In [165]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import missingno as msno
```

In [166]:
```
df = pd.read_csv('googleplaystore.csv')
df.sample(5)
```

Out[166]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Cont Ra
4284	Keepsafe Photo Vault: Hide Private Photos & Vi...	PHOTOGRAPHY	4.6	1656808	Varies with device	50,000,000+	Free	0	Every
1257	Step Counter - Calorie Counter	HEALTH_AND_FITNESS	4.0	1577	2.2M	500,000+	Free	0	Every
8433	Idle Heroes	FAMILY	4.7	417197	99M	10,000,000+	Free	0	Every
10746	FP Opgaver	TOOLS	NaN	9	61M	1,000+	Free	0	Every
3945	Tik Tok - including musical.ly	SOCIAL	4.4	5637451	59M	100,000,000+	Free	0	T

In [167]:
```
df.info()
```

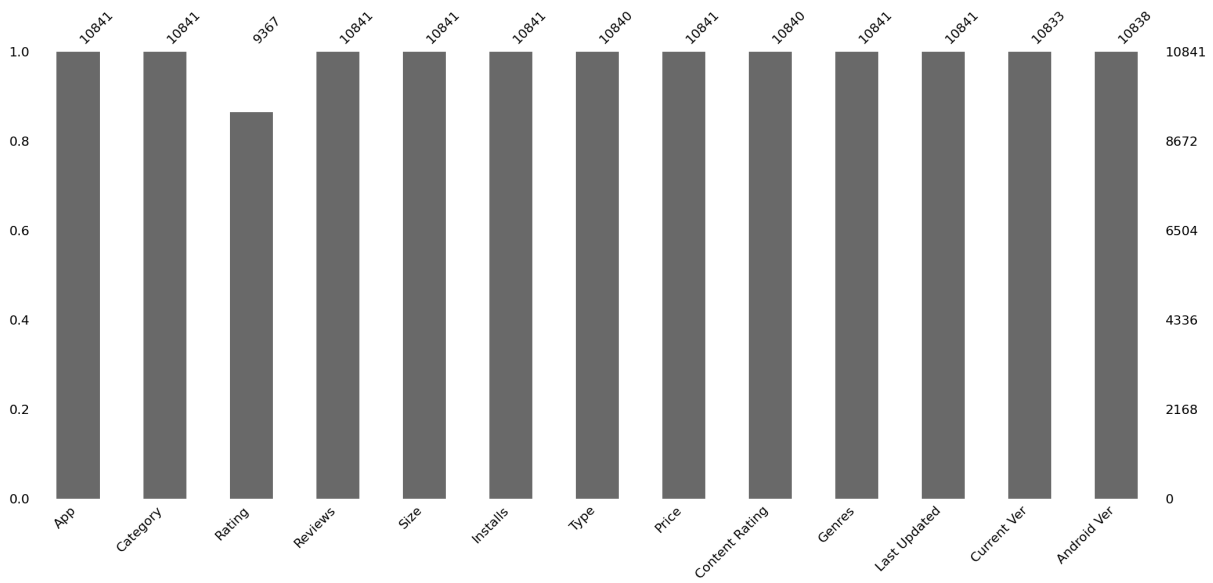
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             10841 non-null  object 
 1   Category        10841 non-null  object 
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object 
 4   Size            10841 non-null  object 
 5   Installs        10841 non-null  object 
 6   Type            10840 non-null  object 
 7   Price           10841 non-null  object 
 8   Content Rating  10840 non-null  object 
 9   Genres          10841 non-null  object 
 10  Last Updated    10841 non-null  object 
 11  Current Ver     10833 non-null  object 
 12  Android Ver     10838 non-null  object 
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

Data Cleaning

1. Which of the following column(s) has/have null values?

In [168]:
```
msno.bar(df)
```

Out[168]: <Axes: >



In [169]:
```
df.isna().sum()
```

Out[169]:
```
App                0
Category           0
Rating          1474
Reviews            0
Size               0
Installs           0
Type               1
Price              0
Content Rating     1
Genres             0
Last Updated       0
Current Ver        8
Android Ver        3
dtype: int64
```

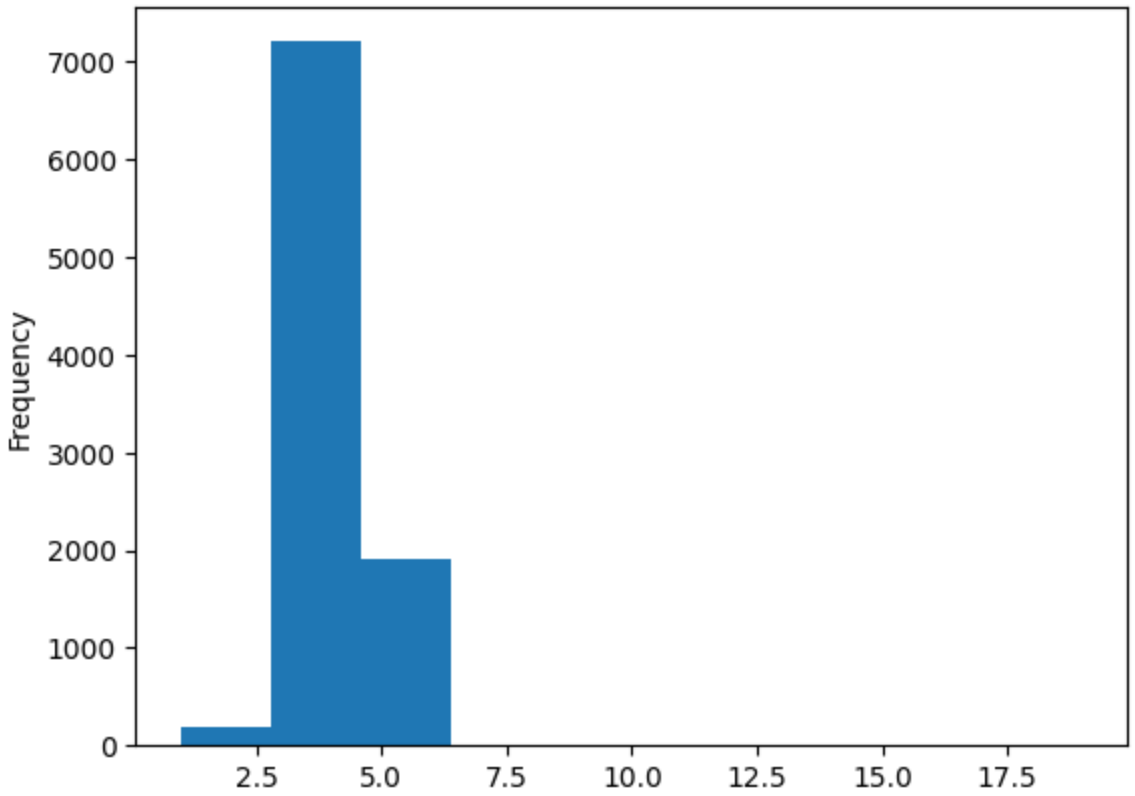
In [170]:
```
df.isna().sum().sort_values(ascending=False)
```

Out[170]:
```
Rating            1474
Current Ver          8
Android Ver          3
Type                 1
Content Rating       1
App                  0
Category             0
Reviews              0
Size                 0
Installs             0
Price                0
Genres               0
Last Updated         0
dtype: int64
```

2. Clean the `Rating` column and the other columns containing null values

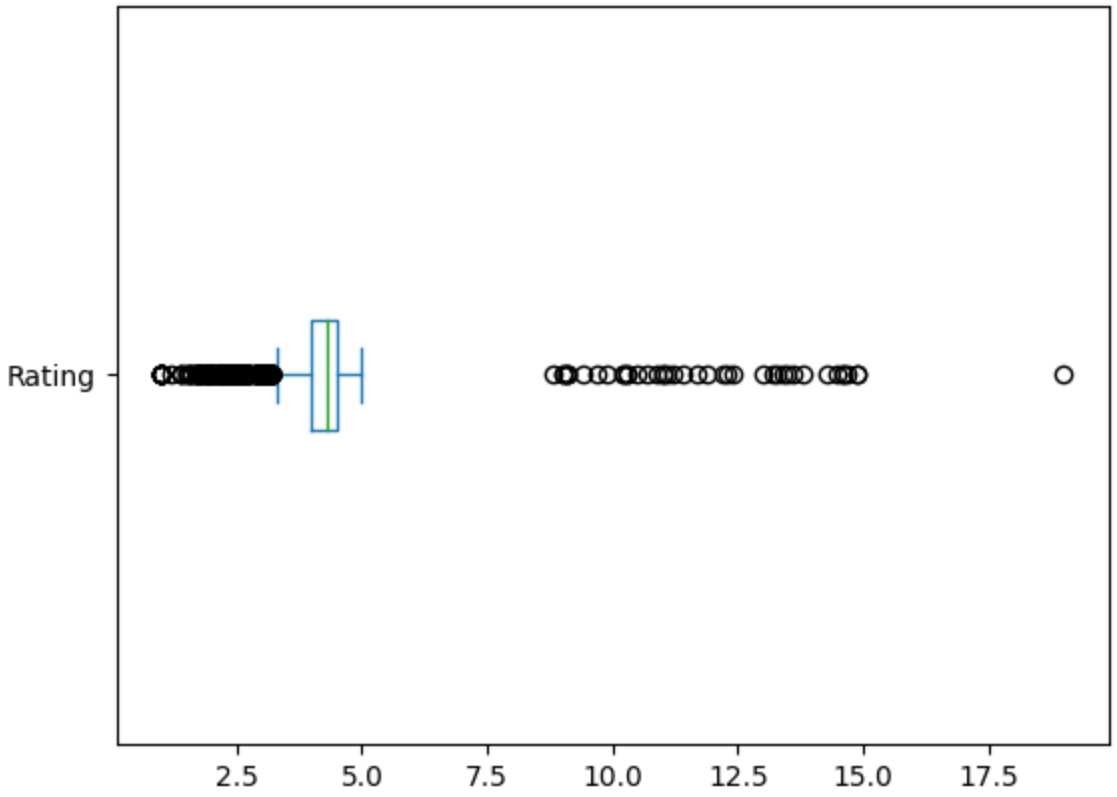
In [171]:
```
df['Rating'].plot(kind='hist')
```

Out[171]:
```
<Axes: ylabel='Frequency'>
```



In [172]:
```
df['Rating'].plot(kind='box', vert=False)
```

Out[172]:
```
<Axes: >
```



In [173]:
```
df['Rating'].describe()
```

Out[173]:
```
count    9367.000000
mean        4.231419
std         0.732847
min         1.000000
25%         4.000000
50%         4.300000
75%         4.500000
max        19.000000
Name: Rating, dtype: float64
```

In [174]:
```
filt = df['Rating'] > 5
df[filt]
```

Out[174]:

	App	Category	Rating	Reviews	Size	Installs	Type	Pr
681	Adult Dirty Emojis	DATING	12.2	80	5.5M	10,000+	Free	
2351	Brilliant Distinctions®	MEDICAL	11.0	78	72M	50,000+	Free	
2535	Patient Portal	MEDICAL	14.9	66	1.2M	50,000+	Free	
4226	How Old am I?	FAMILY	14.9	4635	3.9M	1,000,000+	Free	
4831	Z App	LIFESTYLE	11.7	405	25M	50,000+	Free	
4913	Trane Interactive Smart AC	TOOLS	10.2	48	3.3M	10,000+	Free	
5041	AF Comics Reader - Free	COMICS	13.3	5	4.1M	100+	Free	
5233	Club Penguin Island	FAMILY	10.7	107441	32M	1,000,000+	Free	
5236	AJ Bell Youinvest	FINANCE	12.3	135	5.3M	10,000+	Free	
5239	Baby Game Animal Jam Free	GAME	10.5	9	26M	500+	Free	
5579	Sleep as Android Gear Addon	HEALTH_AND_FITNESS	9.1	961	1.3M	100,000+	Free	
5660	BringGo AU & NZ	MAPS_AND_NAVIGATION	11.2	440	23M	10,000+	Paid	$0
6359	Free Coupons for Burger King	LIFESTYLE	10.9	17	3.1M	5,000+	Free	
6718	Bullshite!	GAME	14.5	48	13M	10,000+	Free	
6767	Bt Notifier - Smartwatch notice	TOOLS	13.2	632	8.2M	500,000+	Free	
6792	B.T.	NEWS_AND_MAGAZINES	13.0	636	1.4M	100,000+	Free	
6903	BW App	LIFESTYLE	8.8	48	12M	5,000+	Free	
7048	bz Basellandschaftliche News	NEWS_AND_MAGAZINES	9.1	33	33M	1,000+	Free	
7238	CF SHOP!	LIFESTYLE	11.9	88	43M	10,000+	Free	
7425	Gold Teeth Photo Editor	PHOTOGRAPHY	11.0	1022	3.8M	100,000+	Free	
7456	Company Kitchen	LIFESTYLE	13.5	81	7.7M	10,000+	Free	
7524	New: CL-150	FAMILY	10.2	4	4.2M	500+	Free	

	App	Category	Rating	Reviews	Size	Installs	Type	Pr
8135	KFC CZ	LIFESTYLE	9.7	1189	18M	100,000+	Free	
8377	Roland DG Mobile Panel	TOOLS	13.8	27	34M	1,000+	Free	
8469	DK 15 Minute Language Course	FAMILY	13.4	21	57M	1,000+	Free	
8508	NY mobilbank DK - Danske Bank	FINANCE	11.4	851	30M	100,000+	Free	
8549	DM airdisk	TOOLS	9.4	11	11M	1,000+	Free	
8552	DM HiDisk	TOOLS	9.9	54	20M	5,000+	Free	
8571	Interactive NPC DM Tool	FAMILY	10.3	5	629k	50+	Paid	$0
8730	Discovery Insure	MAPS_AND_NAVIGATION	14.3	1911	20M	100,000+	Free	
8938	Selfie DV	TOOLS	9.0	29	14M	1,000+	Free	
8942	Porch DV	HOUSE_AND_HOME	12.4	5	13M	1,000+	Free	
9270	My EF Center	FAMILY	14.6	89	35M	10,000+	Free	
9947	EV Connect	MAPS_AND_NAVIGATION	13.6	25	5.8M	1,000+	Free	
9948	HondaLink EV	TOOLS	14.6	44	41M	10,000+	Free	
10052	Advanced EX for RENAULT	TOOLS	11.1	130	143k	5,000+	Paid	$4
10129	EZ Inspections	PRODUCTIVITY	9.0	160	7.6M	10,000+	Free	
10136	EZ-SEE	VIDEO_PLAYERS	9.1	71	10M	10,000+	Free	
10159	My EZ-Link Mobile	LIFESTYLE	14.7	3187	11M	100,000+	Free	
10322	FE Civil Engineering Exam Prep	FAMILY	10.3	9	21M	1,000+	Free	
10472	Life Made WI-Fi Touchscreen Photo Frame	1.9	19.0	3.0M	1,000+	Free	0	F

In [175]:
```
df.loc[filt, 'Rating'] = np.nan
```

In [176]:
```
df['Rating'].mean()
```

Out[176]:
```
4.197726785331332
```

In [177]:
```
df['Rating'].fillna(df['Rating'].mean(), inplace=True)
```

In [178]:
```
df.dropna(inplace=True)
```

3. Clean the column `Reviews` and make it numeric

In [179]:
```
df['Reviews']
```

Out[179]:
```
0           159
1           967
2         87510
3        215644
4           967
          ...
10836        38
10837         4
10838         3
10839       114
10840    398307
Name: Reviews, Length: 10829, dtype: object
```

In [180]:
```
pd.to_numeric(df['Reviews'])
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
File /usr/local/lib/python3.11/site-packages/pandas/_libs/lib.pyx:2280, in pandas._libs.lib.maybe_convert_numeric()

ValueError: Unable to parse string "2M"

During handling of the above exception, another exception occurred:

ValueError                                Traceback (most recent call last)
Cell In[180], line 1
----> 1 pd.to_numeric(df['Reviews'])

File /usr/local/lib/python3.11/site-packages/pandas/core/tools/numeric.py:217, in to_numeric(arg, errors, downcast, dtype_backend)
    215 coerce_numeric = errors not in ("ignore", "raise")
    216 try:
--> 217     values, new_mask = lib.maybe_convert_numeric(  # type: ignore[call-overload]  # noqa
    218         values,
    219         set(),
    220         coerce_numeric=coerce_numeric,
    221         convert_to_masked_nullable=dtype_backend is not lib.no_default
    222         or isinstance(values_dtype, StringDtype),
    223     )
    224 except (ValueError, TypeError):
    225     if errors == "raise":

File /usr/local/lib/python3.11/site-packages/pandas/_libs/lib.pyx:2322, in pandas._libs.lib.maybe_convert_numeric()

ValueError: Unable to parse string "2M" at position 71
```

In [181]:
```
df['Reviews Numeric'] = pd.to_numeric(df['Reviews'], errors='coerce')
```

In [182]:
```
filt = df['Reviews Numeric'].isna()
df[filt]
```

Out[182]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Conten Rating
72	Android Auto - Maps, Media, Messaging & Voice	AUTO_AND_VEHICLES	4.2	2M	16M	10,000,000+	Free	0	Teen
1778	Block Craft 3D: Building Simulator Games For Free	GAME	4.5	1M	57M	50,000,000+	Free	0	Everyone
1781	Trivia Crack	GAME	4.5	6.4M	95M	100,000,000+	Free	0	Everyone

In [183]:
```
filt = df['Reviews'].str.contains('M')
df[filt]
```

Out[183]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Conten Rating
72	Android Auto - Maps, Media, Messaging & Voice	AUTO_AND_VEHICLES	4.2	2M	16M	10,000,000+	Free	0	Teen
1778	Block Craft 3D: Building Simulator Games For Free	GAME	4.5	1M	57M	50,000,000+	Free	0	Everyone
1781	Trivia Crack	GAME	4.5	6.4M	95M	100,000,000+	Free	0	Everyone

In [184]:
```
df.loc[filt, 'Reviews']
```

Out[184]:
```
72        2M
1778      1M
1781    6.4M
Name: Reviews, dtype: object
```

In [185]:
```
df.loc[filt, 'Reviews'].str.replace('M', '')
```

Out[185]:
```
72        2
1778      1
1781    6.4
Name: Reviews, dtype: object
```

In [186]:
```
pd.to_numeric(df.loc[filt, 'Reviews'].str.replace('M', ''))
* 1_000_000
```

Out[186]:
```
72      2000000.0
1778    1000000.0
1781    6400000.0
Name: Reviews, dtype: float64
```

In [189]:
```
new_reviews = (pd.to_numeric(df.loc[filt, 'Reviews']
                             .str.replace('M', ''))
               * 1_000_000).astype(str)
```

In [190]:
```
df.loc[filt, 'Reviews'] = new_reviews
```

In [191]:
```
df.loc[filt, 'Reviews']
```

Out[191]:
```
72      2000000.0
1778    1000000.0
1781    6400000.0
Name: Reviews, dtype: object
```

In [192]:
```
df['Reviews'] = pd.to_numeric(df['Reviews'])
```

4. How many duplicated apps are there?

In [193]:
```
# Case1
# Twitter, 9100
# Twitter, 9100
# Case2
# Fb, 10891
# Fb, 11002

# Case1 counts, Case2 does not count
df.duplicated(keep=False).sum()
```

Out[193]:
```
880
```

In [194]:
```
# Both Case1 and Case2 count
df.duplicated(subset=['App'], keep=False).sum()
```

Out[194]:
```
1979
```

In [195]:
```
# keep=False: keeps the original one
# keep=True: ignore the original one, only show the duplication
filt = df.duplicated(subset=['App'], keep=False)
df.loc[filt].sort_values(by='App')
```

Out[195]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Con Ra
1393	10 Best Foods for You	HEALTH_AND_FITNESS	4.0	2490.0	3.8M	500,000+	Free	0	Every
1407	10 Best Foods for You	HEALTH_AND_FITNESS	4.0	2490.0	3.8M	500,000+	Free	0	Every
2543	1800 Contacts - Lens Store	MEDICAL	4.7	23160.0	26M	1,000,000+	Free	0	Every
2322	1800 Contacts - Lens Store	MEDICAL	4.7	23160.0	26M	1,000,000+	Free	0	Every
2385	2017 EMRA Antibiotic Guide	MEDICAL	4.4	12.0	3.8M	1,000+	Paid	$16.99	Every
...	...	...	...	...	...	...	...	...	
3202	trivago: Hotels & Travel	TRAVEL_AND_LOCAL	4.2	219848.0	Varies with device	50,000,000+	Free	0	Every
3118	trivago: Hotels & Travel	TRAVEL_AND_LOCAL	4.2	219848.0	Varies with device	50,000,000+	Free	0	Every
3103	trivago: Hotels & Travel	TRAVEL_AND_LOCAL	4.2	219848.0	Varies with device	50,000,000+	Free	0	Every
8291	wetter.com - Weather and Radar	WEATHER	4.2	189310.0	38M	10,000,000+	Free	0	Every
3652	wetter.com - Weather and Radar	WEATHER	4.2	189313.0	38M	10,000,000+	Free	0	Every

1979 rows × 14 columns

In [196]:
```
filt1 = df.duplicated(subset=['App'], keep=False)
filt2 = ~df.duplicated(keep=False)
df.loc[filt1&filt2].sort_values(by='App')
```

Out[196]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genr
3083	365Scores - Live Scores	SPORTS	4.6	666521.0	25M	10,000,000+	Free	0	Everyone	Spo
5415	365Scores - Live Scores	SPORTS	4.6	666246.0	25M	10,000,000+	Free	0	Everyone	Spo
1675	8 Ball Pool	GAME	4.5	14198297.0	52M	100,000,000+	Free	0	Everyone	Spo
1703	8 Ball Pool	GAME	4.5	14198602.0	52M	100,000,000+	Free	0	Everyone	Spo
1755	8 Ball Pool	GAME	4.5	14200344.0	52M	100,000,000+	Free	0	Everyone	Spo
...	...	...	...	...	...	...	...	...	...	
565	stranger chat - anonymous chat	DATING	3.5	13204.0	6.1M	1,000,000+	Free	0	Mature 17+	Dati
2590	textPlus: Free Text & Calls	SOCIAL	4.1	382120.0	28M	10,000,000+	Free	0	Everyone	Soc
2637	textPlus: Free Text & Calls	SOCIAL	4.1	382121.0	28M	10,000,000+	Free	0	Everyone	Soc
3652	wetter.com - Weather and Radar	WEATHER	4.2	189313.0	38M	10,000,000+	Free	0	Everyone	Weath
8291	wetter.com - Weather and Radar	WEATHER	4.2	189310.0	38M	10,000,000+	Free	0	Everyone	Weath

1099 rows × 14 columns

5. Drop duplicated apps keeping the ones with the greatest number of reviews

In [197]:
```
filt1 = df.duplicated(subset=['App'], keep=False)
filt2 = ~df.duplicated(keep=False)
df.loc[filt1&filt2].sort_values(by=['App', 'Reviews'])
```

Out[197]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genr
5415	365Scores - Live Scores	SPORTS	4.6	666246.0	25M	10,000,000+	Free	0	Everyone	Spo
3083	365Scores - Live Scores	SPORTS	4.6	666521.0	25M	10,000,000+	Free	0	Everyone	Spo
3953	8 Ball Pool	SPORTS	4.5	14184910.0	52M	100,000,000+	Free	0	Everyone	Spo
1675	8 Ball Pool	GAME	4.5	14198297.0	52M	100,000,000+	Free	0	Everyone	Spo
1703	8 Ball Pool	GAME	4.5	14198602.0	52M	100,000,000+	Free	0	Everyone	Spo
...	...	...	...	...	...	...	...	...	...	
565	stranger chat - anonymous chat	DATING	3.5	13204.0	6.1M	1,000,000+	Free	0	Mature 17+	Dati
2590	textPlus: Free Text & Calls	SOCIAL	4.1	382120.0	28M	10,000,000+	Free	0	Everyone	Soc
2637	textPlus: Free Text & Calls	SOCIAL	4.1	382121.0	28M	10,000,000+	Free	0	Everyone	Soc
8291	wetter.com - Weather and Radar	WEATHER	4.2	189310.0	38M	10,000,000+	Free	0	Everyone	Weath
3652	wetter.com - Weather and Radar	WEATHER	4.2	189313.0	38M	10,000,000+	Free	0	Everyone	Weath

1099 rows × 14 columns

In [198]:
```
#df_copy_5 = df.copy()
# del df['Reviews Numeric']
```

In [199]:
```
df.sort_values(by=['App', 'Reviews'], inplace=True)
```

In [200]:
```
df.drop_duplicates(subset=['App'], keep='last', inplace=True)
```

6. Format the `Category` column

In [201]:
```
df['Category'].value_counts()
```

Out[201]:
```
Category
FAMILY                 1874
GAME                    945
TOOLS                   827
BUSINESS                420
MEDICAL                 395
PRODUCTIVITY            374
PERSONALIZATION         374
LIFESTYLE               369
FINANCE                 345
SPORTS                  325
COMMUNICATION           315
HEALTH_AND_FITNESS      288
PHOTOGRAPHY             281
NEWS_AND_MAGAZINES      254
SOCIAL                  239
BOOKS_AND_REFERENCE     221
TRAVEL_AND_LOCAL        219
SHOPPING                202
DATING                  170
VIDEO_PLAYERS           164
MAPS_AND_NAVIGATION     131
FOOD_AND_DRINK          112
EDUCATION               105
ENTERTAINMENT            86
AUTO_AND_VEHICLES        85
LIBRARIES_AND_DEMO       83
WEATHER                  79
HOUSE_AND_HOME           73
EVENTS                   64
ART_AND_DESIGN           60
PARENTING                60
COMICS                   56
BEAUTY                   53
Name: count, dtype: int64
```

In [202]:
```
df['Category'] = df['Category'].str.replace('_', ' ')
```

In [203]:
```
df['Category'] = df['Category'].str.capitalize()
```

In [204]:
```
df['Category'].value_counts()
```

Out[204]:
```
Category
Family                1874
Game                   945
Tools                  827
Business               420
Medical                395
Productivity           374
Personalization        374
Lifestyle              369
Finance                345
Sports                 325
Communication          315
Health and fitness     288
Photography            281
News and magazines     254
Social                 239
Books and reference    221
Travel and local       219
Shopping               202
Dating                 170
Video players          164
Maps and navigation    131
Food and drink         112
Education              105
Entertainment           86
Auto and vehicles       85
Libraries and demo      83
Weather                 79
House and home          73
Events                  64
Art and design          60
Parenting               60
Comics                  56
Beauty                  53
Name: count, dtype: int64
```

7. Clean and convert the `Installs` column to numeric type

In [205]:
```
df['Installs']
```

Out[205]:
```
8884          500+
324         10,000+
8532     1,000,000+
4541        10,000+
4636        10,000+
            ...
6334       100,000+
4362        10,000+
2575     1,000,000+
7559        10,000+
882      1,000,000+
Name: Installs, Length: 9648, dtype: object
```

In [206]:
```
filt = pd.to_numeric(df['Installs'], errors='coerce').isna()
df[filt].head()
```

Out[206]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	
8884	"i DT" Fútbol. Todos Somos Técnicos.	Sports	4.197727	27.0	3.6M	500+	Free	0	Everyone	
324	#NAME?	Comics	3.500000	115.0	9.1M	10,000+	Free	0	Mature 17+	
8532	+Download 4 Instagram Twitter	Social	4.500000	40467.0	22M	1,000,000+	Free	0	Everyone	
4541	.R	Tools	4.500000	259.0	203k	10,000+	Free	0	Everyone	
4636	/u/app	Communication	4.700000	573.0	53M	10,000+	Free	0	Mature 17+	Cor

In [207]:
```
df['Installs'].str.replace('+', '').str.replace(',', '')
```

Out[207]:
```
8884        500
324       10000
8532    1000000
4541      10000
4636      10000
          ...
6334     100000
4362      10000
2575    1000000
7559      10000
882     1000000
Name: Installs, Length: 9648, dtype: object
```

In [208]:
```
pd.to_numeric(df['Installs'].str.replace('+', '')
              .str.replace(',', ''))
```

Out[208]:
```
8884        500
324       10000
8532    1000000
4541      10000
4636      10000
          ...
6334     100000
4362      10000
2575    1000000
7559      10000
882     1000000
Name: Installs, Length: 9648, dtype: int64
```

In [209]:
```
df['Installs'] = pd.to_numeric(df['Installs']
                               .str.replace('+', '')
                               .str.replace(',', ''))
```

8. Clean and convert the `Size` column to numeric (representing bytes)

In [210]:
```
#df_copy_8 = df.copy()
```

In [211]:
```
#df = df_copy_5
df['Size']
```

Out[211]:
```
8884    3.6M
324     9.1M
8532     22M
4541    203k
4636     53M
        ...
6334     59M
4362     26M
2575     18M
7559    3.2M
882     4.0M
Name: Size, Length: 9648, dtype: object
```

In [212]:
```
filt = df['Size'] == 'Varies with device'
df[filt]
```

Out[212]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	
7338	20 Minuten (CH)	News and magazines	3.5	14153.0	Varies with device	1000000	Free	0	Everyone 10+	
7330	20 minutes (CH)	News and magazines	3.7	4379.0	Varies with device	1000000	Free	0	Teen	
3448	2018Emoji Keyboard 😂 Emoticons Lite - sticker&gif	Personalization	4.2	115773.0	Varies with device	10000000	Free	0	Everyone	Pers
3151	2GIS: directory & navigator	Travel and local	4.5	768833.0	Varies with device	50000000	Free	0	Everyone	Tra
4875	30 Day Ab Challenge FREE	Health and fitness	4.3	48253.0	Varies with device	1000000	Free	0	Everyone	
...	...	...	...	...	...	...	...	...	...	
4811	yHomework - Math Solver	Family	4.2	50771.0	Varies with device	1000000	Free	0	Everyone	
2758	zulily - Shop Daily Deals in Fashion and Home	Shopping	4.5	28560.0	Varies with device	1000000	Free	0	Everyone	
3960	▶ MultiCraft — Free Miner! 👍	Game	4.3	1305050.0	Varies with device	50000000	Free	0	Everyone 10+	
3824	乗換 NAVITIME Timetable & Route Search in Japan T...	Maps and navigation	4.4	50459.0	Varies with device	5000000	Free	0	Everyone	
9222	英漢字典 EC Dictionary	Family	4.3	55408.0	Varies with device	1000000	Free	0	Everyone	

1227 rows × 14 columns

In [213]:
```
df['Size'] = df['Size'].str.replace('Varies with device', '0')
```

In [214]:
```
df[filt]
```

Out[214]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	
7338	20 Minuten (CH)	News and magazines	3.5	14153.0	0	1000000	Free	0	Everyone 10+	M
7330	20 minutes (CH)	News and magazines	3.7	4379.0	0	1000000	Free	0	Teen	M
3448	2018Emoji Keyboard 😂 Emoticons Lite - sticker&gif	Personalization	4.2	115773.0	0	10000000	Free	0	Everyone	Perso
3151	2GIS: directory & navigator	Travel and local	4.5	768833.0	0	50000000	Free	0	Everyone	Trave
4875	30 Day Ab Challenge FREE	Health and fitness	4.3	48253.0	0	1000000	Free	0	Everyone	
...	...	...	...	...	...	...	...	...	...	
4811	yHomework - Math Solver	Family	4.2	50771.0	0	1000000	Free	0	Everyone	E
2758	zulily - Shop Daily Deals in Fashion and Home	Shopping	4.5	28560.0	0	1000000	Free	0	Everyone	S
3960	▶ MultiCraft — Free Miner! 👍	Game	4.3	1305050.0	0	50000000	Free	0	Everyone 10+	A
3824	乗換 NAVITIME Timetable & Route Search in Japan T...	Maps and navigation	4.4	50459.0	0	5000000	Free	0	Everyone	N
9222	英漢字典 EC Dictionary	Family	4.3	55408.0	0	1000000	Free	0	Everyone	E

1227 rows × 14 columns

In [215]:
```
df['Size'].value_counts()
```

Out[215]:
```
Size
0       1227
12M      181
11M      181
13M      177
14M      176
        ...
914k       1
353k       1
784k       1
951k       1
549k       1
Name: count, Length: 457, dtype: int64
```

In [216]:
```
df['Size'].info()
```

```
<class 'pandas.core.series.Series'>
Index: 9648 entries, 8884 to 882
Series name: Size
Non-Null Count  Dtype
--------------  -----
9648 non-null   object
dtypes: object(1)
memory usage: 150.8+ KB
```

In [217]:
```
filt = df['Size'].str.contains('k')
df.loc[filt, 'Size'].head()
```

Out[217]:
```
4541    203k
4897    371k
6671    243k
4871    239k
5035     78k
Name: Size, dtype: object
```

In [218]:
```
pd.to_numeric(df.loc[filt, 'Size'].str.replace('k', '')) * 1024
```

Out[218]:
```
4541    207872.0
4897    379904.0
6671    248832.0
4871    244736.0
5035     79872.0
          ...
5482    670720.0
7370    919552.0
9333    120832.0
8148    902144.0
5832    562176.0
Name: Size, Length: 310, dtype: float64
```

In [219]:
```
df.loc[filt, 'Size'] = (
pd.to_numeric(df.loc[filt, 'Size'].str.replace('k', '')) * 1024)
.astype(str)
```

In [220]:
```
df.loc[filt, 'Size']
```

Out[220]:
```
4541    207872.0
4897    379904.0
6671    248832.0
4871    244736.0
5035     79872.0
          ...
5482    670720.0
7370    919552.0
9333    120832.0
8148    902144.0
5832    562176.0
Name: Size, Length: 310, dtype: object
```

In [221]:
```
filt = df['Size'].str.contains('M')

df.loc[filt, 'Size'] = (
pd.to_numeric(df.loc[filt, 'Size'].str.replace('M', ''))
    * 1024 * 1024).astype(str)
```

In [222]:
```
df.loc[filt, 'Size']
```

Out[222]:
```
8884     3774873.6
324      9542041.6
8532    23068672.0
4636    55574528.0
5940    14680064.0
            ...
6334    61865984.0
4362    27262976.0
2575    18874368.0
7559     3355443.2
882      4194304.0
Name: Size, Length: 8111, dtype: object
```

In [223]:
```
df['Size'] = pd.to_numeric(df['Size'])
```

In [224]:
```
df.head()
```

Out[224]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating
8884	"¡ DT" Fútbol. Todos Somos Técnicos.	Sports	4.197727	27.0	3774873.6	500	Free	0	Everyone
324	#NAME?	Comics	3.500000	115.0	9542041.6	10000	Free	0	Mature 17+
8532	+Download 4 Instagram Twitter	Social	4.500000	40467.0	23068672.0	1000000	Free	0	Everyone
4541	.R	Tools	4.500000	259.0	207872.0	10000	Free	0	Everyone
4636	/u/app	Communication	4.700000	573.0	55574528.0	10000	Free	0	Mature 17+

9. Clean and convert the `Price` column to numeric

In [225]:
```
#df_copy_9 = df
df['Price'].value_counts()
```

Out[225]:
```
Price
0          8853
$0.99       143
$2.99       124
$1.99        73
$4.99        70
           ...
$25.99        1
$389.99       1
$37.99        1
$18.99        1
$1.20         1
Name: count, Length: 93, dtype: int64
```

In [226]:
```
pd.to_numeric(df['Price'].str.replace('$', ''))
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
File /usr/local/lib/python3.11/site-packages/pandas/_libs/lib.pyx:2280, in pandas._libs.lib.maybe_convert_numeric()

ValueError: Unable to parse string "Free"

During handling of the above exception, another exception occurred:

ValueError                                Traceback (most recent call last)
Cell In[226], line 1
----> 1 pd.to_numeric(df['Price'].str.replace('$', ''))

File /usr/local/lib/python3.11/site-packages/pandas/core/tools/numeric.py:217, in to_numeric(arg, errors, downcast, dtype_backend)
    215 coerce_numeric = errors not in ("ignore", "raise")
    216 try:
--> 217     values, new_mask = lib.maybe_convert_numeric(  # type: ignore[call-overload]  # noqa
    218         values,
    219         set(),
    220         coerce_numeric=coerce_numeric,
    221         convert_to_masked_nullable=dtype_backend is not lib.no_default
    222         or isinstance(values_dtype, StringDtype),
    223     )
    224 except (ValueError, TypeError):
    225     if errors == "raise":

File /usr/local/lib/python3.11/site-packages/pandas/_libs/lib.pyx:2322, in pandas._libs.lib.maybe_convert_numeric()

ValueError: Unable to parse string "Free" at position 256
```

In [227]:
```
filt = df['Price'] == 'Free'
df[filt]
```

Out[227]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	
5240	AJ Concept Group App	House and home	4.197727	0.0	17825792.0	10	Free	Free	E
7994	All Star Of CW	Family	4.197727	0.0	2516582.4	10	Free	Free	E
3188	American Airlines	Travel and local	3.700000	16980.0	0.0	5000000	Free	Free	E
4960	Anti Adware	Tools	3.900000	18751.0	3774873.6	1000000	Free	Free	E
3833	Atlan3D Navigation: Korea navigator	Maps and navigation	4.200000	22063.0	0.0	1000000	Free	Free	E
5830	Ay Up	Family	4.500000	11.0	7549747.2	100	Free	Free	E

In [228]:
```
df.loc[filt, 'Price']
```

Out[228]:
```
5240     Free
7994     Free
3188     Free
4960     Free
3833     Free
5830     Free
5895     Free
5951     Free
6372     Free
9159     Free
263      Free
8070     Free
7460     Free
4581     Free
4528     Free
8341     Free
5168     Free
9787     Free
3511     Free
1532     Free
10247    Free
4404     Free
9926     Free
3088     Free
8531     Free
582      Free
911      Free
10692    Free
9676     Free
333      Free
5881     Free
218      Free
958      Free
10503    Free
2112     Free
0        Free
1764     Free
10178    Free
3101     Free
3401     Free
10754    Free
225      Free
10468    Free
5836     Free
Name: Price, dtype: object
```

In [229]:
```
df['Price'] = df['Price'].str.replace('Free', '0')
```

In [230]:
```
df['Price'].value_counts()
```

Out[230]:
```
Price
0           8897
$0.99        143
$2.99        124
$1.99         73
$4.99         70
             ...
$25.99          1
$389.99         1
$37.99          1
$18.99          1
$1.20           1
Name: count, Length: 92, dtype: int64
```

In [231]:
```
df['Price'] = pd.to_numeric(df['Price'].str.replace('$', ''))
```

In [232]:
```
df['Price'].value_counts()
```

Out[232]:
```
Price
0.00      8897
0.99       143
2.99       124
1.99        73
4.99        70
           ...
25.99        1
389.99       1
37.99        1
18.99        1
1.20         1
Name: count, Length: 92, dtype: int64
```

10. Paid or free?

In [233]:
```
df['Distribution'] = 'Free'
```

In [234]:
```
filt = df['Price'] > 0
df.loc[filt, 'Distribution'] = 'Paid'
```

In [235]:
```
df.sample(10)
```

Out[235]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Conter Ratin
7538	CM Launcher 3D Pro💎	Personalization	4.700000	23802.0	177152.0	100000	Paid	4.99	Everyor
2053	Educational Games for Kids	Family	4.500000	7050.0	25165824.0	1000000	Free	0.00	Everyor
546	The League	Dating	3.000000	837.0	9856614.4	100000	Free	0.00	Matur 17
7547	Ruler(cm, inch)	Tools	3.400000	2889.0	2621440.0	500000	Free	0.00	Everyor
8504	DK TEL Dialer	Communication	4.197727	0.0	4404019.2	50	Free	0.00	Everyor
10502	Fun Kid Racing - Motocross	Family	4.100000	59768.0	0.0	10000000	Free	0.00	Everyor
4391	Traps n' Gemstones	Game	4.500000	413.0	9542041.6	1000	Paid	4.99	Everyor
8115	LOCX Applock Lock Apps & Photo	Productivity	4.500000	208543.0	0.0	10000000	Free	0.00	Everyor
7227	CE SODEXO PASS FRANCE	Productivity	4.197727	0.0	22020096.0	50	Free	0.00	Everyor
8325	DF Coaching	Sports	4.000000	1.0	3774873.6	100	Free	0.00	Everyor

Analysis

11. Which app has the most reviews?

In [236]:
```
df.sort_values(by='Reviews', ascending=False)
```

Out[236]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	C
2544	Facebook	Social	4.100000	78158306.0	0.0	1000000000	Free	0.0	
381	WhatsApp Messenger	Communication	4.400000	69119316.0	0.0	1000000000	Free	0.0	Ev
2604	Instagram	Social	4.500000	66577446.0	0.0	1000000000	Free	0.0	
382	Messenger – Text and Video Chat for Free	Communication	4.000000	56646578.0	0.0	1000000000	Free	0.0	Ev
1879	Clash of Clans	Game	4.600000	44893888.0	102760448.0	100000000	Free	0.0	Ev
...	...	...	...	...	...	...	...	...	
5229	AJ+ Beta	News and magazines	4.197727	0.0	0.0	1000	Free	0.0	Ev
5255	AJ and Alyssa	Lifestyle	4.197727	0.0	486400.0	100	Free	0.0	Ev
10692	MARKET FO	Communication	4.197727	0.0	15728640.0	100	Free	0.0	Ev
5270	AJ Wallpapers	Personalization	4.197727	0.0	4089446.4	100	Free	0.0	Ev
9139	DZ Register	Productivity	4.197727	0.0	18874368.0	1	Free	0.0	Ev

9648 rows × 15 columns

In [237]:
```
filt = df['Reviews'] == df['Reviews'].max()
df[filt]
```

Out[237]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	L
2544	Facebook	Social	4.1	78158306.0	0.0	1000000000	Free	0.0	Teen	Social	

12. What category has the highest number of apps uploaded to the store?

In [239]:
```
df['Category'].value_counts()
```

Out[239]:
```
Category
Family                 1874
Game                    945
Tools                   827
Business                420
Medical                 395
Productivity            374
Personalization         374
Lifestyle               369
Finance                 345
Sports                  325
Communication           315
Health and fitness      288
Photography             281
News and magazines      254
Social                  239
Books and reference     221
Travel and local        219
Shopping                202
Dating                  170
Video players           164
Maps and navigation     131
Food and drink          112
Education               105
Entertainment            86
Auto and vehicles        85
Libraries and demo       83
Weather                  79
House and home           73
Events                   64
Art and design           60
Parenting                60
Comics                   56
Beauty                   53
Name: count, dtype: int64
```

13. To which category belongs the most expensive app?

In [242]:
```
df.sort_values(by='Price', ascending=False)
```

Out[242]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Conte Rati
4367	I'm Rich - Trump Edition	Lifestyle	3.6	275.0	7654604.8	10000	Paid	400.00	Everyc
5358	I am Rich!	Finance	3.8	93.0	23068672.0	1000	Paid	399.99	Everyc
5356	I Am Rich Premium	Finance	4.1	1867.0	4928307.2	50000	Paid	399.99	Everyc
5362	I Am Rich Pro	Family	4.4	201.0	2831155.2	5000	Paid	399.99	Everyc
4197	most expensive app (H)	Family	4.3	6.0	1572864.0	100	Paid	399.99	Everyc
...	...	...	...	...	...	...	...	...	
10438	Dolphin and fish coloring book	Family	3.9	2249.0	0.0	500000	Free	0.00	Everyc
3434	Dolphins Live Wallpaper	Personalization	4.2	25807.0	5767168.0	1000000	Free	0.00	Everyc
1242	Domino's Pizza USA	Food and drink	4.7	1032935.0	0.0	10000000	Free	0.00	Everyc
2158	Dominos Game ✔	Family	4.1	2903.0	16777216.0	1000000	Free	0.00	Everyc
882	🔥 Football Wallpapers 4K | Full HD Backgrounds 😍	Entertainment	4.7	11661.0	4194304.0	1000000	Free	0.00	Everyc

9648 rows × 15 columns

14. What's the name of the most expensive game?

In [244]:
```
filt = df['Category'] == 'Game'
df[filt].sort_values(by='Price', ascending=False)
```

Out[244]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating
4203	The World Ends With You	Game	4.6	4108.0	13631488.0	10000	Paid	17.99	Everyone 10+
10782	Trine 2: Complete Story	Game	3.8	252.0	11534336.0	10000	Paid	16.99	Teen
6341	Blackjack Verite Drills	Game	4.6	17.0	4928307.2	100	Paid	14.00	Teen
1838	Star Wars ™: DIRTY	Game	4.5	38207.0	15728640.0	100000	Paid	9.99	Teen
6198	Backgammon NJ for Android	Game	4.4	1644.0	15728640.0	10000	Paid	7.99	Everyone
...	...	...	...	...	...	...	...	...	...
7600	Dreamland Arcade - Steven Universe	Game	4.0	6386.0	25165824.0	500000	Free	0.00	Everyone
10522	Drift Legends	Game	4.2	33788.0	28311552.0	1000000	Free	0.00	Everyone
4434	Drink-O-Tron The Drinking Game	Game	4.1	140.0	47185920.0	50000	Free	0.00	Mature 17+
10508	Drive 4x4 Luxury SUV Jeep	Game	4.2	2183.0	48234496.0	500000	Free	0.00	Everyone
3960	▶ MultiCraft — Free Miner! 👍	Game	4.3	1305050.0	0.0	50000000	Free	0.00	Everyone 10+

945 rows × 15 columns

15. Which is the most popular Finance App?

In [245]:
```
filt = df['Category'] == 'Finance'
df[filt].sort_values(by='Installs', ascending=False)
```

Out[245]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Gen
5601	Google Pay	Finance	4.200000	348132.0	0.0	100000000	Free	0.00	Everyone	Fina
1156	PayPal	Finance	4.300000	659760.0	49283072.0	50000000	Free	0.00	Everyone	Fina
1081	İşCep	Finance	4.500000	381788.0	33554432.0	10000000	Free	0.00	Everyone	Fina
1168	Wells Fargo Mobile	Finance	4.400000	250719.0	38797312.0	10000000	Free	0.00	Everyone	Fina
1169	Capital One® Mobile	Finance	4.600000	510401.0	82837504.0	10000000	Free	0.00	Everyone	Fina
...	...	...	...	...	...	...	...	...	...	
10417	FH Wallet	Finance	4.197727	0.0	10380902.4	1	Free	0.00	Everyone	Fina
9101	amm dz	Finance	4.197727	0.0	14680064.0	1	Paid	5.99	Everyone	Fina
10745	FP Boss	Finance	4.197727	1.0	6081740.8	1	Free	0.00	Everyone	Fina
9905	Eu sou Rico	Finance	4.197727	0.0	2726297.6	0	Paid	30.99	Everyone	Fina
9917	Eu Sou Rico	Finance	4.197727	0.0	1468006.4	0	Paid	394.99	Everyone	Fina

345 rows × 15 columns

16. What Teen Game has the most reviews?

In [247]:
```
filt1 = df['Category'] == 'Game'
filt2 = df['Content Rating'] == 'Teen'
df[filt1&filt2].sort_values(by='Reviews', ascending=False)
```

Out[247]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	G
3912	Asphalt 8: Airborne	Game	4.500000	8389714.0	96468992.0	100000000	Free	0.00	Teen	F
5417	Mobile Legends: Bang Bang	Game	4.400000	8219586.0	103809024.0	100000000	Free	0.00	Teen	A
1988	Hungry Shark Evolution	Game	4.500000	6074627.0	104857600.0	100000000	Free	0.00	Teen	A
10327	Garena Free Fire	Game	4.500000	5534114.0	55574528.0	100000000	Free	0.00	Teen	A
3967	Pixel Gun 3D: Survival shooter & Battle Royale	Game	4.500000	4487182.0	57671680.0	50000000	Free	0.00	Teen	A
...	...	...	...	...	...	...	...	...	...	
4431	Obbligo o Verità? PRO	Game	4.197727	4.0	3040870.4	100	Paid	0.99	Teen	
6335	BJ card game blackjack	Game	4.197727	3.0	22020096.0	500	Free	0.00	Teen	
6555	Sic Bo	Game	4.197727	1.0	11534336.0	100	Paid	1.99	Teen	
6329	Basic Strategy Training BJ 21	Game	4.197727	0.0	24117248.0	500	Free	0.00	Teen	C
7073	Animal Hunting: Sniper Shooting	Game	4.197727	0.0	50331648.0	50	Free	0.00	Teen	A

291 rows × 15 columns

17. Which is the free game with the most reviews?

In [251]:
```
filt1 = df['Distribution'] == 'Free'
filt2 = df['Category'] == 'Game'
df[filt1&filt2].sort_values(by='Reviews', ascending=False)
```

Out[251]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating
1879	Clash of Clans	Game	4.600000	44893888.0	102760448.0	100000000	Free	0.0	Everyone 10+
1917	Subway Surfers	Game	4.500000	27725352.0	79691776.0	1000000000	Free	0.0	Everyone 10+
1878	Clash Royale	Game	4.600000	23136735.0	101711872.0	100000000	Free	0.0	Everyone 10+
1966	Candy Crush Saga	Game	4.400000	22430188.0	77594624.0	500000000	Free	0.0	Everyone
1908	My Talking Tom	Game	4.500000	14892469.0	0.0	500000000	Free	0.0	Everyone
...	...	...	...	...	...	...	...	...	...
7073	Animal Hunting: Sniper Shooting	Game	4.197727	0.0	50331648.0	50	Free	0.0	Teen
8580	DM Adventure	Game	4.197727	0.0	11534336.0	10	Free	0.0	Everyone
5855	Ay Vamos - PJ. Balvin - Piano	Game	4.197727	0.0	30408704.0	5	Free	0.0	Everyone
5824	Cyborg AX-001	Game	4.197727	0.0	0.0	50	Free	0.0	Everyone 10+
6832	Bu Nedir ?	Game	4.197727	0.0	34603008.0	50	Free	0.0	Everyone

863 rows × 15 columns

In [253]:
```
filt1 = df['Distribution'] == 'Paid'
filt2 = df['Category'] == 'Game'
df[filt1&filt2].sort_values(by='Reviews', ascending=False)
```

Out[253]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genr
4034	Hitman Sniper	Game	4.600000	408292.0	30408704.0	10000000	Paid	0.99	Mature 17+	Acti
7417	Grand Theft Auto: San Andreas	Game	4.400000	348962.0	27262976.0	1000000	Paid	6.99	Mature 17+	Acti
5631	Five Nights at Freddy's	Game	4.600000	100805.0	52428800.0	1000000	Paid	2.99	Teen	Acti
8804	DraStic DS Emulator	Game	4.600000	87766.0	12582912.0	1000000	Paid	4.99	Everyone	Acti
10682	Fruit Ninja Classic	Game	4.300000	85468.0	37748736.0	1000000	Paid	0.99	Everyone	Arca
...	...	...	...	...	...	...	...	...	...	
10697	Mu.F.O.	Game	5.000000	2.0	16777216.0	1	Paid	0.99	Everyone	Arca
6555	Sic Bo	Game	4.197727	1.0	11534336.0	100	Paid	1.99	Teen	Ca
6277	Bi-Tank Ads Free	Game	4.197727	0.0	0.0	1	Paid	0.99	Everyone	Arca
5846	YAKALA AY	Game	4.197727	0.0	14680064.0	1	Paid	0.99	Everyone	Arca
4218	D+H Reaction Wall	Game	4.197727	0.0	0.0	1	Paid	0.99	Everyone	Arca

82 rows × 15 columns

18. How many Tb (tebibytes) were transferred (overall) for the most popular Lifestyle app?

In [257]:
```
filt = df['Category'] == 'Lifestyle'
df[filt].sort_values(by='Installs', ascending=False)
```

Out[257]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating
4587	Tinder	Lifestyle	4.000000	2789775.0	71303168.0	100000000	Free	0.00	Mature 17+
1584	Samsung+	Lifestyle	4.500000	82145.0	31457280.0	50000000	Free	0.00	Everyone
5581	Sleep as Android: Sleep cycle tracker, smart a...	Lifestyle	4.300000	246201.0	0.0	10000000	Free	0.00	Everyone
1633	Zara	Lifestyle	4.300000	95905.0	34603008.0	10000000	Free	0.00	Everyone
1595	Galaxy Gift	Lifestyle	4.400000	95557.0	16777216.0	10000000	Free	0.00	Everyone
...	...	...	...	...	...	...	...	...	...
8355	Aproveita DF	Lifestyle	4.197727	0.0	3460300.8	1	Free	0.00	Everyone
7231	CE AR LOG	Lifestyle	4.197727	0.0	24117248.0	1	Free	0.00	Everyone
9201	EB Experience	Lifestyle	4.197727	0.0	1887436.8	1	Free	0.00	Everyone
8509	Dr D K Olukoya	Lifestyle	4.197727	0.0	3460300.8	1	Free	0.00	Teen
9934	I'm Rich/Eu sou Rico/أنا/我很有/غني錢	Lifestyle	4.197727	0.0	41943040.0	0	Paid	399.99	Everyone

369 rows × 15 columns

In [259]:
```
app = df[filt].sort_values(by='Installs', ascending=False).iloc[0]
app
```

Out[259]:
```
App                    Tinder
Category            Lifestyle
Rating                    4.0
Reviews             2789775.0
Size               71303168.0
Installs            100000000
Type                     Free
Price                     0.0
Content Rating     Mature 17+
Genres              Lifestyle
Last Updated         2-Aug-18
Current Ver             9.5.0
Android Ver        4.4 and up
Reviews Numeric     2789775.0
Distribution             Free
Name: 4587, dtype: object
```

In [260]:
```
Size = app['Size'] * app['Installs']
Size
```

Out[260]:
```
7130316800000000.0
```

In [262]:
```
Size_TB = Size / (1024*1024*1024*1024)
Size_TB
```

Out[262]:
```
6484.9853515625
```