IMPORTACIONES

In [2]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings("ignore")
```

CARGAMOS LOS DATOS

audible = pd.read_csv('audible_raw.csv') audible

BUSCAMOS INFORMACION EN NUESTROS DATOS

In [6]:

```
audible.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 87489 entries, 0 to 87488
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   name         87489 non-null  object
 1   author       87489 non-null  object
 2   narrator     87489 non-null  object
 3   time         87489 non-null  object
 4   releasedate  87489 non-null  object
 5   language     87489 non-null  object
 6   stars        87489 non-null  object
 7   price        87489 non-null  object
dtypes: object(8)
memory usage: 5.3+ MB
```

REMPLAZAMOS Writtenby Y Narratedby POR COMILLAS PARA QUE SEA MAS LEGIBLE NUETSRO DATAFRAME

In [8]:

```
audible['author'] = audible['author'].str.replace('Writtenby:','')
audible['narrator'] = audible['narrator'].str.replace('Narratedby:','')
```

In [10]:

```
audible[['narrator','author']].head()
```

Out[10]:

	narrator	author
0	BillLobely	GeronimoStilton
1	RobbieDaymond	RickRiordan
2	DanRussell	JeffKinney
3	SoneelaNankani	RickRiordan
4	JesseBernstein	RickRiordan

In [12]:

```
audible['stars'].sample(10)
```

Out[12]:

```
57485    Not rated yet
67641    Not rated yet
16921    Not rated yet
78995    Not rated yet
37456    Not rated yet
66109    Not rated yet
3320     Not rated yet
25853    Not rated yet
73248    Not rated yet
54989    Not rated yet
Name: stars, dtype: object
```

REMPLAZAMOS LOS VALORES Not rated yet POR VALORES N

In [14]:

```
audible['stars'].replace('Not rated yet',np.nan, inplace=True)
audible['stars'].head()
```

Out[14]:

```
0       5 out of 5 stars34 ratings
1     4.5 out of 5 stars41 ratings
2     4.5 out of 5 stars38 ratings
3     4.5 out of 5 stars12 ratings
4    4.5 out of 5 stars181 ratings
Name: stars, dtype: object
```

CODIGO SIRVE PARA LIMPIAR Y EXTRAER LOS DATOS NUMERICOS Y REMPLAZAR LAS COMAS

In [17]:

```
#Codigo Sirve para limpiar y extraer los datos numericos y remplazar las comas
audible['rating_stars'] = audible['stars'].str.extract(r'(\d+)').astype(float)
audible['n_rating'] = audible['stars'].str.replace(',','').str.extract('(\d+) rating').astype(float)
audible[['rating_stars','n_rating']].head()
```

Out[17]:

	rating_stars	n_rating
0	5.0	34.0
1	4.0	41.0
2	4.0	38.0
3	4.0	12.0
4	4.0	181.0

ELIMINAMOS LA COLUMNA STARS

In [19]:

```
audible.drop(columns = 'stars', axis=1, inplace = True)
audible
```

Out[19]:

	name	author	narrator	time	releasedate	language	price	rating_stars	n_rating
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	2 hrs and 20 mins	04-08-08	English	468.00	5.0	34.0
1	The Burning Maze	RickRiordan	RobbieDaymond	13 hrs and 8 mins	01-05-18	English	820.00	4.0	41.0
2	The Deep End	JeffKinney	DanRussell	2 hrs and 3 mins	06-11-20	English	410.00	4.0	38.0
3	Daughter of the Deep	RickRiordan	SoneelaNankani	11 hrs and 16 mins	05-10-21	English	615.00	4.0	12.0
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	10 hrs	13-01-10	English	820.00	4.0	181.0
...	...	...	...	...	...	...	...	...	...
87484	Last Days of the Bus Club	ChrisStewart	ChrisStewart	7 hrs and 34 mins	09-03-17	English	596.00	NaN	NaN
87485	The Alps	StephenO'Shea	RobertFass	10 hrs and 7 mins	21-02-17	English	820.00	NaN	NaN
87486	The Innocents Abroad	MarkTwain	FloGibson	19 hrs and 4 mins	30-12-16	English	938.00	NaN	NaN
87487	A Sentimental Journey	LaurenceSterne	AntonLesser	4 hrs and 8 mins	23-02-11	English	680.00	NaN	NaN
87488	Havana	MarkKurlansky	FleetCooper	6 hrs and 1 min	07-03-17	English	569.00	NaN	NaN

87489 rows × 9 columns

REMPLAZAMOS VALORES DE LA COLUMNA PRICE Y LOS TRANSFORMAMOS EN FLOAT

In [21]:

```
audible['price'] = audible['price'].str.replace(',','')
audible['price'] = audible['price'].str.replace('Free','')
audible['price'] = audible['price'].replace('', np.nan)
audible['price'] = audible['price'].astype(float)
```

In [23]:

```
audible.sample(10)
```

Out[23]:

	name	author	narrator	time	releasedate	language	price	rating_stars	n_rating
68804	日本一難しいはなし〜必ず眠くなる〜その18	青木幹和	青木幹和	1 hr and 43 mins	23-03-16	japanese	603.0	NaN	NaN
36961	Ahoi, Baby	SusanneMüller-Weiss	SusanneMüller-Weiss	4 hrs and 22 mins	01-07-20	german	398.0	NaN	NaN
72236	The War of Two Queens	JenniferL.Armentrout	StinaNielsen,TimCampbell	26 hrs and 38 mins	15-03-22	English	1055.0	4.0	11.0
58557	Making Money Simple	PeterLazaroff	PeterLazaroff	4 hrs and 42 mins	16-07-19	English	469.0	NaN	NaN
3080	Who Was Charles Darwin?	DeborahHopkinson	KevinPariseau	48 mins	26-03-19	English	351.0	NaN	NaN
20048	Woody	DavidEvanier	AaronAbano	15 hrs and 59 mins	03-11-15	English	221.0	NaN	NaN
2206	Кот да Винчи. Похищение в день рождения	КатяМатюшкина	СеменМендельсон,КсенияБржезовская	2 hrs and 37 mins	15-02-22	russian	166.0	NaN	NaN
64032	The Praise of Folly	DesideriusErasmus	AnnaSimon	3 hrs and 34 mins	24-01-22	English	166.0	NaN	NaN
78463	Somewhere to Belong	JudithMiller	StinaNielsen	10 hrs and 33 mins	19-04-11	English	938.0	NaN	NaN
70661	Wrangler's Corner Series 5	LynetteEason	CharlotteNorth	5 hrs and 16 mins	25-02-20	English	586.0	NaN	NaN

In [ ]:

CAMBIAMOS LA COLUMNA RATING_STARS A CATEGORY

In [27]:

```
audible['rating_stars'] = audible['rating_stars'].astype('category')
audible.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 87489 entries, 0 to 87488
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype   
---  ------        --------------  -----   
 0   name          87489 non-null  object  
 1   author        87489 non-null  object  
 2   narrator      87489 non-null  object  
 3   time          87489 non-null  object  
 4   releasedate   87489 non-null  object  
 5   language      87489 non-null  object  
 6   price         87151 non-null  float64 
 7   rating_stars  15072 non-null  category
 8   n_rating      15072 non-null  float64 
dtypes: category(1), float64(2), object(6)
memory usage: 5.4+ MB
```

CAMBIAMOS LA COLUMNA RELEASEDATE A PANDAS TO_DATETIME

In [29]:

```
audible['releasedate'] = pd.to_datetime(audible.releasedate)
```

In [30]:

```
audible.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 87489 entries, 0 to 87488
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype         
---  ------        --------------  -----         
 0   name          87489 non-null  object        
 1   author        87489 non-null  object        
 2   narrator      87489 non-null  object        
 3   time          87489 non-null  object        
 4   releasedate   87489 non-null  datetime64[ns]
 5   language      87489 non-null  object        
 6   price         87151 non-null  float64       
 7   rating_stars  15072 non-null  category      
 8   n_rating      15072 non-null  float64       
dtypes: category(1), datetime64[ns](1), float64(2), object(5)
memory usage: 5.4+ MB
```

In [31]:

```
audible['time'].sample(10)
```

Out[31]:

```
57709     6 hrs and 2 mins
65076    8 hrs and 33 mins
62470    4 hrs and 46 mins
51353    17 hrs and 8 mins
31617              50 mins
79011              15 mins
44802     14 hrs and 1 min
4255               40 mins
72828               31 hrs
27424    5 hrs and 36 mins
Name: time, dtype: object
```

FILTRAMOS LA COLUMNA QUE CONTIENEN LA PALABRA MINUTE

In [33]:

```
audible.time[audible.time.str.contains('minute')].sample(10)
audible['time'].sample(10)
```

Out[33]:

```
49454     9 hrs and 53 mins
31500     6 hrs and 45 mins
18553     6 hrs and 56 mins
15687               29 mins
65097    40 hrs and 37 mins
63273    12 hrs and 43 mins
61597     7 hrs and 56 mins
11131    13 hrs and 42 mins
85464     8 hrs and 24 mins
40950      1 hr and 35 mins
Name: time, dtype: object
```

ESTANDARIZAMOS LOS DATOS DE LA COLUMNA TIME REMPLZANDO ALGUNOS VALORES

In [36]:

```
audible['time'] = audible['time'].str.replace('hrs','hr')
audible['time'] = audible['time'].str.replace('mins','min')
audible['time'] = audible['time'].str.replace('less than 1 minute','1 min')
```

In [38]:

```
audible['time'].sample(10)
```

Out[38]:

```
67673     6 hr and 8 min
8305     4 hr and 33 min
4335     1 hr and 16 min
47999    21 hr and 1 min
67304    8 hr and 12 min
66323             29 min
12500    1 hr and 11 min
20640    1 hr and 39 min
357                5 min
23630    8 hr and 26 min
Name: time, dtype: object
```

GENERAMOS ESTADISTICAS DESCRIPTIVAS DE NUESTRO DATA FRAME

In [40]:

```
audible.describe()
```

Out[40]:

	releasedate	price	n_rating
count	87489	87151.000000	15072.000000
mean	2018-06-22 01:35:29.780886528	561.177266	21.613190
min	1998-12-27 00:00:00	11.000000	1.000000
25%	2016-08-30 00:00:00	279.000000	1.000000
50%	2020-01-30 00:00:00	585.000000	2.000000
75%	2021-08-04 00:00:00	759.000000	7.000000
max	2025-11-14 00:00:00	7198.000000	12573.000000
std	NaN	334.936411	207.479634

EXTRAEMOS LOS MINUTOS Y LAS HORAS, SE LOS ASIGNAMOS A UNA NUEVA VARIABLE TIME_MINS

In [42]:

```
hours = audible['time'].str.extract('(\d+) hr').fillna(0).astype(int)
mins = audible['time'].str.extract('(\d+) min').fillna(0).astype(int)


# CREAMOS UNA NUEVA COLUMNA SUMANDO LAS NUEVAS VARIABLES
audible['time_mins'] =  hours * 60 + mins
```

In [44]:

```
audible.head(10)
```

Out[44]:

	name	author	narrator	time	releasedate	language	price	rating_stars	n_rating	time_mins
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	2 hr and 20 min	2008-04-08	English	468.0	5.0	34.0	140
1	The Burning Maze	RickRiordan	RobbieDaymond	13 hr and 8 min	2018-01-05	English	820.0	4.0	41.0	788
2	The Deep End	JeffKinney	DanRussell	2 hr and 3 min	2020-06-11	English	410.0	4.0	38.0	123
3	Daughter of the Deep	RickRiordan	SoneelaNankani	11 hr and 16 min	2021-05-10	English	615.0	4.0	12.0	676
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	10 hr	2010-01-13	English	820.0	4.0	181.0	600
5	The Hunger Games: Special Edition	SuzanneCollins	TatianaMaslany	10 hr and 35 min	2018-10-30	English	656.0	5.0	72.0	635
6	Quest for the Diamond Sword	WinterMorgan	LukeDaniels	2 hr and 23 min	2014-11-25	English	233.0	5.0	11.0	143
7	The Dark Prophecy	RickRiordan	RobbieDaymond	12 hr and 32 min	2017-02-05	English	820.0	5.0	50.0	752
8	Merlin Mission Collection	MaryPopeOsborne	MaryPopeOsborne	10 hr and 56 min	2017-02-05	English	1256.0	5.0	5.0	656
9	The Tyrant's Tomb	RickRiordan	RobbieDaymond	13 hr and 22 min	2019-09-24	English	820.0	5.0	58.0	802

ELIMINAMOS LA COLUMNA TIME

In [45]:

```
audible.drop(columns = 'time', axis=1, inplace = True)
```

In [ ]:

In [48]:

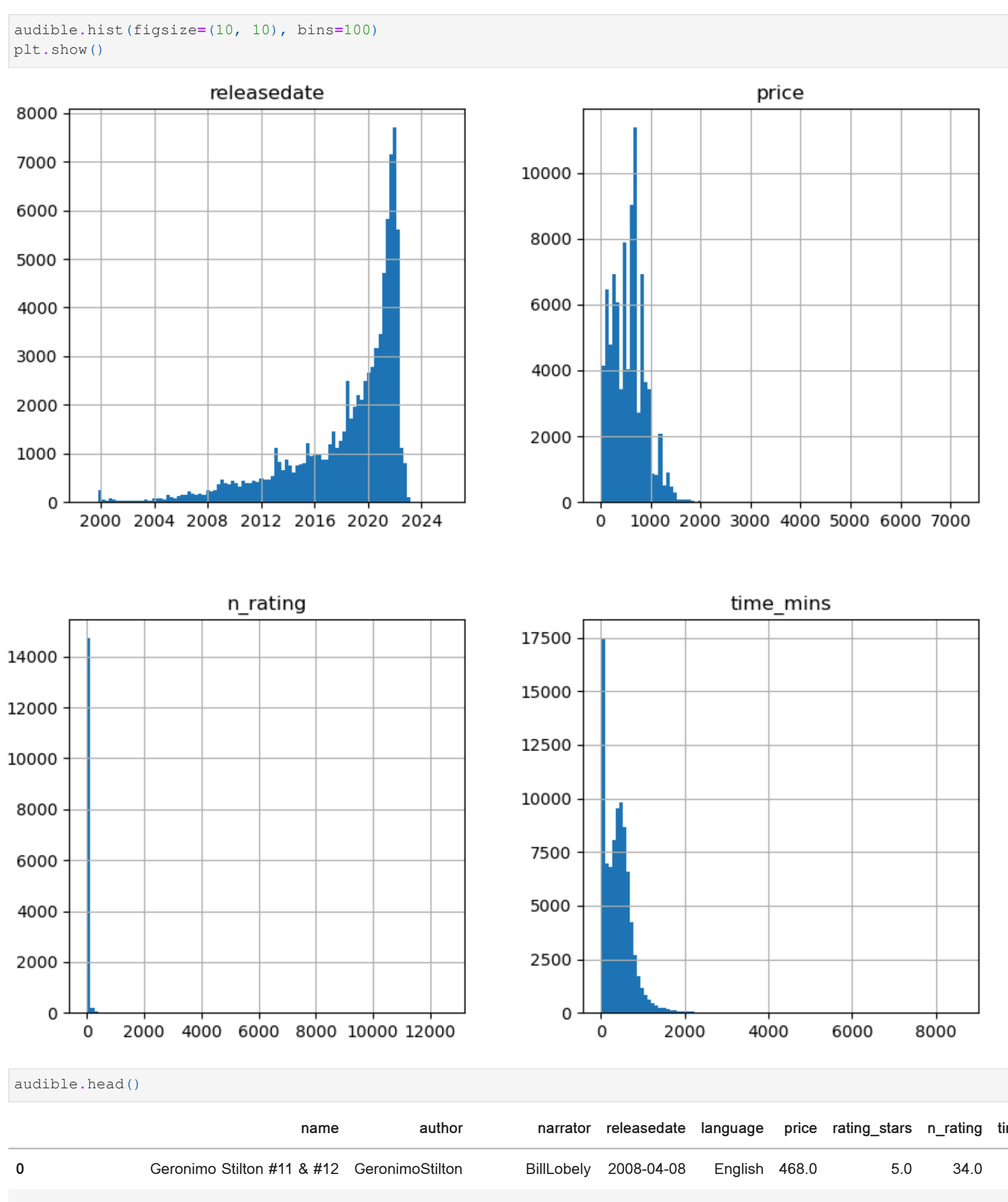
```
audible.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 87489 entries, 0 to 87488
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype         
---  ------        --------------  -----         
 0   name          87489 non-null  object        
 1   author        87489 non-null  object        
 2   narrator      87489 non-null  object        
 3   releasedate   87489 non-null  datetime64[ns]
 4   language      87489 non-null  object        
 5   price         87151 non-null  float64       
 6   rating_stars  15072 non-null  category      
 7   n_rating      15072 non-null  float64       
 8   time_mins     87489 non-null  int32         
dtypes: category(1), datetime64[ns](1), float64(2), int32(1), object(4)
memory usage: 5.1+ MB
```

CREAMOS UN HISTOGRAMA PARA VER NUESTROS DATOS NUMERICO

In [49]:

```
audible.hist(figsize=(10, 10), bins=100)
plt.show()
```



In [50]:

```
audible.head()
```

Out[50]:

	name	author	narrator	releasedate	language	price	rating_stars	n_rating	time_mins
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	2008-04-08	English	468.0	5.0	34.0	140
1	The Burning Maze	RickRiordan	RobbieDaymond	2018-01-05	English	820.0	4.0	41.0	788
2	The Deep End	JeffKinney	DanRussell	2020-06-11	English	410.0	4.0	38.0	123
3	Daughter of the Deep	RickRiordan	SoneelaNankani	2021-05-10	English	615.0	4.0	12.0	676
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	2010-01-13	English	820.0	4.0	181.0	600

VERIFICAMOS LOS VALORES DE LA COLUMNA LANGUAGE

In [53]:

```
audible['language'].unique()
```

Out[53]:

```
array(['English', 'Hindi', 'spanish', 'german', 'french', 'catalan',
       'swedish', 'italian', 'danish', 'finnish', 'dutch', 'hebrew',
       'russian', 'polish', 'galician', 'afrikaans', 'icelandic',
       'romanian', 'japanese', 'tamil', 'portuguese', 'urdu', 'hungarian',
       'czech', 'bulgarian', 'mandarin_chinese', 'basque', 'korean',
       'arabic', 'greek', 'turkish', 'ukrainian', 'slovene', 'norwegian',
       'telugu', 'lithuanian'], dtype=object)
```

CAPITALIZANOS LOS DATOS DE LA COLUMNA LANGUAGE

In [56]:

```
audible['language'] = audible['language'].str.capitalize()
audible['language'].unique()
```

Out[56]:

```
array(['English', 'Hindi', 'Spanish', 'German', 'French', 'Catalan',
       'Swedish', 'Italian', 'Danish', 'Finnish', 'Dutch', 'Hebrew',
       'Russian', 'Polish', 'Galician', 'Afrikaans', 'Icelandic',
       'Romanian', 'Japanese', 'Tamil', 'Portuguese', 'Urdu', 'Hungarian',
       'Czech', 'Bulgarian', 'Mandarin_chinese', 'Basque', 'Korean',
       'Arabic', 'Greek', 'Turkish', 'Ukrainian', 'Slovene', 'Norwegian',
       'Telugu', 'Lithuanian'], dtype=object)
```

VERIFICAMOS SI CONTAMOS CON DATOS DUPLICADOS

In [58]:

```
audible.duplicated().sum()
```

Out[58]:

```
0
```

In [59]:

```
audible.head()
```

Out[59]:

	name	author	narrator	releasedate	language	price	rating_stars	n_rating	time_mins
0	Geronimo Stilton #11 & #12	GeronimoStilton	BillLobely	2008-04-08	English	5.616	5.0	34.0	140
1	The Burning Maze	RickRiordan	RobbieDaymond	2018-01-05	English	9.840	4.0	41.0	788
2	The Deep End	JeffKinney	DanRussell	2020-06-11	English	4.920	4.0	38.0	123
3	Daughter of the Deep	RickRiordan	SoneelaNankani	2021-05-10	English	7.380	4.0	12.0	676
4	The Lightning Thief: Percy Jackson, Book 1	RickRiordan	JesseBernstein	2010-01-13	English	9.840	4.0	181.0	600

VER LOS DUPLICADOS DE LAS COLUMNAS LLAMADAS

In [60]:

```
subset_cols = ['name','author','narrator','time_mins','price']

audible.duplicated(subset = subset_cols).sum()
```

Out[60]:

```
70
```

In [61]:

```
audible[audible.duplicated(subset = subset_cols, keep = False)].sort_values(by = 'name')
```

Out[61]:

	name	author	narrator	releasedate	language	price	rating_stars	n_rating	time_mins
63978	"Das Böse ist des Menschen beste Kraft"	ChristianLiederer	ThomasKrause	2021-12-14	German	2.796	NaN	NaN	144
63965	"Das Böse ist des Menschen beste Kraft"	ChristianLiederer	ThomasKrause	2021-12-23	German	2.796	NaN	NaN	144
24625	90 Minutes in Heaven	DonPiper,CecilMurphey	DonPiper	2015-09-25	English	7.032	NaN	NaN	495
24116	90 Minutes in Heaven	DonPiper,CecilMurphey	DonPiper	2020-10-15	English	7.032	NaN	NaN	495
16971	Adagio in Dm	BillBrown	BillBrown	2021-06-08	English	5.472	NaN	NaN	78
...	...	...	...	...	...	...	...	...	...
18338	What I Talk About When I Talk About Running	HarukiMurakami	RayPorter	2016-07-07	English	4.776	4.0	193.0	263
38021	When Women Ruled the World	KaraCooney	KaraCooney	2018-11-30	English	10.032	4.0	14.0	555
20260	When Women Ruled the World	KaraCooney	KaraCooney	2018-11-30	English	10.032	4.0	13.0	555
22158	Wings of Fire	APJAbdulKalam,ArunTiwari	GirishKarnad	2020-01-04	English	0.900	5.0	1823.0	75
64428	Wings of Fire	APJAbdulKalam,ArunTiwari	GirishKarnad	2020-01-04	English	0.900	5.0	1827.0	75

139 rows × 9 columns

ELIMINAMOS LAS FILAS DUPLICADAS Y MANTENEMOS LAST

In [62]:

```
audible.drop_duplicates(subset = subset_cols, keep = 'last', inplace = True)
audible.duplicated(subset = subset_cols).sum()
```

Out[62]:

```
0
```

VERIFICAMOS LOS VALORES NULOS

In [63]:

```
audible.isna().sum()
```

Out[63]:

```
name                0
author              0
narrator            0
releasedate         0
language            0
price             338
rating_stars    72387
n_rating        72387
time_mins           0
dtype: int64
```

In [64]:

```
audible.sample(10)
```

Out[64]:

	name	author	narrator	releasedate	language	price	rating_stars	n_rating	time_mins
38694	Kashmirnama (Hindi Edition)	AshokKumarPandey	PiyushAgarwal	2019-02-14	Hindi	3.408	4.0	25.0	1280
4337	My First French Lessons	AlexaPolidoro	AlexaPolidoro	2006-11-28	English	2.280	NaN	NaN	51
6143	Moo Pak [Moo Pak] (Audiolibro en Catalán)	GabrielJosipovici	JoanMora	2019-03-01	Catalan	8.292	NaN	NaN	373
40007	Ancient Iran	IrajBashiri	MarkDelgado	2013-02-07	English	8.016	4.0	2.0	428
76345	Weights of Wrath	SmartypantsRomance,M.E.Carter	AlexKydd,StephanieRose	2021-06-22	English	7.032	NaN	NaN	461
30624	Activate Your Money	JanineFirpo	VanessaDaniels	2021-06-22	English	7.032	NaN	NaN	699
41979	Arbitrary Stupid Goal	TamaraShopsin	TamaraShopsin	2017-07-18	English	0.888	NaN	NaN	235
2700	The Turnaway Girls	HayleyChewins	BaileyCarr	2018-09-10	English	10.044	NaN	NaN	318
82729	Into Twilight	CalePlamann	DanielWisniewski	2021-01-12	English	10.032	NaN	NaN	691
27181	I Want to Be Left Behind	BrendaPeterson	BrendaPeterson	2012-09-02	English	5.616	NaN	NaN	633

GUARDAMOS EL NUEVO DATA FRAME

In [41]:

```
audible.to_csv('audible_clean.csv')
```

In [ ]: