 # Make a model to predict the app rating, with other information about the app provided.

1. **Load the data file using pandas.**

In [10]:

```
import pandas as pd
import numpy as np
import seaborn as sns
```

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

In [11]:

```
data = pd.read_csv('googleplaystore.csv')
```

In [12]:

```
data.head()
```

Out[12]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Cu
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	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	
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	
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	V de
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	

In [13]:

```
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
```

In [14]:

```
data.shape
```

Out[14]:

```
(10841, 13)
```

1. **Drop records with nulls in any of the columns.**

In [15]:

```
data.isnull().any()
```

Out[15]:

```
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
```

In [16]:

```
data.isnull().sum()
```

Out[16]:

```
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 [17]:

```
data = data.dropna()
```

In [18]:

```
data.isnull().any()
```

Out[18]:

```
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
```

In [19]:

```
data.shape
```

Out[19]:

```
(9360, 13)
```

1. **a(1) Extract the numeric value from the column**

In [20]:

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

In [21]:

```
data.head()
```

Out[21]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Cur
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	Coloring book moana	ART_AND_DESIGN	3.9	967	14.0	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	
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	
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25.0	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Va de
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. **a(2) Multiply the value by 1,000, if size is mentioned in Mb**

In [22]:

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

In [23]:

```
data
```

Out[23]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genre
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10,000+	Free	0	Everyone	Art & Desig
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500,000+	Free	0	Everyone	Art Design;Preten Pla
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5,000,000+	Free	0	Everyone	Art & Desig
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50,000,000+	Free	0	Teen	Art & Desig
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100,000+	Free	0	Everyone	Art Design;Creativit
...	...	...	...	...	...	...	...	...	...	.
10834	FR Calculator	FAMILY	4.0	7	2600.0	500+	Free	0	Everyone	Educatio
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5,000+	Free	0	Everyone	Educatio
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100+	Free	0	Everyone	Educatio
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	0.0	1,000+	Free	0	Mature 17+	Books Referenc
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10,000,000+	Free	0	Everyone	Lifestyl

9360 rows × 13 columns

In [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   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
```

1. **(b) Reviews is a numeric field that is loaded as a string field. Convert it to numeric (int/float)**

In [25]:

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

In [26]:

```
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
```

1. **3 (a) Treat 1,000,000+ as 1,000,000**

In [27]:

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

In [28]:

```
data.head()
```

Out[28]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated
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	**Coloring book moana**	**ART_AND_DESIGN**	**3.9**	**967.0**	**14000.0**	**500000.0**	**Free**	**0**	**Everyone**	**Art & Design;Pretend Play**	**January 15, 2018**

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated
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
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644.0	25000.0	50000000.0	Free	0	Teen	Art & Design	June 8, 2018
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

In [29]:

```
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
```

1. **3 (b) remove '+', ',' from the field, convert it to integer**

In [30]:

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

In [31]:

```
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   int32  
 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), int32(1), object(9)
memory usage: 987.2+ KB
```

1. **(d) Price field is a string and has $symbol$ ' sign, and convert it to numeric. $.Remove$ '**

In [32]:

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

In [33]:

```
data.head()
```

Out[33]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Cu
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	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	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	
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644.0	25000.0	50000000	Free	0.0	Teen	Art & Design	June 8, 2018	V d
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	

In [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   int32  
 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), int32(1), object(8)
memory usage: 987.2+ KB
```

In [35]:

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

In [36]:

```
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   int32  
 6   Type            9360 non-null   object 
 7   Price           9360 non-null   int32  
 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), int32(2), object(8)
memory usage: 950.6+ KB
```

In [37]:

```
data.shape
```

Out[37]:

```
(9360, 13)
```

1. **Sanity checks:**

(a) Average rating should be between 1 and 5 as only these values are allowed on the play store. Drop the rows that have a value outside this range

In [38]:

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

In [39]:

```
data.shape
```

Out[39]:

```
(9360, 13)
```

1. **(b) Reviews should not be more than installs as only those who installed can review the app. If there are any such records, drop them.**

In [40]:

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

In [41]:

```
data.shape
```

Out[41]:

```
(9353, 13)
```

1. **(c) For free apps (type = “Free”), the price should not be >0. Drop any such rows.**

In [42]:

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

1. **Performing univariate analysis: 5 (a) Boxplot for Price**

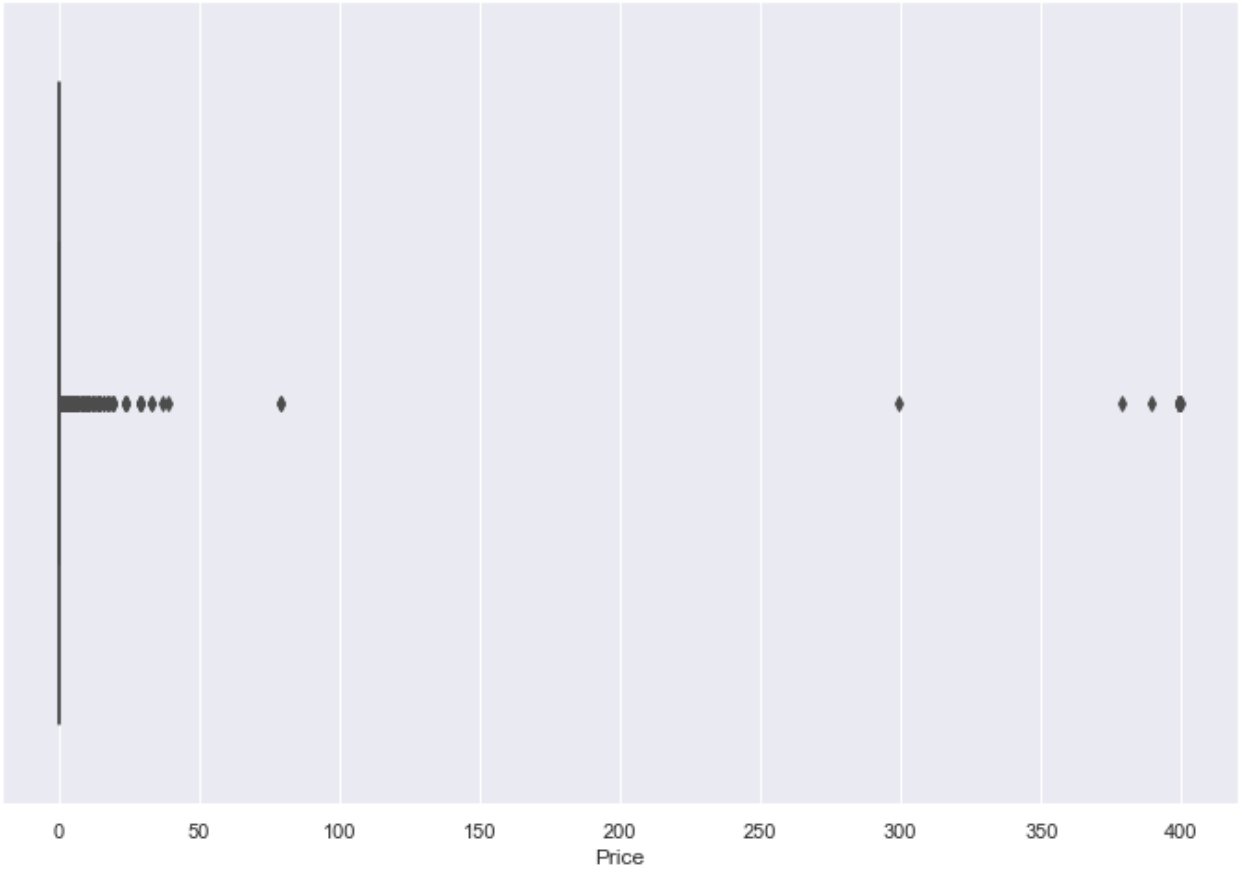
In [43]:

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

```
C:\Users\KANISHK\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pa
ss the following variable as a keyword arg: x. From version 0.12, the only valid position
al argument will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
```

Out[43]:

```
<AxesSubplot:xlabel='Price'>
```



1. **(b) Boxplot for Reviews**

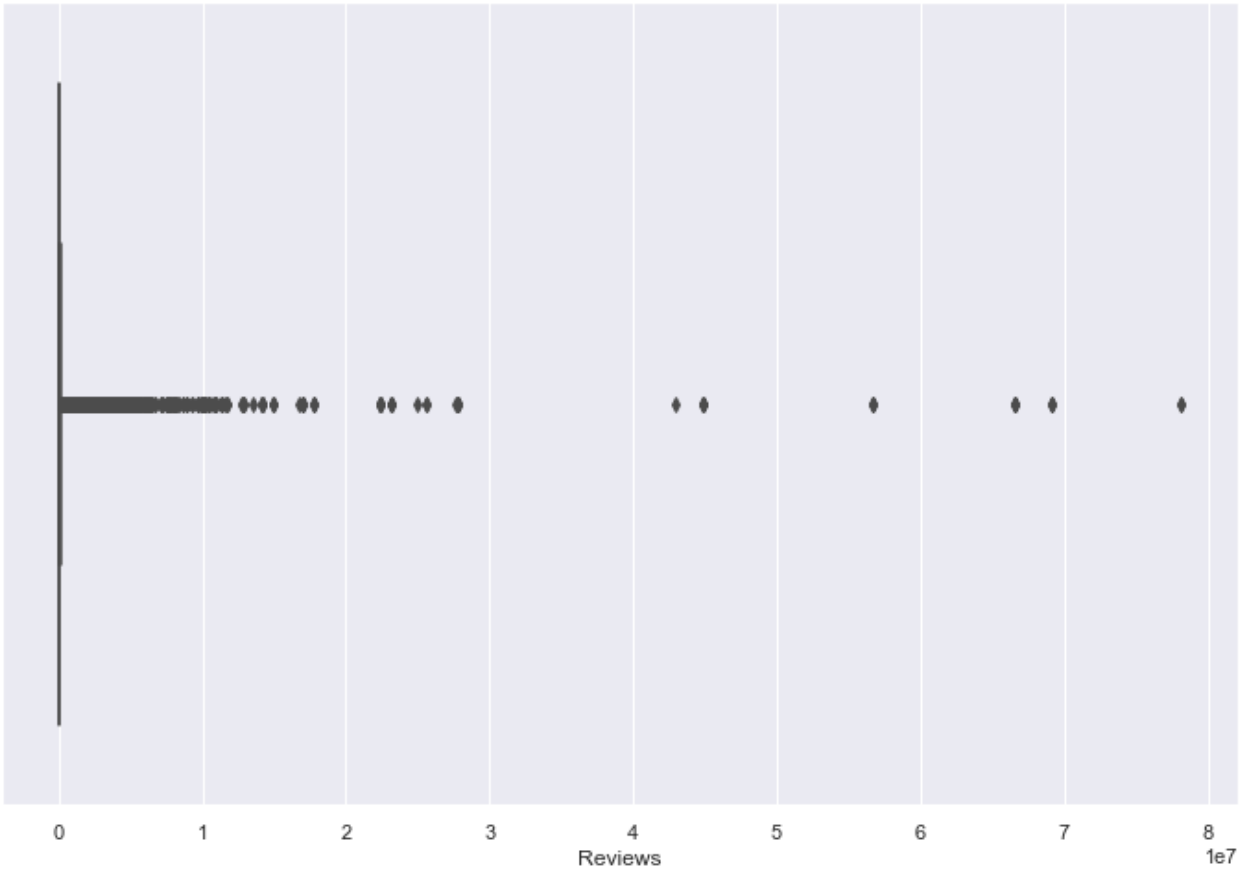
In [44]:

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

```
C:\Users\KANISHK\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pa
ss the following variable as a keyword arg: x. From version 0.12, the only valid position
al argument will be `data`, and passing other arguments without an explicit keyword will
result in an error or misinterpretation.
  warnings.warn(
```

Out[44]:

```
<AxesSubplot:xlabel='Reviews'>
```



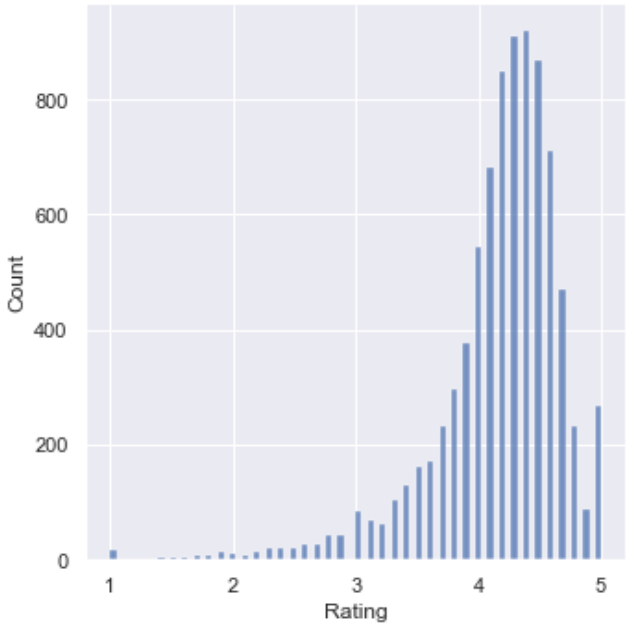
1. (c) Histogram for Rating

In [94]:

```
sns.displot(data['Rating'])
```

Out[94]:

```
<seaborn.axisgrid.FacetGrid at 0xeab35f9400>
```



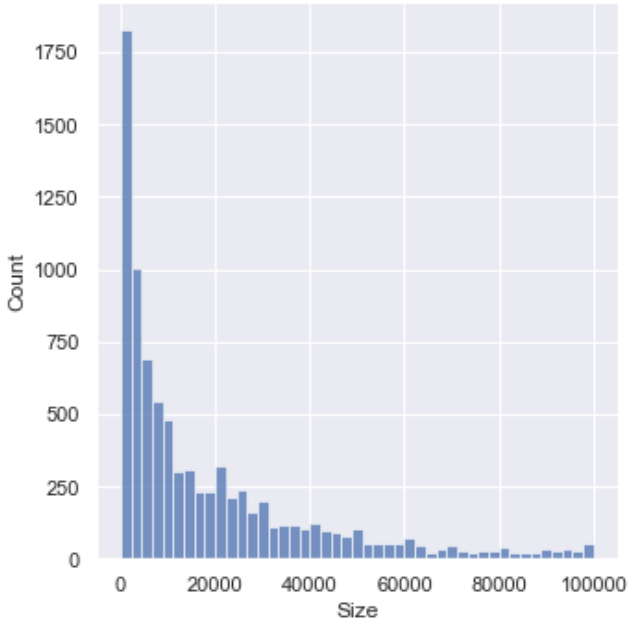
1. (d) Histogram for Size

In [95]:

```
sns.displot(data['Size'])
```

Out[95]:

```
<seaborn.axisgrid.FacetGrid at 0xeab34efc70>
```



1. **Outlier treatment:**

(a). Check out the records with very high price Is 200 indeed a high price?

In [47]:

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

In [48]:

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

In [49]:

```
data.shape
```

Out[49]:

```
(9353, 13)
```

1. **a(2) Drop these as most seem to be junk apps**

In [50]:

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

In [51]:

```
data.shape
```

Out[51]:

```
(9338, 13)
```

1. **(b) Reviews: Very few apps have very high number of reviews. These are all star apps that don't help with the analysis and, in fact, will skew it. Drop records having more than 2 million reviews.**

In [52]:

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

In [53]:

```
data.shape
```

Out[53]:

```
(8885, 13)
```

1. **c(1) Find out the different percentiles – 10, 25, 50, 70, 90, 95, 99**

In [54]:

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

Out[54]:

	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

1. **c(2) Decide a threshold as cutoff for outlier and drop records having values more than that**

In [55]:

```
# dropping more than 10000000 Installs value
data.drop(data[data['Installs'] > 10000000].index, inplace = True)
```

In [56]:

```
data.shape
```

Out[56]:

```
(8496, 13)
```

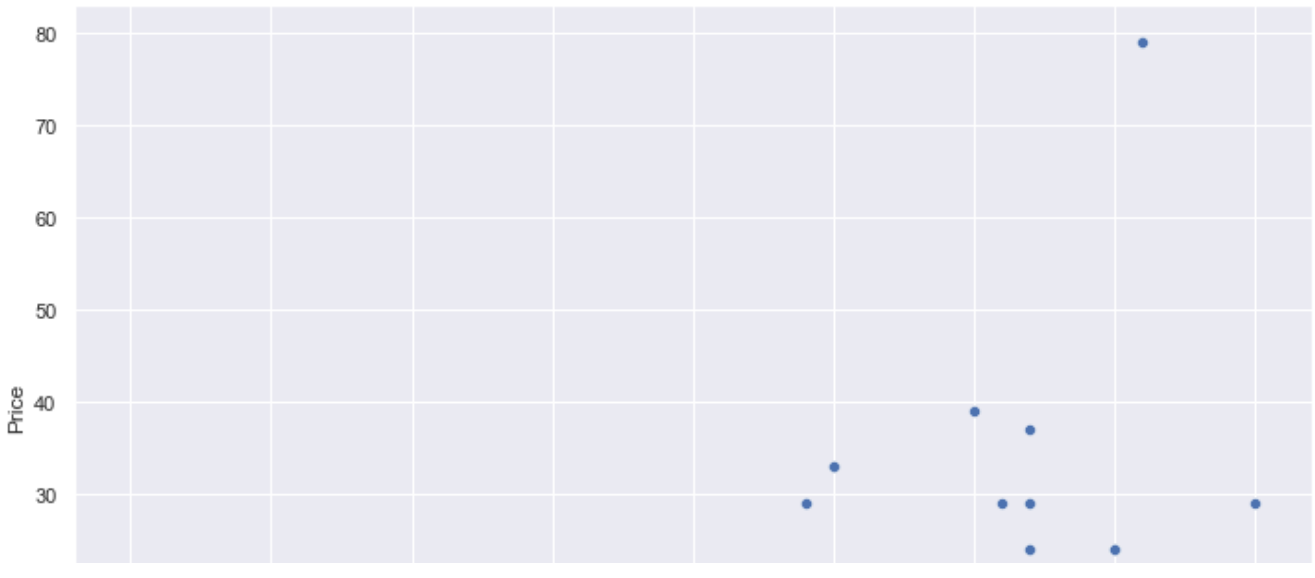
1. **(a) Make scatter plot/joinplot for Rating vs. Price**

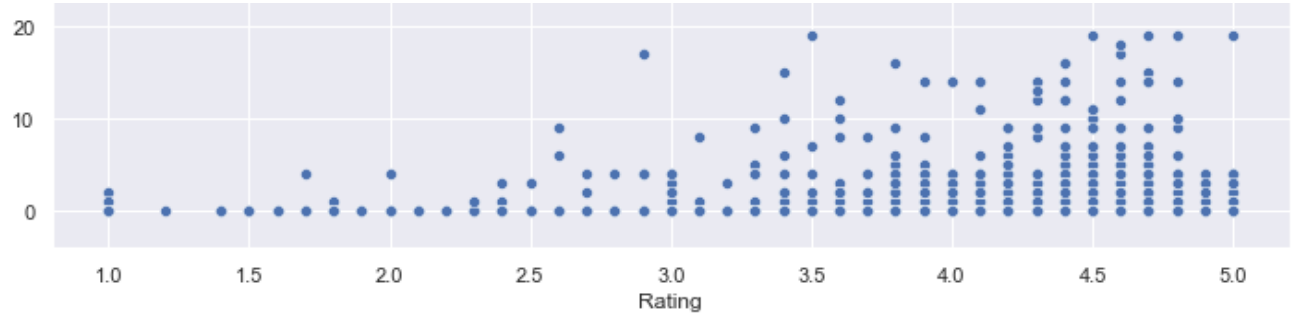
In [57]:

```
sns.scatterplot(x='Rating',y='Price',data=data)
```

Out[57]:

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





Yes, Paid apps are higher ratings comapre to free apps.

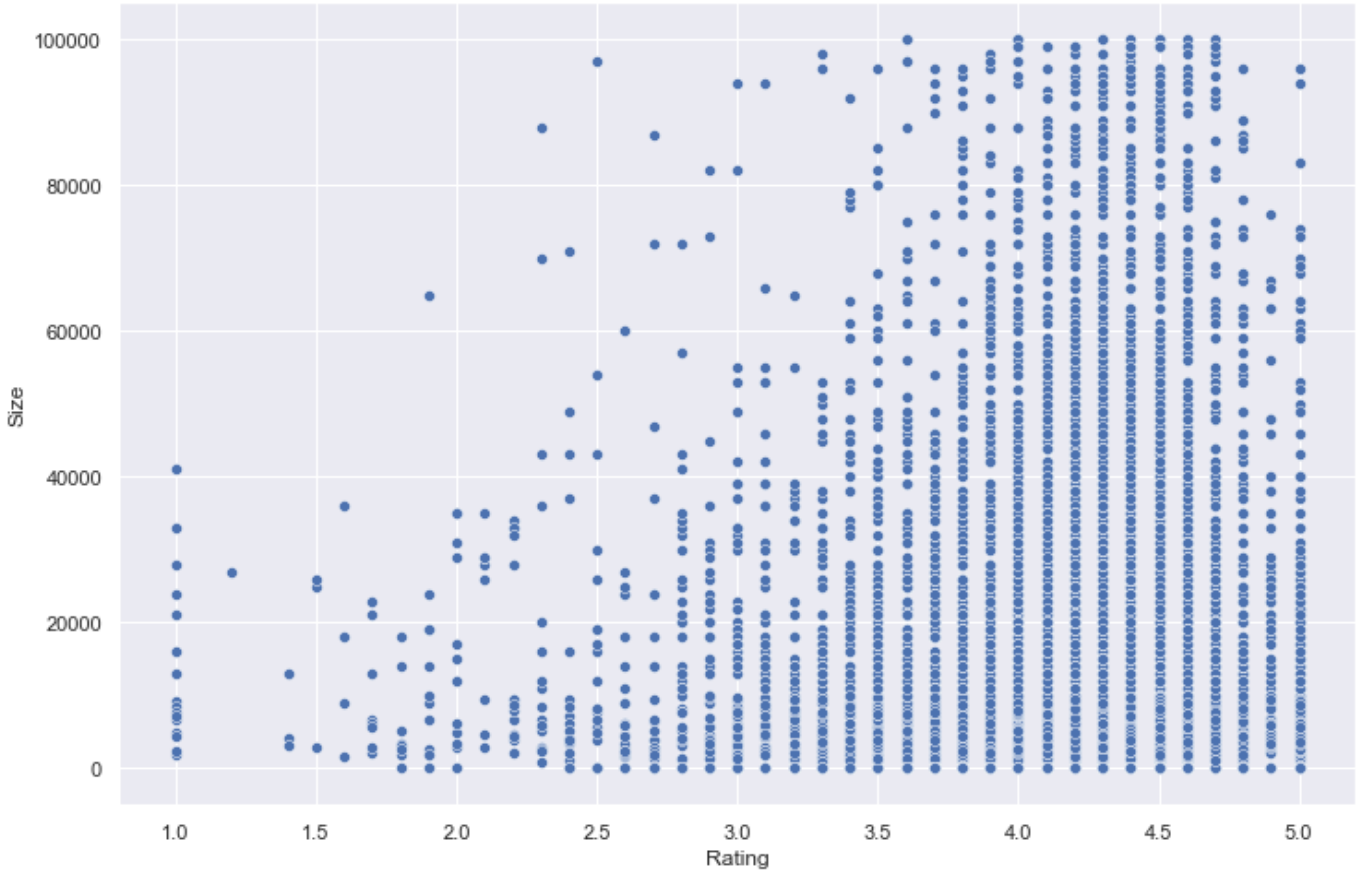
1. **(b) Make scatter plot/joinplot for Rating vs. Size**

In [58]:

```
sns.scatterplot(x='Rating',y='Size',data=data)
```

Out[58]:

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



Yes it is clear that heavior apps are rated better.

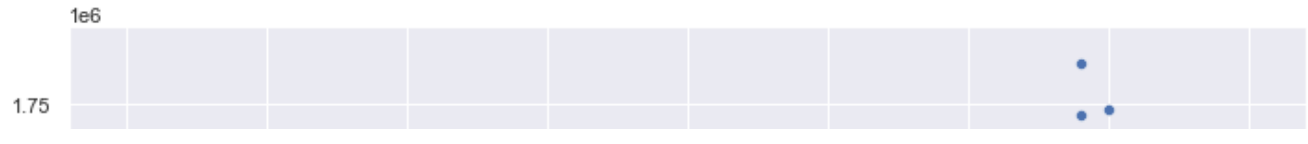
1. **(c) Make scatter plot/joinplot for Rating vs. Reviews**

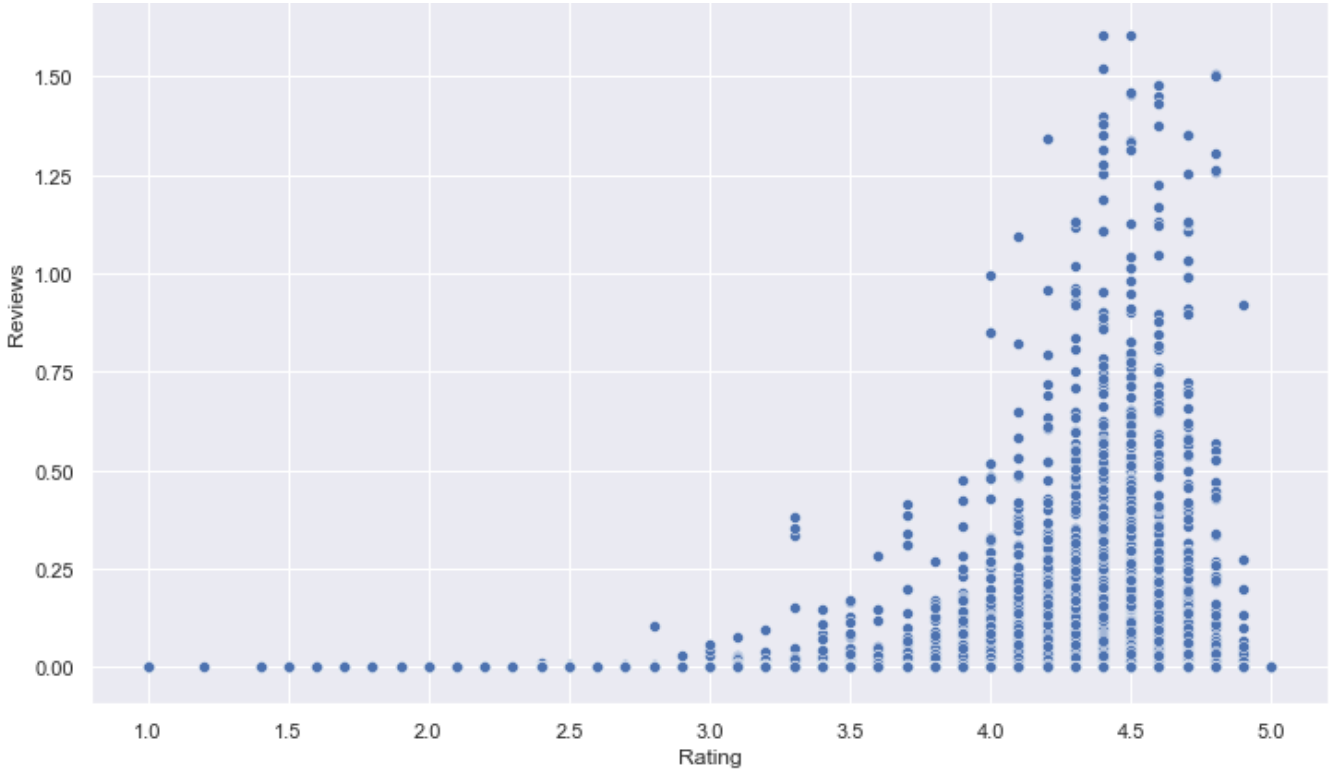
In [59]:

```
sns.scatterplot(x='Rating',y='Reviews',data=data)
```

Out[59]:

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





It is cristal clear that more reviews makes app rating better.

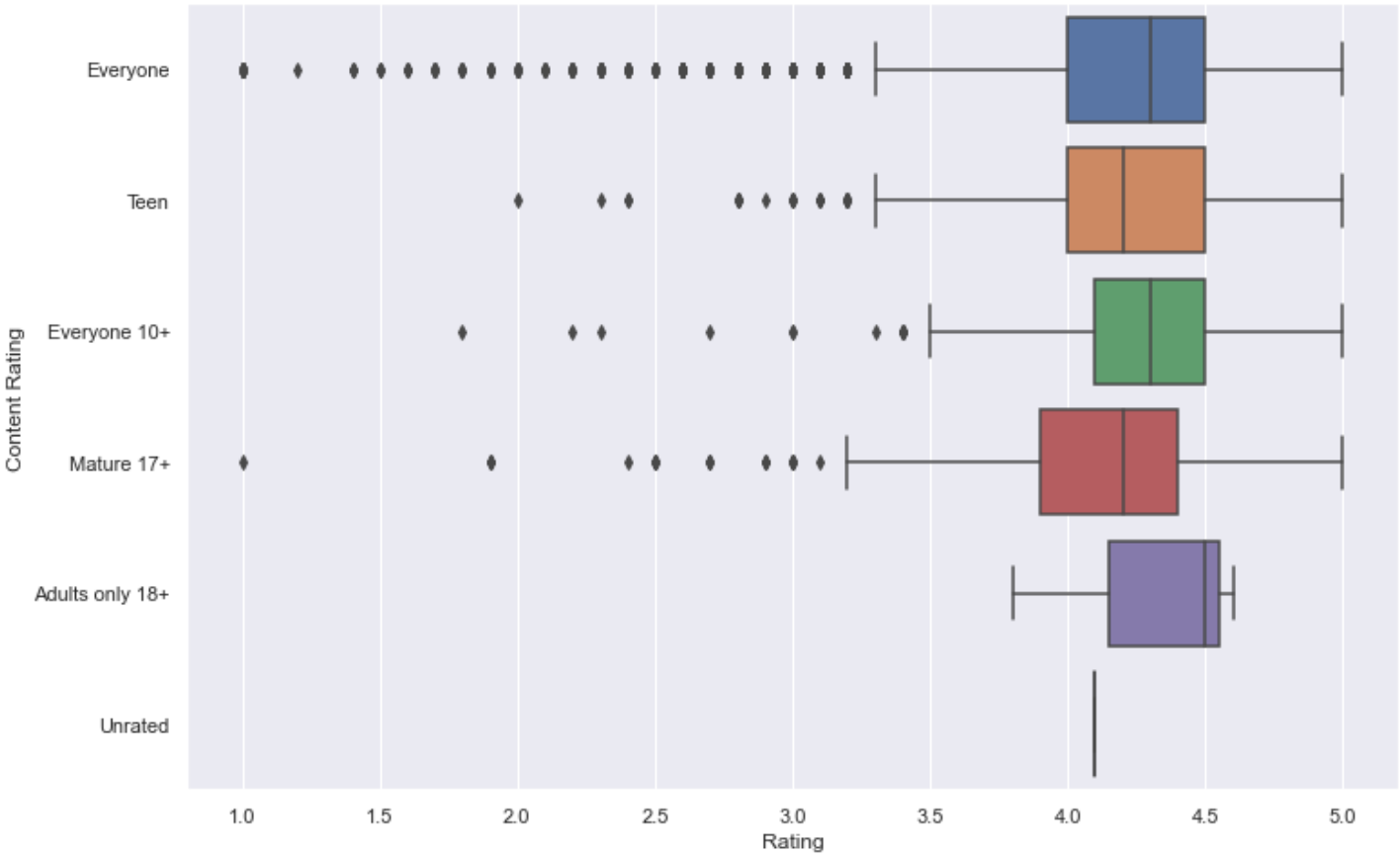
1. **(d) Make boxplot for Rating vs. Content Rating**

In [60]:

```
sns.boxplot(x="Rating", y="Content Rating", data=data)
```

Out[60]:

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



Apps which has more bad ratings compare to other sections as it has so much outliers value, while 18+ apps have better ratings.

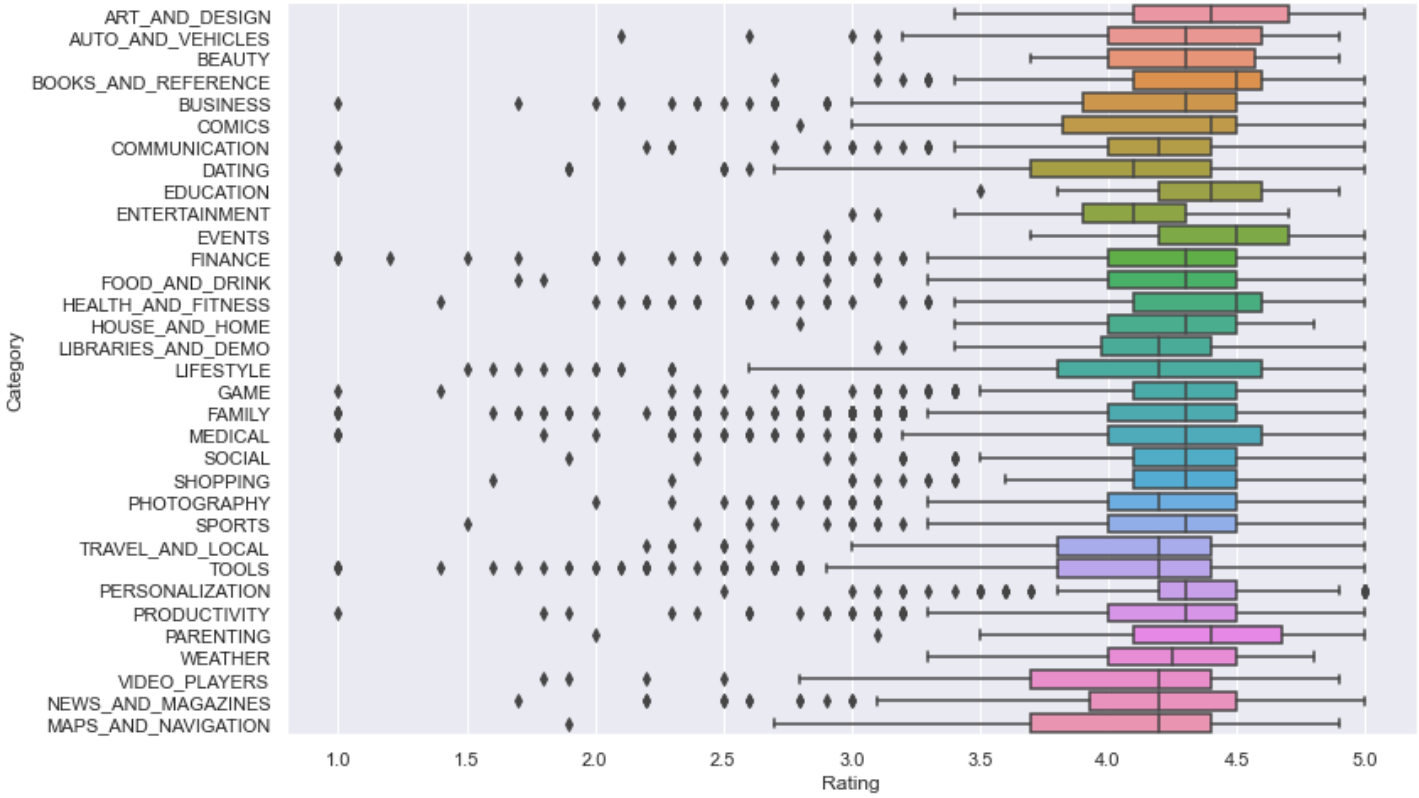
1. **(e) Make boxplot for Ratings vs. Category**

In [61]:

```
sns.boxplot(x="Rating", y="Category", data=data)
```

Out[61]:

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



Events category has better ratings compare to others category.

1. **Data preprocessing**
2. **(a) Reviews and Install have some values that are still relatively very high. Before building a linear regression model, you need to reduce the skew. Apply log transformation (np.log1p) to Reviews and Installs.**

In [62]:

```
inp1 = data
```

In [63]:

```
inp1.head()
```

Out[63]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Cu
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	**Coloring book moana**	**ART_AND_DESIGN**	**3.9**	**967.0**	**14000.0**	**500000**	**Free**	**0**	**Everyone**	**Art & Design;Pretend Play**	**January 15, 2018**	
	U Launcher Lite – FREE										**August**	

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Cu
2	Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510.0	8700.0	5000000	Free	0	Everyone	Art & Design	August 1, 2018	
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	
5	Paper flowers instructions	ART_AND_DESIGN	4.4	167.0	5600.0	50000	Free	0	Everyone	Art & Design	March 26, 2017	

In [64]:

```
inp1.skew()
```

```
C:\Users\KANISHK\AppData\Local\Temp/ipykernel_2568/3545313420.py:1: FutureWarning: Droppi
ng of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated;
in a future version this will raise TypeError.  Select only valid columns before calling
the reduction.
  inp1.skew()
```

Out[64]:

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

In [65]:

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

In [66]:

```
reviewskew.skew()
```

Out[66]:

```
-0.20039949659264134
```

In [67]:

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

Out[67]:

```
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: int32
```

In [68]:

```
installsskew.skew()
```

Out[68]:

```
-0.5097286542754812
```

In [69]:

```
inp1.head()
```

Out[69]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	C
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	Coloring book moana	ART_AND_DESIGN	3.9	6.875232	14000.0	500000	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	
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	
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	
5	Paper flowers instructions	ART_AND_DESIGN	4.4	5.123964	5600.0	50000	Free	0	Everyone	Art & Design	March 26, 2017	

1. **(b) Drop columns App, Last Updated, Current Ver, and Android Ver. These variables are not useful for our task.**

In [70]:

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

In [71]:

```
inp1.head()
```

Out[71]:

	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

In [72]:

```
inp1.shape
```

Out[72]:

```
(8496, 8)
```

1. **(c) Get dummy columns for Category, Genres, and Content Rating. This needs to be done as the models do not understand categorical data, and all data should be numeric. Dummy encoding is one way to convert character fields to numeric. Name of dataframe should be inp2**

character fields to numeric. Name of dataframe should be inp2.

In [73]:

```
inp2 = inp1
```

In [74]:

```
inp2.head()
```

Out[74]:

	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

In [75]:

```
#get unique values in Column "Category"
inp2.Category.unique()
```

Out[75]:

```
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)
```

In [76]:

```
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()
```

Out[76]:

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

5 rows × 40 columns

In [77]:

```
inp2.shape
```

Out[77]:

```
(8496, 40)
```

In [78]:

```
#get unique values in Column "Genres"
inp2["Genres"].unique()
```

Out[78]:

```
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)
```

In [79]:

```
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]
```

In [80]:

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

Out[80]:

```
array(['Art & Design', 'Other', 'Auto & Vehicles', 'Beauty',
```

```
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)
```

In [81]:

```
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)
```

In [82]:

```
inp2.head()
```

Out[82]:

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

5 rows × 91 columns

In [83]:

```
inp2.shape
```

Out[83]:

```
(8496, 91)
```

In [84]:

```
#get unique values in Column "Content Rating"
inp2["Content Rating"].unique()
```

Out[84]:

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

In [85]:

```
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()
```

Out[85]:

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

5 rows × 96 columns

In [86]:

```
inp2.shape
```

Out[86]:

```
(8496, 96)
```

1. **Train test split and apply 70-30 split. Name the new dataframes df_train and df_test.**

In [87]:

```
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
```

1. **Separate the dataframes into X_train, y_train, X_test, and y_test.**

In [88]:

```
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)
```

1. **Model building**

(a) Use linear regression as the technique

In [89]:

```
reg_all = LR()
reg_all.fit(Xtrain,ytrain)
```

Out[89]:

```
LinearRegression()
```

(b) Report the R2 on the train set

In [90]:

```
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
```

1. **Make predictions on test set and report R2**

1. Make predictions on test set and report R2

In [91]:

```
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
```

In [1]:

```
pip install nbconvert
```

```
Requirement already satisfied: nbconvert in c:\users\kanishk\anaconda3\lib\site-packages
(6.1.0)
Requirement already satisfied: bleach in c:\users\kanishk\anaconda3\lib\site-packages (fr
om nbconvert) (4.0.0)
Requirement already satisfied: jupyter-core in c:\users\kanishk\anaconda3\lib\site-packag
es (from nbconvert) (4.8.1)
Requirement already satisfied: pandocfilters>=1.4.1 in c:\users\kanishk\anaconda3\lib\sit
e-packages (from nbconvert) (1.4.3)
Requirement already satisfied: traitlets>=5.0 in c:\users\kanishk\anaconda3\lib\site-pack
ages (from nbconvert) (5.1.0)
Requirement already satisfied: nbformat>=4.4 in c:\users\kanishk\anaconda3\lib\site-packa
ges (from nbconvert) (5.1.3)
Requirement already satisfied: jinja2>=2.4 in c:\users\kanishk\anaconda3\lib\site-package
s (from nbconvert) (2.11.3)
Requirement already satisfied: defusedxml in c:\users\kanishk\anaconda3\lib\site-packages
(from nbconvert) (0.7.1)
Requirement already satisfied: pygments>=2.4.1 in c:\users\kanishk\anaconda3\lib\site-pac
kages (from nbconvert) (2.10.0)
Requirement already satisfied: entrypoints>=0.2.2 in c:\users\kanishk\anaconda3\lib\site-
packages (from nbconvert) (0.3)
Requirement already satisfied: mistune<2,>=0.8.1 in c:\users\kanishk\anaconda3\lib\site-p
ackages (from nbconvert) (0.8.4)
Requirement already satisfied: jupyterlab-pygments in c:\users\kanishk\anaconda3\lib\site
-packages (from nbconvert) (0.1.2)
Requirement already satisfied: testpath in c:\users\kanishk\anaconda3\lib\site-packages (
from nbconvert) (0.5.0)
Requirement already satisfied: nbclient<0.6.0,>=0.5.0 in c:\users\kanishk\anaconda3\lib\s
ite-packages (from nbconvert) (0.5.3)
Requirement already satisfied: MarkupSafe>=0.23 in c:\users\kanishk\anaconda3\lib\site-pa
ckages (from jinja2>=2.4->nbconvert) (1.1.1)
Requirement already satisfied: nest-asyncio in c:\users\kanishk\anaconda3\lib\site-packag
es (from nbclient<0.6.0,>=0.5.0->nbconvert) (1.5.1)
Requirement already satisfied: async-generator in c:\users\kanishk\anaconda3\lib\site-pac
kages (from nbclient<0.6.0,>=0.5.0->nbconvert) (1.10)
Requirement already satisfied: jupyter-client>=6.1.5 in c:\users\kanishk\anaconda3\lib\si
te-packages (from nbclient<0.6.0,>=0.5.0->nbconvert) (6.1.12)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\kanishk\anaconda3\lib\sit
e-packages (from jupyter-client>=6.1.5->nbclient<0.6.0,>=0.5.0->nbconvert) (2.8.2)
Requirement already satisfied: tornado>=4.1 in c:\users\kanishk\anaconda3\lib\site-packag
es (from jupyter-client>=6.1.5->nbclient<0.6.0,>=0.5.0->nbconvert) (6.1)
Requirement already satisfied: pyzmq>=13 in c:\users\kanishk\anaconda3\lib\site-packages
(from jupyter-client>=6.1.5->nbclient<0.6.0,>=0.5.0->nbconvert) (22.2.1)
Requirement already satisfied: pywin32>=1.0 in c:\users\kanishk\anaconda3\lib\site-packag
es (from jupyter-core->nbconvert) (228)
Requirement already satisfied: ipython-genutils in c:\users\kanishk\anaconda3\lib\site-pa
ckages (from nbformat>=4.4->nbconvert) (0.2.0)
Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in c:\users\kanishk\anaconda3\lib\
site-packages (from nbformat>=4.4->nbconvert) (3.2.0)
Requirement already satisfied: pyrsistent>=0.14.0 in c:\users\kanishk\anaconda3\lib\site-
packages (from jsonschema!=2.5.0,>=2.4->nbformat>=4.4->nbconvert) (0.18.0)
Requirement already satisfied: setuptools in c:\users\kanishk\anaconda3\lib\site-packages
(from jsonschema!=2.5.0,>=2.4->nbformat>=4.4->nbconvert) (58.0.4)
Requirement already satisfied: six>=1.11.0 in c:\users\kanishk\anaconda3\lib\site-package
s (from jsonschema!=2.5.0,>=2.4->nbformat>=4.4->nbconvert) (1.16.0)
Requirement already satisfied: attrs>=17.4.0 in c:\users\kanishk\anaconda3\lib\site-packa
ges (from jsonschema!=2.5.0,>=2.4->nbformat>=4.4->nbconvert) (21.2.0)
Requirement already satisfied: packaging in c:\users\kanishk\anaconda3\lib\site-packages
(from bleach->nbconvert) (21.0)
Requirement already satisfied: webencodings in c:\users\kanishk\anaconda3\lib\site-packag
```

```
es (from bleach->nbconvert) (0.5.1)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\kanishk\anaconda3\lib\site-pa
ckages (from packaging->bleach->nbconvert) (3.0.4)
Note: you may need to restart the kernel to use updated packages.
```

In [ ]: