APP SUCCESS PREDICTION ON PLAY STORE

MACHINE LEARNING PROJECT

SUBMITTED BY-

Keshav -

Date: 28-11-2018

TABLE OF CONTENTS

ABSTRACT

This project aims to analyse data about various applications available on Google Play Store. Based on data like application category, size, price, number of installs, content rating, review count , reviews, prediction of how successful an android application will be on the Google Play Store is made. This is achieved by predicting the likely application rating on google play store. For doing this, Linear Regression model, SVM model and Random Forest regression model are used for predicting the rating. Also models are evaluated by comparing the predicted results against the actual results by the use of mean squared error & mean absolute error.

INTRODUCTION

NEED OF THE SYSTEM

For developing a good android application, it is better to be aware of the characteristics that makes an application successful on that platform. This system helps one know how well their application will work on Google Play Store based on features of the application and what improvements can be made to make that application a hit on Playstore platform. It will also help developers in improving existing applications to achieve higher customer satisfaction levels and better reviews and ratings on Play Store.

APPLICATIONS OF PROPOSED SYSTEM

- It can be used to predict rating of an application available on Google Play Store, based on current ratings of other applications.
- It can be used to predict the success of a new application on Google Play Store. One can simply add this new application's details in the testing set and get the results.

CHALLENGES IN DEVELOPMENT

- The columns 'category' and 'genre' store almost the same data. If two explanatory variables in a model are highly linearly related, it poses a problem called multicollinearity. Together, these columns have nearly the same effect on the final result. So considering them both can affect the result. Therefore, we dropped 'genre' column from the dataset.
- The dataset contained columns like 'Last Updated', 'Current version', 'Android Version', which do not play any part in the app ratings on Play Store. So we dropped these columns by using dataset.drop (labels=[]) function.
- In data preprocessing stage, an error was incurred because of the rows which had NULL values. Thus, we applied 'dataset.dropna()' function to remove the rows which had NULL values in them.
- We encountered an error of approximately 65% while using SVM model. It was so because we were initially performing feature scaling in SVM model. We overcame this error by removing feature scaling and re-applying the model. Without feature scaling, an error of approximately 20% was there.

WORKING OF PROPOSED SYSTEM

Proposed system uses Machine learning algorithms to predict the rating of the application of google play store based on their features. Branch of Machine learning used here is supervised Learning which needs a human to “supervise” and tell the computer what it should be trained to predict for, or give it the right answer. We feed the computer with training data containing various features, and we also tell it the right answer. Supervised learning can solve two problems- Classification & Regression. For the said problem, regression is used so as to predict application rating. Machine learning problem in supervised learning can be solved in three stageswhich are - Data Preparation, Training & Testing, & evaluation of used models.

For the regression problem, most commonly used regression models are used which are - Linear Regression, SVM Model & Random forest regression model.

Linear regression is the simplest of regression model which is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). Support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyzedata used for classification and regression analysis. When used alone, decision trees are prone to overfitting. However, random forests help by correcting the possible overfitting that could occur. Random forests work by using multiple decision trees — using a multitude of different decision trees with different predictions, a random forest combines the results of those individual trees to give the final outcomes.

DATA COLLECTION AND DATA PREPARATION

Dataset for the said problem was collected from Kaggle & contains columns as shown-

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Ra	Genres	Last Updat	Current Ve	Android Ver	
2	Photo Edit	ART_AND_	4.1	159	19M	10,000+	Free	0	Everyone	Art & Desig	January 7,	1.0.0	4.0.3 and up	
3	Coloring b	ART_AND_	3.9	967	14M	500,000+	Free	0	Everyone	Art & Desig	January 15	2.0.0	4.0.3 and up	
4	U Launche	ART_AND_	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Desig	August 1, 2	1.2.4	4.0.3 and up	
5	Sketch - Dr	ART_AND_	4.5	215644	25M	50,000,000	Free	0	Teen	Art & Desig	June 8, 201	Varies with	4.2 and up	
6	Pixel Draw	ART_AND_	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Desig	June 20, 20	1.1	4.4 and up	
7	Paper flow	ART_AND_	4.4	167	5.6M	50,000+	Free	0	Everyone	Art & Desig	March 26,	1	2.3 and up	
8	Smoke Effe	ART_AND_	3.8	178	19M	50,000+	Free	0	Everyone	Art & Desig	April 26, 20	1.1	4.0.3 and up	
9	Infinite Pa	ART_AND_	4.1	36815	29M	1,000,000+	Free	0	Everyone	Art & Desig	June 14, 20	6.1.61.1	4.2 and up	
10	Garden Co	ART_AND_	4.4	13791	33M	1,000,000+	Free	0	Everyone	Art & Desig	September	2.9.2	3.0 and up	
11	Kids Paint	ART_AND_	4.7	121	3.1M	10,000+	Free	0	Everyone	Art & Desig	July 3, 201	2.8	4.0.3 and up	
12	Text on Ph	ART_AND_	4.4	13880	28M	1,000,000+	Free	0	Everyone	Art & Desig	October 2	1.0.4	4.1 and up	
13	Name Art	ART_AND_	4.4	8788	12M	1,000,000+	Free	0	Everyone	Art & Desig	July 31, 20	1.0.15	4.0 and up	
14	Tattoo Na	ART_AND_	4.2	44829	20M	10,000,000	Free	0	Teen	Art & Desig	April 2, 201	3.8	4.1 and up	
15	Mandala C	ART_AND_	4.6	4326	21M	100,000+	Free	0	Everyone	Art & Desig	June 26, 20	1.0.4	4.4 and up	
16	3D Color P	ART_AND_	4.4	1518	37M	100,000+	Free	0	Everyone	Art & Desig	August 3, 2	1.2.3	2.3 and up	
17	Learn To D	ART_AND_	3.2	55	2.7M	5,000+	Free	0	Everyone	Art & Desig	June 6, 201	NaN	4.2 and up	
18	Photo Des	ART_AND_	4.7	3632	5.5M	500,000+	Free	0	Everyone	Art & Desig	July 31, 20	3.1	4.1 and up	
19	350 Diy Ro	ART_AND_	4.5	27	17M	10,000+	Free	0	Everyone	Art & Desig	November	1	2.3 and up	
20	FlipaClip -	ART_AND_	4.3	194216	39M	5,000,000+	Free	0	Everyone	Art & Desig	August 3, 2	2.2.5	4.0.3 and up	
21	ibis Paint X	ART_AND_	4.6	224399	31M	10,000,000	Free	0	Everyone	Art & Desig	July 30, 20	5.5.4	4.1 and up	
22	Logo Make	ART_AND_	4	450	14M	100,000+	Free	0	Everyone	Art & Desig	April 20, 20	4	4.1 and up	
23	Boys Photo	ART_AND_	4.1	654	12M	100,000+	Free	0	Everyone	Art & Desig	March 20,	1.1	4.0.3 and up	
24	Superhero	ART_AND_	4.7	7699	4.2M	500,000+	Free	0	Everyone 1	Art & Desig	July 12, 20	2.2.6.2	4.0.3 and up	

Columns in the dataset are explained-

- App
  Application name
- Category
  Category the app belongs to
- Rating
  Overall user rating of the app
- Reviews
  Number of user reviews for the app
- Size
  Size of the app
- Installs
  Number of user downloads/installs for the app
- Type
  Paid or Free
- Price
  Price of the app
- Content Rating
  Age group the app is targeted at - Children / Mature 21+ / Adult

- Genres
  An app can belong to multiple genres
- Last Updated
  Date when the app was last updated on Play Store
- Current Ver
  Current version of the app available on Play Store
- Android Ver
  Min required Android version

In [7]:

```
import pandas as pd
```

In [8]:

```
import numpy as np
```

In [9]:

```
df=pd.read_csv('googleplaystore.csv')
```

Out[9]:

```
df.head()
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & Scrap Book	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

	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	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

In [10]:

```
df.shape
```

Out[10]:

```
(10841, 13)
```

In [11]:

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

In [12]:

```
df.isnull().sum()
```

Out[12]:

```
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 [13]:

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

In [14]:

```
df=df.dropna()
```

In [15]:

```
df.shape
```

Out[15]:

```
(9360, 13)
```

In [16]:

```
df.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
```

In [17]:

```
#Variables seem to have incorrect type and inconsistent formatting. You need to
 fix them:
```

In [18]:

```
df['Reviews']=df['Reviews'].astype("int64")
```

In [19]:

```
df.tail()
```

Out[19]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10834	FR Calcul	FAMILY	4.0	7	2.6M	500+	Free	0	Everyone	Education	June 18,	1.0.0	4.1 and

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
	ator										2017		up
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6M	100+	Free	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	Varies with device	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19M	10,000,000+	Free	0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

In [20]:

```
print(df['Price'])
```

```
0         0
1         0
2         0
3         0
4         0
        ..
10834     0
10836     0
10837     0
10839     0
10840     0
Name: Price, Length: 9360, dtype: object
```

In [21]:

```
df['Price']=df['Price'].str.replace('$','')
```

In [22]:

```
df['Price']=df['Price'].astype('float64')
```

In [23]:

```
df.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   object
 5   Installs        9360 non-null   object
 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(1), object(10)
memory usage: 1023.8+ KB
```

In [24]:

```
df['Installs']=df['Installs'].str.replace('+','')
```

In [25]:

```
df['Installs']=df['Installs'].str.replace(',','')
```

In [26]:

```
df['Installs'].tail()
```

Out[26]:

```
10834         500
10836        5000
10837         100
10839        1000
10840    10000000
Name: Installs, dtype: object
```

In [27]:

```
df['Installs']=df['Installs'].astype('int64')
```

In [28]:

```
df.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   object 
 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(2), object(9)
memory usage: 1023.8+ KB
```

In [29]:

```
df['Size'].tail(50)
```

Out[29]:

```
10768    24M
10770    41M
10771   2.4M
10776    24M
10777   2.2M
10778    38M
10779    75M
10780    50M
10781    44M
10782    11M
10783    72M
10784    84M
10785   9.5M
10786   2.8M
10787    48M
10789    48M
10790    20M
10791    38M
10792    16M
10793    78M
```

```
10795                  4.0M
10796                  7.8M
10797                   46M
10799                  6.8M
10800                   12M
10801                   19M
10802                   28M
10803                   81M
10804                   17M
10805                   15M
10809                   24M
10810                   21M
10812                   13M
10814                   31M
10815                  4.9M
10817                  8.0M
10819                  3.6M
10820                  8.6M
10826    Varies with device
10827                   13M
10828                   13M
10829                  7.4M
10830                  2.3M
10832                  582k
10833                  619k
10834                  2.6M
10836                   53M
10837                  3.6M
10839    Varies with device
10840                   19M
Name: Size, dtype: object
```

In [30]:

```
df.sort_values(by='Size')
```

Out[30]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
9973	German Vocabulary Trainer	FAMILY	3.3	1218	1.0M	100000	Free	0.00	Everyone	Education	August 24, 2012	1.0	2.1 and up
64	BL	TOOLS	4.3	33	1.0	500	Pa	3.9	Ever	Tools	Febru	2.6.15	2.3

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
46	PowerPoint Remote				M		id	9	yone		ary 25, 2015	0226	and up
8077	go41cx	FAMILY	4.8	171	1.0M	1000	Paid	10.00	Everyone	Education	October 18, 2016	1.7.0	2.2 and up
10043	Remote EX for NISSAN	COMMUNICATION	2.3	223	1.0M	5000	Paid	1.49	Everyone	Communication	July 26, 2014	1.3	3.0 and up
2788	Savory - Deals,Freebies,Sales	SHOPPING	4.4	2375	1.1M	100000	Free	0.00	Everyone	Shopping	November 2, 2016	3.1	2.3.3 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
5548	SNOW - AR Camera	PHOTOGRAPHY	4.3	1017237	Varies with device	50000000	Free	0.00	Everyone	Photography	July 30, 2018	7.6.5	4.3 and up
7024	Poly Art: Paint by Sticker, Color by Number Pu...	FAMILY	4.4	2100	Varies with device	100000	Free	0.00	Everyone	Entertainment;Brain Games	July 28, 2018	2.3	4.1 and up
3268	Google app for Android TV	TOOLS	3.0	66	Varies with device	10000000	Free	0.00	Everyone	Tools	July 19, 2018	Varies with device	Varies with device
3309	Unit Converter Pro	TOOLS	4.5	12718	Varies with	1000000	Free	0.00	Everyone	Tools	March 5, 2018	Varies with device	Varies with devi

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
					device								ce
7657	Co-Optima Mobile	FINANCE	4.4	218	Varies with device	10000	Free	0.00	Everyone	Finance	June 8, 2017	Varies with device	Varies with device

9360 rows × 13 columns

In [32]:

```
if 'M' in size:
        x = size[:-1]
        x = float(x)*1000
        return(x)
    elif 'K' in size:
        x = size[:-1]
        return(x)
    else:
        return None
```

```
  File "<tokenize>", line 5
    elif 'K' in size:
    ^
IndentationError: unindent does not match any outer indentation level
```

In [33]:

```
df.head()
```

Out[33]:

	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	10000	Free	0.0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	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	8.7M	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	25M	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	2.8M	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

In [34]:

```
df.df['Rating']>5.index
```

```
  File "<ipython-input-34-828879e219e3>", line 1
    df.df['Rating']>5.index
                           ^
SyntaxError: invalid syntax
```

In [35]:

```
#
#inp1.Reviews=inp1.Reviews.apply(np.log1p)
```

In [36]:

```
#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.
```

In [37]:

```
df[df['Rating']>5]
```

Out[37]:

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

In [38]:

```
df.head()
```

Out[38]:

	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	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	14M	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	8.7M	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	25M	50000000	Free	0.0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring	ART_AND_DESIGN	4.3	967	2.8M	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

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

In [39]:

```
if 'M' in df['Size']:
    df['Size']=Size[:-1]
    df['Size']=float(df['Size'])*1000
    return(df['Size'])
elif 'k' in df['Size']:
    df['Size']=size[:-1]
    return(df['Size'])
```

```
  File "<ipython-input-39-6b76d1856f8e>", line 4
    return(df['Size'])
    ^
SyntaxError: 'return' outside function
```

In [40]:

```
x=df['Size']
```

In [41]:

```
x.head()
```

Out[41]:

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

In [42]:

```
if 'M' in Size:
    x=Size[:-1]
    x=float(x)*1000
    return (x)
elif 'k' in Size:
    x=Size[:-1]
    return (x)
```

```
  File "<ipython-input-42-a3e813e7f397>", line 4
    return (x)
    ^
SyntaxError: 'return' outside function
```

In [43]:

```
df.head()
```

Out[43]:

	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	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	14M	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	8.7M	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	25M	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	2.8M	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

In [44]:

```
print(df["Rating"]>5)
```

```
0        False
1        False
2        False
3        False
```

```
4        False
         ...
10834    False
10836    False
10837    False
10839    False
10840    False
Name: Rating, Length: 9360, dtype: bool
```

In [45]:

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

In [46]:

```
df[df['Reviews']>df['Installs']].index
```

Out[46]:

```
Int64Index([2454, 4663, 5917, 6700, 7402, 8591, 10697], dtype='int64')
```

In [47]:

```
df.drop(index=df[df['Reviews']>df['Installs']].index)
```

Out[47]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & Scrap Book	ART_AND_DESIGN	4.1	159	19M	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	14M	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,	ART_AND_DESIGN	4.7	87510	8.7M	5000000	Free	0.0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
	Hide ...												
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	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	2.8M	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	7	2.6M	500	Free	0.0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	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	3.6M	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	Varies with device	1000	Free	0.0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope -	LIFESTYLE	4.5	398307	19M	10000000	Free	0.0	Everyone	Lifestyle	July 25,	Varies	Varies

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	2018 Daily Horoscope & Astrology										2018	with device	with device

9353 rows × 13 columns

In [48]:

```
df[df['Reviews']>df['Installs']]
```

Out[48]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
2454	KBA-EZ Health Guide	MEDICAL	5.0	4	25M	1	Free	0.00	Everyone	Medical	August 2, 2018	1.0.72	4.0.3 and up
4663	Alarmy (Sleep If U Can) - Pro	LIFESTYLE	4.8	10249	Varies with device	10000	Paid	2.49	Everyone	Lifestyle	July 30, 2018	Varies with device	Varies with device
5917	Ra Ga Ba	GAME	5.0	2	20M	1	Paid	1.49	Everyone	Arcade	February 8, 2017	1.0.4	2.3 and up
6700	Brick Breaker BR	GAME	5.0	7	19M	5	Free	0.00	Everyone	Arcade	July 23, 2018	1.0	4.1 and up
7402	Trovami se ci riesci	GAME	5.0	11	6.1M	10	Free	0.00	Everyone	Arcade	March 11, 2017	0.1	2.3 and up
859	DN	SOCIAL	5.0	20	4.2	10	Fre	0.0	Teen	Social	July	1.0	4.0

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
1	Blog				M		e	0			23, 2018		and up
10697	Mu.F.O.	GAME	5.0	2	16M	1	Paid	0.99	Everyone	Arcade	March 3, 2017	1.0	2.3 and up

In [49]:

```
df
```

Out[49]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & Scrap Book	ART_AND_DESIGN	4.1	159	19M	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	14M	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	8.7M	5000000	Free	0.0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw &	ART_AND_DESIGN	4.5	215644	25M	50000000	Free	0.0	Teen	Art & Design	June 8, 2018	Varies with devi	4.2 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
	Paint											ce	
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	7	2.6M	500	Free	0.0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	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	3.6M	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	Varies with device	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	19M	10000000	Free	0.0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

9360 rows × 13 columns

In [50]:

```
df.drop(index=df[df['Reviews']>df['Installs']].index)
```

Out[50]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & Scrap Book	ART_AND_DESIGN	4.1	159	19M	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	14M	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	8.7M	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	25M	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	2.8M	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	7	2.6M	500	Free	0.0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	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	3.6M	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	Varies with device	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	19M	10000000	Free	0.0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

9353 rows × 13 columns

In [51]:

```
df[df['Reviews']>df['Installs']]
```

Out[51]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
2454	KBA-EZ Health Guide	MEDICAL	5.0	4	25M	1	Free	0.00	Everyone	Medical	August 2, 2018	1.0.72	4.0.3 and up
4663	Alarmy (Sleep If U Can) - Pro	LIFESTYLE	4.8	10249	Varies with device	10000	Paid	2.49	Everyone	Lifestyle	July 30, 2018	Varies with device	Varies with device
5917	Ra Ga Ba	GAME	5.0	2	20M	1	Paid	1.49	Everyone	Arcade	February 8, 2017	1.0.4	2.3 and up
6700	Brick Breaker BR	GAME	5.0	7	19M	5	Free	0.00	Everyone	Arcade	July 23, 2018	1.0	4.1 and up
7402	Trovami se ci riesci	GAME	5.0	11	6.1M	10	Free	0.00	Everyone	Arcade	March 11, 2017	0.1	2.3 and up
8591	DN Blog	SOCIAL	5.0	20	4.2M	10	Free	0.00	Teen	Social	July 23, 2018	1.0	4.0 and up
10697	Mu.F.O.	GAME	5.0	2	16M	1	Paid	0.99	Everyone	Arcade	March 3, 2017	1.0	2.3 and up

In [52]:

```
df
```

Out[52]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo	ART_AND_DES	4.1	159	19	1000	Fr	0.0	Ever	Art &	Janu	1.0.0	4.0.3

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
	Editor & Candy Camera & Grid & Scrap Book	IGN			M	0	ee		yone	Design	ary 7, 2018		and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14 M	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	8.7 M	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	25 M	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	2.8 M	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	7	2.6 M	500	Free	0.0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53M	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	3.6M	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	Varies with device	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	19M	10000000	Free	0.0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

9360 rows × 13 columns

In [53]:

```
df=df.drop(index=df[df['Reviews']>df['Installs']].index)
```

In [54]:

```
df
```

Out[54]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy	ART_AND_DESIGN	4.1	159	19M	10000	Free	0.0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
	Camera & Grid & Scrap Book												
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	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	8.7M	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	25M	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	2.8M	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10834	FR Calculator	FAMILY	4.0	7	2.6M	500	Free	0.0	Everyone	Education	June 18, 2017	1.0.0	4.1 and up
10	Sya9a	FAMILY	4.5	38	53	5000	Fr	0.0	Ever	Education	July	1.48	4.1

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
836	Maroc - FR				M		ee		yone		25, 2017		and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3.6 M	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	Varies with device	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	19 M	10000000	Free	0.0	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

9353 rows × 13 columns

In [55]:

```
df[df['Reviews']>df['Installs']].index
```

Out[55]:

```
Int64Index([], dtype='int64')
```

In [56]:

```
df[df['Rating']>5]
```

Out[56]:

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

In [57]:

```
df.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   object
 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(2), object(9)
memory usage: 1023.0+ KB
```

In [58]:

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

In [59]:

```
df[(df.Type=='Free') & (df.Price>0.0)]
```

Out[59]:

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

In [60]:

```
df[df.Rating>5].index
```

Out[60]:

```
Int64Index([], dtype='int64')
```

In [61]:

```
def mb_to_kb(a):
  if a.endswith("M"):
    return float(a[:-1])*1000
  elif a.endswith("k"):
    return float(a[:-1])
  else:
    return a
```

In [62]:

```
df["Size"]=df["Size"].apply(lambda a:mb_to_kb(a))
```

In [63]:

```
df[df['Size']=='Varies with device'].index
```

Out[63]:

```
Int64Index([   37,    42,    52,    67,    68,    73,    85,    88,    89,
               92,
            ...
            10647, 10679, 10681, 10707, 10712, 10713, 10725, 10765, 10826,
            10839],
           dtype='int64', length=1636)
```

In [64]:

```
df
```

Out[64]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & Scrap Book	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	ART_AND_DES	4.3	967	280	1000	Fr	0.0	Ever	Art &	June	1.1	4.4

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
	Draw-Number Art Coloring Book	IGN			0	00	ee		yone	Design;Creativity	20, 2018		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	Varies with device	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

9353 rows × 13 columns

In [65]:

```
df=df.drop(index=df[df['Size']=='Varies with device'].index)
```

In [66]:

```
df
```

Out[66]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & Scrap Book	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

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10833	Chemin (fr)	BOOKS_AND_REFERENCE	4.8	44	619	1000	Free	0.0	Everyone	Books & Reference	March 23, 2014	0.8	2.2 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
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

7717 rows × 13 columns

In [67]:

```
df[df['Size']=='Varies with device']
```

Out[67]:

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

In [68]:

```
df.tail(50)
```

Out[68]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10766	FreedomPop Diagnostics	TOOLS	2.9	452	7000	100000	Free	0.00	Everyone	Tools	July 17, 2017	1.03.123.0713	4.0.3 and up
10767	NFP 2018	EVENTS	4.8	8	16000	500	Free	0.00	Everyone	Events	January 9, 2018	1.0.3	4.2 and up
10768	AAFP	MEDICAL	3.8	63	24000	10000	Free	0.00	Everyone	Medical	June 22, 2018	2.3.1	5.0 and up
10770	Modern Counter Terrorist FPS Shoot	GAME	4.0	795	41000	100000	Free	0.00	Teen	Action	August 29, 2017	1.2	2.3 and up
10771	FQ METER	PRODUCTIVITY	3.9	17	2400	1000	Free	0.00	Everyone	Productivity	April 23, 2017	1.1	4.0 and up
10776	Monster Ride Pro	GAME	5.0	1	24000	10	Free	0.00	Everyone	Racing	March 5, 2018	2.0	2.3 and up
10777	BEBON COOL GAME PAD V1.0	GAME	3.9	404	2200	100000	Free	0.00	Everyone	Arcade	August 30, 2017	1.2	4.0 and up
10778	Union League	FAMILY	4.0	939	38000	10000	Free	0.00	Everyone	Role Playing	June 8, 2018	1.0.0.96	4.1 and up
10779	Fortune Quest: Savior	FAMILY	3.6	135	75000	10000	Free	0.00	Everyone 10+	Role Playing	June 1, 2018	1.022	4.4 and up
1078	Modern Counter	FAMILY	4.1	17	5000	1000	Free	0.00	Everyone	Strategy	March 16,	15	4.1 and

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	3: FPS Multiplayers battlegro 3				0						2018		up
10781	Modern Strike Online	GAME	4.3	834117	44000	10000000	Free	0.00	Teen	Action	July 30, 2018	1.25.4	4.1 and up
10782	Trine 2: Complete Story	GAME	3.8	252	11000	10000	Paid	16.99	Teen	Action	February 27, 2015	2.22	5.0 and up
10783	Modern Counter Terror Attack – Shooting Game	GAME	4.2	340	72000	50000	Free	0.00	Mature 17+	Action	October 27, 2017	1.0	4.1 and up
10784	Big Hunter	GAME	4.3	245455	84000	10000000	Free	0.00	Everyone 10+	Action	May 31, 2018	2.8.6	4.0 and up
10785	sugar, sugar	FAMILY	4.2	1405	9500	10000	Paid	1.20	Everyone	Puzzle	June 5, 2018	2.7	2.3 and up
10786	ChopAssistant	TOOLS	4.2	455	2800	50000	Free	0.00	Everyone	Tools	February 28, 2017	1.6	6.0 and up
10787	Modern Counter Global Strike 3D	GAME	4.1	297	48000	50000	Free	0.00	Teen	Action	March 28, 2018	1.2	4.1 and up
10789	Modern Counter Global Strike	GAME	4.0	368	48000	50000	Free	0.00	Everyone 10+	Action	March 28, 2018	1.7	4.1 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
	3D V2												
10790	HipChat - beta version	COMMUNICATION	4.1	1035	20000	50000	Free	0.00	Everyone	Communication	August 7, 2018	3.20.001	4.1 and up
10791	Winter Wonderland	GAME	4.0	1287	38000	50000	Free	0.00	Everyone	Word	December 18, 2013	1.0	2.2 and up
10792	Soccer Clubs Logo Quiz	GAME	4.2	21661	16000	1000000	Free	0.00	Everyone	Trivia	May 24, 2018	1.3.81	4.0 and up
10793	Sid Story	GAME	4.4	28510	78000	500000	Free	0.00	Teen	Card	August 1, 2018	2.6.6	4.0.3 and up
10795	Reindeer VPN - Proxy VPN	TOOLS	4.2	7339	4000	100000	Free	0.00	Everyone	Tools	May 10, 2018	1.74	4.1 and up
10796	Inf VPN - Global Proxy & Unlimited Free WIFI VPN	TOOLS	4.7	61445	7800	1000000	Free	0.00	Everyone	Tools	July 26, 2018	1.9.734	4.1 and up
10797	Fuel Rewards® program	LIFESTYLE	4.6	32433	46000	1000000	Free	0.00	Everyone	Lifestyle	June 26, 2018	2.9.1	5.0 and up
10799	Fr Daoud Lamei	SOCIAL	4.7	2036	6800	100000	Free	0.00	Everyone	Social	May 20, 2018	1.72	4.0.3 and up
1080	FR Roster	TOOLS	4.1	174	1200	5000	Free	0.00	Everyone	Tools	July 30,	6.04	4.4 and

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0					0						2018		up
10801	Fr Ignacio Outreach	FAMILY	4.9	52	19000	1000	Free	0.00	Everyone	Education	January 19, 2018	1.0	4.4 and up
10802	FR: My Famous Lover	FAMILY	4.0	185	28000	10000	Free	0.00	Teen	Entertainment	August 6, 2015	1.3.0	3.0 and up
10803	Fatal Raid - No.1 Mobile FPS	GAME	4.3	56496	81000	1000000	Free	0.00	Teen	Action	August 7, 2018	1.5.447	4.0 and up
10804	Poker Pro.Fr	GAME	4.2	5442	17000	100000	Free	0.00	Teen	Card	May 22, 2018	4.1.3	2.3 and up
10805	Scoreboard FR	LIFESTYLE	4.3	3	15000	100	Free	0.00	Everyone	Lifestyle	August 7, 2018	2.1	4.2 and up
10809	Castle Clash: RPG War and Strategy FR	FAMILY	4.7	376223	24000	1000000	Free	0.00	Everyone	Strategy	July 18, 2018	1.4.2	4.1 and up
10810	Fr Lupupa Sermons	BUSINESS	4.8	19	21000	100	Free	0.00	Everyone	Business	June 12, 2018	1.0	4.4 and up
10812	Fr Agnel Pune	FAMILY	4.1	80	13000	1000	Free	0.00	Everyone	Education	June 13, 2018	2.0.20	4.0.3 and up
10814	FR: My Secret Pets!	FAMILY	4.0	785	31000	50000	Free	0.00	Teen	Entertainment	June 3, 2015	1.3.1	3.0 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10815	Golden Dictionary (FR-AR)	BOOKS_AND_REFERENCE	4.2	5775	4900	500000	Free	0.00	Everyone	Books & Reference	July 19, 2018	7.0.4.6	4.2 and up
10817	HTC Sense Input - FR	TOOLS	4.0	885	8000	100000	Free	0.00	Everyone	Tools	October 30, 2015	1.0.612928	5.0 and up
10819	Fanfic-FR	BOOKS_AND_REFERENCE	3.3	52	3600	5000	Free	0.00	Teen	Books & Reference	August 5, 2017	0.3.4	4.1 and up
10820	Fr. Daoud Lamei	FAMILY	5.0	22	8600	1000	Free	0.00	Teen	Education	June 27, 2018	3.8.0	4.1 and up
10827	Fr Agnel Ambarnath	FAMILY	4.2	117	13000	5000	Free	0.00	Everyone	Education	June 13, 2018	2.0.20	4.0.3 and up
10828	Manga-FR - Anime Vostfr	COMICS	3.4	291	13000	10000	Free	0.00	Everyone	Comics	May 15, 2017	2.0.1	4.0 and up
10829	Bulgarian French Dictionary Fr	BOOKS_AND_REFERENCE	4.6	603	7400	10000	Free	0.00	Everyone	Books & Reference	June 19, 2016	2.96	4.1 and up
10830	News Minecraft.fr	NEWS_AND_MAGAZINES	3.8	881	2300	100000	Free	0.00	Everyone	News & Magazines	January 20, 2014	1.5	1.6 and up
10832	FR Tides	WEATHER	3.8	1195	582	100000	Free	0.00	Everyone	Weather	February 16, 2014	6.0	2.1 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
10833	Chemin (fr)	BOOKS_AND_REFERENCE	4.8	44	619	1000	Free	0.00	Everyone	Books & Reference	March 23, 2014	0.8	2.2 and up
10834	FR Calculator	FAMILY	4.0	7	2600	500	Free	0.00	Everyone	Education	June 18, 2017	1.0.0	4.1 and up
10836	Sya9a Maroc - FR	FAMILY	4.5	38	53000	5000	Free	0.00	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600	100	Free	0.00	Everyone	Education	July 6, 2018	1.0	4.1 and up
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307	19000	10000000	Free	0.00	Everyone	Lifestyle	July 25, 2018	Varies with device	Varies with device

In [69]:

```
#boxplot for price
```

In [70]:

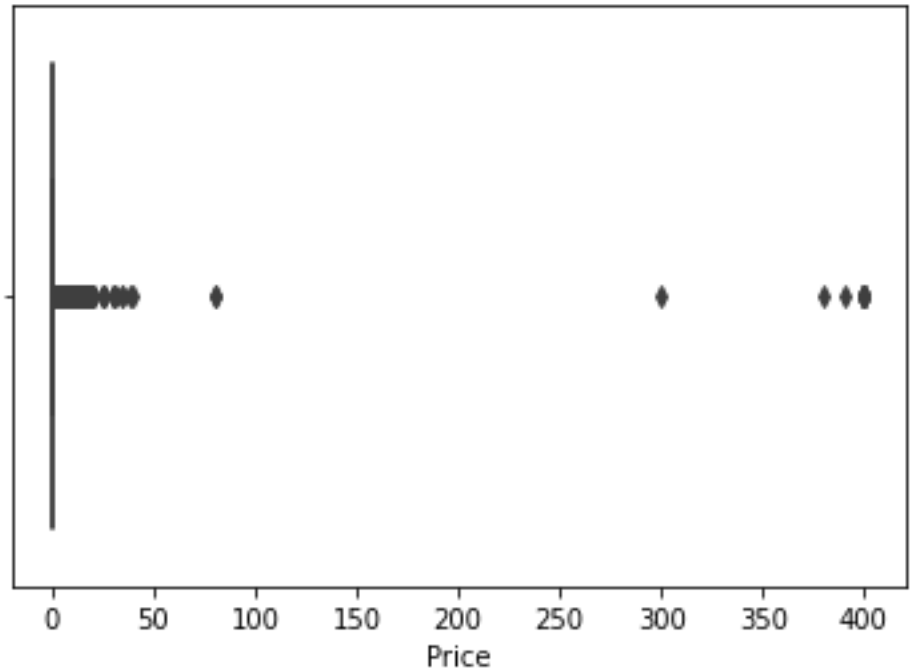
```
import seaborn as sns
```

In [71]:

```
sns.boxplot(df['Price'])
```

Out[71]:

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



In [72]:

```
df.describe()
```

Out[72]:

	Rating	Reviews	Installs	Price
count	7717.000000	7.717000e+03	7.717000e+03	7717.000000
mean	4.173293	2.951275e+05	8.430620e+06	1.128725
std	0.544362	1.864640e+06	5.017636e+07	17.414784
min	1.000000	1.000000e+00	5.000000e+00	0.000000
25%	4.000000	1.090000e+02	1.000000e+04	0.000000
50%	4.300000	2.351000e+03	1.000000e+05	0.000000
75%	4.500000	3.910900e+04	1.000000e+06	0.000000
max	5.000000	4.489389e+07	1.000000e+09	400.000000

In [73]:

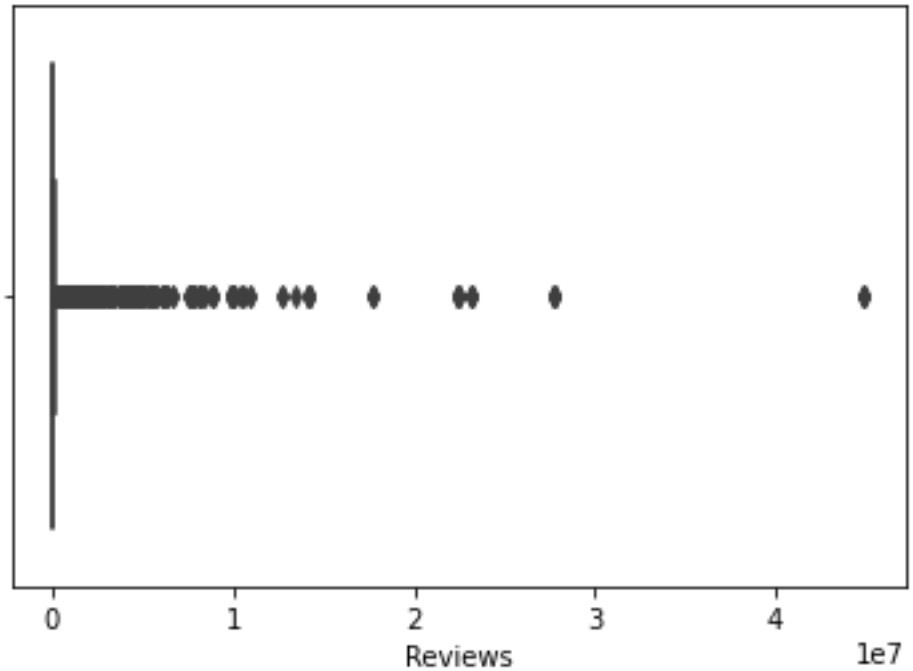
```
#Boxplot for Reviews
```

In [74]:

```
sns.boxplot(df['Reviews'])
```

Out[74]:

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



In [75]:

```
Q1=1.090000e+02
Q3=3.910900e+04
IQR=Q3-Q1
print('Q1 value:',Q1)
print('Q3 value:',Q3)
print(IQR)
outliers=df[(df['Reviews']>Q3+1.5*IQR)|(df['Reviews']<Q1-1.5*IQR)]
print(outliers)
```

```
Q1 value: 109.0
Q3 value: 39109.0
39000.0
                                                App           Category  \
3                                 Sketch - Draw & Paint     ART_AND_DESIGN
18                         FlipaClip - Cartoon animation     ART_AND_DESIGN
19                                            ibis Paint X     ART_AND_DESIGN
45     Canva: Poster, banner, card maker & graphic de...     ART_AND_DESIGN
70     Fines of the State Traffic Safety Inspectorate...  AUTO_AND_VEHICLES
...                                                  ...                ...
10740                                          PhotoFunia        PHOTOGRAPHY
10781                               Modern Strike Online               GAME
10784                                          Big Hunter               GAME
```

```
10809              Castle Clash: RPG War and Strategy FR             FAMILY
10840      iHoroscope - 2018 Daily Horoscope & Astrology          LIFESTYLE

       Rating  Reviews   Size  Installs  Type  Price Content Rating  \
3         4.5   215644  25000  50000000  Free    0.0           Teen
18        4.3   194216  39000   5000000  Free    0.0       Everyone
19        4.6   224399  31000  10000000  Free    0.0       Everyone
45        4.7   174531  24000  10000000  Free    0.0       Everyone
70        4.8   116986  35000   5000000  Free    0.0       Everyone
...       ...      ...    ...       ...   ...    ...            ...
10740     4.3   316378   4400  10000000  Free    0.0       Everyone
10781     4.3   834117  44000  10000000  Free    0.0           Teen
10784     4.3   245455  84000  10000000  Free    0.0   Everyone 10+
10809     4.7   376223  24000   1000000  Free    0.0       Everyone
10840     4.5   398307  19000  10000000  Free    0.0       Everyone

                Genres    Last Updated         Current Ver         Android Ver

3         Art & Design    June 8, 2018  Varies with device          4.2 and up

18        Art & Design  August 3, 2018               2.2.5        4.0.3 and up

19        Art & Design   July 30, 2018               5.5.4          4.1 and up

45        Art & Design   July 31, 2018               1.6.1          4.1 and up

70     Auto & Vehicles  August 2, 2018               1.9.7        4.0.3 and up

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

10740      Photography    June 3, 2017             4.0.7.0          2.3 and up

10781           Action   July 30, 2018              1.25.4          4.1 and up

10784           Action    May 31, 2018               2.8.6          4.0 and up

10809         Strategy   July 18, 2018               1.4.2          4.1 and up

10840        Lifestyle   July 25, 2018  Varies with device  Varies with device


[1326 rows x 13 columns]
```

In [76]:

```
df1=sns.load_dataset('tips')
```

In [77]:

```
df1.describe()
```

Out[77]:

	total_bill	tip	size
count	244.000000	244.000000	244.000000
mean	19.785943	2.998279	2.569672
std	8.902412	1.383638	0.951100
min	3.070000	1.000000	1.000000
25%	13.347500	2.000000	2.000000
50%	17.795000	2.900000	2.000000
75%	24.127500	3.562500	3.000000
max	50.810000	10.000000	6.000000

In [78]:

```
Q1=2.000000
Q3=3.562500
IQR=Q3-Q1
print('Q1 value:',Q1)
print('Q3 value:',Q3)
print(IQR)
outliers=df1[(df1['tip']>Q3+1.5*IQR)|(df1['tip']<Q1-1.5*IQR)]
print(outliers)
```

```
Q1 value: 2.0
Q3 value: 3.5625
1.5625
     total_bill    tip     sex smoker   day    time  size
23        39.42   7.58    Male     No   Sat  Dinner     4
47        32.40   6.00    Male     No   Sun  Dinner     4
59        48.27   6.73    Male     No   Sat  Dinner     4
141       34.30   6.70    Male     No  Thur   Lunch     6
170       50.81  10.00    Male    Yes   Sat  Dinner     3
183       23.17   6.50    Male    Yes   Sun  Dinner     4
212       48.33   9.00    Male     No   Sat  Dinner     4
```

```
214       28.17   6.50  Female    Yes   Sat  Dinner     3
239       29.03   5.92    Male     No   Sat  Dinner     3
```

In [79]:

```
#histogram for rating
```

In [80]:

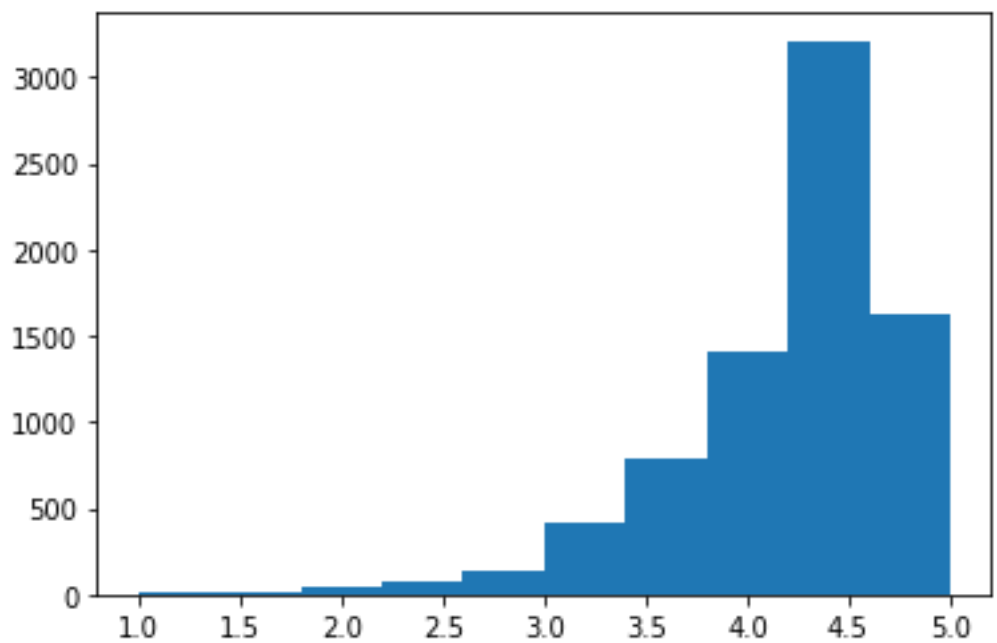
```
import matplotlib.pyplot as plt
```

In [81]:

```
plt.hist(df['Rating'])
```

Out[81]:

```
(array([  17.,   18.,   39.,   72.,  132.,  408.,  781., 1406., 3212.,
        1632.]),
 array([1. , 1.4, 1.8, 2.2, 2.6, 3. , 3.4, 3.8, 4.2, 4.6, 5. ]),
 <BarContainer object of 10 artists>)
```

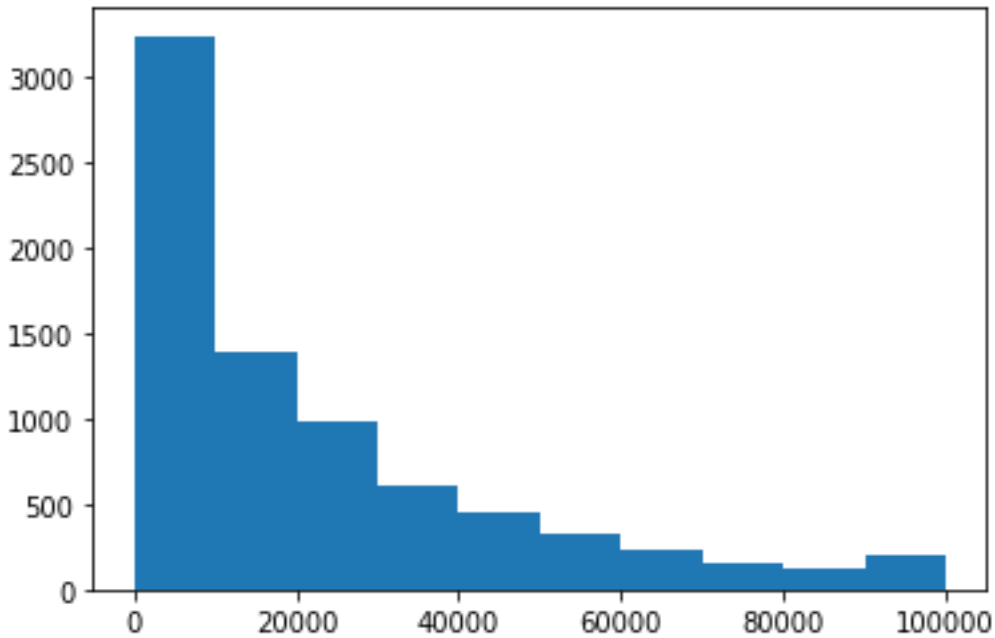


In [82]:

```
#histogram for size
plt.hist(df['Size'])
```

Out[82]:

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



In [83]:

```
#Apps with price more than $200
df[df.Price>200]
```

Out[83]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
4197	most expensive app (H)	FAMILY	4.3	6	1500	100	Paid	399.99	Everyone	Entertainment	July 16, 2018	1.0	7.0 and up
4362	💎 I'm rich	LIFESTYLE	3.8	718	26000	10000	Paid	399.99	Everyone	Lifestyle	March 11, 2018	1.0.0	4.4 and up
4367	I'm Rich - Trump Edition	LIFESTYLE	3.6	275	7300	10000	Paid	400.00	Everyone	Lifestyle	May 3, 2018	1.0.1	4.1 and up
5351	I am rich	LIFESTYLE	3.8	3547	1800	100000	Paid	399.99	Everyone	Lifestyle	January 12, 2018	2.0	4.0.3 and up
5354	I am Rich Plus	FAMILY	4.0	856	8700	10000	Paid	399.99	Everyone	Entertainment	May 19, 2018	3.0	4.4 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
5355	I am rich VIP	LIFESTYLE	3.8	411	2600	10000	Paid	299.99	Everyone	Lifestyle	July 21, 2018	1.1.1	4.3 and up
5356	I Am Rich Premium	FINANCE	4.1	1867	4700	50000	Paid	399.99	Everyone	Finance	November 12, 2017	1.6	4.0 and up
5357	I am extremely Rich	LIFESTYLE	2.9	41	2900	1000	Paid	379.99	Everyone	Lifestyle	July 1, 2018	1.0	4.0 and up
5358	I am Rich!	FINANCE	3.8	93	22000	1000	Paid	399.99	Everyone	Finance	December 11, 2017	1.0	4.1 and up
5359	I am rich(premium)	FINANCE	3.5	472	965	5000	Paid	399.99	Everyone	Finance	May 1, 2017	3.4	4.4 and up
5362	I Am Rich Pro	FAMILY	4.4	201	2700	5000	Paid	399.99	Everyone	Entertainment	May 30, 2017	1.54	1.6 and up
5364	I am rich (Most expensive app)	FINANCE	4.1	129	2700	1000	Paid	399.99	Teen	Finance	December 6, 2017	2	4.0.3 and up
5366	I Am Rich	FAMILY	3.6	217	4900	10000	Paid	389.99	Everyone	Entertainment	June 22, 2018	1.5	4.2 and up
5369	I am Rich	FINANCE	4.3	180	3800	5000	Paid	399.99	Everyone	Finance	March 22, 2018	1.0	4.2 and up
5373	I AM RICH PRO PLUS	FINANCE	4.0	36	41000	1000	Paid	399.99	Everyone	Finance	June 25, 2018	1.0.2	4.1 and up

In [84]:

```
#to drop junk apps with price more than $200
df=df.drop(index=df[df.Price>200].index)
```

In [85]:

```
df
```

Out[85]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & Scrap Book	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 Colori	ART_AND_DESIGN	4.3	967	2800	100000	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
	ng Book												
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10833	Chemin (fr)	BOOKS_AND_REFERENCE	4.8	44	619	1000	Free	0.0	Everyone	Books & Reference	March 23, 2014	0.8	2.2 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
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

7702 rows × 13 columns

In [86]:

```
#Apps with more than 2 million reviews
df[df.Reviews>2000000]
```

Out[86]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
345	Yahoo Mail – Stay Organized	COMMUNICATION	4.3	4187998	16000	100000000	Free	0.0	Everyone	Communication	July 18, 2018	5.29.3	4.4 and up
347	imo free video calls and chat	COMMUNICATION	4.3	4785892	11000	500000000	Free	0.0	Everyone	Communication	June 8, 2018	9.8.000000010501	4.0 and up
366	UC Browser Mini -Tiny Fast Private & Secure	COMMUNICATION	4.4	3648120	3300	100000000	Free	0.0	Teen	Communication	July 18, 2018	11.4.0	4.0 and up
378	UC Browser - Fast Download Private & Secure	COMMUNICATION	4.5	17712922	40000	500000000	Free	0.0	Teen	Communication	August 2, 2018	12.8.5.1121	4.0 and up
383	imo free video calls and chat	COMMUNICATION	4.3	4785988	11000	500000000	Free	0.0	Everyone	Communication	June 8, 2018	9.8.000000010501	4.0 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
91	Need	GAME	4.4	3344	22	50000	Fr	0.0	Ever	Racing	July	2.12.1	4.1

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
42	for Speed ™ No Limits			300	000	000	ee		yone 10+		24, 2018		and up
9166	Modern Combat 5: eSports FPS	GAME	4.3	2903386	58000	100000000	Free	0.0	Mature 17+	Action	July 24, 2018	3.2.1c	4.0 and up
10186	Farm Heroes Saga	FAMILY	4.4	7615646	71000	100000000	Free	0.0	Everyone	Casual	August 7, 2018	5.2.6	2.3 and up
10190	Fallout Shelter	FAMILY	4.6	2721923	25000	10000000	Free	0.0	Teen	Simulation	June 11, 2018	1.13.12	4.1 and up
10327	Garena Free Fire	GAME	4.5	5534114	53000	100000000	Free	0.0	Teen	Action	August 3, 2018	1.21.0	4.0.3 and up

219 rows × 13 columns

In [87]:

```
#Drop records having more than 2 million reviews.
```

In [88]:

```
df[df.Reviews>2000000].index
```

Out[88]:

```
Int64Index([  345,   347,   366,   378,   383,   395,   413,   419,   420,
              452,
            ...
             8399,  8445,  8894,  8896,  9140,  9142,  9166, 10186, 10190,
            10327],
           dtype='int64', length=219)
```

In [89]:

```
df=df.drop(index=df[df.Reviews>2000000].index)
```

In [90]:

```
df
```

Out[90]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & Scrap Book	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

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10833	Chemin (fr)	BOOKS_AND_REFERENCE	4.8	44	619	1000	Free	0.0	Everyone	Books & Reference	March 23, 2014	0.8	2.2 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
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

7483 rows × 13 columns

In [91]:

```
df.index
```

Out[91]:

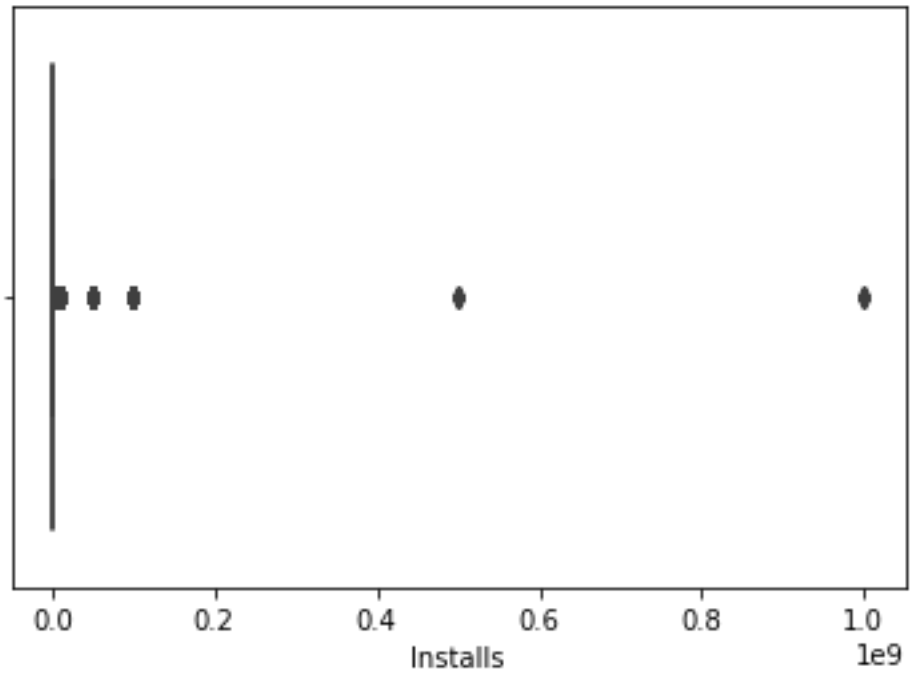
```
Int64Index([     0,      1,      2,      3,      4,      5,      6,      7,      8,
            9,
            ...
            10827, 10828, 10829, 10830, 10832, 10833, 10834, 10836, 10837,
            10840],
           dtype='int64', length=7483)
```

In [92]:

```
#Boxplot for Installs
sns.boxplot(df['Installs'])
```

Out[92]:

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



In [93]:

```
df.describe()
```

Out[93]:

	Rating	Reviews	Installs	Price
count	7483.000000	7.483000e+03	7.483000e+03	7483.000000
mean	4.165789	7.260651e+04	3.947465e+06	0.379595
std	0.549946	2.123720e+05	2.781831e+07	2.381384
min	1.000000	1.000000e+00	5.000000e+00	0.000000
25%	4.000000	9.900000e+01	1.000000e+04	0.000000
50%	4.300000	2.026000e+03	1.000000e+05	0.000000
75%	4.500000	3.238600e+04	1.000000e+06	0.000000

	Rating	Reviews	Installs	Price
max	5.000000	1.986068e+06	1.000000e+09	79.990000

In [94]:
```
df['Installs'].quantile(0.25)
```
Out[94]:
```
10000.0
```
In [95]:
```
df['Installs'].quantile(0.10)
```
Out[95]:
```
1000.0
```
In [96]:
```
df['Installs'].quantile(0.50)
```
Out[96]:
```
100000.0
```
In [97]:
```
df['Installs'].quantile(0.70)
```
Out[97]:
```
1000000.0
```
In [98]:
```
df['Installs'].quantile(0.90)
```
Out[98]:
```
10000000.0
```
In [99]:
```
df['Installs'].quantile(0.95)
```
Out[99]:
```
10000000.0
```
In [100]:
```
df['Installs'].quantile(0.99)
```
Out[100]:
```
50000000.0
```
In [101]:
```
Q1=1.000000e+04
Q3=1.000000e+06
IQR=Q3-Q1
print(Q1)
print(Q3)
print(IQR)
outliers=df[(df['Installs']>Q3+1.5*IQR)|(df['Installs']<Q1-1.5*IQR)]
print(outliers)
```

```
10000.0
1000000.0
990000.0
                                                App        Category  \
2      U Launcher Lite – FREE Live Cool Themes, Hide ...  ART_AND_DESIGN
3                                 Sketch - Draw & Paint  ART_AND_DESIGN
12                       Tattoo Name On My Photo Editor  ART_AND_DESIGN
18                        FlipaClip - Cartoon animation  ART_AND_DESIGN
19                                            ibis Paint X  ART_AND_DESIGN
...                                                   ...             ...
10731                     FeaturePoints: Free Gift Cards          FAMILY
10740                                          PhotoFunia     PHOTOGRAPHY
10781                               Modern Strike Online            GAME
10784                                          Big Hunter            GAME
10840      iHoroscope - 2018 Daily Horoscope & Astrology       LIFESTYLE

       Rating  Reviews   Size  Installs  Type  Price Content Rating  \
2         4.7    87510   8700   5000000  Free    0.0       Everyone
3         4.5   215644  25000  50000000  Free    0.0           Teen
12        4.2    44829  20000  10000000  Free    0.0           Teen
18        4.3   194216  39000   5000000  Free    0.0       Everyone
19        4.6   224399  31000  10000000  Free    0.0       Everyone
...       ...      ...    ...       ...   ...    ...            ...
10731     3.9   121321  46000   5000000  Free    0.0       Everyone
10740     4.3   316378   4400  10000000  Free    0.0       Everyone
10781     4.3   834117  44000  10000000  Free    0.0           Teen
10784     4.3   245455  84000  10000000  Free    0.0   Everyone 10+
10840     4.5   398307  19000  10000000  Free    0.0       Everyone

              Genres      Last Updated         Current Ver         Android Ver

2       Art & Design    August 1, 2018               1.2.4        4.0.3 and up

3       Art & Design      June 8, 2018  Varies with device          4.2 and up

12      Art & Design     April 2, 2018                 3.8          4.1 and up

18      Art & Design    August 3, 2018               2.2.5        4.0.3 and up

19      Art & Design     July 30, 2018               5.5.4          4.1 and up

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

10731  Entertainment  October 22, 2016                 8.7        4.0.3 and up
```

```
10740    Photography      June 3, 2017             4.0.7.0          2.3 and up

10781         Action     July 30, 2018              1.25.4          4.1 and up

10784         Action      May 31, 2018               2.8.6          4.0 and up

10840      Lifestyle     July 25, 2018  Varies with device  Varies with device


[1529 rows x 13 columns]
```

In [102]:

```
#max value 100000000 is considere as cutoff for outlier and drop records having
 values more than that
```

In [103]:

```
df[df.Installs>100000000]
```

Out[103]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
3473	Dropbox	PRODUCTIVITY	4.4	1861310	61000	500000000	Free	0.0	Everyone	Productivity	August 1, 2018	Varies with device	Varies with device
3569	Dropbox	PRODUCTIVITY	4.4	1861309	61000	500000000	Free	0.0	Everyone	Productivity	August 1, 2018	Varies with device	Varies with device
3736	Google News	NEWS_AND_MAGAZINES	3.9	877635	13000	1000000000	Free	0.0	Teen	News & Magazines	August 1, 2018	5.2.0	4.4 and up
3765	Google News	NEWS_AND_MAGAZINES	3.9	877635	13000	1000000000	Free	0.0	Teen	News & Magazines	August 1, 2018	5.2.0	4.4 and up
3816	Google News	NEWS_AND_MAGAZINES	3.9	877643	13000	1000000000	Free	0.0	Teen	News & Magazines	August 1, 2018	5.2.0	4.4 and up

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
4048	Dropbox	PRODUCTIVITY	4.4	1860844	61000	500000000	Free	0.0	Everyone	Productivity	August 1, 2018	Varies with device	Varies with device
5596	Samsung Health	HEALTH_AND_FITNESS	4.3	480208	70000	500000000	Free	0.0	Everyone	Health & Fitness	July 31, 2018	5.17.2.009	5.0 and up
9844	Google News	NEWS_AND_MAGAZINES	3.9	878065	13000	1000000000	Free	0.0	Teen	News & Magazines	August 1, 2018	5.2.0	4.4 and up

In [104]:

```
#Decide a threshold as cutoff for outlier and drop records having values more t
han that
```

In [105]:

```
df[df.Installs>100000000].index
```

Out[105]:

```
Int64Index([3473, 3569, 3736, 3765, 3816, 4048, 5596, 9844], dtype='int64')
```

In [106]:

```
df=df.drop(index=df[df.Installs>100000000].index)
```

In [107]:

```
df.index
```

Out[107]:

```
Int64Index([    0,     1,     2,     3,     4,     5,     6,     7,     8,
                9,
            ...
            10827, 10828, 10829, 10830, 10832, 10833, 10834, 10836, 10837,
            10840],
           dtype='int64', length=7475)
```

In [ ]:

In [108]:

```
#Bivariate Analysis
```

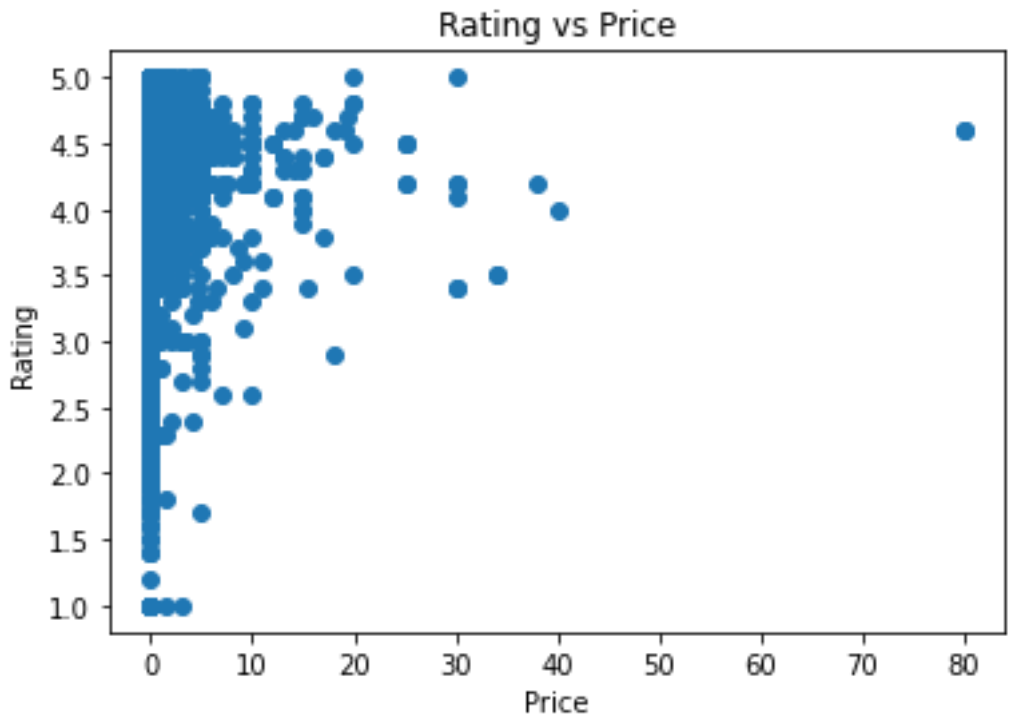
In [109]:

```
#Scatter plot for Rating vs Price
```

```
plt.scatter(df['Price'],df['Rating'])
plt.xlabel('Price')
plt.ylabel('Rating')
plt.title('Rating vs Price')
```

Out[109]:

```
Text(0.5, 1.0, 'Rating vs Price')
```



In [110]:

```
#No, Rating doesn't increase with price
```

In [111]:

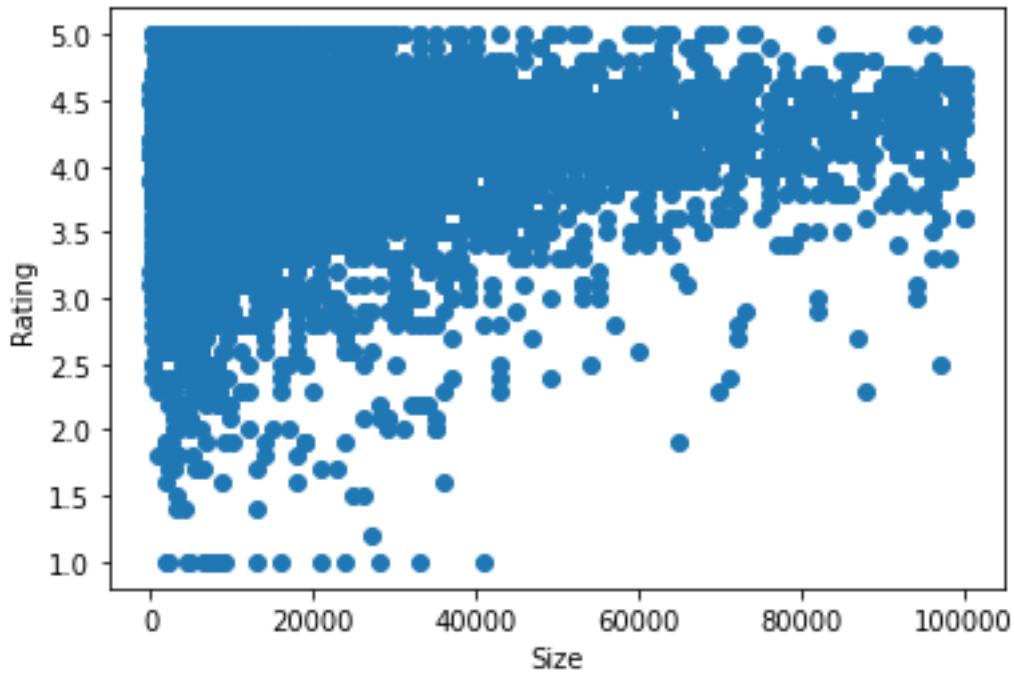
```
#Scatter plot for Rating vs Size
```

In [112]:

```
plt.scatter(df['Size'],df['Rating'])
plt.xlabel('Size')
plt.ylabel('Rating')
```

Out[112]:

```
Text(0, 0.5, 'Rating')
```



In [113]:

```
# No it coudn't prove heavier apps are always rated better
```

In [114]:

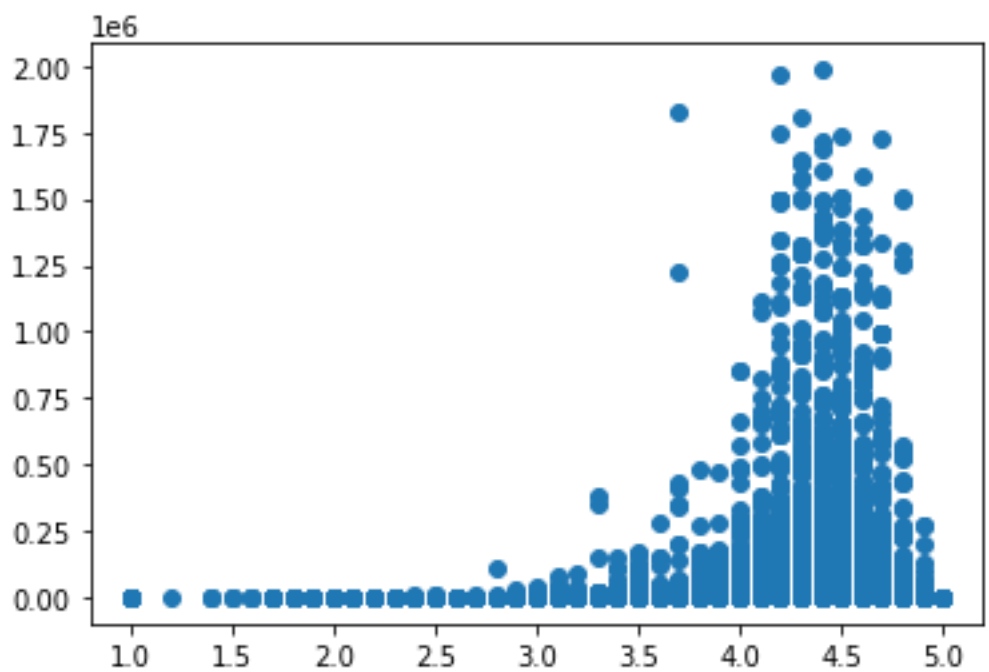
```
#Scatter plot for Rating vs Reviews
```

In [115]:

```
plt.scatter(df['Rating'],df['Reviews'])
```

Out[115]:

```
<matplotlib.collections.PathCollection at 0x7f7e72cab8d0>
```



In [116]:

```
#No, it doesn't proves more review means a better rating always.
```

In [117]:

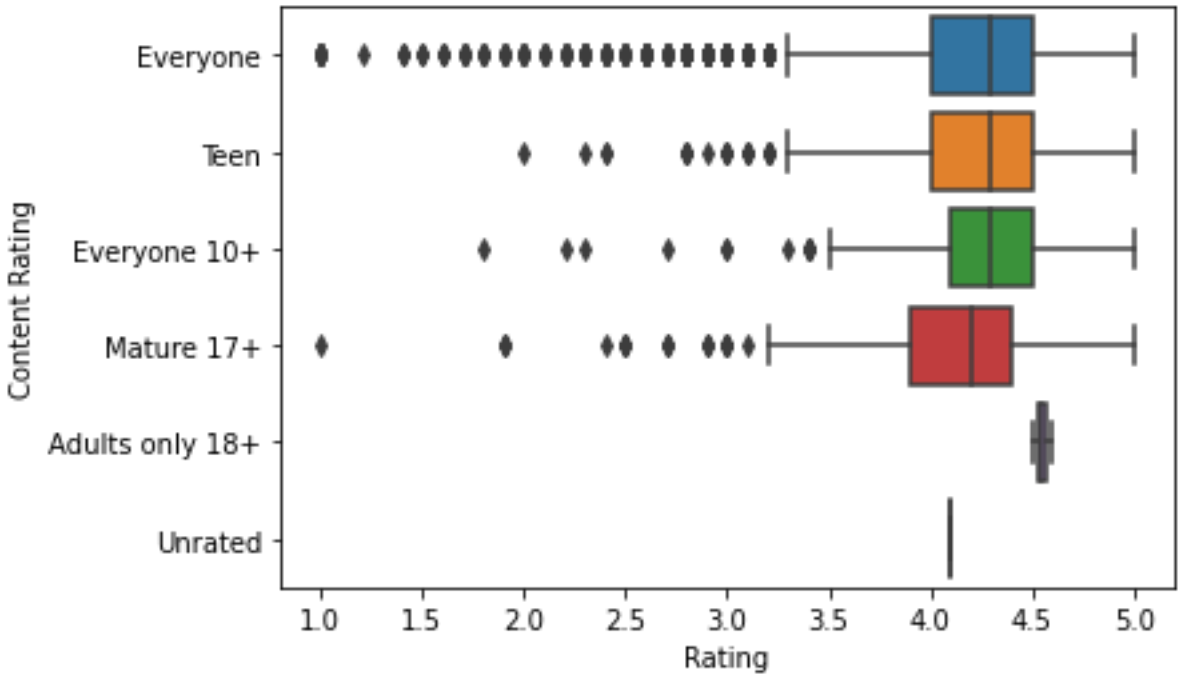
```
#boxplot for Rating vs Content rating
```

In [118]:

```
sns.boxplot(df['Rating'],df['Content Rating'])
```

Out[118]:

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



In [119]:

```
#Everyone type of apps are liked better by the users.
```

In [120]:

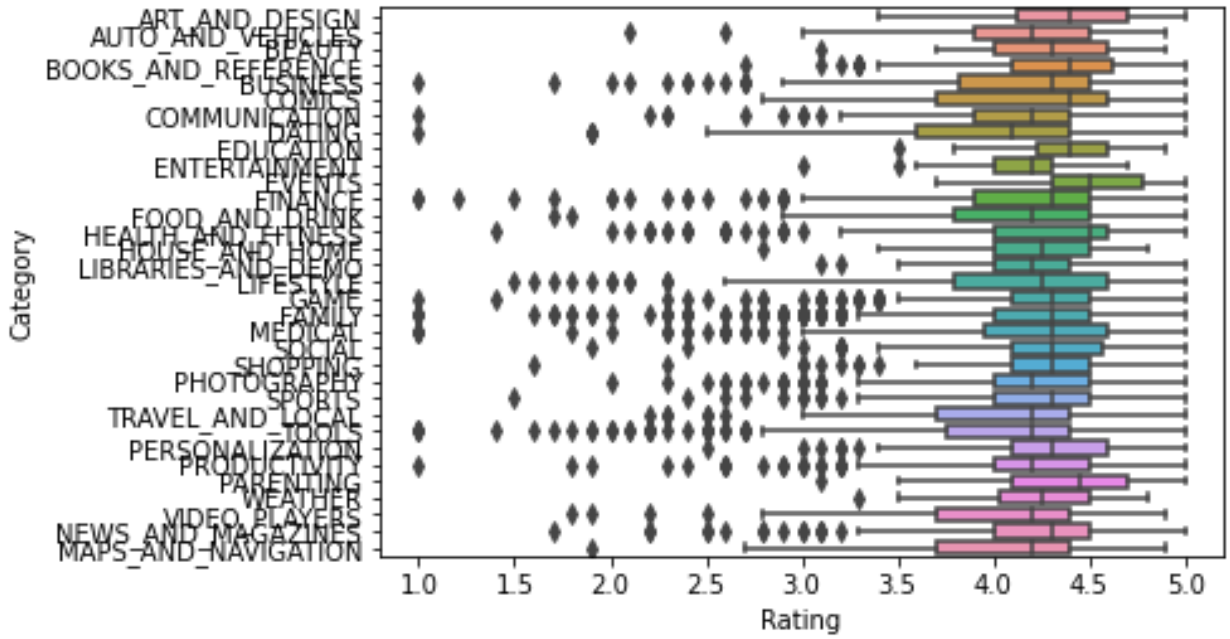
```
#boxplot for Ratings vs Category
```

In [121]:

```
sns.boxplot(df['Rating'],df['Category'])
```

Out[121]:

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



In [122]:

```
df['Installs'].describe()
```

Out[122]:

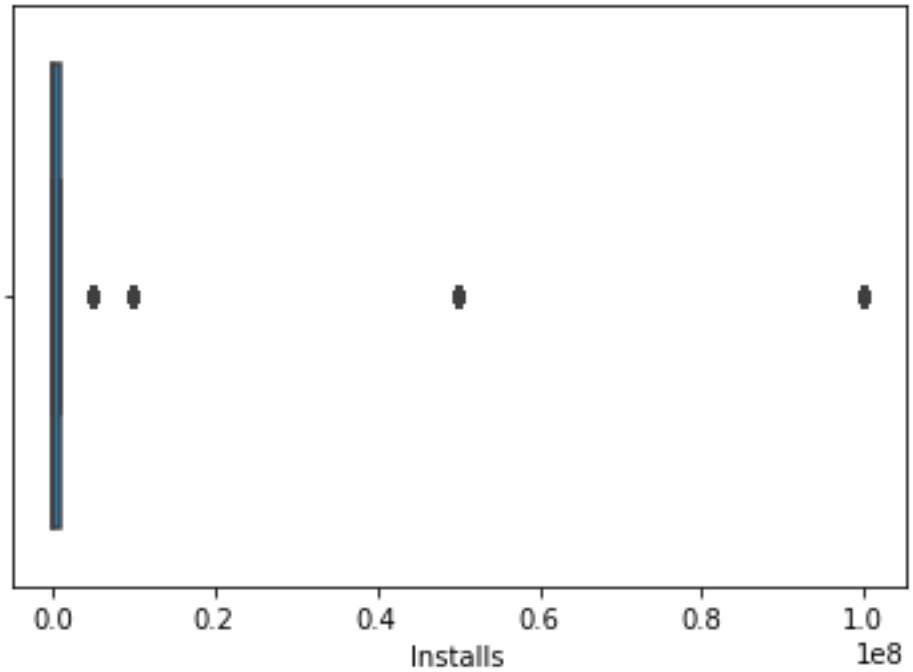
```
count    7.475000e+03
mean     3.149014e+06
std      1.055387e+07
min      5.000000e+00
25%      1.000000e+04
50%      1.000000e+05
75%      1.000000e+06
max      1.000000e+08
Name: Installs, dtype: float64
```

In [123]:

```
sns.boxplot(df['Installs'])
```

Out[123]:

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

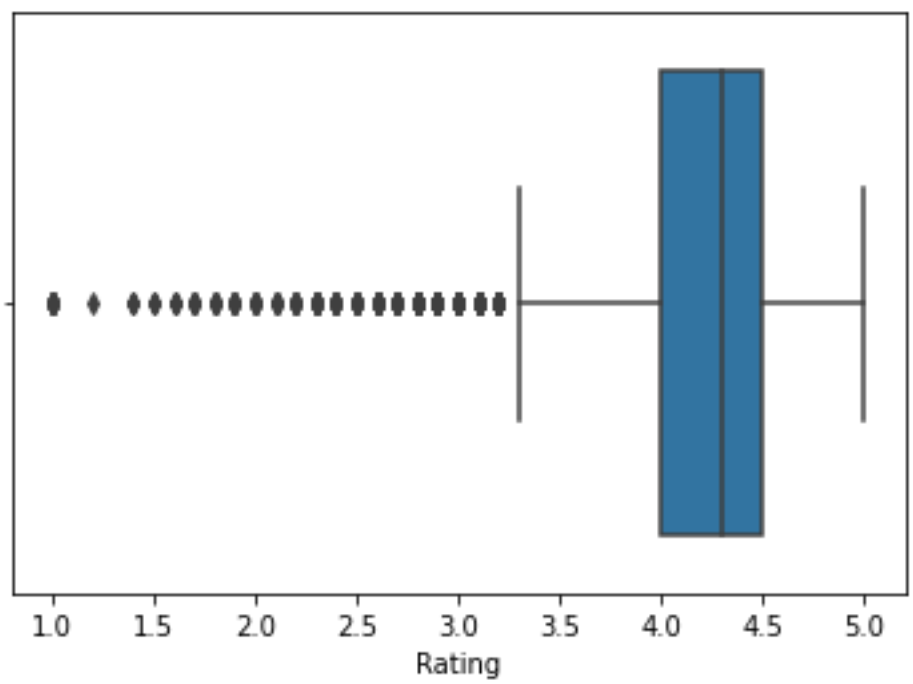


In [124]:

```
sns.boxplot(df['Rating'])
```

Out[124]:

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



In [125]:

```
df.index
```

Out[125]:

```
Int64Index([     0,      1,      2,      3,      4,      5,      6,      7,      8,
```

```
                9,
            ...
            10827, 10828, 10829, 10830, 10832, 10833, 10834, 10836, 10837,
            10840],
           dtype='int64', length=7475)
```

In [126]:

```
#8.Data preprocessing
```

In [127]:

```
inp1=df.copy()
```

In [128]:

```
inp1.shape
```

Out[128]:

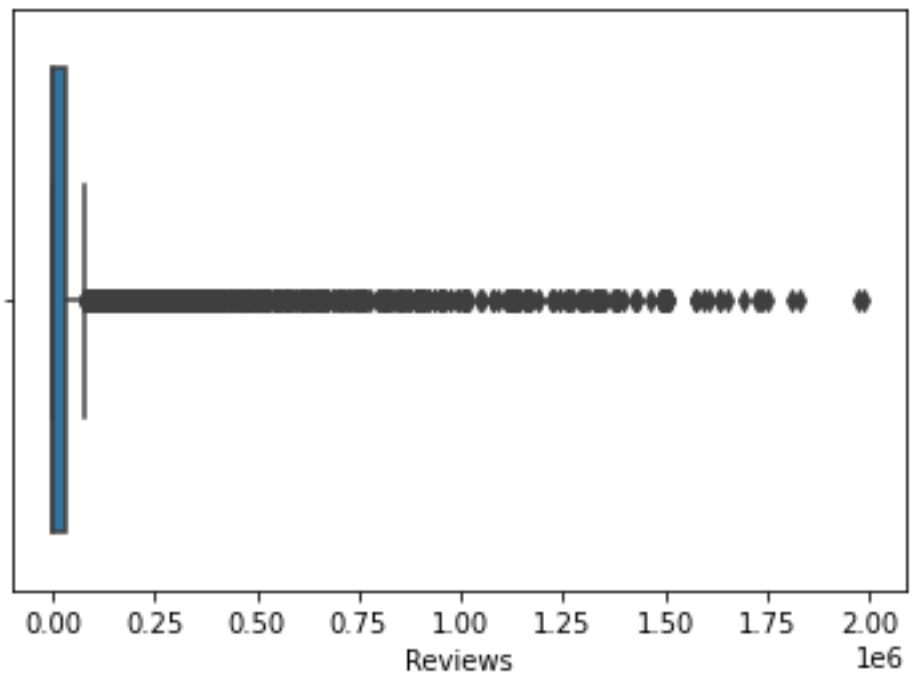
```
(7475, 13)
```

In [129]:

```
sns.boxplot(inp1.Reviews)
```

Out[129]:

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



In [130]:

```
inp1.Reviews.describe()
```

Out[130]:

```
count    7.475000e+03
mean     7.140332e+04
std      2.085558e+05
min      1.000000e+00
25%      9.850000e+01
```

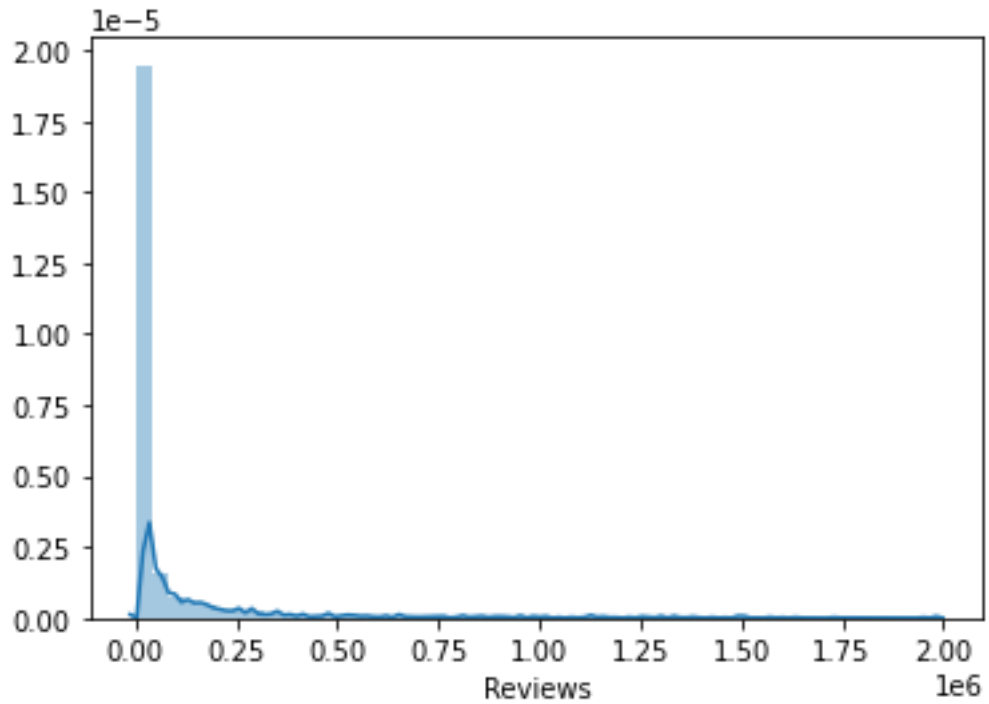
```
50%       2.014000e+03
75%       3.229900e+04
max       1.986068e+06
Name: Reviews, dtype: float64
```

In [131]:

```
sns.distplot(inp1.Reviews)
```

Out[131]:

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



In [132]:

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

In [133]:

```
inp1.shape
```

Out[133]:

```
(7475, 13)
```

In [134]:

```
inp1.Reviews.describe()
```

Out[134]:

```
count    7475.000000
mean        7.489245
std         3.482734
min         0.693147
25%         4.600145
50%         7.608374
75%        10.382822
max        14.501668
```

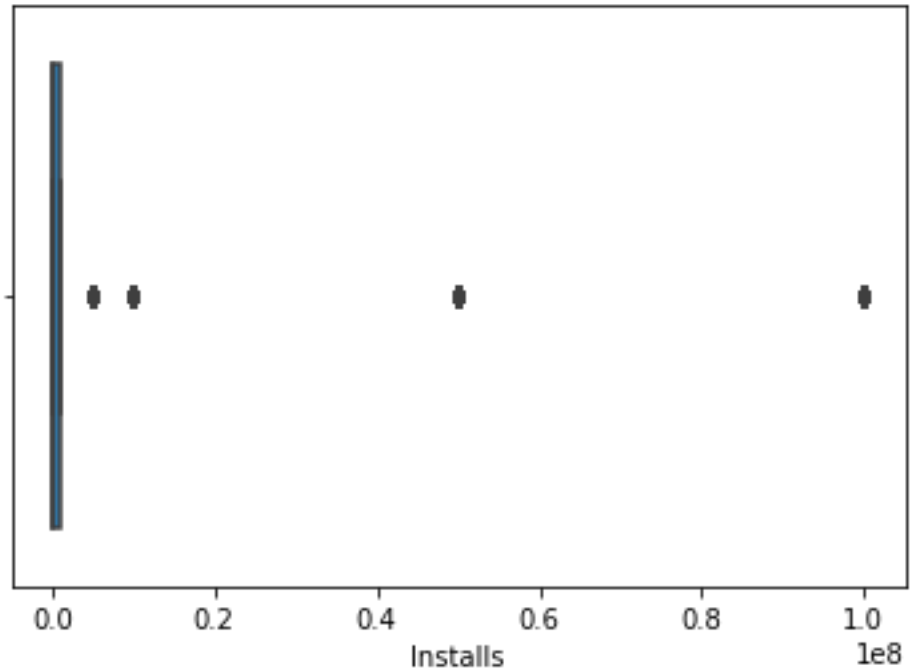
```
Name: Reviews, dtype: float64
```

In [135]:

```
sns.boxplot(df.Installs)
```

Out[135]:

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



In [136]:

```
inp1.Installs.describe()
```

Out[136]:

```
count    7.475000e+03
mean     3.149014e+06
std      1.055387e+07
min      5.000000e+00
25%      1.000000e+04
50%      1.000000e+05
75%      1.000000e+06
max      1.000000e+08
Name: Installs, dtype: float64
```

In [137]:

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

In [138]:

```
inp1.Installs.describe()
```

Out[138]:

```
count    7475.000000
mean       11.457308
std         3.536551
```

```
min         1.791759
25%         9.210440
50%        11.512935
75%        13.815512
max        18.420681
Name: Installs, dtype: float64
```

In [139]:

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

In [140]:

```
inp1.columns
```

Out[140]:

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

In [141]:

```
inp1.head()
```

Out[141]:

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & Scrap Book	ART_AND_DESIGN	4.1	5.075174	19000	9.210440	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	6.875232	14000	13.122365	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	11.379520	8700	15.424949	Free	0.0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up

	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	12.281389	25000	17.727534	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	6.875232	2800	11.512935	Free	0.0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

In [142]:

```
inp1=inp1.drop(['App','Last Updated','Current Ver','Android Ver'],axis=1)
```

In [143]:

```
inp1.columns
```

Out[143]:

```
Index(['Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type', 'Price',
       'Content Rating', 'Genres'],
      dtype='object')
```

In [144]:

```
#To get dummies values for Category,Genres and Content Rating
```

In [145]:

```
inp1_dummies=pd.get_dummies(inp1[['Category','Genres','Content Rating']])
```

In [146]:

```
inp1_dummies.head()
```

Out[146]:

	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	...	Genres_Video Players & Editors; Creativity	Genres_Video Players & Editors; Music & Video	Genres_Weather	Genres_Word	Content Rating_Adults only 18+	Content Rating_Everyone	Content Rating_Everyone 10+	Content Rating_Mature 17+	Content Rating_Teen	Content Rating_Unrated
0	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
4	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0

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	. . .	Genres_Video Players & Editors; Creativity	Genres_Video Players & Editors; Music & Video	Genres_Weather	Genres_Word	Content Rating_Adults only 18+	Content Rating_Everyone	Content Rating_Everyone 10+	Content Rating_Mature 17+	Content Rating_Teen	Content Rating_Unrated
										.										

5 rows × 151 columns

In [147]:

```
inp1.head()
```

Out[147]:

	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
0	ART_AND_DESIGN	4.1	5.075174	19000	9.210440	Free	0.0	Everyone	Art & Design
1	ART_AND_DESIGN	3.9	6.875232	14000	13.122365	Free	0.0	Everyone	Art & Design;Pretend Play
2	ART_AND_DESIGN	4.7	11.379520	8700	15.424949	Free	0.0	Everyone	Art & Design

	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres
3	ART_AND_DESIGN	4.5	12.281389	25000	17.727534	Free	0.0	Teen	Art & Design
4	ART_AND_DESIGN	4.3	6.875232	2800	11.512935	Free	0.0	Everyone	Art & Design;Creativity

In [148]:

```
inp1_num=inp1[['Rating','Reviews','Size','Installs','Price','Type']]
```

In [149]:

```
inp1_num.head()
```

Out[149]:

	Rating	Reviews	Size	Installs	Price	Type
0	4.1	5.075174	19000	9.210440	0.0	Free
1	3.9	6.875232	14000	13.122365	0.0	Free
2	4.7	11.379520	8700	15.424949	0.0	Free
3	4.5	12.281389	25000	17.727534	0.0	Free
4	4.3	6.875232	2800	11.512935	0.0	Free

In [150]:

```
inp2=pd.concat([inp1_num,inp1_dummies],axis=1)
```

In [151]:

```
inp2.shape
```

Out[151]:

```
(7475, 157)
```

In [152]:

```
inp2.head()
```

Out[152]:

	Rating	Reviews	Size	Installs	Price	Type	Category_ART_AND_DESIGN	Category_AUTO_AND_VEHICLES	Category_BEAUTY	Category_BOOKS_AND_REFERENCE	. . .	Genres_Video Players & Editors; Creativity	Genres_Video Players & Editors; Music & Video	Genres_Weather	Genres_Word	Content Rating_Adults only 18+	Content Rating_Everyone	Content Rating_Everyone 10+	Content Rating_Mature 17+	Content Rating_Teen	Content Rating_Unrated
0	4.1	5.075174	19000	9.2104440	0.0	Free	1	0	0	0	. . .	0	0	0	0	0	1	0	0	0	0
1	3.9	6.875232	14000	13.122365	0.0	Free	1	0	0	0	. . .	0	0	0	0	0	1	0	0	0	0
2	4.7	11.379520	8700	15.424949	0.0	Free	1	0	0	0	. . .	0	0	0	0	0	1	0	0	0	0
3	4	1	2	1	0	F	1	0	0	0	.	0	0	0	0	0	0	0	0	1	0

	Rating	Reviews	Size	Installs	Price	Type	Category_ART_AND_DESIGN	Category_AUTO_AND_VEHICLES	Category_BEAUTY	Category_BOOKS_AND_REFERENCE	...	Genres_VideoPlayers&Editors;Creativity	Genres_VideoPlayers&Editors;Music&Video	Genres_Weather	Genres_Word	Content Rating_Adults only 18+	Content Rating_Everyone	Content Rating_Everyone 10+	Content Rating_Mature 17+	Content Rating_Teen	Content Rating_Unrated
	.5	2.281389	5000	7.727534	.0	free					...										
4	4.3	6.875232	2800	11.512935	0.0	Free	1	0	0	0	...	0	0	0	0	0	1	0	0	0	0

5 rows × 157 columns

In [153]:

```
#To perform Linear Regression
```

In [154]:

```
X = inp2[['Reviews', 'Size', 'Installs', 'Price']]
y = inp2[['Rating']]
```

In [155]:

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

In [156]:

```
X_train, X_test, y_train, y_test = train_test_split(X,y, train_size = 0.7, rand
```

```
om_state = 120)
```

In [157]:

```
print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(5232, 4)
(2243, 4)
(5232, 1)
(2243, 1)
```

In [158]:

```
X_train.head()
```

Out[158]:

	Reviews	Size	Installs	Price
6096	10.271355	6300	15.424949	0.00
4172	5.545177	1600	8.517393	14.99
1499	8.799662	5500	13.815512	0.00
2871	10.062242	30000	13.815512	0.00
6348	6.877296	3600	9.210440	0.00

In [159]:

```
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
lm.fit(X_train, y_train)
```

Out[159]:

```
LinearRegression()
```

In [160]:

```
print(lm.coef_)
```

```
[[ 1.64124303e-01 -1.91747302e-07 -1.43029432e-01 -4.50065175e-03]]
```

In [161]:

```
print(lm.intercept_)
```

```
[4.58384813]
```

In [162]:

```
prediction = lm.predict(X_test)
print(prediction)
```

```
[[4.20366177]
 [4.21542551]
 [3.96471795]
 ...
 [4.51791915]
 [3.96765691]
 [4.08091251]]
```

In [163]:

```
from sklearn.metrics import mean_squared_error, r2_score
```

In [164]:

```
mse = mean_squared_error(y_test,prediction)
print(mse)
```

```
0.2863659787311721
```

In [165]:

```
r_square = r2_score(y_test,prediction)
print(r_square)
```

```
0.10307459340283753
```

In [166]:

```
print('R square value is:',r_square)
print('Mean squared value is:',mse)
```

```
R square value is: 0.10307459340283753
Mean squared value is: 0.2863659787311721
```

In [167]:

```
import pandas as pd
```

In [168]:

```
inp2_new = pd.DataFrame({'Actual':[y_test],'Predicted':[prediction]})
inp2_new
```

Out[168]:

	Actual	Predicted
0	Rating 8664 4.4 10155 4.2 4406...	[[4.203661771464269], [4.215425511384417], [3....

In [169]:

```
y
```

Out[169]:

	Rating
0	4.1
1	3.9

	Rating
2	4.7
3	4.5
4	4.3
...	...
10833	4.8
10834	4.0
10836	4.5
10837	5.0
10840	4.5

7475 rows × 1 columns

In [170]:

```
prediction
```

Out[170]:

```
array([[4.20366177],
       [4.21542551],
       [3.96471795],
       ...,
       [4.51791915],
       [3.96765691],
       [4.08091251]])
```

In [ ]:

Dataset contained many null values so our first step was to remove those rows with null values in it. Also the dataset contained many information that are irrelevant in predicting the rating of app. Thus, second step is to remove those unnecessary & unrelated column.

```
#remove missing values from dataset
dataset.dropna(inplace = True)

#dropping of unrelated and unnecessary columns from the dataset
dataset.drop(labels = ['Last Updated','Current Ver','Android Ver','App','Genres'], axis = 1, inplace = True)
```

Categories column contains more than one value & also it was in the string format. For any machine learning model to be applied, dataset must be in the format of integers. So the next step is to clean the Category column.

```
# Cleaning Categories into integers
CategoryString = dataset["Category"]
categoryVal = dataset["Category"].unique()
categoryValCount = len(categoryVal)
category_dict = {}
for i in range(0,categoryValCount):
    category_dict[categoryVal[i]] = i
dataset["Category_c"] = dataset["Category"].map(category_dict).astype(int)

#cleaning size of installation
```

Next step is to clean the size of installation column as it contains the values in KB & MB, so converting the all the values in bytes will be much easier for the dataset to be processed under any machine learning model.

```
#cleaning size of installation
def change_size(size):
    if 'M' in size:
        x = size[:-1]
        x = float(x)*1000000
        return(x)
    elif 'k' == size[-1:]:
        x = size[:-1]
        x = float(x)*1000
        return(x)
    else:
        return None

dataset["Size"] = dataset["Size"].map(change_size)
#filling Size which had NA
dataset.Size.fillna(method = 'ffill', inplace = True)
```

Number of installs column contains the value in the form of x+ which denotes no of installs of said app is greater than x. Thus cleaning of installs column as-

```
#Cleaning no of installs column
dataset['Installs'] = [int(i[:-1].replace(',','')) for i in dataset['Installs']]
```

Type of app denotes whether the app is free or paid which is binary values so, converting those values in 0 &1.

```
#Converting Type column into binary column
def type_cat(types):
    if types == 'Free':
        return 0
    else:
        return 1

dataset['Type'] = dataset['Type'].map(type_cat)
```

Also, Price column depicts the price of application on the play store, whose value is either 0 for free app & some amount in dollars for paid apps. Converting those amounts in float values as-

```
#Cleaning prices
def price_clean(price):
    if price == '0':
        return 0
    else:
        price = price[1:]
        price = float(price)
        return price

dataset['Price'] = dataset['Price'].map(price_clean).astype(float)
```

Now, the first step for any machine learning algorithm completes. Dataset is preprocessed & is ready to be applied on regression model.

TRAINING & TESTING OF MODEL

For training of our selected models- Linear Regression,SVM, Random Forest Regression models.

LINEAR REGRESSION

```
"""Step 2 - Training & Testing of the model
Model 1 : Linear Regression"""

#splitting the dataset into training & test set

X = dataset.iloc[:, 1:].values
y = dataset.iloc[:, 0].values

from sklearn.model_selection import train_test_split
X_train , X_test , y_train , y_test = train_test_split(X , y , test_size = 0.3 , random_state = 0)

#fitting simple linear regression into training set

from sklearn.linear_model import LinearRegression
linear_regressor = LinearRegression()
linear_regressor.fit(X_train , y_train)

#predicting the test set results
y_pred = linear_regressor.predict(X_test)
```

SVM REGRESSION (With Feature Scaling)

Firstly, Feature scaling is applied on the dataset in order to scale the data in all columns. Some machine learning algorithms uses Euclidean Distance between the features for training purpose. With wide range in the values of column, sometimes result gets purely dependent on the column with larger values.

```
"""Step 2 - Training & Testing of the model
Model 2 : SVM Regression(with feature scaling)"""

X = dataset.iloc[:, 1:].values
y = dataset.iloc[:, 0].values

#feature scaling
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
sc_y = StandardScaler()
X = sc_X.fit_transform(X)
y = y.reshape(-1, 1)
y = sc_y.fit_transform(y)

from sklearn.model_selection import train_test_split
X_train , X_test , y_train , y_test = train_test_split(X , y , test_size = 0.3 , random_state = 0)

#fitting the SVR model to the dataset
from sklearn.svm import SVR
regressor = SVR(kernel = 'rbf')   #chhosing the gaussian kernel..ie rbf kernel
regressor.fit(X_train , y_train)

y_pred_svm = regressor.predict(X_test)
```

SVM REGRESSION (Without Feature Scaling)

```python

"""Step 2 - Training & Testing of the model
Model 2 : SVM Regression(without feature scaling)"""

X = dataset.iloc[:, 1:].values
y = dataset.iloc[:, 0].values

from sklearn.model_selection import train_test_split
X_train , X_test , y_train , y_test = train_test_split(X , y , test_size = 0.3 , random_state = 0)

#fitting the SVR model to the dataset
from sklearn.svm import SVR
regressor = SVR(kernel = 'rbf')   #chhosing the gaussian kernel..ie rbf kernel
regressor.fit(X_train , y_train)

y_pred_svm = regressor.predict(X_test)
```

RANDOM FOREST REGRESSION

```python

"""Step 2 - Training & Testing of the model
Model 2 : Random forest Regression"""

X = dataset.iloc[:, 1:].values
y = dataset.iloc[:, 0].values

from sklearn.model_selection import train_test_split
X_train , X_test , y_train , y_test = train_test_split(X , y , test_size = 0.3 , random_state = 0)

#fitting the Random forest regression model to the dataset
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators = 100 , random_state = 0)   #n_estimators is no of trees in forest
regressor.fit(X_train , y_train)

#predicting the new result with polynomial regression
y_pred = regressor.predict(X_test)

#random forest regression with 300 trees
from sklearn.ensemble import RandomForestRegressor
regressor_2 = RandomForestRegressor(n_estimators = 300 , random_state = 0)   #n_estimators is no of trees in fore
regressor_2.fit(X_train , y_train)

#predicting the new result with polynomial regression
y_pred_2 = regressor_2.predict(X_test)
```

RESULTS AND DISCUSSIONS

LINEAR REGRESSION

```
def Evaluationmatrix(y_test, y_pred):
    print ('Mean Squared Error: '+ str(metrics.mean_squared_error(y_test,y_pred)))
    print ('Mean absolute Error: '+ str(metrics.mean_absolute_error(y_test,y_pred)))

Evaluationmatrix(y_test,y_pred)
```

```
IPython console
Preprocessing/A   Linear_R/A   SVM_R1/A   SVM_R2/

   ...: linear_regressor.fit(X_train , y_train)
Out[3]: LinearRegression(copy_X=True, fit_intercept=

In [4]: y_pred = linear_regressor.predict(X_test)

In [5]: def Evaluationmatrix(y_test, y_pred):
   ...:     print ('Mean Squared Error: '+
str(metrics.mean_squared_error(y_test,y_pred)))
   ...:     print ('Mean absolute Error: '+
str(metrics.mean_absolute_error(y_test,y_pred)))

In [6]: Evaluationmatrix(y_test,y_pred)
Mean Squared Error: 0.26077708107509806
Mean absolute Error: 0.3528789627513037
```

After the Linear regression model is trained, it is then tested on test set. Predicted results & actual results are compared using the evaluation parameter for regression model such as mean squared error & mean absolute error. As shown in the picture,

Mean squared error - 0.2607

Mean absolute error - 0.3528

SVM REGRESSION (With Feature Scaling)

```
def Evaluationmatrix(y_test, y_pred):
    print ('Mean Squared Error: '+ str(metrics.mean_squared_error(y_test,y_pred)))
    print ('Mean absolute Error: '+ str(metrics.mean_absolute_error(y_test,y_pred)))

Evaluationmatrix(y_test,y_pred_svm)
```

```
IPython console
Preprocessing/A   Linear_R/A   SVM_R1/A   SVM_R2/A
In [6]: y_pred_svm = regressor.predict(X_test)

In [7]: def Evaluationmatrix(y_test, y_pred):
   ...:     print ('Mean Squared Error: '+
str(metrics.mean_squared_error(y_test,y_pred)))
   ...:     print ('Mean absolute Error: '+
str(metrics.mean_absolute_error(y_test,y_pred)))
   ...:
   ...:
   ...: Evaluationmatrix(y_test,y_pred_svm)
Mean Squared Error: 1.0190424013531676
Mean absolute Error: 0.654663949977556
```

After the SVM model is trained, it is then tested on test set. Predicted results & actual results are compared using the evaluation parameter for regression model such as mean squared error & mean absolute error. In SVM model here, firstly feature scaling is used for scaling the all the values in columns. As shown in the picture,

Mean squared error - 1.0190

Mean absolute error - 0.6546

SVM REGRESSION (Without Feature Scaling)

```
def Evaluationmatrix(y_test, y_pred):
    print ('Mean Squared Error: '+ str(metrics.mean_squared_error(y_test,y_pred)))
    print ('Mean absolute Error: '+ str(metrics.mean_absolute_error(y_test,y_pred)))

Evaluationmatrix(y_test,y_pred_svm)
```

```
IPython console
Preprocessing/A   Linear_R/A   SVM_R1/A   SVM_R2/A

In [5]: y_pred_svm = regressor.predict(X_test)

In [6]: def Evaluationmatrix(y_test, y_pred):
   ...:     print ('Mean Squared Error: '+
str(metrics.mean_squared_error(y_test,y_pred)))
   ...:     print ('Mean absolute Error: '+
str(metrics.mean_absolute_error(y_test,y_pred)))
   ...:
   ...:
   ...: Evaluationmatrix(y_test,y_pred_svm)
Mean Squared Error: 0.26232867994406256
Mean absolute Error: 0.34716344856088266
```

After the SVM model is trained, it is then tested on test set. Predicted results & actual results are compared using the evaluation parameter for regression model such as mean squared error & mean absolute error. In SVM model here, feature scaling is not used. As shown in the picture,
Mean squared error - 0.2623
Mean absolute error - 0.3471

RANDOM FOREST REGRESSION

```
def Evaluationmatrix(y_test, y_pred):
    print ('Mean Squared Error: '+ str(metrics.mean_squared_error(y_test,y_pred)))
    print ('Mean absolute Error: '+ str(metrics.mean_absolute_error(y_test,y_pred)))

Evaluationmatrix(y_test,y_pred)
Evaluationmatrix(y_test,y_pred_2)
```

```
IPython console
Preprocessing/A   Linear_R/A   SVM_R1/A   SVM_R2/A

In [6]: def Evaluationmatrix(y_test, y_pred):
   ...:     print ('Mean Squared Error: '+
str(metrics.mean_squared_error(y_test,y_pred)))
   ...:     print ('Mean absolute Error: '+
str(metrics.mean_absolute_error(y_test,y_pred)))
   ...:
   ...:
   ...: Evaluationmatrix(y_test,y_pred)
Mean Squared Error: 0.23598689141030588
Mean absolute Error: 0.3129456255935423
```

After the Random forest model is trained, it is then tested on test set. Predicted results & actual results are compared using the evaluation parameter for regression model such as mean squared error & mean absolute error. Regression model shown here is trained with 100 trees. As shown in the picture,

Mean squared error - 0.23598

Mean absolute error - 0.31294

```
Preprocessing/A   Linear_R/A   SVM_R1/A   SVM_R2/

           min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, n_estimator
           oob_score=False, random_state=0, verbose=

In [8]: y_pred_2 = regressor_2.predict(X_test)

In [9]: Evaluationmatrix(y_test,y_pred_2)
Mean Squared Error: 0.23509671674550342
Mean absolute Error: 0.31283733578664136

In [10]:
```

After the Random forest model is trained, it is then tested on test set. Predicted results & actual results are compared using the evaluation parameter for regression model such as mean squared error & mean absolute error. Regression model shown here is trained with 300 trees. As shown in the picture,

Mean squared error - 0.23509

Mean absolute error - 0.3128

CONCLUSIONS

Results produced by all the models are-

LINEAR REGRESSION

Mean squared error - 0.2607
Mean absolute error - 0.3528

SVM REGRESSION (With Feature Scaling)

Mean squared error - 1.0190
Mean absolute error - 0.6546

SVM REGRESSION (Without Feature Scaling)

Mean squared error - 0.2623
Mean absolute error - 0.3471

RANDOM FOREST REGRESSION

Trees = 100

Mean squared error - 0.23598
Mean absolute error - 0.31294

Trees = 300

Mean squared error - 0.23509
Mean absolute error - 0.3128

As observed from the results, Random forest regression model produces better results in comparison of other models. By increasing the no of trees(n_estimators parameter) in random forest regression, increases the performance and makes the predictions more stable. One of the big problems in machine learning is overfitting, but most of the time this won't happen that easy to a random forest model. That's because if there are enough trees in the forest, the regressor won't overfit the model. The main limitation of Random Forest is that a large number of treescan make the algorithm to slow and ineffective for real-time predictions. In general, these algorithms are fast to train, but quite slow to create predictions once they are trained. A more accurate prediction requires more trees, which results in a slower model. In most real-world applications the random forest algorithm is fast enough, but there can certainly be situations where run-time performance is important and other approaches would be preferred.

REFERENCES

1. https://www.kaggle.com/lava18/google-play-store-apps
2. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
3. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
5. https://scikit-learn.org/stable/modules/preprocessing.html
6. https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd