Audible Data Cleaning Project

- Case Study Description:

- 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.

- Standardize the name column to ensure consistent title casing.

- Separate combined names in the author column if there are multiple authors.

- 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."

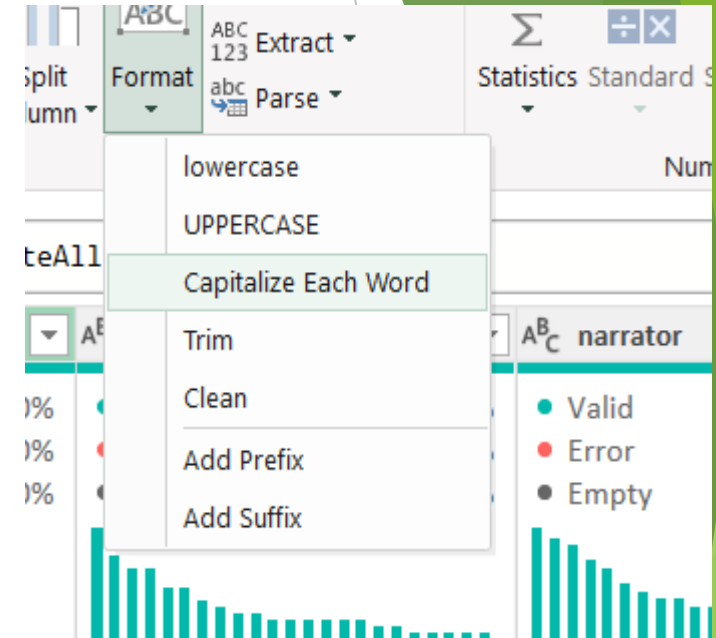- Ensure all currency values in the price column are formatted consistently with two decimal places.

Standardize the name column to ensure consistent title casing.

Load data into Power Query.

Select the "Name" column.

Go to the Transform tab.

Click Format → choose Capitalize Each Word.

Separate combined names in the author column if there are multiple authors.

1. Load data into Power Query.

2. Select the Author column.

3. On the Home or Transform tab:Click Split Column → By DelimiterChoose Comma (",") as the delimiter Choose Each occurrence of the delimiter Choose Split into columns

4. 4. Power Query will auto-create:Author.1, Author.2, Author.3, etc.

Split Column by Delimiter

Specify the delimiter used to split the text column.

Select or enter delimiter

Comma ▾

Split at

○ Left-most delimiter

○ Right-most delimiter

◉ Each occurrence of the delimiter

▷ Advanced options

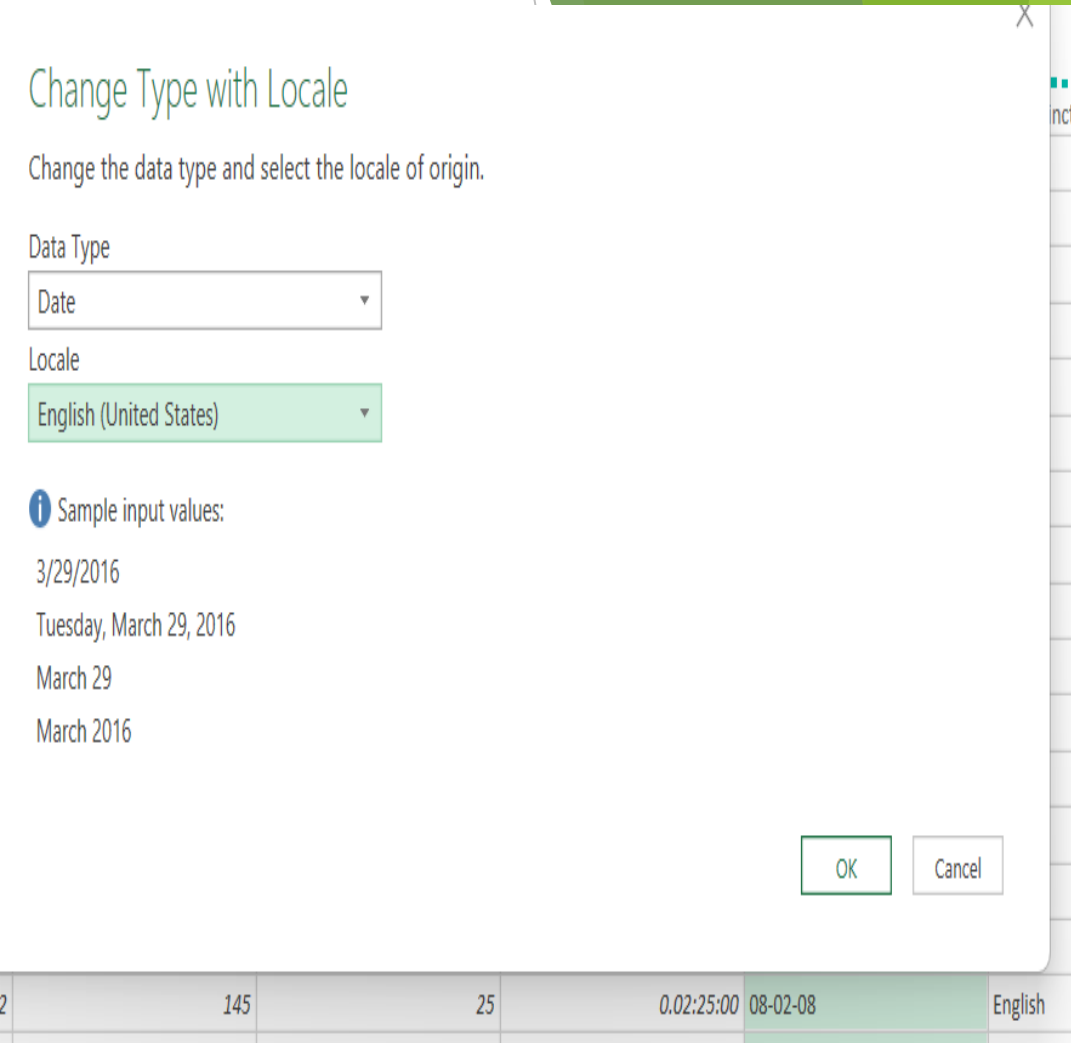Quote Character

" ▾

☐ Split using special characters

Insert special character ▾

OK   Cancel

X

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

- Select releasedate
- Go to Transform → Data Type → Using Locale
- Choose:Data Type: Date
- Locale: English (United State)

Change Type with Locale

Change the data type and select the locale of origin.

Data Type

Date

Locale

English (United States)

ⓘ Sample input values:

3/29/2016

Tuesday, March 29, 2016

March 29

March 2016

OK   Cancel

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

▶ Step 1: Select Time (Clean) Go to Transform → Extract → Text Before Delimiter Use delimiter: hr Result: If the value was "4 hrs and 21 mins", get "4 ".Rename this to Hours

▶ Step 2: Select the time column Go to:Transform → Extract → Text Between Delimiters

▶ Set the delimiters: Start delimiter: min

▶ End delimiter: and

▶ Click Advanced options Scan for the start delimiter: From the end of the input

▶ Scan for the end delimiter: From the start delimiter to the start of input

▶ This reverses the search:It finds "min" from the end, and then looks backward to find "and" — effectively isolating the minutes value that appears before "mins".

Akshay Solanki

5

Text Before Delimiter

Enter the delimiter that marks the end of what you would like to extract.

Delimiter

hr

▲ Advanced options

Scan for the delimiter

From the end of the input ▾

Number of delimiters to skip ⓘ

0

OK   Cancel

Text Between Delimiters

Enter the delimiters that mark the beginning and end of what you would like to extract.

Start delimiter

min

End delimiter

and

▲ Advanced options

Scan for the start delimiter

From the end of the input ▾

Number of start delimiters to skip ⓘ

0

Scan for the end delimiter

From the start delimiter, toward the... ▾

From the start delimiter, toward the end of the input

From the start delimiter, toward the start of the input

OK   Cancel

- ► Ensure you have two columns:Hours (numeric)Minutes (numeric)Add a Custom Column:Go to Add Column → Custom Column

- ► Enter the formula = [Hours] * 60 + [Minutes]

- ► Change the new column's data type to Whole Number (if not automatic)

- ► With total_minute column, go to Add Column → Custom Column

- ► Use this formula:

- ► = #duration(0, 0, [total_minute], 0)

- ► This converts total minutes into an Excel/Power Query duration value

Custom Column

Add a column that is computed from the other columns.

New column name

total_minute

Custom column formula ⓘ

= [hour]*60+[minutes]

Available columns

name
Author 1
Author 2
Author 3
narrator
time
releasedate

<< Insert

Custom Column

Add a column that is computed from the other columns.

New column name

duration

Custom column formula ⓘ
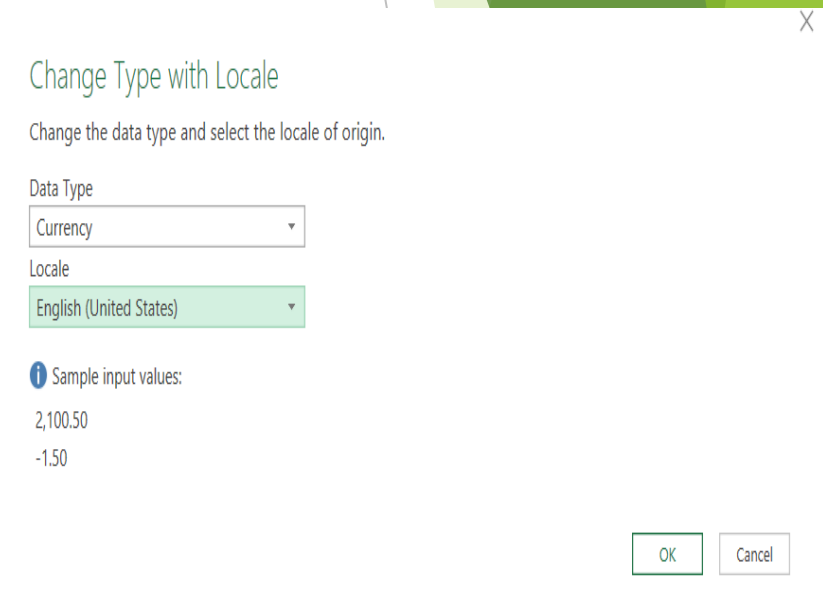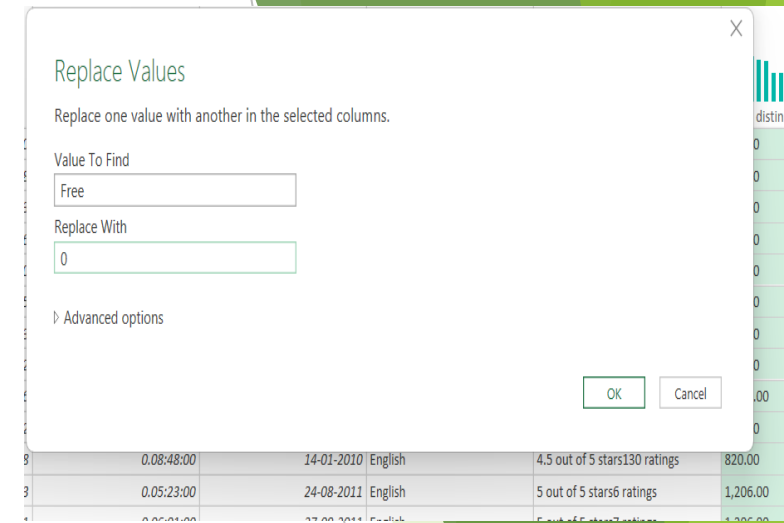
= #duration(0, 0, [total_minute], 0)

Available columns

time
hour
minutes
releasedate
language
stars
price
total_minute

<< Insert

Learn about Power Query formulas

Ensure the price column is in a numeric format, and identify any non-numeric values. & Ensure all currency values in the price column are formatted consistently with two decimal places.

▶ Replace "Free" with 0:

▶ Select the price column.Go to Transform → Replace Values. Find: "Free" Replace with: 0.

▶ Select the price column in Power Query.Go to Transform → Data Type → Using Locale...In the popup window: Data Type: Choose Currency Locale: Select English (United States) or the locale matching your currency format (e.g., "en-US" for USD) Click OK

Convert text ratings in the stars column to numeric values.

- ▶ Select the stars column in Power Query.
- ▶ Go to Add Column → Extract → Text Before Delimiter.
- ▶ Enter delimiter: " out of ".
- ▶ Convert the new column to Decimal Number type.

Text Before Delimiter

Enter the delimiter that marks the end of what you would like to extract.

Delimiter

out of

▲ Advanced options

Scan for the delimiter

From the end of the input ▼

Number of delimiters to skip ⓘ

0

OK   Cancel

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

- ▶ Select the narratedby column in Power Query.
- ▶ Go to Home → Split Column → By Delimiter.
- ▶ Choose the delimiter that separates narrators (e.g., comma).
- ▶ Select Each occurrence of the delimiter.
- ▶ Click OK to split into multiple columns.

Akshay Solanki

Split Column by Delimiter

Specify the delimiter used to split the text column.

Select or enter delimiter

Comma ▾

Split at

○ Left-most delimiter

○ Right-most delimiter

◉ Each occurrence of the delimiter

▷ Advanced options

Quote Character

" ▾

☐ Split using special characters

Insert special character ▾

OK   Cancel

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

▶ Select the releasedate and language columns in Power Query.

▶ Go to Add Column → Merge Columns.

▶ Choose delimiter: , (comma and space).

▶ Name the new column releaseinfo.

▶ Format releasedate to DD-MM-YYYY before merging if needed.

▶ Click OK to create the merged column.

Merge Columns

Choose how to merge the selected columns.

Separator

--Custom--

,

New column name (optional)

releaseinfo

OK   Cancel

X

Thank You