Step 13: As per the analysis this dataset doesn't have missing values and scrapped data seems to not readable, we used data wrangling techniques to make it more readable and for "Not rated" yet data we changed values to "0" so that we can for graphs in future. Moreover, we changed the time format into minutes from "hours & minutes combination" to help with data for visualization.

Step 1: By visualization it seems data sets have impurities

	name	author	narrator	time	releasedate	language	stars	price
0	Geronimo Stilton #11 & #12	Writtenby:GeronimoStilton	Narratedby:BillLobely	2 hrs and 20 mins	04-08-08	English	5 out of 5 stars34 ratings	468.00
1	The Burning Maze	Writtenby:RickRiordan	Narratedby:RobbieDaymond	13 hrs and 8 mins	01-05-18	English	4.5 out of 5 stars41 ratings	820.00
2	The Deep End	Writtenby:JeffKinney	Narratedby:DanRussell	2 hrs and 3 mins	06-11-20	English	4.5 out of 5 stars38 ratings	410.00
3	Daughter of the Deep	Writtenby:RickRiordan	Narratedby:SoneelaNankani	11 hrs and 16 mins	05-10-21	English	4.5 out of 5 stars12 ratings	615.00
4	The Lightning Thief: Percy Jackson, Book 1	Writtenby:RickRiordan	Narratedby:JesseBernstein	10 hrs	13-01-10	English	4.5 out of 5 stars181 ratings	820.00

Step2: we will clean the "author" column data first

```
original_author_column_data = df['author'].copy()
```

```
#In author column cleaning data
df['author'] = df['author'].str.replace('Writtenby:', '', regex=False)
```

```
author_rows_updated = (original_author_column_data != df['author']) & original_author_column_data.notna()
```

```
num_of_author_rows_updated = author_rows_updated.sum()
print(f"Number of rows updated in author column: {num_of_author_rows_updated}")
```

```
Number of rows updated in author column: 87489
```

```
df.head()
```

	name	author	narrator	time	releasedate	language	stars	price
0	Geronimo Stilton #11 & #12	GeronimoStilton	Narratedby:BillLobely	2 hrs and 20 mins	04-08-08	English	5 out of 5 stars34 ratings	468.00
1	The Burning Maze	RickRiordan	Narratedby:RobbieDaymond	13 hrs and 8 mins	01-05-18	English	4.5 out of 5 stars41 ratings	820.00
2	The Deep End	JeffKinney	Narratedby:DanRussell	2 hrs and 3 mins	06-11-20	English	4.5 out of 5 stars38 ratings	410.00
3	Daughter of the Deep	RickRiordan	Narratedby:SoneelaNankani	11 hrs and 16 mins	05-10-21	English	4.5 out of 5 stars12 ratings	615.00
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	Narratedby:JesseBernstein	10 hrs	13-01-10	English	4.5 out of 5 stars181 ratings	820.00

Step3: after that we will clean "narrator" column data same way

```
original_narrator_column_data = df['narrator'].copy()
```

```
#Similarly cleaning data in narrator column
df['narrator'] = df['narrator'].str.replace('Narratedby:', '', regex=False)
```

```
narrator_rows_updated = (original_narrator_column_data != df['narrator']) & original_narrator_column_data.notna()
```

```
num_of_narrator_rows_updated = narrator_rows_updated.sum()
print(f"Number of rows updated in aurthor column: {num_of_narrator_rows_updated}")
```

```
Number of rows updated in aurthor column: 87489
```

```
df.head()
```

	name	author	narrator	time	releasedate	language	stars	price
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	2 hrs and 20 mins	04-08-08	English	5 out of 5 stars34 ratings	468.00
1	The Burning Maze	RickRiordan	RobbieDaymond	13 hrs and 8 mins	01-05-18	English	4.5 out of 5 stars41 ratings	820.00
2	The Deep End	JeffKinney	DanRussell	2 hrs and 3 mins	06-11-20	English	4.5 out of 5 stars38 ratings	410.00
3	Daughter of the Deep	RickRiordan	SoneelaNankani	11 hrs and 16 mins	05-10-21	English	4.5 out of 5 stars12 ratings	615.00
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	10 hrs	13-01-10	English	4.5 out of 5 stars181 ratings	820.00

Step4: adding space in stars column between “5 out of 5 stars” and “34 ratings” to make it more readable

```
#Similarly cleaning data in stars column
df['stars'] = df['stars'].str.replace(r'(stars)(\d+)', r'\1 \2', regex=True)
```

```
stars_rows_updated = (original_stars_column_data != df['stars']) & original_stars_column_data.notna()
```

```
num_of_stars_rows_updated = stars_rows_updated.sum()
print(f"Number of rows updated in star column: {num_of_stars_rows_updated}")
```

```
Number of rows updated in star column: 15072
```

```
df.head()
```

	name	author	narrator	time	releasedate	language	stars	price
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	2 hrs and 20 mins	04-08-08	English	5 out of 5 stars 34 ratings	468.00
1	The Burning Maze	RickRiordan	RobbieDaymond	13 hrs and 8 mins	01-05-18	English	4.5 out of 5 stars 41 ratings	820.00
2	The Deep End	JeffKinney	DanRussell	2 hrs and 3 mins	06-11-20	English	4.5 out of 5 stars 38 ratings	410.00
3	Daughter of the Deep	RickRiordan	SoneelaNankani	11 hrs and 16 mins	05-10-21	English	4.5 out of 5 stars 12 ratings	615.00
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	10 hrs	13-01-10	English	4.5 out of 5 stars 181 ratings	820.00

Step5: used regex here above for adding space

Regex explanation

stars - to match the word

() - to capture the part of the matching

\d is for any digit between 0 to 9

+ means one or more digits

\1: Refers to the first capture group "stars".

\2: Refers to numeric part (the digits after "stars")

r'\1 \2' is telling pandas to keep the first part ("stars") and then insert a space.

Step6: data looks better to visualise however there seems to be some issues

```
df
```

	name	author	narrator	time	releasedate	language	stars	price
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	2 hrs and 20 mins	04-08-08	English	5 out of 5 stars 34 ratings	468.00
1	The Burning Maze	RickRiordan	RobbieDaymond	13 hrs and 8 mins	01-05-18	English	4.5 out of 5 stars 41 ratings	820.00
2	The Deep End	JeffKinney	DanRussell	2 hrs and 3 mins	06-11-20	English	4.5 out of 5 stars 38 ratings	410.00
3	Daughter of the Deep	RickRiordan	SoneelaNankani	11 hrs and 16 mins	05-10-21	English	4.5 out of 5 stars 12 ratings	615.00
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	10 hrs	13-01-10	English	4.5 out of 5 stars 181 ratings	820.00
...	...	...	...	...	...	...	...	...
87484	Last Days of the Bus Club	ChrisStewart	ChrisStewart	7 hrs and 34 mins	09-03-17	English	Not rated yet	596.00
87485	The Alps	StephenO'Shea	RobertFass	10 hrs and 7 mins	21-02-17	English	Not rated yet	820.00
87486	The Innocents Abroad	MarkTwain	FloGibson	19 hrs and 4 mins	30-12-16	English	Not rated yet	938.00
87487	A Sentimental Journey	LaurenceSterne	AntonLesser	4 hrs and 8 mins	23-02-11	English	Not rated yet	680.00
87488	Havana	MarkKurlansky	FleetCooper	6 hrs and 1 min	07-03-17	English	Not rated yet	569.00

87489 rows × 8 columns

Step7: Create a function to change time from “hrs and minutes” to only “minutes” to make it more readable

```
import re
```

```
#converting time in minutes we create a function

def convert_time_to_minutes(time_str):
    hours = 0
    minutes = 0
    hours_match = re.search(r'(\d+)\s*hrs?', time_str)
    minutes_match = re.search(r'(\d+)\s*min', time_str)

    if hours_match:
        hours = int(hours_match.group(1))

    if minutes_match:
        minutes = int(minutes_match.group(1))

    #to convert hours into minutes
    total_minutes = (hours * 60) + minutes
    return total_minutes
```

```
df['time_in_minutes'] = df['time'].apply(convert_time_to_minutes)
```

Step8: time in minutes now

```
[82]: columns = list(df.columns)
```

```
[84]: columns
```

```
[84]: ['name',
       'author',
       'narrator',
       'releasedate',
       'language',
       'stars',
       'price',
       'time_in_minutes']
```

```
[86]: columns.remove('time_in_minutes')
```

```
[88]: columns.insert(3, 'time_in_minutes')
```

```
[90]: df = df[columns]
```

```
[92]: df
```

[92]:

	name	author	narrator	time_in_minutes	releasedate	language	stars	price
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	140	04-08-08	English	5 out of 5 stars 34 ratings	468.00
1	The Burning Maze	RickRiordan	RobbieDaymond	788	01-05-18	English	4.5 out of 5 stars 41 ratings	820.00
2	The Deep End	JeffKinney	DanRussell	123	06-11-20	English	4.5 out of 5 stars 38 ratings	410.00
3	Daughter of the Deep	RickRiordan	SoneelaNankani	676	05-10-21	English	4.5 out of 5 stars 12 ratings	615.00
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	600	13-01-10	English	4.5 out of 5 stars 181 ratings	820.00
...	...	...	...	...	...	...	...	...
87484	Last Days of the Bus Club	ChrisStewart	ChrisStewart	454	09-03-17	English	Not rated yet	596.00

Step9: Created two more functions to sort stars ratings to different columns

```
[94]: #function to extract the star rating
def extract_star_rating(stars_str):
    match = re.search(r'(\d+(\.\d+)?) out of 5 stars', stars_str)  # Find the rating part (e.g., 5 or 4.5)
    if match:
        return float(match.group(1))  # Convert the match to a float and return
    return None  # Return None if no match is found
```

```
[96]: #function to extract the number of ratings
def extract_ratings(stars_str):
    match = re.search(r'(\d+)\s+ratings', stars_str)  # Find the ratings count (e.g., 34 or 41)
    if match:
        return int(match.group(1))  # Convert the match to an integer and return
    return None  # Return None if no match is found
```

Step 10: added star_rating column here

```
df.loc[:, 'star_rating'] = df['stars'].apply(extract_star_rating)
```

```
C:\Users\abres\AppData\Local\Temp\ipykernel_11616\4150790645.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df.loc[:, 'star_rating'] = df['stars'].apply(extract_star_rating)
```

```
df
```

	name	author	narrator	time_in_minutes	releasedate	language	stars	price	star_rating
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	140	04-08-08	English	5 out of 5 stars 34 ratings	468.00	5.0
1	The Burning Maze	RickRiordan	RobbieDaymond	788	01-05-18	English	4.5 out of 5 stars 41 ratings	820.00	4.5
2	The Deep End	JeffKinney	DanRussell	123	06-11-20	English	4.5 out of 5 stars 38 ratings	410.00	4.5
3	Daughter of the Deep	RickRiordan	SoneelaNankani	676	05-10-21	English	4.5 out of 5 stars 12 ratings	615.00	4.5
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	600	13-01-10	English	4.5 out of 5 stars 181 ratings	820.00	4.5
...	...	...	...	...	...	...	...	...	...
87484	Last Days of the Bus Club	ChrisStewart	ChrisStewart	454	09-03-17	English	Not rated yet	596.00	NaN
87485	The Alps	StephenO'Shea	RobertFass	607	21-02-17	English	Not rated yet	820.00	NaN
87486	The Innocents Abroad	MarkTwain	FloGibson	1144	30-12-16	English	Not rated yet	938.00	NaN
87487	A Sentimental Journey	LaurenceSterne	AntonLesser	248	23-02-11	English	Not rated yet	680.00	NaN
87488	Havana	MarkKurlansky	FleetCooper	361	07-03-17	English	Not rated yet	569.00	NaN

Step 11: Creating rating column too

```
df.loc[:, 'ratings'] = df['stars'].apply(extract_ratings)
```

```
C:\Users\abres\AppData\Local\Temp\ipykernel_11616\4155449710.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df.loc[:, 'ratings'] = df['stars'].apply(extract_ratings)
```

```
df
```

	name	author	narrator	time_in_minutes	releasedate	language	stars	price	star_rating	ratings
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	140	04-08-08	English	5 out of 5 stars 34 ratings	468.00	5.0	34.0
1	The Burning Maze	RickRiordan	RobbieDaymond	788	01-05-18	English	4.5 out of 5 stars 41 ratings	820.00	4.5	41.0
2	The Deep End	JeffKinney	DanRussell	123	06-11-20	English	4.5 out of 5 stars 38 ratings	410.00	4.5	38.0
3	Daughter of the Deep	RickRiordan	SoneelaNankani	676	05-10-21	English	4.5 out of 5 stars 12 ratings	615.00	4.5	12.0
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	600	13-01-10	English	4.5 out of 5 stars 181 ratings	820.00	4.5	181.0
...	...	...	...	...	...	...	...	...	...	...
87484	Last Days of the Bus Club	ChrisStewart	ChrisStewart	454	09-03-17	English	Not rated yet	596.00	NaN	NaN
87485	The Alps	StephenO'Shea	RobertFass	607	21-02-17	English	Not rated yet	820.00	NaN	NaN
87486	The Innocents Abroad	MarkTwain	FloGibson	1144	30-12-16	English	Not rated yet	938.00	NaN	NaN
87487	A Sentimental Journey	LaurenceSterne	AntonLesser	248	23-02-11	English	Not rated yet	680.00	NaN	NaN

Step 12: Removed “NaN” value and added “0” to make data more readable for “not rated yet” data and exported.

```
df
```

	name	author	narrator	time_in_minutes	releasedate	language	star_rating	price	ratings
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	140	04-08-08	English	5.0	468.00	34.0
1	The Burning Maze	RickRiordan	RobbieDaymond	788	01-05-18	English	4.5	820.00	41.0
2	The Deep End	JeffKinney	DanRussell	123	06-11-20	English	4.5	410.00	38.0
3	Daughter of the Deep	RickRiordan	SoneelaNankani	676	05-10-21	English	4.5	615.00	12.0
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	600	13-01-10	English	4.5	820.00	181.0
...	...	...	...	...	...	...	...	...	...
87484	Last Days of the Bus Club	ChrisStewart	ChrisStewart	454	09-03-17	English	0.0	596.00	0.0
87485	The Alps	StephenO'Shea	RobertFass	607	21-02-17	English	0.0	820.00	0.0
87486	The Innocents Abroad	MarkTwain	FloGibson	1144	30-12-16	English	0.0	938.00	0.0
87487	A Sentimental Journey	LaurenceSterne	AntonLesser	248	23-02-11	English	0.0	680.00	0.0
87488	Havana	MarkKurlansky	FleetCooper	361	07-03-17	English	0.0	569.00	0.0

87489 rows × 9 columns

```
df.to_csv("final_updated_audio_data_after_cleaning.csv")
```