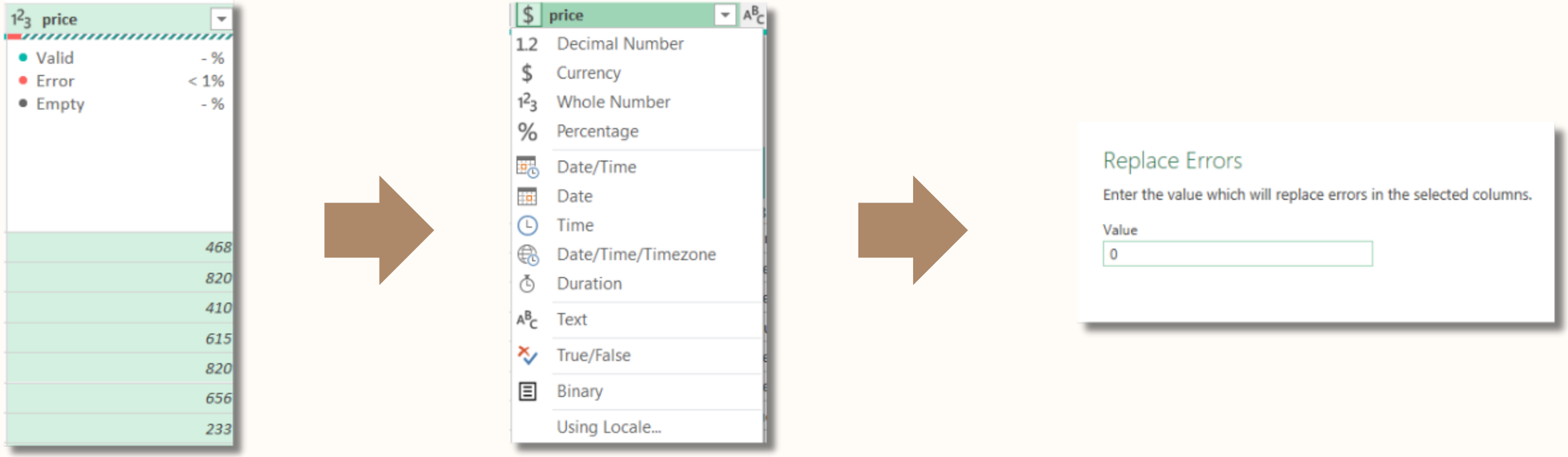
Duration02
Valid 100%
Error 0%
Empty 0%
455 distinct, 234 unique
0.02:20:00
0.13:08:00
0.02:03:00
0.11:16:00
0.10:00:00
0.10:35:00
0.02:23:00
0.12:32:00

- **Step 5**: Merge the 3 col. “**Hour**”, “**Minutes**”, and “**Seconds**” to form col. “**Duration**”

DATA CLEANING TASKS

TASK 5

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

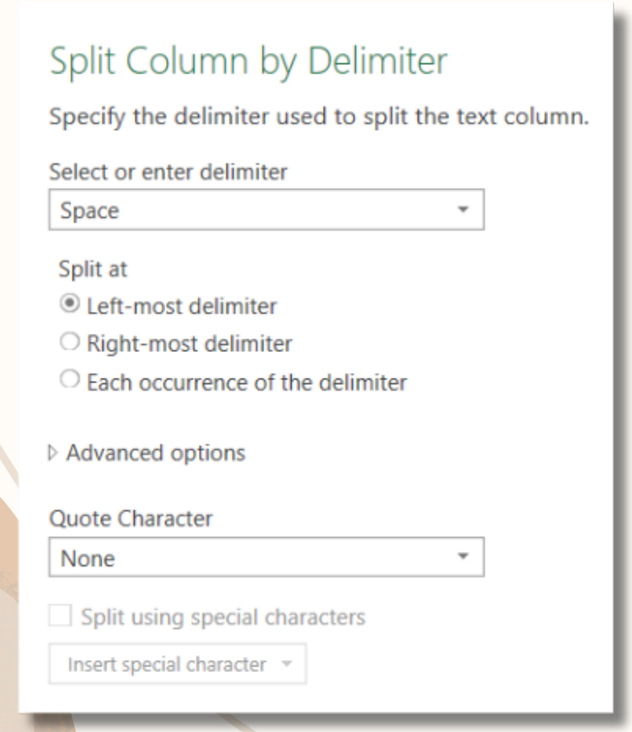


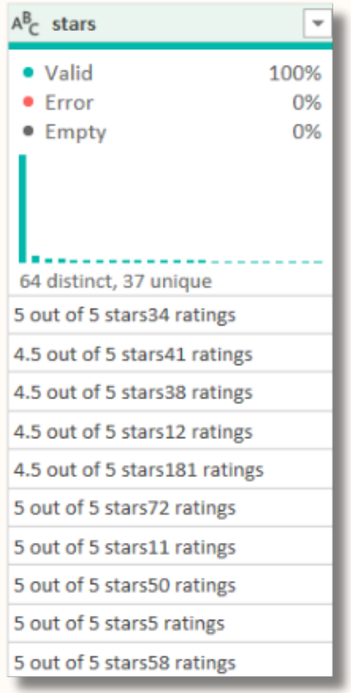
- From the drop-down menu of the col. I made sure the "**price**" col was in numeric format
- Found error values like the text "**Free**" which were then replaced with 0.

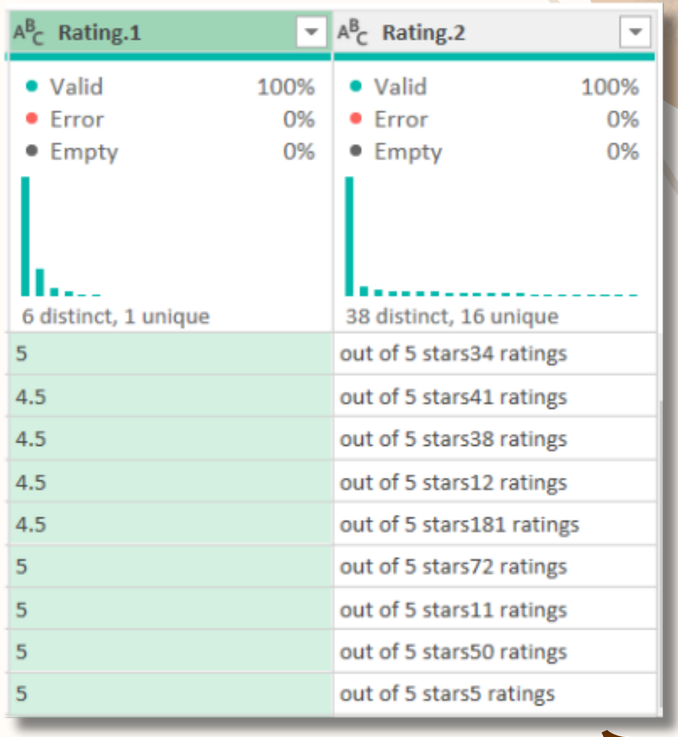
DATA CLEANING TASKS

TASK 6

- **Convert text ratings in the stars column to numeric values.**





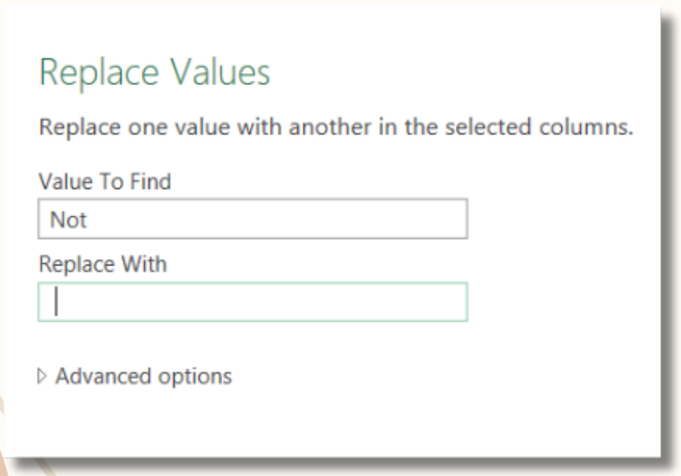


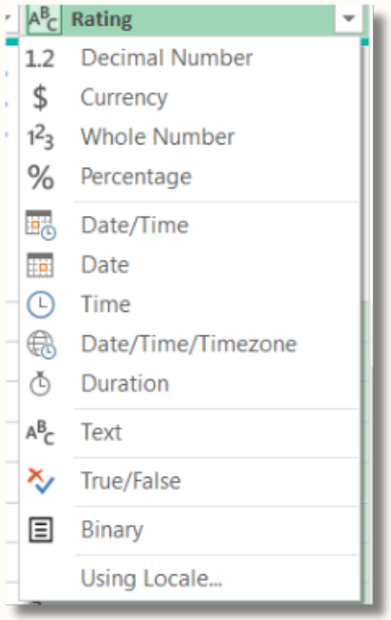
- **Step 1**: First I split the "**stars**" col. using a space delimiter.
- The "**stars**" col. was then split into two halves, one with the text "**Rating.1**" and the other "**Rating.2**" with the authors.

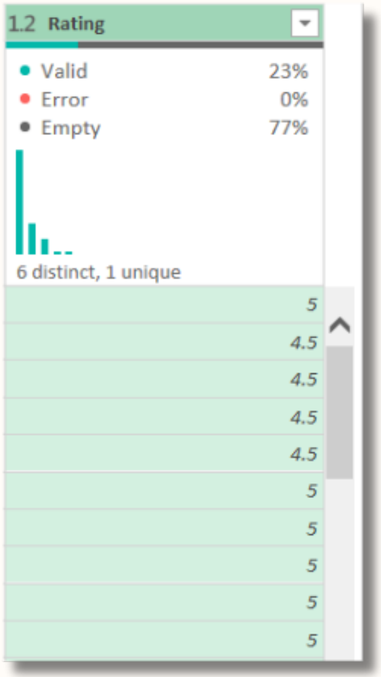
DATA CLEANING TASKS

TASK 6

- **Convert text ratings in the stars column to numeric values.**





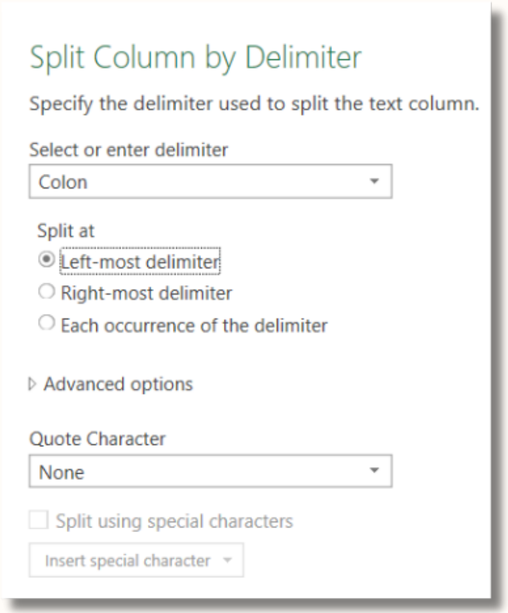


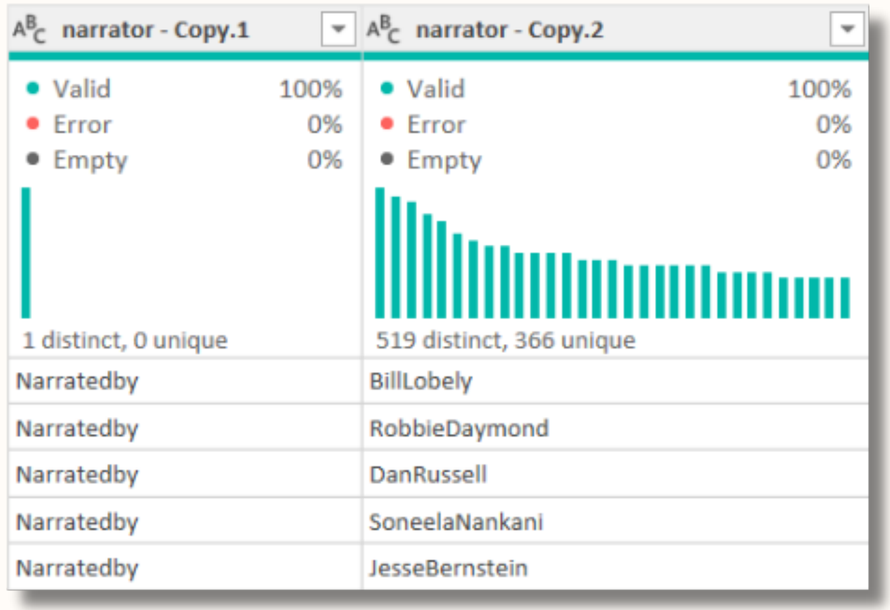
- **Step 2**: Converting the “Not yet rated values” to null.
- **Step 3**: Converted the entire col into numeric values.

DATA CLEANING TASKS

TASK 7

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



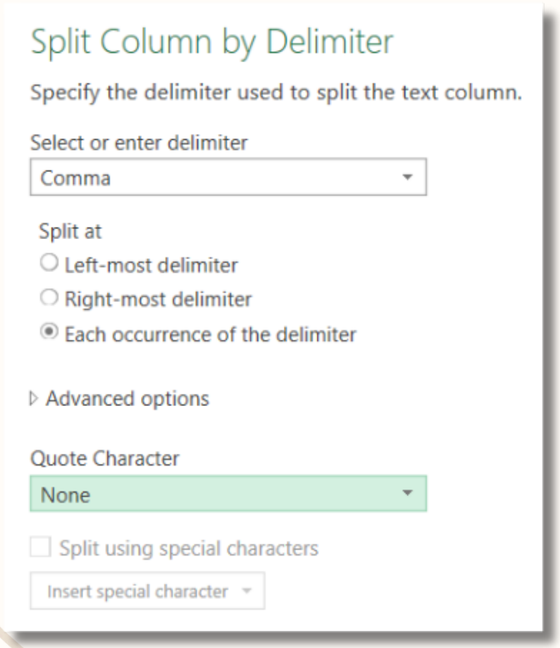


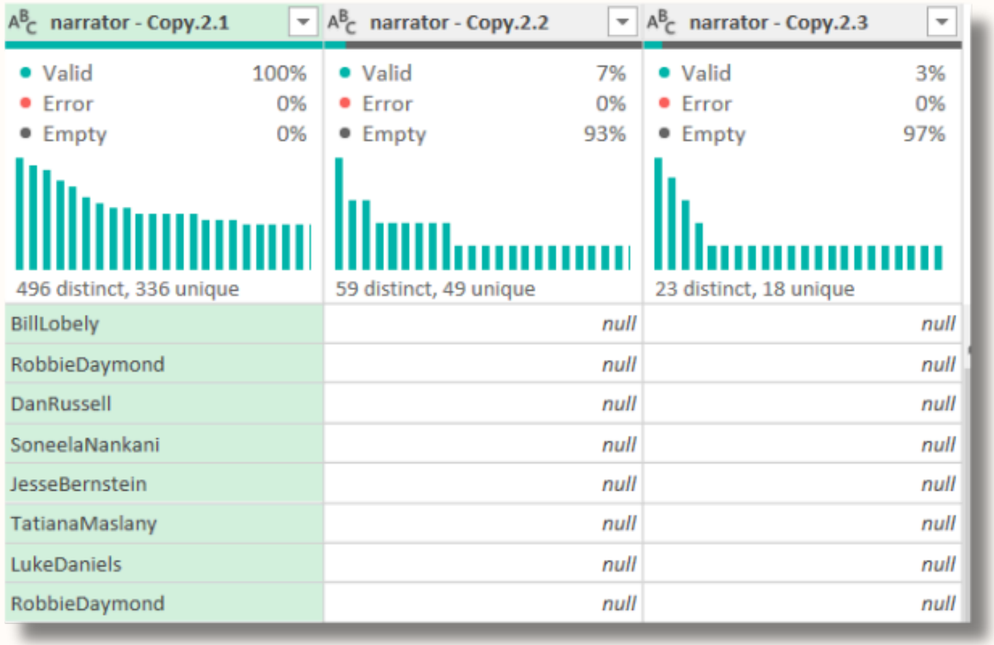
- **Step 1**: First I split the author col. using a colon delimiter.
- The narrator col. was then split into two halves, one with the text "**Narratedby**" and the other col. with the **narrators**.

DATA CLEANING TASKS

TASK 7

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





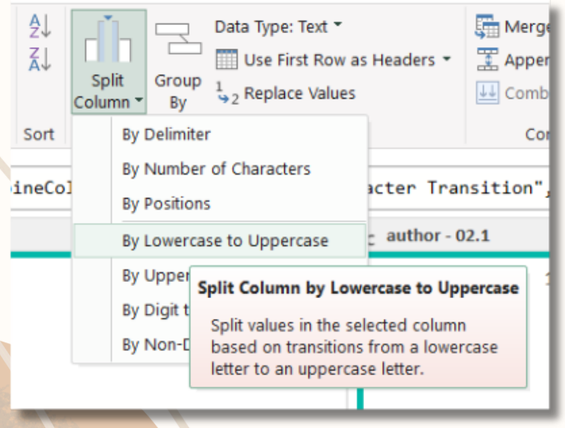
narrator - Copy.2.1	narrator - Copy.2.2	narrator - Copy.2.3
Valid 100%	Valid 7%	Valid 3%
Error 0%	Error 0%	Error 0%
Empty 0%	Empty 93%	Empty 97%
496 distinct, 336 unique	59 distinct, 49 unique	23 distinct, 18 unique
BillLobely	null	null
RobbieDaymond	null	null
DanRussell	null	null
SoneelaNankani	null	null
JesseBernstein	null	null
TatianaMaslany	null	null
LukeDaniels	null	null
RobbieDaymond	null	null

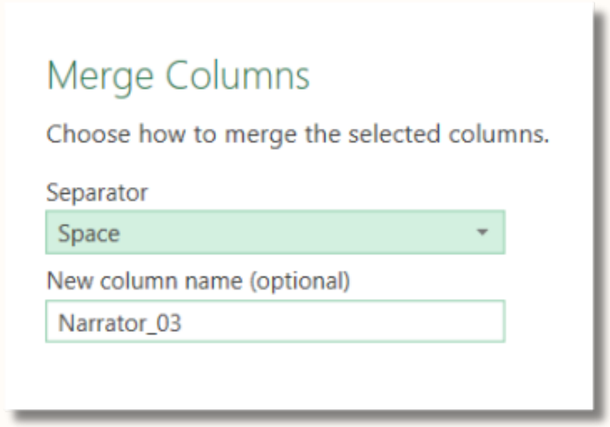
- **Step 2**: Then I split the “**narrator 02.1"** col. using a coma delimiter.
- The “**narrator 02.1"** col. was then split into three halves, with each different narrator in each different col.

DATA CLEANING TASKS

TASK 1

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





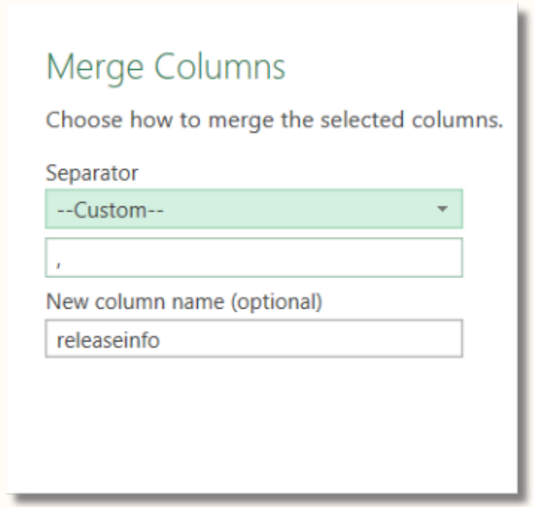
Narrator_01	Narrator_02	Narrator_03
Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
7 distinct, 6 unique	8 distinct, 6 unique	6 distinct, 2 unique
Len Forgione	Winston Bromhead	I'ke Mitchell
Katie Leung	Rebecca Lee	Alexander Capon
Len Forgione	Winston Bromhead	I'ke Mitchell
Len Forgione	Winston Bromhead	I'ke Mitchell
Len Forgione	Dazjon Freeman	Ben D'Amico
Jessica Almasy	Keylor Leigh	Barrett Leddy
Len Forgione	Dazjon Freeman	Ben D'Amico
Len Forgione	Dazjon Freeman	Ben D'Amico
Michael Stauffer	Andri Perl	Amina Abdulkadir

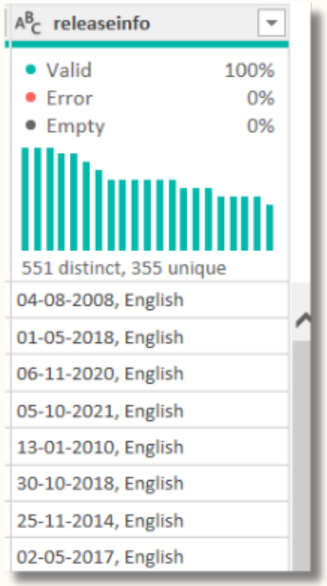
- **Step 3**: Then I split each **narrator** col. using a delimiter lowercase to uppercase. We split the **First** and **last names** by using this delimeiter.
- The split author col. was then merged to form proper **narrator names**

DATA CLEANING TASKS

TASK 8

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



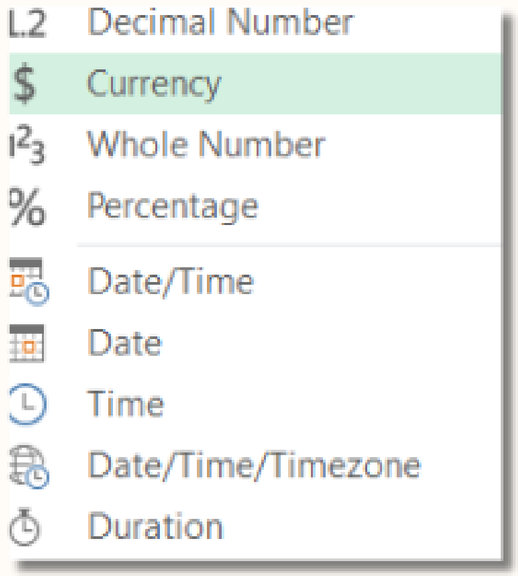


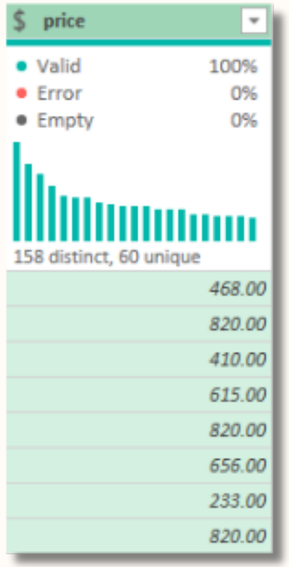
- Merged the **releasedate** and **language** col into a single col to form a single col **releaseinfo**.
- The **releaseinfo** col show date in (DD-MM-YYYY) format.

DATA CLEANING TASKS

TASK 9

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





- From the drop-down menu, it is ensured that the currency col is formatted to two decimal places.
- There were some rows with error which were then replaced with “0”

Thank you!

You can check my **github** link for the detailed projrct:

https://github.com/Binay005X/Audible-Data-Cleaning