```
#Importing the Libraries:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

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)
```

```
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]
```

```
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
```

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

```
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
```

```
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'>
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
```

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

```
data.info()
```

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

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

```
dtypes: float64(4), int64(1), object(8)
memory usage: 1023.8+ KB
```

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

```python
data.info()
```

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

```python
data.shape
```

```
(9360, 13)
```

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

```python
data.shape
```

```
(9360, 13)
```

```python
data.shape
```

```
(9360, 13)
```

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

```python
data.shape
```

```
(9353, 13)
```

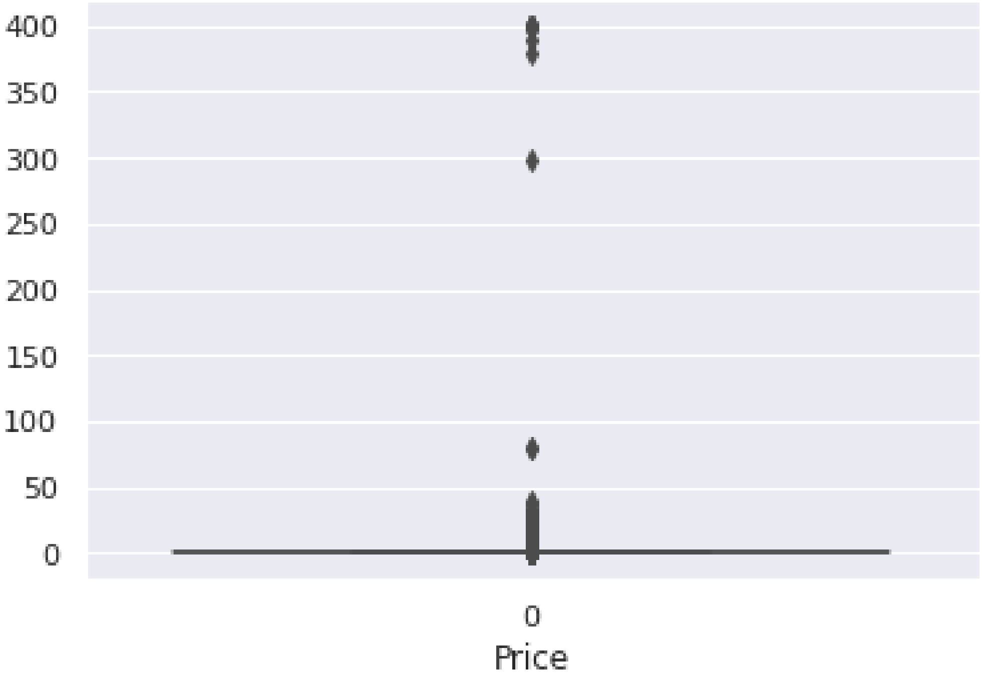
```python
data.shape
```

```
(9353, 13)
```

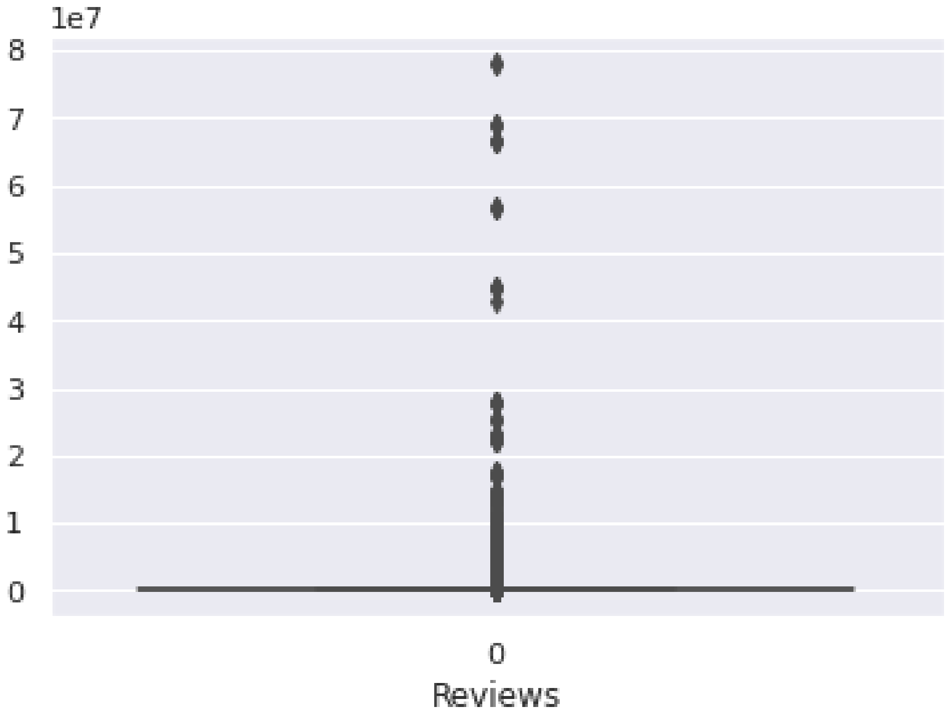
```
data.drop(data[(data['Type'] =='Free') & (data['Price'] > 0 )].index,
inplace = True)

data.shape
```

```
(9353, 13)
```

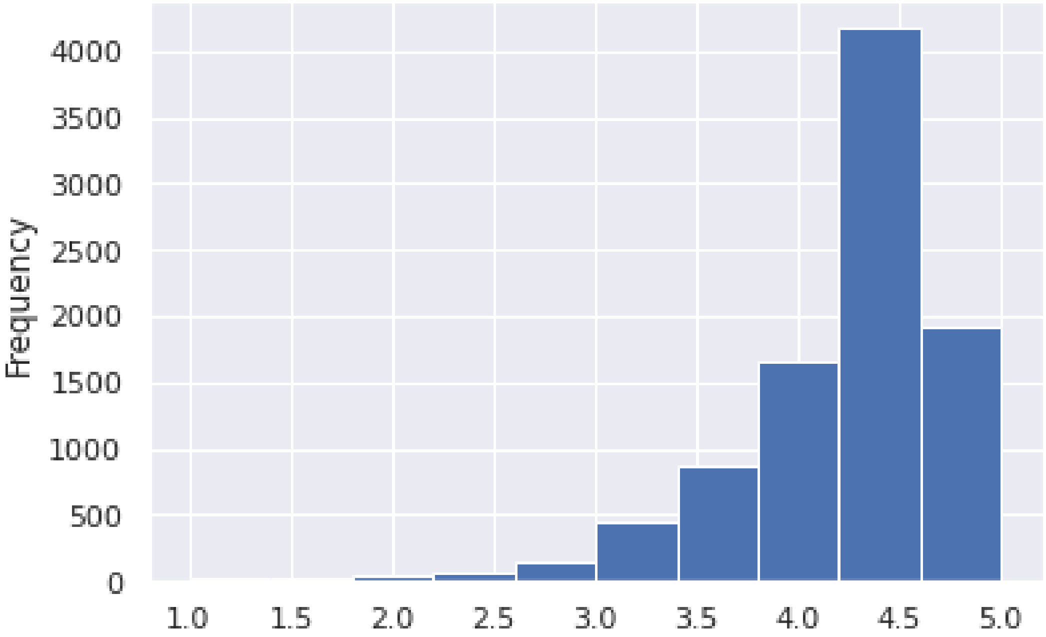
```
import seaborn as sns
sns.set(rc={'figure.figsize':(12,8)})
sns.boxplot(data.Price)
plt.xlabel('Price')
plt.show()
```



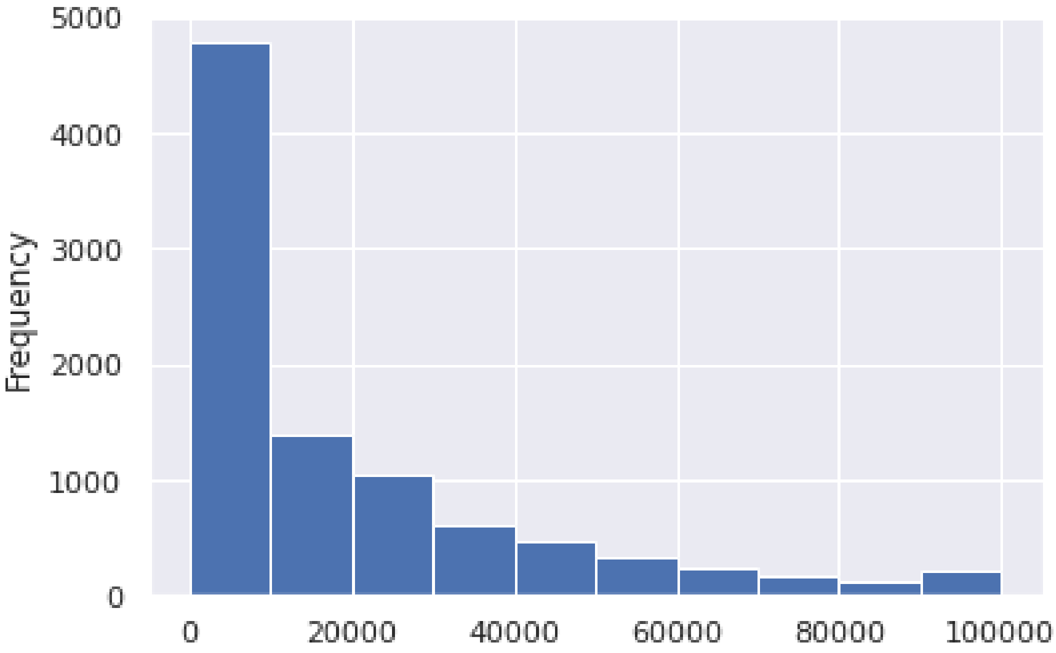
```
sns.boxplot(data['Reviews'])
plt.xlabel('Reviews')
plt.show()
```



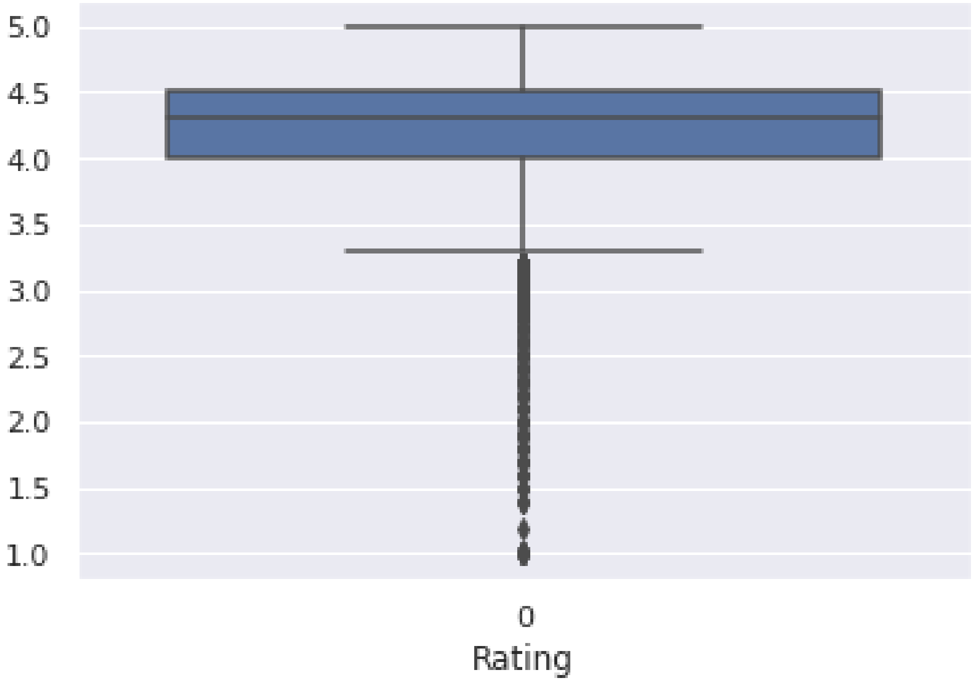
```
data.Rating.plot.hist()
plt.show()
```



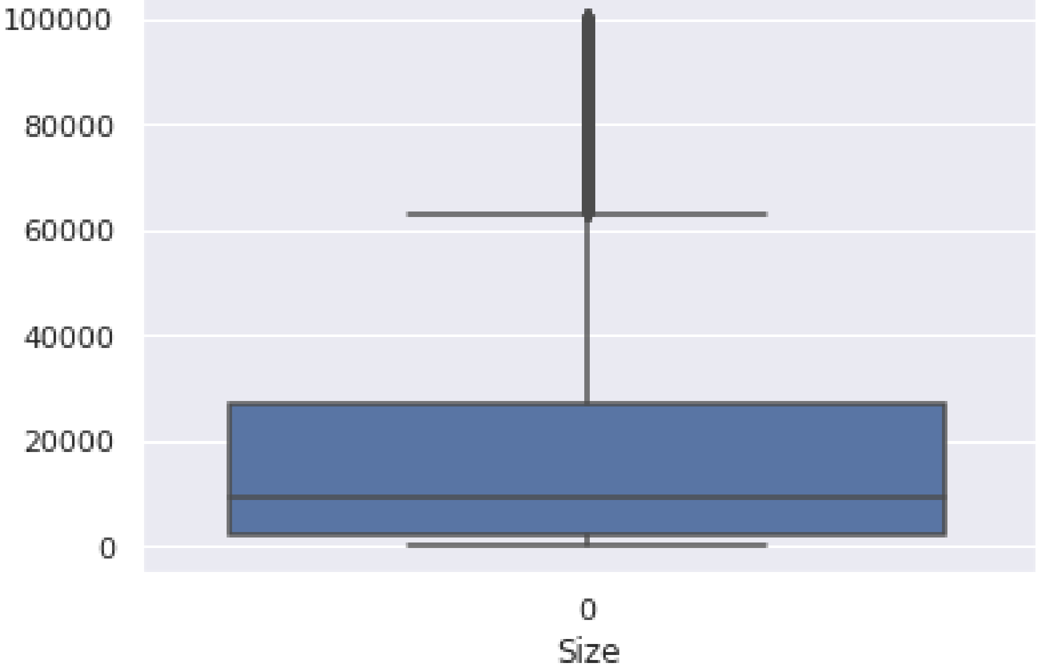
```
#Histogram for Size
data['Size'].plot.hist()
plt.show()
```



```
sns.boxplot(data['Rating'])
plt.xlabel('Rating')
plt.show()
```



```
sns.boxplot(data['Size'])
plt.xlabel('Size')
plt.show()
```



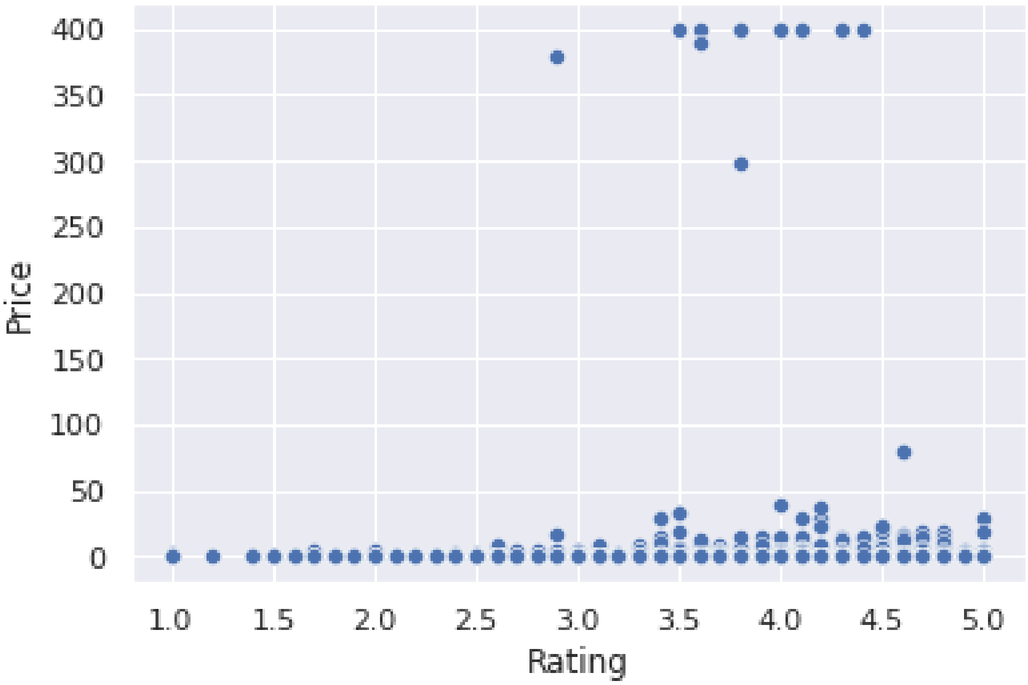
```
more = data.apply(lambda x : True
            if x['Price'] > 200 else False, axis = 1)

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

data.shape

(9353, 13)

#dropping the records with more than 2 million reviews
data = data[data.Reviews<=2000000]
data.shape

(8900, 13)

#dropping the apps that have a very high number of installs

data.quantile([.1, .25, .5, .70, .90, .95, .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     4255.50   9500.0     500000.0    0.0
0.70     4.5    35724.30  23000.0    1000000.0    0.0
0.90     4.7   295695.50  50000.0   10000000.0    0.0
0.95     4.8   637256.75  68000.0   10000000.0    1.0
0.99     5.0  1461766.93  95000.0  100000000.0    9.0

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

```
/tmp/ipykernel_350/3810112732.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  data.drop(data[data['Installs'] > 10000000].index, inplace = True)
```

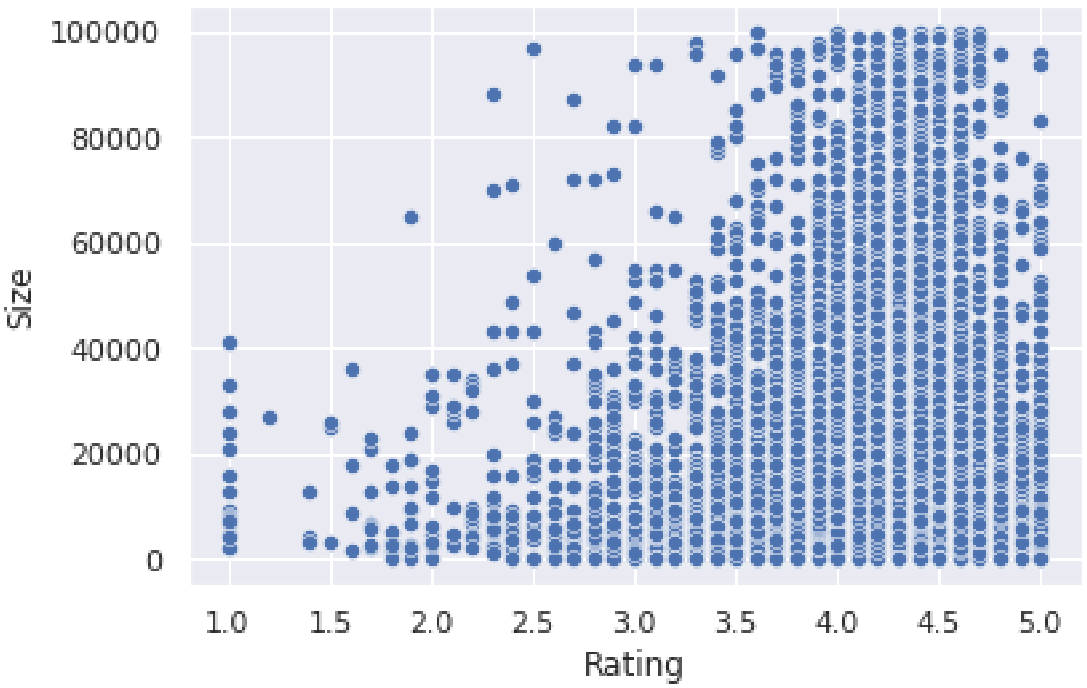
```
data.shape
```

```
(8511, 13)
```

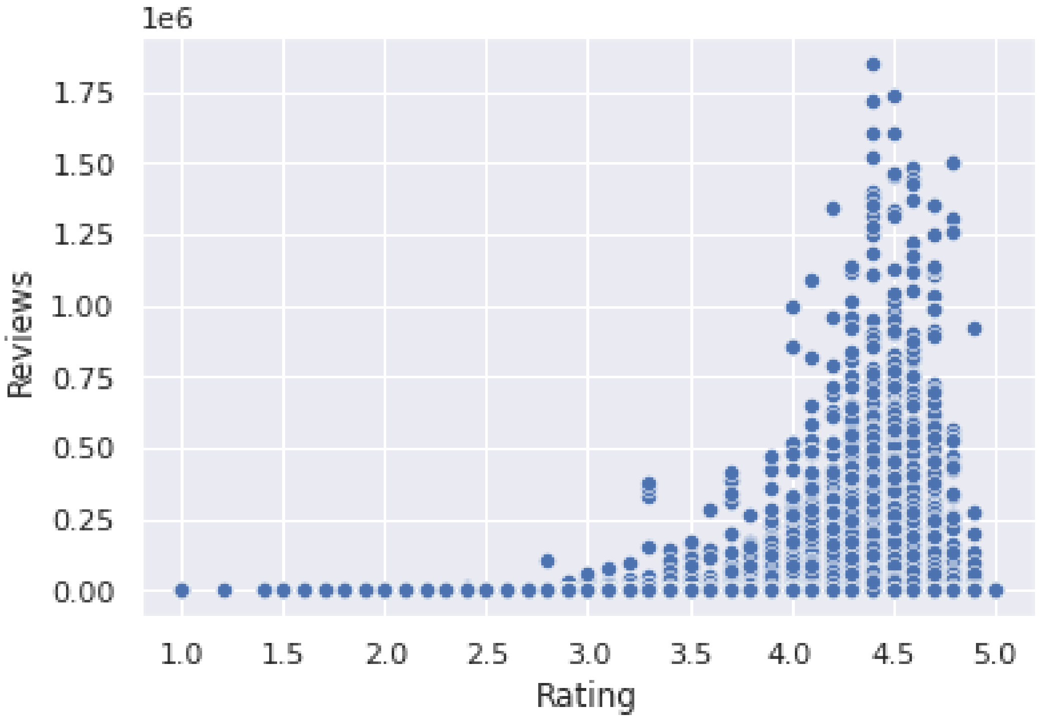
```
# 7(i) Make scatter plot/joinplot for Rating vs. Price
sns.scatterplot(x='Rating',y='Price',data=data)
plt.xlabel('Rating')
plt.ylabel('Price')
plt.show()
```



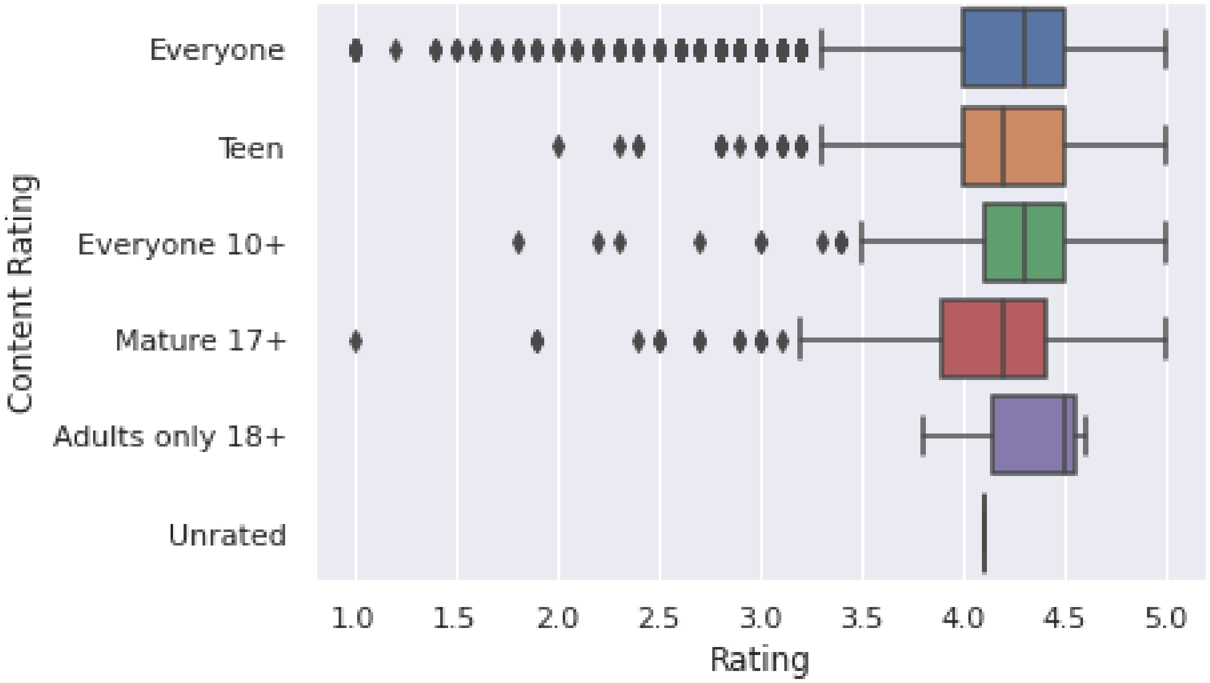
```
sns.scatterplot(x='Rating',y='Size',data=data)
plt.xlabel('Rating')
plt.ylabel('Size')
plt.show()
```



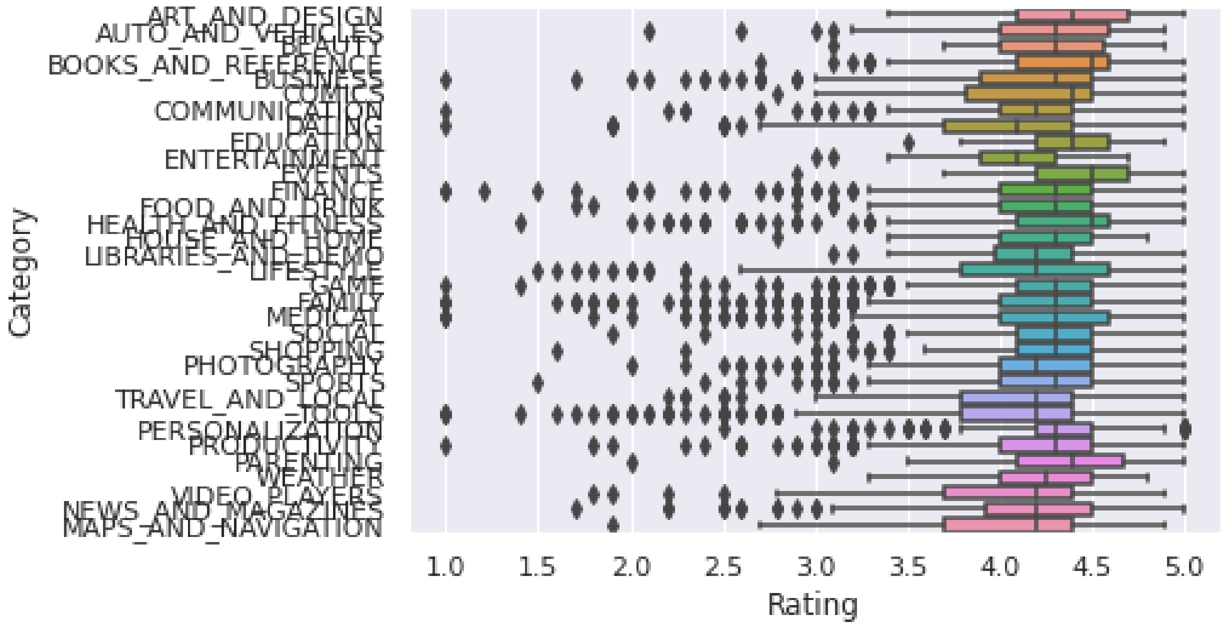
```
#Yes it is clear that heavior apps are rated better.

sns.scatterplot(x='Rating',y='Reviews',data=data)
plt.xlabel('Rating')
plt.ylabel('Reviews')
plt.show()
```



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

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

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



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

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

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



```
#Events category has best ratings compare to others.

inp1 = data

inp1.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   
4              Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN  
4.3   
5                         Paper flowers instructions  ART_AND_DESIGN  
4.4   

   Reviews     Size  Installs  Type  Price Content Rating  \
0    159.0  19000.0     10000  Free      0       Everyone   
1    967.0  14000.0    500000  Free      0       Everyone   
```

```
2  87510.0   8700.0   5000000  Free      0       Everyone
4    967.0   2800.0    100000  Free      0       Everyone
5    167.0   5600.0     50000  Free      0       Everyone

                      Genres      Last Updated Current Ver   Android
Ver
0               Art & Design   January 7, 2018       1.0.0  4.0.3 and
up
1  Art & Design;Pretend Play  January 15, 2018       2.0.0  4.0.3 and
up
2               Art & Design    August 1, 2018       1.2.4  4.0.3 and
up
4    Art & Design;Creativity     June 20, 2018         1.1    4.4 and
up
5               Art & Design    March 26, 2017         1.0    2.3 and
up
```

```
inp1.skew()
```

```
/tmp/ipykernel_350/3545313420.py:1: FutureWarning: Dropping of
nuisance columns in DataFrame reductions (with 'numeric_only=None') is
deprecated; in a future version this will raise TypeError.  Select
only valid columns before calling the reduction.
  inp1.skew()

Rating      -1.746804
Reviews      4.580662
Size         1.657389
Installs     1.546381
Price       23.326873
dtype: float64
```

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

```
/tmp/ipykernel_350/3380202217.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  inp1['Reviews'] = reviewskew
```

```
reviewskew.skew()
```

```
-0.19795445828004443
```

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

```
0           10000
1          500000
2         5000000
4          100000
5           50000
           ...
10834         500
10836        5000
10837         100
10839        1000
10840    10000000
Name: Installs, Length: 8511, dtype: int64
```

```
installsskew.skew()
```

```
-0.5063710333961149
```

```
inp1.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
4                Pixel Draw - Number Art Coloring Book  ART_AND_DESIGN
4.3
5                           Paper flowers instructions  ART_AND_DESIGN
4.4

     Reviews     Size  Installs  Type  Price Content Rating  \
0   5.075174  19000.0     10000  Free      0       Everyone
1   6.875232  14000.0    500000  Free      0       Everyone
2  11.379520   8700.0   5000000  Free      0       Everyone
4   6.875232   2800.0    100000  Free      0       Everyone
5   5.123964   5600.0     50000  Free      0       Everyone

                      Genres      Last Updated Current Ver   Android
Ver
0               Art & Design   January 7, 2018       1.0.0  4.0.3 and
up
1  Art & Design;Pretend Play  January 15, 2018       2.0.0  4.0.3 and
up
2               Art & Design    August 1, 2018       1.2.4  4.0.3 and
up
4    Art & Design;Creativity     June 20, 2018         1.1    4.4 and
up
```

```
5                Art & Design    March 26, 2017         1.0    2.3 and
up
```

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

```
/tmp/ipykernel_350/1867454288.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  inp1.drop(["Last Updated","Current Ver","Android
Ver","App","Type"],axis=1,inplace=True)
```

```
inp1.head()
```

```
         Category  Rating    Reviews     Size  Installs  Price Content
Rating  \
0  ART_AND_DESIGN     4.1   5.075174  19000.0     10000      0
Everyone
1  ART_AND_DESIGN     3.9   6.875232  14000.0    500000      0
Everyone
2  ART_AND_DESIGN     4.7  11.379520   8700.0   5000000      0
Everyone
4  ART_AND_DESIGN     4.3   6.875232   2800.0    100000      0
Everyone
5  ART_AND_DESIGN     4.4   5.123964   5600.0     50000      0
Everyone

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

```
inp1.shape
```

```
(8511, 8)
```

```
inp2 = inp1
```

```
inp2.head()
```

```
         Category  Rating    Reviews     Size  Installs  Price Content
Rating  \
0  ART_AND_DESIGN     4.1   5.075174  19000.0     10000      0
Everyone
1  ART_AND_DESIGN     3.9   6.875232  14000.0    500000      0
Everyone
```

```
2  ART_AND_DESIGN     4.7  11.379520   8700.0   5000000      0
Everyone
4  ART_AND_DESIGN     4.3   6.875232   2800.0    100000      0
Everyone
5  ART_AND_DESIGN     4.4   5.123964   5600.0     50000      0
Everyone

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

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

```
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)
```

```
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()
```

```
/tmp/ipykernel_350/2005045317.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  inp2.Category = pd.Categorical(inp2.Category)

   Rating    Reviews     Size  Installs  Price Content Rating  \
0     4.1   5.075174  19000.0     10000      0       Everyone
1     3.9   6.875232  14000.0    500000      0       Everyone
2     4.7  11.379520   8700.0   5000000      0       Everyone
```

```
4     4.3   6.875232   2800.0    100000      0       Everyone
5     4.4   5.123964   5600.0     50000      0       Everyone

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

   Category_AUTO_AND_VEHICLES  Category_BEAUTY  ...
Category_PERSONALIZATION  \
0                           0                0  ...
0
1                           0                0  ...
0
2                           0                0  ...
0
4                           0                0  ...
0
5                           0                0  ...
0

   Category_PHOTOGRAPHY  Category_PRODUCTIVITY  Category_SHOPPING  \
0                     0                      0                  0
1                     0                      0                  0
2                     0                      0                  0
4                     0                      0                  0
5                     0                      0                  0

   Category_SOCIAL  Category_SPORTS  Category_TOOLS  \
0                0                0               0
1                0                0               0
2                0                0               0
4                0                0               0
5                0                0               0

   Category_TRAVEL_AND_LOCAL  Category_VIDEO_PLAYERS  Category_WEATHER

0                          0                       0                 0

1                          0                       0                 0

2                          0                       0                 0

4                          0                       0                 0

5                          0                       0                 0
```

```
[5 rows x 40 columns]
```

```
inp2.shape
```

```
(8511, 40)
```

```
#Let's apply Dummy EnCoding on Column "Genres"
```

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

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

```
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]

inp2["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)

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()

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

   Category_ART_AND_DESIGN  Category_AUTO_AND_VEHICLES
Category_BEAUTY  \
0                        1                           0
0
1                        1                           0
0
2                        1                           0
0
4                        1                           0
0
5                        1                           0
0

   Category_BOOKS_AND_REFERENCE  ...  Genres_Simulation  Genres_Social
\
0                             0  ...                  0              0

1                             0  ...                  0              0

2                             0  ...                  0              0

4                             0  ...                  0              0

5                             0  ...                  0              0


   Genres_Sports  Genres_Strategy  Genres_Tools  Genres_Travel & Local
\
0              0                0             0                      0

1              0                0             0                      0

2              0                0             0                      0

4              0                0             0                      0

5              0                0             0                      0
```

```
   Genres_Trivia  Genres_Video Players & Editors  Genres_Weather
Genres_Word
0              0                               0               0
0
1              0                               0               0
0
2              0                               0               0
0
4              0                               0               0
0
5              0                               0               0
0

[5 rows x 91 columns]
```

```
inp2.shape
```

```
(8511, 91)
```

```
#Let's apply Dummy EnCoding on Column "Content Rating"
```

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

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

```
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()
```

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

   Category_AUTO_AND_VEHICLES  Category_BEAUTY
```

```
Category_BOOKS_AND_REFERENCE  \
0                           0                0
0
1                           0                0
0
2                           0                0
0
4                           0                0
0
5                           0                0
0

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

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

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

   Content Rating_Mature 17+  Content Rating_Teen  Content
Rating_Unrated
0                          0                    0
0
1                          0                    0
0
2                          0                    0
0
4                          0                    0
0
5                          0                    0
```

```
0

[5 rows x 96 columns]
```

```
inp2.shape
```

```
(8511, 96)
```

```
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 = inp2
X_train = d1.drop('Rating',axis=1)
y_train = d1['Rating']

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

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
/tmp/ipykernel_350/572442865.py in <cell line: 5>()
      3 y_train = d1['Rating']
      4 
----> 5 Xtrain, Xtest, ytrain, ytest = tts(X,y, test_size=0.3, random_state=5)

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

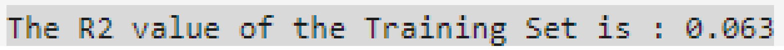
```
reg_all = LR()
reg_all.fit(X_train,y_train)
```

```
LinearRegression()
```

```
R2_train = round(reg_all.score(X_train,y_train),3)
print("The R2 value of the Training Set is : {}".format(R2_train))
```

```
The R2 value of the Training Set is : 0.072
```

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



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