Project :App Rating Prediction..

1. Load the data file using pandas

In [83]:
```
import pandas as pd
import numpy as np
import seaborn as sns
```

In [84]:
```
data = pd.read_csv('googleplaystore.csv')
```

In [85]:
```
data.head()
```

Out[85]:

	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

In [86]:
```
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 [87]:
```
data.shape
```

Out[87]:
```
(10841, 13)
```

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

In [88]:
```
data.isnull().any()
```

Out[88]:
```
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 [89]:
```
data.isnull().sum()
```

Out[89]:
```
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
```

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

In [90]:
```
data = data.dropna()
```

In [91]:
```
data.isnull().any()
```

Out[91]:
```
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 [92]:
```
data.shape
```

Out[92]:
```
(9360, 13)
```

1. Variables seem to have incorrect type and inconsistent formatting. You need to fix them:

    A. Size column has sizes in Kb as well as Mb. To analyze, you'll need to convert these to numeric.

        a. Extract the numeric value from the column

        b. Multiply the value by 1,000, if size is mentioned in Mb

In [93]:
```
data["Size"] = [ float(i.split('M')[0]) if 'M' in i else float(0) for i in data["Size"] ]
```

In [94]:
```
data.head()
```

Out[94]:

	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

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
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

In [95]:
```
data["Size"] = 1000 * data["Size"]
```

In [96]:
```
data
```

Out[96]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	A
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	
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	
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	
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	

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	A
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	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	7	2600.0	500+	Free	0	Everyone	Education	June 18, 2017	1.0.0	
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100+	Free	0	Everyone	Education	July 6, 2018	1.0	
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	
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	

9360 rows × 13 columns

4.

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

In [97]:
```
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
```

In [98]:

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

In [99]:

```
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. A. Installs field is currently stored as string and has values like 1,000,000+.

    a. Treat 1,000,000+ as 1,000,000

    b. remove '+', ',' from the field, convert it to integer

In [100…
```
data["Installs"] = [float(i.replace('+','').replace(',','')) if '+' in i else float(0) for i in data["Installs"]]
```

In [101…
```
data.head()
```

Out[101…

	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

In [102…
```
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
```

In [103…

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

In [104…

```
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. A. Price field is a string and has *symbol*. *Remove*" sign, and convert it to numeric.

In [105…
```
data['Price'] = [ float(i.split('$')[1]) if '$' in i else float(0) for i in data['Price'] ]
```

In [106…
```
data.head()
```

Out[106…

	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

In [107…
```
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 [108…

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

In [109…

```
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
```

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 [110…
```
data.shape
```

Out[110…
```
(9360, 13)
```

In [111…
```
data.drop(data[(data["Reviews"] < 1 ) & (data["Reviews"] > 5)].index, inplace = True)
```

In [112…
```
data.shape
```

Out[112…
```
(9360, 13)
```

5.

2. 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 [113…
```
data.shape
```

Out[113…
```
(9360, 13)
```

In [114…
```
data.drop(data[data['Installs'] < data['Reviews'] ].index, inplace = True)
```

In [115…
```
data.shape
```

Out[115…
```
(9353, 13)
```

5.

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

In [116…

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

In [117…
```
data.shape
```

Out[117…
```
(9353, 13)
```

1. Performing univariate analysis:

   . Boxplot for Price

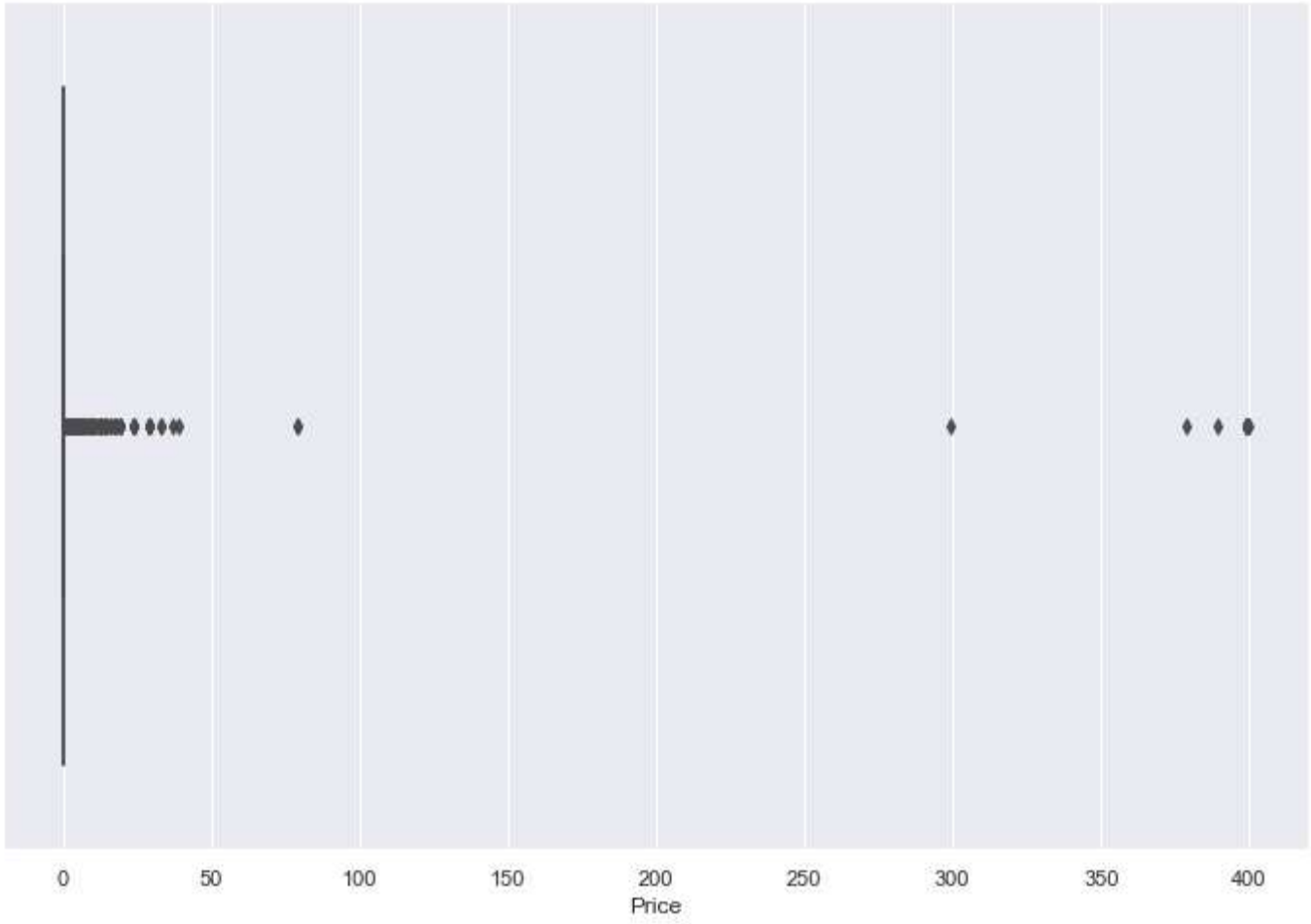
   . Are there any outliers? Think about the price of usual apps on Play Store.

In [118…
```
sns.set(rc={'figure.figsize':(12,8)})
```

In [119…
```
sns.boxplot(data['Price'])
```

```
C:\Users\satis\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a key word 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.
  warnings.warn(
<AxesSubplot:xlabel='Price'>
```

Out[119…



1. . Boxplot for Reviews

. Are there any apps with very high number of reviews? Do the values seem right?

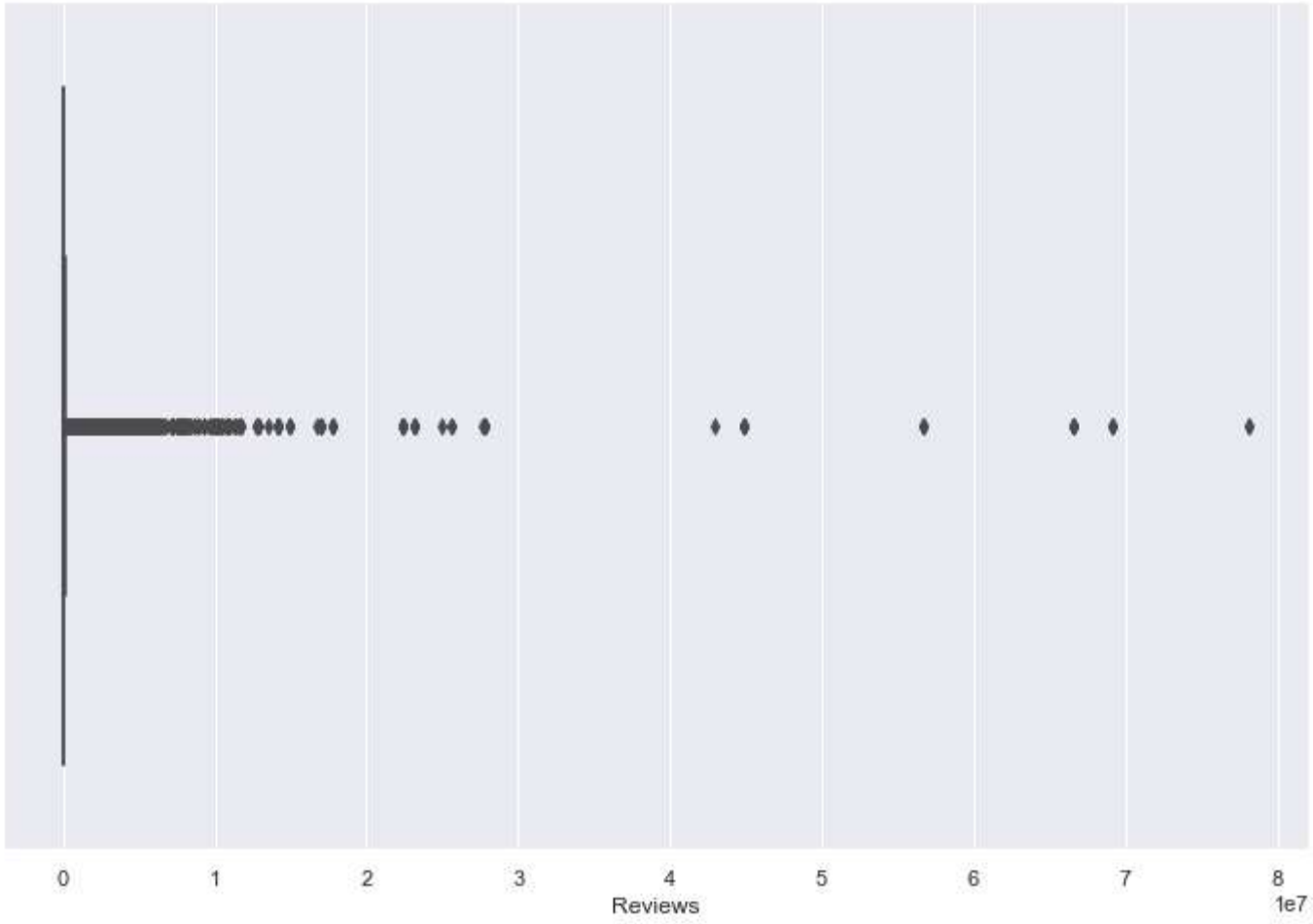
In [120…

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

```
C:\Users\satis\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a key
word 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.
  warnings.warn(
```

Out[120…

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



1. . Histogram for Rating

. How are the ratings distributed? Is it more toward higher ratings?

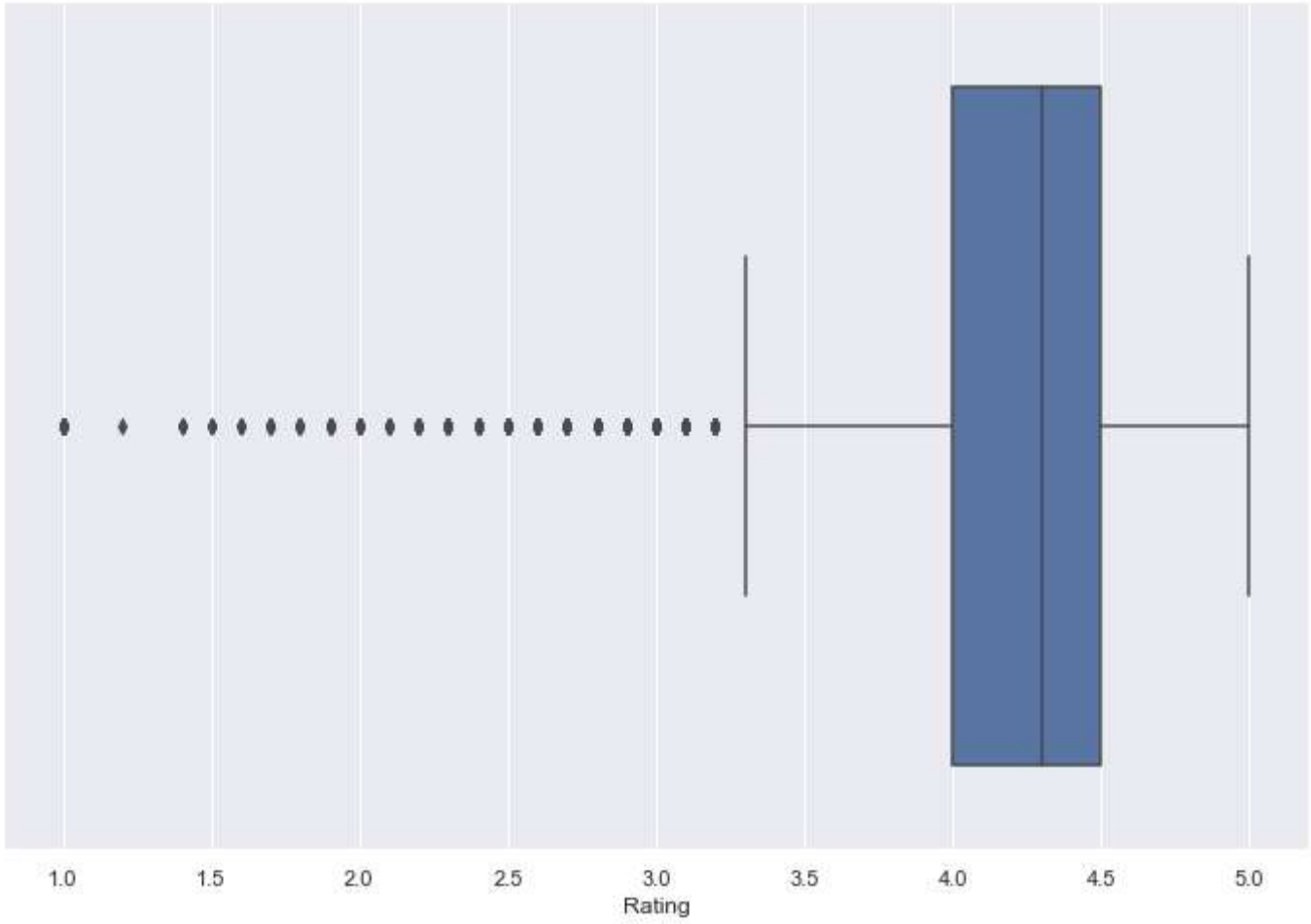
In [121…

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

```
C:\Users\satis\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a key
word 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.
  warnings.warn(
```

Out[121…

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



1. . Histogram for Size

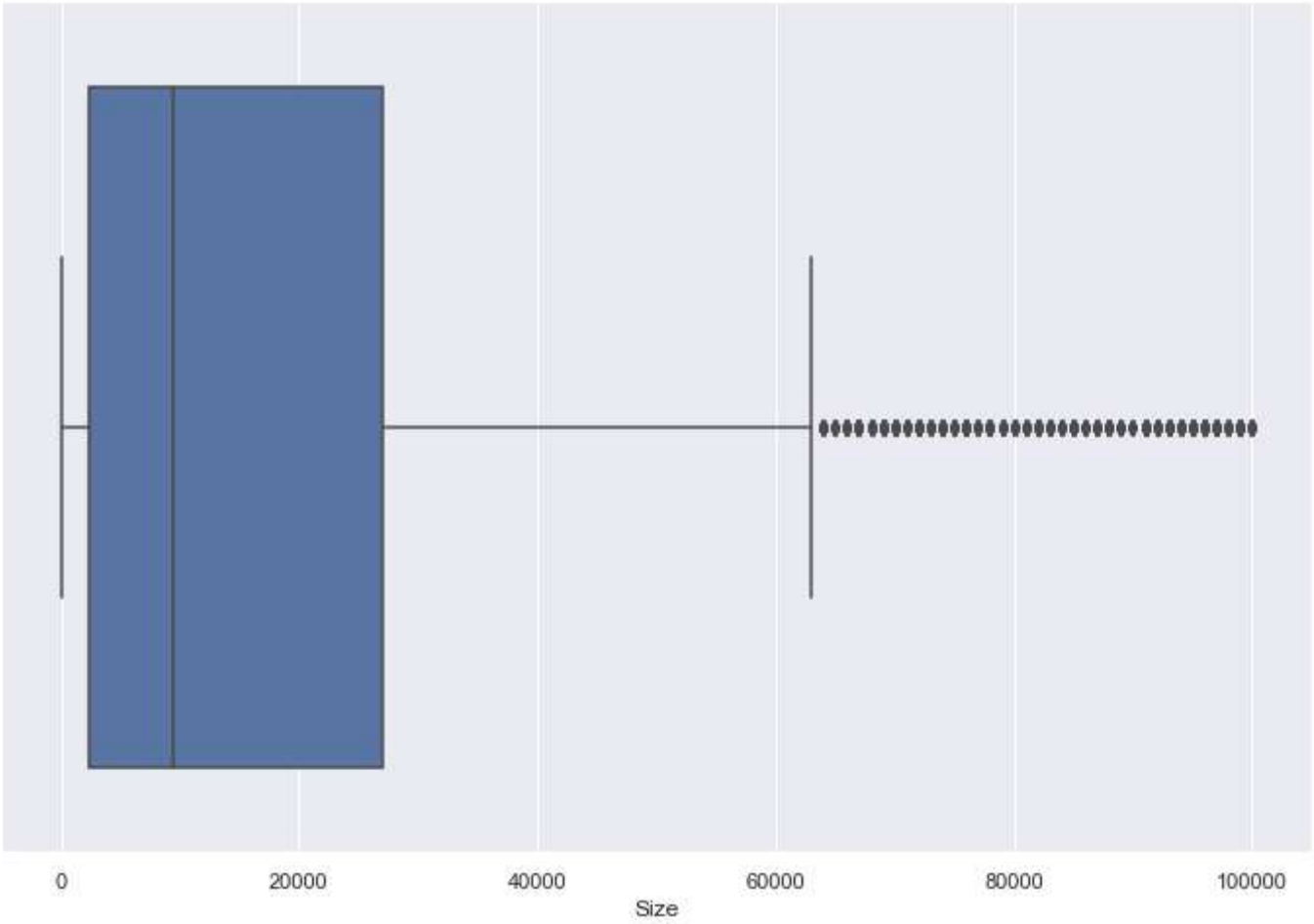
. Note down your observations for the plots made above. Which of these seem to have outliers?

In [122…

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

```
C:\Users\satis\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a key
word 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.
  warnings.warn(
<AxesSubplot:xlabel='Size'>
```

Out[122…



1. Outlier treatment:

    A. Price: From the box plot, it seems like there are some apps with very high price.

        `A price of $200 for an application on the Play Store is very high and suspicious!`

        a. Check out the records with very high price

            i. Is 200 indeed a high price?

        b. Drop these as most seem to be junk apps

In [123…

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

In [124…
```
more_count = len(more[more == True].index)
```

In [125…
```
data.shape
```

Out[125… (9353, 13)

In [126…
```
data.drop(data[data['Price']>200].index, inplace = True)
```

In [127…
```
data.shape
```

Out[127… (9338, 13)

6.

2. 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 [128…
```
data.drop(data[data['Reviews']>2000000].index, inplace = True)
```

In [129…
```
data.shape
```

Out[129… (8885, 13)

1. A. Installs: There seems to be some outliers in this field too.

   Apps having very high number of installs should be dropped from the analysis.

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

   2. Decide a threshold as cutoff for outlier and drop records having values more than that

In [130…
```
data.quantile([.1, .25, .5, .70, .90, .95, .99], axis = 0)
```

Out[130…

	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

In [131…
```
data.drop(data[data['Installs'] > 10000000].index, inplace = True)    # dropping more than 10000000 Installs valueI
```

In [132…
```
data.shape
```

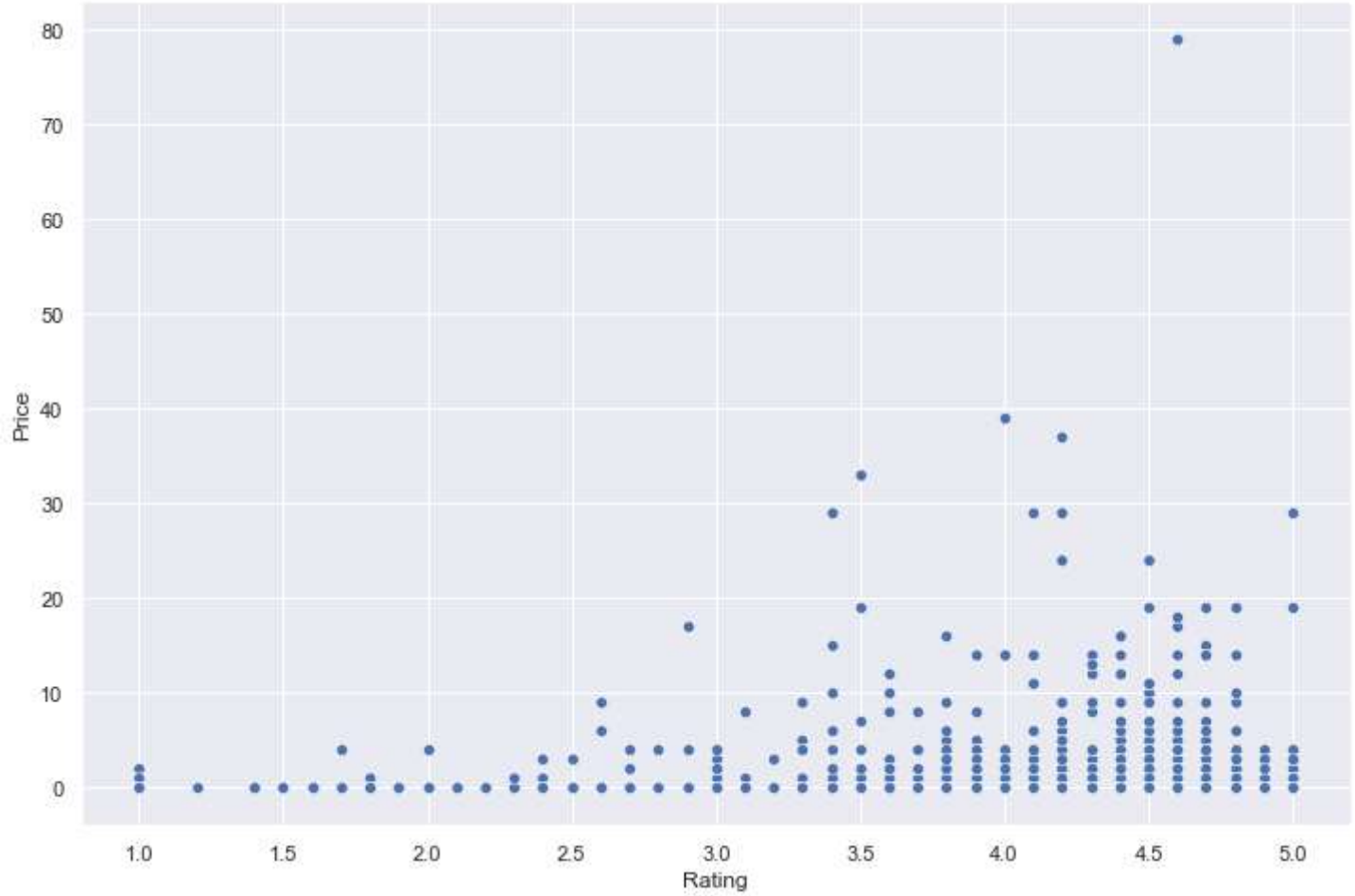
Out[132…
```
(8496, 13)
```

1. Bivariate analysis:

1. 1.Make scatter plot/joinplot for Rating vs. Price What pattern do you observe? Does rating increase with price?

In [133…
```
sns.scatterplot(x='Rating', y='Price',data=data)
```

Out[133…
```
<AxesSubplot:xlabel='Rating', ylabel='Price'>
```



Yes.Paid apps are higher ratings when compare to free apps.

7.

2. Make scatter plot/joinplot for Rating vs. Size

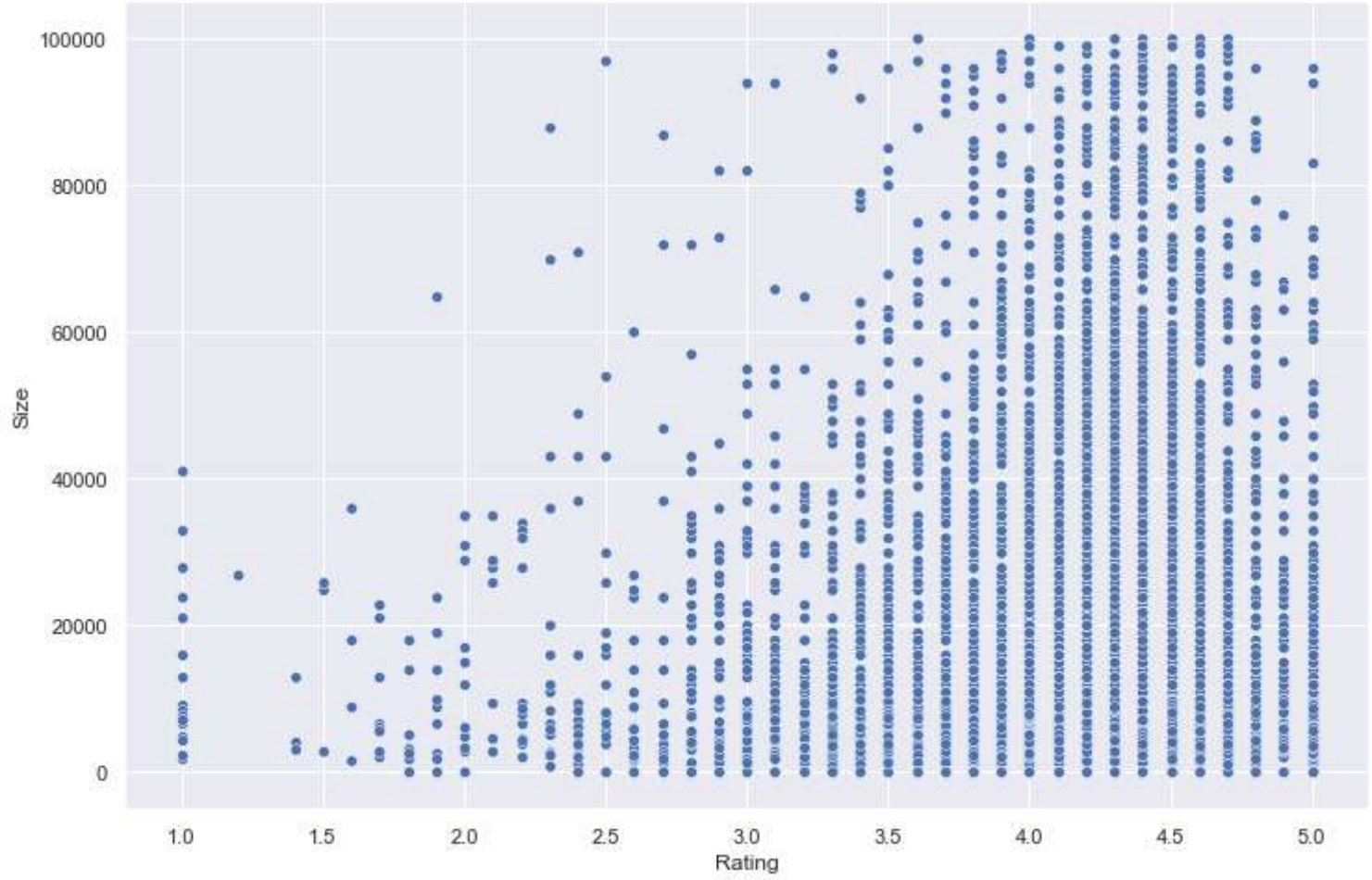
Are heavier apps rated better?

In [134…

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

Out[134…

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



Yes. heavier apps rated better.

7.

1. Make scatter plot/joinplot for Rating vs. Reviews

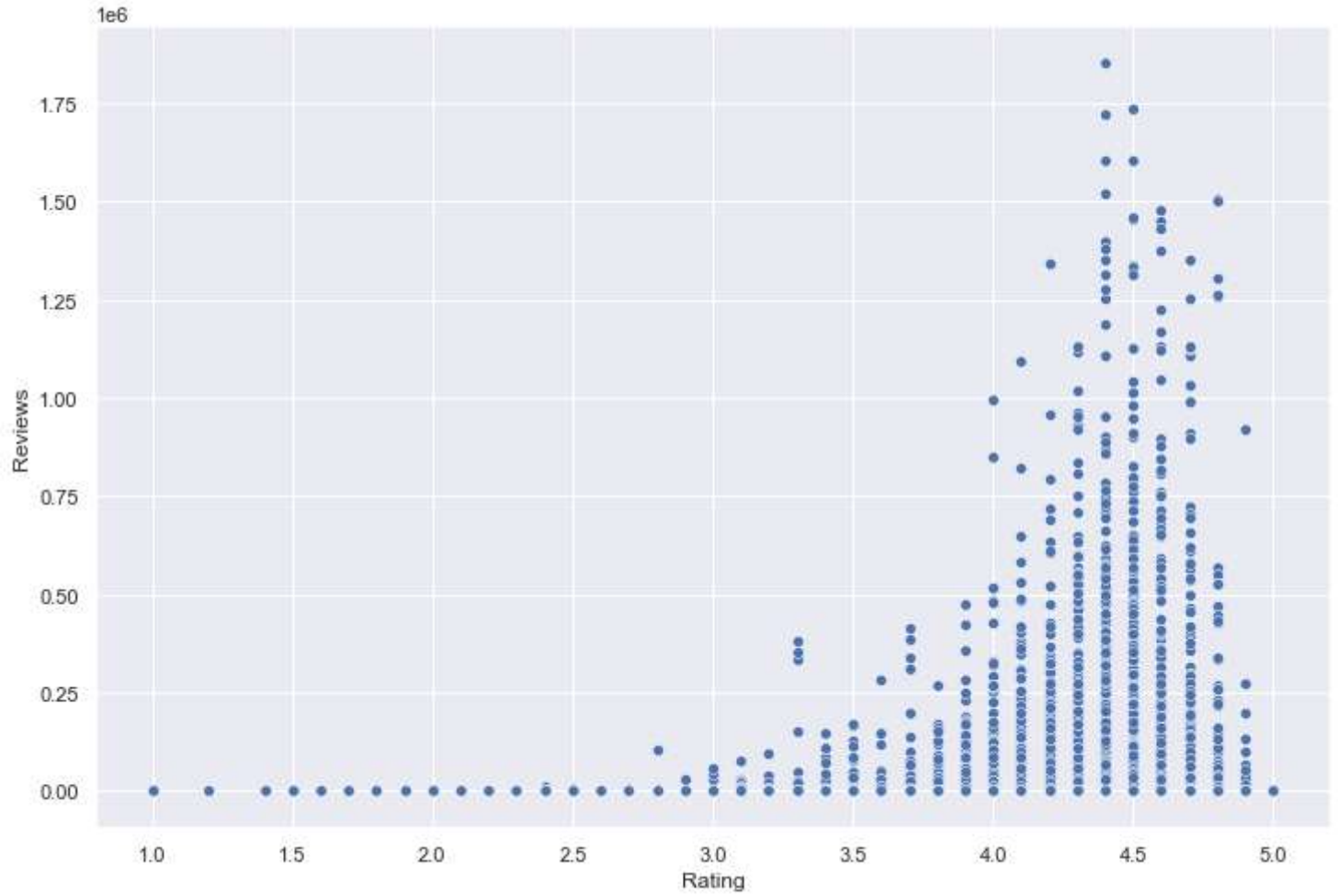
   Does more review mean a better rating always?

In [135…

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

Out[135…

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



Yes.More review mean a better rating always.

7.

1. Make boxplot for Rating vs. Content Rating

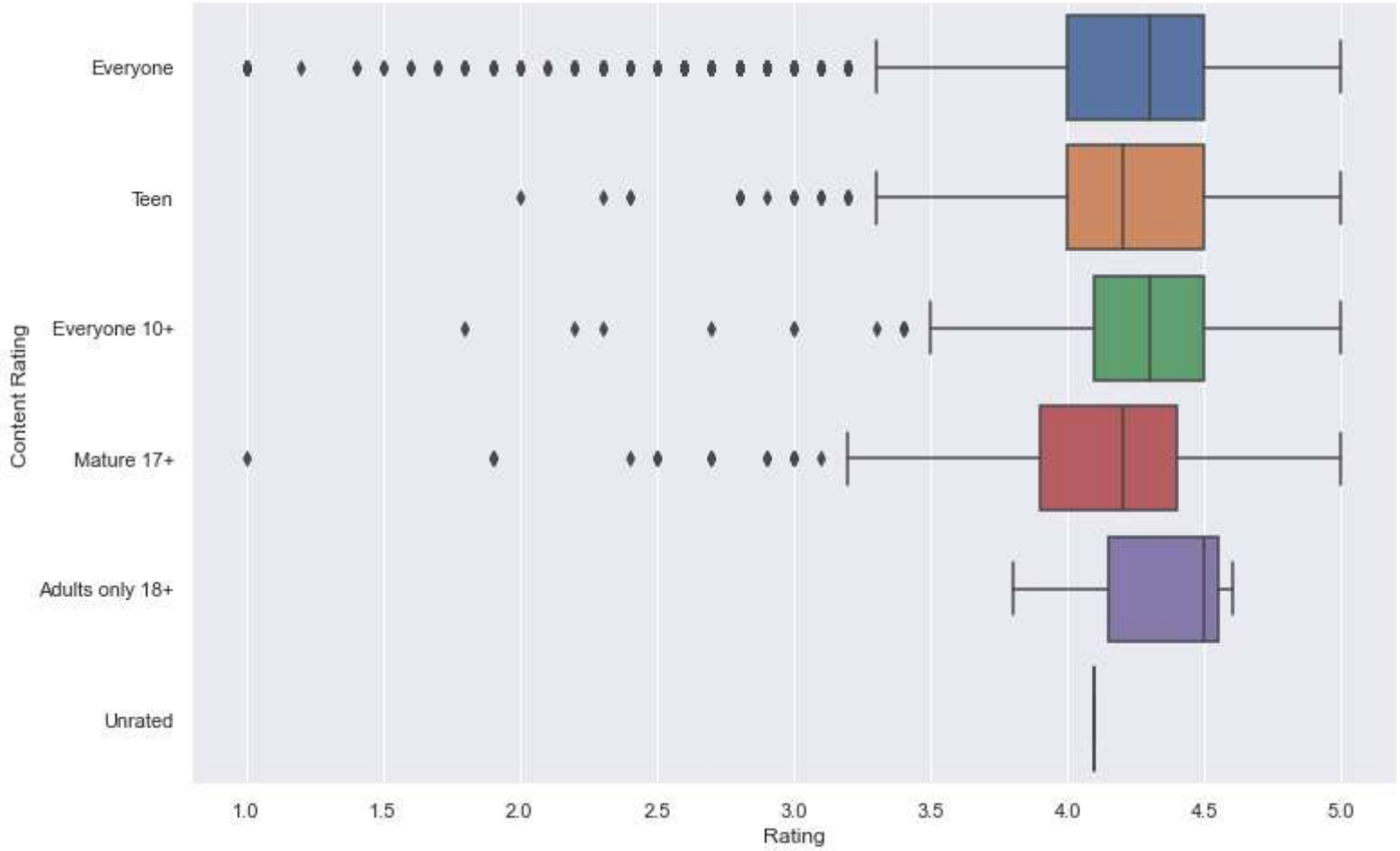
   Is there any difference in the ratings? Are some types liked better?

In [136…

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

Out[136…

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



Apps which has content rating for Everyone has more bad ratings when compare to other sections as it has so much outliers value, while Adults only 18+ have better ratings.

7.

5. Make boxplot for Ratings vs. Category

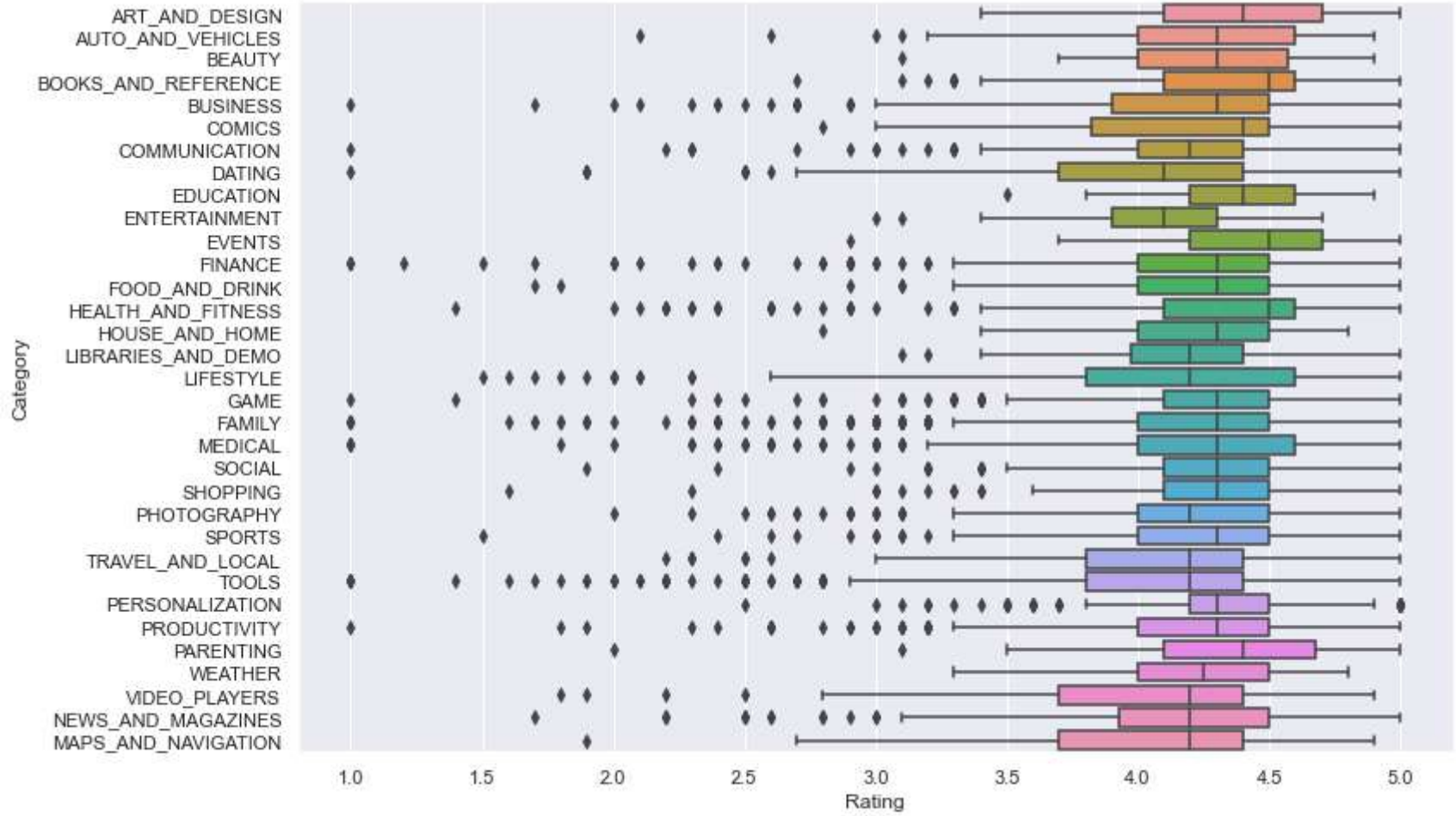
   Which genre has the best ratings?

In [137…

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

Out[137…

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



each and every category has best ratings when compare to others.

1. Data preprocessing

For the steps below, create a copy of the dataframe to make all the edits. Name it inp1.

In [138…

```
inp1 = data
```

In [139…

```
inp1.head()
```

Out[139…

	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

8.

1. 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 [140…

```
inp1.skew()
```

Out[140…

```
C:\Users\satis\AppData\Local\Temp/ipykernel_40384/3545313420.py:1: FutureWarning: Dropping of nuisance columns in DataFra
me 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()
Rating      -1.749753
Reviews      4.576494
Size         1.655917
Installs     1.543697
Price       18.074542
dtype: float64
```

In [141…
```
reviewskew = np.log1p(inp1['Reviews'])
inp1['Reviews'] =reviewskew
```

In [142…
```
reviewskew.skew()
```

Out[142…
```
-0.20039949659264134
```

In [143…
```
Installsskew = np.log1p(inp1['Installs'])
inp1['Installs'] =Installsskew
```

In [144…
```
Installsskew.skew()
```

Out[144…
```
-0.5097286542754812
```

In [145…
```
inp1.head()
```

Out[145…

	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	9.210440	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	13.122365	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	15.424949	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	6.875232	2800.0	11.512935	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	10.819798	Free	0	Everyone	Art & Design	March 26, 2017	1.0	2.3 and up

8.

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

In [146...
```
inp1.drop(["Last Updated","Current Ver", "Android Ver", "App", "Type"],axis=1,inplace=True)
```

In [147...
```
inp1.head()
```

Out[147...

	Category	Rating	Reviews	Size	Installs	Price	Content Rating	Genres
0	ART_AND_DESIGN	4.1	5.075174	19000.0	9.210440	0	Everyone	Art & Design
1	ART_AND_DESIGN	3.9	6.875232	14000.0	13.122365	0	Everyone	Art & Design;Pretend Play
2	ART_AND_DESIGN	4.7	11.379520	8700.0	15.424949	0	Everyone	Art & Design
4	ART_AND_DESIGN	4.3	6.875232	2800.0	11.512935	0	Everyone	Art & Design;Creativity
5	ART_AND_DESIGN	4.4	5.123964	5600.0	10.819798	0	Everyone	Art & Design

In [148...
```
inp1.shape
```

Out[148...
```
(8496, 8)
```

8.

1. 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.

In [149…
```
inp2 = inp1
```

In [150…
```
inp2.head()
```

Out[150…

	Category	Rating	Reviews	Size	Installs	Price	Content Rating	Genres
0	ART_AND_DESIGN	4.1	5.075174	19000.0	9.210440	0	Everyone	Art & Design
1	ART_AND_DESIGN	3.9	6.875232	14000.0	13.122365	0	Everyone	Art & Design;Pretend Play
2	ART_AND_DESIGN	4.7	11.379520	8700.0	15.424949	0	Everyone	Art & Design
4	ART_AND_DESIGN	4.3	6.875232	2800.0	11.512935	0	Everyone	Art & Design;Creativity
5	ART_AND_DESIGN	4.4	5.123964	5600.0	10.819798	0	Everyone	Art & Design

Now will apply Dummy encoding on column "Category"

In [151…
```
inp2.Category.unique()              # get unique values to the column "Category"
```

Out[151…
```
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 [152…
```
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[152...

	Rating	Reviews	Size	Installs	Price	Content Rating	Genres	Category_ART_AND_DESIGN	Category_AUTO_AND_VEHICLES	Categor
0	4.1	5.075174	19000.0	9.210440	0	Everyone	Art & Design	1	0	
1	3.9	6.875232	14000.0	13.122365	0	Everyone	Art & Design;Pretend Play	1	0	
2	4.7	11.379520	8700.0	15.424949	0	Everyone	Art & Design	1	0	
4	4.3	6.875232	2800.0	11.512935	0	Everyone	Art & Design;Creativity	1	0	
5	4.4	5.123964	5600.0	10.819798	0	Everyone	Art & Design	1	0	

5 rows × 40 columns

In [154...

```
inp2.shape
```

Out[154...

```
(8496, 40)
```

Now will apply Dummy encoding on column "Genres"

In [155...

```
inp2.Genres.unique()              # get unique values to the column "Genres"
```

Out[155...

```
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)
```

Here, In Genres there are so many categories. So, we will reduce some categories which have few samples under them and put them in common category as "Other"

In [156…

```
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 [157…
```
inp2["Genres"].unique()
```

Out[157…
```
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 [158…
```
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)
inp2.head()
```

Out[158…

	Rating	Reviews	Size	Installs	Price	Content Rating	Category_ART_AND_DESIGN	Category_AUTO_AND_VEHICLES	Category_BEAUTY	Categ
0	4.1	5.075174	19000.0	9.210440	0	Everyone	1	0	0	
1	3.9	6.875232	14000.0	13.122365	0	Everyone	1	0	0	
2	4.7	11.379520	8700.0	15.424949	0	Everyone	1	0	0	
4	4.3	6.875232	2800.0	11.512935	0	Everyone	1	0	0	
5	4.4	5.123964	5600.0	10.819798	0	Everyone	1	0	0	

5 rows × 91 columns

In [164…

```
inp2.shape
```

Out[164...
```
(8496, 91)
```

Now will apply Dummy encoding on column "Content Rating"

In [167...
```
inp2['Content Rating'].unique()              # get unique values to the column "Content Rating"
```

Out[167...
```
array(['Everyone', 'Teen', 'Everyone 10+', 'Mature 17+',
       'Adults only 18+', 'Unrated'], dtype=object)
```

In [171...
```
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[171...

	Rating	Reviews	Size	Installs	Price	Category_ART_AND_DESIGN	Category_AUTO_AND_VEHICLES	Category_BEAUTY	Category_BOOK
0	4.1	5.075174	19000.0	9.210440	0	1	0	0	
1	3.9	6.875232	14000.0	13.122365	0	1	0	0	
2	4.7	11.379520	8700.0	15.424949	0	1	0	0	
4	4.3	6.875232	2800.0	11.512935	0	1	0	0	
5	4.4	5.123964	5600.0	10.819798	0	1	0	0	

5 rows × 96 columns

In [172...
```
inp2.shape
```

Out[172...
```
(8496, 96)
```

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

2. Separate the dataframes into X_train, y_train, X_test, and y_test.

In [174...
```
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
```

In [176...
```
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

   Use linear regression as the technique

   Report the R2 on the train set

In [178...
```
reg_all = LR()
reg_all.fit(Xtrain,ytrain)
```

Out[178...
```
LinearRegression()
```

In [181...
```
R2_train = round (reg_all.score(Xtrain,ytrain),3)
print("The R2 value of Training Set is :{}".format(R2_train))
```

```
The R2 value of Training Set is :0.157
```

1. Make predictions on test set and report R2.

In [183...
```
R2_test = round (reg_all.score(xtest,ytest),3)
print("The R2 value of Training Set is :{}".format(R2_test))
```

```
The R2 value of Training Set is :0.141
```