In [258]:

```python
import pandas as pd
import numpy as np
import seaborn as sns
import missingno as msno
```

In [259]:

```python
df = pd.read_csv('googleplaystore.csv')
df.sample(5)
```

Out[259]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
4059	Smash Hit	GAME	4.5	4147718	79M	100,000,000+	Free	0	Everyone	Arcade	26-Nov-15	1.4.0	2.3 and up
8434	Wallpapers DH 4K	PERSONALIZATION	3.8	23	5.9M	1,000+	Free	0	Everyone	Personalization	16-Jul-18	2	4.0.3 and up
6672	Zetup, print in one click	TOOLS	NaN	40	24M	1,000+	Free	0	Everyone	Tools	26-Jan-18	1.11.6	5.0 and up
8492	Salah Widget (DK+Malmo)	LIFESTYLE	4.5	532	6.4M	10,000+	Free	0	Everyone	Lifestyle	28-Jun-15	0.9.9.14	2.3 and up
8833	DS-11 form	BUSINESS	NaN	3	28M	100+	Free	0	Everyone	Business	27-Apr-18	1.7.7	4.1 and up

In [260]:

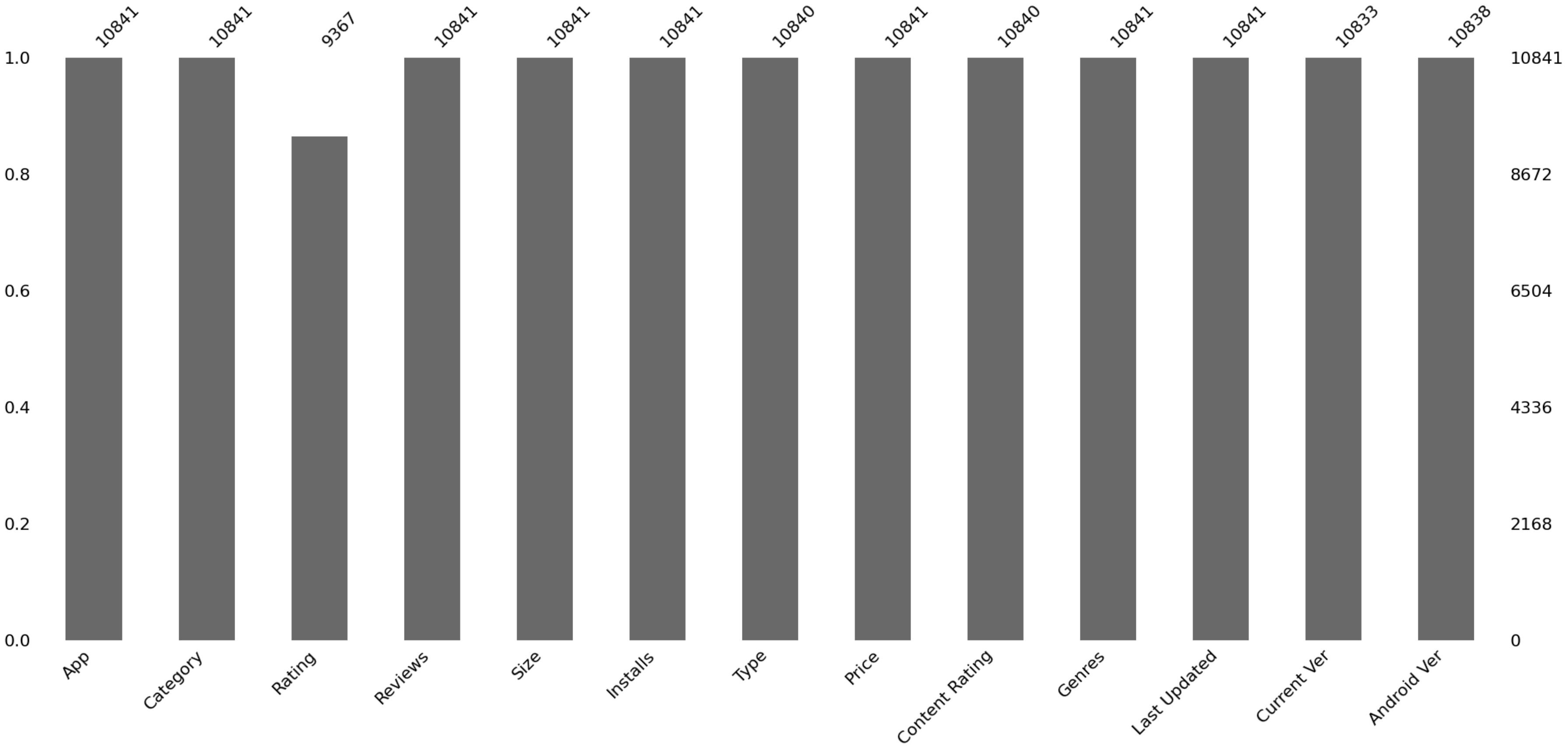
```
df.info()
```

```
<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

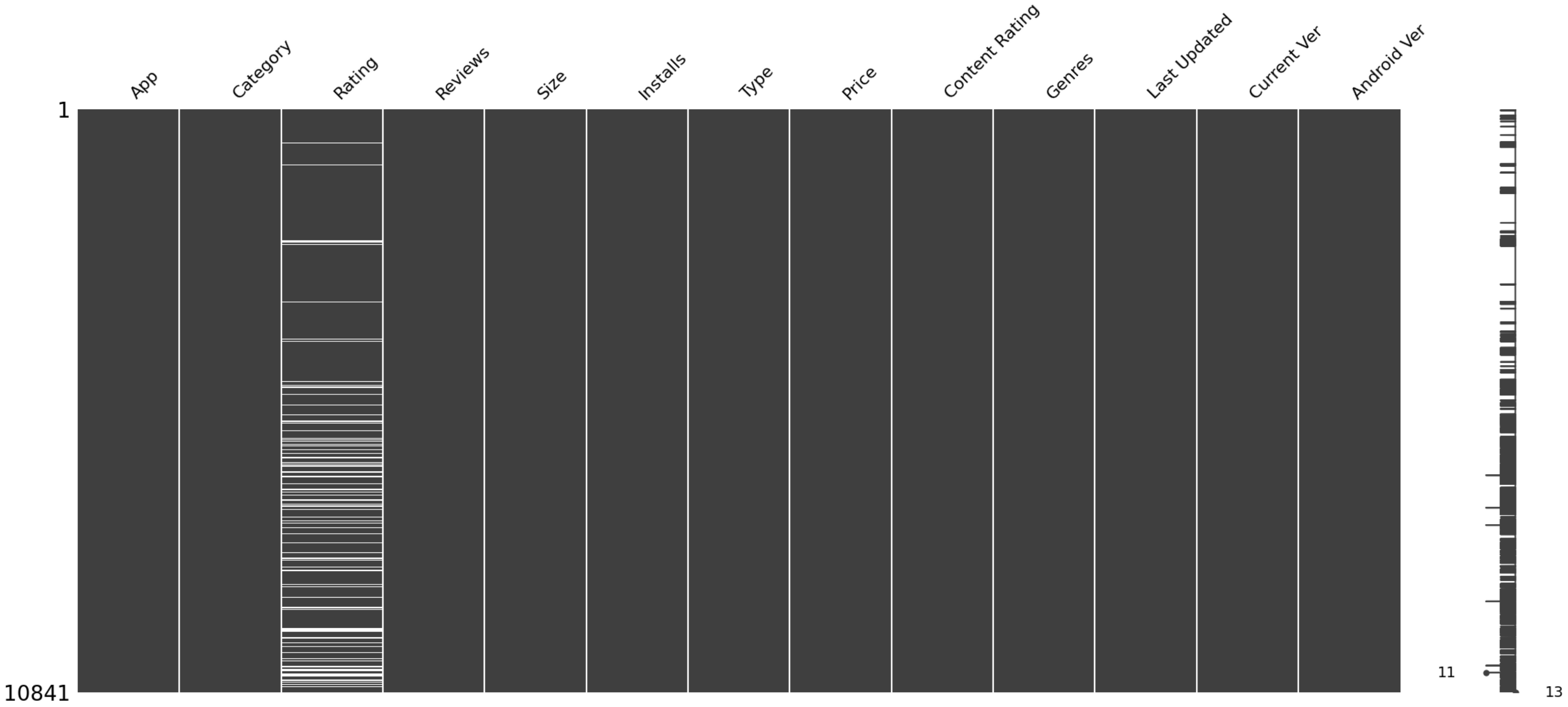
1. Which of the following column(s) has/have null values?

```
In [261]: msno.bar(df)
Out[261]: <Axes: >
```



```
In [262]: msno.matrix(df)
```

```
Out[262]: <Axes: >
```



In [263]:

```
df.isna().sum().sort_values(ascending=False)
```

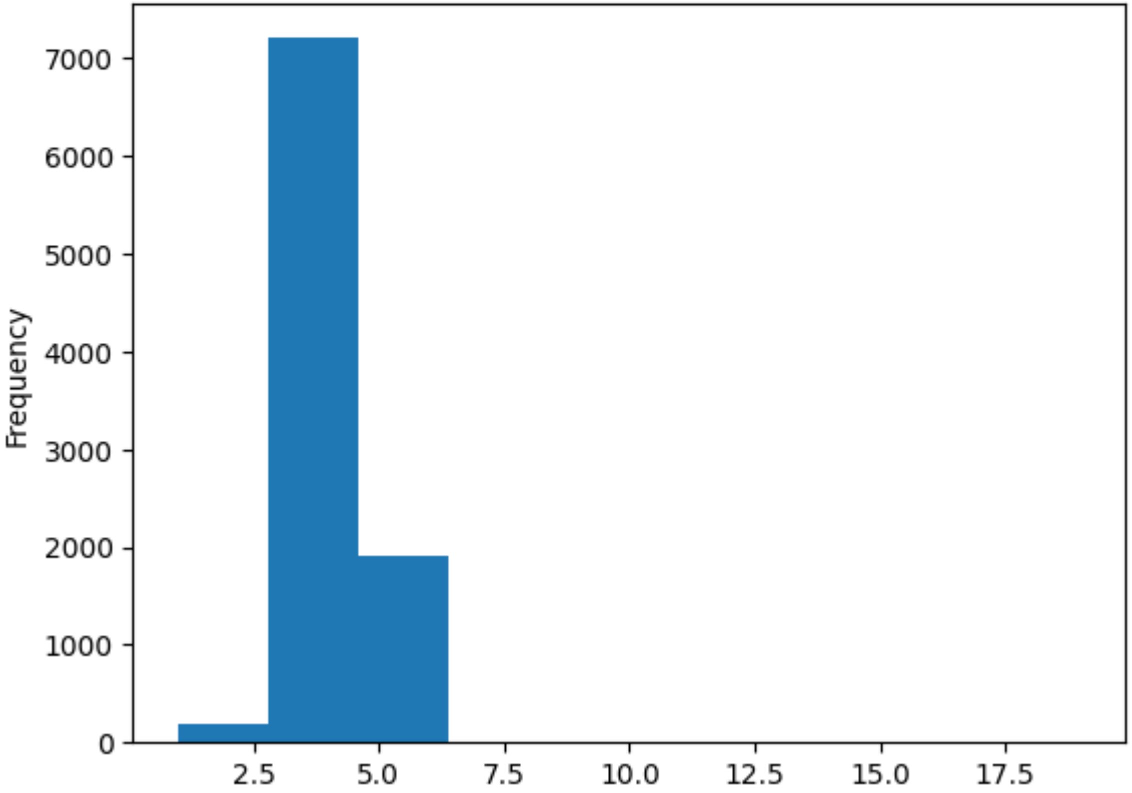
Out[263]:

```
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

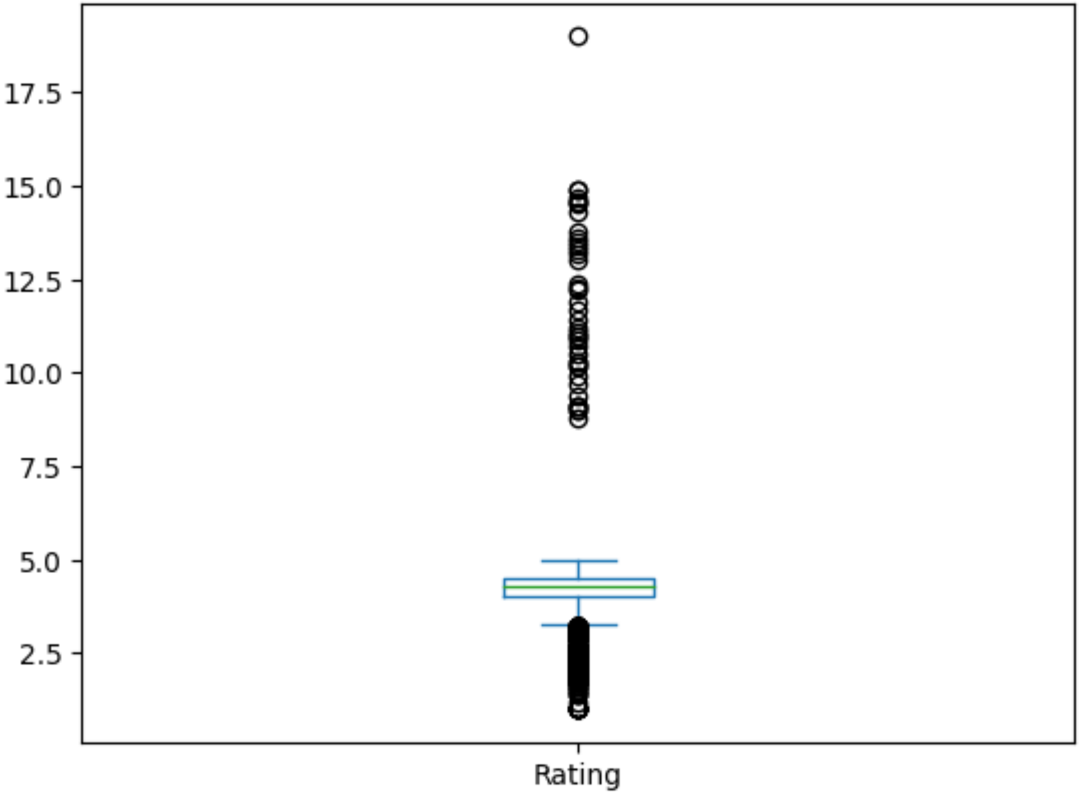
```
In [264]: df['Rating'].plot(kind='hist')
```

```
Out[264]: <Axes: ylabel='Frequency'>
```



```
In [265]: df['Rating'].plot(kind='box')
```

```
Out[265]: <Axes: >
```



In [266]:
```
df['Rating'].describe()
```

Out[266]:
```
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 [267]:
```
# Remove the invalid values from Rating (if any). Just set them as NaN
df.loc[df['Rating']>5,'Rating']=np.nan
df.head()
```

Out[267]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	Free	Everyone	Art & Design	7-Jan-18	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	15-Jan-18	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	1-Aug-18	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	8-Jun-18	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	20-Jun-18	1.1	4.4 and up

```
In [268]: df['Rating'].mean()

Out[268]: 4.197726785331332
```

```
In [269]: # Fill the null values in the Rating column using the mean()
          df['Rating'].fillna(df['Rating'].mean(),inplace=True)
```

```
In [270]: df.dropna(inplace=True)
          df.shape

Out[270]: (10829, 13)
```

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

```
In [271]: df['Reviews']

Out[271]: 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 [272]:

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

Out[272]:

```
0           159.0
1           967.0
2         87510.0
3        215644.0
4           967.0
           ...   
10836        38.0
10837         4.0
10838         3.0
10839       114.0
10840    398307.0
Name: Reviews Numeric, Length: 10829, dtype: float64
```

In [273]:

```
df.loc[df['Reviews Numeric'].isna()]
```

Out[273]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Revie Nume
72	Android Auto - Maps, Media, Messaging & Voice	AUTO_AND_VEHICLES	4.2	2M	16M	10,000,000+	Free	0	Teen	Auto & Vehicles	11-Jul-18	Varies with device	5.0 and up	Na
1778	Block Craft 3D: Building Simulator Games For Free	GAME	4.5	1M	57M	50,000,000+	Free	0	Everyone	Simulation	5-Mar-18	2.10.2	4.0.3 and up	Na
1781	Trivia Crack	GAME	4.5	6.4M	95M	100,000,000+	Free	0	Everyone	Trivia	3-Aug-18	2.79.0	4.1 and up	Na

In [274]:

```
# df.loc[[72,1778,1781],'Reviews']=[2_000_000,1_000_000,6_400_000]
```

In [275]:
```
df.loc[df['Reviews'].str.contains('M')]
```

Out[275]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Revie Nume
72	Android Auto - Maps, Media, Messaging & Voice	AUTO_AND_VEHICLES	4.2	2M	16M	10,000,000+	Free	0	Teen	Auto & Vehicles	11-Jul-18	Varies with device	5.0 and up	N
1778	Block Craft 3D: Building Simulator Games For Free	GAME	4.5	1M	57M	50,000,000+	Free	0	Everyone	Simulation	5-Mar-18	2.10.2	4.0.3 and up	N
1781	Trivia Crack	GAME	4.5	6.4M	95M	100,000,000+	Free	0	Everyone	Trivia	3-Aug-18	2.79.0	4.1 and up	N

In [276]:
```
df.loc[df['Reviews'].str.contains('M'),'Reviews'].str.replace('M','')
```

Out[276]:
```
72        2
1778      1
1781    6.4
Name: Reviews, dtype: object
```

In [277]:
```
(pd.to_numeric(df.loc[df['Reviews'].str.contains('M'),'Reviews'].str.replace('M',''))*1_000_000).astype(str
```

Out[277]:
```
72      2000000.0
1778    1000000.0
1781    6400000.0
Name: Reviews, dtype: object
```

In [278]:
```
df.loc[df['Reviews'].str.contains('M'),'Reviews'] = (pd.to_numeric(df.loc[df['Reviews'].str.contains('M'),\
                                                     'Reviews'].str.replace('M',''))*1_000_000).astype(str)
```

In [279]:

```
df.loc[[72,1778,1781]]
```

Out[279]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Revi Num
72	Android Auto - Maps, Media, Messaging & Voice	AUTO_AND_VEHICLES	4.2	2000000.0	16M	10,000,000+	Free	0	Teen	Auto & Vehicles	11-Jul-18	Varies with device	5.0 and up	
1778	Block Craft 3D: Building Simulator Games For Free	GAME	4.5	1000000.0	57M	50,000,000+	Free	0	Everyone	Simulation	5-Mar-18	2.10.2	4.0.3 and up	
1781	Trivia Crack	GAME	4.5	6400000.0	95M	100,000,000+	Free	0	Everyone	Trivia	3-Aug-18	2.79.0	4.1 and up	

```
In [280]: df
```

Out[280]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	A
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.100000	159	19M	10,000+	Free	Free	Everyone	Art & Design	7-Jan-18	1.0.0	
1	Coloring book moana	ART_AND_DESIGN	3.900000	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	15-Jan-18	2.0.0	
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.700000	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	1-Aug-18	1.2.4	
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.500000	215644	25M	50,000,000+	Free	0	Teen	Art & Design	8-Jun-18	Varies with device	
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.300000	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	20-Jun-18	1.1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
10836	Sya9a Maroc - FR	FAMILY	4.500000	38	53M	5,000+	Free	0	Everyone	Education	25-Jul-17	1.48	
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.000000	4	3.6M	100+	Free	0	Everyone	Education	6-Jul-18	1	
10838	Parkinson Exercices FR	MEDICAL	4.197727	3	9.5M	1,000+	Free	0	Everyone	Medical	20-Jan-17	1	

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	A
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.500000	114	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	19-Jan-15	Varies with device	
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.500000	398307	19M	10,000,000+	Free	0	Everyone	Lifestyle	25-Jul-18	Varies with device	

In [281]:
```
df.drop('Reviews Numeric',axis='columns',inplace=True)
```

In [282]:
```
df.head()
```

Out[282]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	Free	Everyone	Art & Design	7-Jan-18	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	15-Jan-18	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	1-Aug-18	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	8-Jun-18	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	20-Jun-18	1.1	4.4 and up

In [283]:
```
df['Reviews'] = pd.to_numeric(df['Reviews'])
```

4. How many duplicated apps are there?

```
In [284]: df.duplicated(subset=['App'],keep=False).sum()
```

```
Out[284]: 1979
```

In [285]:

```python
df.loc[df.duplicated(subset=['App'],keep=False)].sort_values(by='App')
```

Out[285]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
1393	10 Best Foods for You	HEALTH_AND_FITNESS	4.0	2490.0	3.8M	500,000+	Free	0	Everyone 10+	Health & Fitness	17-Feb-17	1.9	2.3.3 and up
1407	10 Best Foods for You	HEALTH_AND_FITNESS	4.0	2490.0	3.8M	500,000+	Free	0	Everyone 10+	Health & Fitness	17-Feb-17	1.9	2.3.3 and up
2543	1800 Contacts - Lens Store	MEDICAL	4.7	23160.0	26M	1,000,000+	Free	0	Everyone	Medical	27-Jul-18	7.4.1	5.0 and up
2322	1800 Contacts - Lens Store	MEDICAL	4.7	23160.0	26M	1,000,000+	Free	0	Everyone	Medical	27-Jul-18	7.4.1	5.0 and up
2385	2017 EMRA Antibiotic Guide	MEDICAL	4.4	12.0	3.8M	1,000+	Paid	$16.99	Everyone	Medical	27-Jan-17	1.0.5	4.0.3 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3202	trivago: Hotels & Travel	TRAVEL_AND_LOCAL	4.2	219848.0	Varies with device	50,000,000+	Free	0	Everyone	Travel & Local	2-Aug-18	Varies with device	Varies with device
3118	trivago: Hotels & Travel	TRAVEL_AND_LOCAL	4.2	219848.0	Varies with device	50,000,000+	Free	0	Everyone	Travel & Local	2-Aug-18	Varies with device	Varies with device
3103	trivago: Hotels & Travel	TRAVEL_AND_LOCAL	4.2	219848.0	Varies with device	50,000,000+	Free	0	Everyone	Travel & Local	2-Aug-18	Varies with device	Varies with device
8291	wetter.com - Weather and Radar	WEATHER	4.2	189310.0	38M	10,000,000+	Free	0	Everyone	Weather	6-Aug-18	Varies with device	Varies with device
3652	wetter.com - Weather and Radar	WEATHER	4.2	189313.0	38M	10,000,000+	Free	0	Everyone	Weather	6-Aug-18	Varies with device	Varies with device

1979 rows × 13 columns

```
In [286]: df.loc[df.duplicated(subset=['App'],keep=False) & ~df.duplicated(keep=False)].sort_values(by='App')
```

Out[286]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
3083	365Scores - Live Scores	SPORTS	4.6	666521.0	25M	10,000,000+	Free	0	Everyone	Sports	29-Jul-18	5.5.9	4.1 and up
5415	365Scores - Live Scores	SPORTS	4.6	666246.0	25M	10,000,000+	Free	0	Everyone	Sports	29-Jul-18	5.5.9	4.1 and up
1675	8 Ball Pool	GAME	4.5	14198297.0	52M	100,000,000+	Free	0	Everyone	Sports	31-Jul-18	4.0.0	4.0.3 and up
1703	8 Ball Pool	GAME	4.5	14198602.0	52M	100,000,000+	Free	0	Everyone	Sports	31-Jul-18	4.0.0	4.0.3 and up
1755	8 Ball Pool	GAME	4.5	14200344.0	52M	100,000,000+	Free	0	Everyone	Sports	31-Jul-18	4.0.0	4.0.3 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
565	stranger chat - anonymous chat	DATING	3.5	13204.0	6.1M	1,000,000+	Free	0	Mature 17+	Dating	7-Jul-18	2.4.1	4.1 and up
2590	textPlus: Free Text & Calls	SOCIAL	4.1	382120.0	28M	10,000,000+	Free	0	Everyone	Social	26-Jul-18	7.3.1	4.1 and up
2637	textPlus: Free Text & Calls	SOCIAL	4.1	382121.0	28M	10,000,000+	Free	0	Everyone	Social	26-Jul-18	7.3.1	4.1 and up
3652	wetter.com - Weather and Radar	WEATHER	4.2	189313.0	38M	10,000,000+	Free	0	Everyone	Weather	6-Aug-18	Varies with device	Varies with device
8291	wetter.com - Weather and Radar	WEATHER	4.2	189310.0	38M	10,000,000+	Free	0	Everyone	Weather	6-Aug-18	Varies with device	Varies with device

1099 rows × 13 columns

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

In [287]:

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

Out[287]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
3083	365Scores - Live Scores	SPORTS	4.6	666521.0	25M	10,000,000+	Free	0	Everyone	Sports	29-Jul-18	5.5.9	4.1 and up
5415	365Scores - Live Scores	SPORTS	4.6	666246.0	25M	10,000,000+	Free	0	Everyone	Sports	29-Jul-18	5.5.9	4.1 and up
1675	8 Ball Pool	GAME	4.5	14198297.0	52M	100,000,000+	Free	0	Everyone	Sports	31-Jul-18	4.0.0	4.0.3 and up
1703	8 Ball Pool	GAME	4.5	14198602.0	52M	100,000,000+	Free	0	Everyone	Sports	31-Jul-18	4.0.0	4.0.3 and up
1755	8 Ball Pool	GAME	4.5	14200344.0	52M	100,000,000+	Free	0	Everyone	Sports	31-Jul-18	4.0.0	4.0.3 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
565	stranger chat - anonymous chat	DATING	3.5	13204.0	6.1M	1,000,000+	Free	0	Mature 17+	Dating	7-Jul-18	2.4.1	4.1 and up
2590	textPlus: Free Text & Calls	SOCIAL	4.1	382120.0	28M	10,000,000+	Free	0	Everyone	Social	26-Jul-18	7.3.1	4.1 and up
2637	textPlus: Free Text & Calls	SOCIAL	4.1	382121.0	28M	10,000,000+	Free	0	Everyone	Social	26-Jul-18	7.3.1	4.1 and up
3652	wetter.com - Weather and Radar	WEATHER	4.2	189313.0	38M	10,000,000+	Free	0	Everyone	Weather	6-Aug-18	Varies with device	Varies with device
8291	wetter.com - Weather and Radar	WEATHER	4.2	189310.0	38M	10,000,000+	Free	0	Everyone	Weather	6-Aug-18	Varies with device	Varies with device

1099 rows × 13 columns

In [288]:

```
df_copy = df.copy()
```

In [289]:

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

```
In [290]: # help(df.duplicated)
```

```
In [291]: df.drop_duplicates(subset='App',keep='last',inplace=True)
          df.head()
```

Out[291]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
8884	"i DT" Fútbol. Todos Somos Técnicos.	SPORTS	4.197727	27.0	3.6M	500+	Free	0	Everyone	Sports	7-Oct-17	0.22	4.1 and up
324	#NAME?	COMICS	3.500000	115.0	9.1M	10,000+	Free	0	Mature 17+	Comics	13-Jul-18	5.0.12	5.0 and up
8532	+Download 4 Instagram Twitter	SOCIAL	4.500000	40467.0	22M	1,000,000+	Free	0	Everyone	Social	2-Aug-18	5.03	4.1 and up
4541	.R	TOOLS	4.500000	259.0	203k	10,000+	Free	0	Everyone	Tools	16-Sep-14	1.1.06	1.5 and up
4636	/u/app	COMMUNICATION	4.700000	573.0	53M	10,000+	Free	0	Mature 17+	Communication	3-Jul-18	4.2.4	4.1 and up

6. Format the `Category` column

```
In [292]: df['Category'] = df['Category'].str.capitalize()
          df['Category'] = df['Category'].str.replace('_',' ')
          df['Category']
```

```
Out[292]: 8884           Sports
          324            Comics
          8532           Social
          4541            Tools
          4636    Communication
                      ...
          6334    Video players
          4362        Lifestyle
          2575           Social
          7559            Tools
          882     Entertainment
          Name: Category, Length: 9648, dtype: object
```

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

```
In [293]: df['Installs'] = df['Installs'].str.replace('+','')
          df['Installs'] = df['Installs'].str.replace(',','')
          df['Installs']
```

```
Out[293]: 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 [294]:
```
df['Installs'] = pd.to_numeric(df['Installs'])
df['Installs']
```

Out[294]:
```
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
```

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

In [295]:
```
df_copy = df.copy()
```

In [296]:
```
df = df_copy.copy()
```

```
In [297]: pd.to_numeric(df['Size'].str.replace('M','').str.replace('k',''))
```

```
---------------------------------------------------------------------------
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 "Varies with device"

During handling of the above exception, another exception occurred:

ValueError                                Traceback (most recent call last)
Cell In[297], line 1
----> 1 pd.to_numeric(df['Size'].str.replace('M','').str.replace('k',''))

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 "Varies with device" at position 25
```

```
In [298]: df.loc[df['Size']=='Varies with device','Size']= '0'
          df['Size']
```

```
Out[298]: 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 [299]: pd.to_numeric(df['Size'].str.replace('M','').str.replace('k',''))
```

```
Out[299]: 8884      3.6
          324       9.1
          8532     22.0
          4541    203.0
          4636     53.0
                  ...
          6334     59.0
          4362     26.0
          2575     18.0
          7559      3.2
          882       4.0
          Name: Size, Length: 9648, dtype: float64
```

In [300]:

```
df.loc[df['Size'].str.contains('k')]
```

Out[300]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
4541	.R	Tools	4.500000	259.0	203k	10000	Free	0	Everyone	Tools	16-Sep-14	1.1.06	1.5 and up
4897	30-Day Ab Challenge Tracker	Health and fitness	3.500000	224.0	371k	10000	Free	0	Everyone	Health & Fitness	9-Jul-14	1.2.6	4.1 and up
6671	4-T's Bar-BQ & Catering	Shopping	4.197727	0.0	243k	10	Free	0	Everyone	Shopping	16-Jan-17	1.0.1	4.1 and up
4871	A-B repeater	Video players	4.400000	32.0	239k	5000	Free	0	Everyone	Video Players & Editors	18-May-18	1.8	3.0 and up
5035	AE Checkout Plugin	Shopping	3.800000	208.0	78k	10000	Free	0	Everyone	Shopping	11-Feb-15	1.3.1	2.3 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5482	meStudying: AP English Lit	Family	5.000000	1.0	655k	10	Paid	$4.99	Everyone	Education	31-Aug-13	1.3	2.0.1 and up
7370	mySharpBranded CI Test	Tools	4.197727	0.0	898k	1	Free	0	Everyone	Tools	4-Oct-17	1.0.7	4.2 and up
9333	qEG APP / Química EG SRL	Tools	4.197727	0.0	118k	10	Free	0	Everyone	Tools	24-Jul-18	1	4.0 and up
8148	signály.cz	Social	4.197727	38.0	881k	1000	Free	0	Everyone	Social	9-May-13	1.1	2.2 and up
5832	¡Ay Caramba!	Family	4.197727	0.0	549k	1	Paid	$1.99	Everyone	Education	13-Jun-14	1.2	3.0 and up

310 rows × 13 columns

```
In [301]: df.loc[df['Size'].str.contains('k'),'Size'].str.replace('k','')

Out[301]: 4541    203
          4897    371
          6671    243
          4871    239
          5035     78
                 ...
          5482    655
          7370    898
          9333    118
          8148    881
          5832    549
          Name: Size, Length: 310, dtype: object

In [302]: df.loc[df['Size'].str.contains('k'),'Size'] = (pd.to_numeric(df.loc[df['Size'].str.contains('k'),\
                                                                     'Size'].str.replace('k',''))*1024).astyp

In [303]: df.loc[df['Size'].str.contains('M'),'Size']

Out[303]: 8884    3.6M
          324     9.1M
          8532     22M
          4636     53M
          5940     14M
                  ...
          6334     59M
          4362     26M
          2575     18M
          7559    3.2M
          882     4.0M
          Name: Size, Length: 8111, dtype: object

In [304]: df.loc[df['Size'].str.contains('M'),'Size'] = (pd.to_numeric(df.loc[df['Size'].str.contains('M'),\
                                                                     'Size'].str.replace('M',''))*1024*1024).a
```

```
In [305]: df['Size'] = pd.to_numeric(df['Size'])
          df['Size']
```

```
Out[305]: 8884     3774873.6
          324      9542041.6
          8532    23068672.0
          4541      207872.0
          4636    55574528.0
                     ...
          6334    61865984.0
          4362    27262976.0
          2575    18874368.0
          7559     3355443.2
          882      4194304.0
          Name: Size, Length: 9648, dtype: float64
```

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

```
In [310]: df['Price'] = df['Price'].str.replace('$','')
          df['Price'] = df['Price'].str.replace('Free','0')
```

```
In [311]: df['Price'] = pd.to_numeric(df['Price'])
          df['Price']
```

```
Out[311]: 8884      0.00
          324       0.00
          8532      0.00
          4541      0.00
          4636      0.00
                   ...
          6334      0.00
          4362    399.99
          2575      0.00
          7559      0.00
          882       0.00
          Name: Price, Length: 9648, dtype: float64
```

10. Paid or free?

In [314]:
```
df['Distribution']= 'Paid'
df.head()
```

Out[314]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Distı
8884	"i DT" Fútbol. Todos Somos Técnicos.	Sports	4.197727	27.0	3774873.6	500	Free	0.0	Everyone	Sports	7-Oct-17	0.22	4.1 and up	
324	#NAME?	Comics	3.500000	115.0	9542041.6	10000	Free	0.0	Mature 17+	Comics	13-Jul-18	5.0.12	5.0 and up	
8532	+Download 4 Instagram Twitter	Social	4.500000	40467.0	23068672.0	1000000	Free	0.0	Everyone	Social	2-Aug-18	5.03	4.1 and up	
4541	.R	Tools	4.500000	259.0	207872.0	10000	Free	0.0	Everyone	Tools	16-Sep-14	1.1.06	1.5 and up	
4636	/u/app	Communication	4.700000	573.0	55574528.0	10000	Free	0.0	Mature 17+	Communication	3-Jul-18	4.2.4	4.1 and up	

In [315]:
```
df.loc[df['Price']==0,'Distribution']='Free'
```

In [316]:
```
df.head()
```

Out[316]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Distı
8884	"i DT" Fútbol. Todos Somos Técnicos.	Sports	4.197727	27.0	3774873.6	500	Free	0.0	Everyone	Sports	7-Oct-17	0.22	4.1 and up	
324	#NAME?	Comics	3.500000	115.0	9542041.6	10000	Free	0.0	Mature 17+	Comics	13-Jul-18	5.0.12	5.0 and up	
8532	+Download 4 Instagram Twitter	Social	4.500000	40467.0	23068672.0	1000000	Free	0.0	Everyone	Social	2-Aug-18	5.03	4.1 and up	
4541	.R	Tools	4.500000	259.0	207872.0	10000	Free	0.0	Everyone	Tools	16-Sep-14	1.1.06	1.5 and up	
4636	/u/app	Communication	4.700000	573.0	55574528.0	10000	Free	0.0	Mature 17+	Communication	3-Jul-18	4.2.4	4.1 and up	

Analysis

11. Which app has the most reviews?

In [319]:
```
df.loc[df['Reviews']==df['Reviews'].max()]
```

Out[319]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Distribution
2544	Facebook	Social	4.1	78158306.0	0.0	1000000000	Free	0.0	Teen	Social	3-Aug-18	Varies with device	Varies with device	Free

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

In [323]:

```
df['Category'].value_counts().sort_values(ascending=False)
```

Out[323]:

```
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 [324]:
```
df.loc[df['Price']==df['Price'].max()]
```

Out[324]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Distribution
4367	I'm Rich - Trump Edition	Lifestyle	3.6	275.0	7654604.8	10000	Paid	400.0	Everyone	Lifestyle	3-May-18	1.0.1	4.1 and up	Paid

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

In [329]:
```
df_game = df.loc[df['Category']=='Game']
df_game.sort_values(by='Price',ascending=False)
```

Out[329]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Distributic
4203	The World Ends With You	Game	4.6	4108.0	13631488.0	10000	Paid	17.99	Everyone 10+	Arcade	14-Dec-15	1.0.4	4.0 and up	Pa
10782	Trine 2: Complete Story	Game	3.8	252.0	11534336.0	10000	Paid	16.99	Teen	Action	27-Feb-15	2.22	5.0 and up	Pa
6341	Blackjack Verite Drills	Game	4.6	17.0	4928307.2	100	Paid	14.00	Teen	Casino	9-Jul-17	1.1.10	3.0 and up	Pa
1838	Star Wars ™: DIRTY	Game	4.5	38207.0	15728640.0	100000	Paid	9.99	Teen	Role Playing	19-Oct-15	1.0.6	4.1 and up	Pa
6198	Backgammon NJ for Android	Game	4.4	1644.0	15728640.0	10000	Paid	7.99	Everyone	Board	5-Apr-17	4.1	2.3.3 and up	Pa
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
7600	Dreamland Arcade - Steven Universe	Game	4.0	6386.0	25165824.0	500000	Free	0.00	Everyone	Arcade	19-Nov-17	0.99	5.1 and up	Fre
10522	Drift Legends	Game	4.2	33788.0	28311552.0	1000000	Free	0.00	Everyone	Racing	29-Mar-18	1.8.5	4.1 and up	Fre
4434	Drink-O-Tron The Drinking Game	Game	4.1	140.0	47185920.0	50000	Free	0.00	Mature 17+	Card	31-May-17	1.64	4.0.3 and up	Fre
10508	Drive 4x4 Luxury SUV Jeep	Game	4.2	2183.0	48234496.0	500000	Free	0.00	Everyone	Racing	10-Jul-18	1.12	2.3 and up	Fre
3960	► MultiCraft — Free Miner! 👍	Game	4.3	1305050.0	0.0	50000000	Free	0.00	Everyone 10+	Adventure	29-Jul-18	1.1.11.11	4.1 and up	Fre

945 rows × 14 columns

15. Which is the most popular Finance App?

In [333]:
```
df_finance = df.loc[df['Category']=='Finance']
df_finance.sort_values(by='Installs',ascending=False)
```

Out[333]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Distribu
5601	Google Pay	Finance	4.200000	348132.0	0.0	100000000	Free	0.00	Everyone	Finance	26-Jul-18	2.70.206190089	Varies with device	I
1156	PayPal	Finance	4.300000	659760.0	49283072.0	50000000	Free	0.00	Everyone	Finance	18-Jul-18	6.28.0	4.4 and up	I
1081	İşCep	Finance	4.500000	381788.0	33554432.0	10000000	Free	0.00	Everyone	Finance	2-Aug-18	3.22.0	4.1 and up	I
1168	Wells Fargo Mobile	Finance	4.400000	250719.0	38797312.0	10000000	Free	0.00	Everyone	Finance	31-Jul-18	6.8.0.109	5.0 and up	I
1169	Capital One® Mobile	Finance	4.600000	510401.0	82837504.0	10000000	Free	0.00	Everyone	Finance	1-Aug-18	5.38.1	5.0 and up	I
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
10417	FH Wallet	Finance	4.197727	0.0	10380902.4	1	Free	0.00	Everyone	Finance	26-Jul-18	1.0.0	4.1 and up	I
9101	amm dz	Finance	4.197727	0.0	14680064.0	1	Paid	5.99	Everyone	Finance	8-Jul-18	1	4.2 and up	I
10745	FP Boss	Finance	4.197727	1.0	6081740.8	1	Free	0.00	Everyone	Finance	27-Jul-18	1.0.2	5.0 and up	I
9905	Eu sou Rico	Finance	4.197727	0.0	2726297.6	0	Paid	30.99	Everyone	Finance	9-Jan-18	1	4.0 and up	I
9917	Eu Sou Rico	Finance	4.197727	0.0	1468006.4	0	Paid	394.99	Everyone	Finance	11-Jul-18	1	4.0.3 and up	I

345 rows × 14 columns

16. What Teen Game has the most reviews?

```
In [336]: df_game.loc[df_game['Content Rating']=='Teen'].sort_values(by='Reviews',ascending=False)
```

Out[336]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Distributic
3912	Asphalt 8: Airborne	Game	4.500000	8389714.0	96468992.0	100000000	Free	0.00	Teen	Racing	4-Jul-18	3.7.1a	4.0.3 and up	Fre
5417	Mobile Legends: Bang Bang	Game	4.400000	8219586.0	103809024.0	100000000	Free	0.00	Teen	Action	24-Jul-18	1.2.97.3042	4.0.3 and up	Fre
1988	Hungry Shark Evolution	Game	4.500000	6074627.0	104857600.0	100000000	Free	0.00	Teen	Arcade	25-Jul-18	6.0.0	4.1 and up	Fre
10327	Garena Free Fire	Game	4.500000	5534114.0	55574528.0	100000000	Free	0.00	Teen	Action	3-Aug-18	1.21.0	4.0.3 and up	Fre
3967	Pixel Gun 3D: Survival shooter & Battle Royale	Game	4.500000	4487182.0	57671680.0	50000000	Free	0.00	Teen	Action	4-Jul-18	15.1.2	4.0.3 and up	Fre
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4431	Obbligo o Verità? PRO	Game	4.197727	4.0	3040870.4	100	Paid	0.99	Teen	Board	26-Apr-18	1,01	3.0 and up	Pa
6335	BJ card game blackjack	Game	4.197727	3.0	22020096.0	500	Free	0.00	Teen	Card	2-Dec-16	1	2.3 and up	Fre
6555	Sic Bo	Game	4.197727	1.0	11534336.0	100	Paid	1.99	Teen	Card	27-Aug-13	1.0.0	2.2 and up	Pa
6329	Basic Strategy Training BJ 21	Game	4.197727	0.0	24117248.0	500	Free	0.00	Teen	Casino	7-Mar-16	1.1	2.3 and up	Fre

App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Distributic
Animal Hunting:												4.0 and	

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

In [337]:
```
df_game.loc[df_game['Distribution']=='Free'].sort_values(by='Reviews',ascending=False)
```

Out[337]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Distribu
1879	Clash of Clans	Game	4.600000	44893888.0	102760448.0	100000000	Free	0.0	Everyone 10+	Strategy	15-Jul-18	10.322.16	4.1 and up	
1917	Subway Surfers	Game	4.500000	27725352.0	79691776.0	1000000000	Free	0.0	Everyone 10+	Arcade	12-Jul-18	1.90.0	4.1 and up	
1878	Clash Royale	Game	4.600000	23136735.0	101711872.0	100000000	Free	0.0	Everyone 10+	Strategy	27-Jun-18	2.3.2	4.1 and up	
1966	Candy Crush Saga	Game	4.400000	22430188.0	77594624.0	500000000	Free	0.0	Everyone	Casual	5-Jul-18	1.129.0.2	4.1 and up	
1908	My Talking Tom	Game	4.500000	14892469.0	0.0	500000000	Free	0.0	Everyone	Casual	19-Jul-18	4.8.0.132	4.1 and up	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
7073	Animal Hunting: Sniper Shooting	Game	4.197727	0.0	50331648.0	50	Free	0.0	Teen	Action	6-Jul-18	1	4.0 and up	
8580	DM Adventure	Game	4.197727	0.0	11534336.0	10	Free	0.0	Everyone	Adventure	18-Jun-18	1.0.4	2.3 and up	
5855	Ay Vamos - PJ. Balvin - Piano	Game	4.197727	0.0	30408704.0	5	Free	0.0	Everyone	Arcade	9-Jul-18	1	4.1 and up	
5824	Cyborg AX-001	Game	4.197727	0.0	0.0	50	Free	0.0	Everyone 10+	Action	25-Jun-18	Varies with device	Varies with device	
6832	Bu Nedir ?	Game	4.197727	0.0	34603008.0	50	Free	0.0	Everyone	Trivia	15-Apr-18	3.1.6z	4.0.3 and up	

863 rows × 14 columns

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

```
In [341]: df_lifestyle = df.loc[df['Category']=='Lifestyle']
          df_lifestyle.sort_values(by='Installs',ascending=False)
```

Out[341]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver	Distributio
4587	Tinder	Lifestyle	4.000000	2789775.0	71303168.0	100000000	Free	0.00	Mature 17+	Lifestyle	2-Aug-18	9.5.0	4.4 and up	Fre
1584	Samsung+	Lifestyle	4.500000	82145.0	31457280.0	50000000	Free	0.00	Everyone	Lifestyle	5-Jul-18	10.19.0.0	4.4 and up	Fre
5581	Sleep as Android: Sleep cycle tracker, smart a...	Lifestyle	4.300000	246201.0	0.0	10000000	Free	0.00	Everyone	Lifestyle	23-Jul-18	Varies with device	Varies with device	Fre
1633	Zara	Lifestyle	4.300000	95905.0	34603008.0	10000000	Free	0.00	Everyone	Lifestyle	30-Jul-18	4.1.0	4.1 and up	Fre
1595	Galaxy Gift	Lifestyle	4.400000	95557.0	16777216.0	10000000	Free	0.00	Everyone	Lifestyle	23-Jul-18	7.0.8	4.0.3 and up	Fre
...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
8355	Aproveita DF	Lifestyle	4.197727	0.0	3460300.8	1	Free	0.00	Everyone	Lifestyle	2-Aug-18	1	4.0 and up	Fre
7231	CE AR LOG	Lifestyle	4.197727	0.0	24117248.0	1	Free	0.00	Everyone	Lifestyle	11-Jul-18	1.0.1	4.1 and up	Fre
9201	EB Experience	Lifestyle	4.197727	0.0	1887436.8	1	Free	0.00	Everyone	Lifestyle	19-Oct-17	1	4.0.3 and up	Fre
8509	Dr D K Olukoya	Lifestyle	4.197727	0.0	3460300.8	1	Free	0.00	Teen	Lifestyle	25-Jul-18	1	4.1 and up	Fre
9934	I'm Rich/Eu sou Rico/أنا غني/我很有錢	Lifestyle	4.197727	0.0	41943040.0	0	Paid	399.99	Everyone	Lifestyle	1-Dec-17	MONEY	4.1 and up	Pai

```
369 rows × 14 columns
```

In [345]:
```
top_app_lifestyle = df_lifestyle.sort_values(by='Installs',ascending=False).iloc[0]
```

In [346]:
```
top_app_lifestyle
```

Out[346]:
```
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
Distribution               Free
Name: 4587, dtype: object
```

In [351]:
```
overall = top_app_lifestyle['Installs']*top_app_lifestyle['Size']/(1024**4)
overall
```

Out[351]:
```
6484.9853515625
```

In [ ]: