Google_playstore_project

November 12, 2021

1 Google Play store Analysis

Objective: Make a model to predict the app rating, with other information about the app provided

[349]:
```
#Import the libraries required
import pandas as pd
import numpy as np
```

[350]:
```
#1. Load the data file using pandas
df = pd.read_csv('googleplaystore.csv')
df.head()
```

[350]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide …	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	

```
3    4.2 and up
4    4.4 and up
```

[351]:
```
#2.Check for null values in the data. Get the number of null values for each␣
↪column
df.isna().sum()
```

[351]:
```
App                0
Category           0
Rating          1474
Reviews            0
Size               0
Installs           0
Type               1
Price              0
Content Rating     1
Genres             0
Last Updated       0
Current Ver        8
Android Ver        3
dtype: int64
```

[352]:
```
df.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
```

[353]:
```
df.describe()
```

[353]:
```
             Rating
count  9367.000000
mean      4.193338
std       0.537431
min       1.000000
25%       4.000000
50%       4.300000
75%       4.500000
max      19.000000
```

[354]:
```
#3. Drop records(rows) with nulls in any of the columns and name the dataframe␣
↪as df_new
df_new = df.dropna(axis=0, how='any')
```

[355]:
```
df_new.head()
```

[355]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite - FREE Live Cool Themes, Hide …	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

[356]:
```
df_new.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   object
 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(1), object(12)
memory usage: 1023.8+ KB
```

[357]:
```
df_new.isna().sum()
```

[357]:
```
App               0
Category          0
Rating            0
Reviews           0
Size              0
Installs          0
Type              0
Price             0
Content Rating    0
Genres            0
Last Updated      0
Current Ver       0
Android Ver       0
dtype: int64
```

[358]:
```
#4. Correcting incorrect type and inconsistent formatting
#4.1. Changing Size to Kb and numeric data
df_new.Size.unique()
```

[358]:
```
array(['19M', '14M', '8.7M', '25M', '2.8M', '5.6M', '29M', '33M', '3.1M',
       '28M', '12M', '20M', '21M', '37M', '5.5M', '17M', '39M', '31M',
       '4.2M', '23M', '6.0M', '6.1M', '4.6M', '9.2M', '5.2M', '11M',
       '24M', 'Varies with device', '9.4M', '15M', '10M', '1.2M', '26M',
       '8.0M', '7.9M', '56M', '57M', '35M', '54M', '201k', '3.6M', '5.7M',
       '8.6M', '2.4M', '27M', '2.7M', '2.5M', '7.0M', '16M', '3.4M',
```

```
       '8.9M', '3.9M', '2.9M', '38M', '32M', '5.4M', '18M', '1.1M',
       '2.2M', '4.5M', '9.8M', '52M', '9.0M', '6.7M', '30M', '2.6M',
       '7.1M', '22M', '6.4M', '3.2M', '8.2M', '4.9M', '9.5M', '5.0M',
       '5.9M', '13M', '73M', '6.8M', '3.5M', '4.0M', '2.3M', '2.1M',
       '42M', '9.1M', '55M', '23k', '7.3M', '6.5M', '1.5M', '7.5M', '51M',
       '41M', '48M', '8.5M', '46M', '8.3M', '4.3M', '4.7M', '3.3M', '40M',
       '7.8M', '8.8M', '6.6M', '5.1M', '61M', '66M', '79k', '8.4M',
       '3.7M', '118k', '44M', '695k', '1.6M', '6.2M', '53M', '1.4M',
       '3.0M', '7.2M', '5.8M', '3.8M', '9.6M', '45M', '63M', '49M', '77M',
       '4.4M', '70M', '9.3M', '8.1M', '36M', '6.9M', '7.4M', '84M', '97M',
       '2.0M', '1.9M', '1.8M', '5.3M', '47M', '556k', '526k', '76M',
       '7.6M', '59M', '9.7M', '78M', '72M', '43M', '7.7M', '6.3M', '334k',
       '93M', '65M', '79M', '100M', '58M', '50M', '68M', '64M', '34M',
       '67M', '60M', '94M', '9.9M', '232k', '99M', '624k', '95M', '8.5k',
       '41k', '292k', '80M', '1.7M', '10.0M', '74M', '62M', '69M', '75M',
       '98M', '85M', '82M', '96M', '87M', '71M', '86M', '91M', '81M',
       '92M', '83M', '88M', '704k', '862k', '899k', '378k', '4.8M',
       '266k', '375k', '1.3M', '975k', '980k', '4.1M', '89M', '696k',
       '544k', '525k', '920k', '779k', '853k', '720k', '713k', '772k',
       '318k', '58k', '241k', '196k', '857k', '51k', '953k', '865k',
       '251k', '930k', '540k', '313k', '746k', '203k', '26k', '314k',
       '239k', '371k', '220k', '730k', '756k', '91k', '293k', '17k',
       '74k', '14k', '317k', '78k', '924k', '818k', '81k', '939k', '169k',
       '45k', '965k', '90M', '545k', '61k', '283k', '655k', '714k', '93k',
       '872k', '121k', '322k', '976k', '206k', '954k', '444k', '717k',
       '210k', '609k', '308k', '306k', '175k', '350k', '383k', '454k',
       '1.0M', '70k', '812k', '442k', '842k', '417k', '412k', '459k',
       '478k', '335k', '782k', '721k', '430k', '429k', '192k', '460k',
       '728k', '496k', '816k', '414k', '506k', '887k', '613k', '778k',
       '683k', '592k', '186k', '840k', '647k', '373k', '437k', '598k',
       '716k', '585k', '982k', '219k', '55k', '323k', '691k', '511k',
       '951k', '963k', '25k', '554k', '351k', '27k', '82k', '208k',
       '551k', '29k', '103k', '116k', '153k', '209k', '499k', '173k',
       '597k', '809k', '122k', '411k', '400k', '801k', '787k', '50k',
       '643k', '986k', '516k', '837k', '780k', '20k', '498k', '600k',
       '656k', '221k', '228k', '176k', '34k', '259k', '164k', '458k',
       '629k', '28k', '288k', '775k', '785k', '636k', '916k', '994k',
       '309k', '485k', '914k', '903k', '608k', '500k', '54k', '562k',
       '847k', '948k', '811k', '270k', '48k', '523k', '784k', '280k',
       '24k', '892k', '154k', '18k', '33k', '860k', '364k', '387k',
       '626k', '161k', '879k', '39k', '170k', '141k', '160k', '144k',
       '143k', '190k', '376k', '193k', '473k', '246k', '73k', '253k',
       '957k', '420k', '72k', '404k', '470k', '226k', '240k', '89k',
       '234k', '257k', '861k', '467k', '676k', '552k', '582k', '619k'],
      dtype=object)
```

[359]:
```
df_new.Size.loc[(df_new.Size.str.contains('Varies with device'))].value_counts()
```

[359]:
```
Varies with device    1637
Name: Size, dtype: int64
```

[360]:
```
df_new.Size.loc[(df_new.Size.str.endswith('+'))].value_counts()
```

[360]:
```
Series([], Name: Size, dtype: int64)
```

[361]:
```
#Taking a copy of the size column to make changes
df_new_copy = df_new['Size'].copy(deep=True)
df_new_copy.head()
```

[361]:
```
0     19M
1     14M
2    8.7M
3     25M
4    2.8M
Name: Size, dtype: object
```

[362]:
```
#Finding the indexes of rows ending with M, k, varies with device
size_M = df_new_copy.str.endswith('M')
size_k = df_new_copy.str.endswith('k')
size_text = df_new_copy.str.match('Varies with device')
```

[363]:
```
#if the records ends with M => multiply by 1000, ends with k => pick numeric␣
↪data having 'Varies with device' => fill with NA
df_new_copy[size_M] = (df_new_copy[size_M].apply(lambda x: float(x[:-1]) *␣
↪1000))
df_new_copy[size_k] = (df_new_copy[size_k].apply(lambda x: float(x[:-1])))
df_new_copy[size_text] = np.nan
```

[364]:
```
df_new_copy.head()
```

[364]:
```
0    19000
1    14000
2     8700
3    25000
4     2800
Name: Size, dtype: object
```

[365]:
```
df_new_copy.values
```

[365]:
```
array([19000.0, 14000.0, 8700.0, …, 3600.0, nan, 19000.0], dtype=object)
```

[366]:
```
#Copying the dataframe to a new variable 'playstore' and performing the␣
↪following steps to avoid 'Settingwithcopywarning'
playstore = df_new.copy()
```

[367]:
```
result = df_new_copy.copy()
```

[368]:
```
playstore.loc[:,'Size'] = result
```

[369]:
```
playstore['Size'].astype(float) #changing the datatype from 'object' to 'float'
```

[369]:
```
0        19000.0
1        14000.0
2         8700.0
3        25000.0
4         2800.0
          ...
10834     2600.0
10836    53000.0
10837     3600.0
10839        NaN
10840    19000.0
Name: Size, Length: 9360, dtype: float64
```

[370]:
```
#Checking how many 'NA' values in Size column
playstore['Size'].isna().sum()
```

[370]:
```
1637
```

[371]:
```
#Replacing 'NA' values in size column with mean of the values in Size column
playstore['Size'].fillna(playstore['Size'].mean(),inplace=True)
```

[372]:
```
#Changing the datatype to 'integer'
playstore['Size']=playstore['Size'].astype(int)
```

[373]:
```
playstore['Size']
```

[373]:
```
0        19000
1        14000
2         8700
3        25000
4         2800
         ...
10834     2600
10836    53000
10837     3600
10839    22970
10840    19000
Name: Size, Length: 9360, dtype: int64
```

[374]:
```
playstore['Size'].isna().sum()
```

[374]:
```
0
```

[375]:
```
#4.2. converting Reviews column from string to numeric field
playstore['Reviews'] = playstore['Reviews'].astype(int)
```

[376]:
```
playstore['Reviews']
```

[376]:
```
0           159
1           967
2         87510
3        215644
4           967
          …
10834         7
10836        38
10837         4
10839       114
10840    398307
Name: Reviews, Length: 9360, dtype: int64
```

[377]:
```
#4.3.Changing Installs field and removing + and ,
playstore['Installs'].head()
```

[377]:
```
0        10,000+
1       500,000+
2     5,000,000+
3    50,000,000+
4       100,000+
Name: Installs, dtype: object
```

[378]:
```
playstore.Installs.unique()
```

[378]:
```
array(['10,000+', '500,000+', '5,000,000+', '50,000,000+', '100,000+',
       '50,000+', '1,000,000+', '10,000,000+', '5,000+', '100,000,000+',
       '1,000,000,000+', '1,000+', '500,000,000+', '100+', '500+', '10+',
       '5+', '50+', '1+'], dtype=object)
```

[379]:
```
playstore['Installs']=playstore.Installs.apply(lambda x: x.strip('+'))
playstore['Installs']=playstore.Installs.apply(lambda x: x.replace(',',''))
playstore['Installs']=playstore.Installs.replace('Free',np.nan)
playstore['Installs'].head()
```

[379]:
```
0       10000
1      500000
2     5000000
3    50000000
4      100000
```

```
Name: Installs, dtype: object
```

[380]:
```
#converting datatype of 'Installs' to integer type
playstore['Installs'] = playstore['Installs'].astype(int)
playstore.Installs.head()
```

[380]:
```
0       10000
1      500000
2     5000000
3    50000000
4      100000
Name: Installs, dtype: int64
```

[381]:
```
#4.4 Removing '$' sign from 'Price' field and changing datatype to 'integer'
playstore.Price.unique()
```

[381]:
```
array(['0', '$4.99', '$3.99', '$6.99', '$7.99', '$5.99', '$2.99', '$3.49',
       '$1.99', '$9.99', '$7.49', '$0.99', '$9.00', '$5.49', '$10.00',
       '$24.99', '$11.99', '$79.99', '$16.99', '$14.99', '$29.99',
       '$12.99', '$2.49', '$10.99', '$1.50', '$19.99', '$15.99', '$33.99',
       '$39.99', '$3.95', '$4.49', '$1.70', '$8.99', '$1.49', '$3.88',
       '$399.99', '$17.99', '$400.00', '$3.02', '$1.76', '$4.84', '$4.77',
       '$1.61', '$2.50', '$1.59', '$6.49', '$1.29', '$299.99', '$379.99',
       '$37.99', '$18.99', '$389.99', '$8.49', '$1.75', '$14.00', '$2.00',
       '$3.08', '$2.59', '$19.40', '$3.90', '$4.59', '$15.46', '$3.04',
       '$13.99', '$4.29', '$3.28', '$4.60', '$1.00', '$2.95', '$2.90',
       '$1.97', '$2.56', '$1.20'], dtype=object)
```

[382]:
```
playstore['Price']=playstore.Price.apply(lambda x: x.strip('$'))
playstore.Price.unique()
```

[382]:
```
array(['0', '4.99', '3.99', '6.99', '7.99', '5.99', '2.99', '3.49',
       '1.99', '9.99', '7.49', '0.99', '9.00', '5.49', '10.00', '24.99',
       '11.99', '79.99', '16.99', '14.99', '29.99', '12.99', '2.49',
       '10.99', '1.50', '19.99', '15.99', '33.99', '39.99', '3.95',
       '4.49', '1.70', '8.99', '1.49', '3.88', '399.99', '17.99',
       '400.00', '3.02', '1.76', '4.84', '4.77', '1.61', '2.50', '1.59',
       '6.49', '1.29', '299.99', '379.99', '37.99', '18.99', '389.99',
       '8.49', '1.75', '14.00', '2.00', '3.08', '2.59', '19.40', '3.90',
       '4.59', '15.46', '3.04', '13.99', '4.29', '3.28', '4.60', '1.00',
       '2.95', '2.90', '1.97', '2.56', '1.20'], dtype=object)
```

[383]:
```
#converting to numeric type
playstore['Price'] = playstore['Price'].astype(float)
playstore.Price.unique()
```

[383]:
```
array([  0.   ,   4.99,   3.99,   6.99,   7.99,   5.99,   2.99,   3.49,
         1.99,   9.99,   7.49,   0.99,   9.   ,   5.49,  10.   ,  24.99,
        11.99,  79.99,  16.99,  14.99,  29.99,  12.99,   2.49,  10.99,
         1.5 ,  19.99,  15.99,  33.99,  39.99,   3.95,   4.49,   1.7 ,
         8.99,   1.49,   3.88, 399.99,  17.99, 400.   ,   3.02,   1.76,
         4.84,   4.77,   1.61,   2.5 ,   1.59,   6.49,   1.29, 299.99,
       379.99,  37.99,  18.99, 389.99,   8.49,   1.75,  14.   ,   2.   ,
         3.08,   2.59,  19.4 ,   3.9 ,   4.59,  15.46,   3.04,  13.99,
         4.29,   3.28,   4.6 ,   1.   ,   2.95,   2.9 ,   1.97,   2.56,
         1.2 ])
```

[384]:
```
#5.1.Average rating should be between 1 and 5
playstore.head()
```

[384]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000	10000	Free	0.0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000	500000	Free	0.0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite - FREE Live Cool Themes, Hide …	ART_AND_DESIGN	4.7	87510	8700	5000000	Free	0.0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000	50000000	Free	0.0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

[385]:
```
playstore.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   int64
 4   Size            9360 non-null   int64
 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(2), int64(3), object(8)
memory usage: 1023.8+ KB
```

[386]:
```
#There is no record having rating outside the limits of 1 to 5
playstore[((playstore['Rating'] < 1) & (playstore['Rating'] > 5))]
```

[386]:
```
Empty DataFrame
Columns: [App, Category, Rating, Reviews, Size, Installs, Type, Price, Content
Rating, Genres, Last Updated, Current Ver, Android Ver]
Index: []
```

[387]:
```
playstore[((playstore['Rating'] >= 1) & (playstore['Rating'] <= 5))]
```

[387]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000	10000	Free	0.0	Everyone
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000	500000	Free	0.0	Everyone
2	U Launcher Lite - FREE Live Cool Themes, Hide …	ART_AND_DESIGN	4.7	87510	8700	5000000	Free	0.0	Everyone
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000	50000000	Free	0.0	Teen
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800	100000	Free	0.0	Everyone
…	…	…							
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							

```
…       …      …     …       …   …    …            …
10834     4.0      7   2600      500  Free    0.0       Everyone
10836     4.5     38  53000     5000  Free    0.0       Everyone
10837     5.0      4   3600      100  Free    0.0       Everyone
10839     4.5    114  22970     1000  Free    0.0     Mature 17+
10840     4.5 398307  19000 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
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]
```

[388]:
```
#5.2. review should not be more than installs
playstore[playstore.Reviews > playstore.Installs]
```

[388]:
```
                                App   Category  Rating  Reviews   Size  \
2454             KBA-EZ Health Guide    MEDICAL     5.0        4  25000
4663   Alarmy (Sleep If U Can) - Pro  LIFESTYLE     4.8    10249  22970
5917                          Ra Ga Ba       GAME     5.0        2  20000
6700                Brick Breaker BR       GAME     5.0        7  19000
7402            Trovami se ci riesci       GAME     5.0       11   6100
8591                         DN Blog     SOCIAL     5.0       20   4200
10697                        Mu.F.O.       GAME     5.0        2  16000
```

```
       Installs  Type  Price Content Rating     Genres      Last Updated  \
2454          1  Free   0.00       Everyone    Medical    August 2, 2018
4663      10000  Paid   2.49       Everyone  Lifestyle     July 30, 2018
5917          1  Paid   1.49       Everyone     Arcade  February 8, 2017
6700          5  Free   0.00       Everyone     Arcade     July 23, 2018
7402         10  Free   0.00       Everyone     Arcade    March 11, 2017
8591         10  Free   0.00           Teen     Social     July 23, 2018
10697         1  Paid   0.99       Everyone     Arcade     March 3, 2017

              Current Ver         Android Ver
2454                1.0.72        4.0.3 and up
4663   Varies with device  Varies with device
5917                1.0.4          2.3 and up
6700                  1.0          4.1 and up
7402                  0.1          2.3 and up
8591                  1.0          4.0 and up
10697                 1.0          2.3 and up
```

[389]:
```
#There are seven rows with Reviews more than Installs and hence dropping those␣
 ↪rows
playstore.drop(playstore[(playstore.Reviews > playstore.Installs)].index,␣
 ↪inplace=True)
playstore.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9353 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             9353 non-null   object
 1   Category        9353 non-null   object
 2   Rating          9353 non-null   float64
 3   Reviews         9353 non-null   int64
 4   Size            9353 non-null   int64
 5   Installs        9353 non-null   int64
 6   Type            9353 non-null   object
 7   Price           9353 non-null   float64
 8   Content Rating  9353 non-null   object
 9   Genres          9353 non-null   object
 10  Last Updated    9353 non-null   object
 11  Current Ver     9353 non-null   object
 12  Android Ver     9353 non-null   object
dtypes: float64(2), int64(3), object(8)
memory usage: 1023.0+ KB
```

[390]:
```
#5.3.price should not greater than zero for free apps
#There is no record having price > 0 with type = 'Free'
```

```
playstore[(playstore.Type == 'Free') &  (playstore.Price > 0)]
```

[390]:
```
Empty DataFrame
Columns: [App, Category, Rating, Reviews, Size, Installs, Type, Price, Content
Rating, Genres, Last Updated, Current Ver, Android Ver]
Index: []
```

[391]:
```
playstore[(playstore.Type == 'Free') &  (playstore.Price == 0)]
```

[391]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000	10000	Free	0.0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000	500000	Free	0.0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide …	ART_AND_DESIGN	4.7	87510	8700	5000000	Free	0.0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000	50000000	Free	0.0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
…	…	…	…	…	…	…	…	…	…	…	…	…	…
10834	FR Calculator	FAMILY	4.0	7	2600	500	Free	0.0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53000	5000	Free	0.0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600	100	Free	0.0	Everyone	Education	July 6, 2018	1.0	4.1 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	22970	1000	Free	0.0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000	10000000	Free	0.0	Everyone	Lifestyle	July 25, 2018	Varies with device	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

[8711 rows x 13 columns]
```

[392]:
```
playstore.Type.unique()
```

[392]:
```
array(['Free', 'Paid'], dtype=object)
```

[393]:
```
playstore[(playstore.Type == 'Paid')]
```

[393]:
```
                                                App             Category  \
234    TurboScan: scan documents and receipts in PDF             BUSINESS
235                    Tiny Scanner Pro: PDF Doc Scan             BUSINESS
290    TurboScan: scan documents and receipts in PDF             BUSINESS
291                    Tiny Scanner Pro: PDF Doc Scan             BUSINESS
427                                Puffin Browser Pro        COMMUNICATION
…                                                 …                    …
10682                          Fruit Ninja Classic                 GAME
10690                                      FO Bixby      PERSONALIZATION
10760                                 Fast Tract Diet  HEALTH_AND_FITNESS
10782                          Trine 2: Complete Story                 GAME
10785                                      sugar, sugar               FAMILY

       Rating  Reviews   Size  Installs  Type  Price Content Rating  \
234       4.7    11442   6800    100000  Paid   4.99       Everyone
235       4.8    10295  39000    100000  Paid   4.99       Everyone
290       4.7    11442   6800    100000  Paid   4.99       Everyone
291       4.8    10295  39000    100000  Paid   4.99       Everyone
427       4.0    18247  22970    100000  Paid   3.99       Everyone
…         …      …      …        …       …      …                …
10682     4.3    85468  36000   1000000  Paid   0.99       Everyone
10690     5.0        5    861       100  Paid   0.99       Everyone
10760     4.4       35   2400      1000  Paid   7.99       Everyone
10782     3.8      252  11000     10000  Paid  16.99           Teen
10785     4.2     1405   9500     10000  Paid   1.20       Everyone
```

```
                 Genres       Last Updated   Current Ver   Android Ver
234            Business     March 25, 2018         1.5.2    4.0 and up
235            Business     April 11, 2017         3.4.6    3.0 and up
290            Business     March 25, 2018         1.5.2    4.0 and up
291            Business     April 11, 2017         3.4.6    3.0 and up
427       Communication       July 5, 2018   7.5.3.20547    4.1 and up
…                   …                  …             …             …
10682           Arcade       June 8, 2018  2.4.1.485300  4.0.3 and up
10690   Personalization     April 25, 2018           0.2    7.0 and up
10760  Health & Fitness     August 8, 2018         1.9.3    4.2 and up
10782           Action  February 27, 2015          2.22    5.0 and up
10785           Puzzle       June 5, 2018           2.7    2.3 and up

[642 rows x 13 columns]
```

[394]:
```
playstore.shape
```

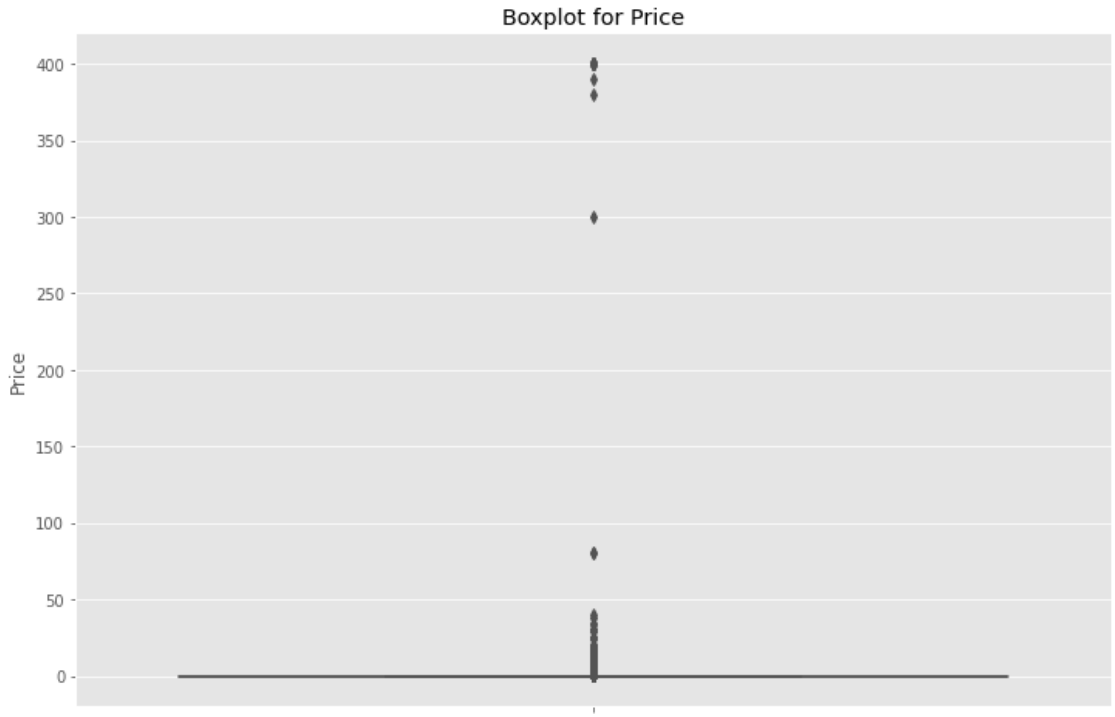
[394]:
```
(9353, 13)
```

[395]:
```
#6. Performing Univariate analysis
```

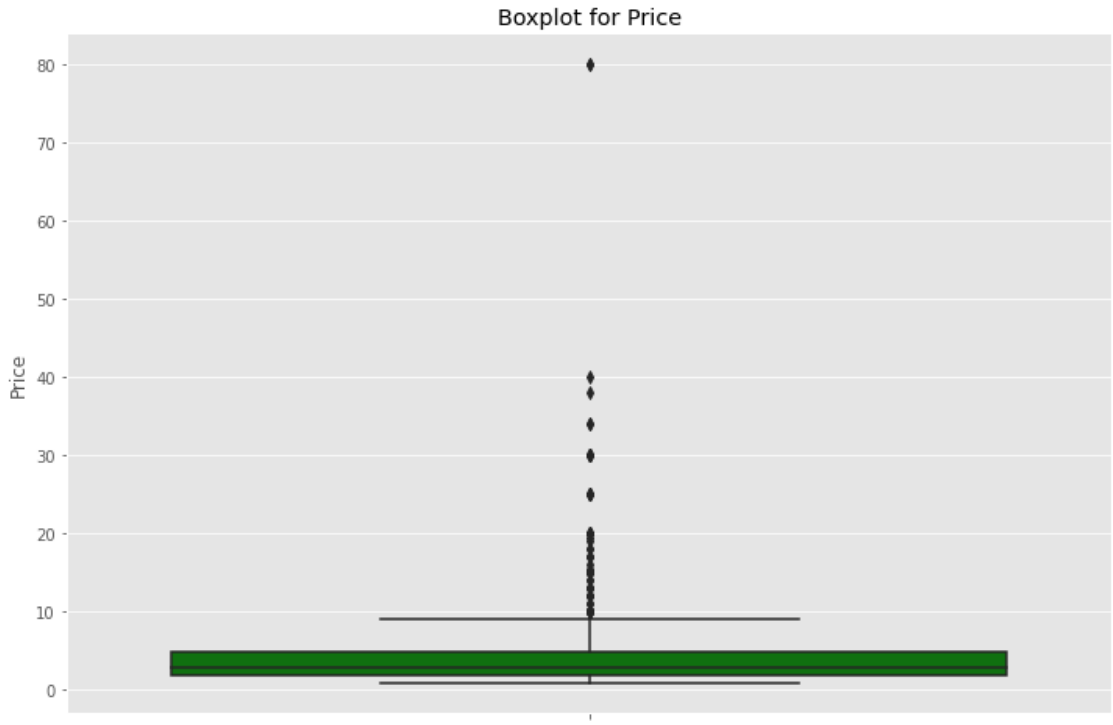
[396]:
```
#Importing the libraries for visualisation
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib import style
```

[397]:
```
style.use("ggplot")
```

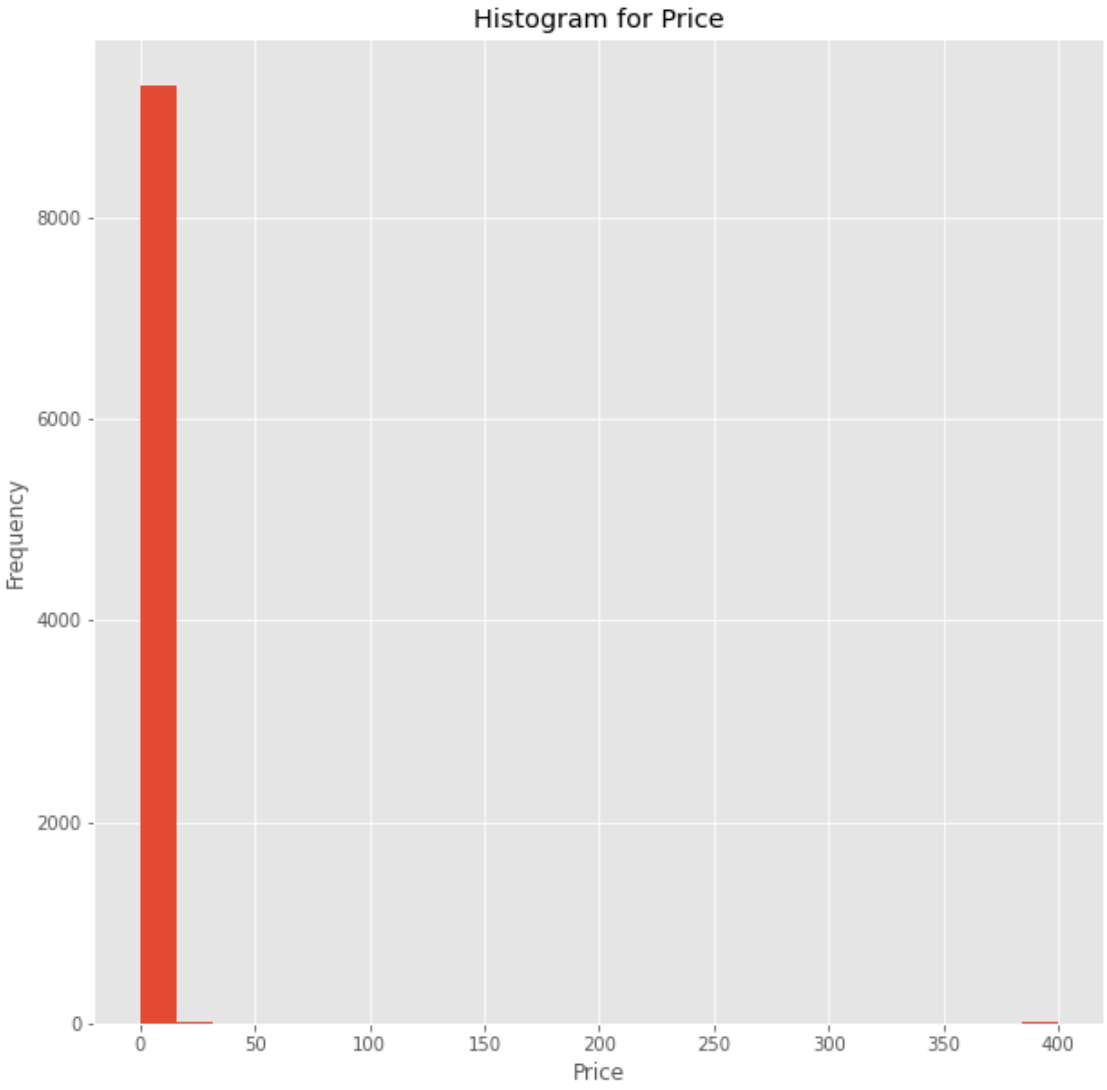
[398]:
```
#6.1. Boxplot for Price
plt.figure(figsize=(12,8))
sns.boxplot(y=playstore['Price'])
plt.title('Boxplot for Price')
plt.show()
```



[458]:
```
#Since most of the apps are free, the price values are zero for a number of␣
 ↪rows, we are unable to see the boxplot in the above
#plot, hence excluded the apps with zero price inorder to get a good␣
 ↪understanding of the outliers
plt.figure(figsize=(12,8))
exp = playstore[(playstore.Price > 0)]
sns.boxplot(y=exp.Price, color = 'green')
plt.title('Boxplot for Price')
plt.show()
```



[400]:
```
#Due to large variation in Price range, we are unable to view the boxplot␣
 ↪perfectly and hence using Histogtram for Price
#to check for outliers
plt.figure(figsize=(10,10))
plt.hist(playstore.Price, bins=25)
plt.xlabel('Price')
plt.ylabel('Frequency')
plt.title('Histogram for Price')
plt.show()

#There are some apps with price between $350 to $400 which is way more than the␣
 ↪average
```



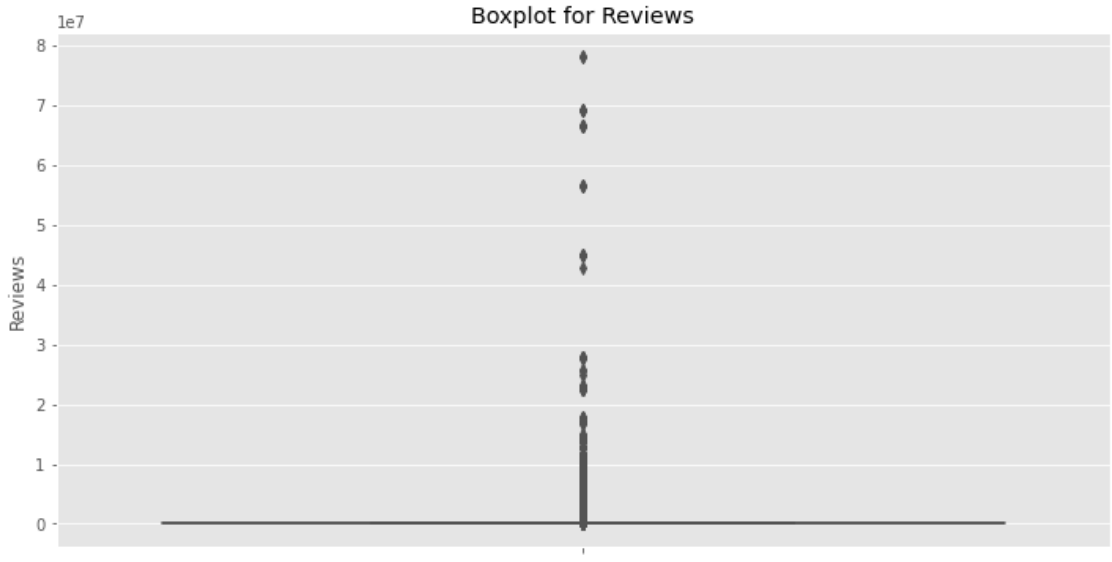
[401]:
```
playstore[playstore['Price'] > 200]['Price'].value_counts()
```

[401]:
```
399.99    11
389.99     1
379.99     1
299.99     1
400.00     1
Name: Price, dtype: int64
```

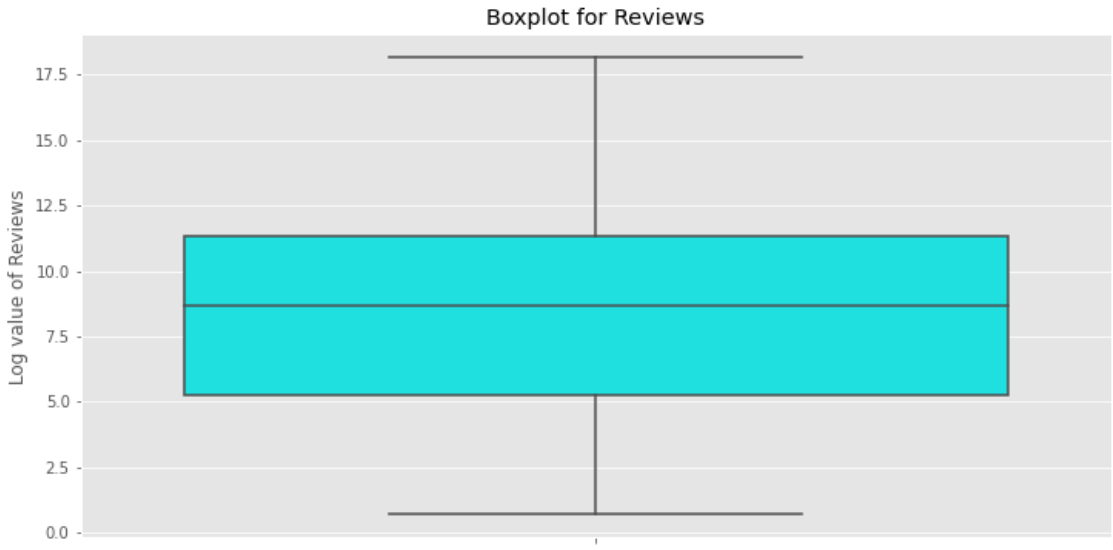
[402]:
```
plt.figure(figsize=(12,6))
sns.scatterplot(data = playstore['Price'], color = 'green')
plt.ylabel('Price')
plt.title('A Scatterplot of Price to show the variation in Price')
```

```
plt.show()
```

[403]:
```
#6.2. Boxplot for Reviews
plt.figure(figsize=(12,6))
sns.boxplot(y=playstore['Reviews'])
plt.title('Boxplot for Reviews')
plt.show()
```



[404]:
```
#Due to large range of Reviews values, the bosplot is not properly displayed,␣
 ↪so taking
#logarithmic values of Reviews to get the visaulaisation
changeReview = playstore['Reviews'].copy()
logReview = np.log1p(changeReview)

plt.figure(figsize=(12,6))
sns.boxplot(y=logReview, color = 'cyan')
plt.ylabel('Log value of Reviews')
plt.title('Boxplot for Reviews')
plt.show()
```



[405]:
```
playstore.Reviews.max()
```

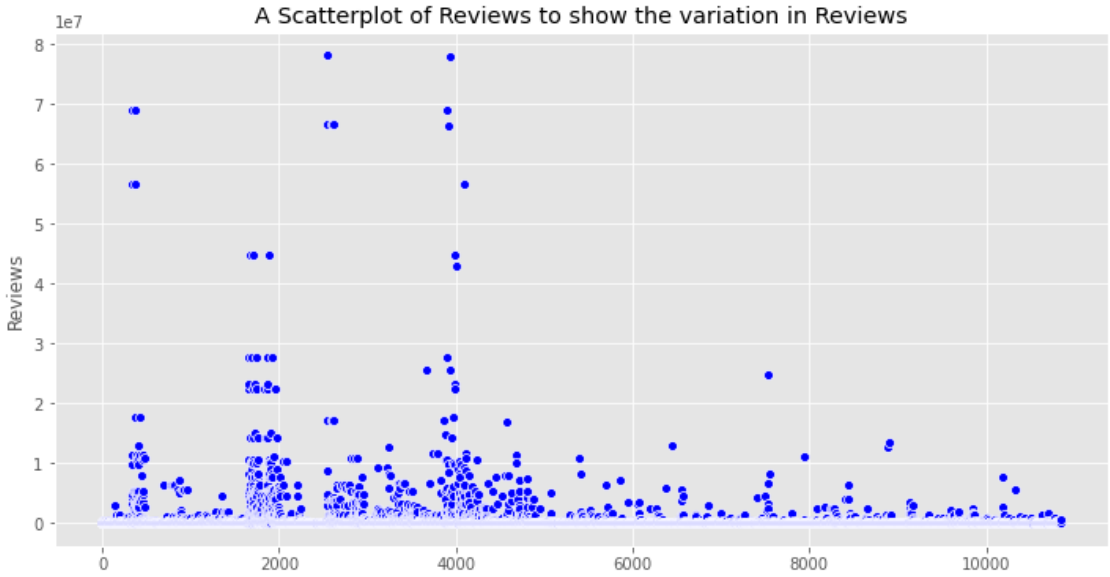
[405]: 78158306

[406]:
```
playstore.Reviews.min()
```

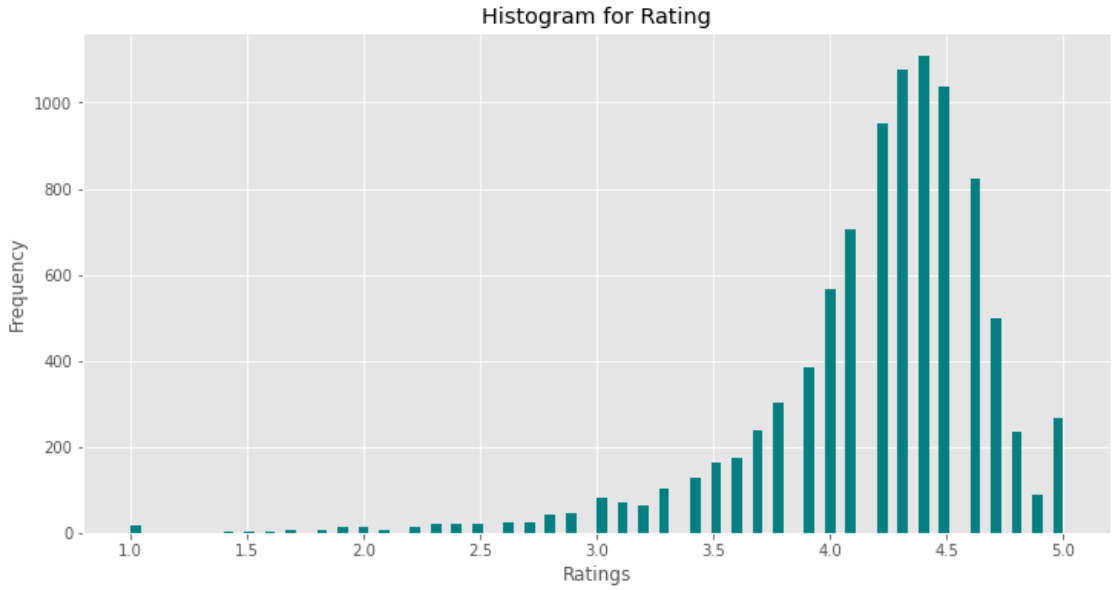
[406]: 1

[407]:
```
#Due to large variation in Reviews range, we are unable to view the boxplot␣
 ↪perfectly and hence using Scatterplot for Reviews
#to check for outliers

plt.figure(figsize=(12,6))
sns.scatterplot(data = playstore['Reviews'], color = 'blue')
plt.ylabel('Reviews')
plt.title('A Scatterplot of Reviews to show the variation in Reviews')
```

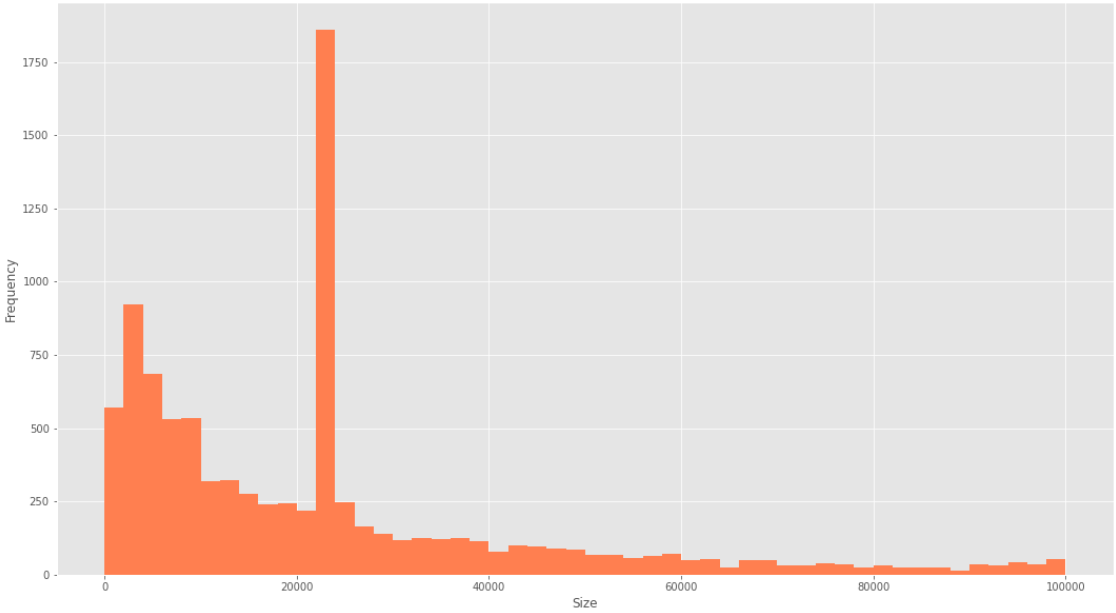
```
plt.show()

#There are some apps with high Reviews which is way more than the average
```



[408]:
```
#6.3. Histogram for Rating
plt.figure(figsize=(12,6))
plt.hist(playstore.Rating, bins=90, color='teal')
plt.xlabel('Ratings')
plt.ylabel('Frequency')
plt.title('Histogram for Rating')
plt.show()

#The ratings are highly distributed towards the higher value of ratings
```



[409]:
```
#6.4. Histogram for Size
plt.figure(figsize=(18,10))
plt.hist(playstore.Size, bins=50, color='coral')
plt.xlabel('Size')
plt.ylabel('Frequency')
plt.show()

#The size value between 20000 and 250000 is most frequently used compared to␣
↪other apps with varying sizes
```



[410]:
```
#Outlier Treatment
```

[411]:
```
#1.Price
playstore[playstore.Price > 200]
```

[411]:

	App	Category	Rating	Reviews	Size
4197	most expensive app (H)	FAMILY	4.3	6	1500
4362	I'm rich	LIFESTYLE	3.8	718	26000
4367	I'm Rich - Trump Edition	LIFESTYLE	3.6	275	7300
5351	I am rich	LIFESTYLE	3.8	3547	1800
5354	I am Rich Plus	FAMILY	4.0	856	8700
5355	I am rich VIP	LIFESTYLE	3.8	411	2600
5356	I Am Rich Premium	FINANCE	4.1	1867	4700
5357	I am extremely Rich	LIFESTYLE	2.9	41	2900
5358	I am Rich!	FINANCE	3.8	93	22000
5359	I am rich(premium)	FINANCE	3.5	472	965
5362	I Am Rich Pro	FAMILY	4.4	201	2700
5364	I am rich (Most expensive app)	FINANCE	4.1	129	2700
5366	I Am Rich	FAMILY	3.6	217	4900
5369	I am Rich	FINANCE	4.3	180	3800
5373	I AM RICH PRO PLUS	FINANCE	4.0	36	41000

	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
```

[412]:
```
#Removing records having Price greater than $200 as most of them are junk apps
playstore.drop(playstore[(playstore.Price > 200)].index, inplace=True)
playstore.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9338 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             9338 non-null   object
 1   Category        9338 non-null   object
 2   Rating          9338 non-null   float64
 3   Reviews         9338 non-null   int64
 4   Size            9338 non-null   int64
 5   Installs        9338 non-null   int64
 6   Type            9338 non-null   object
 7   Price           9338 non-null   float64
 8   Content Rating  9338 non-null   object
 9   Genres          9338 non-null   object
 10  Last Updated    9338 non-null   object
 11  Current Ver     9338 non-null   object
 12  Android Ver     9338 non-null   object
dtypes: float64(2), int64(3), object(8)
memory usage: 1021.3+ KB
```

[413]:
```
playstore[playstore.Price > 200]
```

[413]:
```
Empty DataFrame
Columns: [App, Category, Rating, Reviews, Size, Installs, Type, Price, Content
Rating, Genres, Last Updated, Current Ver, Android Ver]
Index: []
```

[414]:
```
#6.2. Reviews
```

[415]:
```
#checking for apps with more than 2 million reviews
playstore[playstore.Reviews > 2000000]
```

[415]:

	App	Category	Rating	\
139	Wattpad Free Books	BOOKS_AND_REFERENCE	4.6	
335	Messenger - Text and Video Chat for Free	COMMUNICATION	4.0	
336	WhatsApp Messenger	COMMUNICATION	4.4	
338	Google Chrome: Fast & Secure	COMMUNICATION	4.3	
340	Gmail	COMMUNICATION	4.3	
…	…	…	…	
9166	Modern Combat 5: eSports FPS	GAME	4.3	
9841	Google Earth	TRAVEL_AND_LOCAL	4.3	
10186	Farm Heroes Saga	FAMILY	4.4	
10190	Fallout Shelter	FAMILY	4.6	
10327	Garena Free Fire	GAME	4.5	

	Reviews	Size	Installs	Type	Price	Content Rating	\
139	2914724	22970	100000000	Free	0.0	Teen	
335	56642847	22970	1000000000	Free	0.0	Everyone	
336	69119316	22970	1000000000	Free	0.0	Everyone	
338	9642995	22970	1000000000	Free	0.0	Everyone	
340	4604324	22970	1000000000	Free	0.0	Everyone	
…	…	…	…	…	…	…	
9166	2903386	58000	100000000	Free	0.0	Mature 17+	
9841	2339098	22970	100000000	Free	0.0	Everyone	
10186	7615646	71000	100000000	Free	0.0	Everyone	
10190	2721923	25000	10000000	Free	0.0	Teen	
10327	5534114	53000	100000000	Free	0.0	Teen	

	Genres	Last Updated	Current Ver	\
139	Books & Reference	August 1, 2018	Varies with device	
335	Communication	August 1, 2018	Varies with device	
336	Communication	August 3, 2018	Varies with device	
338	Communication	August 1, 2018	Varies with device	
340	Communication	August 2, 2018	Varies with device	
…	…	…	…	
9166	Action	July 24, 2018	3.2.1c	
9841	Travel & Local	June 18, 2018	9.2.17.13	
10186	Casual	August 7, 2018	5.2.6	
10190	Simulation	June 11, 2018	1.13.12	
10327	Action	August 3, 2018	1.21.0	

	Android Ver
139	Varies with device
335	Varies with device
336	Varies with device
338	Varies with device
340	Varies with device
…	…
9166	4.0 and up

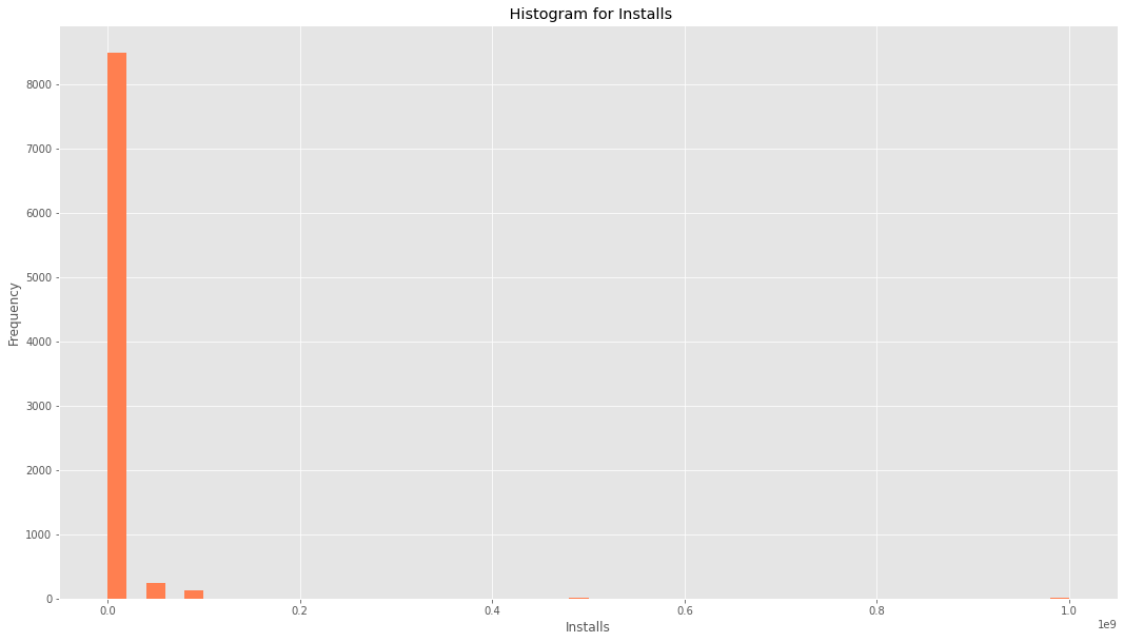
```
9841           4.1 and up
10186          2.3 and up
10190          4.1 and up
10327        4.0.3 and up

[453 rows x 13 columns]
```

[416]:
```
#Removing the apps with more than 2 million reviews as they are mostly star␣
 ↪apps and including their data in training the model
#may skew the model
playstore.drop(playstore[(playstore.Reviews > 2000000)].index, inplace=True)
playstore.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8885 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             8885 non-null   object
 1   Category        8885 non-null   object
 2   Rating          8885 non-null   float64
 3   Reviews         8885 non-null   int64
 4   Size            8885 non-null   int64
 5   Installs        8885 non-null   int64
 6   Type            8885 non-null   object
 7   Price           8885 non-null   float64
 8   Content Rating  8885 non-null   object
 9   Genres          8885 non-null   object
 10  Last Updated    8885 non-null   object
 11  Current Ver     8885 non-null   object
 12  Android Ver     8885 non-null   object
dtypes: float64(2), int64(3), object(8)
memory usage: 971.8+ KB
```

[417]:
```
#6.3 Installs
```

[418]:
```
#To find the outliers in the 'installs' field using Histogram
plt.figure(figsize=(18,10))
plt.hist(playstore.Installs, bins=50, color='coral')
plt.xlabel('Installs')
plt.ylabel('Frequency')
plt.title(' Histogram for Installs')
plt.show()
```



[419]:
```
#Finding different percentiles to find the threshold to drop the outlier values
def Percentile_installs(x):
    return playstore.Installs.quantile(x)
percent = [10,25,50,70,90,95,99]
for count,value in enumerate(percent):
    print('Percentile', value, ':', Percentile_installs(value/100))
```

```
Percentile 10 : 1000.0
Percentile 25 : 10000.0
Percentile 50 : 500000.0
Percentile 70 : 1000000.0
Percentile 90 : 10000000.0
Percentile 95 : 10000000.0
Percentile 99 : 100000000.0
```

[420]:
```
def PercentileNumpy_installs(x):
    return np.percentile(playstore['Installs'],x)
percent = [10,25,50,70,90,95,99]
for count,value in enumerate(percent):
    print('Percentile', value, ':', PercentileNumpy_installs(value))
```

```
Percentile 10 : 1000.0
Percentile 25 : 10000.0
Percentile 50 : 500000.0
Percentile 70 : 1000000.0
Percentile 90 : 10000000.0
Percentile 95 : 10000000.0
```

```
Percentile 99 : 100000000.0
```

[421]:
```
#Considering 90th percentile as threshold
playstore[(playstore.Installs > 10000000)]
```

[421]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000	50000000	Free	0.0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
143	Amazon Kindle	BOOKS_AND_REFERENCE	4.2	814080	22970	100000000	Free	0.0	Teen	Books & Reference	July 27, 2018	Varies with device	Varies with device
152	Google Play Books	BOOKS_AND_REFERENCE	3.9	1433233	22970	1000000000	Free	0.0	Teen	Books & Reference	August 3, 2018	Varies with device	Varies with device
188	Indeed Job Search	BUSINESS	4.3	674730	22970	50000000	Free	0.0	Everyone	Business	May 21, 2018	Varies with device	Varies with device
192	Docs To Go Free Office Suite	BUSINESS	4.1	217730	22970	50000000	Free	0.0	Everyone	Business	April 2, 2018	Varies with device	Varies with device
…	…	…	…	…	…	…	…	…	…	…	…	…	…
10429	Talking Tom Bubble Shooter	FAMILY	4.4	687136	54000	50000000	Free	0.0	Everyone	Casual	May 25, 2018	1.5.3.20	4.1 and up
10513	Flight Simulator: Fly Plane 3D	FAMILY	4.0	660613	21000	50000000	Free	0.0	Everyone	Simulation	March 1, 2017	1.32	2.3 and up
10549	Toy Truck Rally 3D	GAME	4.0	301895	25000	50000000	Free	0.0	Everyone	Racing	May 23, 2018	1.4.4	4.1 and up
10647	Motorola FM Radio	VIDEO_PLAYERS	3.9	54815	22970	100000000	Free	0.0	Everyone	Video Players & Editors	May 2, 2018	Varies with device	Varies with device
10707	Photo Editor Collage Maker Pro	PHOTOGRAPHY	4.5	1519671	22970	100000000	Free	0.0	Everyone	Photography	February 1, 2018	Varies with device	Varies with device

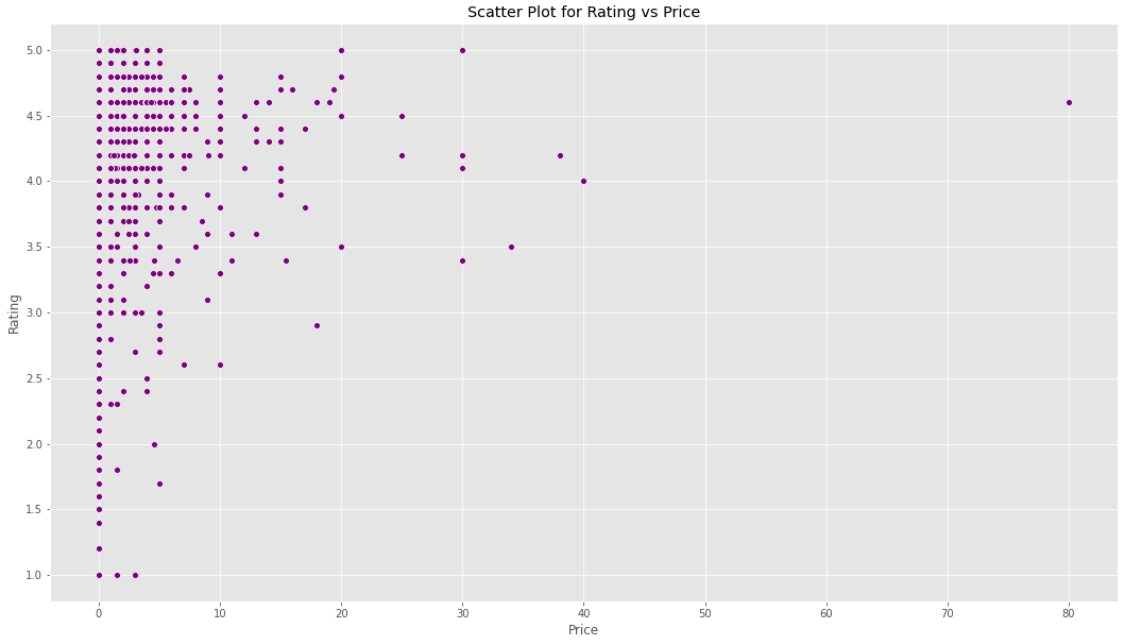
[389 rows x 13 columns]

[422]:
```
#Dropping rows with more than 100000000 Installs
playstore.drop(playstore[(playstore.Installs > 10000000)].index, inplace=True)
playstore.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8496 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             8496 non-null   object
 1   Category        8496 non-null   object
 2   Rating          8496 non-null   float64
 3   Reviews         8496 non-null   int64
 4   Size            8496 non-null   int64
 5   Installs        8496 non-null   int64
 6   Type            8496 non-null   object
 7   Price           8496 non-null   float64
 8   Content Rating  8496 non-null   object
 9   Genres          8496 non-null   object
 10  Last Updated    8496 non-null   object
 11  Current Ver     8496 non-null   object
 12  Android Ver     8496 non-null   object
dtypes: float64(2), int64(3), object(8)
memory usage: 929.2+ KB
```

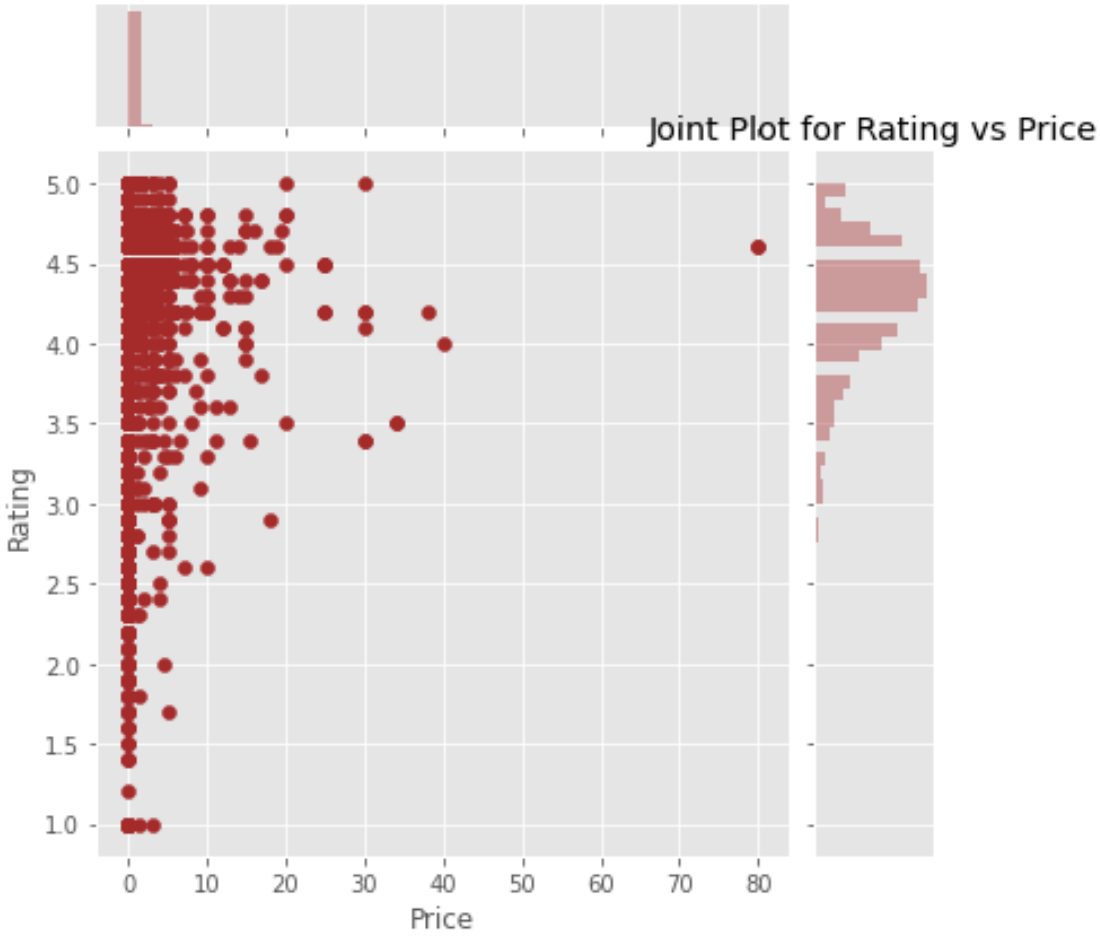
[423]:
```
#7.Bivariate Analysis
```

[424]:
```
#7.1. Rating vs Price - Scatter plot
plt.figure(figsize=(18,10))
sns.scatterplot(y=playstore.Rating,x=playstore.Price, color = 'purple')
plt.title('Scatter Plot for Rating vs Price')
plt.show()

# When Price increases, Rating is mostly on the higher side
```

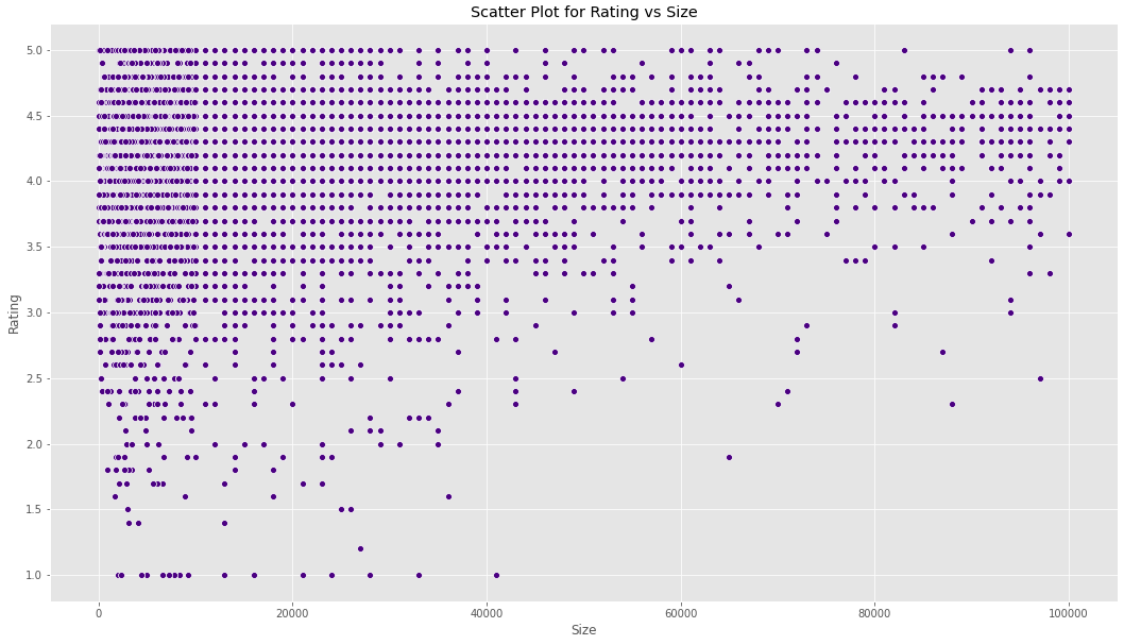


[425]:
```
#7.1. Rating vs Price - Joint plot
plt.figure(figsize=(15,15))
sns.jointplot(y=playstore.Rating,x=playstore.Price, kind='scatter',␣
 ↪color='brown')
plt.title('Joint Plot for Rating vs Price')
plt.show()
```

```
<Figure size 1080x1080 with 0 Axes>
```

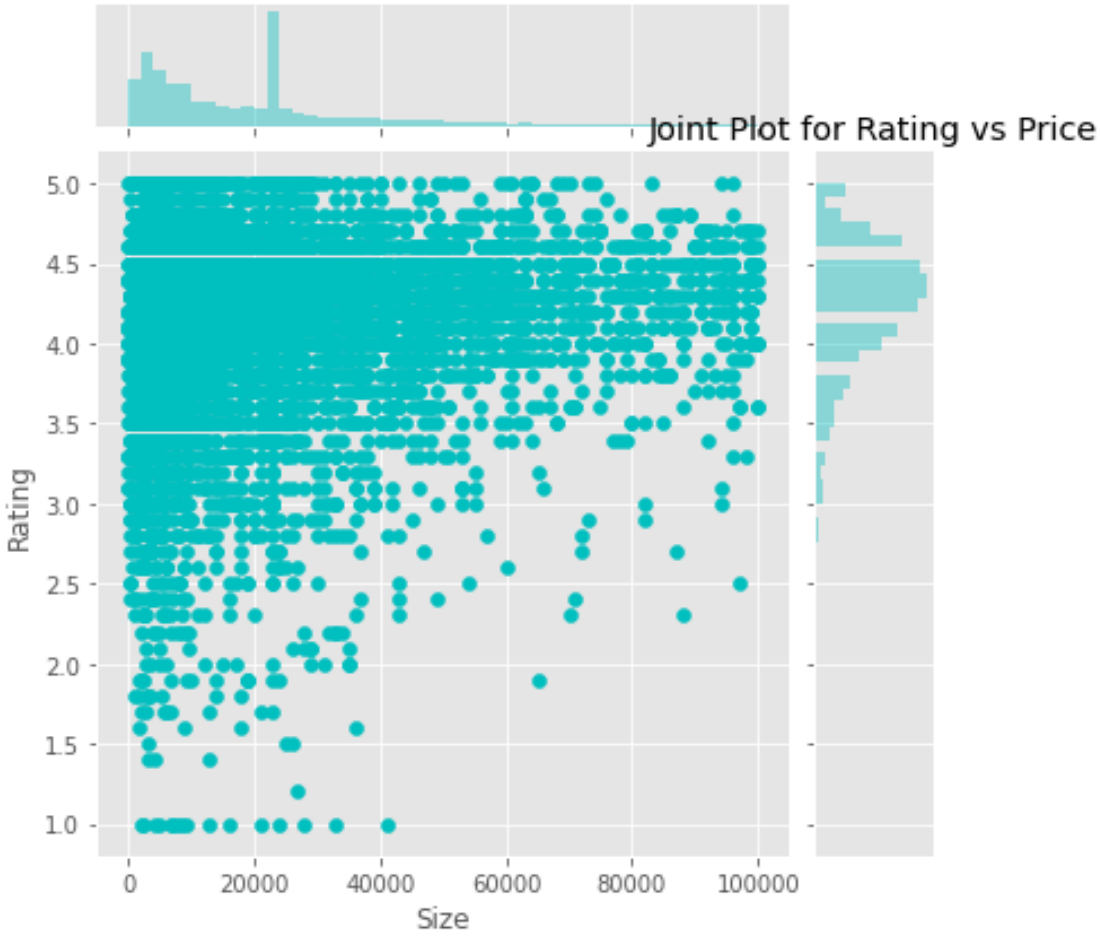


[426]:
```
#7.2. Rating vs Size - Scatter plot
plt.figure(figsize=(18,10))
sns.scatterplot(y=playstore.Rating,x=playstore.Size, color='indigo')
plt.title('Scatter Plot for Rating vs Size')
plt.show()

# The heavier apps are rated better compared to that of the lighter apps
```



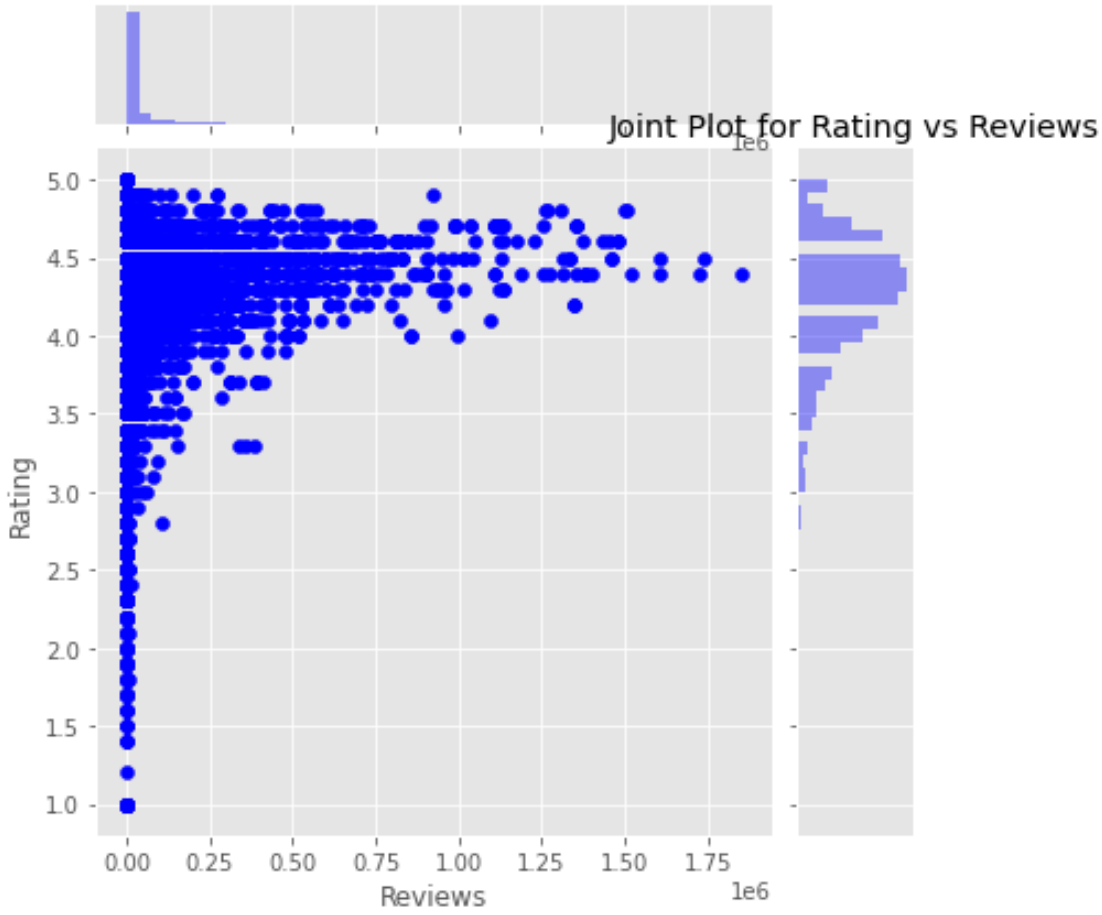
[427]:
```
#7.2. Rating vs Size - Joint plot
plt.figure(figsize=(6,10))
sns.jointplot(y=playstore.Rating,x=playstore.Size, color ='c')
plt.title('Joint Plot for Rating vs Price')
plt.show()
```

```
<Figure size 432x720 with 0 Axes>
```

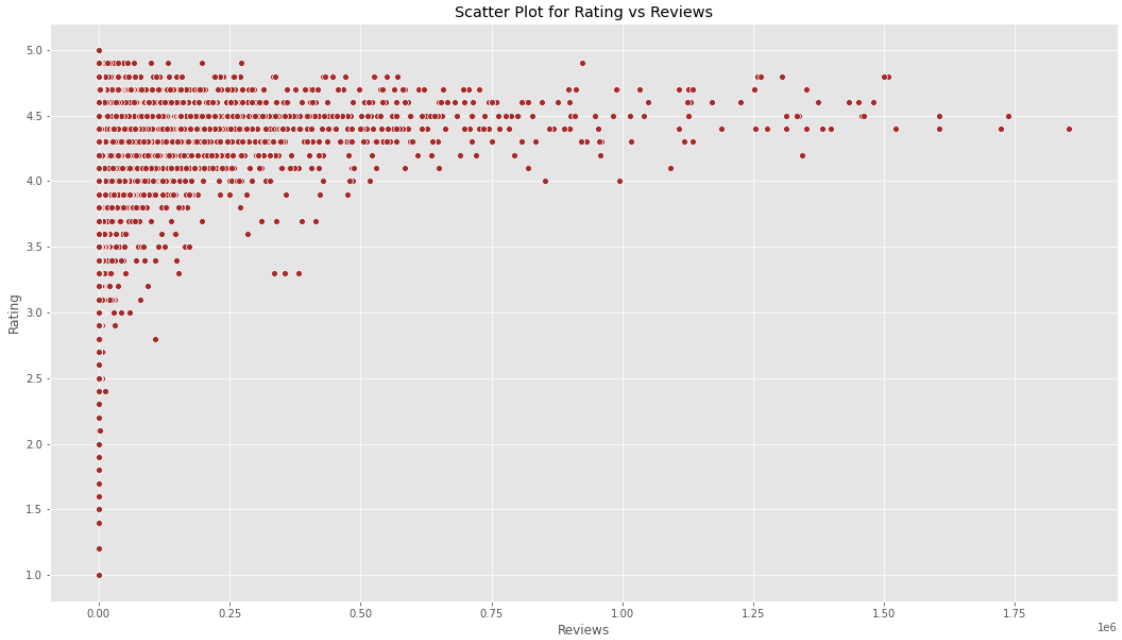


[428]:
```
#7.3. Rating vs Reviews - Joint plot
plt.figure(figsize=(10,10))
sns.jointplot(y=playstore.Rating,x=playstore.Reviews, kind='scatter', color='b')
plt.title('Joint Plot for Rating vs Reviews')
plt.show()
```

```
<Figure size 720x720 with 0 Axes>
```



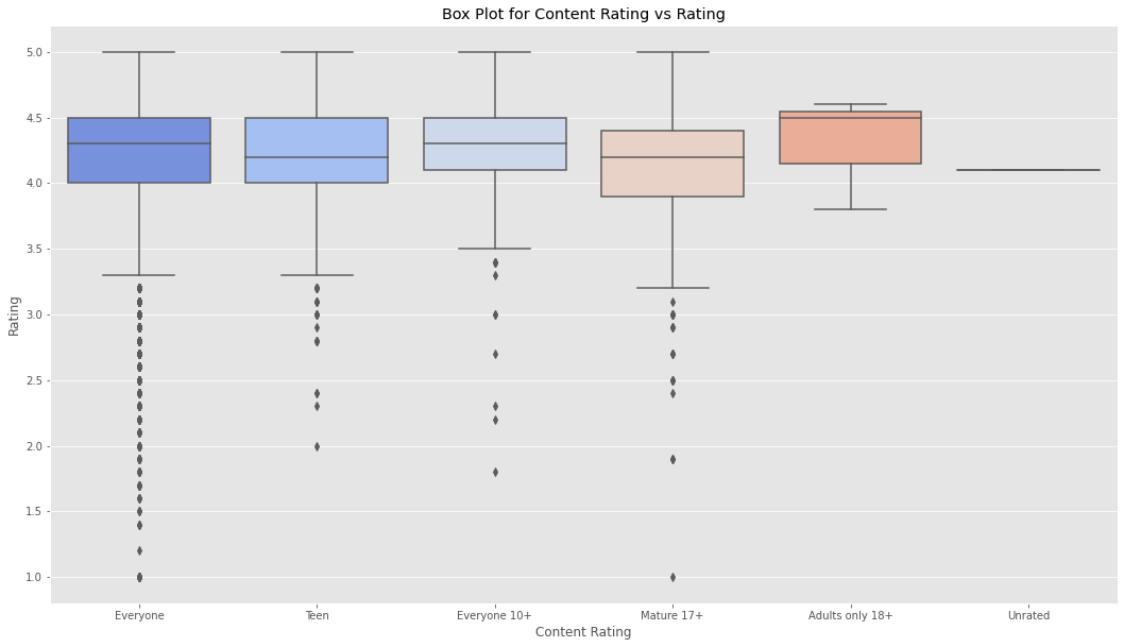
[429]:
```
#7.3. Rating vs Reviews - Scatter plot
plt.figure(figsize=(18,10))
sns.scatterplot(y=playstore.Rating,x=playstore.Reviews, color='brown')
plt.title('Scatter Plot for Rating vs Reviews')
plt.show()

# When there is a large number of reviews, the rating is always high
```



[430]:
```
playstore['Content Rating'].unique()
```

[430]:
```
array(['Everyone', 'Teen', 'Everyone 10+', 'Mature 17+',
       'Adults only 18+', 'Unrated'], dtype=object)
```

[431]:
```
#7.4 Boxplot for Rating vs Content Rating
plt.figure(figsize=(18,10))
sns.boxplot(y=playstore.Rating,x=playstore['Content Rating'],␣
 ↪palette='coolwarm')
plt.title('Box Plot for Content Rating vs Rating')
plt.show()

# The ratings are nearly same( between 4.0 and 4.5) for Everyone, Teen and␣
 ↪Everyone 10+ contents. But, there is a difference
#in ratings for Mature 17+ and Adults only 18+ contents
```



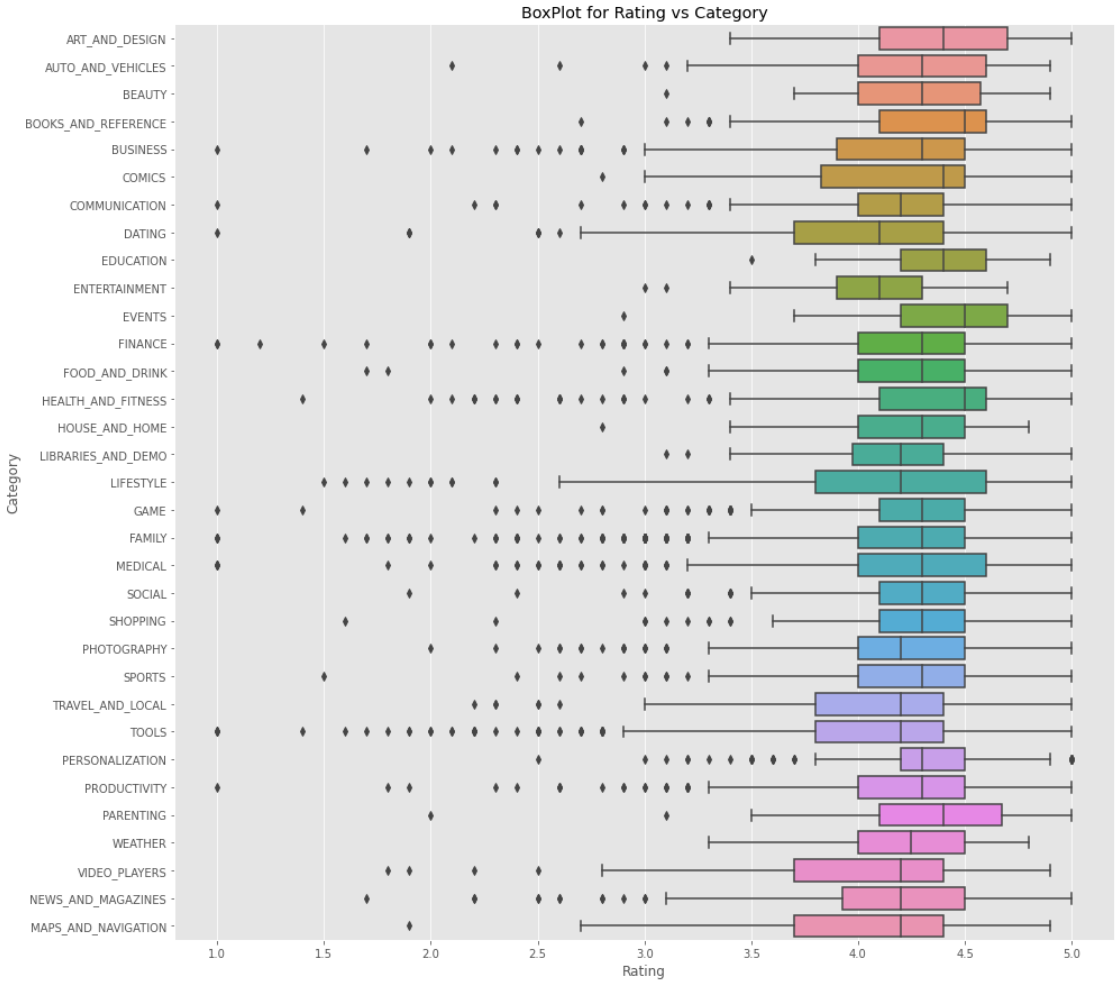
[432]:
```
type(playstore['Category'][0])
```

[432]:
```
str
```

[433]:
```
playstore.Category.unique()
```

[433]:
```
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)
```

[434]:
```
#7.5 Boxplot for Rating vs Category
plt.figure(figsize=(15,15))
sns.boxplot(x=playstore.Rating,y=playstore['Category'])
plt.title('BoxPlot for Rating vs Category')
plt.show()

# We can see that 'Personalization', 'Events', 'Education' and 'Art and design'␣
↪has the best ratings compared to others.
```



[435]:
```
# Other Visualisations
res=playstore.groupby('Category')['Rating'].sum()
res
```

[435]:
```
Category
ART_AND_DESIGN          262.5
AUTO_AND_VEHICLES       305.9
BEAUTY                  179.7
BOOKS_AND_REFERENCE     742.9
BUSINESS               1207.1
COMICS                  241.0
COMMUNICATION           997.5
DATING                  774.3
EDUCATION               661.5
ENTERTAINMENT           546.3
EVENTS                  199.6
FAMILY                 6958.8
```

```
FINANCE                1285.4
FOOD_AND_DRINK          454.2
GAME                   3570.6
HEALTH_AND_FITNESS     1238.2
HOUSE_AND_HOME          319.0
LIBRARIES_AND_DEMO      267.5
LIFESTYLE              1254.6
MAPS_AND_NAVIGATION     476.2
MEDICAL                1461.2
NEWS_AND_MAGAZINES      916.8
PARENTING               215.0
PERSONALIZATION        1253.2
PHOTOGRAPHY            1026.2
PRODUCTIVITY           1189.7
SHOPPING                879.7
SOCIAL                  904.6
SPORTS                 1286.8
TOOLS                  2705.4
TRAVEL_AND_LOCAL        837.1
VIDEO_PLAYERS           542.6
WEATHER                 296.2
Name: Rating, dtype: float64
```

[436]:
```
# Exploring Crosstab
pd.crosstab(playstore.Category,playstore.Rating,margins=True)
```

[436]:

Rating	1.0	1.2	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	…	\
Category											…	
ART_AND_DESIGN	0	0	0	0	0	0	0	0	0	0	…	
AUTO_AND_VEHICLES	0	0	0	0	0	0	0	0	0	1	…	
BEAUTY	0	0	0	0	0	0	0	0	0	0	…	
BOOKS_AND_REFERENCE	0	0	0	0	0	0	0	0	0	0	…	
BUSINESS	1	0	0	0	0	1	0	0	1	1	…	
COMICS	0	0	0	0	0	0	0	0	0	0	…	
COMMUNICATION	1	0	0	0	0	0	0	0	0	0	…	
DATING	1	0	0	0	0	0	0	3	0	0	…	
EDUCATION	0	0	0	0	0	0	0	0	0	0	…	
ENTERTAINMENT	0	0	0	0	0	0	0	0	0	0	…	
EVENTS	0	0	0	0	0	0	0	0	0	0	…	
FAMILY	3	0	0	0	1	2	2	3	1	0	…	
FINANCE	2	1	0	1	0	1	0	0	2	1	…	
FOOD_AND_DRINK	0	0	0	0	0	1	1	0	0	0	…	
GAME	1	0	1	0	0	0	0	0	0	0	…	
HEALTH_AND_FITNESS	0	0	1	0	0	0	0	0	1	1	…	
HOUSE_AND_HOME	0	0	0	0	0	0	0	0	0	0	…	
LIBRARIES_AND_DEMO	0	0	0	0	0	0	0	0	0	0	…	
LIFESTYLE	0	0	0	1	1	1	1	1	2	2	…	

MAPS_AND_NAVIGATION	0	0	0	0	0	0	0	1	0	0	…
MEDICAL	3	0	0	0	0	0	1	0	1	0	…
NEWS_AND_MAGAZINES	0	0	0	0	0	1	0	0	0	0	…
PARENTING	0	0	0	0	0	0	0	0	1	0	…
PERSONALIZATION	0	0	0	0	0	0	0	0	0	0	…
PHOTOGRAPHY	0	0	0	0	0	0	0	0	1	0	…
PRODUCTIVITY	1	0	0	0	0	0	1	1	0	0	…
SHOPPING	0	0	0	0	1	0	0	0	0	0	…
SOCIAL	0	0	0	0	0	0	0	1	0	0	…
SPORTS	0	0	0	1	0	0	0	0	0	0	…
TOOLS	3	0	1	0	1	1	1	2	2	2	…
TRAVEL_AND_LOCAL	0	0	0	0	0	0	0	0	0	0	…
VIDEO_PLAYERS	0	0	0	0	0	0	1	1	0	0	…
WEATHER	0	0	0	0	0	0	0	0	0	0	…
All	16	1	3	3	4	8	8	13	12	8	…

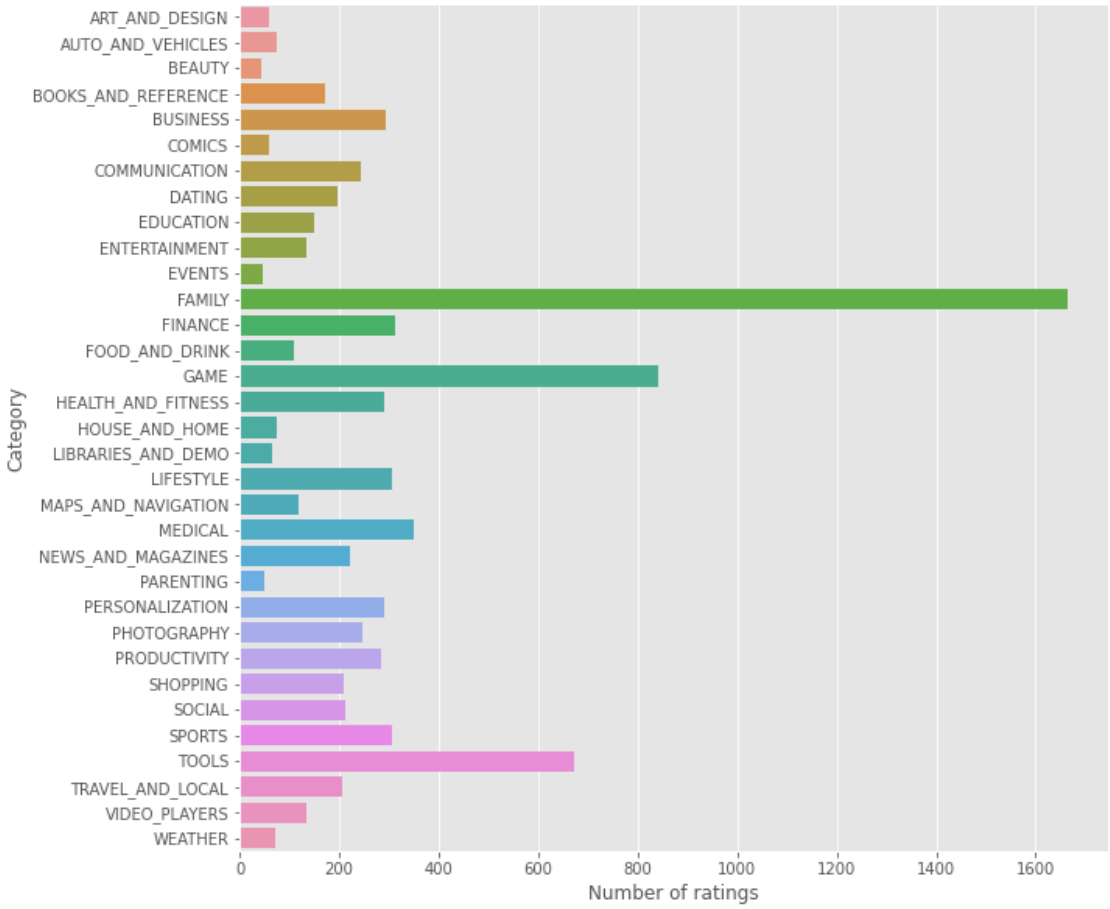
Rating	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0	All
Category										
ART_AND_DESIGN	7	4	8	3	3	13	5	0	1	60
AUTO_AND_VEHICLES	7	7	6	5	11	0	5	5	0	73
BEAUTY	4	3	3	5	3	6	1	1	0	42
BOOKS_AND_REFERENCE	12	9	18	24	22	20	11	4	6	171
BUSINESS	28	25	37	20	19	16	10	3	18	293
COMICS	4	1	9	8	4	5	2	0	2	58
COMMUNICATION	32	38	29	16	9	1	6	0	5	243
DATING	24	7	19	7	4	7	7	1	6	195
EDUCATION	19	20	21	20	24	17	2	2	0	151
ENTERTAINMENT	19	19	8	11	7	1	0	0	0	133
EVENTS	2	3	5	5	6	2	4	1	6	45
FAMILY	151	189	180	160	136	82	50	17	67	1663
FINANCE	23	33	34	31	29	27	8	2	8	311
FOOD_AND_DRINK	8	13	6	16	11	10	2	0	2	109
GAME	93	111	104	104	82	56	12	3	8	842
HEALTH_AND_FITNESS	15	13	38	49	45	14	23	6	12	290
HOUSE_AND_HOME	5	11	6	13	8	2	1	0	0	76
LIBRARIES_AND_DEMO	7	7	7	5	5	3	0	0	2	64
LIFESTYLE	15	22	22	23	26	15	5	8	29	306
MAPS_AND_NAVIGATION	19	15	12	11	5	3	1	2	0	118
MEDICAL	34	27	34	39	30	22	14	7	27	349
NEWS_AND_MAGAZINES	16	28	12	27	13	13	2	5	7	222
PARENTING	2	1	7	4	6	7	4	1	1	50
PERSONALIZATION	38	38	35	36	25	25	10	2	10	290
PHOTOGRAPHY	25	33	22	28	21	9	2	1	6	248
PRODUCTIVITY	37	31	29	28	32	16	4	2	8	285
SHOPPING	28	24	25	29	14	12	1	0	6	208
SOCIAL	25	27	22	20	16	10	11	4	7	213
SPORTS	38	38	42	29	34	10	6	5	4	305

```
TOOLS                 69   73   53   53   41   28   11    4   17   673
TRAVEL_AND_LOCAL      16   20   40   18    9    7    3    0    3   205
VIDEO_PLAYERS         10   17   16    9    6    6    5    1    0   135
WEATHER               15    4   10   12    3    3    3    0    0    70
All                  847  911  919  868  709  468  231   87  268  8496

[34 rows x 40 columns]
```

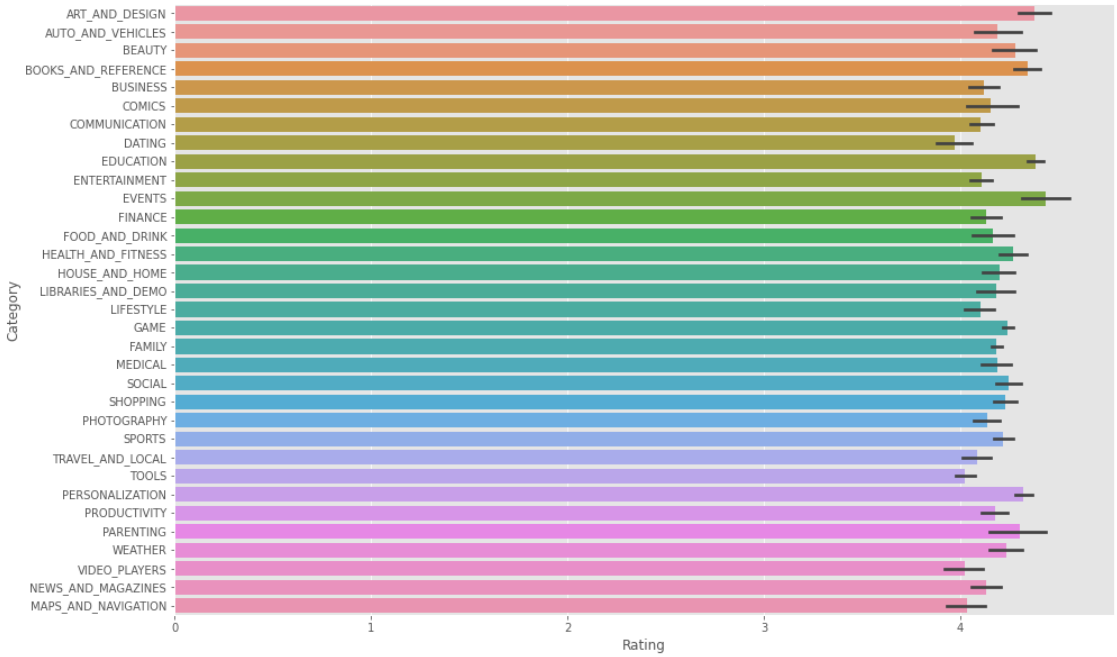
[437]:
```
res=playstore.groupby('Category')['Rating'].count()
res.index
```

[437]:
```
Index(['ART_AND_DESIGN', 'AUTO_AND_VEHICLES', 'BEAUTY', 'BOOKS_AND_REFERENCE',
       'BUSINESS', 'COMICS', 'COMMUNICATION', 'DATING', 'EDUCATION',
       'ENTERTAINMENT', 'EVENTS', 'FAMILY', 'FINANCE', 'FOOD_AND_DRINK',
       'GAME', 'HEALTH_AND_FITNESS', 'HOUSE_AND_HOME', 'LIBRARIES_AND_DEMO',
       'LIFESTYLE', 'MAPS_AND_NAVIGATION', 'MEDICAL', 'NEWS_AND_MAGAZINES',
       'PARENTING', 'PERSONALIZATION', 'PHOTOGRAPHY', 'PRODUCTIVITY',
       'SHOPPING', 'SOCIAL', 'SPORTS', 'TOOLS', 'TRAVEL_AND_LOCAL',
       'VIDEO_PLAYERS', 'WEATHER'],
      dtype='object', name='Category')
```

[438]:
```
plt.figure(figsize=(10,10))
sns.barplot(y=res.index,x=res)
plt.xlabel('Number of ratings')
plt.show()
```



[439]:
```
plt.figure(figsize=(15,10))
sns.barplot(y=playstore.Category,x=playstore.Rating)
plt.show()
```



[440]:
```
#8. Data Preprocessing
# Creating copy of dataframe and naming it as inp1
inp1 = playstore.copy(deep=True)
```

[441]:
```
inp1.Reviews.head()
```

[441]:
```
0      159
1      967
2    87510
4      967
5      167
Name: Reviews, dtype: int64
```

[442]:
```
#8.1. Applying log transformation to Reviews to reduce the skew
inp1.Reviews = np.log1p(inp1.Reviews)
inp1.Reviews.head()
```

[442]:
```
0     5.075174
1     6.875232
2    11.379520
4     6.875232
5     5.123964
Name: Reviews, dtype: float64
```

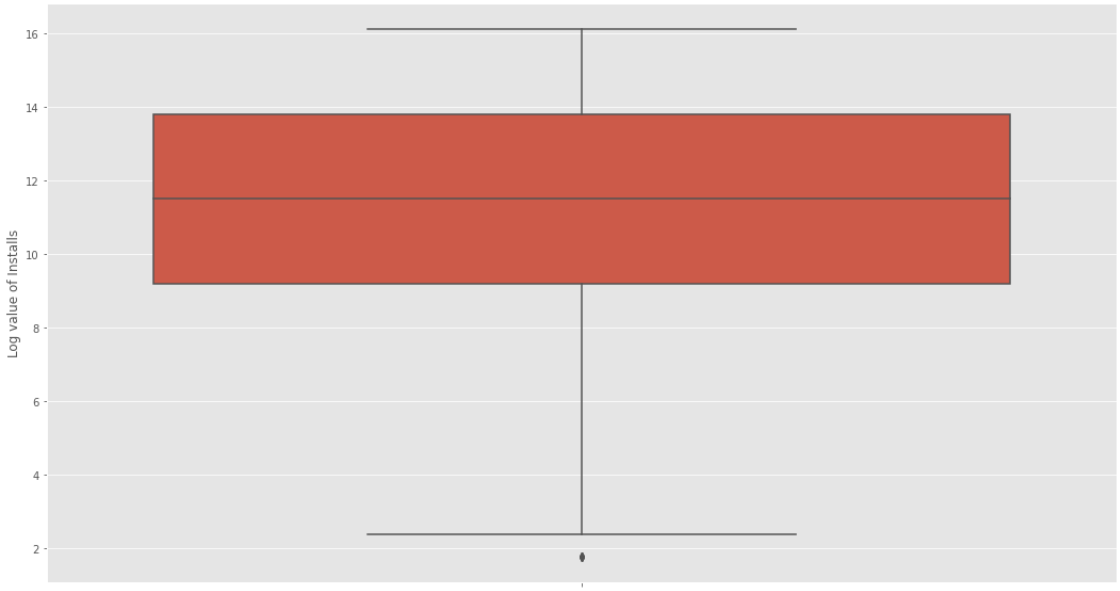
[443]:
```
inp1.Installs.head()
```

[443]:
```
0      10000
1     500000
2    5000000
4     100000
5      50000
Name: Installs, dtype: int64
```

[444]:
```
#8.1. Applying log transformation to Installa to reduce the skew
inp1.Installs = np.log1p(inp1.Installs)
inp1.Installs.head()
```

[444]:
```
0     9.210440
1    13.122365
2    15.424949
4    11.512935
5    10.819798
Name: Installs, dtype: float64
```

[445]:
```
plt.figure(figsize=(18,10))
sns.boxplot(y=inp1.Installs)
plt.ylabel('Log value of Installs')
plt.show()
```



[446]:
```
inp1.head()
```

[446]:

	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   9.210440  Free    0.0       Everyone
1   6.875232  14000  13.122365  Free    0.0       Everyone
2  11.379520   8700  15.424949  Free    0.0       Everyone
4   6.875232   2800  11.512935  Free    0.0       Everyone
5   5.123964   5600  10.819798  Free    0.0       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
```

[447]:
```
#8.2 Dropping columns App, Last Updated, Current Ver and Android Ver as they␣
↪don't contribute much to our model
inp1.drop(columns=['App','Last Updated', 'Current Ver', 'Android Ver'],␣
↪inplace=True)
```

[448]:
```
inp1.head()
```

[448]:
```
         Category  Rating    Reviews   Size    Installs  Type  Price  \
0  ART_AND_DESIGN     4.1   5.075174  19000   9.210440  Free    0.0
1  ART_AND_DESIGN     3.9   6.875232  14000  13.122365  Free    0.0
2  ART_AND_DESIGN     4.7  11.379520   8700  15.424949  Free    0.0
4  ART_AND_DESIGN     4.3   6.875232   2800  11.512935  Free    0.0
5  ART_AND_DESIGN     4.4   5.123964   5600  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
```

[449]:
```
#8.3. Getting dummy data for 'Category', 'Type', 'Content Rating', 'Genres'␣
↪columns to convert categorical data to numerical data
inp2 = pd.get_dummies(data=inp1,columns=['Category','Type','Content␣
↪Rating','Genres'], drop_first = True)
inp2.head()
```

[449]:

	Rating	Reviews	Size	Installs	Price	Category_AUTO_AND_VEHICLES	\
0	4.1	5.075174	19000	9.210440	0.0	0	
1	3.9	6.875232	14000	13.122365	0.0	0	
2	4.7	11.379520	8700	15.424949	0.0	0	
4	4.3	6.875232	2800	11.512935	0.0	0	
5	4.4	5.123964	5600	10.819798	0.0	0	

	Category_BEAUTY	Category_BOOKS_AND_REFERENCE	Category_BUSINESS	\
0	0	0	0	
1	0	0	0	
2	0	0	0	
4	0	0	0	
5	0	0	0	

	Category_COMICS	…	Genres_Tools	Genres_Tools;Education	\
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 157 columns]
```

[450]:
```
#Define features and outcomes
x = inp2.drop(labels='Rating',axis=1)
y = inp2.Rating
```

[451]:
```
#9 and 10. Split data with 70-30 for train and test data
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,random_state = 0,test_size=0.
↪30)
```

[452]:
```
#11. Model building
from sklearn.linear_model import LinearRegression
model = LinearRegression()
#Fit the model
model.fit(x_train,y_train)
print('Slope:', model.coef_)
print('Intercept:', model.intercept_)
```

```
Slope: [ 1.67831834e-01  6.80429728e-08 -1.47533686e-01  1.28080981e-03
  5.12041763e-01  5.54473726e-01  5.21870544e-01  4.26117991e-01
  7.76852309e-01  3.86135823e-01  3.42730363e-01  5.55237668e-01
  4.76053168e-01  5.84292758e-01  5.95183848e-01  4.19341107e-01
  4.36180304e-01  8.21340007e-01  4.56974536e-01  4.40317945e-01
  4.81437731e-01  4.12910537e-01  3.83305238e-01  4.64270547e-01
  4.15013625e-01  7.75692347e-01  5.09633035e-01  4.17390522e-01
  4.41607358e-01  4.55582116e-01  4.48961224e-01  7.97578047e-01
  5.91307624e-01  4.03777842e-01  5.95143013e-01  4.62141532e-01
 -7.95633871e-02  4.81879167e-02  5.32195101e-02  3.27150051e-02
  4.86294860e-02  3.95673538e-02  3.08237122e-01 -2.56453639e-02
  3.02619861e-01 -8.59312621e-13  2.63927504e-02  1.89866834e-02
  4.42328858e-01  5.06437241e-01  1.18378650e+00  7.85336969e-01
  3.44732357e-01  5.12041763e-01  5.54473726e-01 -9.39592328e-03
  3.41535557e-01  3.98337679e-01  9.50994536e-01  5.21870544e-01
  2.65207211e-01  4.26117991e-01 -2.54477731e-01  6.37194559e-02
  6.65473648e-01  6.50869843e-02  1.08655279e-01  3.33364489e-01
  6.43283545e-01  4.71495190e-01  3.52199741e-01  3.71514257e-01
  3.02750953e-01 -5.31130713e-02  8.29965381e-01  3.86135823e-01
  4.30629395e-01  3.42730363e-01  4.60903233e-01  4.19383693e-01
  3.57827729e-01  8.74013148e-01  4.39456623e-01  2.66172217e-01
  5.44233367e-01 -2.83299481e-02  5.24751041e-01  4.92609363e-01
  6.39547143e-02  4.90887442e-01  3.00885313e-01  2.41616877e-01
  3.35058582e-01  5.25911278e-01  6.83501659e-01  6.08907097e-01
  3.27230761e-01 -2.45029841e-01  5.84292758e-01  4.19341107e-01
  4.36180304e-01  4.56974536e-01 -1.38617031e-01  5.16218574e-01
```

```
  4.40317945e-01  4.81437731e-01  4.12910537e-01  3.10017008e-01
  2.63011835e-13  3.83305238e-01  4.64270547e-01 -1.92488130e-01
  6.87933740e-01  7.30716903e-01  4.15013625e-01  2.82201962e-01
 -5.27141279e-02 -1.29446101e-03  5.47498974e-01  5.09633035e-01
  4.17390522e-01  4.41607358e-01  3.36366248e-01  5.38061295e-01
  3.86864665e-01  3.75911868e-01 -5.52613511e-14 -1.13233014e-01
  3.91611039e-01  9.06326584e-01  1.38929564e-01  2.78979923e-01
  0.00000000e+00  3.16377281e-01  4.55582116e-01  1.93462666e-01
  4.59032616e-01  6.37189157e-02  1.89171209e-01  4.48961224e-01
  5.53369904e-02  2.53596633e-01  1.10812174e-01  5.04171714e-01
  1.23075326e-01  8.17040712e-01  1.85019663e-01  4.06287961e-01
  4.03777843e-01  0.00000000e+00 -2.56888745e-01  1.10753680e-01
  2.82404618e-02  7.44781443e-02  4.62141532e-01  2.36407082e-01]
Intercept: 3.691842104087253
```

[453]:
```
#R2 on train set
R2_train = model.score(x_train,y_train)
print('R2 on train set:', R2_train)
```

```
R2 on train set: 0.15438672866386827
```

[457]:
```
#12.Making predictions on the test set
y_test_pred = model.predict(x_test)
y_test_pred
```

[457]:
```
array([4.1117247 , 4.2477706 , 4.5028666 , …, 4.46065532, 4.10131754,
       3.98984321])
```

[455]:
```
#R2 on test set
R2_test = model.score(x_test,y_test)
print('R2 on test set:', R2_test)
```

```
R2 on test set: 0.15410630338250908
```