1. Load the data file using pandas.


```
[3]: import pandas as pd
     import numpy as np
     import seaborn as sns
```

```
[6]: data = pd.read_csv('googleplaystore.csv')
```

```
[7]: data.head()
```

[7]:

	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	0	Everyone	Art & Design	January 7, 2018	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	January 15, 2018	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	August 1, 2018	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	June 8, 2018	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	June 20, 2018	1.1	4.4 and up

```
data.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.shape
```

```
(10841, 13)
```

2. Check for null values in the data. Get the number of null values for each column.

```
[11]: data.isnull().any()
```

```
[11]: App              False
      Category         False
      Rating            True
      Reviews          False
      Size             False
      Installs         False
      Type              True
      Price            False
      Content Rating    True
      Genres           False
      Last Updated     False
      Current Ver       True
      Android Ver       True
      dtype: bool
```

```
[12]: data.isnull().sum()
```

```
[12]: 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
```

3. Drop records with nulls in any of the columns.

```
[15]: data = data.dropna()
```

```
[16]: data.isnull().any()
```

```
[16]: App              False
      Category         False
      Rating           False
      Reviews          False
      Size             False
      Installs         False
      Type             False
      Price            False
      Content Rating   False
      Genres           False
      Last Updated     False
      Current Ver      False
      Android Ver      False
      dtype: bool
```

```
[17]: data.shape
```

```
[17]: (9360, 13)
```

4.1 Size column has sizes in Kb as well as Mb

```
[18]: data["Size"] = [ float(i.split('M')[0]) if 'M' in i else float(0) for i in data["Size"]
      ]
```

```
[19]: data.head()
```

[19]:

	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	19.0	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14.0	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25.0	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

```
[20]: data["Size"] = 1000 * data["Size"]
```

```
[21]: data
```

[21]:

	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	19000.0	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	7	2600.0	500+	Free	0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	0.0	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10,000,000+	Free	0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

9360 rows × 13 columns

4.2 Reviews is a numeric field that is loaded as a string field

[22]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9360 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             9360 non-null   object 
 1   Category        9360 non-null   object 
 2   Rating          9360 non-null   float64
 3   Reviews         9360 non-null   object 
 4   Size            9360 non-null   float64
 5   Installs        9360 non-null   object 
 6   Type            9360 non-null   object 
 7   Price           9360 non-null   object 
 8   Content Rating  9360 non-null   object 
 9   Genres          9360 non-null   object 
 10  Last Updated    9360 non-null   object 
 11  Current Ver     9360 non-null   object 
 12  Android Ver     9360 non-null   object 
dtypes: float64(2), object(11)
memory usage: 1023.8+ KB
```

[23]:
```
data["Reviews"] = data["Reviews"].astype(float)
```

[24]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9360 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             9360 non-null   object 
 1   Category        9360 non-null   object 
 2   Rating          9360 non-null   float64
 3   Reviews         9360 non-null   float64
 4   Size            9360 non-null   float64
 5   Installs        9360 non-null   object 
 6   Type            9360 non-null   object 
 7   Price           9360 non-null   object 
 8   Content Rating  9360 non-null   object 
 9   Genres          9360 non-null   object 
 10  Last Updated    9360 non-null   object 
 11  Current Ver     9360 non-null   object 
 12  Android Ver     9360 non-null   object 
dtypes: float64(3), object(10)
memory usage: 1023.8+ KB
```

4.3 Installs field is currently stored as string and convert it to integer:

[30]:
```
data["Installs"] = [float(i.replace('+','').replace(',', '')) if '+' in i or ',' in i else float(0) for i in data["Installs"]]
```

[31]:
```
data.head()
```

[31]:

	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	159.0	19000.0	10000.0	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967.0	14000.0	500000.0	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510.0	8700.0	5000000.0	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644.0	25000.0	50000000.0	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967.0	2800.0	100000.0	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

[32]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9360 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             9360 non-null   object 
 1   Category        9360 non-null   object 
 2   Rating          9360 non-null   float64
 3   Reviews         9360 non-null   float64
 4   Size            9360 non-null   float64
 5   Installs        9360 non-null   float64
 6   Type            9360 non-null   object 
 7   Price           9360 non-null   object 
 8   Content Rating  9360 non-null   object 
 9   Genres          9360 non-null   object 
 10  Last Updated    9360 non-null   object 
 11  Current Ver     9360 non-null   object 
 12  Android Ver     9360 non-null   object 
dtypes: float64(4), object(9)
memory usage: 1023.8+ KB
```

[33]:
```
data["Installs"] = data["Installs"].astype(int)
```

[34]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9360 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             9360 non-null   object 
 1   Category        9360 non-null   object 
 2   Rating          9360 non-null   float64
 3   Reviews         9360 non-null   float64
 4   Size            9360 non-null   float64
 5   Installs        9360 non-null   int64  
 6   Type            9360 non-null   object 
 7   Price           9360 non-null   object 
 8   Content Rating  9360 non-null   object 
 9   Genres          9360 non-null   object 
 10  Last Updated    9360 non-null   object 
 11  Current Ver     9360 non-null   object 
 12  Android Ver     9360 non-null   object 
dtypes: float64(3), int64(1), object(9)
memory usage: 1023.8+ KB
```

4.4 Remove $ sign and Convert into numeric

[71]:
```
data["Price"] = [ float(i.split('$')[1]) if '$' in i else float(0) for i in data["Price"] ]
```

[72]:
```
data.head()
```

[72]:

	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	159.0	19000.0	10000	Free	0.0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967.0	14000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510.0	8700.0	5000000	Free	0.0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644.0	25000.0	50000000	Free	0.0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967.0	2800.0	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

[73]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9360 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             9360 non-null   object 
 1   Category        9360 non-null   object 
 2   Rating          9360 non-null   float64
 3   Reviews         9360 non-null   float64
 4   Size            9360 non-null   float64
 5   Installs        9360 non-null   int64  
 6   Type            9360 non-null   object 
 7   Price           9360 non-null   float64
 8   Content Rating  9360 non-null   object 
 9   Genres          9360 non-null   object 
 10  Last Updated    9360 non-null   object 
 11  Current Ver     9360 non-null   object 
 12  Android Ver     9360 non-null   object 
dtypes: float64(4), int64(1), object(8)
memory usage: 1023.8+ KB
```

[74]:
```
data["Price"] = data["Price"].astype(int)
```

[75]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9360 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   App             9360 non-null   object 
 1   Category        9360 non-null   object 
 2   Rating          9360 non-null   float64
 3   Reviews         9360 non-null   float64
 4   Size            9360 non-null   float64
 5   Installs        9360 non-null   int64  
 6   Type            9360 non-null   object 
 7   Price           9360 non-null   int64  
 8   Content Rating  9360 non-null   object 
 9   Genres          9360 non-null   object 
 10  Last Updated    9360 non-null   object 
 11  Current Ver     9360 non-null   object 
 12  Android Ver     9360 non-null   object 
dtypes: float64(3), int64(2), object(8)
memory usage: 1023.8+ KB
```

4- 5.1 Average rating should be between 1 & 5:

```
[76]: data.shape
```

```
[76]: (9360, 13)
```

```
[78]: data.drop(data[(data['Reviews'] < 1) & (data['Reviews'] > 5 )].index, inplace = True)
```

```
[79]: data.shape
```

```
[79]: (9360, 13)
```

4- 5.2 Reviews should not be more than installs as only those who installed can review the app

```
[79]: data.shape
```

```
[79]: (9360, 13)
```

```
[80]: data.drop(data[data['Installs'] < data['Reviews'] ].index, inplace = True)
```

```
[81]: data.shape
```

```
[81]: (9353, 13)
```

4- 5.3 For free apps (type = “Free”), the price should not be >0

```
[83]: data.drop(data[(data['Type'] =='Free') & (data['Price'] > 0 )].index, inplace = True)
```

```
[84]: data.shape
```

```
[84]: (9353, 13)
```

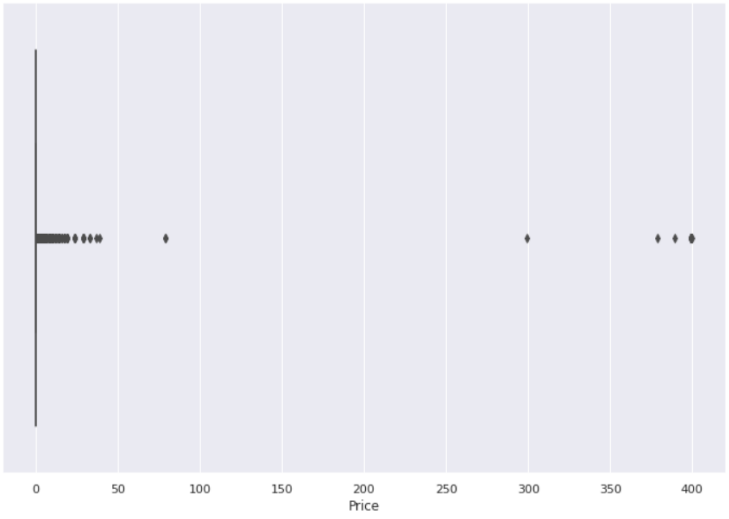
5.1 Boxplot for price

```
[85]: sns.set(rc={'figure.figsize':(12,8)})
```

```
[86]: sns.boxplot(data['Price'])
```

```
/usr/local/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

```
[86]: <AxesSubplot:xlabel='Price'>
```

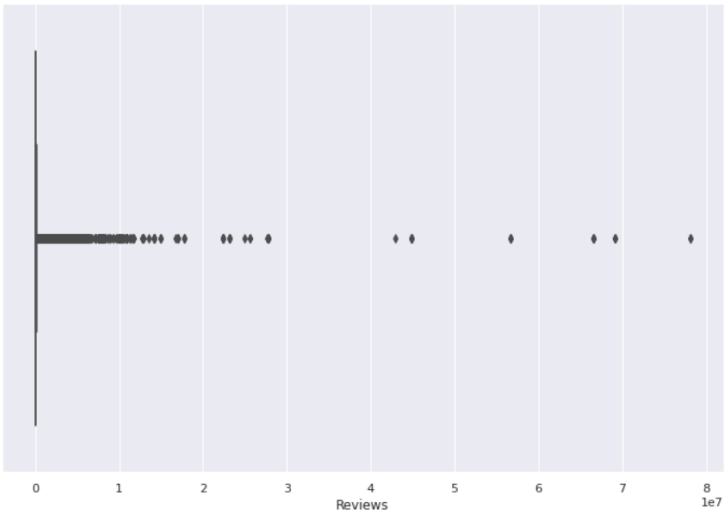


5.2 Boxplot for Reviews:

```
[87]: sns.boxplot(data['Reviews'])
```

```
/usr/local/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

```
[87]: <AxesSubplot:xlabel='Reviews'>
```

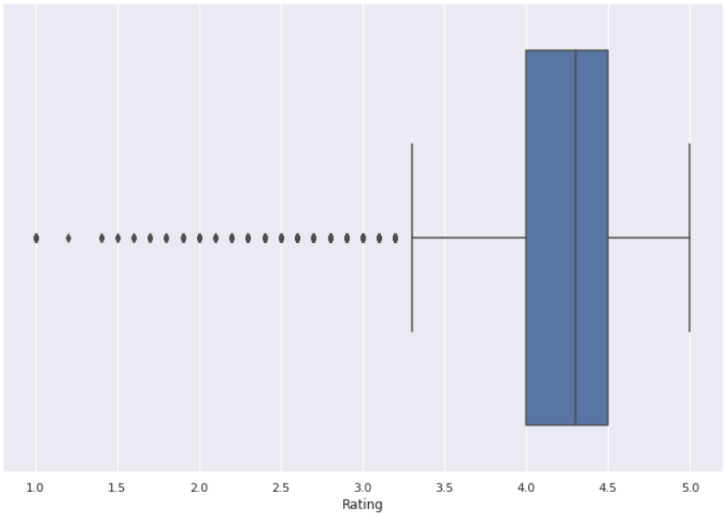


5.3 Histogram for Rating:

```
[88]: sns.boxplot(data['Rating'])
```

```
/usr/local/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

```
[88]: <AxesSubplot:xlabel='Rating'>
```

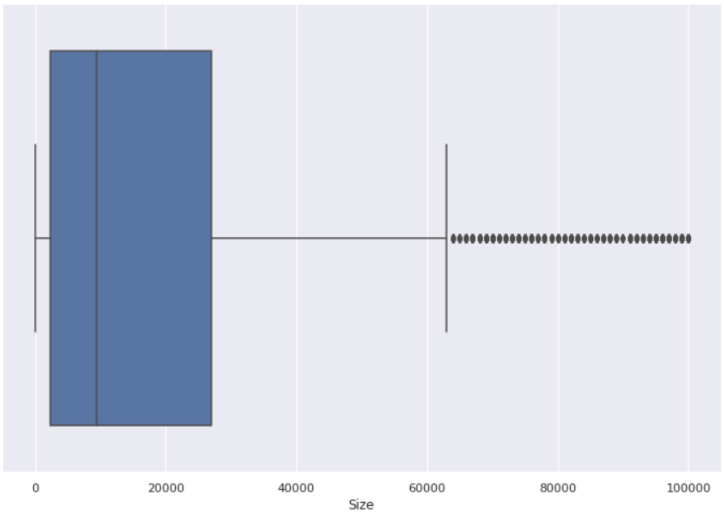


5.4 Histogram for size:

[89]:
```
sns.boxplot(data['Size'])
```

```
/usr/local/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

[89]: <AxesSubplot:xlabel='Size'>



6.1 Outlier treatment - Price:

[90]:
```
more = data.apply(lambda x : True
if x['Price'] > 200 else False, axis = 1)
```

[91]:
```
more_count = len(more[more == True].index)
```

[92]:
```
data.shape
```

[92]: (9353, 13)

[93]:
```
data.drop(data[data['Price'] > 200].index, inplace = True)
```

[94]:
```
data.shape
```

[94]: (9338, 13)

6.2 Reviews:

[95]:
```
data.drop(data[data['Reviews'] > 2000000].index, inplace = True)
```

[96]:
```
data.shape
```

[96]: (8885, 13)

6.3 Installs:

[98]:
```
data.quantile([.1, .25, .5, .70, .90, .95, .99], axis = 0)
```

[98]:

	Rating	Reviews	Size	Installs	Price
0.10	3.5	18.00	0.0	1000.0	0.0
0.25	4.0	159.00	2600.0	10000.0	0.0
0.50	4.3	4290.00	9500.0	500000.0	0.0
0.70	4.5	35930.40	23000.0	1000000.0	0.0
0.90	4.7	296771.00	50000.0	10000000.0	0.0
0.95	4.8	637298.00	68000.0	10000000.0	1.0
0.99	5.0	1462800.88	95000.0	100000000.0	7.0

[99]:
```
data.drop(data[data['Installs'] > 10000000].index, inplace = True)
```

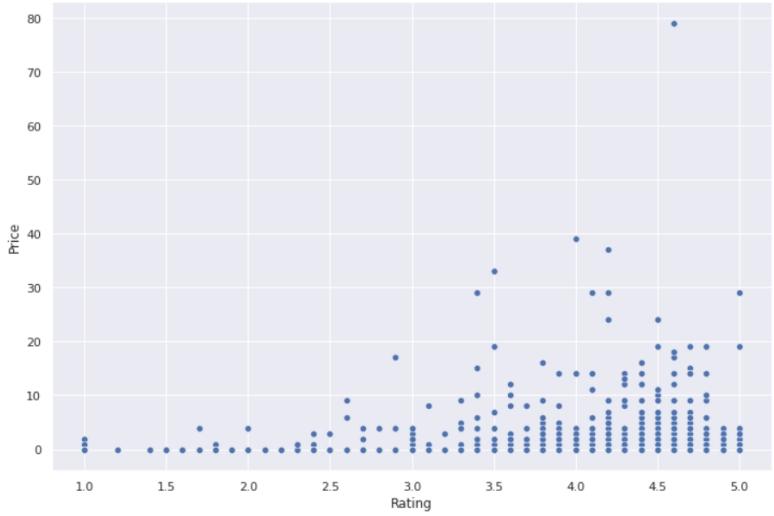
```
[100]: data.shape
```

```
[100]: (8496, 13)
```

7.1 Scatter plot/joinplot for Rating vs. Price

```
[101]: sns.scatterplot(x='Rating',y='Price',data=data)
```

```
[101]: <AxesSubplot:xlabel='Rating', ylabel='Price'>
```



Answer: Yes, Paid apps are higher ratings compare to free apps.

7.2 Scatter plot/joinplot for Rating vs. Size

```
[102]: sns.scatterplot(x='Rating',y='Size',data=data)
```

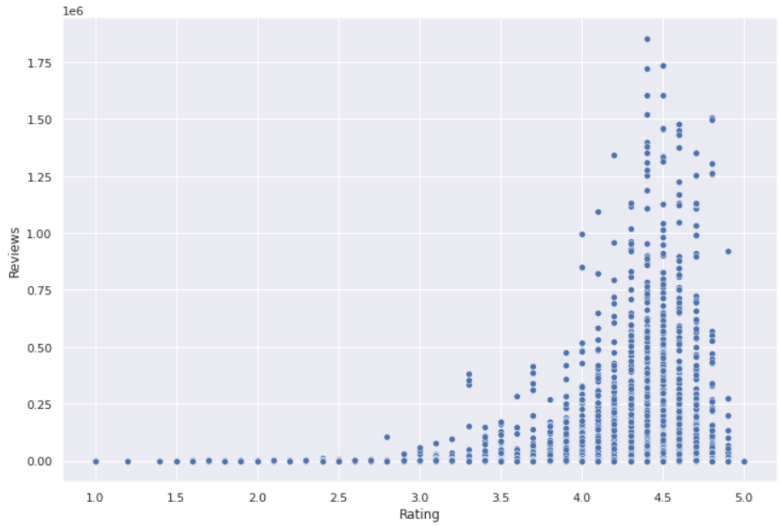
```
[102]: <AxesSubplot:xlabel='Rating', ylabel='Size'>
```

Answer: Yes, it is clear that heavier apps are rated better.

7.3 Scatter plot/joinplot for Rating vs. Reviews

```
[103]: sns.scatterplot(x='Rating',y='Reviews',data=data)
```

```
[103]: <AxesSubplot:xlabel='Rating', ylabel='Reviews'>
```

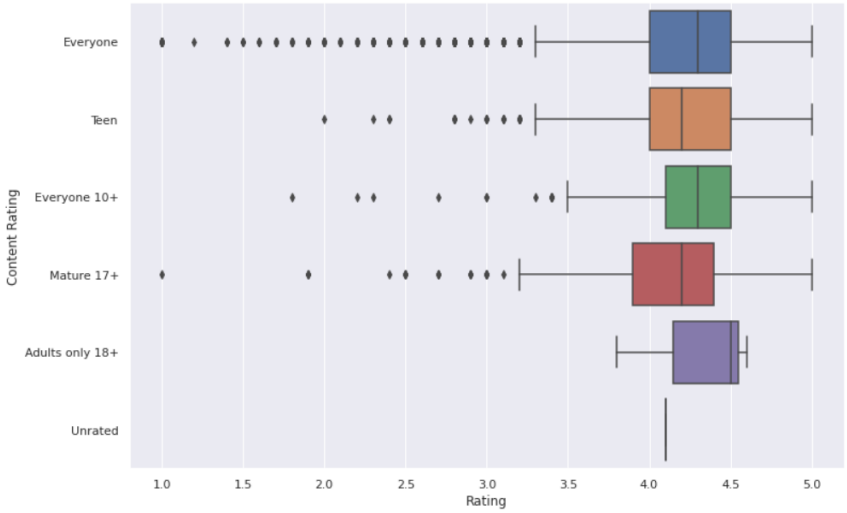


Answer: It is clear that more reviews makes app rating better.

7.4 Boxplot for Rating vs. Content Rating

```
[104]: sns.boxplot(x="Rating", y="Content Rating", data=data)
```

```
[104]: <AxesSubplot:xlabel='Rating', ylabel='Content Rating'>
```

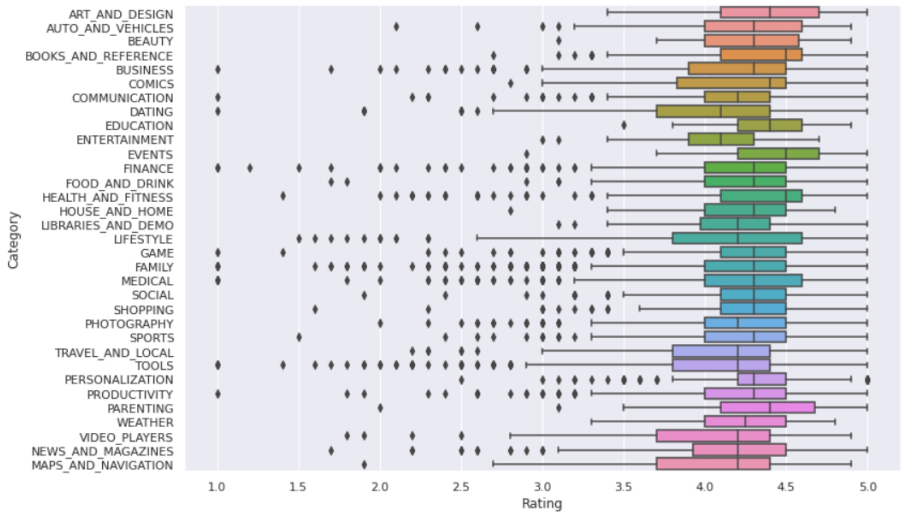


Answer: Apps which are for everyone have more bad ratings compared to other sections as it has many outliers' value, while 18+ apps have better ratings.

7.5 Boxplot for Ratings vs. Category

[105]:
```
sns.boxplot(x="Rating", y="Category", data=data)
```

[105]:
```
<AxesSubplot:xlabel='Rating', ylabel='Category'>
```



Answer: Events category has best ratings compare to others.

8.1 Data Preprocessing Reviews and Install

[106]:
```
inp1 = data
```

[108]:
```
inp1.head()
```

[108]:

	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	159.0	19000.0	10000	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967.0	14000.0	500000	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510.0	8700.0	5000000	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967.0	2800.0	100000	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
5	Paper flowers instructions	ART_AND_DESIGN	4.4	167.0	5600.0	50000	Free	0	Everyone	Art & Design	March 26, 2017	1.0	2.3 and up

[109]:
```
inp1.skew()
```

[109]:
```
Rating      -1.749753
Reviews      4.576494
Size         1.655917
Installs     1.543697
Price       18.074542
dtype: float64
```

[110]:
```
reviewskew = np.log1p(inp1['Reviews'])
inp1['Reviews'] = reviewskew
```

```
[110]: reviewskew = np.log1p(inp1['Reviews'])
       inp1['Reviews'] = reviewskew

[111]: reviewskew.skew()

[111]: -0.20039949659264134

[112]: installsskew = np.log1p(inp1['Installs'])
       inp1['Installs']

[112]: 0          10000
       1         500000
       2        5000000
       4         100000
       5          50000
                 ...
       10834        500
       10836       5000
       10837        100
       10839       1000
       10840    10000000
       Name: Installs, Length: 8496, dtype: int64

[113]: installsskew.skew()

[113]: -0.5097286542754812

[114]: inp1.head()
```

[114]:

	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	5.075174	19000.0	10000	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	6.875232	14000.0	500000	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	11.379520	8700.0	5000000	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	6.875232	2800.0	100000	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
5	Paper flowers instructions	ART_AND_DESIGN	4.4	5.123964	5600.0	50000	Free	0	Everyone	Art & Design	March 26, 2017	1.0	2.3 and up

8.2 Data Preprocessing Drop columns App, Last Updated, Current Ver, and Android Ver

```
[115]: inp1.drop(["Last Updated","Current Ver","Android Ver","App","Type"],axis=1,inplace=True)

[116]: inp1.head()
```

[116]:

	Category	Rating	Reviews	Size	Installs	Price	Content Rating	Genres
0	ART_AND_DESIGN	4.1	5.075174	19000.0	10000	0	Everyone	Art & Design
1	ART_AND_DESIGN	3.9	6.875232	14000.0	500000	0	Everyone	Art & Design;Pretend Play
2	ART_AND_DESIGN	4.7	11.379520	8700.0	5000000	0	Everyone	Art & Design
4	ART_AND_DESIGN	4.3	6.875232	2800.0	100000	0	Everyone	Art & Design;Creativity
5	ART_AND_DESIGN	4.4	5.123964	5600.0	50000	0	Everyone	Art & Design

```
[117]: inp1.shape

[117]: (8496, 8)
```

8.3 Data Preprocessing Get dummy columns for Category, Genres, and Content Rating.

```
[118]: inp2 = inp1

[119]: inp2.head()
```

[119]:

	Category	Rating	Reviews	Size	Installs	Price	Content Rating	Genres
0	ART_AND_DESIGN	4.1	5.075174	19000.0	10000	0	Everyone	Art & Design
1	ART_AND_DESIGN	3.9	6.875232	14000.0	500000	0	Everyone	Art & Design;Pretend Play
2	ART_AND_DESIGN	4.7	11.379520	8700.0	5000000	0	Everyone	Art & Design
4	ART_AND_DESIGN	4.3	6.875232	2800.0	100000	0	Everyone	Art & Design;Creativity
5	ART_AND_DESIGN	4.4	5.123964	5600.0	50000	0	Everyone	Art & Design

Appy dummy encoding on column ‘Category’

[120]:
```
inp2.Category.unique()
```

[120]:
```
array(['ART_AND_DESIGN', 'AUTO_AND_VEHICLES', 'BEAUTY',
       'BOOKS_AND_REFERENCE', 'BUSINESS', 'COMICS', 'COMMUNICATION',
       'DATING', 'EDUCATION', 'ENTERTAINMENT', 'EVENTS', 'FINANCE',
       'FOOD_AND_DRINK', 'HEALTH_AND_FITNESS', 'HOUSE_AND_HOME',
       'LIBRARIES_AND_DEMO', 'LIFESTYLE', 'GAME', 'FAMILY', 'MEDICAL',
       'SOCIAL', 'SHOPPING', 'PHOTOGRAPHY', 'SPORTS', 'TRAVEL_AND_LOCAL',
       'TOOLS', 'PERSONALIZATION', 'PRODUCTIVITY', 'PARENTING', 'WEATHER',
       'VIDEO_PLAYERS', 'NEWS_AND_MAGAZINES', 'MAPS_AND_NAVIGATION'],
      dtype=object)
```

[121]:
```
inp2.Category = pd.Categorical(inp2.Category)
x = inp2[['Category']]
del inp2['Category']
dummies = pd.get_dummies(x, prefix = 'Category')
inp2 = pd.concat([inp2,dummies], axis=1)
inp2.head()
```

[121]:

	Rating	Reviews	Size	Installs	Price	Content Rating	Genres	Category_ART_AND_DESIGN	Category_AUTO_AND_VEHICLES	Category_BEAUTY	...
0	4.1	5.075174	19000.0	10000	0	Everyone	Art & Design	1	0	0	...
1	3.9	6.875232	14000.0	500000	0	Everyone	Art & Design;Pretend Play	1	0	0	...
2	4.7	11.379520	8700.0	5000000	0	Everyone	Art & Design	1	0	0	...
4	4.3	6.875232	2800.0	100000	0	Everyone	Art & Design;Creativity	1	0	0	...
5	4.4	5.123964	5600.0	50000	0	Everyone	Art & Design	1	0	0	...

5 rows × 40 columns

[122]:
```
inp2.shape
```

[122]:
```
(8496, 40)
```

Apply dummy encoding on column ‘Genres’

[123]:
```
inp2["Genres"].unique()
```

[123]:
```
array(['Art & Design', 'Art & Design;Pretend Play',
       'Art & Design;Creativity', 'Auto & Vehicles', 'Beauty',
       'Books & Reference', 'Business', 'Comics', 'Comics;Creativity',
       'Communication', 'Dating', 'Education', 'Education;Creativity',
       'Education;Education', 'Education;Music & Video',
       'Education;Action & Adventure', 'Education;Pretend Play',
       'Education;Brain Games', 'Entertainment',
       'Entertainment;Brain Games', 'Entertainment;Creativity',
       'Entertainment;Music & Video', 'Events', 'Finance', 'Food & Drink',
       'Health & Fitness', 'House & Home', 'Libraries & Demo',
       'Lifestyle', 'Lifestyle;Pretend Play', 'Card', 'Casual', 'Puzzle',
       'Action', 'Arcade', 'Word', 'Racing', 'Casual;Creativity',
       'Sports', 'Board', 'Simulation', 'Role Playing', 'Adventure',
       'Strategy', 'Simulation;Education', 'Action;Action & Adventure',
       'Trivia', 'Casual;Brain Games', 'Simulation;Action & Adventure',
       'Educational;Creativity', 'Puzzle;Brain Games',
       'Educational;Education', 'Card;Brain Games',
       'Educational;Brain Games', 'Educational;Pretend Play',
       'Casual;Action & Adventure', 'Entertainment;Education',
       'Casual;Education', 'Casual;Pretend Play', 'Music;Music & Video',
       'Racing;Action & Adventure', 'Arcade;Pretend Play',
       'Adventure;Action & Adventure', 'Role Playing;Action & Adventure',
       'Simulation;Pretend Play', 'Puzzle;Creativity',
       'Sports;Action & Adventure', 'Educational;Action & Adventure',
       'Arcade;Action & Adventure', 'Entertainment;Action & Adventure',
       'Puzzle;Action & Adventure', 'Strategy;Action & Adventure',
       'Music & Audio;Music & Video', 'Health & Fitness;Education',
       'Adventure;Education', 'Board;Brain Games',
       'Board;Action & Adventure', 'Board;Pretend Play',
       'Casual;Music & Video', 'Role Playing;Pretend Play',
       'Entertainment;Pretend Play', 'Video Players & Editors;Creativity',
       'Card;Action & Adventure', 'Medical', 'Social', 'Shopping',
```

```
        'Photography', 'Travel & Local',
        'Travel & Local;Action & Adventure', 'Tools', 'Tools;Education',
        'Personalization', 'Productivity', 'Parenting',
        'Parenting;Music & Video', 'Parenting;Brain Games',
        'Parenting;Education', 'Weather', 'Video Players & Editors',
        'Video Players & Editors;Music & Video', 'News & Magazines',
        'Maps & Navigation', 'Health & Fitness;Action & Adventure',
        'Music', 'Educational', 'Casino', 'Adventure;Brain Games',
        'Lifestyle;Education', 'Books & Reference;Education',
        'Puzzle;Education', 'Role Playing;Brain Games',
        'Strategy;Education', 'Racing;Pretend Play',
        'Communication;Creativity', 'Strategy;Creativity'], dtype=object)
```

[125]:
```
lists = []
for i in inp2.Genres.value_counts().index:
    if inp2.Genres.value_counts()[i]<20:
        lists.append(i)
inp2.Genres = ['Other' if i in lists else i for i in inp2.Genres]
```

[126]:
```
inp2["Genres"].unique()
```

[126]:
```
array(['Art & Design', 'Other', 'Auto & Vehicles', 'Beauty',
       'Books & Reference', 'Business', 'Comics', 'Communication',
       'Dating', 'Education', 'Education;Education',
       'Education;Pretend Play', 'Entertainment',
       'Entertainment;Music & Video', 'Events', 'Finance', 'Food & Drink',
       'Health & Fitness', 'House & Home', 'Libraries & Demo',
       'Lifestyle', 'Card', 'Casual', 'Puzzle', 'Action', 'Arcade',
       'Word', 'Racing', 'Sports', 'Board', 'Simulation', 'Role Playing',
       'Adventure', 'Strategy', 'Trivia', 'Educational;Education',
       'Casual;Pretend Play', 'Medical', 'Social', 'Shopping',
       'Photography', 'Travel & Local', 'Tools', 'Personalization',
       'Productivity', 'Parenting', 'Weather', 'Video Players & Editors',
       'News & Magazines', 'Maps & Navigation', 'Educational', 'Casino'],
      dtype=object)
```

Since, there are too many categories under Genres. Hence, we will try to reduce some categories which have very few samples under them and put them under one new common category i.e. "Other".

[127]:
```
inp2.Genres = pd.Categorical(inp2['Genres'])
x = inp2[["Genres"]]
del inp2['Genres']
dummies = pd.get_dummies(x, prefix = 'Genres')
inp2 = pd.concat([inp2,dummies], axis=1)
```

[128]:
```
inp2.head()
```

[128]:

	Rating	Reviews	Size	Installs	Price	Content Rating	Category_ART_AND_DESIGN	Category_AUTO_AND_VEHICLES	Category_BEAUTY	Category_BOOKS_AN
0	4.1	5.075174	19000.0	10000	0	Everyone	1	0	0	
1	3.9	6.875232	14000.0	500000	0	Everyone	1	0	0	
2	4.7	11.379520	8700.0	5000000	0	Everyone	1	0	0	
4	4.3	6.875232	2800.0	100000	0	Everyone	1	0	0	
5	4.4	5.123964	5600.0	50000	0	Everyone	1	0	0	

5 rows × 91 columns

Apply dummy encoding on column ‘Content Rating’

[129]:
```
inp2.shape
```

[129]:
```
(8496, 91)
```

[130]:
```
inp2["Content Rating"].unique()
```

[130]:
```
array(['Everyone', 'Teen', 'Everyone 10+', 'Mature 17+',
       'Adults only 18+', 'Unrated'], dtype=object)
```

[131]:
```
inp2['Content Rating'] = pd.Categorical(inp2['Content Rating'])
x = inp2[['Content Rating']]
del inp2['Content Rating']
dummies = pd.get_dummies(x, prefix = 'Content Rating')
inp2 = pd.concat([inp2,dummies], axis=1)
inp2.head()
```

[131]:

	Rating	Reviews	Size	Installs	Price	Category_ART_AND_DESIGN	Category_AUTO_AND_VEHICLES	Category_BEAUTY	Category_BOOKS_AND_REFEREI
0	4.1	5.075174	19000.0	10000	0	1	0	0	
1	3.9	6.875232	14000.0	500000	0	1	0	0	
2	4.7	11.379520	8700.0	5000000	0	1	0	0	
4	4.3	6.875232	2800.0	100000	0	1	0	0	
5	4.4	5.123964	5600.0	50000	0	1	0	0	

5 rows × 96 columns

[132]:
```
inp2.shape
```

[132]:
```
(8496, 96)
```

9 & 10. Train test split

[134]:
```
from sklearn.model_selection import train_test_split as tts
from sklearn.linear_model import LinearRegression as LR
from sklearn.metrics import mean_squared_error as mse
```

[135]:
```
d1 = inp2
X = d1.drop('Rating',axis=1)
y = d1['Rating']
Xtrain, Xtest, ytrain, ytest = tts(X,y, test_size=0.3, random_state=5)
```

11. Model Building

[136]:
```
reg_all = LR()
reg_all.fit(Xtrain,ytrain)
```

[136]:
```
LinearRegression()
```

[137]:
```
R2_train = round(reg_all.score(Xtrain,ytrain),3)
print("The R2 value of the Training Set is : {}".format(R2_train))
```

```
The R2 value of the Training Set is : 0.074
```

12. Predictions on test set and report R2.

[138]:
```
R2_test = round(reg_all.score(Xtest,ytest),3)
print("The R2 value of the Testing Set is : {}".format(R2_test))
```

```
The R2 value of the Testing Set is : 0.063
```