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

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

data.head()
```

```
                                                 App        Category
Rating  \
0     Photo Editor & Candy Camera & Grid & ScrapBook  ART_AND_DESIGN
4.1
1                                 Coloring book moana  ART_AND_DESIGN
3.9
2  U Launcher Lite – FREE Live Cool Themes, Hide ...  ART_AND_DESIGN
4.7
3                               Sketch - Draw & Paint  ART_AND_DESIGN
4.5
4              Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN
4.3

  Reviews  Size     Installs  Type Price Content Rating  \
0     159   19M      10,000+  Free     0       Everyone
1     967   14M     500,000+  Free     0       Everyone
2   87510  8.7M   5,000,000+  Free     0       Everyone
3  215644   25M  50,000,000+  Free     0           Teen
4     967  2.8M     100,000+  Free     0       Everyone

                      Genres      Last Updated         Current Ver  \
0               Art & Design   January 7, 2018               1.0.0
1  Art & Design;Pretend Play  January 15, 2018               2.0.0
2               Art & Design    August 1, 2018               1.2.4
3               Art & Design      June 8, 2018  Varies with device
4    Art & Design;Creativity     June 20, 2018                 1.1

    Android Ver
0  4.0.3 and up
1  4.0.3 and up
2  4.0.3 and up
3    4.2 and up
4    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)
```

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

```
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
```

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

```
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
```

```
data = data.dropna()

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

```
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
```

```
data.shape
```

```
(9360, 13)
```

4(i).

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

data.head()
```

```
                                                App        Category
Rating  \
0     Photo Editor & Candy Camera & Grid & ScrapBook  ART_AND_DESIGN
4.1
1                                Coloring book moana  ART_AND_DESIGN
3.9
2  U Launcher Lite – FREE Live Cool Themes, Hide ...  ART_AND_DESIGN
4.7
3                             Sketch - Draw & Paint  ART_AND_DESIGN
4.5
4               Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN
4.3

  Reviews  Size     Installs  Type Price Content Rating  \
0     159  19.0      10,000+  Free     0       Everyone
1     967  14.0     500,000+  Free     0       Everyone
2   87510   8.7   5,000,000+  Free     0       Everyone
3  215644  25.0  50,000,000+  Free     0           Teen
4     967   2.8     100,000+  Free     0       Everyone
```

```
                      Genres      Last Updated         Current Ver  \
0               Art & Design   January 7, 2018               1.0.0
1  Art & Design;Pretend Play  January 15, 2018               2.0.0
2               Art & Design    August 1, 2018               1.2.4
3               Art & Design      June 8, 2018  Varies with device
4    Art & Design;Creativity     June 20, 2018                 1.1

    Android Ver
0  4.0.3 and up
1  4.0.3 and up
2  4.0.3 and up
3    4.2 and up
4    4.4 and up
```

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

```
data
```

```
                                                 App
Category  \
0         Photo Editor & Candy Camera & Grid & ScrapBook
ART_AND_DESIGN
1                                    Coloring book moana
ART_AND_DESIGN
2      U Launcher Lite – FREE Live Cool Themes, Hide ...
ART_AND_DESIGN
3                                  Sketch - Draw & Paint
ART_AND_DESIGN
4                  Pixel Draw - Number Art Coloring Book
ART_AND_DESIGN
...                                                  ...
...
10834                                      FR Calculator
FAMILY
10836                                   Sya9a Maroc - FR
FAMILY
10837                    Fr. Mike Schmitz Audio Teachings
FAMILY
10839                      The SCP Foundation DB fr nn5n
BOOKS_AND_REFERENCE
10840      iHoroscope - 2018 Daily Horoscope & Astrology
LIFESTYLE

       Rating Reviews     Size     Installs  Type Price Content Rating
\
0         4.1     159  19000.0      10,000+  Free     0       Everyone

1         3.9     967  14000.0     500,000+  Free     0       Everyone

2         4.7   87510   8700.0   5,000,000+  Free     0       Everyone
```

```
3          4.5  215644  25000.0  50,000,000+  Free     0            Teen
4          4.3     967   2800.0     100,000+  Free     0       Everyone
...        ...     ...      ...          ...   ...   ...             ...
10834      4.0       7   2600.0         500+  Free     0       Everyone
10836      4.5      38  53000.0       5,000+  Free     0       Everyone
10837      5.0       4   3600.0         100+  Free     0       Everyone
10839      4.5     114      0.0       1,000+  Free     0     Mature 17+
10840      4.5  398307  19000.0  10,000,000+  Free     0       Everyone

                          Genres      Last Updated         Current Ver
\
0                    Art & Design   January 7, 2018               1.0.0
1      Art & Design;Pretend Play  January 15, 2018               2.0.0
2                    Art & Design    August 1, 2018               1.2.4
3                    Art & Design      June 8, 2018  Varies with device
4         Art & Design;Creativity     June 20, 2018                 1.1
...                           ...               ...                 ...
10834                  Education     June 18, 2017               1.0.0
10836                  Education     July 25, 2017                1.48
10837                  Education      July 6, 2018                 1.0
10839          Books & Reference  January 19, 2015  Varies with device
10840                  Lifestyle     July 25, 2018  Varies with device

               Android Ver
0            4.0.3 and up
1            4.0.3 and up
2            4.0.3 and up
3              4.2 and up
4              4.4 and up
...                   ...
10834          4.1 and up
```

```
10836          4.1 and up
10837          4.1 and up
10839  Varies with device
10840  Varies with device

[9360 rows x 13 columns]
```

4(ii).

```
data.info()

<class 'pandas.core.frame.DataFrame'>
Index: 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

data['Reviews'] = data['Reviews'].astype(float)

data.info()

<class 'pandas.core.frame.DataFrame'>
Index: 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(iii).

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

data.head()
```

```
                                                 App        Category
Rating  \
0     Photo Editor & Candy Camera & Grid & ScrapBook  ART_AND_DESIGN
4.1
1                                 Coloring book moana  ART_AND_DESIGN
3.9
2  U Launcher Lite – FREE Live Cool Themes, Hide ...  ART_AND_DESIGN
4.7
3                               Sketch - Draw & Paint  ART_AND_DESIGN
4.5
4              Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN
4.3

    Reviews     Size    Installs  Type Price Content Rating  \
0     159.0  19000.0     10000.0  Free     0       Everyone
1     967.0  14000.0    500000.0  Free     0       Everyone
2   87510.0   8700.0   5000000.0  Free     0       Everyone
3  215644.0  25000.0  50000000.0  Free     0           Teen
4     967.0   2800.0    100000.0  Free     0       Everyone

                      Genres      Last Updated         Current Ver  \
0               Art & Design   January 7, 2018               1.0.0
1  Art & Design;Pretend Play  January 15, 2018               2.0.0
2               Art & Design    August 1, 2018               1.2.4
3               Art & Design      June 8, 2018  Varies with device
4    Art & Design;Creativity     June 20, 2018                 1.1

    Android Ver
0  4.0.3 and up
1  4.0.3 and up
2  4.0.3 and up
3    4.2 and up
4    4.4 and up
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 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
```

```
import pandas as pd

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

data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 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
```

4(iv).

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

data.head()
```

```
                                                 App        Category
Rating  \
0     Photo Editor & Candy Camera & Grid & ScrapBook  ART_AND_DESIGN
4.1
1                                 Coloring book moana  ART_AND_DESIGN
3.9
2  U Launcher Lite – FREE Live Cool Themes, Hide ...  ART_AND_DESIGN
4.7
3                               Sketch - Draw & Paint  ART_AND_DESIGN
4.5
4              Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN
4.3

    Reviews     Size  Installs  Type  Price Content Rating  \
0     159.0  19000.0     10000  Free    0.0       Everyone
1     967.0  14000.0    500000  Free    0.0       Everyone
2   87510.0   8700.0   5000000  Free    0.0       Everyone
3  215644.0  25000.0  50000000  Free    0.0           Teen
4     967.0   2800.0    100000  Free    0.0       Everyone

                      Genres      Last Updated         Current Ver  \
0               Art & Design   January 7, 2018               1.0.0
1  Art & Design;Pretend Play  January 15, 2018               2.0.0
2               Art & Design    August 1, 2018               1.2.4
3               Art & Design      June 8, 2018  Varies with device
4    Art & Design;Creativity     June 20, 2018                 1.1

    Android Ver
0  4.0.3 and up
1  4.0.3 and up
2  4.0.3 and up
3    4.2 and up
4    4.4 and up
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 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
```

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

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 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
```

4(V-a).

```
data.shape
```

```
(9360, 13)
```

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

```
data.shape
```

```
(9360, 13)
```

4(V-b).

```
data.shape
```

```
(9360, 13)
```

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

```
data.shape
```

```
(9353, 13)
```

4(V-c).

```
data.shape
```

```
(9353, 13)
```

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

```
data.shape
```

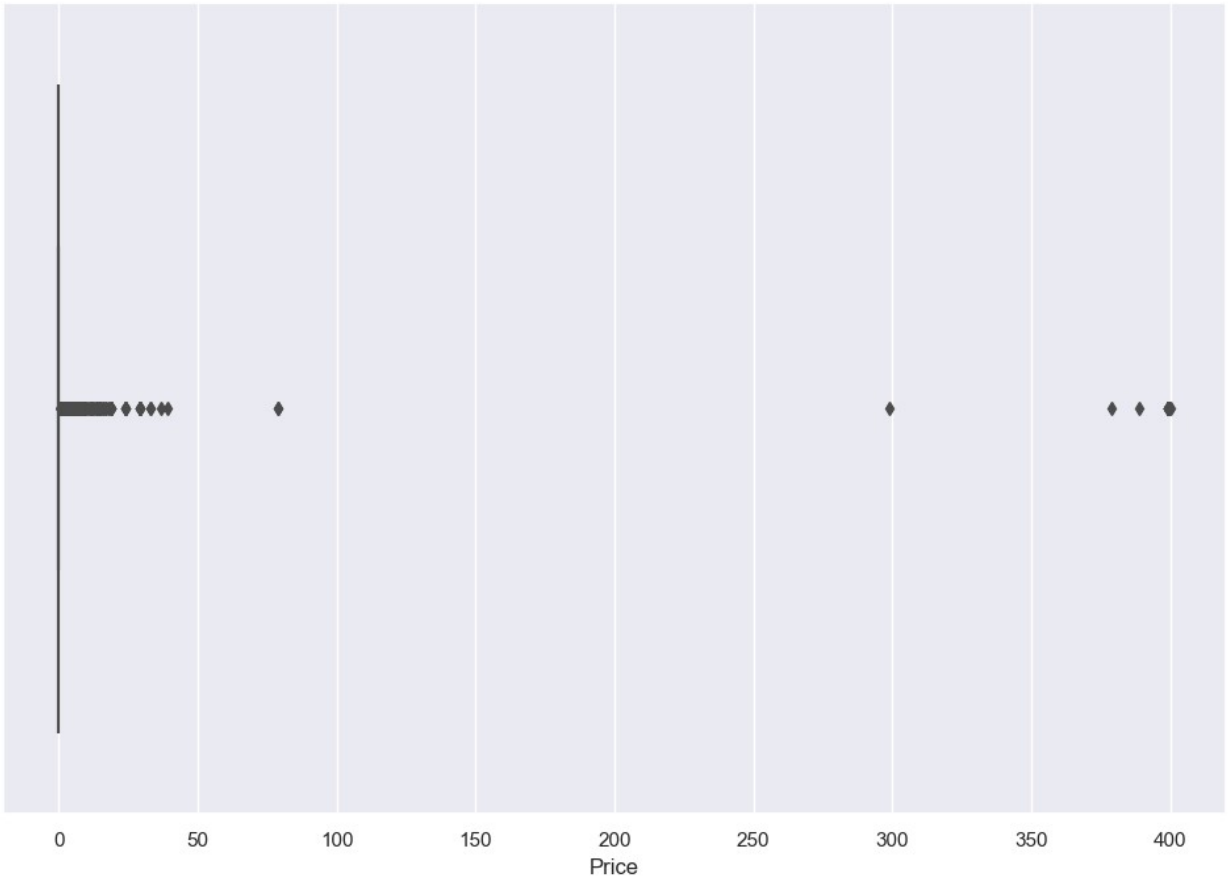
```
(9353, 13)
```

5(i).

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

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

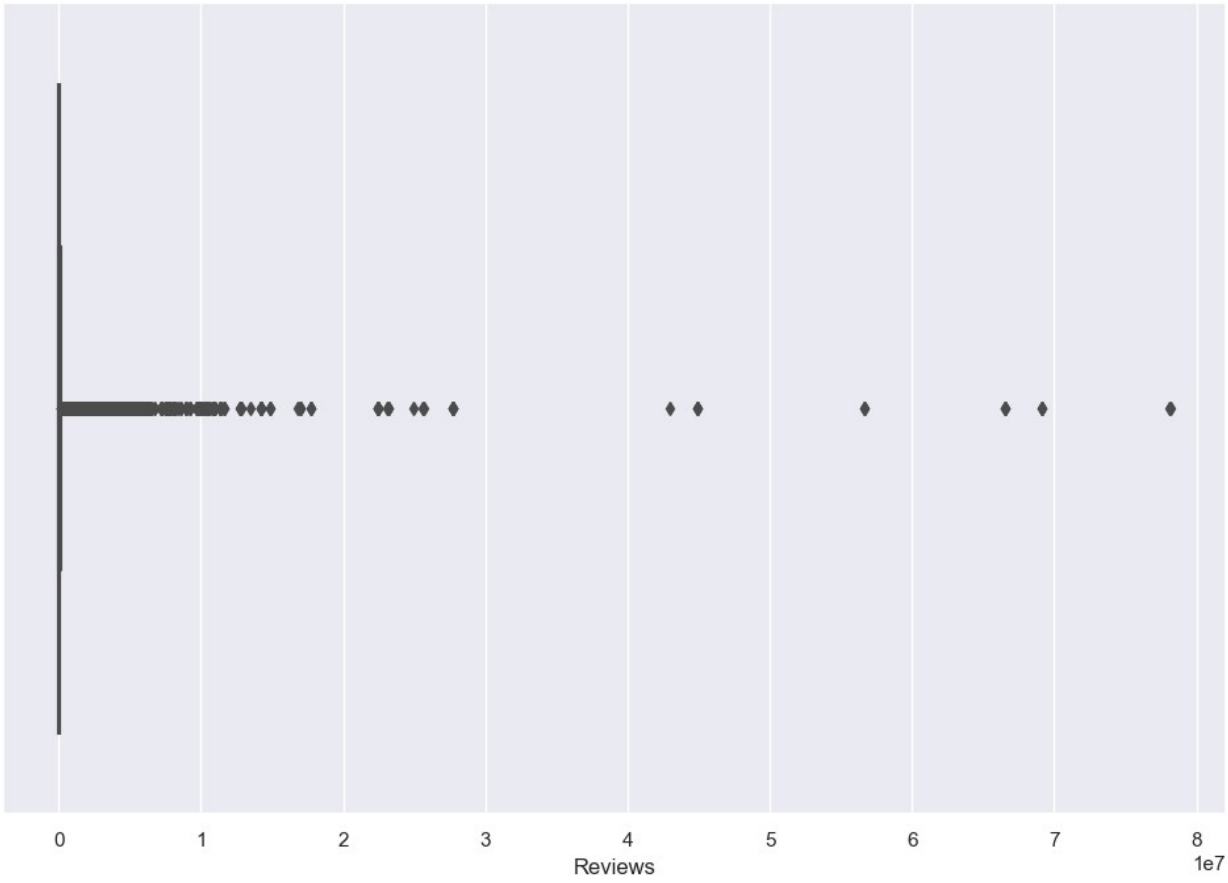
```
<Axes: xlabel='Price'>
```



5(ii).

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

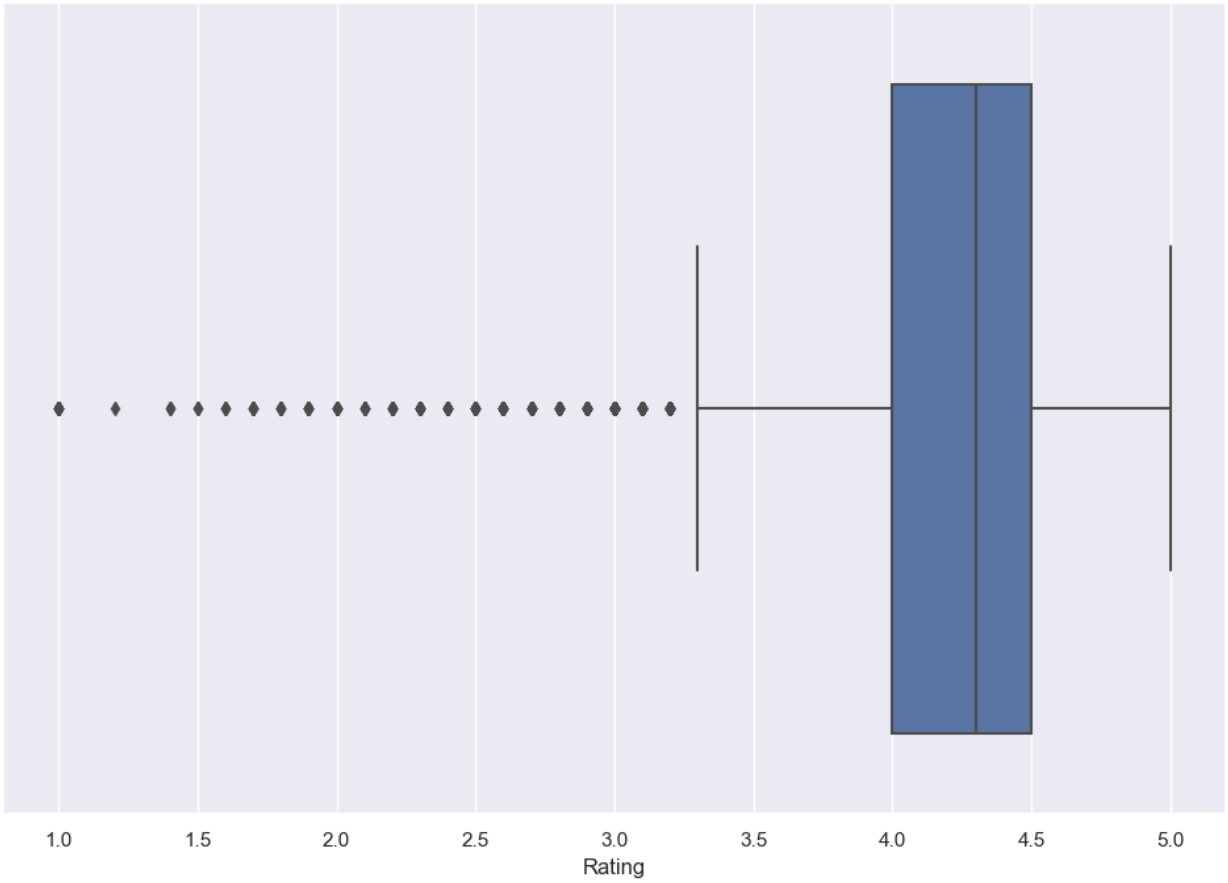
```
<Axes: xlabel='Reviews'>
```



5(iii).

```
sns.boxplot(x=data['Rating'])
```

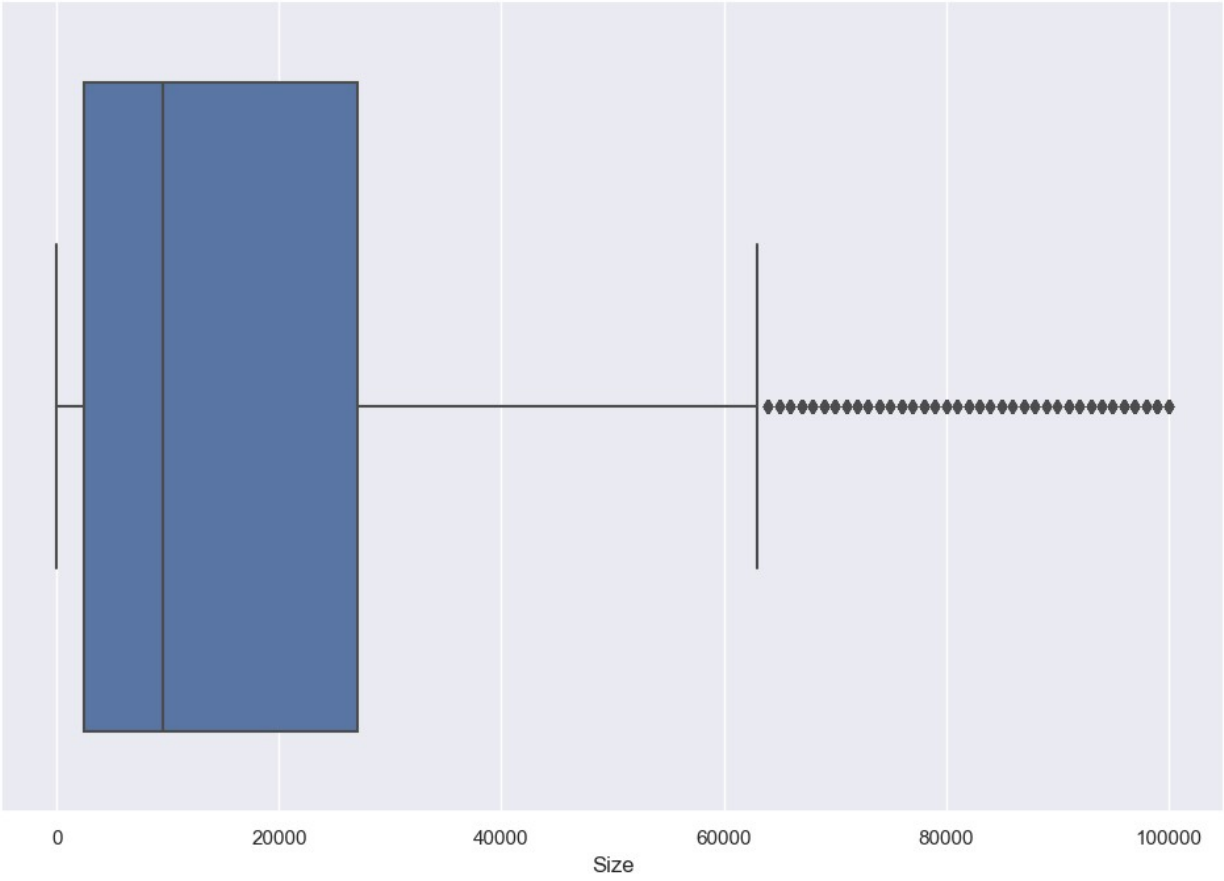
```
<Axes: xlabel='Rating'>
```



5(iv).

```
sns.boxplot(x=data['Size'])
```

```
<Axes: xlabel='Size'>
```



6(i).

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

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

data.shape

(9353, 13)

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

data.shape

(9338, 13)
```

6(ii).

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

data.shape

(8885, 13)
```

6(iii).

```
numeric_data = data.select_dtypes(include='number')

import pandas as pd
numeric_data.quantile([0.1, 0.25, 0.50, 0.70, 0.90, 0.95, 0.99], axis
= 0)
```

	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

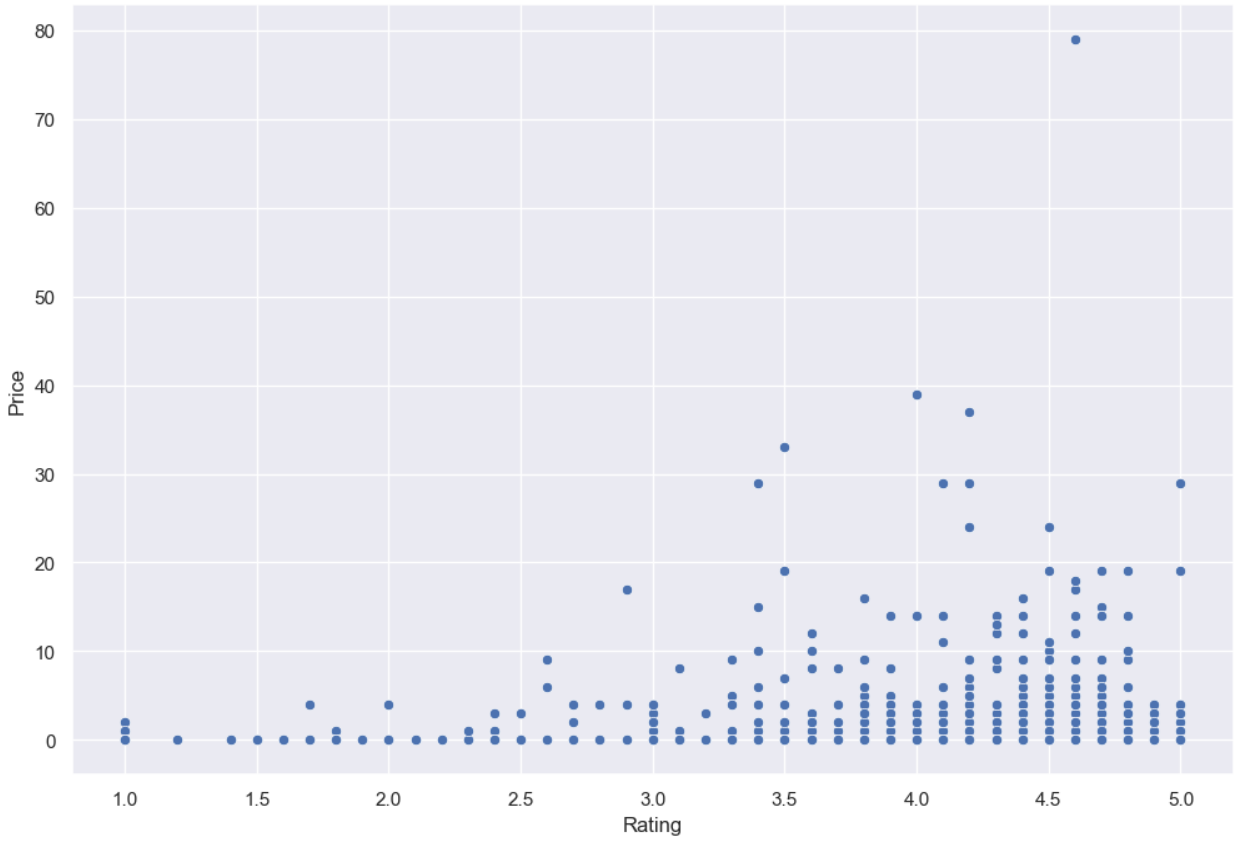
```
#Dropping more than 10000000 Installs Values
data.drop(data[data['Installs']>10000000].index, inplace=True)

data.shape
```

```
(8496, 13)
```

7(i).

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

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

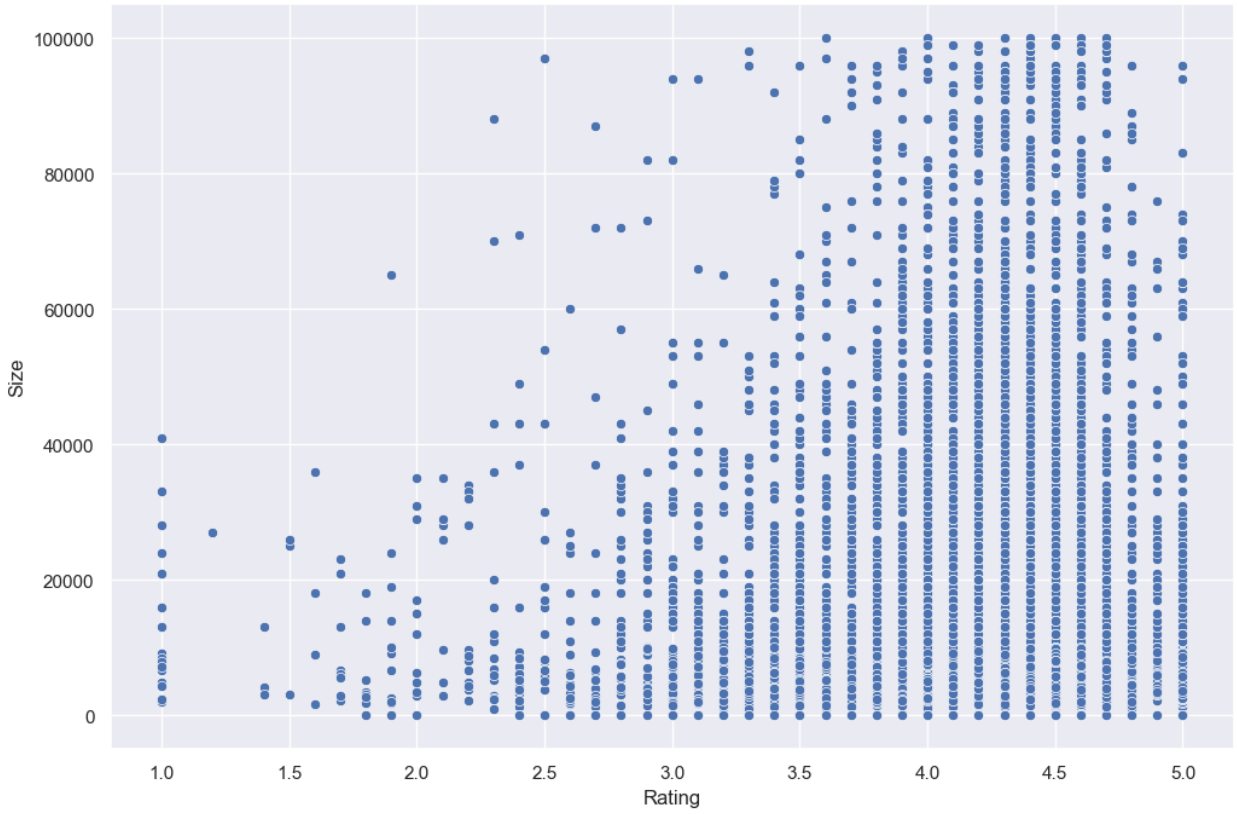


Yes, Paid apps are having a higher ratings compared to frre apps

7(ii).

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

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

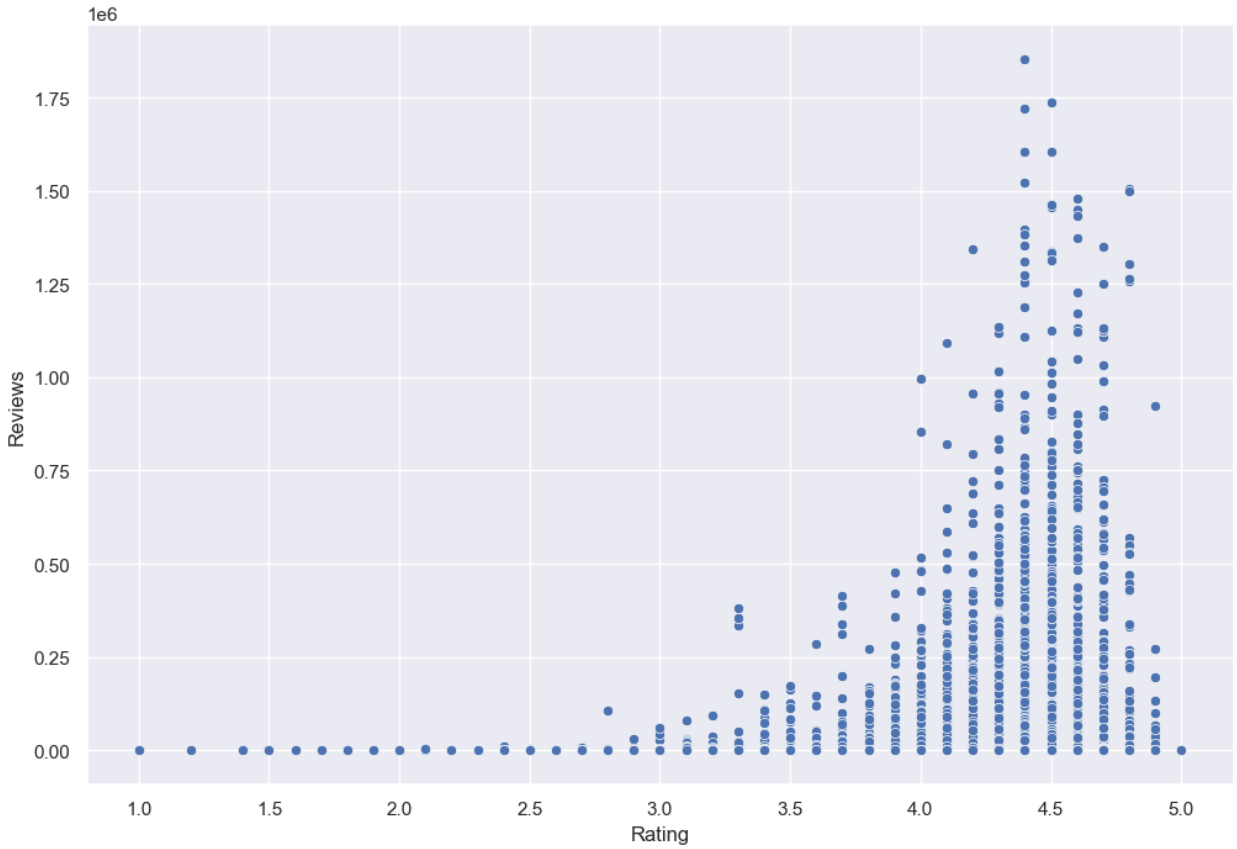


Yes, It is clear that heavier apps are rated better.

7(iii).

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

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

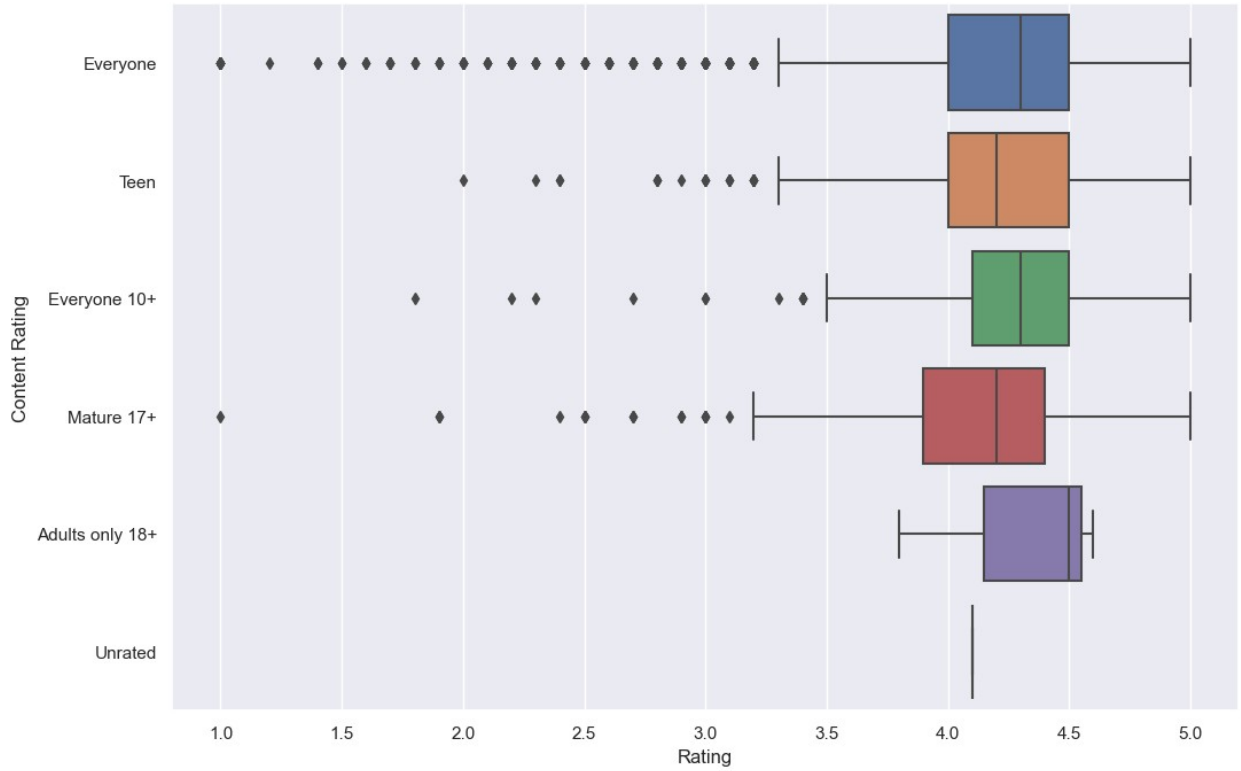


By seeing above graph its clear that more reviews makes a app rating better.

7(iv).

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

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

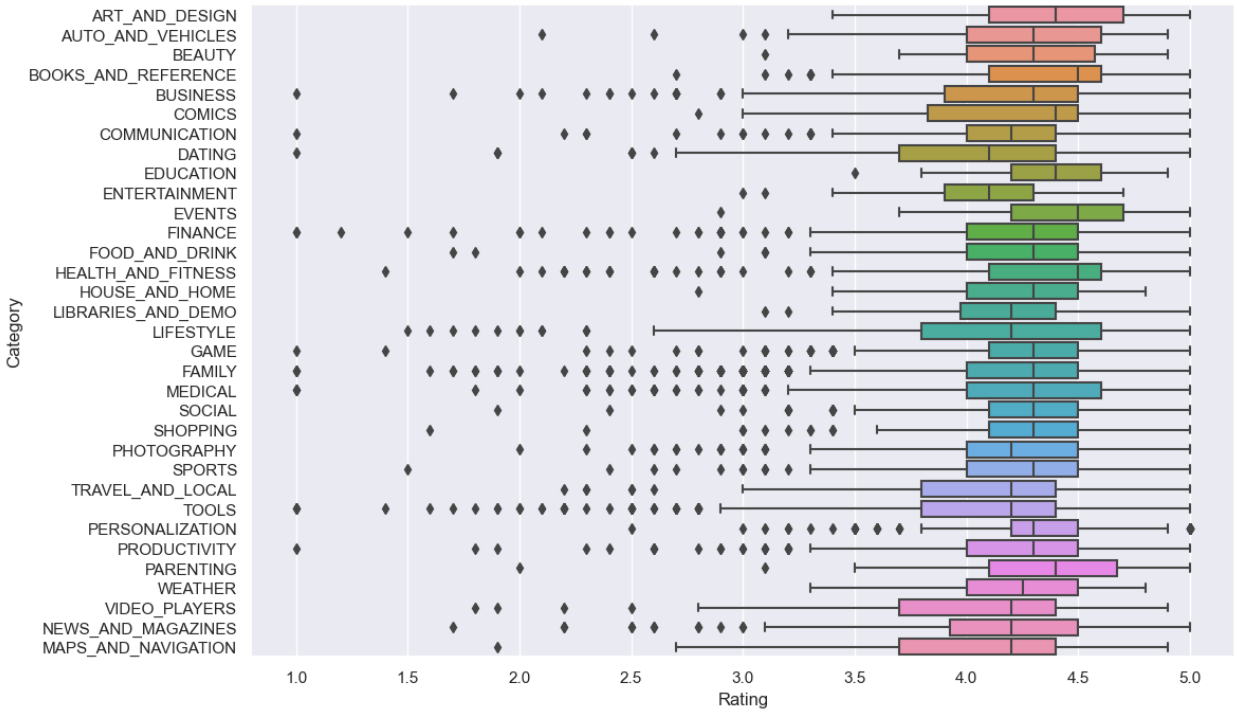


Apps which are rated as Everyone has many bad ratings compared to othere as it has so many outliers and 18+ apps have better ratings!

7(v).

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

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



Apps which are marketed under category of EVENTS has good ratings compare to others!!

8(i).

```
input1=numeric_data
```

```
input1.head()
```

	Rating	Reviews	Size	Installs	Price
0	4.1	159.0	19000.0	10000	0
1	3.9	967.0	14000.0	500000	0
2	4.7	87510.0	8700.0	5000000	0
3	4.5	215644.0	25000.0	50000000	0
4	4.3	967.0	2800.0	100000	0

```
import pandas as pd
import numpy as np
input1.skew()
```

```
Rating      -1.792695
Reviews      3.819200
Size         1.639726
Installs    19.524193
Price       18.478466
dtype: float64
```

```
reviewskew=np.log1p(input1['Reviews'])
input1['Reviews']=reviewskew
```

```
reviewskew.skew()
```

```
-0.18828800378554247
```

```
installsskew=np.log1p(input1['Installs'])
input1['Installs']
```

```
0           10000
1          500000
2         5000000
3        50000000
4          100000
           ...
10834         500
10836        5000
10837         100
10839        1000
10840    10000000
Name: Installs, Length: 8885, dtype: int32
```

```
installsskew.skew()
```

```
-0.409593704318801
```

```
input1.head()
```

	Rating	Reviews	Size	Installs	Price
0	4.1	5.075174	19000.0	10000	0
1	3.9	6.875232	14000.0	500000	0
2	4.7	11.379520	8700.0	5000000	0
3	4.5	12.281389	25000.0	50000000	0
4	4.3	6.875232	2800.0	100000	0

```
input1=data
```

8(ii).

```
input1.drop(['Last Updated','Current Ver','Android
Ver','App','Type'],axis=1,inplace=True)
```

```
input1.head()
```

	Category	Rating	Reviews	Size	Installs	Price	Content Rating
0	ART_AND_DESIGN	4.1	159.0	19000.0	10000	0	Everyone
1	ART_AND_DESIGN	3.9	967.0	14000.0	500000	0	Everyone
2	ART_AND_DESIGN	4.7	87510.0	8700.0	5000000	0	Everyone
4	ART_AND_DESIGN	4.3	967.0	2800.0	100000	0	

```
Everyone
5  ART_AND_DESIGN     4.4    167.0   5600.0     50000      0
Everyone

                       Genres
0               Art & Design
1  Art & Design;Pretend Play
2               Art & Design
4    Art & Design;Creativity
5               Art & Design

input1.shape

(8496, 8)
```

8(iii).

```
input2=input1

input2.head()

   Rating  Reviews     Size  Installs  Price Content Rating  \
0     4.1    159.0  19000.0     10000      0       Everyone
1     3.9    967.0  14000.0    500000      0       Everyone
2     4.7  87510.0   8700.0   5000000      0       Everyone
4     4.3    967.0   2800.0    100000      0       Everyone
5     4.4    167.0   5600.0     50000      0       Everyone

                       Genres
0               Art & Design
1  Art & Design;Pretend Play
2               Art & Design
4    Art & Design;Creativity
5               Art & Design
```

Applying Dummy Encoding on category column.

```
#Get unique values in column 'Category'
input2.Category.unique()

---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_2712\2121897957.py in ?()
      1 #Get unique values in column 'Category'
----> 2 input2.Category.unique()

~\.matplotlib\Lib\site-packages\pandas\core\generic.py in ?(self, name)
```

```
   6200             and name not in self._accessors
   6201             and
self._info_axis._can_hold_identifiers_and_holds_name(name)
   6202         ):
   6203             return self[name]
-> 6204         return object.__getattribute__(self, name)

AttributeError: 'DataFrame' object has no attribute 'Category'
```

```
input2.Category=pd.Categorical(input2.Category)

x=input2[['Category']]
del input2['Category']

dummies=pd.get_dummies(x,prefix='Category')
input2=pd.concat([input2,dummies],axis=1)

input2.head()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call
last)
~\AppData\Local\Temp\ipykernel_2712\3477892310.py in ?()
----> 1 input2.Category=pd.Categorical(input2.Category)
      2
      3 x=input2[['Category']]
      4 del input2['Category']

~\.matplotlib\Lib\site-packages\pandas\core\generic.py in ?(self,
name)
   6200             and name not in self._accessors
   6201             and
self._info_axis._can_hold_identifiers_and_holds_name(name)
   6202         ):
   6203             return self[name]
-> 6204         return object.__getattribute__(self, name)

AttributeError: 'DataFrame' object has no attribute 'Category'
```

```
input2.shape
```

```
(8496, 7)
```

Applying Dummy Encoding on 'Genres' Column.

```
#Getting unique Values in column 'Geners'
input2['Genres'].unique()
```

```
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)
```

As there are too many categories under genres. Hence, we should try to reduce some categories which have a very few samples under them and put them in new column called 'Others'

```
lists=[]
for i in input2.Genres.value_counts().index:
    if input2.Genres.value_counts()[i]<20:
        lists.append(i)
input2.Genres=['Other'if i in lists else i for i in input2.Genres]

input2['Genres'].unique()

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)

input2.Genres=pd.Categorical(input2['Genres'])
x=input2[['Genres']]
del input2['Genres']
dummies=pd.get_dummies(x,prefix='Genres')
input2=pd.concat([input2,dummies],axis=1)

input2.head()

   Rating  Reviews     Size  Installs  Price  Category_ART_AND_DESIGN
\
0     4.1    159.0  19000.0     10000      0                    True

1     3.9    967.0  14000.0    500000      0                    True

2     4.7  87510.0   8700.0   5000000      0                    True
```

```
4     4.3    967.0   2800.0    100000      0                       True

5     4.4    167.0   5600.0     50000      0                       True


   Category_AUTO_AND_VEHICLES  Category_BEAUTY
Category_BOOKS_AND_REFERENCE  \
0                        False            False
False
1                        False            False
False
2                        False            False
False
4                        False            False
False
5                        False            False
False

   Category_BUSINESS  ...  Genres_Trivia  Genres_Video Players &
Editors  \
0              False  ...           False
False
1              False  ...           False
False
2              False  ...           False
False
4              False  ...           False
False
5              False  ...           False
False

   Genres_Weather  Genres_Word  Content Rating_Adults only 18+  \
0           False        False                            False
1           False        False                            False
2           False        False                            False
4           False        False                            False
5           False        False                            False

   Content Rating_Everyone  Content Rating_Everyone 10+  \
0                     True                        False
1                     True                        False
2                     True                        False
4                     True                        False
5                     True                        False

   Content Rating_Mature 17+  Content Rating_Teen  Content
Rating_Unrated
0                      False                False
False
1                      False                False
```

```
False
2                       False                 False
False
4                       False                 False
False
5                       False                 False
False

[5 rows x 96 columns]

input2.shape

(8496, 7)
```

9 and 10

```
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

d1=input2
X=d1.drop('Rating',axis=1)
y=d1['Rating']

XTrain, Xtest, ytrain, ytest=tts(X,y, test_size=0.3, random_state=5)
```

11

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

---------------------------------------------------------------------------
NameError                                 Traceback (most recent call
last)
Cell In[126], line 2
      1 reg_all=LR()
----> 2 reg_all.fit(Xtrain,ytrain)

NameError: name 'Xtrain' is not defined
```