```
In [346]: import os
          import pandas as pd
          import seaborn as sns
          import numpy as np
          import matplotlib.pyplot as plt
          import warnings
          warnings.filterwarnings('ignore')
```

```
In [347]: os.getcwd()
```

```
Out[347]: '/home/labsuser/Project/Dataset'
```

```
In [348]: os.chdir('/home/labsuser/Project/Dataset')
```

```
In [349]: os.getcwd()
```

```
Out[349]: '/home/labsuser/Project/Dataset'
```

In [350]:

```
#1) Load the data file using pandas
App_Rating = pd.read_csv('googleplaystore.csv')
App_Rating
```

Out[350]:

	App	Category	Rating	Reviews	Size	Installs	Type	Pric
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	
...	...	...	...	...	...	...	...	
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	
10838	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	

10841 rows × 13 columns

In [351]:

```
#2) Check for null values in the data. Get the number of null values for each column.
App_Rating.isnull().sum(axis=0)
```

Out[351]:

```
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 [352]:

```
#3)Drop records with nulls in any of the columns.
App_Rating_Final = App_Rating.dropna()
App_Rating_Final
```

Out[352]:

	App	Category	Rating	Reviews	Size	Installs	Type	Pric
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	
...	...	...	...	...	...	...	...	
10834	FR Calculator	FAMILY	4.0	7	2.6M	500+	Free	
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	

9360 rows × 13 columns

```
In [353]: App_Rating_Final.isnull().any()
```

```
Out[353]: 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 [354]:

```
#4)1)1)Size column has sizes in Kb as well as Mb. To analyze, you’ll need to convert these to numeric:
#Extract the numeric value from the column
#Multiply the value by 1,000, if size is mentioned in Mb
App_Rating_Final['Size'] = App_Rating_Final['Size'].apply(lambda x: np.float(x.replace('M', '')) * 1e3 if type(x) != float and 'M' in x else x)
App_Rating_Final
```

Out[354]:

	App	Category	Rating	Reviews	Size	Installs	Type	Pric
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000	10,000+	Free	
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000	500,000+	Free	
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700	5,000,000+	Free	
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000	50,000,000+	Free	
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800	100,000+	Free	
...	...	...	...	...	...	...	...	
10834	FR Calculator	FAMILY	4.0	7	2600	500+	Free	
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53000	5,000+	Free	
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600	100+	Free	
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000	10,000,000+	Free	

9360 rows × 13 columns

In [355]:
```
App_Rating_Final['Size'] = App_Rating_Final['Size'].apply(lambda x: np.NaN
if x == 'Varies with device' else x)
App_Rating_Final
```

Out[355]:

	App	Category	Rating	Reviews	Size	Installs	Type	Pric
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000	10,000+	Free	
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000	500,000+	Free	
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700	5,000,000+	Free	
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000	50,000,000+	Free	
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800	100,000+	Free	
...	...	...	...	...	...	...	...	.
10834	FR Calculator	FAMILY	4.0	7	2600	500+	Free	
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53000	5,000+	Free	
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600	100+	Free	
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	NaN	1,000+	Free	
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000	10,000,000+	Free	

9360 rows × 13 columns

In [356]:

```
App_Rating_Final['Size'] = App_Rating_Final['Size'].apply(lambda x: np.float(x.replace('k', '')) if type(x) != float and 'k' in x else x)
App_Rating_Final
```

Out[356]:

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

9360 rows × 13 columns

```
In [357]: App_Rating_Final['Size'].astype(np.float)
```

```
Out[357]: 0        19000.0
          1        14000.0
          2         8700.0
          3        25000.0
          4         2800.0
                     ...
          10834     2600.0
          10836    53000.0
          10837     3600.0
          10839        NaN
          10840    19000.0
          Name: Size, Length: 9360, dtype: float64
```

In [358]:

```
#4)2)Reviews is a numeric field that is loaded as a string field. Convert i
t to numeric (int/float).
App_Rating_Final['Reviews'] = App_Rating_Final['Reviews'].astype(int)
App_Rating_Final
```

Out[358]:

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

9360 rows × 13 columns

In [359]:

```
#4)3)Installs field is currently stored as string and has values like 1,000,000+.
#Treat 1,000,000+ as 1,000,000, remove ‘+’, ‘,’ from the field, convert it to integer
App_Rating_Final["Installs"] = [float(i.replace('+','').replace(',', '')) if '+' in i or ',' in i else float(0) for i in App_Rating_Final["Installs"]]
```

In [360]:

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

Out[360]:

	App	Category	Rating	Reviews	Size	Installs	Type	Pric
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10000	Free	(
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500000	Free	(
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5000000	Free	(
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50000000	Free	(
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100000	Free	(
...	...	...	...	...	...	...	...	..
10834	FR Calculator	FAMILY	4.0	7	2600.0	500	Free	(
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5000	Free	(
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100	Free	(
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	NaN	1000	Free	(
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10000000	Free	(

9360 rows × 13 columns

In [361]:

```python
#4)4)Price field is a string and has $ symbol. Remove ‘$’ sign, and convert it to numeric.
App_Rating_Final["Price"] = [float(i.replace('$','')) if '$' in i else float(0) for i in App_Rating_Final["Price"]]
App_Rating_Final
```

Out[361]:

	App	Category	Rating	Reviews	Size	Installs	Type	Pric
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10000	Free	0.(
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500000	Free	0.(
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5000000	Free	0.(
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50000000	Free	0.(
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100000	Free	0.(
...	...	...	...	...	...	...	...	..
10834	FR Calculator	FAMILY	4.0	7	2600.0	500	Free	0.(
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5000	Free	0.(
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100	Free	0.(
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	NaN	1000	Free	0.(
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10000000	Free	0.(

9360 rows × 13 columns

In [362]:

```
#4)5Sanity checks:Average rating should be between 1 and 5 as only these va
lues are allowed on the play store.
#5)1)Drop the rows that have a value outside this range.
App_Rating_Final_Rating = App_Rating_Final [App_Rating_Final ['Rating'].bet
ween(1, 5)]
App_Rating_Final = App_Rating_Final.reset_index(drop=True)
App_Rating_Final
```

Out[362]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10000	Free	0.0
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500000	Free	0.0
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5000000	Free	0.0
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50000000	Free	0.0
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100000	Free	0.0
...	...	...	...	...	...	...	...	...
9355	FR Calculator	FAMILY	4.0	7	2600.0	500	Free	0.0
9356	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5000	Free	0.0
9357	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100	Free	0.0
9358	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	NaN	1000	Free	0.0
9359	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10000000	Free	0.0

9360 rows × 13 columns

In [363]:

```
#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.
App_Rating_Final_Sanity = App_Rating_Final[(App_Rating_Final['Reviews'] > App_Rating_Final['Installs'])]
App_Rating_Final = App_Rating_Final.drop(App_Rating_Final_Sanity.index[range(0,7)])
App_Rating_Final
```

Out[363]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10000	Free	0.0
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500000	Free	0.0
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5000000	Free	0.0
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50000000	Free	0.0
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100000	Free	0.0
...	...	...	...	...	...	...	...	...
9355	FR Calculator	FAMILY	4.0	7	2600.0	500	Free	0.0
9356	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5000	Free	0.0
9357	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100	Free	0.0
9358	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	NaN	1000	Free	0.0
9359	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10000000	Free	0.0

9353 rows × 13 columns

In [364]:

```
#5)3)For free apps (type = “Free”), the price should not be >0. Drop any such rows.
def type_cat(Type):
    if Type == 'Free':
        return 0
    else:
        return 1
    App_Rating_Final['Type'] = App_Rating_Final['Type']
    App_Rating_Final = App_Rating_Final.drop(App_Rating_Final[(App_Rating_Final['Reviews']) > (App_Rating_Final['Installs'])].index)
App_Rating_Final
```

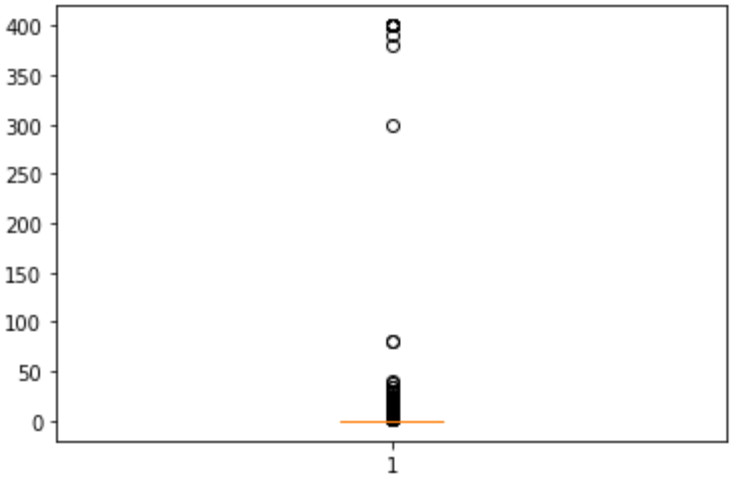
Out[364]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10000	Free	0.0
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500000	Free	0.0
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5000000	Free	0.0
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50000000	Free	0.0
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100000	Free	0.0
...	...	...	...	...	...	...	...	...
9355	FR Calculator	FAMILY	4.0	7	2600.0	500	Free	0.0
9356	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5000	Free	0.0
9357	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100	Free	0.0
9358	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	NaN	1000	Free	0.0
9359	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10000000	Free	0.0

9353 rows × 13 columns

```
In [365]: #5. Performing univariate analysis:
          #Boxplot for Price
          plt.boxplot(App_Rating_Final.Price)
```

```
Out[365]: {'whiskers': [<matplotlib.lines.Line2D at 0x7f606f22c3d0>,
            <matplotlib.lines.Line2D at 0x7f606f22c690>],
           'caps': [<matplotlib.lines.Line2D at 0x7f606f22cb10>,
            <matplotlib.lines.Line2D at 0x7f606f22ce50>],
           'boxes': [<matplotlib.lines.Line2D at 0x7f606f22c090>],
           'medians': [<matplotlib.lines.Line2D at 0x7f606f1bb0d0>],
           'fliers': [<matplotlib.lines.Line2D at 0x7f606f1bb550>],
           'means': []}
```

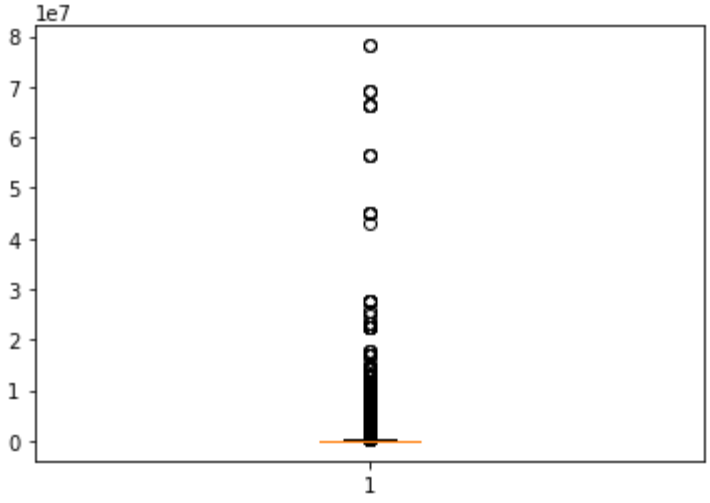


In [366]:

```
#5. Performing univariate analysis:
#Boxplot for Reviews
plt.boxplot(App_Rating_Final.Reviews)
```

Out[366]:

```
{'whiskers': [<matplotlib.lines.Line2D at 0x7f606f1b0a10>,
  <matplotlib.lines.Line2D at 0x7f606f1b0d50>],
 'caps': [<matplotlib.lines.Line2D at 0x7f606f1b9210>,
  <matplotlib.lines.Line2D at 0x7f606f1b9550>],
 'boxes': [<matplotlib.lines.Line2D at 0x7f606f1b0810>],
 'medians': [<matplotlib.lines.Line2D at 0x7f606f1b9790>],
 'fliers': [<matplotlib.lines.Line2D at 0x7f606f1b9c10>],
 'means': []}
```

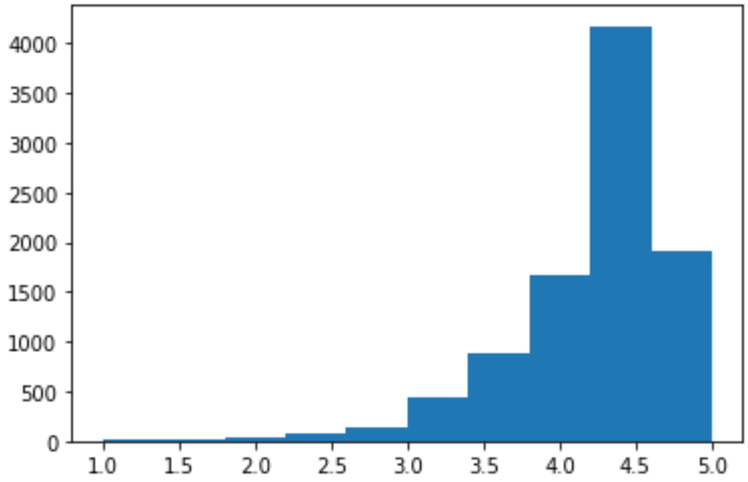


In [367]:

```
#5. Histogram for Rating:How are the ratings distributed? Is it more toward
higher ratings(Answer:Yes)?
plt.hist(App_Rating_Final.Rating)
```

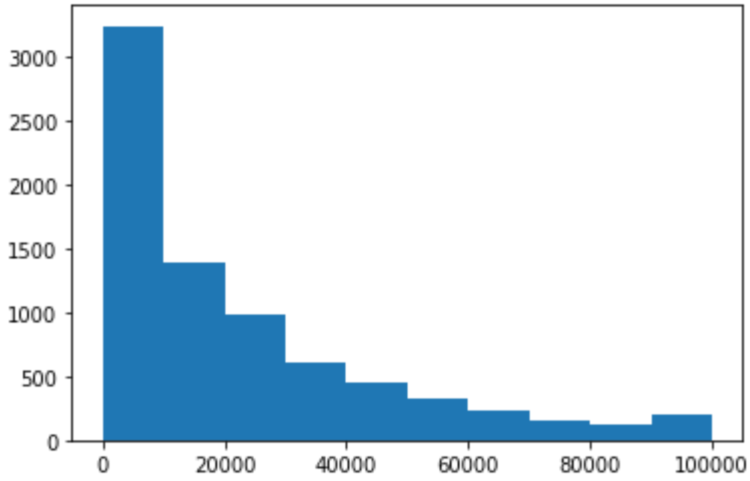
Out[367]:

```
(array([  17.,   18.,   41.,   74.,  137.,  445.,  879., 1660., 4172.,
        1910.]),
 array([1. , 1.4, 1.8, 2.2, 2.6, 3. , 3.4, 3.8, 4.2, 4.6, 5. ]),
 <BarContainer object of 10 artists>)
```



In [368]:
```
#Histogram for Size
plt.hist(App_Rating_Final.Size)
```

Out[368]:
```
(array([3245., 1398.,  991.,  606.,  449.,  325.,  226.,  161.,  117.,
         199.]),
 array([8.500000e+00, 1.000765e+04, 2.000680e+04, 3.000595e+04,
        4.000510e+04, 5.000425e+04, 6.000340e+04, 7.000255e+04,
        8.000170e+04, 9.000085e+04, 1.000000e+05]),
 <BarContainer object of 10 artists>)
```



In [369]:
```
#Note down your observations for the plots made above. Which of these seem
to have outliers?
#Boxplots for price and Reviews clearly mentions that there are outliers
```

In [370]:

```
#6.1) Outlier treatment:
#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 an
d suspicious!
#Check out the records with very high price
#Is 200 indeed a high price?
#Drop these as most seem to be junk apps
App_High_Price = App_Rating_Final[(App_Rating_Final['Price'] > 200)]
App_Rating_Final= App_Rating_Final.drop(App_Rating_Final [(App_Rating_Final
['Price']) > 200 ].index)
App_Rating_Final
```

Out[370]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10000	Free	0.0
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500000	Free	0.0
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5000000	Free	0.0
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50000000	Free	0.0
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100000	Free	0.0
...	...	...	...	...	...	...	...	...
9355	FR Calculator	FAMILY	4.0	7	2600.0	500	Free	0.0
9356	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5000	Free	0.0
9357	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100	Free	0.0
9358	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	NaN	1000	Free	0.0

In [371]:

```
#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. Dro
p records having more than 2 million reviews
App_High_Review = App_Rating_Final[(App_Rating_Final['Reviews'] > 2000000)]
App_Rating_Final= App_Rating_Final.drop(App_Rating_Final [(App_Rating_Final
['Reviews']) > 2000000 ].index)
App_Rating_Final
```

Out[371]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10000	Free	0.0
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500000	Free	0.0
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5000000	Free	0.0
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50000000	Free	0.0
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100000	Free	0.0
...	...	...	...	...	...	...	...	...
9355	FR Calculator	FAMILY	4.0	7	2600.0	500	Free	0.0
9356	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5000	Free	0.0
9357	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100	Free	0.0
9358	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	NaN	1000	Free	0.0
9359	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10000000	Free	0.0

8885 rows × 13 columns

In [372]:

```
#6)3)Installs:  There seems to be some outliers in this field too. Apps having very high number of installs should be dropped from the analysis.
#Find out the different percentiles – 10, 25, 50, 70, 90, 95, 99
#Decide a threshold as cutoff for outlier and drop records having values more than that
App_Rating_Final_Quantiles = App_Rating_Final.Installs.quantile([0.1, 0.25, 0.5, 0.70, 0.9, 0.95, 0.99])
App_Rating_Final_Quantiles
```

Out[372]:

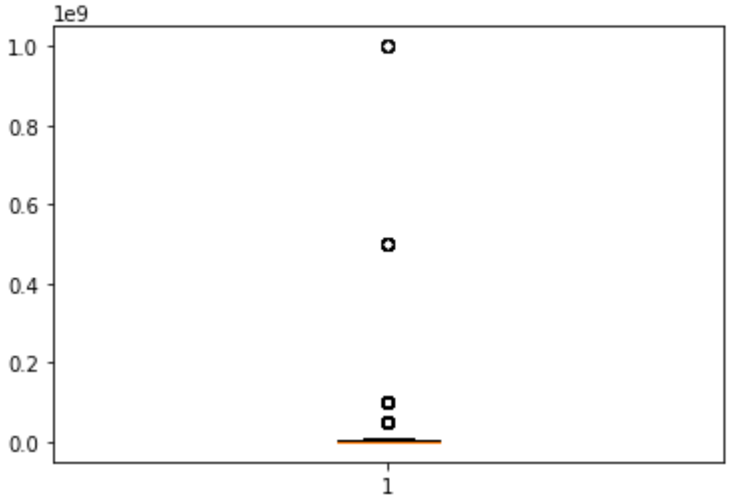
```
0.10          1000.0
0.25         10000.0
0.50        500000.0
0.70       1000000.0
0.90      10000000.0
0.95      10000000.0
0.99     100000000.0
Name: Installs, dtype: float64
```

In [373]:

```
plt.boxplot(App_Rating_Final.Installs)
```

Out[373]:

```
{'whiskers': [<matplotlib.lines.Line2D at 0x7f606ef7cc10>,
  <matplotlib.lines.Line2D at 0x7f606ef7cf50>],
 'caps': [<matplotlib.lines.Line2D at 0x7f606ef81410>,
  <matplotlib.lines.Line2D at 0x7f606ef81750>],
 'boxes': [<matplotlib.lines.Line2D at 0x7f606ef7c910>],
 'medians': [<matplotlib.lines.Line2D at 0x7f606ef81990>],
 'fliers': [<matplotlib.lines.Line2D at 0x7f606ef81e10>],
 'means': []}
```



In [374]:

```
#6)3)2)Decide a threshold as cutoff for outlier and drop records having values more than that
App_High_Intalls = App_Rating_Final[(App_Rating_Final['Installs'] >=100000000)]
App_Rating_Final= App_Rating_Final.drop(App_Rating_Final [(App_Rating_Final['Installs']) >=100000000 ].index)
App_Rating_Final
```

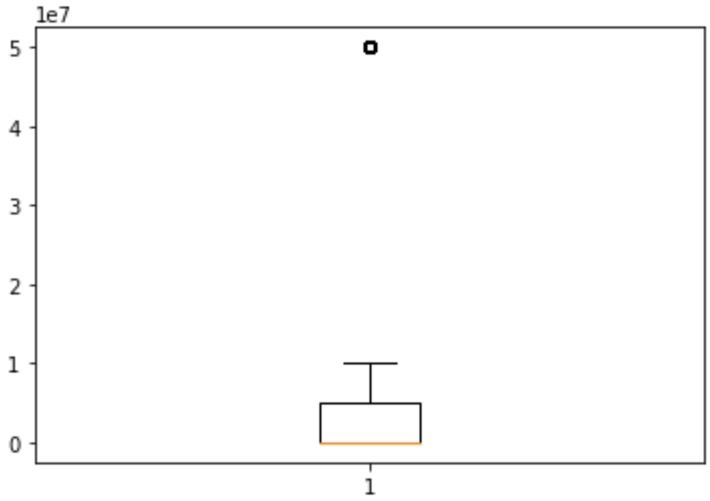
Out[374]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10000	Free	0.0
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500000	Free	0.0
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5000000	Free	0.0
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50000000	Free	0.0
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100000	Free	0.0
...	...	...	...	...	...	...	...	...
9355	FR Calculator	FAMILY	4.0	7	2600.0	500	Free	0.0
9356	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5000	Free	0.0
9357	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100	Free	0.0
9358	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	NaN	1000	Free	0.0
9359	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10000000	Free	0.0

8743 rows × 13 columns

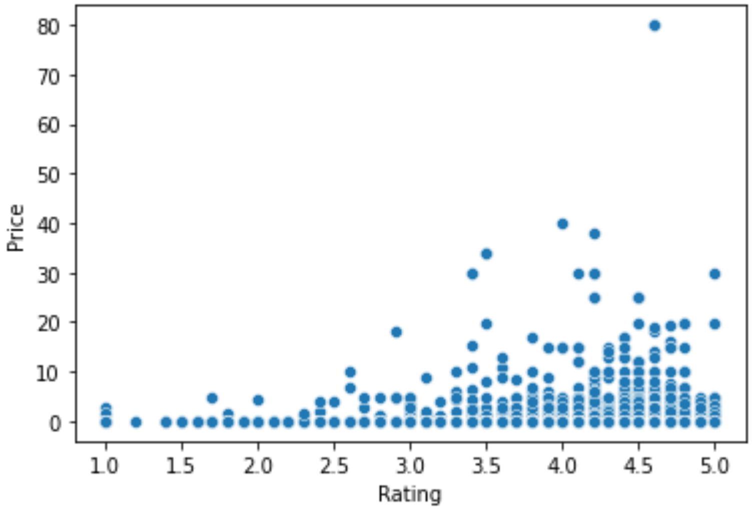
In [375]:
```
plt.boxplot(App_Rating_Final.Installs)
```

Out[375]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x7f606ef69ad0>,
  <matplotlib.lines.Line2D at 0x7f606ef69e10>],
 'caps': [<matplotlib.lines.Line2D at 0x7f606ef702d0>,
  <matplotlib.lines.Line2D at 0x7f606ef70610>],
 'boxes': [<matplotlib.lines.Line2D at 0x7f606ef697d0>],
 'medians': [<matplotlib.lines.Line2D at 0x7f606ef70850>],
 'fliers': [<matplotlib.lines.Line2D at 0x7f606ef70cd0>],
 'means': []}
```



In [376]:
```
#7)1)1)Make scatter plot/joinplot for Rating vs. Price
#What pattern do you observe? Does rating increase with price?
#(Answer:Yes, rating does seem to increase with price)
sns.scatterplot(data=App_Rating_Final, x="Rating", y="Price")
```

Out[376]:
```
<AxesSubplot:xlabel='Rating', ylabel='Price'>
```

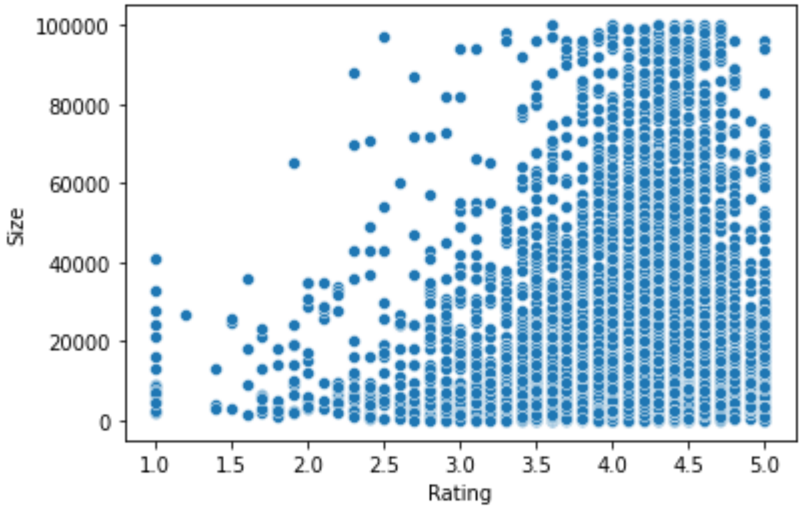


In [377]:

```
#7)2)Make scatter plot/joinplot for Rating vs. Size
#Are heavier apps rated better?
#(Answer:Yes)
sns.scatterplot(data=App_Rating_Final, x="Rating", y="Size")
```

Out[377]:

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

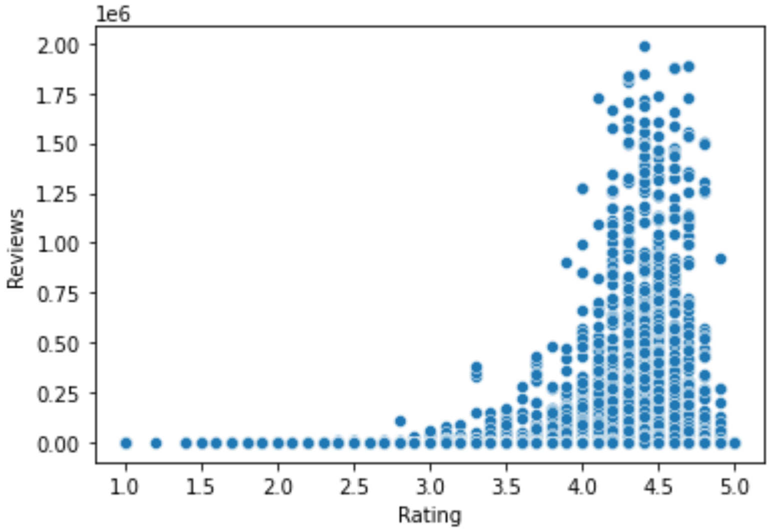


In [378]:

```
#7)3)Make scatter plot/joinplot for Rating vs. Reviews
#Does more review mean a better rating always?(Answer:No, not always, bad r
eviews also contribute to the ratings)
sns.scatterplot(data=App_Rating_Final, x="Rating", y="Reviews")
```

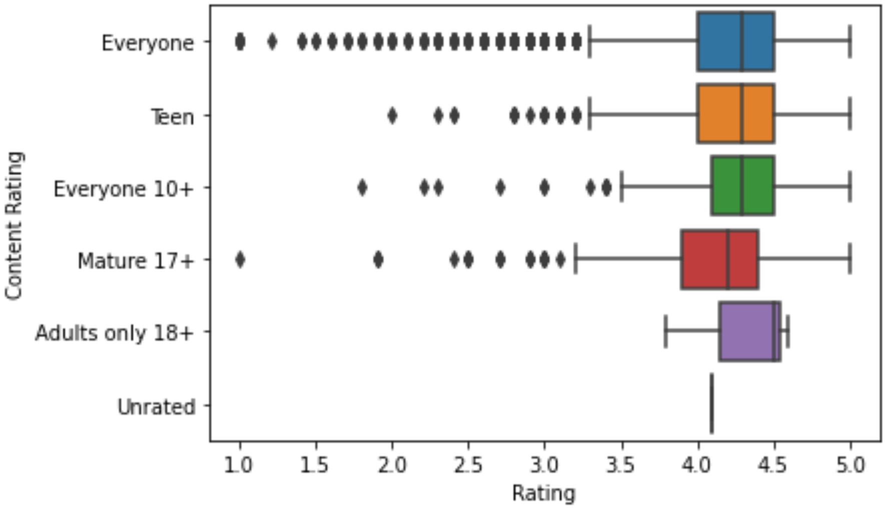
Out[378]:

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



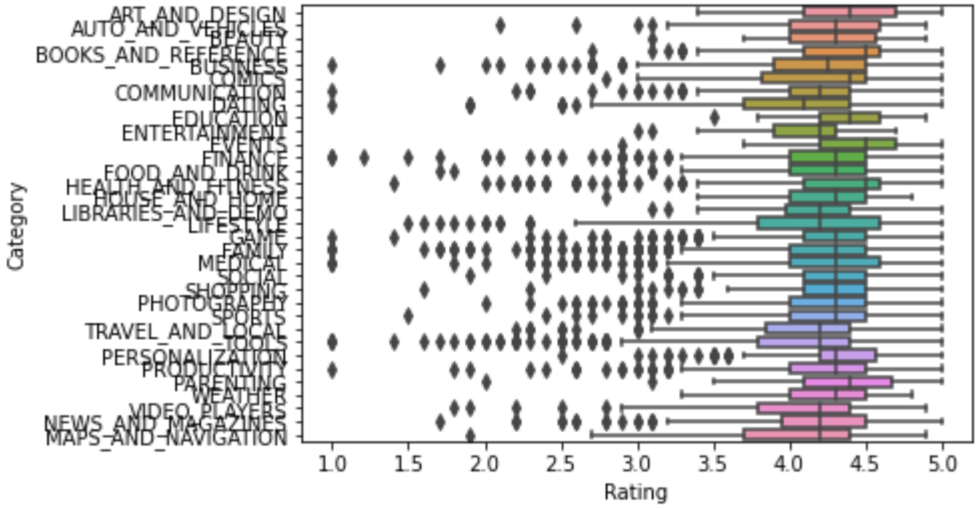
In [379]:
```
#7)4)Make boxplot for Rating vs. Content Rating
#Is there any difference in the ratings? Are some types liked better?
#(Answer:Yes, apps that can be used by Everyone is liked more)
sns.boxplot(data=App_Rating_Final, x="Rating", y="Content Rating")
```

Out[379]:
```
<AxesSubplot:xlabel='Rating', ylabel='Content Rating'>
```



In [380]:
```
#7)5)Make boxplot for Ratings vs. Category
#Which genre has the best ratings?
#(Answer: Maps_and_Navigation)
sns.boxplot(data=App_Rating_Final, x="Rating", y="Category")
```

Out[380]:
```
<AxesSubplot:xlabel='Rating', ylabel='Category'>
```



In [381]:

```
#8)1)Data preprocessing
#Reviews and Install have some values that are still relatively very high.
Before building a linear regression model, you need to reduce the skew. App
ly log transformation (np.log1p) to Reviews and Installs.
inp1 = App_Rating_Final
np.log1p(inp1.Reviews)
np.log1p(inp1.Installs)
inp1
```

Out[381]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10000	Free	0.0
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500000	Free	0.0
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5000000	Free	0.0
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50000000	Free	0.0
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100000	Free	0.0
...	...	...	...	...	...	...	...	...
9355	FR Calculator	FAMILY	4.0	7	2600.0	500	Free	0.0
9356	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5000	Free	0.0
9357	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100	Free	0.0
9358	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	NaN	1000	Free	0.0
9359	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10000000	Free	0.0

8743 rows × 13 columns

In [382]:
```
inp1 = inp1.drop(labels = ['App', 'Last Updated', 'Current Ver', 'Android V
er','Type'],axis = 1)
inp1
```

Out[382]:

	Category	Rating	Reviews	Size	Installs	Price	Content Rating	
0	ART_AND_DESIGN	4.1	159	19000.0	10000	0.0	Everyone	Art
1	ART_AND_DESIGN	3.9	967	14000.0	500000	0.0	Everyone	Desig
2	ART_AND_DESIGN	4.7	87510	8700.0	5000000	0.0	Everyone	Art
3	ART_AND_DESIGN	4.5	215644	25000.0	50000000	0.0	Teen	Art
4	ART_AND_DESIGN	4.3	967	2800.0	100000	0.0	Everyone	Design;
...	...	...	...	...	...	...	...	
9355	FAMILY	4.0	7	2600.0	500	0.0	Everyone	I
9356	FAMILY	4.5	38	53000.0	5000	0.0	Everyone	I
9357	FAMILY	5.0	4	3600.0	100	0.0	Everyone	I
9358	BOOKS_AND_REFERENCE	4.5	114	NaN	1000	0.0	Mature 17+	F
9359	LIFESTYLE	4.5	398307	19000.0	10000000	0.0	Everyone	

8743 rows × 8 columns

In [383]:
```
#8)3)Get dummy columns for Category, Genres, and Content Rating.
#get unique values in Column "Category"
inp2 = inp1
inp2.Category.unique()
```

Out[383]:
```
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 [384]:

```
inp2.Category = pd.Categorical(inp1.Category)

x = inp2[['Category']]
del inp2['Category']

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

Out[384]:

	Rating	Reviews	Size	Installs	Price	Content Rating	Genres	Category_ART_AND_
0	4.1	159	19000.0	10000	0.0	Everyone	Art & Design	
1	3.9	967	14000.0	500000	0.0	Everyone	Art & Design;Pretend Play	
2	4.7	87510	8700.0	5000000	0.0	Everyone	Art & Design	
3	4.5	215644	25000.0	50000000	0.0	Teen	Art & Design	
4	4.3	967	2800.0	100000	0.0	Everyone	Art & Design;Creativity	

5 rows × 40 columns

In [385]:

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

Out[385]:

```
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;Music & Video', 'Entertainment;Brain Games',
       'Entertainment;Creativity', 'Events', 'Finance', 'Food & Drink',
       'Health & Fitness', 'House & Home', 'Libraries & Demo',
       'Lifestyle', 'Lifestyle;Pretend Play', 'Card', 'Casual',
       'Casual;Pretend Play', 'Puzzle', 'Action', 'Arcade', 'Music',
       'Word', 'Racing', 'Casual;Creativity', 'Sports', 'Simulation',
       'Board', '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', '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',
       '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 [386]:

```
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 [387]:
```
inp2["Genres"].unique()
```

Out[387]:
```
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', 'Casual;Pretend Play', 'Puzzle',
       'Action', 'Arcade', 'Music', 'Word', 'Racing', 'Sports',
       'Simulation', 'Board', 'Role Playing', 'Adventure', 'Strategy',
       'Trivia', 'Educational;Education', 'Racing;Action & Adventure',
       'Medical', 'Social', 'Shopping', 'Photography', 'Travel & Local',
       'Tools', 'Personalization', 'Productivity', 'Parenting', 'Weather',
       'Video Players & Editors', 'News & Magazines', 'Maps & Navigation',
       'Educational', 'Casino'], dtype=object)
```

In [388]:
```
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 [389]:
```
inp2.head()
```

Out[389]:

	Rating	Reviews	Size	Installs	Price	Content Rating	Category_ART_AND_DESIGN	Categor
0	4.1	159	19000.0	10000	0.0	Everyone	1	
1	3.9	967	14000.0	500000	0.0	Everyone	1	
2	4.7	87510	8700.0	5000000	0.0	Everyone	1	
3	4.5	215644	25000.0	50000000	0.0	Teen	1	
4	4.3	967	2800.0	100000	0.0	Everyone	1	

5 rows × 93 columns

In [390]:
```
#get unique values in Column "Content Rating"
inp2["Content Rating"].unique()
```

Out[390]:
```
array(['Everyone', 'Teen', 'Everyone 10+', 'Mature 17+',
       'Adults only 18+', 'Unrated'], dtype=object)
```

In [391]:

```
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[391]:

	Rating	Reviews	Size	Installs	Price	Category_ART_AND_DESIGN	Category_AUTO_A
0	4.1	159	19000.0	10000	0.0	1	
1	3.9	967	14000.0	500000	0.0	1	
2	4.7	87510	8700.0	5000000	0.0	1	
3	4.5	215644	25000.0	50000000	0.0	1	
4	4.3	967	2800.0	100000	0.0	1	

5 rows × 98 columns

In [392]:

```
inp2
```

Out[392]:

	Rating	Reviews	Size	Installs	Price	Category_ART_AND_DESIGN	Category_AUTC
0	4.1	159	19000.0	10000	0.0	1	
1	3.9	967	14000.0	500000	0.0	1	
2	4.7	87510	8700.0	5000000	0.0	1	
3	4.5	215644	25000.0	50000000	0.0	1	
4	4.3	967	2800.0	100000	0.0	1	
...	...	...	...	...	...	...	
9355	4.0	7	2600.0	500	0.0	0	
9356	4.5	38	53000.0	5000	0.0	0	
9357	5.0	4	3600.0	100	0.0	0	
9358	4.5	114	NaN	1000	0.0	0	
9359	4.5	398307	19000.0	10000000	0.0	0	

8743 rows × 98 columns

In [393]:
```
inp2.dropna(inplace=True)
inp2
```

Out[393]:
```
      Rating  Reviews     Size   Installs  Price  Category_ART_AND_DESIGN  Category_AUTC

   0     4.1      159  19000.0      10000    0.0                        1
   1     3.9      967  14000.0     500000    0.0                        1
   2     4.7    87510   8700.0    5000000    0.0                        1
   3     4.5   215644  25000.0   50000000    0.0                        1
   4     4.3      967   2800.0     100000    0.0                        1
 ...     ...      ...      ...        ...    ...                      ...
9354     4.8       44    619.0       1000    0.0                        0
9355     4.0        7   2600.0        500    0.0                        0
9356     4.5       38  53000.0       5000    0.0                        0
9357     5.0        4   3600.0        100    0.0                        0
9359     4.5   398307  19000.0   10000000    0.0                        0

7423 rows × 98 columns
```

In [394]:
```
#9)Train test split  and apply 70-30 split. Name the new dataframes df_trai
n and df_test.
#10)Separate the dataframes into X_train, y_train, X_test, and y_test.
from sklearn.model_selection import train_test_split
df_train = inp2.drop('Rating',axis=1)
df_test = inp2['Rating']
X_train, X_test, y_train, y_test =  train_test_split(df_train,df_test,test_
size=.30)
```

In [395]:

```
X_train
```

Out[395]:

	Reviews	Size	Installs	Price	Category_ART_AND_DESIGN	Category_AUTO_AND_
6655	254	38000.0	10000	0.00	0	
3277	157495	6400.0	10000000	0.00	0	
4946	361734	58000.0	5000000	0.00	0	
8515	249	21000.0	10000	0.00	0	
8080	38419	100000.0	1000000	0.99	0	
...	...	...	...	...	...	
7897	99	71000.0	10000	0.00	0	
6548	8	3700.0	500	0.00	0	
8034	3606	17000.0	100000	0.00	0	
8376	8011	13000.0	500000	0.00	0	
201	6903	14000.0	1000000	0.00	0	

5196 rows × 97 columns

In [396]:

```
#11)Model building; Use linear regression as the technique;Report the R2 on the train set
from sklearn.linear_model import LinearRegression
App_Rating_LR = LinearRegression().fit(X_train,y_train)
App_Rating_LR.intercept_
```

Out[396]:

```
4.212979580663019
```

In [397]:
```
App_Rating_LR.coef_
```

Out[397]:
```
array([ 4.78221787e-07,  4.41422344e-07, -3.64280698e-09,  4.92495029e-03,
        7.63914217e-02, -2.56420487e-03,  6.09424041e-02,  6.87139859e-02,
       -1.67973375e-02,  5.27752415e-01, -5.71433146e-02, -8.82677622e-02,
        6.37536908e-02,  2.57679678e-02,  1.52924577e-01, -1.79219889e-02,
       -1.34013474e-02, -6.28209835e-02,  1.43527620e-01,  9.01552571e-03,
        1.33451434e-02,  3.41298393e-02, -3.42026941e-01, -7.83840275e-02,
        9.49829731e-03, -8.36793236e-03, -1.52549573e-01,  9.29345775e-02,
        7.37066164e-03, -6.93689166e-03,  4.69360346e-02,  7.01995713e-03,
       -6.28668471e-02, -8.00232642e-02, -1.70302042e-01, -2.35159292e-01,
        5.55096305e-02, -1.54251620e-01, -1.81647118e-01, -1.31533304e-01,
        1.59516580e-01, -2.56420487e-03,  6.09424041e-02, -4.22656266e-02,
        6.87139859e-02, -1.67973375e-02, -3.07095422e-01, -3.08625209e-02,
       -1.03907766e-01,  1.33665798e-01, -4.35974953e-01, -5.71433146e-02,
       -8.82677622e-02,  1.77420608e-01,  1.02310520e-01,  2.09129678e-01,
       -1.40831533e-01,  1.29120825e-01, -6.12933862e-02, -4.00121841e-02,
        1.52924577e-01, -1.34013474e-02, -6.28209835e-02,  9.01552571e-03,
        1.33451434e-02,  3.41298393e-02,  2.78709095e-01, -7.83840275e-02,
        9.49829731e-03, -1.95462417e-01, -8.36793236e-03,  1.43609062e-01,
        4.04821345e-01,  9.29345775e-02,  7.37066164e-03, -6.93689166e-03,
        2.04396333e-01, -1.83219786e-01,  3.06341191e-02,  2.76615486e-02,
        4.69360346e-02, -3.22676708e-02,  7.01995713e-03,  7.70746245e-02,
       -9.32061842e-02, -8.00232642e-02,  4.52667429e-02, -4.27567354e-01,
        1.00528340e-01,  5.55096305e-02,  1.93900060e-01,  2.58929796e-01,
       -8.97027474e-02, -6.68011798e-02, -5.62871637e-02, -4.61387046e-02,
        0.00000000e+00])
```

In [398]:
```
#12. Make predictions on test set and report R2.
predicted_rating = pd.DataFrame(App_Rating_LR.predict(X_test),columns=['predicted_rating'])
predicted_rating
```

Out[398]:

	predicted_rating
0	4.267659
1	4.149090
2	4.093163
3	4.088377
4	4.148062
...	...
2222	4.164187
2223	3.964606
2224	4.113524
2225	4.361657
2226	4.280286

2227 rows × 1 columns

In [399]:
```
X_test
```

Out[399]:

	Reviews	Size	Installs	Price	Category_ART_AND_DESIGN	Category_AUTO_AND_VE
6350	24557	24000.0	1000000	0.0	0	
3926	17350	12000.0	500000	0.0	0	
7100	25	7900.0	5000	0.0	0	
5598	4	1700.0	100	0.0	0	
5593	112	13000.0	1000	0.0	0	
...	...	...	...	...	...	
1410	22584	16000.0	1000000	0.0	0	
6742	1147	2700.0	100000	0.0	0	
7160	48	54000.0	5000	0.0	0	
4906	486	5900.0	100000	0.0	1	
2136	44062	54000.0	1000000	0.0	0	

2227 rows × 97 columns

In [400]:
```
y_test
```

Out[400]:
```
6350    4.7
3926    4.5
7100    4.4
5598    5.0
5593    4.4
       ... 
1410    4.3
6742    4.2
7160    4.4
4906    3.4
2136    3.9
Name: Rating, Length: 2227, dtype: float64
```

In [401]:

```
test_rating_final = pd.concat([X_test.reset_index(drop=True),y_test.reset_index(drop=True),predicted_rating],axis=1)
test_rating_final
```

Out[401]:

	Reviews	Size	Installs	Price	Category_ART_AND_DESIGN	Category_AUTO_AND_VE
0	24557	24000.0	1000000	0.0	0	
1	17350	12000.0	500000	0.0	0	
2	25	7900.0	5000	0.0	0	
3	4	1700.0	100	0.0	0	
4	112	13000.0	1000	0.0	0	
...	...	...	...	...	...	
2222	22584	16000.0	1000000	0.0	0	
2223	1147	2700.0	100000	0.0	0	
2224	48	54000.0	5000	0.0	0	
2225	486	5900.0	100000	0.0	1	
2226	44062	54000.0	1000000	0.0	0	

2227 rows × 99 columns

In [402]:

```
test_rating_final['err_pct'] = abs(test_rating_final.Rating-test_rating_final.predicted_rating)/test_rating_final.Rating
```

In [403]:

```
test_rating_final
```

Out[403]:

	Reviews	Size	Installs	Price	Category_ART_AND_DESIGN	Category_AUTO_AND_VE
0	24557	24000.0	1000000	0.0	0	
1	17350	12000.0	500000	0.0	0	
2	25	7900.0	5000	0.0	0	
3	4	1700.0	100	0.0	0	
4	112	13000.0	1000	0.0	0	
...	...	...	...	...	...	
2222	22584	16000.0	1000000	0.0	0	
2223	1147	2700.0	100000	0.0	0	
2224	48	54000.0	5000	0.0	0	
2225	486	5900.0	100000	0.0	1	
2226	44062	54000.0	1000000	0.0	0	

2227 rows × 100 columns

In [404]:

```
# error in model
test_rating_final.err_pct.mean()
```

Out[404]:

```
0.11134232515136859
```

In [405]:

```
# Accuracy in model
1- test_rating_final.err_pct.mean()
```

Out[405]:

```
0.8886576748486315
```

In [406]:

```
from sklearn.metrics import r2_score
r2_score(test_rating_final.Rating,test_rating_final.predicted_rating)
```

Out[406]:

```
0.037877576968752935
```