In [113...
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import os
import seaborn as sns
```

In [114...
```
os.chdir("E:\\Data Science\\Python")
```

In [115...
```
df=pd.read_csv("googleplaystore.csv")
```

In [116...
```
df.isnull().sum()
```

Out[116...
```
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 [117...
```
df
```

Out[117...

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Con Ra
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Every
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Every
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Every
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	-

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Con Ra
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Every
...	...	...	...	...	...	...	...	...	
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Every
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Every
10838	Parkinson Exercices FR	MEDICAL	NaN	3	9.5M	1,000+	Free	0	Every
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	Varies with device	1,000+	Free	0	Ma
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Every

10841 rows × 13 columns

In [118...]
```
df.shape
```

Out[118...]
```
(10841, 13)
```

In [119...]
```
df1=df.dropna()
```

In [120...]
```
df1["Size"]
```

Out[120...]
```
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
```

In [121...]
```
df1=df1[-df1["Size"].str.contains("Var")]
```

In [122...
```
print(df.dtypes)
```

```
App                object
Category           object
Rating            float64
Reviews            object
Size               object
Installs           object
Type               object
Price              object
Content Rating     object
Genres             object
Last Updated       object
Current Ver        object
Android Ver        object
dtype: object
```

In [123...
```
df1["Sizenum"]=df1["Size"].str.rstrip("MKk+")
```

In [124...
```
df1["Sizenum"]=pd.to_numeric(df1["Sizenum"])
```

In [125...
```
df1["Sizenum"].dtype
```

Out[125...
```
dtype('float64')
```

In [126...
```
df1["Sizenum"]=np.where(df1["Size"].str.contains("M"),df1["Sizenum"]*1000,df1["Sizen
```

In [127...
```
df1
```

Out[127...

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Conte Rati
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyc
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyc
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyc
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Te
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyc

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Conte Rati
...	...	...	...	...	...	...	...	...	
10833	Chemin (fr)	BOOKS_AND_REFERENCE	4.8	44	619k	1,000+	Free	0	Everyc
10834	FR Calculator	FAMILY	4.0	7	2.6M	500+	Free	0	Everyc
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyc
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	0	Everyc
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Everyc

7723 rows × 14 columns

In [128...
```
df1["Size"]=df1["Sizenum"]
```

In [129...
```
df1.drop(["Sizenum"],axis=1,inplace=True)
```

In [130...
```
df1["Reviews"]=pd.to_numeric(df1["Reviews"])
```

In [131...
```
df1["Reviews"].dtype
```

Out[131...
```
dtype('int64')
```

In [132...
```
##Installs field is currently stored as string and has values like 1,000,000+. Treat
#the field, convert it to integer
```

In [133...
```
df1["Installs"]=df1["Installs"].str.replace("+","")
```

```
C:\Users\manju\AppData\Local\Temp/ipykernel_24320/2884856660.py:1: FutureWarning: Th
e default value of regex will change from True to False in a future version. In addi
tion, single character regular expressions will *not* be treated as literal strings
when regex=True.
  df1["Installs"]=df1["Installs"].str.replace("+","")
```

In [134...
```
df1["Installs"]=df1["Installs"].str.replace(",","")
```

In [135...
```
df1["Installs"]
```

Out[135…
```
0           10000
1          500000
2         5000000
3        50000000
4          100000
           ...
10833        1000
10834         500
10836        5000
10837         100
10840    10000000
Name: Installs, Length: 7723, dtype: object
```

In [136…
```
df1["Installs"]=pd.to_numeric(df1.Installs)
```

In [137…
```
df1["Installs"].dtype
```

Out[137…
```
dtype('int64')
```

In [138…
```
#Price field is a string and has $ symbol. Remove ‘$’ sign, and convert it to numeri
```

In [139…
```
df1["Price"]=df1["Price"].str.replace("$","")
```

```
C:\Users\manju\AppData\Local\Temp/ipykernel_24320/3018115500.py:1: FutureWarning: Th
e default value of regex will change from True to False in a future version. In addi
tion, single character regular expressions will *not* be treated as literal strings
when regex=True.
  df1["Price"]=df1["Price"].str.replace("$","")
```

In [140…
```
df1["Price"]=pd.to_numeric(df1.Price)
```

In [141…
```
df1["Price"]
```

Out[141…
```
0        0.0
1        0.0
2        0.0
3        0.0
4        0.0
        ...
10833    0.0
10834    0.0
10836    0.0
10837    0.0
10840    0.0
Name: Price, Length: 7723, dtype: float64
```

In [142…
```
#5. Sanity checks:

#Average rating should be between 1 and 5 as only these values are allowed on the pl


#Reviews should not be more than installs as only those who installed can review the

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

In [ ]:

In [143…
```
df1 = df1[(df1.Rating>=1) & (df1.Rating<=5)]
```

In [144…
```
df1["Rating"]
```

Out[144…
```
0        4.1
1        3.9
2        4.7
3        4.5
4        4.3
        ...
10833    4.8
10834    4.0
10836    4.5
10837    5.0
10840    4.5
Name: Rating, Length: 7723, dtype: float64
```

In [145…
```
len(df1.index)
```

Out[145…
```
7723
```

In [146…
```
df1.drop(df1.index[df1.Reviews>df1.Installs],axis=0,inplace = True)
```

In [147…
```
len(df1.index)
```

Out[147…
```
7717
```

In [148…
```
df1[(df1.Type=="Free")&(df1.Price>0)]
```

Out[148…

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

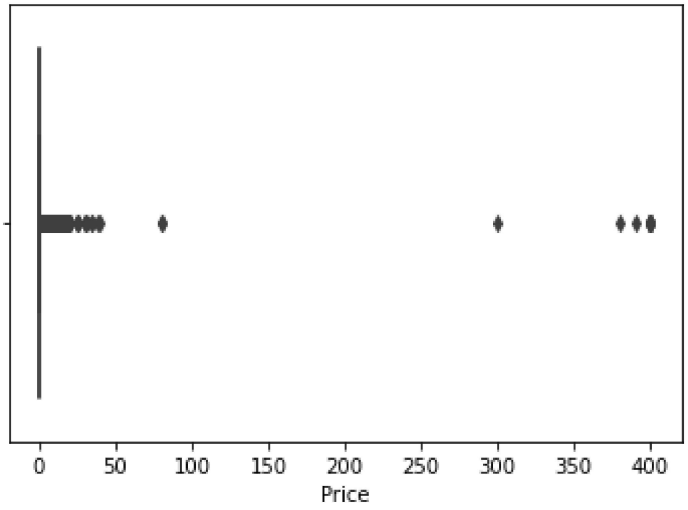
In [149…
```
#no free apps has price greater than 0
```

In [150…
```
#5. Performing univariate analysis:

#Boxplot for Price
```

In [151…
```
sns.boxplot(data=df1,x
            ="Price")
```

Out[151…
```
<AxesSubplot:xlabel='Price'>
```



In [152…
```
#more than 100 is outliers
```

In [104…
```
std=np.std(df1.Price)
```

In [105…
```
mean=np.mean(df1.Price)
```

In [106…
```
outlier=mean+3*std
```

In [107…
```
outlier
```

Out[107…
```
53.36969138940857
```

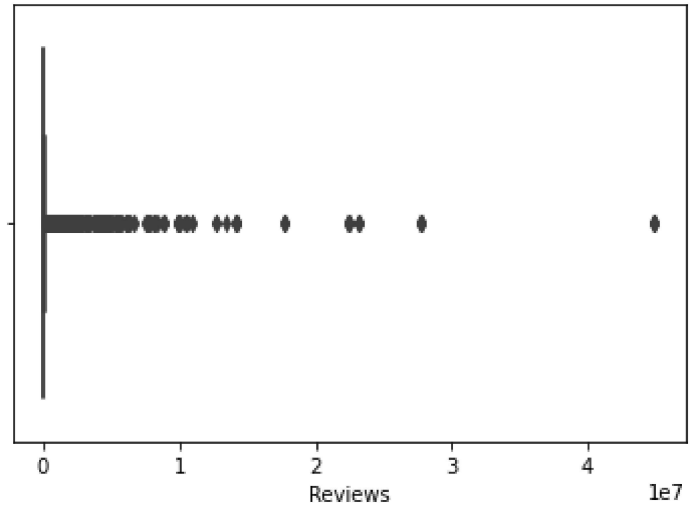
In [111…
```
len(df1[(df1.Price >outlier)])
```

Out[111…
```
17
```

In [ ]:
```
#Boxplot for Reviews -- Are there any apps with very high number of reviews? Do the
```

In [94]:
```
sns.boxplot(data=df1,x="Reviews")
```

Out[94]:
```
<AxesSubplot:xlabel='Reviews'>
```



In [154…
```
df1["Installs"].dtype
```

Out[154…
```
dtype('int64')
```

In [ ]:
```
#Find out the different percentiles - 10, 25, 50, 70, 90, 95, 99
```

In [156…
```
np.percentile(df1["Installs"],10)
```

Out[156…
```
1000.0
```

In [157…
```
np.percentile(df1["Installs"],25)
```

Out[157…
```
10000.0
```

In [158…
```
np.percentile(df1["Installs"],50)
```

Out[158…
```
100000.0
```

In [159…
```
np.percentile(df1["Installs"],70)
```

Out[159…
```
1000000.0
```

In [160…
```
np.percentile(df1["Installs"],90)
```

Out[160…
```
10000000.0
```

In [161…
```
np.percentile(df1["Installs"],95)
```

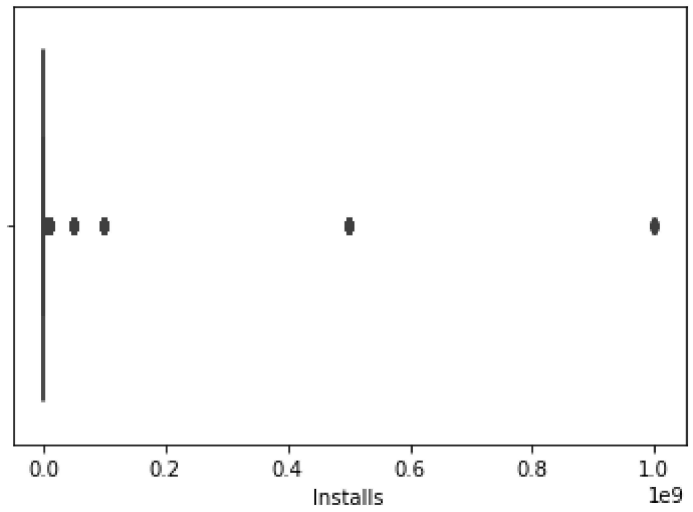
Out[161…
```
50000000.0
```

In [162…
```
np.percentile(df1["Installs"],99)
```

Out[162…
```
100000000.0
```

In [163…
```
sns.boxplot(data=df1,x="Installs")
```

Out[163…
```
<AxesSubplot:xlabel='Installs'>
```



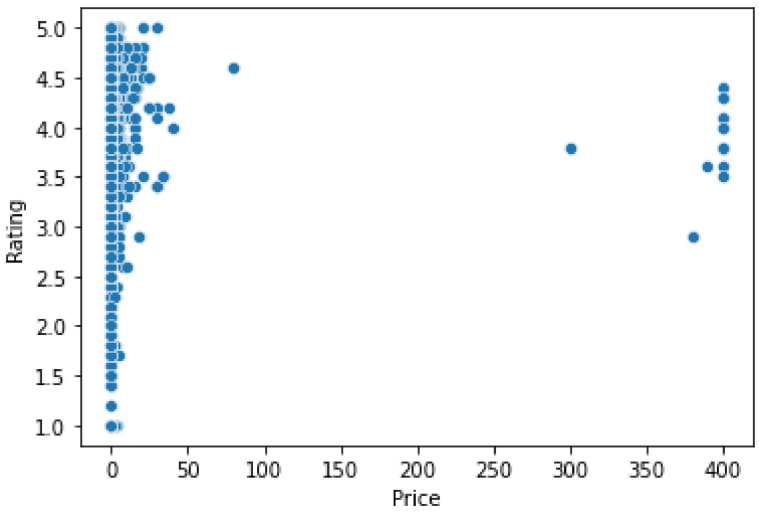
In [170…
```
len(df1[(df1.Installs>=100000000.0)])
```

Out[170…
```
241
```

In [175…
```
df1.drop(df1.index[df1.Installs>=100000000.0],inplace=True)
```

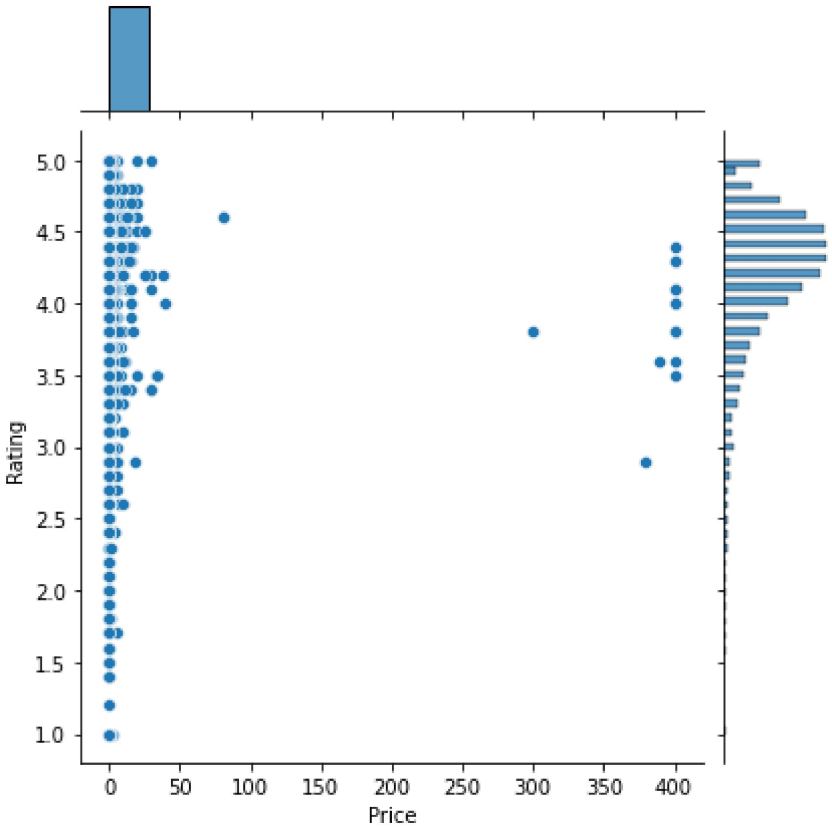
In [177…
```
#7. Bivariate analysis:
#Make scatter plot/joinplot for Rating vs. Price
#What pattern do you observe? Does rating increase with price?
sns.scatterplot(x="Price",y="Rating",data=df1)
```

Out[177…
```
<AxesSubplot:xlabel='Price', ylabel='Rating'>
```



In [178…
```
sns.jointplot(x="Price",y="Rating",data=df1)
```

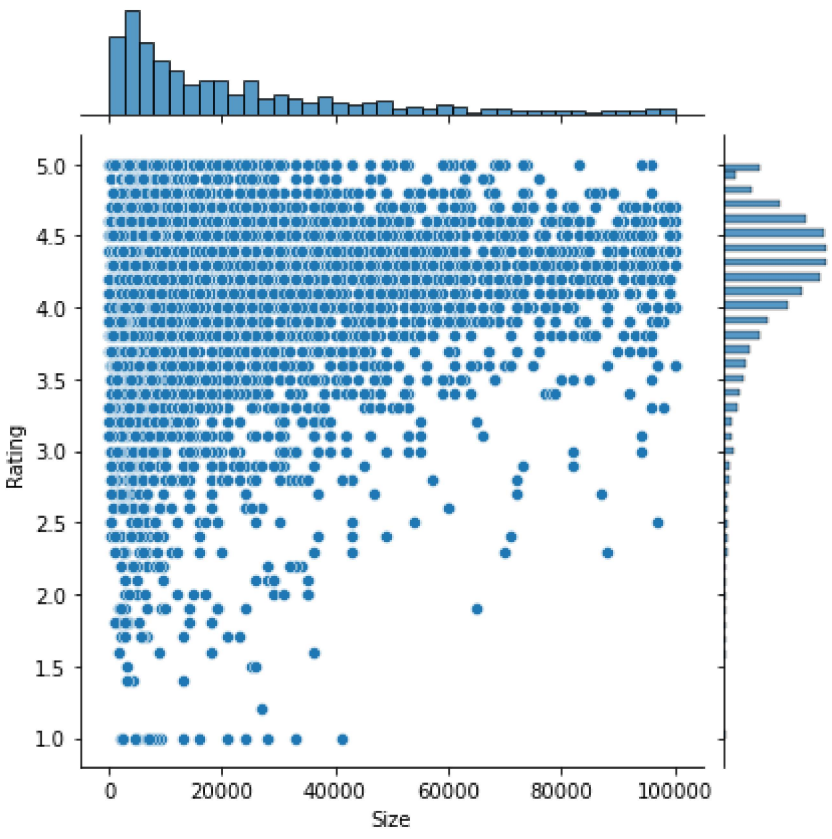
Out[178…
```
<seaborn.axisgrid.JointGrid at 0x25ed94f0bb0>
```



In [179…

```
#Make scatter plot/joinplot for Rating vs. Size
#Are heavier apps rated better?No
sns.jointplot(data=df1,x="Size",y="Rating")
```

Out[179… <seaborn.axisgrid.JointGrid at 0x25edb2d85e0>



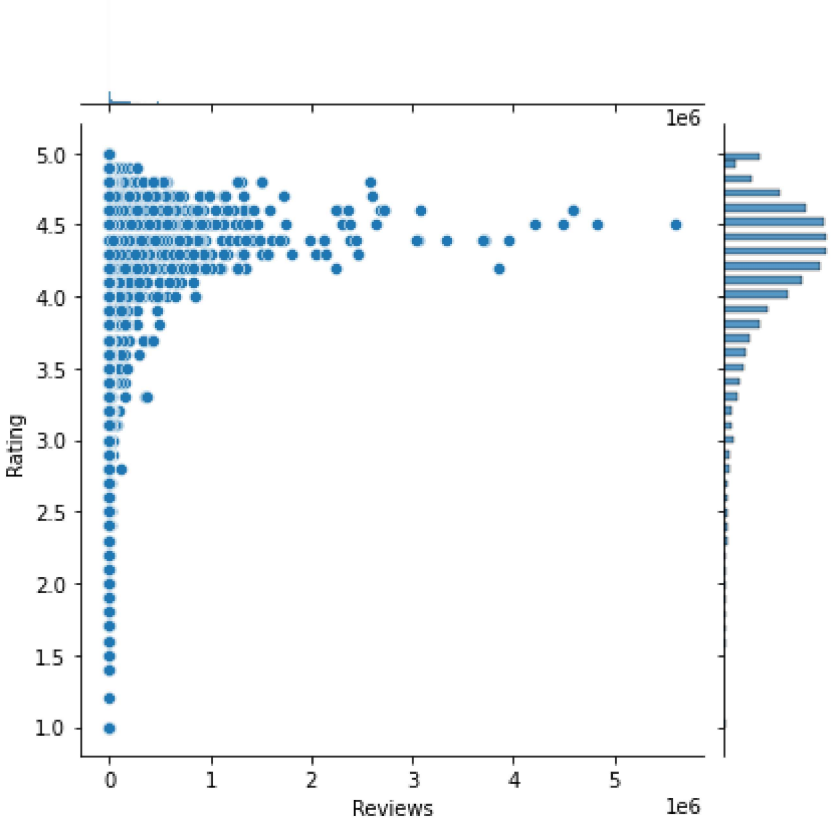
In [180…

```
#Make scatter plot/joinplot for Rating vs. Reviews
#Does more review mean a better rating always?
```

```
sns.jointplot(data=df1,x="Reviews",y="Rating")
```

Out[180…

```
<seaborn.axisgrid.JointGrid at 0x25edaeef8b0>
```

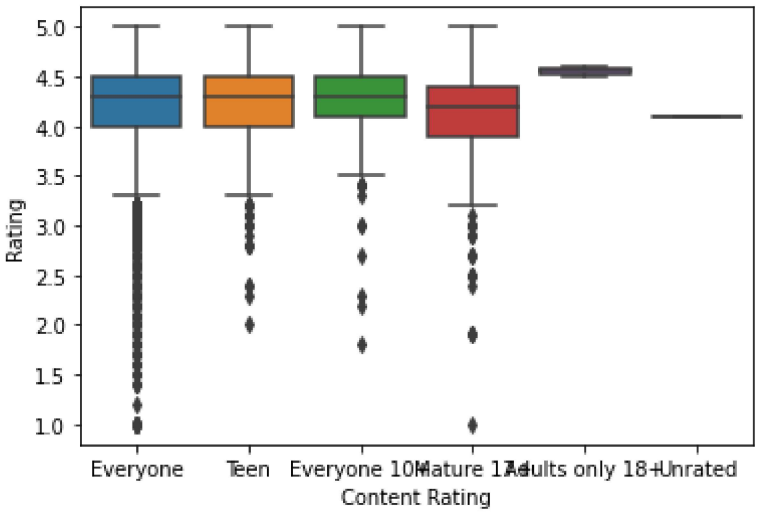


In [182…

```
#Make boxplot for Rating vs. Content Rating
#Is there any difference in the ratings? Are some types liked better?
sns.boxplot(data=df1,x="Content Rating",y="Rating")
```

Out[182…

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

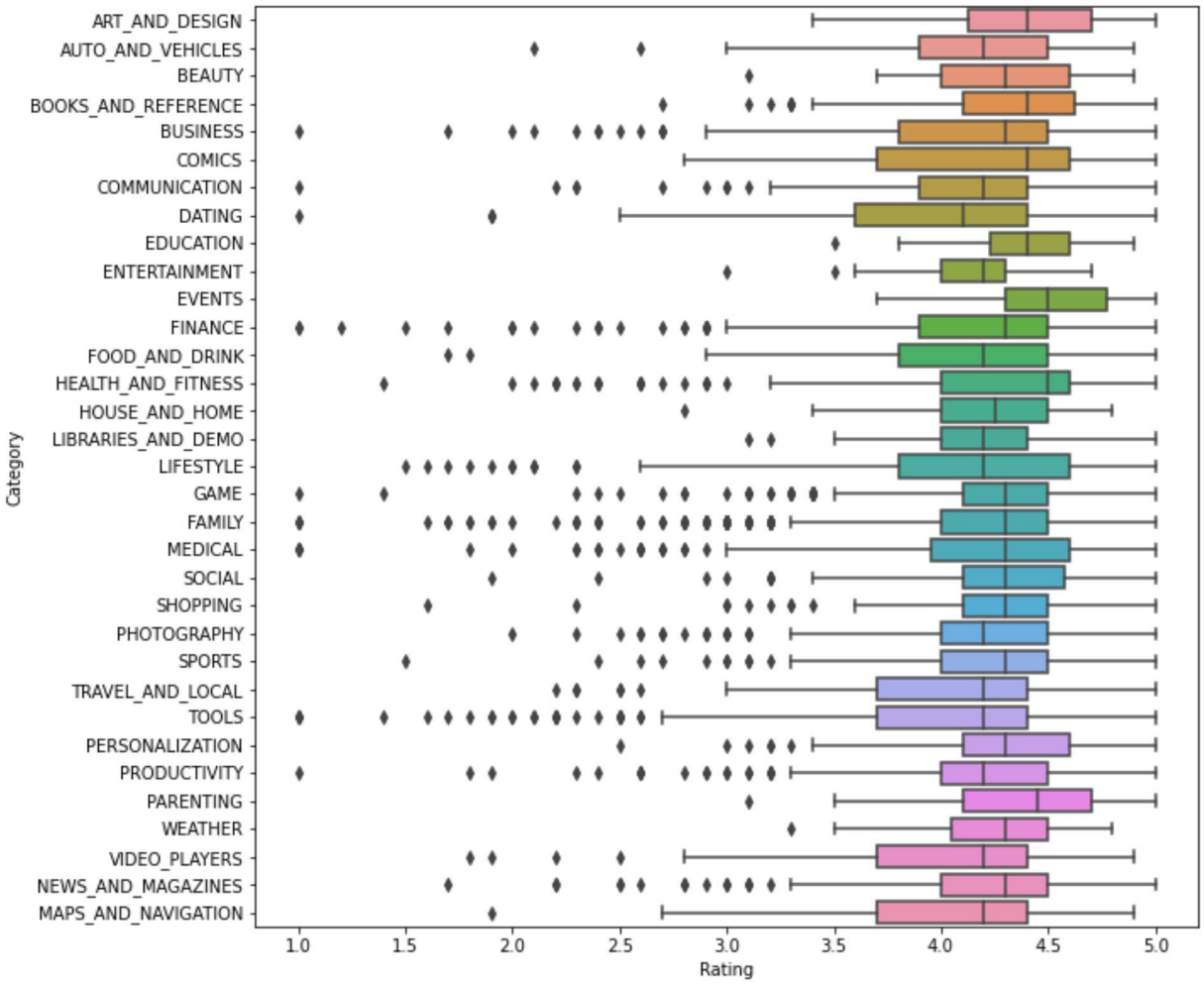


In [185…

```
#Make boxplot for Ratings vs. Category
#Which genre has the best ratings?
plt.figure(figsize=(10,10))
sns.boxplot(data=df1,x="Rating",y="Category")
```

Out[185…

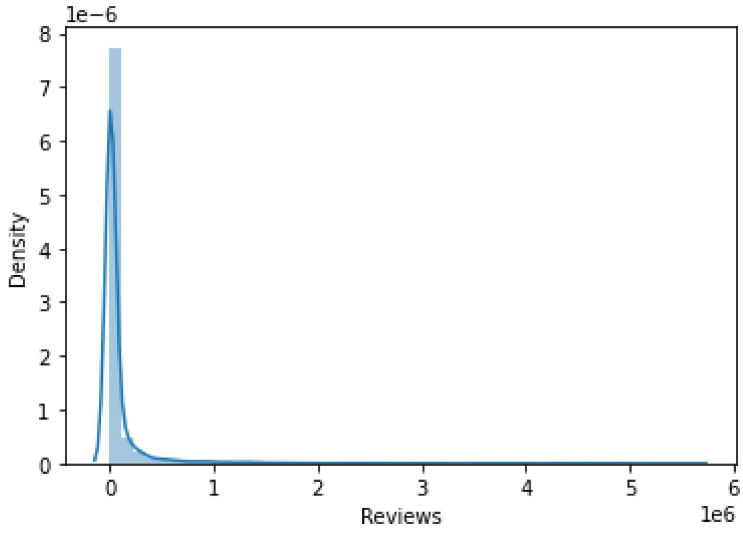
```
<AxesSubplot:xlabel='Rating', ylabel='Category'>
```



In [198…
```
sns.distplot(df1.Reviews)
```

```
D:\Softwares\New folder\lib\site-packages\seaborn\distributions.py:2619: FutureWarni
ng: `distplot` is a deprecated function and will be removed in a future version. Ple
ase adapt your code to use either `displot` (a figure-level function with similar fl
exibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[198…
```
<AxesSubplot:xlabel='Reviews', ylabel='Density'>
```



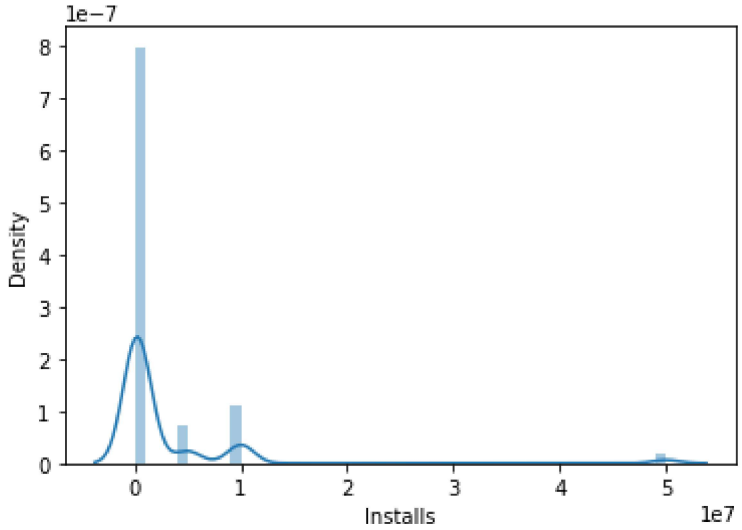
In [199…
```
sns.distplot(df1.Installs)
```

```
D:\Softwares\New folder\lib\site-packages\seaborn\distributions.py:2619: FutureWarni
```

```
ng: `distplot` is a deprecated function and will be removed in a future version. Ple
ase adapt your code to use either `displot` (a figure-level function with similar fl
exibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[199…

```
<AxesSubplot:xlabel='Installs', ylabel='Density'>
```



In [219…

Out[219…

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Conte Rati
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.0	10000	Free	0.0	Everyc
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.0	500000	Free	0.0	Everyc
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.0	5000000	Free	0.0	Everyc
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.0	50000000	Free	0.0	Te
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.0	100000	Free	0.0	Everyc
...	...	...	...	...	...	...	...	...	
10833	Chemin (fr)	BOOKS_AND_REFERENCE	4.8	44	619.0	1000	Free	0.0	Everyc
10834	FR Calculator	FAMILY	4.0	7	2600.0	500	Free	0.0	Everyc

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Conte Rati
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53000.0	5000	Free	0.0	Everyc
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.0	100	Free	0.0	Everyc
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000.0	10000000	Free	0.0	Everyc

7476 rows × 13 columns

In [212...

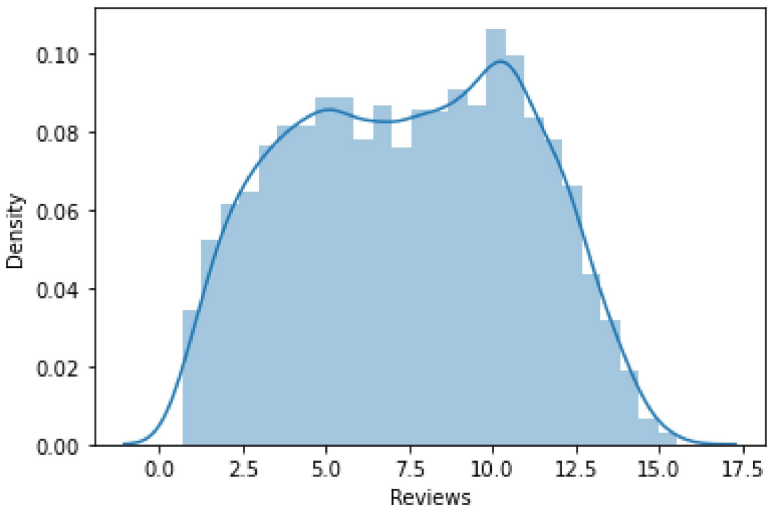
In [213...

In [214...

```
D:\Softwares\New folder\lib\site-packages\seaborn\distributions.py:2619: FutureWarni
ng: `distplot` is a deprecated function and will be removed in a future version. Ple
ase adapt your code to use either `displot` (a figure-level function with similar fl
exibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[214...

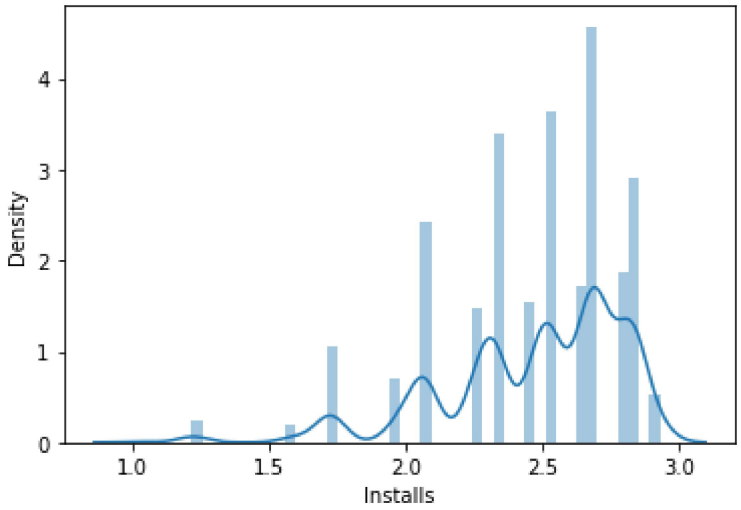
```
<AxesSubplot:xlabel='Reviews', ylabel='Density'>
```



In [217...

```
D:\Softwares\New folder\lib\site-packages\seaborn\distributions.py:2619: FutureWarni
ng: `distplot` is a deprecated function and will be removed in a future version. Ple
ase adapt your code to use either `displot` (a figure-level function with similar fl
exibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[217… <AxesSubplot:xlabel='Installs', ylabel='Density'>



In [ ]:

In [220…
```
inp1=df2.copy()
```

In [221…
```
inp1
```

Out[221…

	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	
0	ART_AND_DESIGN	4.1	5.075174	19000.0	2.323411	Free	0.0	Everyone	Art
1	ART_AND_DESIGN	3.9	6.875232	14000.0	2.647760	Free	0.0	Everyone	Desig
2	ART_AND_DESIGN	4.7	11.379520	8700.0	2.798801	Free	0.0	Everyone	Art
3	ART_AND_DESIGN	4.5	12.281389	25000.0	2.929995	Free	0.0	Teen	Art
4	ART_AND_DESIGN	4.3	6.875232	2800.0	2.526763	Free	0.0	Everyone	Design;
…	…	…	…	…	…	…	…	…	
10833	BOOKS_AND_REFERENCE	4.8	3.806662	619.0	2.067970	Free	0.0	Everyone	
10834	FAMILY	4.0	2.079442	2600.0	1.976385	Free	0.0	Everyone	E
10836	FAMILY	4.5	3.663562	53000.0	2.253121	Free	0.0	Everyone	E
10837	FAMILY	5.0	1.609438	3600.0	1.725463	Free	0.0	Everyone	E
10840	LIFESTYLE	4.5	12.894981	19000.0	2.840136	Free	0.0	Everyone	

7476 rows × 9 columns

In [ ]:
```
Reviews and Install have some values that are still relatively very high. Before bui
Installs.
```

In [ ]:

```
# Data preprocessing-For the steps below, create a copy of the dataframe to make all
#Reviews and Install have some values that are still relatively very high. Before bu
```

In [222…
```
inp1.Reviews=inp1.Reviews.apply(np.log1p)
```

In [223…
```
inp1.Installs=inp1.Installs.apply(np.log1p)
```

In [224…
```
inp1.head(2)
```

Out[224…

	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
0	ART_AND_DESIGN	4.1	1.804211	19000.0	1.200992	Free	0.0	Everyone	Art & Design
1	ART_AND_DESIGN	3.9	2.063723	14000.0	1.294113	Free	0.0	Everyone	Art & Design;Pretend Play

In [ ]:
```
#Drop columns App, Last Updated, Current Ver, and Android Ver. These variables are n
inp1.drop(["App","Last Updated","Current Ver","Android Ver"],axis=1,inplace=True)
```

In [226…
```
#Get dummy columns for Category, Genres, and Content Rating. This needs to be done a
#character fields to numeric. Name of dataframe should be inp2.
inp2=pd.get_dummies(inp1)
```

In [227…
```
inp2.head(2)
```

Out[227…

	Rating	Reviews	Size	Installs	Price	Category_ART_AND_DESIGN	Category_AUTO_AND_VEH
0	4.1	1.804211	19000.0	1.200992	0.0	1	
1	3.9	2.063723	14000.0	1.294113	0.0	1	

2 rows × 158 columns

In [229…
```
inp2.shape
```

Out[229…
```
(7476, 158)
```

In [232…
```
y=inp2.iloc[:,0]
X=inp2.iloc[:,1:]
```

In [233…
```
from sklearn.model_selection import train_test_split
```

In [234…
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

In [237…
```
from sklearn.linear_model import LinearRegression
```

In [238…
```
lr=LinearRegression()
```

In [239…
```
lr.fit(X_train,y_train)
```

Out[239… LinearRegression()

In [240…
```
predict=lr.predict(X_test)
```

In [244…
```
from sklearn.metrics import r2_score,accuracy_score
```

In [242…
```
r2=r2_score(y_test,predict)
```

r2

In [243…
```
r2
```

Out[243… 0.11794017047329708

In [ ]:
```
accuracy_score()
```