```
import pandas as pd
```

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

```
import os
```

```
os.getcwd()
```

```
'C:\\Users\\admin\\Documents'
```

```
import warnings
```

```
warnings.filterwarnings('ignore')
```

```
data.head(5)
```

```
                                                 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
```

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

```
data.isna().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
```

```
# Drop records with nulls in any of the columns.
```

```
data.dropna(inplace=True)
```

```
#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
```

```
data["Size"]
```

```
0                        19M
1                        14M
2                       8.7M
3                        25M
4                       2.8M
                ...
10834                   2.6M
10836                    53M
10837                   3.6M
10839    Varies with device
10840                    19M
Name: Size, Length: 9360, dtype: object
```

```
data = data[-data["Size"].str.contains("Var")]
```

```
data["Size"]
```

```
0         19M
1         14M
2        8.7M
3         25M
4        2.8M
```

```
         ...
10833    619k
10834    2.6M
10836     53M
10837    3.6M
10840     19M
Name: Size, Length: 7723, dtype: object
```

```
data.loc[:, "Sizenum"]= data["Size"].str.rstrip('MKk+')
```

```
data.Sizenum = pd.to_numeric(data["Sizenum"])
```

```
data.Sizenum.dtype
```

```
dtype('float64')
```

```
data.columns
```

```
Index(['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 
'Type',
       'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current 
Ver',
       'Android Ver', 'Sizenum'],
      dtype='object')
```

```
import numpy as np
```

```
data["Sizenum"]= np.where(data["Size"].str.contains("M"), 
data["Sizenum"]*1000, data["Sizenum"])
```

```
data ["Size"] = data ["Sizenum"]
```

```
data.head(5)
```

```
                                                 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  19000.0      10,000+  Free     0       Everyone   
1     967  14000.0     500,000+  Free     0       Everyone   
2   87510   8700.0   5,000,000+  Free     0       Everyone   
3  215644  25000.0  50,000,000+  Free     0           Teen   
```

```
4      967   2800.0      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  Sizenum
0  4.0.3 and up  19000.0
1  4.0.3 and up  14000.0
2  4.0.3 and up   8700.0
3    4.2 and up  25000.0
4    4.4 and up   2800.0
```

```
data.drop("Sizenum", axis=1, inplace= True)
```

```
data.head(5)
```

```
                                                 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  19000.0       10,000+  Free     0       Everyone
1      967  14000.0      500,000+  Free     0       Everyone
2    87510   8700.0    5,000,000+  Free     0       Everyone
3   215644  25000.0   50,000,000+  Free     0           Teen
4      967   2800.0      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
```

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

```
data.Reviews= pd.to_numeric(data.Reviews)
```

```
data["Reviews"].dtype
```

```
dtype('int64')
```

```
#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
```

```
data["Installs"]= data.Installs.str.replace("+","")
```

```
data["Installs"]= data.Installs.str.replace(",","")
```

```
data.Installs = pd.to_numeric(data.Installs)
```

```
data.Installs.dtype
```

```
dtype('int64')
```

```
#Price field is a string and has $ symbol. Remove ‘$’ sign, and
convert it to numeric.
```

```
data.Price = data.Price.str.replace("$","")
```

```
data.Price = pd.to_numeric(data.Price)
```

```
data.tail(5)
```

```
                                                 App
Category  \
10833                                    Chemin (fr)
BOOKS_AND_REFERENCE
10834                                  FR Calculator
FAMILY
10836                               Sya9a Maroc - FR
FAMILY
10837               Fr. Mike Schmitz Audio Teachings
FAMILY
10840  iHoroscope - 2018 Daily Horoscope & Astrology
LIFESTYLE

       Rating  Reviews     Size  Installs  Type  Price Content Rating
```

```
\
10833     4.8        44    619.0       1000  Free    0.0       Everyone

10834     4.0         7   2600.0        500  Free    0.0       Everyone

10836     4.5        38  53000.0       5000  Free    0.0       Everyone

10837     5.0         4   3600.0        100  Free    0.0       Everyone

10840     4.5    398307  19000.0  10000000  Free    0.0       Everyone


                  Genres    Last Updated         Current Ver  \
10833  Books & Reference  March 23, 2014                 0.8
10834          Education   June 18, 2017               1.0.0
10836          Education   July 25, 2017                1.48
10837          Education    July 6, 2018                 1.0
10840          Lifestyle   July 25, 2018  Varies with device

              Android Ver
10833          2.2 and up
10834          4.1 and up
10836          4.1 and up
10837          4.1 and up
10840  Varies with device
```

```
#Sanity checks:

#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.

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

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

data= data[(data["Rating"]>=1)&(data["Rating"]<=5)]

data= data[data["Reviews"]<= data ["Installs"]]

len(data.index)
```

```
7717
```

```
data[(data["Type"]=="Free")&(data["Price"]>0)]
```

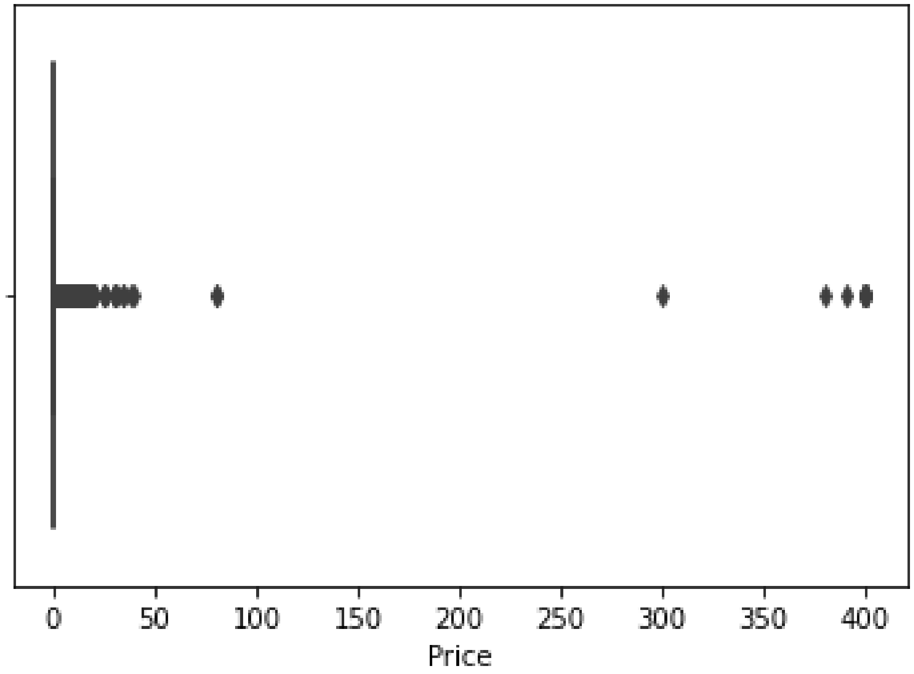
```
Empty DataFrame
Columns: [App, Category, Rating, Reviews, Size, Installs, Type, Price,
Content Rating, Genres, Last Updated, Current Ver, Android Ver]
Index: []
```

```
#Performing univariate analysis:

#Boxplot for Price

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

#Boxplot for Reviews

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

#Histogram for Rating

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

#Histogram for Size

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

```
import seaborn as sns
```

```
sns.boxplot(x="Price", data=data)
```

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



```
#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 and suspicious!

#Check out the records with very high price

#Is 200 indeed a high price?

#Drop these as most seem to be junk apps

data[data["Price"]>=200]
```

```
                                 App   Category  Rating  Reviews
Size  \
4197          most expensive app (H)     FAMILY     4.3        6
1500.0
4362                      □ I'm rich  LIFESTYLE     3.8      718
26000.0
4367        I'm Rich - Trump Edition  LIFESTYLE     3.6      275
7300.0
5351                       I am rich  LIFESTYLE     3.8     3547
1800.0
5354                  I am Rich Plus     FAMILY     4.0      856
8700.0
5355                   I am rich VIP  LIFESTYLE     3.8      411
2600.0
5356               I Am Rich Premium    FINANCE     4.1     1867
4700.0
5357             I am extremely Rich  LIFESTYLE     2.9       41
2900.0
5358                      I am Rich!    FINANCE     3.8       93
22000.0
5359              I am rich(premium)    FINANCE     3.5      472
965.0
5362                   I Am Rich Pro     FAMILY     4.4      201
2700.0
5364  I am rich (Most expensive app)    FINANCE     4.1      129
2700.0
5366                       I Am Rich     FAMILY     3.6      217
4900.0
5369                       I am Rich    FINANCE     4.3      180
3800.0
5373              I AM RICH PRO PLUS    FINANCE     4.0       36
41000.0

      Installs  Type   Price Content Rating         Genres        Last
Updated  \
4197       100  Paid  399.99       Everyone  Entertainment       July
```

```
16, 2018
4362     10000  Paid  399.99       Everyone      Lifestyle     March
11, 2018
4367     10000  Paid  400.00       Everyone      Lifestyle        May
3, 2018
5351    100000  Paid  399.99       Everyone      Lifestyle   January
12, 2018
5354     10000  Paid  399.99       Everyone  Entertainment       May
19, 2018
5355     10000  Paid  299.99       Everyone      Lifestyle      July
21, 2018
5356     50000  Paid  399.99       Everyone        Finance  November
12, 2017
5357      1000  Paid  379.99       Everyone      Lifestyle       July
1, 2018
5358      1000  Paid  399.99       Everyone        Finance  December
11, 2017
5359      5000  Paid  399.99       Everyone        Finance        May
1, 2017
5362      5000  Paid  399.99       Everyone  Entertainment       May
30, 2017
5364      1000  Paid  399.99           Teen        Finance   December
6, 2017
5366     10000  Paid  389.99       Everyone  Entertainment      June
22, 2018
5369      5000  Paid  399.99       Everyone        Finance     March
22, 2018
5373      1000  Paid  399.99       Everyone        Finance      June
25, 2018

     Current Ver   Android Ver
4197         1.0    7.0 and up
4362       1.0.0    4.4 and up
4367       1.0.1    4.1 and up
5351         2.0  4.0.3 and up
5354         3.0    4.4 and up
5355       1.1.1    4.3 and up
5356         1.6    4.0 and up
5357         1.0    4.0 and up
5358         1.0    4.1 and up
5359         3.4    4.4 and up
5362        1.54    1.6 and up
5364           2  4.0.3 and up
5366         1.5    4.2 and up
5369         1.0    4.2 and up
5373       1.0.2    4.1 and up
```

```python
len(data[data["Price"]>=200])
```

```
15
```

```
data= data.drop(data.index[data["Price"]>=200])
```

```
#  Outliers on Price have been removed
```

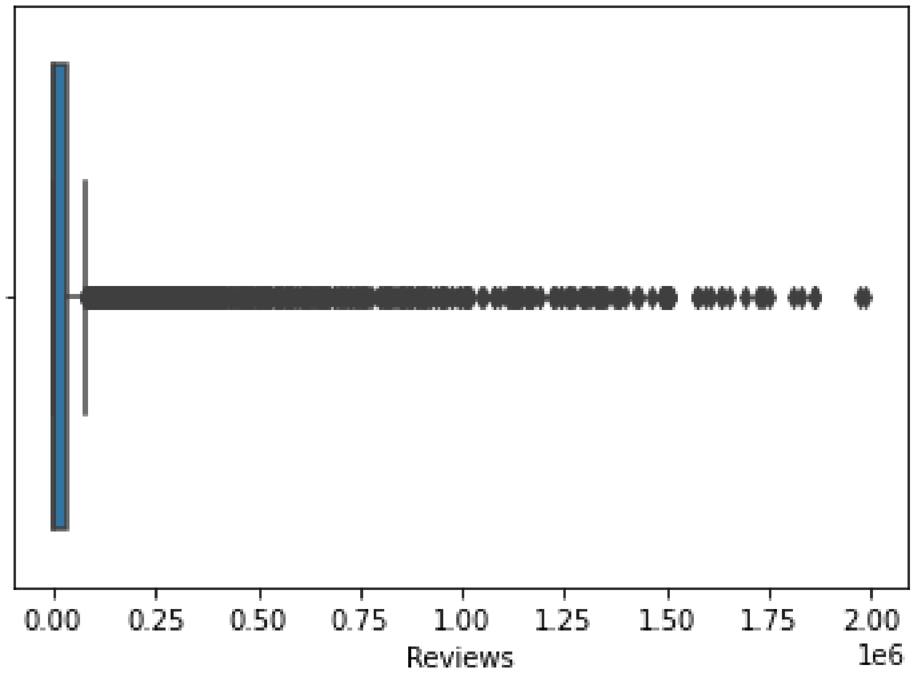
```
#Reviews: Very few apps have very high number of reviews. These are
all star apps that don’t help with the analysis and, in fact, will
skew it. Drop records having more than 2 million reviews.
```

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

```
7483
```

```
sns.boxplot(x="Reviews", data=data)
```

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

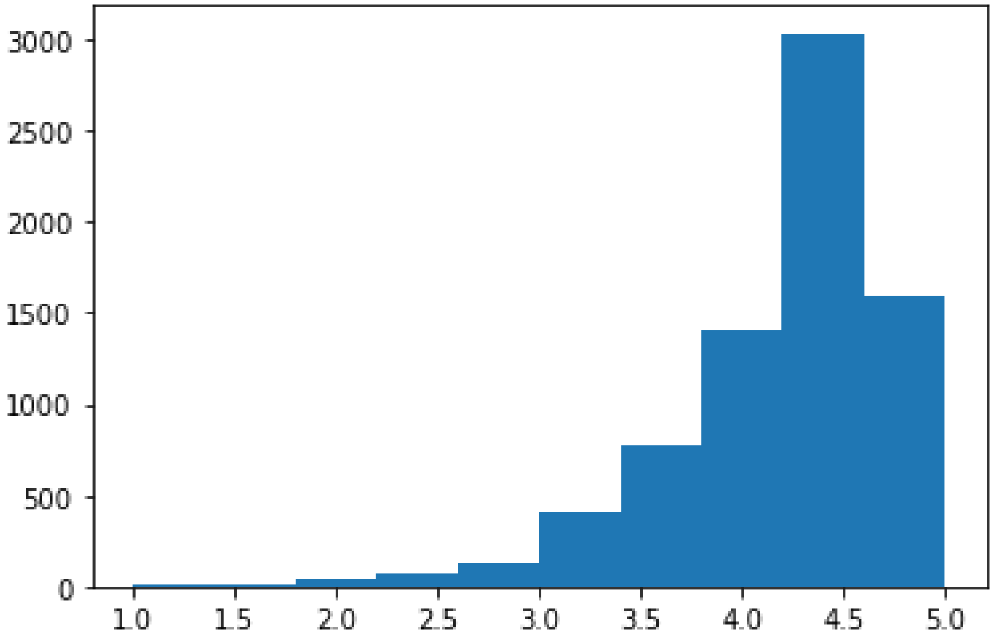


```
#Are there any apps with very high number of reviews? Do the values
seem right?
# No there is no app with higher number of reviews, as we have droped
values which are higher in number.
```

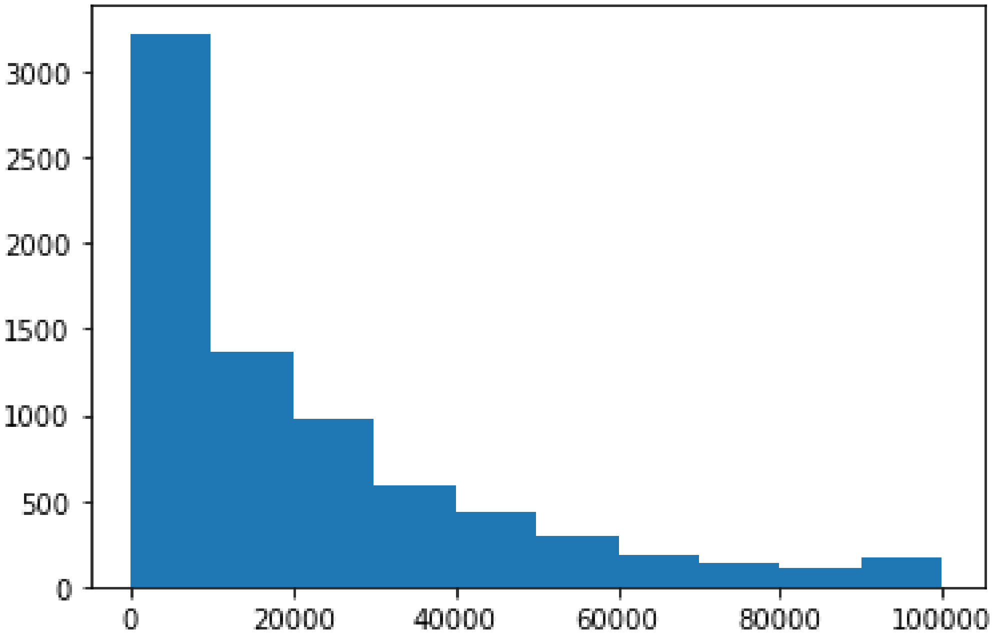
```
import matplotlib.pyplot as plt
```

```
plt.hist(data["Rating"])
```

```
(array([  17.,   18.,   39.,   72.,  131.,  408.,  774., 1399., 3036.,
        1589.]),
 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>)
```

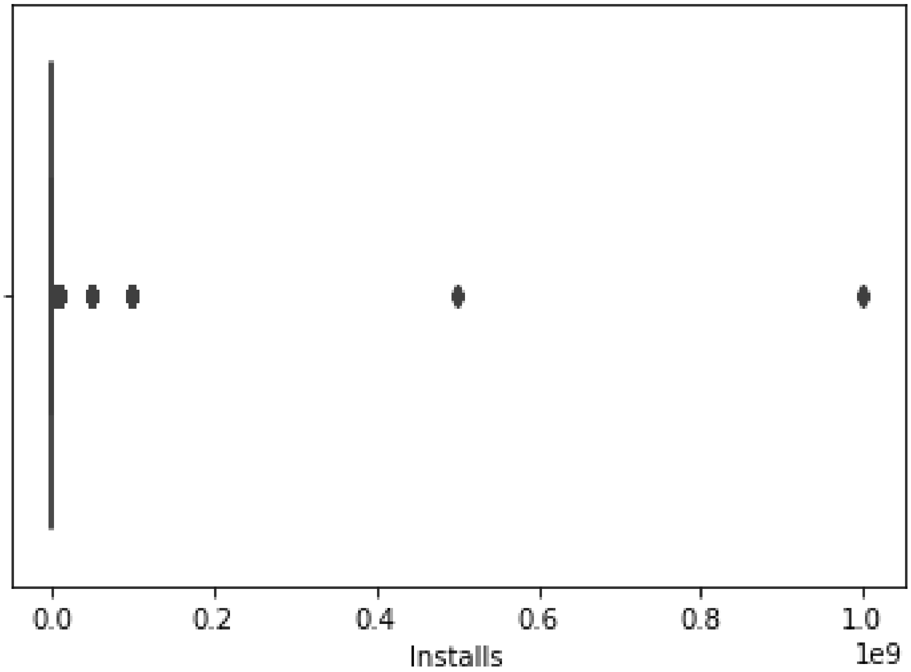


```
#How are the ratings distributed? Is it more toward higher ratings?
# The ratings are mostly given between 3 and 5, Where, we can see the
highest number of rating is 4.5 from nearly 3000 people, This shows a
good graph for the app among the Users.

plt.hist(data["Size"])

(array([3223., 1371.,  976.,  589.,  436.,  293.,  190.,  134.,  107.,
         164.]),
 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>)
```



```
# Note down your observations for the plots made above. Which of these
seem to have outliers?
# There is no Outliers as we have removed all the outliers

#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

sns.boxplot(x="Installs", data=data)
```

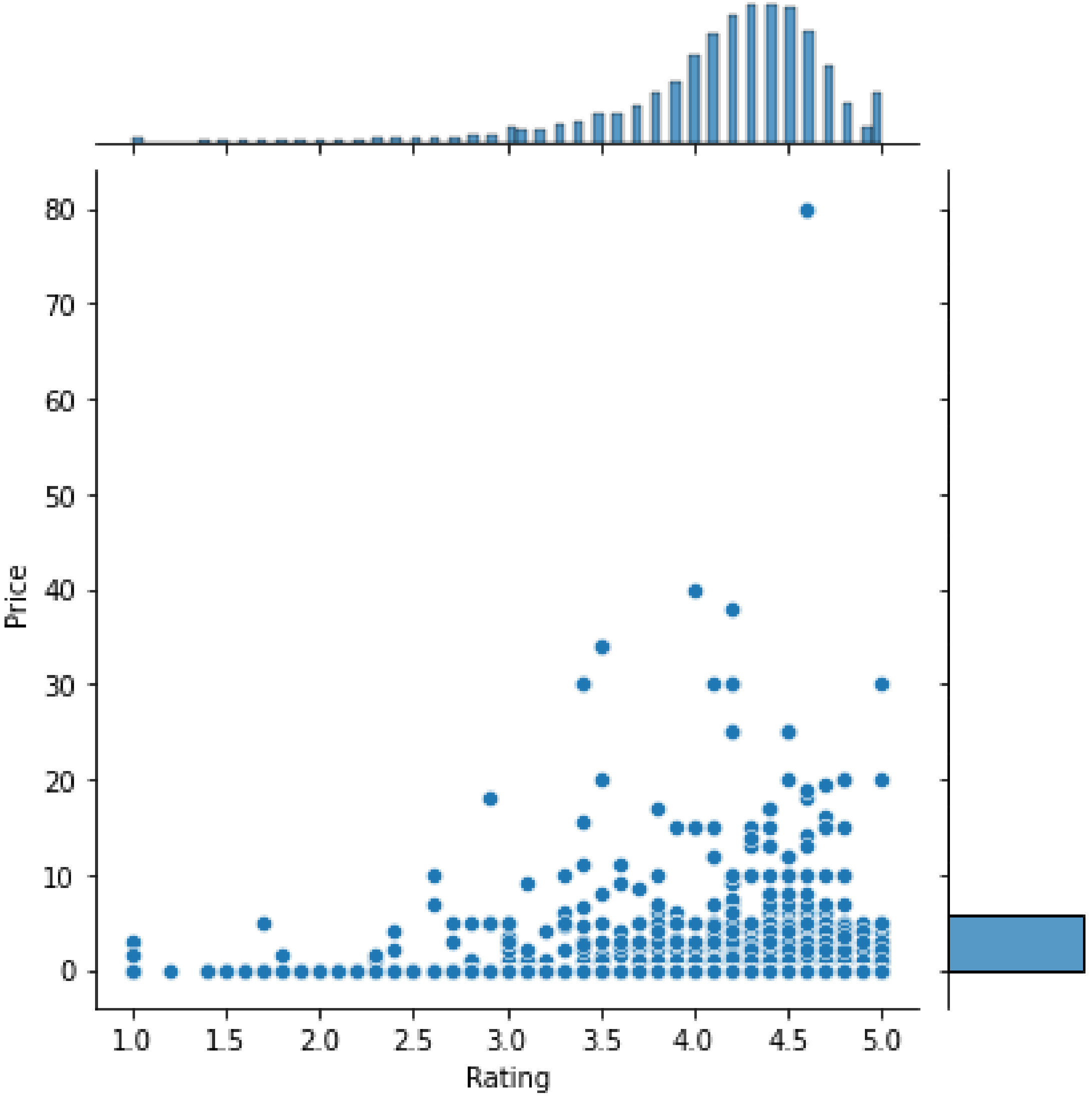
```
<AxesSubplot:xlabel='Installs'>
```



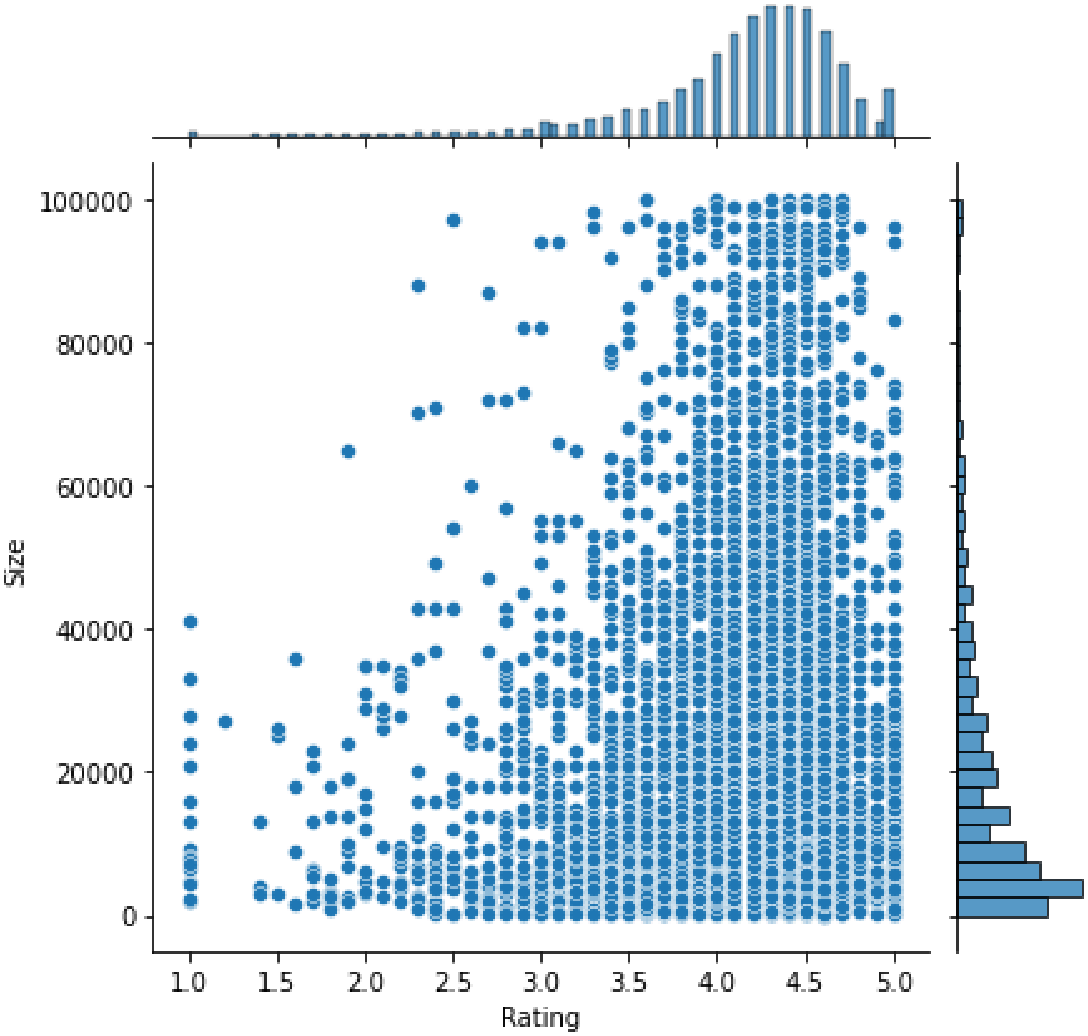
```
import numpy as np

np. percentile(data["Installs"],10)

1000.0

np. percentile(data["Installs"],25)

10000.0

np. percentile(data["Installs"],50)

100000.0

np. percentile(data["Installs"],70)

1000000.0

np. percentile(data["Installs"],90)

10000000.0

np. percentile(data["Installs"],95)

10000000.0

Installs_99percentile= np. percentile(data["Installs"],99)

data.drop(data.index[data.Installs>=Installs_99percentile])

                                                     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
4                  Pixel Draw - Number Art Coloring Book
ART_AND_DESIGN
5                             Paper flowers instructions
ART_AND_DESIGN
...                                                  ...
...
10833                                        Chemin (fr)
BOOKS_AND_REFERENCE
10834                                      FR Calculator
FAMILY
10836                                   Sya9a Maroc - FR
FAMILY
10837                   Fr. Mike Schmitz Audio Teachings
FAMILY
10840      iHoroscope - 2018 Daily Horoscope & Astrology
LIFESTYLE

       Rating  Reviews     Size  Installs  Type  Price Content Rating
\
0         4.1      159  19000.0     10000  Free    0.0       Everyone

1         3.9      967  14000.0    500000  Free    0.0       Everyone

2         4.7    87510   8700.0   5000000  Free    0.0       Everyone

4         4.3      967   2800.0    100000  Free    0.0       Everyone

5         4.4      167   5600.0     50000  Free    0.0       Everyone

...       ...      ...      ...       ...   ...    ...            ...

10833     4.8       44    619.0      1000  Free    0.0       Everyone

10834     4.0        7   2600.0       500  Free    0.0       Everyone

10836     4.5       38  53000.0      5000  Free    0.0       Everyone

10837     5.0        4   3600.0       100  Free    0.0       Everyone

10840     4.5   398307  19000.0  10000000  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

4        Art & Design;Creativity     June 20, 2018                 1.1

5                   Art & Design    March 26, 2017                 1.0

...                          ...               ...                 ...

10833          Books & Reference    March 23, 2014                 0.8

10834                  Education     June 18, 2017               1.0.0

10836                  Education     July 25, 2017                1.48

10837                  Education      July 6, 2018                 1.0

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
4              4.4 and up
5              2.3 and up
...                   ...
10833          2.2 and up
10834          4.1 and up
10836          4.1 and up
10837          4.1 and up
10840  Varies with device

[7307 rows x 13 columns]
```

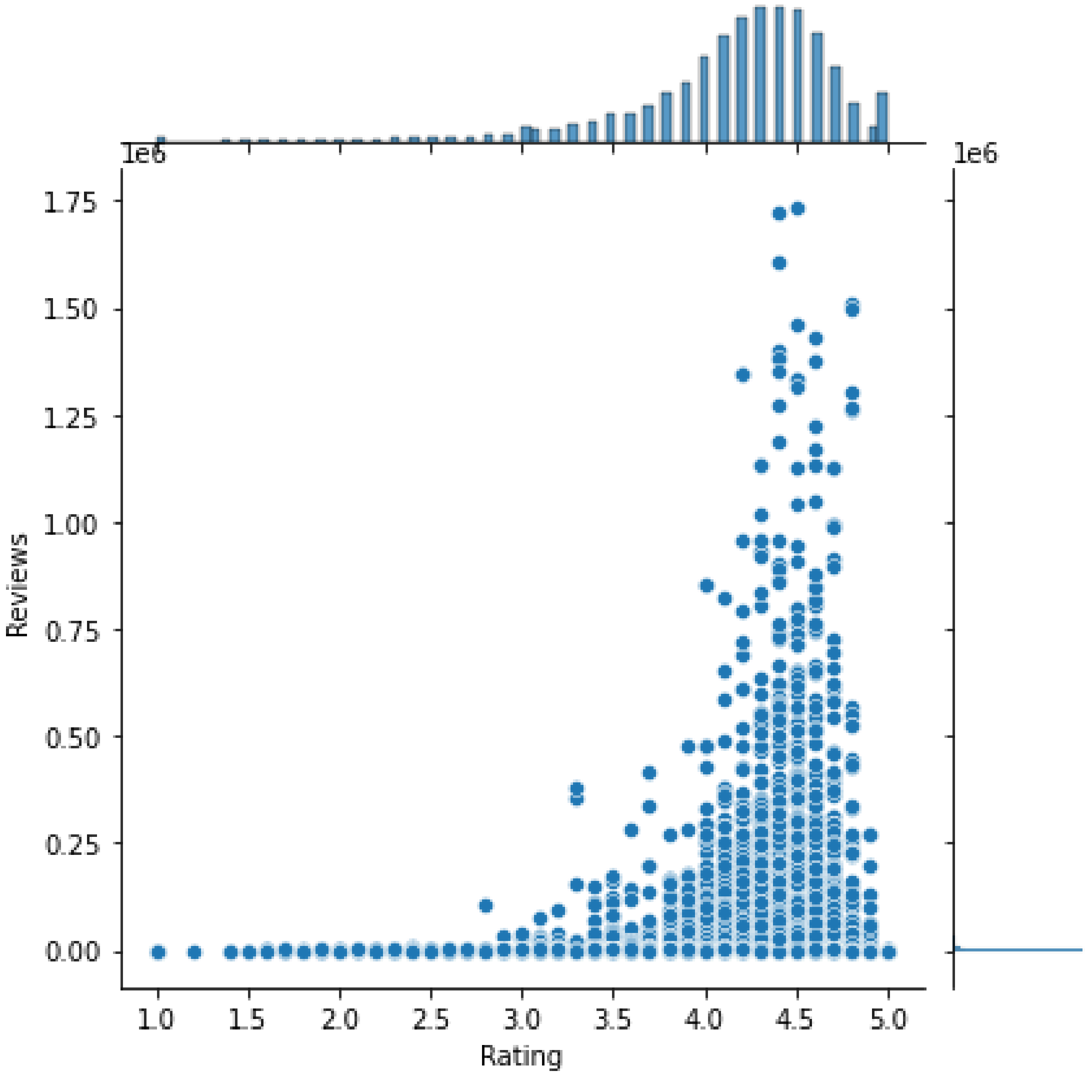
```
data.drop(data.index[data.Installs>=Installs_99percentile], inplace =
True)

# Bivariate analysis

# Make scatter plot/joinplot for Rating vs. Price

sns.jointplot("Rating", "Price", data= data)
plt.savefig("Jointplot_Rating_Price_Proj.png")
```



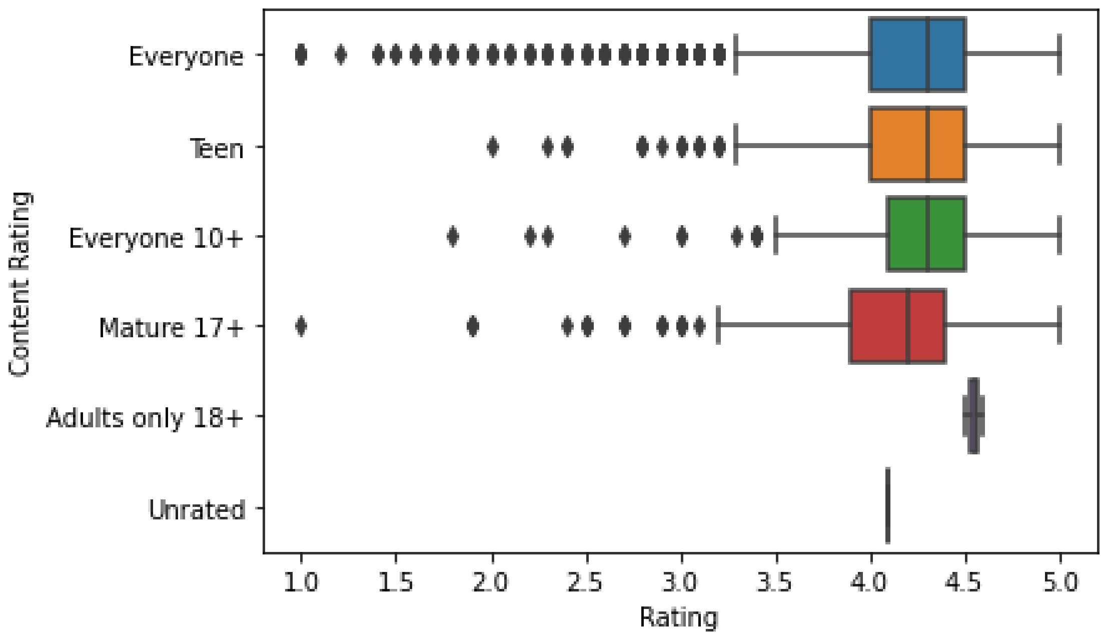
```
# What pattern do you observe? Does rating increase with price?
# Yes, rating increases along with the price

# Make scatter plot/joinplot for Rating vs. Size

sns.jointplot("Rating", "Size", data= data)
plt.savefig("Jointplot_Rating_Size_Proj.png")
```



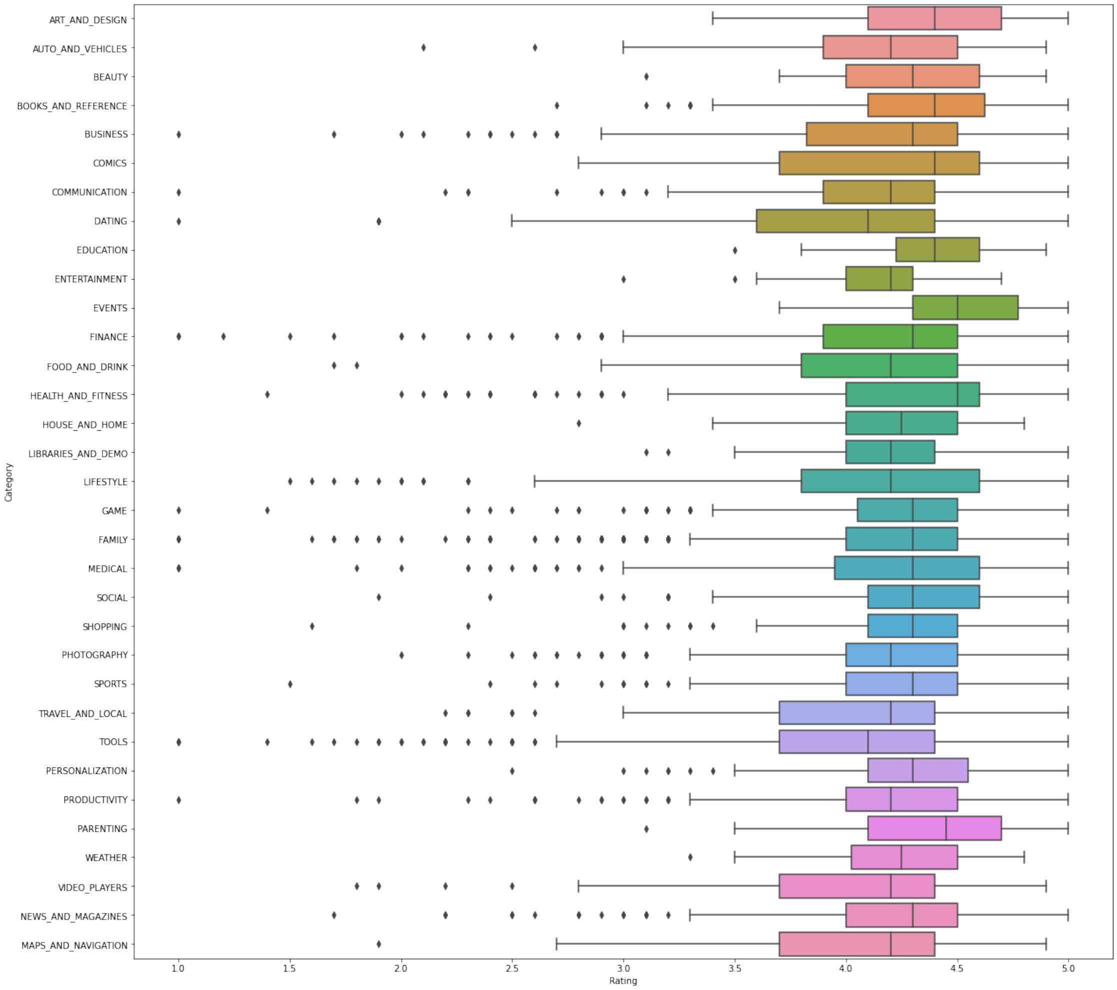
```
# Are heavier apps rated better?
# Yes, heavier apps rated better

# Make scatter plot/joinplot for Rating vs. Reviews

sns.jointplot("Rating", "Reviews", data= data)
plt.savefig("Jointplot_Rating_Reviews_Proj.png")
```



```
# Does more review mean a better rating always?
# Yes, more reviews has a bettrer rating

# Make boxplot for Rating vs. Content Rating

sns.boxplot(x= "Rating", y= "Content Rating", data= data)
plt.savefig("Boxplot_Rating_ContentRating_Proj.png")
```



```
# Is there any difference in the ratings? Are some types liked better?
# No, There is no much difference in the ratings

# Make boxplot for Ratings vs. Category

plt.figure(figsize=(20,20))
sns.boxplot(x= "Rating", y= "Category", data= data)
plt.savefig("Boxplot_Rating_Category_Proj.png")
```



```
# Which genre has the best ratings?
# Events has the best rating

# 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. Apply log transformation (np.log1p) to Reviews and Installs.

inp1= data.copy()

inp1.Reviews=inp1.Reviews.apply(np.log1p)
inp1.Installs=inp1.Installs.apply(np.log1p)

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

inp1.drop(columns=['App','Last Updated', 'Current Ver', 'Android
Ver'], inplace= True)

inp1.head(5)
```

```
         Category  Rating    Reviews     Size   Installs  Type  Price  
\
0  ART_AND_DESIGN     4.1   5.075174  19000.0   9.210440  Free    0.0   

1  ART_AND_DESIGN     3.9   6.875232  14000.0  13.122365  Free    0.0   

2  ART_AND_DESIGN     4.7  11.379520   8700.0  15.424949  Free    0.0   

4  ART_AND_DESIGN     4.3   6.875232   2800.0  11.512935  Free    0.0   

5  ART_AND_DESIGN     4.4   5.123964   5600.0  10.819798  Free    0.0   


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

```
# 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.

inp2= pd.get_dummies(inp1)

inp2.head(5)
```

```
   Rating    Reviews     Size   Installs  Price  
Category_ART_AND_DESIGN  \
0     4.1   5.075174  19000.0   9.210440    0.0                        
1   
1     3.9   6.875232  14000.0  13.122365    0.0                        
1   
2     4.7  11.379520   8700.0  15.424949    0.0                        
1   
4     4.3   6.875232   2800.0  11.512935    0.0                        
1   
5     4.4   5.123964   5600.0  10.819798    0.0                        
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_Strategy;Education  Genres_Tools  \
0                  0  ...                          0             0
1                  0  ...                          0             0
2                  0  ...                          0             0
4                  0  ...                          0             0
5                  0  ...                          0             0

   Genres_Travel & Local  Genres_Travel & Local;Action & Adventure  \
0                      0                                         0
1                      0                                         0
2                      0                                         0
4                      0                                         0
5                      0                                         0

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

   Genres_Video Players & Editors;Creativity  \
0                                          0
1                                          0
2                                          0
4                                          0
5                                          0

   Genres_Video Players & Editors;Music & Video  Genres_Weather
Genres_Word
0                                             0               0
0
1                                             0               0
0
2                                             0               0
0
4                                             0               0
0
5                                             0               0
0

[5 rows x 158 columns]
```

```
set(inp2.columns)
```

```
{'Category_ART_AND_DESIGN',
 'Category_AUTO_AND_VEHICLES',
 'Category_BEAUTY',
 'Category_BOOKS_AND_REFERENCE',
 'Category_BUSINESS',
 'Category_COMICS',
 'Category_COMMUNICATION',
 'Category_DATING',
 'Category_EDUCATION',
 'Category_ENTERTAINMENT',
 'Category_EVENTS',
 'Category_FAMILY',
 'Category_FINANCE',
 'Category_FOOD_AND_DRINK',
 'Category_GAME',
 'Category_HEALTH_AND_FITNESS',
 'Category_HOUSE_AND_HOME',
 'Category_LIBRARIES_AND_DEMO',
 'Category_LIFESTYLE',
 'Category_MAPS_AND_NAVIGATION',
 'Category_MEDICAL',
 'Category_NEWS_AND_MAGAZINES',
 'Category_PARENTING',
 'Category_PERSONALIZATION',
 'Category_PHOTOGRAPHY',
 'Category_PRODUCTIVITY',
 'Category_SHOPPING',
 'Category_SOCIAL',
 'Category_SPORTS',
 'Category_TOOLS',
 'Category_TRAVEL_AND_LOCAL',
 'Category_VIDEO_PLAYERS',
 'Category_WEATHER',
 'Content Rating_Adults only 18+',
 'Content Rating_Everyone',
 'Content Rating_Everyone 10+',
 'Content Rating_Mature 17+',
 'Content Rating_Teen',
 'Content Rating_Unrated',
 'Genres_Action',
 'Genres_Action;Action & Adventure',
 'Genres_Adventure',
 'Genres_Adventure;Action & Adventure',
 'Genres_Adventure;Brain Games',
 'Genres_Adventure;Education',
 'Genres_Arcade',
 'Genres_Arcade;Action & Adventure',
 'Genres_Arcade;Pretend Play',
 'Genres_Art & Design',
 'Genres_Art & Design;Creativity',
 'Genres_Art & Design;Pretend Play',
 'Genres_Auto & Vehicles',
 'Genres_Beauty',
 'Genres_Board',
 'Genres_Board;Action & Adventure',
 'Genres_Board;Brain Games',
 'Genres_Board;Pretend Play',
 'Genres_Books & Reference',
 'Genres_Books & Reference;Education',
 'Genres_Business',
 'Genres_Card',
 'Genres_Card;Action & Adventure',
 'Genres_Card;Brain Games',
 'Genres_Casino',
 'Genres_Casual',
 'Genres_Casual;Action & Adventure',
 'Genres_Casual;Brain Games',
 'Genres_Casual;Creativity',
 'Genres_Casual;Education',
 'Genres_Casual;Music & Video',
 'Genres_Casual;Pretend Play',
 'Genres_Comics',
 'Genres_Comics;Creativity',
 'Genres_Communication',
 'Genres_Dating',
 'Genres_Education',
 'Genres_Education;Action & Adventure',
 'Genres_Education;Brain Games',
 'Genres_Education;Creativity',
 'Genres_Education;Education',
 'Genres_Education;Music & Video',
 'Genres_Education;Pretend Play',
 'Genres_Educational',
 'Genres_Educational;Action & Adventure',
 'Genres_Educational;Brain Games',
 'Genres_Educational;Creativity',
 'Genres_Educational;Education',
 'Genres_Educational;Pretend Play',
 'Genres_Entertainment',
 'Genres_Entertainment;Action & Adventure',
 'Genres_Entertainment;Brain Games',
 'Genres_Entertainment;Creativity',
 'Genres_Entertainment;Education',
 'Genres_Entertainment;Music & Video',
 'Genres_Entertainment;Pretend Play',
 'Genres_Events',
 'Genres_Finance',
 'Genres_Food & Drink',
 'Genres_Health & Fitness',
 'Genres_Health & Fitness;Action & Adventure',
 'Genres_Health & Fitness;Education',
 'Genres_House & Home',
 'Genres_Libraries & Demo',
 'Genres_Lifestyle',
 'Genres_Lifestyle;Pretend Play',
 'Genres_Maps & Navigation',
 'Genres_Medical',
 'Genres_Music',
 'Genres_Music & Audio;Music & Video',
 'Genres_Music;Music & Video',
 'Genres_News & Magazines',
 'Genres_Parenting',
 'Genres_Parenting;Brain Games',
 'Genres_Parenting;Education',
 'Genres_Parenting;Music & Video',
 'Genres_Personalization',
 'Genres_Photography',
 'Genres_Productivity',
 'Genres_Puzzle',
 'Genres_Puzzle;Action & Adventure',
 'Genres_Puzzle;Brain Games',
 'Genres_Puzzle;Creativity',
 'Genres_Puzzle;Education',
 'Genres_Racing',
 'Genres_Racing;Action & Adventure',
 'Genres_Racing;Pretend Play',
 'Genres_Role Playing',
 'Genres_Role Playing;Action & Adventure',
 'Genres_Role Playing;Brain Games',
 'Genres_Role Playing;Pretend Play',
 'Genres_Shopping',
 'Genres_Simulation',
 'Genres_Simulation;Action & Adventure',
 'Genres_Simulation;Education',
 'Genres_Simulation;Pretend Play',
 'Genres_Social',
 'Genres_Sports',
 'Genres_Sports;Action & Adventure',
 'Genres_Strategy',
 'Genres_Strategy;Action & Adventure',
 'Genres_Strategy;Creativity',
 'Genres_Strategy;Education',
 'Genres_Tools',
 'Genres_Travel & Local',
 'Genres_Travel & Local;Action & Adventure',
 'Genres_Trivia',
 'Genres_Video Players & Editors',
 'Genres_Video Players & Editors;Creativity',
 'Genres_Video Players & Editors;Music & Video',
 'Genres_Weather',
 'Genres_Word',
 'Installs',
 'Price',
 'Rating',
 'Reviews',
 'Size',
 'Type_Free',
 'Type_Paid'}
```

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

```
from sklearn.linear_model import LinearRegression
```

```
from sklearn.model_selection import train_test_split
```

```
df_train, df_test=train_test_split(inp2, test_size=0.30,
random_state=42)
```

```
# Separate the dataframes into X_train, y_train, X_test, and y_test.
```

```
y_train=df_train.pop('Rating')
X_train= df_train
y_test= df_test.pop('Rating')
X_test= df_test
```

```
# Model building
# Use linear regression as the technique
# Report the R2 on the train set
```

```
lm=LinearRegression()
```

```
lm.fit(X_train, y_train)
```

```
LinearRegression()
```

```
from sklearn.metrics import r2_score
```

```
y_train_predict=lm.predict(X_train)
r2_score(y_train,y_train_predict)
```

```
0.16036440979501376
```

```
# Make predictions on test set and report R2.
```

```
X_test_predict=lm.predict(X_test)
r2_score(y_test,X_test_predict)
```

```
0.11710848240929339
```